U.PORTO

2º CICLO DE ESTUDOS
INFORMÁTICA MÉDICA

Predicting Healthy Eating in the Portuguese Population with Bayesian Networks

Maria Carolina Rafael Pombinho

M

2026

FACULDADE DE **MEDICINA**

Investigation Team

Orientador: Pedro Pereira Rodrigues

Coorientador: Andreia Sofia da Costa Teixeira e Maria Luísa de Morais e Sousa de Castro Guedes

Discente: Maria Carolina Rafael Pombinho

Acknowledgements

The completion of this dissertation would not have been possible without the support and contribution of several people, to whom I wish to express my sincere gratitude.

First and foremost, I would like to thank my supervisor, Professor Pedro Pereira Rodrigues, and my co-supervisor, Professor Andreia Teixeira, for their guidance, availability, and critical insight throughout every stage of this work. Their scientific rigour and constructive feedback were essential in shaping both the methodological approach and the final form of this dissertation.

I am also deeply grateful to Professor Maria Luísa Castro for her clinical guidance and valuable contributions, which were instrumental in grounding the interpretation of the results in their clinical and public health context.

My thanks extend to the Faculty of Medicine of the University of Porto and to the research team responsible for the 2019 study, whose data made this analysis possible, and to the researchers who had previously worked with this dataset for generously sharing their feedback and experience.

To my parents, Sónia Rafael and Luís Pombinho, for their unconditional support and for always believing in me.

To my life partner, Bruno Silva, for the patience, encouragement, and steady presence throughout this journey.

To my friends and colleagues, for the shared moments and the constant motivation along the way.

To all of you, my heartfelt thank you.

Abstract

Dietary behaviour is a key modifiable determinant of chronic non-communicable diseases, which account for approximately 74% of all deaths globally and represent a major public health burden in Portugal. Despite this, the conditional dependency structure among the determinants of healthy eating in the Portuguese adult population remains incompletely characterised.

This study developed a Bayesian Network model to identify predictors of healthy eating and estimate individual posterior probabilities of adopting a healthy dietary pattern. Data were collected in 2019 through a face-to-face questionnaire administered in relation to another study (n = 891 adults). Healthy eating was operationalised as a binary composite outcome requiring simultaneous daily consumption of fruit, vegetables and legumes within recommended portions, regular meal structure, and non-excessive alcohol intake (prevalence ≈ 14%). Structure learning used the Hill-Climbing algorithm with BDeu scoring and bootstrap resampling (R = 200); performance was evaluated through stratified 5-fold cross-validation with F1-optimised thresholds, using PR-AUC and Brier Skill Score as primary metrics.

Smoking status and region of residence emerged as the only direct parents of healthy eating in the consensus graph, stable across all sensitivity analyses. Body mass index, age, and educational level showed no direct arc to the target. *What-if* analysis revealed a two-order-of-magnitude probability gradient, from near-zero in active smokers in the Algarve (0.003) to 0.334 in ex-smokers in the Centro region. The model achieved a PR-AUC of 0.241 — 68% above the prevalence baseline — a Brier Skill Score of 0.040, and an Expected Calibration Error of 0.030.

Both baselines achieved marginally higher PR-AUC (Logistic Regression 0.293; Naïve Bayes 0.304), a difference that does not detract from the network's distinctive contribution: its interpretable dependency structure and *what-if* inference.

Bayesian Networks provide an interpretable and calibrated framework for modelling dietary determinants, surfacing structural insights not recoverable from standard regression approaches. Findings support targeted nutritional surveillance by regions and the integration of dietary assessment into smoking cessation programs.

Keywords: **Dietary Behavior; Risk Factors; Bayesian Networks; Healthy Eating**

Resumo

O comportamento alimentar é um determinante modificável fundamental das doenças crónicas não transmissíveis, que representam aproximadamente 74% de todas as mortes a nível mundial e constituem um importante problema de saúde pública em Portugal. Apesar disso, a estrutura de dependência condicional entre os determinantes da alimentação saudável na população adulta portuguesa permanece incompletamente caraterizada.

Este estudo desenvolveu um modelo de rede bayesiana para identificar preditores da alimentação saudável e estimar probabilidades posteriores individuais de adoção de um padrão alimentar saudável. Os dados foram recolhidos em 2019 através de um questionário administrado presencialmente por um estudo anterior (n = 891 adultos). A alimentação saudável foi operacionalizada como um *outcome* binário composto, exigindo o cumprimento simultâneo de consumo diário de fruta, vegetais e leguminosas dentro das porções recomendadas, estrutura regular de refeições e consumo não excessivo de álcool (prevalência ≈ 14%). A aprendizagem estrutural utilizou o algoritmo *Hill-Climbing* com pontuação *BDeu* e reamostragem *bootstrap* (R = 200); o desempenho foi avaliado através de validação cruzada estratificada em 5 *folds* com limiares otimizados por F1, utilizando a *PR-AUC* e o *Brier Skill Score* como métricas primárias.

O estatuto tabágico e a região de residência emergiram como os únicos pais diretos da alimentação saudável no grafo de consenso, estáveis em todas as análises de sensibilidade. O índice de massa corporal, a idade e o nível de escolaridade não apresentaram arco direto para o nó alvo, indicando efeitos. A análise *what-if* revelou um gradiente de probabilidade de duas ordens de grandeza, desde valores próximos de zero em fumadores ativos no Algarve (0,003) até 0,334 em ex-fumadores na região Centro. O modelo atingiu uma *PR-AUC* de 0,241, 68% acima da *baseline* de prevalência, um *Brier Skill Score* de 0,040 e um erro de calibração esperado de 0,030.

Ambos os modelos de referência obtiveram uma *PR-AUC* ligeiramente superior (Regressão Logística 0,293; Naïve Bayes 0,304), diferença que não retira o contributo distintivo da rede: a estrutura de dependências interpretável e a inferência *what-if*.

As redes bayesianas fornecem um enquadramento interpretável e calibrado para modelar os determinantes do comportamento alimentar, revelando *insights* estruturais não recuperáveis através de abordagens de regressão padrão. Os resultados apoiam a vigilância nutricional direcionada por região e a integração da avaliação dietética nos programas de cessação tabágica.

Palavras-chave: Comportamento alimentar; Fatores de risco; Redes Bayesianas; Alimentação saudável.

Table of Contents

Acknowledgements IV

Abstract V

Resumo VIII

Table of Contents IX

List of Tables XV

List of Figures XIX

List of abbreviations and acronyms XXI

1. Introduction 23

1.1. Context and Background 23

1.2. Structure of the Dissertation 24

2. Literature review 27

2.1. Healthy eating and dietary patterns 27

2.2. Determinants of healthy eating 27

2.3. Healthy eating in Portugal 31

2.4. Bayesian networks in health research and nutritional epidemiology 32

2.4.1. What are Bayesian networks 32

2.4.2. Bayesian networks and their applicability in nutritional epidemiology 33

2.5. Summary: gaps and contribution of this thesis 34

3. Methods 36

3.1. Study design and analytical workflow 36

3.2. Data Analysis 37

3.3. Data source and participants 38

3.4. Data collection instrument 38

3.5. Definition of variables and data preparation 39

3.5.1. Full variables 39

3.5.2. Reduced Variables 46

3.6. Bayesian network structure learning 48

3.7. Parameter learning 50

3.7.1. Treatment of missing data 51

3.7.2. Estimation procedure 52
3.8. Classification threshold optimisation 52
3.9. Validation and comparison with baseline models 54
3.9.1. Handling of fold-specific structural failures 54
3.9.2. Primary and secondary performance metrics 54
3.9.3. Calibration assessment 54
3.9.4. Baseline comparators 55
3.10. Sensitivity and robustness analyses 55
3.10.1. Sensitivity to arc strength threshold 56
3.10.2. Sensitivity to the variable set 56
3.10.3. Sensitivity to discretisation 56
3.10.4. Sensitivity to fold stratification seed 57
3.10.5. Age-Stratified models 57
3.10.6. Tree-Augmented Naïve Bayes (TAN) 58
3.10.7. Probabilistic sensitivity (*What-if* analysis) 58
3.11. Ethical considerations 58
4. Results 61
4.1. Sample description 61
4.2. Preliminary results and the threshold problem 62
4.2.1. Initial configurations (default threshold = 0.5) 62
4.2.2. The threshold problem 63
4.2.3. Threshold optimisation with K-Fold validation 63
4.3. Reduced variable set analysis 64
4.3.1. Structural analysis (bootstrap, arc strengths) 64
4.3.2. Predictive performance 66
4.4. Full variable set analysis 66
4.4.1. Conservative graph (Strength ≥ 0.7): model collapse 66
4.4.2. Permissive configurations (F≥0.4, F≥0.3, F≥0.2) 67
4.4.3. Arc strengths to the target node 68
4.4.4. Full *vs* reduced variable set 69
4.4.5. Classification threshold 70

4.4.6. Final model selection 71
4.5. Final model: structural description 72
4.5.1. Consensus DAG 72
4.5.2. Conditional probability tables 73
4.5.3. Target node: *alimentacao_saudavel_bin* 73
4.6. Final model: predictive performance 75
4.6.1. Discrimination 76
4.6.2. Calibration 76
4.6.3. Confusion matrix 77
4.6.4. Per-Fold results 77
4.7. *What-if* Analysis 79
4.7.1. Evaluated scenarios 81
4.8. Comparison with baseline classifiers 82
4.8.1. Discriminative performance 83
4.8.2. Probabilistic calibration 83
4.8.3. The value of the Bayesian Network beyond classification 84
4.9. Sensitivity and robustness analyses 85
4.9.1. Age-stratified Bayesian network models 85
4.9.2. Sensitivity to age stratification 87
4.9.3. Sensitivity to arc strength threshold 89
4.9.4. SMOTE 89
4.9.5. Threshold comparison: F1 *vs* MCC *vs* Youden 91
4.9.6. Weighted bootstrap structure learning 92
4.9.7. Alternative fold stratification seed 93
4.9.8. Tree-Augmented Naïve Bayes 94
5. Discussion 97
5.1. Main findings 97
5.2. Clinical interpretation and public health 99
5.2.1. Smoking status and health behaviour clustering 99
5.2.2. Geographic gradient and the structural nature of region as a determinant 100
5.3. Joint effect of smoking status and region of residence 102

5.3.1. Cautious interpretation: small cell counts and parameter smoothing 103
5.3.2. Robust patterns despite limitations 103
5.3.3. Combined effects: direction over magnitude 104
5.3.4. Public health implications 105
5.4. Variables disconnected from the target: statistical limitation or substantive finding? 105
5.5. Model performance in context 107
5.6. Class imbalance and methodological robustness 109
5.7. Comparison with previous studies 112
5.7.1. Sociodemographic determinants of healthy eating in the portuguese population 112
5.7.2. Smoking status and dietary quality 114
5.7.3. Bayesian network applications to dietary outcomes 115
5.7.4. Synthesis 115
5.8. Specific contributions of this study 116
5.8.1. Methodological contribution 116
5.8.2. Substantive contribution 117
5.8.3. Applied contribution 118
5.9. Strengths and limitations 118
5.9.1. Strengths 119
5.9.2. Limitations of the data source 119
5.9.3. Limitations associated with the operationalisation of the target variable 121
5.9.4. Limitations for clinical and public health application 122
5.10. Implications for practice and public health policy 123
5.10.1. Implications for population surveillance and risk stratification 124
5.10.2. Implications for primary care and clinical practice 124
5.10.3. Implications for structural and policy interventions 125
5.10.4. The limits of public health policy and the need for coordinated action 126
5.11. Future work 127
6. Conclusion 129
6.1. Summary of the study 129
6.2. Answer to the research question 129
6.3. Final considerations 130

References 133
Annexes 140
Annex A 141
Appendices 151
Appendix A - Distribution of Variables 152
Full variable set (n = 847) 153
Reduced variable set (n = 891) 156

List of Tables

Table 4.1 - Description of the recruited sample, prior to exclusions (n = 891). 61

Table 4.2 - Preliminary predictive performance across four configurations (classification threshold = 0.5) (n=891). 63

Table 4.3 - Predictive performance of the reduced variable set (16 variables) across three consensus arc strength thresholds. Values are mean ± SD across five valid folds (n = 891). PR-AUC baseline = prevalence = 0.139. 64

Table 4.4 - Arc strength from each variable to the target node (16 variables, R = 200). Highlighted: sole variable in the conservative graph (n=891). 65

Table 4.5- Predictive performance of the Bayesian Network across three consensus arc strength thresholds. Values are mean ± SD across valid folds. F ≥ 0.7 and F ≥ 0.4: five valid folds (n = 847 evaluated); F ≥ 0.3: four valid folds (n = 678 evaluated, Fold 2 excluded). PR-AUC baseline = prevalence = 0.143. 67

Table 4.6 - Arc strength from each variable to the target node (full variable set, R = 200). Only variables with strength > 0.05 (n=847). 69

Table 4.7- Comparison between the best exploratory configuration (all variables, F ≥ 0.3 n=678) and the best reduced model configuration (16 variables, F ≥ 0.4 n=891). 70

Table 4.8 - P(*alimentacao_saudavel_bin* = Yes | *tabagismo_cat3*, *nuts2_cat5*) (n = 847). 74

Table 4.9 - Confusion Matrix (n = 678). 77

Table 4.10 - Per-fold cross-validation results for the F ≥ 0.3 configuration. Fold 2 excluded due to bootstrap failure. Means computed over four valid folds (n = 678 evaluated, Fold 2 excluded). 77

Table 4.11 - *What-if* scenario analysis: posterior probabilities of healthy eating under single and combined evidence conditions. Δ = change relative to marginal probability (0.142). Values derived from exact inference on the compiled *gRain* junction tree (F ≥ 0.3) (n = 847). 80

Table 4.12 - Number of cases in each combination of the *what-if* class (n=847). 80

Table 4.13 - Predictive performance comparison across models (BN F ≥ 0.3, Logistic Regression L1, Naïve Bayes). All models evaluated on the same stratified **4**-fold cross-validation splits with F1-optimised threshold. Primary metrics (PR-AUC, Brier Skill) in bold n = 678, 4 valid folds). 83

Table 4.14 - Variables excluded per age stratum due to insufficient variability (n= 847). 86

Table 4.15- Predictive performance of age-stratified Bayesian Network models (cross-validation). 86

Table 4.16 - Network structure and direct predictors by age stratum. 88

Table 4.17 - SMOTENC sensitivity analysis: performance metrics by configuration and variable set. BSS: Brier Skill Score; ECE: Expected Calibration Error; FP: false positives. ‡ Balanced test set (≈50%) not directly comparable with original model (n=847). 90

Table 4.18 - Classification threshold and performance metrics by optimisation criterion (mean ± SD, 4 valid folds, 32-variable model). Specificity maximisation subject to recall ≥ 40% (n = 678, 4 valid folds). The thresholds reported here (F1 = MCC = Youden = 0.208 ± 0.091) derive from the threshold-criteria comparison and are not the operating threshold of the final model; the final model applies the F1-optimised threshold of 0.180 reported in Table 4.10. 91

Table 4.19 - Performance comparison: original model (seed 42, 4 valid folds n=678) vs weighted bootstrap (5 valid folds n= 847), 32-variable dataset. FP: false positives (individuals with inadequate diet classified as healthy). Bold values indicate superior performance. 92

Table 4.20 - Aggregated confusion matrix — weighted bootstrap (n = 847, 5 valid folds, F1-score threshold). Rows: real class. Columns: predicted class. 92

Table 4.21 - Comparison of random seeds tested for fold stratification. FP: false positives under F1-score threshold. Metrics for seeds 123 and 99 were not computed due to PDAG collapse in one fold (seed 42, n= 678; seed 44, n=847). 93

Table 4.22 - Aggregated confusion matrix — seed 44 (n = 847, 5 valid folds, F1-score threshold). Rows: real class. Columns: predicted class. 93

Table 4.23 - Performance comparison: original HC bootstrap model vs. TAN configurations. FP: false positives (individuals with inadequate diet classified as healthy). BSS: Brier Skill Score. 94

List of Figures

Figure 3.1 - Study workflow. 37
Figure 4.1- BN strength >= 0.7 (fold 5). 67
Figure 4.2- BN strength >= 0.2 (fold5). 68
Figure 4.3 - Per-threshold results. 70
Figure 4.4 - Consensus DAG of the Bayesian Network (arc strength threshold F ≥ 0.3, n = 847). Network complexity across thresholds. 73
Figure 4.5 - Reduced Variable Set Bootstrap Results. 74
Figure 4.6 - Full Variable Set Bootstrap Result. 75
Figure 4.7 - Evaluation panel for the BN model (F ≥ 0.3, Fold 5). 76
Figure 4.8 - Calibration assessment. 77
Figure 4.9 - Comparison between folds. 79
Figure 4.10 - *What-if* Analysis: Impact of interventions on P(Healthy Eating). 82
Figure 4.11 - Predictive performance: Bayesian Network *vs* baseline classifiers. 85
Figure 4.12- Comparative interpretation across strata. 87
Figure 5.1 - Smokers by region. 100
Figure 5.2 - Healthy eaters by region. 102
Figure 5.3 - Probability of healthy eating by region and smoking status. 103

List of abbreviations and acronyms

AHEI - Alternative Healthy Eating Index

AI - Artificial Intelligence

AM Lisboa - Área Metropolitana de Lisboa / Lisbon Metropolitan Area

AMED - Alternate Mediterranean Diet score

AUC-PR / PR-AUC - Area Under the Precision–Recall Curve

AUC-ROC / ROC-AUC - Area Under the Receiver Operating Characteristic Curve

BDeu - Bayesian Dirichlet equivalent uniform score

BMI - Body Mass Index

BN - Bayesian Network

BSS - Brier Skill Score

COPD - Chronic Obstructive Pulmonary Disease

CPDAG - Completed Partially Directed Acyclic Graph

CPT - Conditional Probability Table

CV - Cross-Validation

DAG - Directed Acyclic Graph

ECE - Expected Calibration Error

EM - Expectation–Maximisation

F1 - F1-score

FCI - Fast Causal Inference

FMUP - Faculdade de Medicina da Universidade do Porto / Faculty of Medicine, University of Porto

FN - False Negative

FP - False Positive

GDPR - General Data Protection Regulation

HC - Hill-Climbing

HPDI - Healthful Plant-Based Diet Index

IAN-AF - Inquérito Alimentar Nacional e de Atividade Física / Portuguese National Food and Physical Activity Survey

ISS - Imaginary Sample Size

LASSO, L1 - Least Absolute Shrinkage and Selection Operator

LR - Logistic Regression

MAP - Maximum A Posteriori

MCC - Matthews Correlation Coefficient

MD - Mediterranean Diet

MDSS - Mediterranean Diet Serving Score

ML / MLE - Maximum Likelihood Estimation

MNAR - Missing Not At Random

NB - Naïve Bayes

NCDs - Noncommunicable Diseases

NHANES - National Health and Nutrition Examination Survey

NUTS - Nomenclature of Territorial Units for Statistics

PDAG - Partially Directed Acyclic Graph

PNPAS - Programa Nacional para a Promoção da Alimentação Saudável

ROC - Receiver Operating Characteristic

SD - Standard Deviation

SES - Socioeconomic Status

SMOTE - Synthetic Minority Over-sampling Technique

SMOTENC - Synthetic Minority Over-sampling Technique for Nominal and Continuous features

TAN - Tree-Augmented Naïve Bayes

TN - True Negative

TP - True Positive

UNESCO - United Nations Educational, Scientific and Cultural Organisation

WHO - World Health Organisation

1. Introduction

1.1. Context and background

Noncommunicable diseases (NCDs), including heart disease, stroke, cancer, diabetes and chronic lung disease, are collectively responsible for 74% of all deaths worldwide (1). Approximately 86% of the total burden of disease can be attributed to this type of disease, and inadequate dietary habits are among the main modifiable risk factors.

In Portugal, data from the Global Burden of Disease Study (2015) showed that inadequate dietary habits are responsible for approximately 9% of disability-adjusted life-years (DALYs; 300,000 DALYs) and represent the leading modifiable risk factor for mortality. Just over 237,000 years of life could be saved if Portuguese people improved their eating habits (2).

It is well established that diet has a significant influence on health status. In terms of risks, foods with excessive caloric content, particularly those high in salt, processed meats, sugar, and trans fats (industrially processed), constitute the main problem (3). Alternative Mediterranean Diet (AMED), Alternative Healthy Eating Index (AHEI), and Healthful Plant-Based Diet Index (HPDI), which were selected due to their widespread use in nutritional epidemiology and demonstrated predictive validity for chronic disease outcomes; although they differ in dietary components and scoring approaches, these indices have been inversely associated with the risk of all cancers and certain types of cancer (4).

Despite evidence highlighting the critical role of modifiable risk factors in improving health outcomes, one study conducted among Portuguese patients in a primary care setting indicated that individuals often place a higher value on diagnostic and laboratory tests than on adopting lifestyle changes and seem unaware that behavioural risks are the primary modifiable risk factors (5).

Addressing modifiable risk factors, particularly dietary habits, can significantly improve public health outcomes (6), highlighting the need for data-driven public health policies and targeted interventions capable of identifying individuals and population groups at higher risk. When discussing complex topics like chronic disease and healthy eating habits, uncertainty is inevitable. In this context, Bayes' theorem emerges as a powerful tool. First discovered over 250 years ago by English clergyman Thomas Bayes (1701–1761) and later refined by French polymath Pierre-Simon Laplace (1749–1827), Bayes' formula helps address uncertainty, including the challenges faced by clinicians in medical decision-making.

Unlike traditional hypothesis testing, which yields fixed binary outcomes (*e.g.*, positive or negative), the Bayesian approach allows for continuous probability updates as new data becomes available. This method

offers two key advantages: it does not inherently favour the null hypothesis, and it evolves dynamically with additional information (6).

The question the dissertation aims to answer is: “How can Bayesian Networks contribute by describing dependencies and predicting the existence of healthy eating patterns in the Portuguese population?”

Objectives:

1. Describe the dependencies between variables related to healthy eating using Bayesian Networks, through probabilistic structure learning and conditional dependency analysis.
2. Develop a predictive model based on Bayesian Networks to estimate the probability of an individual maintaining healthy eating based on available consultation variables.
3. Evaluate the performance and robustness of the proposed Bayesian Network model using appropriate validation and accuracy metrics.

1.2. Structure of the dissertation

This dissertation is organised into six chapters, each addressing a distinct phase of the research process, from theoretical grounding to empirical modelling and critical reflection on findings.

Chapter 1 introduces the research problem by situating healthy eating within the broader context of non-communicable disease burden in Portugal, establishing the rationale for a probabilistic modelling approach, defining the central research question, and stating the study objectives. The present section outlines the overall organisation of the dissertation.

Chapter 2 provides the theoretical and empirical foundations of the study through a structured literature review. It begins by defining healthy eating and dietary patterns, with reference to the WHO framework and the Mediterranean Diet, and proceeds to examine the sociodemographic, behavioural, and psychosocial determinants of dietary behaviour documented in the literature. The chapter then introduces Bayesian Networks, covering their theoretical underpinnings, structure learning methods, and existing applications in nutritional epidemiology, before identifying the research gap that motivates the present work.

Chapter 3 describes the methodological framework adopted in this study. It details the study design and data source, the variable engineering process and operationalization of the outcome variable, the structure learning procedure including the Hill-Climbing algorithm, the BDeu scoring criterion, bootstrap-based stability assessment, and the construction of the consensus network as well as the parameter estimation approach, the validation framework, comparisons with baseline models, sensitivity and robustness analyses, and the ethical considerations governing the research.

Chapter 4 presents the results of the study in a structured sequence. It begins with descriptive characteristics of the analytical sample, followed by the learned Bayesian Network structure and its key conditional probability tables. Predictive performance metrics across validation folds are then reported,

alongside variable importance analyses and *what-if* scenario simulations. The chapter concludes with the findings of the sensitivity and robustness analyses.

Chapter 5 discusses the results considering existing literature, contextualising the main findings within prior evidence on dietary determinants and Bayesian Network applications. It reflects on the specific contributions of the study, examines its strengths and limitations, and considers the implications of the findings for public health practice and policy. Directions for future research are also proposed.

Chapter 6 concludes the dissertation by summarising the study, providing a direct answer to the research question, and offering final considerations on the role of Bayesian Networks as a methodological tool in nutritional epidemiology.

2. Literature review

2.1. Healthy eating and dietary patterns

The World Health Organisation (WHO) frames healthy eating as a predominantly plant-based pattern, emphasising fruits, vegetables, and legumes together with limits on components that increase cardiometabolic risk, namely free sugars, unhealthy fats, and salt. WHO guidance explicitly highlights a minimum intake of ≥400 g (≈5 portions) of fruit and vegetables/day, together with limits on free sugars and salt, which provides an internationally recognised benchmark for “healthy eating” against which food-based proxies can be justified (7,8).

Conceptually, the above criteria also map onto the Mediterranean dietary pattern, which is widely used in nutritional epidemiology as an exemplar of a healthy dietary pattern (9–12); this diet is the food pattern that best represents the characteristics of the food culture in Portugal (13).

The classic Mediterranean diet is characterised by high intakes of fruits, vegetables, legumes, and cereals, a beneficial fat profile, and (traditionally) moderate ethanol consumption (14). Adherence to this diet has been prospectively linked to increased longevity, as evidenced by Greek multicenter prospective cohort studies (15), and has been associated with an improved quality of life across both Mediterranean (specifically Italian) and non-Mediterranean self-reporting cohorts (16).

Consistent with this perspective, a Mediterranean diet is commonly operationalised at the food-group level (rather than by single nutrients), precisely because dietary patterns reflect the synergistic and cumulative effects of habitual food combinations (9–12).

The updated Mediterranean Diet Pyramid places greater emphasis on foods to be consumed at each main meal, particularly fruits and vegetables. It also incorporates a component for specified frequencies of fermented beverages (wine/beer) for adults. Key "base-of-the-pyramid" behaviours, including meal regularity (main meals), consumption of fruit, and intake of legumes and salads, are central to MDSS-based adherence frameworks and are effectively captured by this model (11).

Finally, regarding alcohol, a healthy diet can include alcohol if we treated it as a "non-excessive consumption" criterion (i.e., not exceeding a moderate threshold) rather than as an abstinence requirement. This is consistent with the way traditional Mediterranean adherence scores incorporate moderate fermented beverage intake, while acknowledging contemporary WHO risk communications stating that no level of alcohol consumption is risk-free, even at low levels (5,7,12).

2.2. Determinants of healthy eating

Now that we understand how healthy eating was classified, this segment will analyse the determinants of healthy eating, exploring how these factors influence attitudes and intentions, ultimately translating into behaviour. Furthermore, we will examine how the surrounding environment also shapes this behaviour. We will have a multilevel approach to personal, social, and environmental behaviour.

Across studies, facilitators of healthy eating clustered into three domains. Personal facilitators included meal planning and preparation, food label literacy, mindful eating, anticipating the physical effects of food, and allowing flexibility through occasional indulgences. Social facilitators encompass supportive family environments, culturally adaptive healthy cooking practices, peer influence, exposure to healthy food content on social media, and religious or cultural norms that promote healthy eating. Environmental facilitators are primarily related to the availability and affordability of healthy foods, sufficient time for meal preparation, and structured mealtimes (17–20).

At the individual level, key determinants include nutrition knowledge and skills, beliefs and attitudes, psychological factors, physiological preferences, and habitual behaviours. At the social level, influences from family, friends, peers, and coworkers, alongside exposure to social marketing of unhealthy foods and prevailing sociocultural norms, emerge as consistent and powerful drivers of dietary behaviour. In contrast, determinants at the environmental level, such as food price and affordability, availability, product characteristics, time constraints, and convenience, are predominantly described as barriers, rather than facilitators, to healthy eating (21).

Personal barriers included stress, limited cooking skills, cravings, inadequate nutrition knowledge, unhealthy relationships with food, competing priorities, and the perception that negative health consequences are temporally distant.

A conference paper published in 2012 showed that weight concern, knowledge, concern for diseases, social influence, and food choice motive influence healthy eating attitude and predict 65% of the healthy eating attitude variance. The highest prediction is weight concern, being the most significant influencer of healthy eating attitude in the study target population. (21).

Qualitative evidence further reinforces the perceived importance of healthy eating. Participants commonly associate healthy eating with improved physical, mental, and emotional well-being. However, despite acknowledging its importance, many reported deprioritising healthy eating due to time constraints or a perception that the consequences of unhealthy eating were not immediately salient, particularly among individuals in good health (22).

Culinary skills and food literacy constitute key facilitators, as they allow individuals to understand food labelling and prepare home-cooked meals, thus limiting dependence on ultra-processed products (22–24). However, eating behaviour is also shaped by psychological factors such as stress and anxiety, which often act as triggers for hedonic hunger and the excessive consumption of foods high in sugar and fat as a form of emotional compensation (13,23,25).

Further insights into heterogeneity in eating behaviour are provided by studies identifying two distinct consumer clusters based on food motivation profiles: Emotional Eating Consumers and Health-Driven Consumers. Although both groups expressed strong concern for environmental and political aspects of food consumption, their secondary motivations diverged substantially. Emotional Eating Consumers were primarily influenced by emotional factors, whereas Health-Driven Consumers prioritised health-related considerations (26).

Healthy eating is influenced by a complex web of sociodemographic and economic determinants, among which gender and age play a prominent role, as women and older individuals tend to show greater adherence to balanced dietary patterns such as the Mediterranean Diet (11,27,28). Our food choices are strongly influenced by social factors. Research has highlighted the robust tendency for people to adjust their food intake to align with perceived eating norms, and to convey positive impressions of themselves to other people (29).

Educational level and household income also emerge as robust predictors of nutritional quality, as the high cost of fresh foods such as fish and olive oil constitutes a significant barrier for economically vulnerable groups (28,29).

Complementing these findings, a meta-ethnographic synthesis suggests that the determinants of healthy eating in similar categories appear across strata, but salience/constraints differ; structural barriers are more pronounced in lower Socioeconomic Status (SES) contexts, affecting lower socioeconomic groups to a comparable extent as the general population (5).

Lower adherence is consistently associated with fewer years of formal education, lower income, and greater perceived financial difficulty, highlighting the strong socioeconomic gradient in Mediterranean Diet (MD) adherence. Older adults show higher adherence than younger generations, suggesting a progressive erosion of traditional dietary patterns among younger cohorts.

Despite some encouraging public health developments such as fiscal and regulatory measures targeting salt and sugar-sweetened beverages, adherence to the MD has declined across southern Europe over recent decades, particularly among socioeconomically disadvantaged groups. Economic constraints, including the higher cost of foods characteristic of the MD, appear to play a central role in this trend. Lower MD adherence is also independently associated with poorer quality of life, reinforcing the bidirectional relationship between diet quality, social inequalities, and health outcomes (27,28).

Parenting, social influences, and the food environment influence the development of eating behaviour (30). There's a lot of focus on the benefits of the social aspect of the Mediterranean Diet. The socio-cultural context of the Mediterranean Diet has been driven by public health experts within the region and has paralleled its UNESCO recognition. The 2013 UNESCO report highlights that the context of food consumption is a fundamental aspect of the Mediterranean Diet. The act of eating together is seen to promote cultural identity and ensure social continuity (31).

Conviviality is more than a shared meal, because it relates to “how” we eat. The “how” involves socialising when eating and, as complementary aspects, self-awareness of satiety and hunger signals, eating slowly, and chewing well (32).

However, studies show that people eat more when eating with friends and family, relative to when eating alone. Known as the ‘social facilitation of eating ', the augmentation of food intake in the presence of co-eaters can promote overconsumption and misalignment with formal nutritional guidelines. Can act as a barrier to moderation and portion control, even in contexts where nutritional recommendations advocate restraint (29).

The food environment plays a decisive role through the high availability of fast food and aggressive exposure to digital marketing, which promotes the desire for unhealthy snacks, particularly among younger individuals (25,26,33). In the Portuguese context, adherence to healthy diets still faces practical barriers, such as lack of time for meal planning and cooking, as well as the perception that foods such as legumes may be fattening. Finally, the literature emphasises that public policies focused on product reformulation and the taxation of sugar-sweetened beverages are essential instruments for modifying these food environments and promoting long-term public health (33).

Eating behaviour is strongly influenced by social context. Dietary choices also tend to converge with close social connections. One reason for this is that conforming to the behaviour of others is adaptive and we find it rewarding. Norms of appropriate eating are set by the behaviour of other people, but also shared cultural expectations and environmental cues (34).

New life stages like starting college were frequently associated with deteriorations in dietary quality. Participants described increased reliance on convenience foods, reduced frequency of home-cooked meals, meal skipping, irregular eating patterns, and episodes of overeating or binge eating linked to stress and disrupted schedules. Nonetheless, some individuals reported increased dietary awareness, particularly when eating alone or when gaining greater autonomy over food purchasing and preparation (35).

Variations in dietary patterns have been linked to distinct behavioural profiles, reflecting underlying differences in motivations and health-related behaviours. In a population-based cohort of UK adults, differences in eating and dietary behaviours were identified between smokers and non-smokers, with potential implications for food intake and weight status. Smoking was associated with eating and dietary behaviour patterns consistent with inhibited food intake, low diet quality, and altered food preferences. These behavioural patterns may help explain the weight gain commonly observed during smoking cessation, suggesting that changes in appetite regulation and food choice play a role in post-cessation weight dynamics (36).

Suboptimal eating habits have also been identified in patients with chronic pain, including irregular mealtimes and weekly or daily consumption of confectionery and fast food (37). In this population, over three-quarters of individuals reported engaging in pain-induced comfort eating. Although chronic pain intensity did not significantly predict pain-induced comfort eating directly, a significant indirect effect was observed through stress, while experiential avoidance and pain catastrophizing did not mediate this relationship (38). These

findings highlight the central role of psychological stress in shaping eating behaviours among individuals with chronic pain.

Among Brazilian adults, the prevalence of prolonged screen time was 21.8% for television viewing and 22.2% for the use of other electronic devices during leisure time, which was associated with unhealthy eating patterns, characterised by both higher consumption of unhealthy foods and lower consumption of healthy foods. These findings indicate that prolonged screen exposure is negatively associated with healthy food consumption and promotes the intake of unhealthy foods in adult populations (25).

A systematic review and meta-analysis of ten studies further demonstrated that eating while distracted, such as while listening to music, reading, or watching television, led to a small but significant increase in food intake among adults. Previous research has shown that distracted individuals tend to eat larger portions and consume more food compared to when they are focused, as distraction can reduce fat perception in high-fat foods, thereby increasing the risk of overeating energy-dense products. Overall, this systematic review and meta-analysis advances understanding of the time-dependent effects of screen exposure on food intake in healthy adults, demonstrating that screen-based distraction increases food consumption regardless of screen content (39).

2.3. Healthy eating in Portugal

Following industrialisation, a gradual distancing of the Portuguese population from the Mediterranean dietary pattern was observed. Thus, to preserve and promote this heritage, the Mediterranean Food Wheel was developed. Its circular format conveys complementarity and dietary balance, unlike the pyramid format, which can convey a message of hierarchy among foods (40).

Studies indicate that health concerns are the primary driver of food purchasing decisions in Portugal. However, data on actual dietary habits reveal a concerning discrepancy between intention and behaviour (41). Using data from the Portuguese National Food, Nutrition and Physical Activity Survey (IAN-AF 2015–2016), with results published/presented in 2017, conducted by *Universidade do Porto,* analyzed the food consumption patterns of 6,553 individuals across Portugal, aged between 3 months and 84 years. The results highlight several dietary imbalances. The health status of the Portuguese population was characterised by a high prevalence of chronic diseases such as hypertension, obesity, and diabetes. Unhealthy lifestyles, including inadequate nutrition, are known to be central determinants for these diseases (42).

Dietary intake data indicate substantial deviations from nutritional recommendations in the Portuguese population. More than half of individuals fail to meet the World Health Organisation guideline of consuming at least 400 g/day of fruit and vegetables, with inadequacy particularly pronounced among children and adolescents. Meat consumption exceeds recommendations, with a considerable proportion of adolescents and adults consuming more than 100 g/day of red meat, while intake of fruit, vegetables, and cereals remains below the levels proposed by the Portuguese Food Wheel. Consumption of sugar-sweetened beverages is

notable, especially among adolescents, among whom prevalence and intake levels are markedly higher. Alcohol intake remains substantial among adults and the elderly.

Overall, these patterns reveal an imbalance characterised by excessive intake of meat, alcohol, and sugar-sweetened drinks, alongside insufficient consumption of plant-based foods (42).

Evidence from nationally representative data indicates that adherence to the Mediterranean Diet (MD) among Portuguese adults is generally low and socially patterned. Most adults exhibit low to moderate adherence, with only a small proportion achieving high adherence, although recent reports suggest a modest improvement over time.

Another concerning finding is that 21% of the food consumed in Portugal is not included in the official food wheel. These are drinks, nectars, juices, coffees, teas, wine, beer, and other alcoholic beverages. Men consume more alcohol than women, and older people more than adults. Five per cent of the elderly drink over a litre of wine daily (43,44).

These findings emphasise the gap between perceived health-conscious food choices and actual consumption patterns. The overconsumption of processed foods and animal-based products, coupled with insufficient intake of plant-based foods, poses significant public health challenges.

The Portuguese population appears to be consuming high amounts of salt. Excessive salt intake in the diet is a well-documented issue, and it has been observed that consuming 100 g of a composite dish or a cereal-based product (such as bread or plain crackers) can account for around 30% of the recommended daily salt intake (45).

Improving healthy eating in the Portuguese population requires integrated food and nutrition policies that both support informed individual choices and shape healthier food environments. Dietary habits have a direct impact on health and play a significant role in environmental sustainability, influencing resource use and climate change. Evidence from Portugal and other European countries shows that interventions targeting food availability, such as salt reduction in staple foods and improvements in school food environments, can effectively modify consumption patterns. Moving forward, policies should balance nutritional and environmental goals while respecting consumer autonomy and enabling sustainable, healthy dietary choices (41).

2.4. Bayesian networks in health research and nutritional epidemiology

2.4.1. What are Bayesian networks

We will now describe the basic concepts around what are Bayesian Networks (BNs). Bayesian networks make use of probability theory and graph theory to search through a state space and make decisions

under uncertainty. Therefore, the main theoretical concepts that we need to understand to implement a Bayesian network fall under probability theory (specifically Bayes' theorem) and graph theory.

In general, P (A|B) is read as "the probability of A given B", where A is the dependent variable, and B is the independent variable. Bayes' theorem allows us to calculate the probability of our hypothesis H given our evidence and background. That is, given the context c and evidence E, we can know the probability of hypothesis H by using the conditional probability of H given c, the conditional probability of E given H and c, and the probability of E given c.

The formula is often simplified to the form.

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

This is an acceptable substitution since the background context often doesn't change and thus we can assume it to be constant throughout our analysis.

Bayes' theorem can then be rewritten using the definition of conditional probability.

$$P(x|y) = \frac{P(x, y)}{P(y)}$$

A great benefit of Bayes' theorem is that you can calculate the probability of event A given event B when you can more easily calculate the probability of event B given event A. That is, we can derive P (A|B) from P (B|A).

In the next chapter, we will discuss how this mathematical theorem can aid in Nutritional Epidemiology (46).

2.4.2. Bayesian networks and their applicability in nutritional epidemiology

Bayesian Networks (BNs) have received increasing attention during the last two decades for their ability to be applied to challenging issues and aid those making decisions to reason about cause and outcome under conditions of uncertainty.

Machine learning (ML) and artificial intelligence (AI) approaches have been proposed for a diverse range of health reasons. A significant benefit of BNs over methods of pure ML from data is that BNs do not explicitly require massively large datasets. BNs can incorporate the accumulated knowledge of experts in circumstances where data are limited, and still produce meaningful and accurate decision-support systems (47). Experts find numerous sources of possible uncertainty, unclear and incomplete measurements, and events for which it is difficult to establish causal relationships. Accordingly, probabilistic graphs are useful tools to

understand, in a visual and intuitive way, how each factor affects the rest of the system (48), unlike traditional statistical methodologies, which often rely on linear relationships and strong parametric assumptions. Dynamic Bayesian Networks (DBNs) can capture nonlinear, time-dependent interactions among multiple covariates (49).

BNs have been extensively applied in public health and disease modelling, especially for obesity and diabetes prediction (50–56). However, despite their success in these areas, there is a notable gap in research applying Bayesian Networks to predict and analyse dietary patterns in the general population.

Most BN-based dietary studies focus on undernutrition and food insecurity in developing countries, where the primary concern is the lack of access to essential nutrients (56). There is limited research utilising Bayesian Networks to model and predict healthy eating habits in well-developed food environments, such as Portugal.

2.5. Summary: gaps and contribution of this thesis

This study seeks to bridge this research gap by exploring the potential of Bayesian Networks in predicting healthy eating behaviours in the Portuguese population, providing new insights into dietary trends, risk factors, and targeted interventions.

This study aims to enhance the understanding of the key factors influencing healthy eating habits in the Portuguese population through the application of Bayesian Networks. By identifying the most significant variables and determinants of healthy eating, this research seeks to generate insights that can inform and support the development of evidence-based policies for promoting balanced dietary practices.

A core objective is the construction of a Bayesian Network capable of predicting the probability of an individual adopting healthy eating habits based on multiple variables, allowing for more precise and data-driven strategic interventions in public health. Additionally, this study explores the potential of Bayesian Networks as an effective tool for modelling eating behaviours and assessing nutritional risks, highlighting their relevance in chronic disease prevention and healthcare cost reduction.

Beyond its immediate applications, this research also aims to stimulate further exploration of Bayesian Networks in dietary behaviour analysis, reinforcing their role as a valuable methodology for improving public health strategies and enhancing decision-making in nutritional epidemiology.

3. Methods

3.1. Study design and analytical workflow

This study followed a structured, multi-phase analytical design integrating literature review, rigorous data preprocessing and robust predictive validation. Each phase built directly upon the outputs of the previous one, forming a coherent analytical pipeline from problem definition to model evaluation.

The research began with a comprehensive literature review to contextualise the problem of healthy eating in Portugal and to identify relevant determinants supported by nutritional epidemiology and behavioural health research.

Building on these findings, variable engineering was conducted according to a predefined variable map, including recoding rules, discretisation criteria (clinical where available; distribution-based otherwise), and semantic harmonisation to ensure interpretability of ordinal relationships.

Once the analytical dataset was consolidated, and prior to model learning, the data were partitioned using stratified k-fold cross-validation (k = 5), preserving the class distribution of the binary outcome variable (*alimentacao_saudavel_bin*) across folds. This strategy ensured unbiased evaluation of predictive performance.

Within this validation framework, Bayesian Network structure learning was performed using a score-based Hill-Climbing algorithm with the Bayesian Dirichlet equivalent uniform (BDeu) score, better explained in section 3.5. Structural robustness was assessed through bootstrap resampling (≥ 200 networks), and a consensus graph was constructed by retaining arcs with a strength threshold of 0.3 (experiments were conducted to identify the optimal threshold), depending on sensitivity analysis. Arc direction was considered only when directional consistency was sufficiently supported. With the structure established, parameter estimation was conducted using Maximum A Posteriori (MAP) estimation, given the absence of missing data requiring expectation-maximisation procedures.

Predictive performance was then evaluated on validation folds using discrimination and probabilistic metrics appropriate for imbalanced binary outcomes. Primary metrics included the Area Under the Precision–Recall Curve (PR-AUC), which quantifies the trade-off between precision and recall across decision thresholds and is particularly informative for the minority class; the Brier Skill Score (BSS), which measures the improvement in probabilistic accuracy relative to a reference model; and calibration assessment, which evaluates the agreement between predicted probabilities and observed outcome frequencies. Secondary metrics included ROC-AUC (overall discrimination between classes), accuracy (proportion of correct classifications), precision (proportion of positive predictions that are correct), recall (proportion of true positives correctly

identified), and F1-score (harmonic mean of precision and recall). Complementarily, confusion matrices were analysed to assess classification behaviour under the selected decision threshold.

Beyond the primary evaluation, model sensitivity and robustness were further explored through alternative discretisation strategies, variation of structural strength thresholds, and *what-if* intervention analyses on modifiable nodes. Additional experiments included SMOTENC, weighted bootstraps, alternative seeds, and age-stratified models.

Finally, model performance was compared with alternative approaches, ensuring methodological comparability across the study. A visual exemplification of the work elaborated is presented in Figure 3.1.

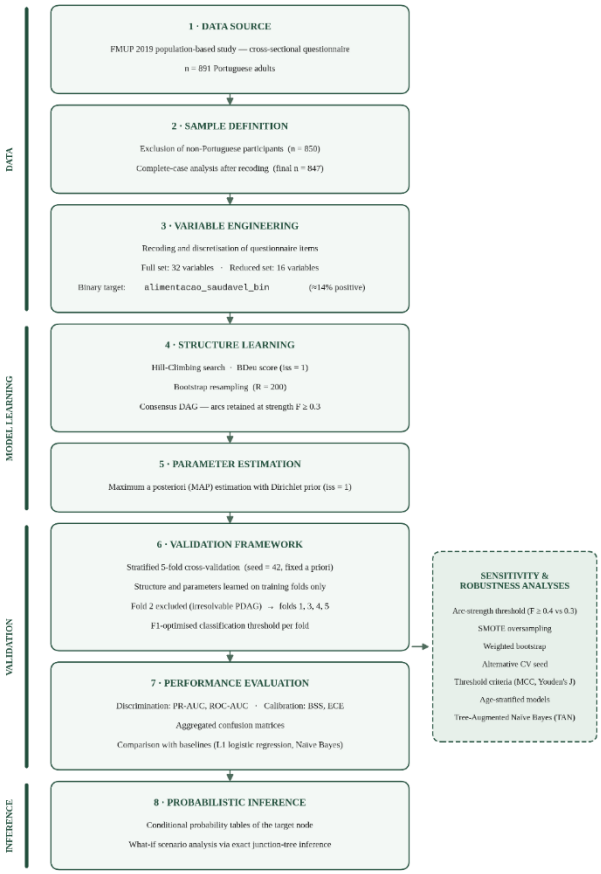

Figure 3.1 - Study workflow.

3.2. Data Analysis

Variable recoding, discretisation and the construction of composite scores were performed in *IBM SPSS Statistics (version 30.0.0.0)*, following the variable map described in Section 3.4. The exclusion of the

41 participants of foreign nationality and of the 3 participants with missing values in *ecras_interf_2s_cat5* was likewise carried out in *SPSS*, yielding the final analytical sample of n = 847. The preliminary analyses reported in Section 4.2 were performed on an earlier edition of the dataset, prior to the exclusion of foreign nationals (n = 891). The recoded datasets were then exported to *R (version 4.6.0)*, where all remaining analyses and all figures were produced. Structure learning, bootstrap resampling (R = 200) and parameter estimation were performed with the *bnlearn package*, and exact inference with *gRain*. Network structures were rendered with *Rgraphviz* and with *ggraph* and *igraph.* Model evaluation figures were produced with *ggplot2* and composed with *patchwork*. The choropleth maps of mainland Portugal by NUTS II region were produced with *ggplot2* and *sf*, using official NUTS II geometries retrieved through giscoR, with dplyr for data joins and ggrepel for regional labelling. Sensitivity analyses used *UBL* for synthetic minority oversampling and *pROC* for threshold-based diagnostics; baseline classifiers were fitted with glmnet and e1071. The sole exception is the analytical workflow diagram (Figure 3.1), which was produced in *Python 3* using *Matplotlib*. No null-hypothesis significance testing was performed: structural inclusion was governed by the bootstrap arc-strength threshold ($F \geq 0.3$), fixed a priori, and classification by the F1-optimised probability cut-off estimated within each training fold.

3.3. Data source and participants

The data used in this study derive from a national survey conducted by the Faculty of Medicine of the University of Porto between January and April 2019 (5,8).

Data collection took place from January 16 to April 30, 2019, through a structured questionnaire administered in Portuguese. It was administered through face-to-face interviews conducted at participants' homes by trained interviewers. The average duration of each interview was approximately 20–25 minutes.

Participation was voluntary. Exclusion criteria included the presence of cognitive or physical impairments preventing participation in a face-to-face interview, residence in collective dwellings (e.g., institutions), and inability to speak or understand Portuguese.

The initial sample consisted of 891 participants.

3.4. Data collection instrument

The interviewers were trained to clarify the meaning of each question to ensure that participants correctly understood all the questions and were trained to follow standard operating procedures for all contacts made. For quality control, all interviews were monitored by a data collection supervisor, and at least 20% were randomly supervised by members of the investigation team.

A structured questionnaire containing three sections, after an introductory section presenting study aims and motivation, was used: (1) questions about health status; (2) main research section; and (3) sociodemographic data. In the main research section, the interviewees answered: (A) a Likert scale (1 = strongly disagree, 2 = partially disagree, 3 = indifferent, 4 = partially agree, and 5 = strongly agree) on different lifestyle interventions (diet, physical activity, alcohol consumption, tobacco use, illicit drugs, sleep habits, screen, stress, and being sedentary); and (B) nine closed and direct questions about their pattern of lifestyle, including all of the different interventions identified above. The questionnaire can be examined in Annex A.

In the Likert scale, respondents reported for each lifestyle intervention: (1) personal belief about each lifestyle intervention (can prevent and help control *vs*. can cause and aggravate some diseases, as appropriate); (2) personal behavior self-assessment (if, in the personal case, each lifestyle intervention is healthy or not; (3) level of importance given to the family doctor to address each lifestyle intervention; and (4) if, in the personal case, something was usual or not, the family doctor asks/advises on each lifestyle intervention.

3.5. Definition of variables and data preparation

The original database consisted of the complete set of responses to the questionnaire presented in Annex A (5,8). In an initial phase, data cleaning procedures were performed, and the study variables were created, which are described below (Section 3.4.1). Subsequently, based on the evaluation of the available variables, the existing literature, and feedback from researchers who had previously worked with this database, the final variables to be included in the analysis were selected, as described in the following subsection (Section 3.4.2). The distribution of each variable in the dataset can be examined in Appendix A.

3.5.1. Full variables

alimentacao_saudavel_bin

The variable *alimentacao_saudavel_bin* represents healthy eating and is the target variable of the model, being of ordinal binary type where 1 means healthy diet and 0 means non-healthy diet. Originating from a composite index or score defined from items of the food consumption and alcohol questionnaire where a healthy diet is considered when all the following conditions are simultaneously true:

- Number of main meals per day must be between 2 to 3 according to question Q2.
- Fruit must have daily frequency and between 2 to 6 portions per day according to questions Q2.2 and Q2.2.1.
- Vegetables salads and legumes must have daily frequency and between 2 to 6 portions per day according to questions Q2.3 and Q2.3.1.

- Alcohol consumption must be moderate following the criteria that women and men aged 65 or older should consume a maximum of 1 drink per day while men under 65 years may consume a maximum of 2 drinks per day applying both to weekdays through question Q4.1.1 and to weekends through question Q4.2.1 with those who do not drink on these days being considered within the limit.

egs_cat3

The variable *egs_cat3* represents the perception of general health status and is of ordinal type originating from the question "Overall, how do you consider your health?" which originally had 5 categories being Very Good, Good, Reasonable, Bad and Very bad but due to the original distribution having few very bad and bad cases it was decided to recode into only 3 categories being Very Good or Good, Reasonable and Very Bad or Bad.

imc_cat6

The variable *imc_cat6* represents Body Mass Index (BMI) categorised into six WHO levels. It is of ordinal type. It originates from two questions: Q1.1, which asks for weight in kilograms, and Q1.2, which asks for height in centimetres. BMI is calculated using the formula BMI = weight / height², with height expressed in metres. The resulting value is then categorised into six classes according to WHO cut-offs: 1 for underweight (BMI < 18.5); 2 for normal weight (18.5 ≤ BMI < 25); 3 for overweight (25 ≤ BMI < 30); 4 for obesity grade I (30 ≤ BMI < 35); 5 for obesity grade II (35 ≤ BMI < 40); and 6 for obesity grade III (BMI ≥ 40).

hipertensao_12m_bin

The variable *hipertensao_12m_bin* represents arterial hypertension and is of nominal type, originating from the question "Indicate during the last 12 months you have had any of the following health problems," being recoded into binary, where 0 means without hypertension, and 1 means with hypertension.

hipercol_12m_bin

The variable *hipercol_12m_bin* represents hypercholesterolemia or elevated cholesterol and is of nominal type, originating from the question "Indicate during the last 12 months you have had any of the following health problems," being recoded into binary, where 0 means without hypercholesterolemia and 1 means with hypercholesterolemia.

diabetes_12m_bin

The variable *diabetes_12m_bin* represents diabetes and is of nominal type, originating from the question "Indicate during the last 12 months you have had any of the following health problems," being recoded into binary, where 0 means without diabetes and 1 means with diabetes.

doenca_cardiaca_12m_bin

The variable *doenca_cardiaca_12m_bin* represents cardiac problems and is of nominal type, originating from the question "Indicate during the last 12 months you have had any of the following health problems," being recoded into binary, where 0 means without cardiac problems and 1 means with cardiac problems.

avc_12m_bin

The variable *avc_12m_bin* represents stroke or cerebrovascular accident and is of nominal type, originating from the question "Indicate during the last 12 months you have had any of the following health problems," being recoded into binary, where 0 means without stroke and 1 means with stroke.

asma_dpoc_12m_bin

The variable *asma_dpoc_12m_bin* represents asthma and/or COPD, which is chronic bronchitis, and is of nominal type, originating from the question "Indicate during the last 12 months you have had any of the following health problems", being recoded into binary, where 0 means without asthma or COPD and 1 means with asthma or COPD.

gastrite_ulcera_12m_bin

The variable *gastrite_ulcera_12m_bin* represents gastritis or peptic ulcer and is of nominal type, originating from the question "Indicate during the last 12 months you have had any of the following health problems", being recoded into binary, where 0 means without gastritis or ulcer and 1 means with gastritis or ulcer.

ansiedade_12m_bin

The variable *ansiedade_12m_bin* represents anxiety and is of nominal type, originating from the question "Indicate during the last 12 months you have had any of the following health problems," being recoded into binary, where 0 means without anxiety and 1 means with anxiety.

depressao_12m_bin

The variable *depressao_12m_bin* represents depression and is of nominal type, originating from the question "Indicate during the last 12 months you have had any of the following health problems," being re-coded into binary, where 0 means without depression and 1 means with depression.

cancro_12m_bin

The variable *cancro_12m_bin* represents cancer and is of nominal type, originating from the question "Indicate during the last 12 months you have had any of the following health problems," being re-coded into binary, where 0 means without cancer and 1 means with cancer.

dores_osteomusculares_12m_bin

The variable *dores_osteomusculares_12m_bin* represents musculoskeletal pain and is of nominal type, originating from the question "Indicate during the last 12 months you have had any of the following health problems," being re-coded into binary, where 0 means without musculoskeletal pain and 1 means with musculoskeletal pain.

sexo_cat

The variable *sexo_cat* represents biological sex and is of nominal type, originating from the question about sex in the questionnaire, with the original categories being maintained, where 1 means male and 2 means female.

ansiedade_2s_cat3

The variable *ansiedade_2s_cat3* represents anxiety in 3 levels and is of ordinal type originating from questions P3Q9 which asks "In the last 2 weeks, how often have you done less because anxious/nervous?", P3Q9.1 which asks "In the last 2 weeks, how often have you felt calm and peaceful?" and P3Q9.2 which asks "In the last 2 weeks, how often did you feel anxious/nervous?" where the original categories of the items were 1 for Every day, 2 for Most part of the days, 3 for Some days, 4 for A few days and 5 for Never being recoded so that higher means more anxiety through direction alignment where Q9_rev equals 6 minus Q9 to reverse, Q9_1 equals Q9_1 without reversing and Q9_2_rev equals 6 minus Q9_2 to reverse, then calculating the total score ansiedade_score as Q9_rev plus Q9_1 plus Q9_2_rev with range from 3 to 15 where higher means more anxiety and being then discretized for BN into 3 states being 1 for Low, 2 for Moderate and 3 for High.

med_ansiedade_2s_bin

The variable *med_ansiedade_2s_bin* represents medication for anxiety in the last 2 weeks and is of binary nominal type originating from question Q9.3 "In the last 2 weeks, have you had to take medication to control anxiety/nervousness?" where the original categories were 1 for Every day, 2 for Most part of the days, 3 for Some days, 4 for A few days and 5 for Never being recoded into binary "used at least once" where 0 means No and 1 means Yes through the rule where Q9.3 equal to 5 results in 0 and Q9.3 belonging to the set 1, 2, 3 or 4 results in 1.

escolaridade_cat4

The variable *escolaridade_cat4* represents education in 4 levels and is of ordinal type originating from the question about detailed education where 1 means none, 2 means 1st cycle, 3 means 2nd cycle, 4 means 3rd cycle, 5 means secondary and 6 to 9 means university with the original categories as above and the recoding made through the rule where 1 means None, 2 means 1st to 3rd cycle, 3 means Secondary and 4 means Higher education through the rule 1 goes to 1, 2 and 3 and 4 go to 2, 5 and 3 go to 3 and 6 to 9 go to 4 .

idade_cat14

The variable *idade_cat14* represents age by 5-year age groups and is of ordinal type originating from the question "How old are you? ____ years" which was continuous age in years with the original categories being continuous in complete years and the recoding made into 14 groups being 1 for 20 to 24 years, 2 for 25 to 29 years, 3 for 30 to 34 years, 4 for 35 to 39 years, 5 for 40 to 44 years, 6 for 45 to 49 years, 7 for 50 to 54 years, 8 for 55 to 59 years, 9 for 60 to 64 years, 10 for 65 to 69 years, 11 for 70 to 74 years, 12 for 75 to 79 years, 13 for 80 to 84 years and 14 for 85 or more years through 5-year groups which are standard in public health and capture variation throughout the life cycle maintaining clear interpretation.

drogas_ilicitas_cat3

The variable *drogas_ilicitas_cat3* represents recoded illicit drug consumption and is of nominal type originating from questions Q6 "Do you use illicit drugs?", Q6.0.1 "Have you ever consumed?" and Q6.0.2 "How long have you stopped consuming?" where the original categories were p3Q6 with 1 for Yes and 2 for No, p3Q6.0.1 with 1 for No and 2 for Yes and p3Q6.0.2 with 1 for less than 1 year and 2 for 1 year or more with the recoding made into 3 groups where 1 means currently consumes or stopped less than 1 year ago, 2 means never consumed and 3 means stopped more than 1 year ago.

estado_civil_cat4

The variable *estado_civil_cat4* represents marital status in four categories being single, married, divorced or separated and widowed and is of nominal type originating from question Q4 "What is your marital status?" from Part III of sociodemographic data where the original categories were 1 for Single, 2 for Married, 3 for Married but legally separated, 4 for Divorced, 5 for Widowed and 6 for Does not know with the recoding made into 4 groups where 1 means Single from original category 1, 2 means Married from original categories 2, 3 means Divorced or Separated from original categories 3 and 4 and 4 means Widowed from original category 5.

tabagismo_cat3

The variable *tabagismo_cat3* represents smoking habit in current category being smokes, never smoked or ex-smoker and is of nominal type originating from question Q5 "Do you smoke?" and sub-questions Q5.0.1 "Have you ever smoked?" and Q5.0.2 "How long have you quit smoking?" where the original categories were Q5 with 1 for Yes and 2 for No, Q5.0.1 with 1 for No and 2 for Yes and Q5.0.2 with 1 for less than 1 year and 2 for 1 year or more with the recoding made into 3 groups where 1 means currently smokes, 2 means never smoked and 3 means ex-smoker.

uma_carga_cat3

The variable *uma_carga_cat3* represents smoking burden in pack-years categorized into tertiles and is of ordinal type originating from question Q5.3 about how many cigarettes per day being calculated through

the formula pack-years equals number of cigarettes per day multiplied by years of consumption divided by 20 being then categorized into tertiles among smokers where 1 means low burden, 2 means moderate burden and 3 means high burden but only defined for smokers and should not impute 0 to non-smokers or ex-smokers marking as outside the domain.

Fumador_passivo_bin

The variable *Fumador_passivo_bin* represents exposure to passive smoking being yes or no and is of binary nominal type originating from question Q5.4 "How often are you indoors where other people smoke?" with responses daily, occasionally or never where the original categories were 1 for Daily, 2 for Occasionally and 3 for Never with the recoding made into 2 groups where 1 means Yes combining Daily and Occasionally and 0 means No for Never.

nivel_af_cardio_cat4

The variable *nivel_af_cardio_cat4* represents level of cardiovascular physical activity in 4 categories and is of ordinal type originating from questions about frequency and duration of moderate and vigorous physical activity with the total minutes of cardio activity per week being calculated and categorized into 4 levels being 1 for very low when less than 10 minutes per week, 2 for low when 10 to 149 minutes per week, 3 for moderate when 150 to 299 minutes per week and 4 for high when 300 or more minutes per week.

af_oms_bin

The variable *af_oms_bin* represents compliance with WHO physical activity recommendations and is of binary nominal type, being calculated from the same origin as *nivel_af_cardio_cat4* but recoded into binary, where 0 means do not meet WHO when less than 150 minutes per week and 1 means meets WHO when 150 or more minutes per week.

sono_2s_cat3

The variable *sono_2s_cat3* represents overall sleep quality in the last two weeks and is of ordinal type. It originates from two questions. Q7 asks: "In the last 2 weeks, how often have you had sleep problems (difficulty falling asleep, little sleep or too much sleep)?" Q7.2 asks: "In the last 2 weeks, how often have you got restful sleep?" The original response categories for both questions were: 1 for every day, 2 for most part of the days, 3 for some days, 4 for a few days, and 5 for never.

Two intermediate variables were created during recoding. The variable sono_prob is set equal to Q7 and is maintained as is, since high values already indicate fewer sleep problems. The variable sono_desc is calculated as 6 minus Q7.2, reversing the scale so that high values indicate more restful sleep. A composite score, sono_score, is then calculated as the sum of sono_prob and sono_desc. This score ranges from 2 to 10, with higher values corresponding to better overall sleep quality: a score of 2 represents the worst possible sleep, and a score of 10 represents the best possible sleep.

This continuous score can also be categorised into sono_cat. A score from 2 to 4 indicates poor sleep quality. A score from 5 to 7 indicates moderate sleep quality. A score from 8 to 10 indicates good sleep quality.

sono_med_2s_bin

The variable *sono_med_2s_bin* represents use of sleep medication in the last 2 weeks being yes or no and is of binary nominal type originating from question Q7.1 "In the last 2 weeks, how often have you needed to take sleeping medication?" where the original categories were 1 for Every day, 2 for Most part of the days, 3 for Some days, 4 for A few days and 5 for Never with the recoding made into binary rule where 1 means Yes for responses 1 to 4 representing need for medication at any frequency and 0 means No for response 5 representing never needed.

ecras_risco_cat3

The variable *ecras_risco_cat3* represents average daily screen time at a behavioural risk level and is of ordinal type. It originates from two questions. Q8 asks: "In a normal week, how many days do you perform screen activities?" Q8.1 asks: "And on a normal day, how much time do you use for screen activities?"

The original response categories were as follows. For Q8, the number of days per week ranged from 0 to 7. For Q8.1, average daily duration was recorded as: 1 for less than 10 minutes, 2 for 10 to 29 minutes, 3 for 30 to 59 minutes, 4 for 1 to less than 2 hours, 5 for 2 to less than 3 hours, and 6 for 3 hours or more.

Average daily screen time, in minutes per day, was estimated from these two questions. This value was then converted into hours per week. Based on these continuous values, a daily risk level variable was created with three ordered categories. A value of 1 indicates low consumption, corresponding to less than 1 hour per day. A value of 2 indicates medium consumption, corresponding to 1 to 2 hours per day. A value of 3 indicates high consumption, corresponding to more than 2 hours per day.

This variable expresses the degree of daily exposure to screens, taking into account both weekly frequency and average daily duration. Higher values indicate greater exposure and, consequently, greater risk associated with sedentary behaviour, sleep disturbances, and attention deficits.

ecras_interf_2s_cat5

The variable *ecras_interf_2s_cat5* represents degree of interference of screen activities in daily tasks on a scale from 1 to 5 where higher means worse and is of ordinal type originating from question Q8.3 "In the last 2 weeks, how often did you do less than you wanted in your work or daily activities because you were doing screen activities?" where the original categories were 1 for Every day, 2 for Most part of the days, 3 for Some days, 4 for A few days and 5 for Never with the recoding made into 5-level reversed rule where 1 means minimal interference, 2 means little interference, 3 means some interference, 4 means much interference and 5 means maximum interference with the scale being reversed so that higher values represent greater negative impact of screen time on daily activities.

setor_atividade_cat4

The variable *setor_atividade_cat4* represents main professional activity sector being primary, secondary, tertiary or does not work and is of nominal type originating from question Q7 "What is your main profession?" from Part III of sociodemographic data where the original categories were open response subsequently classified into three economic activity sectors with cases "Does not have a profession" or "Does not know" considered "Does not work" and the recoding made into 4 groups where 1 means primary sector including agriculture, fishing, forestry and extraction, 2 means secondary sector including industry, construction, energy and manufacturing, 3 means tertiary sector including commerce, services, education, health and administration and 4 means does not work including without profession, retired, student, unemployed and domestic.

nuts2_cat5

The variable *nuts2_cat5* represents the NUTS II region of Portugal in 5 categories and is of nominal type, originating from the question about municipality of residence, which was subsequently converted into NUTS II region, with the categories being North, Centre, Lisbon Metropolitan Area, Alentejo and Algarve.

3.5.2. Reduced Variables

In parallel to the analysis of the full variable set described in Section 3.4.1, a reduced variable set was constructed with the objective of evaluating whether a more parsimonious model — using only information that is typically available at the beginning of a primary care consultation — could achieve comparable predictive performance. The composition of this reduced set was validated by professionals of General and Family Medicine, who reviewed the proposed variables for clinical relevance and routine availability in family practice. The rationale for this exercise was twofold: first, to assess the marginal contribution of the more detailed behavioural and clinical variables present in the full set; second, to explore the feasibility of a simpler model that could be applied in settings where extensive lifestyle and health information is not immediately accessible.

Most variables in the reduced set were retained with the same recoding rules described in Section 3.4.1, namely: *sexo_cat* (sex), *idade_cat14* (age), *estado_civil_cat4* (marital status), *escolaridade_cat4* (education level), *imc_cat6* (body mass index), *egs_cat3* (self-perceived health status), dores_osteo_musc_12m_bin (musculoskeletal pain), and *alimentacao_saudavel_bin* (the target variable). *Nationality* (Nacionalidade) was added as a binary nominal variable indicating Portuguese or foreign nationality. Eight additional variables were either aggregated from items already used in Section 3.4.1 or recorded differently for the purposes of the reduced set, as described below.

3.5.2.1. Cardiovascular diseases (12m)

The variable *doencas_cardio* represents the presence of cardiovascular disease in the last 12 months, being yes or no, and is of binary nominal type, originating from the self-reported health problems question, which asked participants to indicate whether they had experienced any of the following conditions in the preceding 12 months: hypertension, stroke, hypercholesterolemia, or cardiac disease. The re-coding was made into 2 groups, where 1 means Yes if at least one of these conditions was reported and 0 means No if none were reported.

3.5.2.2. Mental health problems (12m)

The variable *doencas_foro_mental* represents the presence of mental health conditions in the last 12 months, being yes or no, and is of binary nominal type originating from the same self-reported health problems question, covering anxiety and depression. The re-coding was made into 2 groups, where 1 means Yes if at least one of these conditions was reported and 0 means No if neither was reported.

3.5.2.3. Metabolic diseases (12m)

The variable *doencas_metabolicas* represents the presence of metabolic conditions in the last 12 months, being yes or no, and is of binary nominal type, originating from the same self-reported health problems question, covering diabetes and overweight or obesity as a self-reported condition. The re-coding was made into 2 groups, where 1 means Yes if at least one of these conditions was reported and 0 means No if neither was reported.

3.5.2.4. Current smokers

Current smokers. Binary nominal variable derived from *tabagismo_cat3* (Section 3.4.1), recorded into two categories: 1 for current smokers, 0 for never smokers and ex-smokers combined.

3.5.2.5. Regular physical activity

Regular physical activity was defined as walking or doing any physical activity for five or more days of the week and for at least 30 to 59 min on average.

3.5.2.6. High screen activity

Moderate screen activity was defined as, on a normal day, having had a screen time of fewer than 3 h.

3.5.2.7. Main occupation

The variable was derived from main occupation (question Q7) and recoded into a binary nominal variable distinguishing professionally active from non-active individuals. Participants who were not engaged in any paid professional activity homemakers, unemployed individuals, retirees, and students were classified as non-active, while those engaged in a professional activity were classified as active.

This dataset ends up not being in the final model but was thoughtfully tested during the results section (Chapter 4).

3.6. Bayesian network structure learning

The structural learning of the Bayesian Network was conducted using a score-based approach. Learning the optimal network structure from observational data is a known NP-hard problem, as the number of possible directed acyclic graphs increases super-exponentially with the number of variables. Exhaustive search is therefore computationally infeasible for datasets of moderate dimensionality, and a heuristic optimisation strategy is required.

3.6.1.1. Search algorithm and scoring criterion

The Hill-Climbing algorithm was selected as the structural learning procedure. Hill-Climbing is a greedy search method that iteratively improves a candidate network structure by applying local modifications — namely the addition, deletion, or reversal of edges. At each step, neighbouring structures are evaluated according to a predefined scoring function, and the modification that maximises the score is retained until a local optimum is reached. For model evaluation, the Bayesian Dirichlet equivalent uniform (BDeu) score was employed with the imaginary sample size parameter set to iss = 1. The BDeu score is a fully Bayesian scoring criterion that evaluates the posterior probability of a network structure given the observed data under a Dirichlet prior distribution, and assumes likelihood equivalence, meaning that network structures encoding the same set of conditional independences receive identical scores. Structure selection was performed under a Maximum a Posteriori (MAP) principle: among the candidate structures explored during the search, the network retained is the one maximising the posterior probability given the data and the prior.

3.6.1.2. Cross-validation framework and seed selection.

Structural learning was embedded within a stratified 5-fold cross-validation framework to obtain unbiased estimates of out-of-sample performance. Stratification preserved the proportional distribution of the

binary outcome (*alimentacao_saudavel_bin*) across folds, mitigating bias arising from class imbalance. The random seed for fold partitioning was fixed a priori at seed = 42, The random seed for fold partitioning was fixed a priori at seed = 42, a value adopted by convention in the data science community rather than on any statistical grounds. What matters for reproducibility is that the seed is fixed and reported, so that the fold partition — and every result derived from it — can be regenerated exactly (57) . Fixing it before any model evaluation was deliberate, to prevent post-hoc seed selection, a form of data snooping in which the seed is chosen to optimise observed performance. Since any procedure relying on random partitioning exhibits some seed sensitivity, robustness to this choice was assessed by re-estimating the model under alternative seeds, reported as a sensitivity analysis rather than as a basis for selecting a seed (58). These results are reported in Section 4.9. The decision to fix the seed before any model evaluation was deliberate, intended to prevent post-hoc seed selection — a form of data snooping in which the seed is chosen to optimise observed performance. Robustness to seed choice is examined in the sensitivity analyses described in Section 4.9.

3.6.1.3. Bootstrap-based stability assessment

To distinguish robust dependencies from associations driven by sampling variability, a non-parametric bootstrap procedure with R = 200 resamples was implemented within each training fold. For each resampled dataset, the network structure was re-learned using the same Hill-Climbing algorithm with BDeu scoring described above. For each possible directed edge, the proportion of bootstrap networks in which the arc appeared was computed; this proportion, referred to as arc strength (F), provides an empirical estimate of structural robustness. The choice of R = 200 represents a balance between the reliability of stability estimates and the computational cost of repeated structural learning, given that the procedure is nested within five-fold cross-validation and replicated across multiple sensitivity analyses.

3.6.1.4. Consensus graph construction

The consensus Directed Acyclic Graph (DAG) was constructed by retaining only those arcs whose strength exceeded a predefined threshold. The selection of this threshold requires a deliberate trade-off between structural stability favoured by higher thresholds and structural completeness, essential for capturing conditional dependencies of moderate but substantively meaningful strength. Three threshold values were systematically evaluated: a conservative threshold of $F \geq 0.7$, an intermediate threshold of $F \geq 0.4$, and a more permissive threshold of $F \geq 0.3$. The selection criterion across thresholds prioritised, in order: (i) the existence of a non-empty consensus graph for the target node; (ii) structural stability across folds and seeds; and (iii) predictive performance on the held-out folds as measured by PR-AUC and Brier Skill Score. The empirical comparison across thresholds, the structural and predictive consequences of each, and the final threshold adopted for the model are reported in Chapter 4.

3.6.1.5. Handling of partially directed acyclic graphs (PDAGs)

In score-based bootstrap structure learning, the consensus arcs at a given threshold may form a partially directed acyclic graph (PDAG) rather than a fully directed DAG. This occurs when, for one or more pairs of variables, the bootstrap procedure votes for both possible orientations of an arc at approximately equal frequency (direction score ≈ 0.5), leaving the orientation unresolved. PDAGs cannot be used directly for parameter estimation, since the conditional probability tables require a fully oriented DAG. The standard resolution procedure consisted of applying the *cextend()* extension function of the *bnlearn* package, which attempts to derive a coherent DAG consistent with the partial ordering. When this procedure failed to converge, that is, when the partial ordering admitted no coherent fully directed extension, the affected fold was flagged for further inspection. The set of fold-specific outcomes under this procedure, and the consequent decisions regarding fold inclusion are documented in Section 4.

In summary, the structural learning pipeline combined the Hill-Climbing algorithm, the BDeu scoring criterion with iss = 1, MAP-based structure selection, non-parametric bootstrap stability assessment with R = 200 resamples and stratified 5-fold cross-validation under a pre-specified random seed of 42. Three consensus thresholds were systematically evaluated, with the selection of the final threshold deferred to the empirical assessment reported in Chapter 4.

3.7. Parameter learning

Following the determination of the consensus network structure, the conditional probability tables (CPTs) governing each node were estimated from the observed data. Parameter learning was performed using Maximum A Posteriori (MAP) estimation with a Bayesian Dirichlet prior (method = "bayes", iss = 1), consistent with the scoring criterion employed during structural learning (Section 3.5). The imaginary sample size parameter was set to iss = 1, corresponding to a weakly informative uniform Dirichlet prior sufficient to regularise the probability estimates without imposing substantive prior preferences over the parameter space.

MAP estimation was selected over Maximum Likelihood Estimation (MLE) because it incorporates prior information to regularise probability estimates, which is particularly important when certain combinations of parent configurations are sparsely represented in the data. With a binary target variable of low prevalence (~14%) and direct parents that include the five-level NUTS II region variable, several combinations in the conditional probability table correspond to small or empty cells — most notably for the smallest regional strata (Alentejo, Algarve). In these cells, MLE would produce zero-probability estimates and render the model undefined for inference under those evidence configurations. The MAP estimator with a Dirichlet prior avoids this degeneracy by attributing a small, non-zero posterior probability to all cells, allowing the model to remain well-defined for inference across the full evidence space.

3.7.1. Treatment of missing data

The analytical dataset used for parameter estimation was constructed to contain no missing values in any of the variables included in the network, eliminating the need for expectation-maximisation (EM) procedures during parameter learning. This complete-case structure was achieved through two decisions taken during data preparation.

Exclusion of non-Portuguese participants. Forty-one participants of foreign nationality (4.6% of the original sample) were excluded from the analytical sample, reducing the working dataset from n = 891 to n = 850. This decision was motivated by the small absolute number of non-Portuguese participants and by the substantive heterogeneity of dietary patterns and behavioural determinants that nationality of origin may introduce, which the sample size does not permit modelling adequately. The exclusion was applied before all structural learning and parameter estimation procedures. The recoding of the screen-time variable (*ecras_interf_2s_cat5*) produced missing values for three participants; under the complete-case requirement, these three participants were removed, yielding the final analytical sample of n = 847 in the full variable dataset.

Exclusion of *uma_carga_cat3* (pack-year burden). The variable representing categorised smoking burden in pack-years (*uma_carga_cat3*) was excluded from the analytical variable set. The variable is, by construction, undefined for never-smokers and ex-smokers — together representing 74.1% of the analytical sample, yielding 80.8% structurally missing observations that correspond to individuals for whom the concept of pack-year burden is not applicable rather than to unobserved data. This pattern of missingness is therefore Missing Not At Random (MNAR) in the strict sense that the absence of a value is determined by the variable being outside its conceptual domain.

The exclusion of *uma_carga_cat3* was preceded by an exploratory attempt to retain the information it contains by fusing it with *tabagismo_cat3* into a composite smoking variable that combined smoking status and smoking intensity into a single ordered factor. This fusion was rejected on the basis of the empirical results: the arc strength of the composite variable to the target node dropped substantially relative to the strength of the original *tabagismo_cat3* (from F = 0.655 to F = 0.220), and the predictive metrics of the resulting model worsened across all primary indicators. The structural pathway from smoking status to dietary quality was therefore retained through *tabagismo_cat3* alone, and *uma_carga_cat3* was excluded from the final variable set. The decision and the supporting metrics are documented in Section 3.9.2.

Following these two exclusions, no further missing values remained in the analytical dataset, and parameter estimation was performed directly on the complete-case data of n = 847 participants, due to the additional exclusion of three participants with missing values in the *ecras_interf_2s_cat5* variable during the data audit phase.

3.7.2. Estimation procedure

For each node in the consensus network, a CPT was estimated conditional on its parent set as defined by the F ≥ 0.3 consensus DAG (Section 3.5). The CPTs corresponding to the target node *alimentacao_saudavel_bin* and its two direct parents (*tabagismo_cat3* and *nuts2_cat5*) were extracted and retained for model interpretation, *what-if* inference, and predictive evaluation. These tables represent the core probabilistic machinery of the model, encoding how the posterior probability of healthy eating shifts as a function of the joint state of its direct determinants.

The estimated parameters were used in two distinct downstream procedures: (i) exact probabilistic inference via the *gRain library*, in which the fitted Bayesian Network was compiled into a junction tree to support *what-if* scenario analysis (Section 4.7) and individual-level prediction; and (ii) predictive evaluation within the stratified 5-fold cross-validation framework (Section 3.5), where parameter estimation was performed exclusively on training folds to avoid information leakage into the held-out performance estimates.

3.8. Classification threshold optimisation

For a binary outcome modelled by a Bayesian Network, predictive classification requires a decision threshold that converts the continuous posterior probability P(target = Yes | evidence) into a discrete class assignment. The selection of this threshold is non-trivial in the presence of pronounced class imbalance, and the conventional choice of 0.5, which corresponds to the implicit decision rule of equal misclassification cost under balanced classes, is well known to perform poorly when the positive class is rare (59,60).

3.8.1.1. Rationale for threshold optimisation

Under the binary target *alimentacao_saudavel_bin* with positive class prevalence of approximately 14%, the model assigns posterior probabilities of P(Sim) concentrated in the lower region of the unit interval. The conventional 0.5 cut-off would therefore classify all individuals as negative, producing zero true positives, zero recall, and an F1-score of zero — an outcome that, while mathematically defensible under the equal-cost assumption, conveys no operational information about the minority class. For this reason, the decision threshold was treated as an additional hyperparameter to be optimised within the cross-validation framework, rather than fixed a priori at the conventional 0.5 (61).

3.8.1.2. Selection criterion

Among the candidate criteria for threshold selection in imbalanced binary classification, including the F1-score, the Matthews Correlation Coefficient (MCC), Youden's J statistic, and specificity maximisation under a recall constraint, the F1-score was adopted as the primary optimisation criterion. The F1-score is defined as the harmonic mean of precision and recall, and as such balances the model's ability to identify true positives against its propensity to generate false positives. This balance is appropriate for the substantive setting of this study, in which both the failure to identify individuals with healthy dietary patterns (false negatives) and the over-identification of individuals as having healthy diets when they do not (false positives) carry operational consequences for downstream interpretation and use. The MCC was retained as a secondary criterion for internal cross-validation of the threshold selection, given its documented robustness to class imbalance and its independence from the choice of class as positive or negative.

3.8.1.3. Optimisation procedure

For each cross-validation fold, the optimal threshold was identified by sweeping the predicted probability range from 0.01 to 0.99 in fine increments and selecting the threshold value that maximised the F1-score on the training fold. The optimisation was performed independently within each fold to avoid information leakage from held-out data, and the resulting fold-specific threshold was then applied to the corresponding test fold to produce the binary classifications used for confusion matrix construction and for the computation of all threshold-dependent metrics (precision, recall, F1, MCC, specificity, accuracy). The variability of the optimised threshold across folds was retained as an indicator of the stability of the operating point and is reported alongside the predictive metrics in Chapter 4.

3.8.1.4. Comparison with alternative criteria

As part of the sensitivity analyses described in Section 3.9, the F1-optimised threshold was systematically compared with thresholds obtained under the MCC criterion, under Youden's J statistic, and under a specificity-maximisation rule constrained to a minimum recall floor. The objective of this comparison was to verify the robustness of the operating point chosen for the primary model and to assess whether alternative criteria would have produced materially different trade-offs between sensitivity and specificity. The empirical results of this comparison are documented in Chapter 4.

3.8.1.5. Reporting convention

All threshold-dependent metrics reported in Chapter 4, including the confusion matrices, precision, recall, F1-score, and specificity, were computed under the F1-optimised threshold. Threshold-independent metrics including PR-AUC, ROC-AUC, BSS, Brier Skill Score, and Expected Calibration Error were computed directly from the posterior probabilities without reference to any specific operating point and are therefore not affected by the choice of threshold.

3.9. Validation and comparison with baseline models

Model performance was evaluated within the stratified 5-fold cross-validation framework described in Section 3.5, ensuring that both structural learning and parameter estimation were performed exclusively on training folds, with predictive assessment conducted on held-out folds. This design avoids information leakage between training and evaluation and provides an unbiased estimate of out-of-sample performance.

3.9.1. Handling of fold-specific structural failures

As described in Section 3.5, the bootstrap structure-learning procedure may, in specific cross-validation partitions, produce a partially directed acyclic graph (PDAG) that cannot be coherently extended into a fully directed DAG. In such cases and in the absence of clinically informed orientation constraints capable of resolving the ambiguity at the time of model fitting, the affected fold was excluded from the performance summary, and predictive metrics were reported as the mean across the remaining valid folds. The number of valid folds for each model configuration is reported alongside the corresponding performance estimates in Chapter 4.

3.9.2. Primary and secondary performance metrics

Given the pronounced class imbalance in the outcome variable, with healthy eating as the minority class (prevalence ≈ 14%), standard accuracy was considered inadequate as a primary indicator. The primary evaluation metrics were therefore the *Area Under the Precision-Recall Curve* (PR-AUC) and the *Brier Skill Score* (BSS). PR-AUC reflects the model's ability to identify true positives while minimising false positives and is particularly informative when the positive class is rare. The Brier Skill Score expresses the improvement of the model's probabilistic predictions over a non-informative baseline that predicts the marginal class prevalence for every observation; positive values indicate an improvement in probabilistic prediction; negative values indicate worse-than-baseline calibration. Secondary metrics included ROC-AUC, F1-score, precision, recall, and specificity, computed under the F1-optimised decision threshold described in Section 3.7.

3.9.3. Calibration assessment

Calibration of the predicted posterior probabilities was assessed through two complementary instruments. First, the *Expected Calibration Error* (ECE) was computed by partitioning the predicted probabilities into ten equally spaced bins, computing the absolute difference between the mean predicted probability and the observed empirical frequency in each bin, and aggregating across bins weighted by bin

size. Second, calibration was assessed visually through *reliability diagrams*, which plot the observed frequency of the positive class against the predicted probability within each bin and allow direct inspection of systematic over- or under-confidence in the predictions. ECE values closer to zero indicate better-calibrated predictions; values below 0.05 are generally considered acceptable for population-level applications.

3.9.4. Baseline comparators

To contextualise the performance of the Bayesian Network, two baseline classifiers were trained and evaluated under identical cross-validation conditions, on the same training-test partitions defined by the stratified fold assignment under seed 42:

1. ***Logistic Regression with L1 regularisation*** (LASSO) was selected as the primary comparator, given its widespread use in nutritional epidemiology and its interpretability as a standard reference for binary classification under sociodemographic and behavioural predictors. The regularisation parameter was tuned through internal cross-validation on the training fold.
2. ***Naïve Bayes*** with Laplace smoothing ($\lambda = 1$) was included as a probabilistic baseline, representing the simplest member of the graphical model family. The comparison with Naïve Bayes allows direct assessment of the added value of learning the conditional independence structure from data, as performed by the Bayesian Network, rather than assuming full conditional independence between predictors given the target, as Naïve Bayes does by construction.

For methodological comparability, both baseline classifiers were evaluated using the same metrics, the same F1-based threshold optimisation procedure described in Section 3.7, and the same fold-specific exclusion criteria as the Bayesian Network. Performance comparisons across models focused on PR-AUC, BSS, and the threshold-dependent classification metrics under the F1-optimised operating point. The interpretation of these comparisons is discussed in Chapter 5, with explicit attention to the fact that discriminative parity or modest differences in PR-AUC do not exhaust the methodological value of the Bayesian Network framework, which lies in the learned dependency structure and in the support for probabilistic *what-if* reasoning, neither of which is replicable through logistic regression or Naïve Bayes.

3.10. Sensitivity and robustness analyses

A set of complementary analyses was conducted to assess the stability of the model and its principal structural findings across methodological choices that, although defensible, were not uniquely determined by the data or by a priori clinical reasoning. These analyses were designed to verify whether the dependencies and predictive performance reported for the primary model are robust to the specific configurations adopted, or

whether they are artefacts of those choices. The empirical results of each analysis are reported in Chapter 4; the present section describes only the methodological framework of each.

3.10.1. Sensitivity to arc strength threshold

As described in Section 3.5, three values of the consensus arc strength threshold were systematically evaluated: $F \geq 0.7$, $F \geq 0.4$, and $F \geq 0.3$. For each threshold, the structural composition of the consensus graph, the neighbourhood of the target node, the connectivity of the overall network, and the cross-validated predictive performance were compared. This analysis serves both to justify the threshold ultimately adopted for the primary model and to characterise the dependence of the structural and predictive findings on this methodological choice.

3.10.2. Sensitivity to the variable set

Two variable sets of distinct granularities were estimated and compared: the full variable set (Section 3.4.1), which retains the disaggregated behavioural, clinical, and sociodemographic variables, and the reduced variable set (Section 3.4.2), which aggregates these into composite indicators clinically validated as routinely available at the beginning of a primary care consultation. The comparison evaluates whether a more parsimonious model, using only information accessible in routine practice, achieves predictive performance comparable to that of the full variable set, and characterises the trade-off between parsimony and the granularity of the captured dependencies.

Sensitivity to Variable Composition: Composite Smoking Variable. As an additional structural exploration, the smoking burden variable (*uma_carga_cat3*, in pack-years) was tentatively fused with the smoking status variable (*tabagismo_cat3*) into a single composite ordered factor combining smoking status and smoking intensity. This fusion was motivated by the structurally missing pattern of *uma_carga_cat3*, which is by construction undefined for the 74.1% of the sample composed of never-smokers and ex-smokers (Section 3.6.2). The composite variable was substituted for *tabagismo_cat3* in the structure-learning procedure, and the resulting arc strengths to the target, consensus graph composition, and predictive metrics were compared with those obtained under the original *tabagismo_cat3* configuration.

3.10.3. Sensitivity to discretisation

The robustness of the model to discretisation decisions was examined by testing an alternative cut-point for the daily screen-time exposure variable (*ecras_risco_cat3*), for which a higher threshold (>3 hours/day) was evaluated as an alternative to the primary categorisation (>2 hours/day, three levels: <1h, 1–2h, >2h). This alternative cut-point was explored during model development; the primary >2 hours/day threshold was retained for the final model, as it provided a more even distribution across categories and no

compelling structural or predictive advantage was identified for the higher cut-point. This decision is documented here as part of the robustness assessment rather than as a formally benchmarked comparison.

Sensitivity to Class Imbalance: SMOTENC. The pronounced class imbalance of the target variable (prevalence ≈ 14%) motivated the evaluation of synthetic minority oversampling as a candidate strategy for balancing the training distribution. The Synthetic Minority Over-sampling Technique for Nominal and Continuous features (SMOTENC) was applied exclusively to the training partition of each cross-validation fold, generating synthetic minority-class observations through k-nearest-neighbour interpolation (k = 5) with the over-sampling ratio set to 1 (50/50 class balance in the training data). Two experimental conditions were evaluated: in the first, the test partition retained the original real-world prevalence (14%), reflecting deployment conditions; in the second, the test partition was also synthetically balanced, isolating the effect of class imbalance on discrimination from the effect of prevalence mismatch on calibration. Threshold optimisation for the SMOTENC models used Youden's J statistic, given the balanced training distribution. The structural composition of the resulting Bayesian Networks and their predictive performance were compared against the primary model.

Sensitivity to Bootstrap Reweighting: Weighted Bootstrap. As an alternative to synthetic oversampling, a weighted bootstrap procedure was implemented in which observations of the minority class were assigned higher sampling weights during the bootstrap resampling phase of structure learning (weights ≈ 7 for the positive class, ≈ 1.17 for the negative class, corresponding to the inverse of class prevalence). This procedure preserves the original observations — avoiding the synthetic interpolation problems of SMOTENC for categorical variables — while increasing the effective influence of the minority class on structure estimation. The resulting consensus graphs and predictive metrics were compared with those of the primary model.

3.10.4. Sensitivity to fold stratification seed

Although the random seed for stratified fold partitioning was fixed a priori at seed = 42 to prevent post-hoc selection (Section 3.5), the robustness of the principal structural and predictive findings to this specific partitioning was assessed by re-estimating the model under three alternative seeds (123, 99, and 44). For each alternative seed, the structural composition of the consensus graph, the occurrence of PDAGs across folds, the identification of direct parents of the target, and the predictive performance were evaluated under the same methodological framework as the primary model.

3.10.5. Age-Stratified models

To evaluate whether the dependency structure underlying healthy eating is homogeneous across age groups or whether age-specific patterns are masked in the pooled analysis, separate Bayesian Networks were estimated within three age strata defined a priori on clinical grounds: 20–35 years, 35–65 years, and 65 years

or older. Each stratum-specific model was estimated using an identical procedure to the pooled model: Hill-Climbing structure learning, BDeu scoring (ISS = 1), bootstrap stability assessment (R = 200), a consensus threshold of 0.30 for the averaged network, and 5-fold cross-validation for predictive evaluation. The age variable (`*idade_cat14*`) was excluded from both the pooled and the stratum-specific models, ensuring that all networks were estimated over the same variable set and that any structural differences could be attributed to age stratification rather than to differing model inputs. Structural and predictive findings were then compared across strata.

3.10.6. Tree-Augmented Naïve Bayes (TAN)

As an alternative Bayesian Network architecture, a Tree-Augmented Naïve Bayes classifier was evaluated. TAN extends the Naïve Bayes structure by relaxing the assumption of full conditional independence between predictor variables: the target node remains a parent of all predictors, and each predictor is additionally permitted to depend on one other predictor, structured as a maximum-weight spanning tree derived from pairwise conditional mutual information. In contrast to the sparse, data-driven structure produced by the Hill-Climbing bootstrap procedure, TAN imposes a fixed, globally connected architecture. Four TAN configurations were evaluated, varying by variable set (full versus reduced) and imaginary sample size of the Dirichlet prior (iss = 1 versus iss = 5). All four configurations were evaluated under the same cross-validation framework, threshold optimisation procedure, and performance metrics as the primary model.

3.10.7. Probabilistic sensitivity (*What-if* analysis)

Beyond the methodological sensitivity analyses described above, probabilistic sensitivity analyses in the tradition of *what-if* reasoning were conducted by intervening on the direct parents of the target node and on selected modifiable variables, observing the resulting changes in the posterior probability of healthy eating computed via exact inference on the compiled junction tree. These analyses serve both as a robustness check on the conditional probability structure of the primary model and as a means of quantifying the potential impact of behavioural and contextual configurations within the model's probabilistic framework. The detailed scenarios and resulting posterior probabilities are reported in Chapter 4.

3.11. Ethical considerations

This study was conducted in full compliance with the ethical and legal standards applicable to secondary analysis of anonymised population data. The dataset used in this research was originally collected between January and April 2019 through face-to-face interviews administered at participants' homes as part of a nationally representative survey of the Portuguese population. The original data collection received appropriate ethical approval, and all participants provided informed consent prior to participation. The

questionnaire was anonymous, and no personal identifiers were recorded or made available for the purposes of the present study (5,8).

No new data were collected during this research, and no direct contact with human participants took place. Accordingly, the study does not constitute a primary clinical or interventional investigation and did not require independent ethical review beyond the original approval obtained for the source dataset. All data were handled in aggregate form and processed exclusively at the level of coded, non-identifiable variables. The research was conducted in accordance with the General Data Protection Regulation (GDPR) and with the ethical standards of the Faculty of Medicine of the University of Porto (FMUP).

A further ethical consideration concerns the potential for a predictive model of this nature to reflect and reinforce existing socioeconomic inequalities in access to healthy eating. Variables such as educational attainment and household income are known determinants of dietary quality in the Portuguese population, and their inclusion in the model means that predictions may partly reflect structural disadvantage rather than individual agency. This limitation is acknowledged explicitly, and the model is framed as a tool to support evidence-based public health planning rather than as an instrument for individual-level profiling or stratification. The outputs of this research are intended to inform population-level intervention strategies and should not be applied to individual clinical decision-making without appropriate contextualization and professional judgement.

4. Results

This chapter presents the results of applying Bayesian Networks to the prediction of healthy dietary patterns in the Portuguese adult population (n = 847). Section 4.1 describes the analytical sample. Section 4.2 reports the preliminary results that motivated the F1-based threshold optimisation adopted throughout the analysis. Sections 4.3 and 4.4 present the bootstrap-based structural learning under the reduced and full variable sets, respectively, concluding with the formal declaration of the final model in Section 4.4.6. Section 4.5 describes the structural composition of the final model, including the consensus directed acyclic graph and the conditional probability tables. Section 4.6 reports its predictive performance — discrimination, calibration, confusion matrix, and per-fold results. Section 4.7 presents the *what-if* analysis under single and joint evidence scenarios. Section 4.8 compares the Bayesian Network with Logistic Regression (L1) and Naïve Bayes baselines. The chapter closes with Section 4.9, which reports systematic sensitivity and robustness analyses.

4.1. Sample description

Table 4.1 summarises the sociodemographic composition of the sample as originally recruited (n = 891), prior to the exclusions described in Section 3.6.1. It is presented at this stage to document the population from which the analytical sample was drawn; all subsequent results in this chapter refer to the complete-case analytical dataset of n = 847 participants.

Table 4.1 - Description of the recruited sample, prior to exclusions (n = 891).

		n (%)
Gender	*Female*	478 (53.6)
	Male	413 (46.4)
Age (years)	*20-24*	54 (6.4)
	25-29	51 (6.0)
	30-34	56 (6.6)
	35-39	71 (8.4)
	40-44	83 (9.8)
	45-49	77 (9.1)
	50-54	74 (8.7)
	55-59	75 (8.8)
	60-64	71 (8.4)
	65-69	66 (7.8)

	70-74	55 (6.5)
	75-79	47 (5.5)
	80-84	38 (4.5)
	85+	32 (3.8)
Nationality	*Portuguese*	850 (95.4)
	Foreign	41 (4.6)
Marital status	*Married*	501 (56.2)
	Others	390 (43.8)
Education level	*1st cycle or less*	224 (25.1)
	2nd or 3rd cycle	258 (29.0)
	Secondary	294 (33.0)
	Higher education	115 (12.9)
Main occupation	*Active*	564 (63.3)
	Non-active	327 (36.7)

The sample was balanced with respect to sex, with a slight predominance of women (53.6%), and covered the full adult age range, with 26.7% of participants aged 65 years or over. Most participants were of Portuguese nationality (95.4%) and married (56.2%). Educational attainment was concentrated in the lower and intermediate levels: 54.1% had completed no more than the third cycle of basic education, while 12.9% had completed higher education. Just under two thirds (63.3%) were economically active.

The sex distribution of the analytical sample was close to that of the Portuguese adult population (53.6% women vs 52.2% at national level), and the regional composition broadly followed the national NUTS II structure, in which the North (36.5%) and the Lisbon Metropolitan Area (29.1%) are the most populous regions and the Algarve (4.5%) the least (62).

4.2. Preliminary results and the threshold problem

4.2.1. Initial configurations (default threshold = 0.5)

The initial evaluation of the Bayesian Network was conducted using two editions of the dataset and two consensus graph thresholds, yielding four distinct configurations. In all configurations, the default classification threshold of 0.5 was applied to convert predicted probabilities into binary class labels.

Preliminary configurations were estimated on two editions of the dataset preceding the final analytical sample: Ed.1, based on the reduced 16-variable set, and Ed.2, based on the full 32-variable set. Both correspond to the 891-participant stage, prior to the exclusion of foreign nationals, and are reported in Table 4.2.

The optimal thresholds identified by the F1 sweep ranged from approximately 0.046 (conservative graph) to 0.074 (permissive graph), consistent with the low prevalence of the positive class. Results following threshold optimisation are presented in Table 4.2.

Table 4.2 - Preliminary predictive performance across four configurations (classification threshold = 0.5) (n=891).

Configuration	AUC-ROC	PR-AUC	Brier	Recall	F1	TP
Ed.1 F≥0.7	0.608	0.186	0.121	0.000	0.000	0
Ed.1 F≥0.4	0.682	0.214	0.119	0.000	0.000	0
Ed.2 F≥0.7	0.594	0.163	0.118	0.000	0.000	0
Ed.2 F≥0.4	0.603	0.181	0.119	0.000	0.000	0

Across all configurations, the model failed to identify any true positive case (TP = 0), resulting in Precision, Recall, and F1 scores of zero. Accuracy remained high (approximately 85–86%), attributable to the baseline strategy of predicting the majority class. Despite the absence of correct positive classifications, ranking-based metrics revealed signals above chance: AUC-ROC ranged from 0.594 to 0.682, and PR-AUC from 0.163 to 0.214, both exceeding the prevalence baseline of 0.14.

4.2.2. The threshold problem

The root cause of zero Recall was identified as a mismatch between the classification threshold and the predicted probability distribution. The positive class (*alimentacao_saudavel_bin = 1*) had a prevalence of approximately 14%. Given this low base rate, the Bayesian Network produced predicted probabilities for the positive class in the range 0.06 to 0.18. Even for individuals who genuinely met the healthy eating criteria, the model rarely assigned probabilities exceeding 0.20. Consequently, the 0.5 threshold was never reached, and all individuals were classified as negative.

Importantly, the model was discriminating between classes: the distribution of predicted probabilities for true positives was shifted to the right relative to true negatives. AUC-ROC values above 0.5 confirm this ordering capacity. The problem lay not in the model itself, but in the decision, boundary applied to its output. The solution involves optimising the classification threshold, as described in the following section.

4.2.3. Threshold optimisation with K-Fold validation

To address the threshold problem, a stratified 5-fold cross-validation framework was implemented (5-fold cross-validation) with automatic threshold selection via F1-score maximisation (sweep from 0.01 to 0.99). The structural learning procedure was embedded within each fold, meaning each fold learned its own network structure from the training partition (~80%), ensuring independence between estimation and evaluation. Bootstrap was performed with R = 200 resamples per fold using Hill-Climbing with BDeu (iss = 1). Two consensus thresholds were evaluated: strength ≥ 0.7 (conservative) and strength ≥ 0.4 (permissive). Table 4.3 reports the resulting predictive performance.

Table 4.3 - Predictive performance of the reduced variable set (16 variables) across three consensus arc strength thresholds. Values are mean ± SD across five valid folds (n = 891). PR-AUC baseline = prevalence = 0.139.

Metric	Type	F≥0.7	F≥0.4	F≥0.3
PR-AUC	Primary	0.172 ± 0.006	0.182 ± 0.014	0.193 ± 0.017
Brier Skill Score (BSS)	Primary	0.028 ± 0.009	0.028 ± 0.007	0.026 ± 0.011
Brier score	Primary	0.117 ± 0.002	0.117 ± 0.001	0.116 ± 0.002
AUC-ROC	Secondary	0.605 ± 0.016	0.625 ± 0.026	0.633 ± 0.019
F1-score	Secondary	0.292 ± 0.010	0.300 ± 0.017	0.286 ± 0.019
Precision	Secondary	0.173 ± 0.006	0.183 ± 0.021	0.194 ± 0.022
Recall	Secondary	0.927 ± 0.034	0.879 ± 0.135	0.793 ± 0.151
ECE	Calibration	0.009 ± 0.004	0.014 ± 0.009	0.028 ± 0.018
Classification threshold	—	0.046 ± 0.005	0.074 ± 0.054	0.100 ± 0.072

The prevalence baseline for PR-AUC was 0.14 and for the BSS was 0.120. Both configurations exceeded the baseline on both primary metrics. The positive Brier Skill Score (0.028) indicates that predicted probabilities were more accurate than the naive strategy of always predicting the prevalence. The Expected Calibration Error (ECE) for the conservative graph was 0.009, indicating excellent calibration.

The precision-recall trade-off reflects a deliberate prioritisation of sensitivity. With a threshold of approximately 0.046, the conservative model achieved a mean recall of 0.927, at the cost of a high number of false positives (~551 per fold).

4.3. Reduced variable set analysis

4.3.1. Structural analysis (bootstrap, arc strengths)

The bootstrap structural learning procedure (R = 200, Hill-Climbing, BDeu with iss = 1) yielded arc strength estimates for all possible edges to the target node *alimentacao_saudavel_bin*. Arc strength represents the proportion of bootstrap networks in which a given edge appeared and serves as an empirical measure of structural stability. Table 4.4 presents arc strengths for the reduced variable set.

Table 4.4 - Arc strength from each variable to the target node (16 variables, R = 200). Highlighted: sole variable in the conservative graph (n=891).

Variable	Strength	Direction
Smoker	0.985	0.42
work	0.290	0.16
Nacionalidade	0.200	0.50
Sexo	0.195	0.74
SemAtividadeEcra	0.100	0.50
doencas_foro_mental	0.090	0.36
Dores	0.055	0.05
RegularPhysicalActivity	0.040	0.63
doencas_cardio	0.025	0.00
EGS_rec / IMC / etc.	0.000–0.010	—

Only one variable, Smoker (smoking status; strength = 0.985) exceeded an arc strength of 0.7, making it the sole direct predictor in the conservative consensus graph. Its direction value (0.42) indicated uncertainty

regarding arc orientation, which is expected in unsupervised Bayesian Network learning. The variables *work*, *Nacionalidade* and *Sexo* showed moderate strengths (0.19–0.29) without reaching the 0.7 or 0.4 threshold.

The remaining ten variables showed strengths below 0.1, suggesting their associations with the target were weak.

4.3.2. Predictive performance

The reduced variable model was estimated using a set of 16 aggregated variables. Two consensus graph configurations were evaluated: a conservative threshold ($F \geq 0.7$) and a permissive threshold ($F \geq 0.3$). Under $F \geq 0.7$, consensus graphs contained between 7 and 12 arcs across folds. Under $F \geq 0.3$, graphs were denser, with 21 to 29 arcs per fold.

The classification threshold was optimised by maximising the F1-score on training folds. Under $F \geq 0.7$, the mean optimal threshold was 0.046 (±0.005), and under $F \geq 0.3$ it varied more widely across folds (0.05–0.23), reflecting greater variability in the predicted probability distributions produced by the richer graph structures.

Under $F \geq 0.7$, the model achieved a mean PR-AUC of 0.172 (±0.006), representing a 20% improvement over the prevalence baseline of 0.139, with mean AUC-ROC of 0.605 (±0.016) and Brier Skill Score of 0.028 (±0.009). Calibration was strong (ECE = 0.009 ± 0.004). The model operated in a high-recall regime (recall = 0.927 ± 0.034) at the cost of low precision (0.173 ± 0.006), consistent with the pronounced class imbalance.

The $F \geq 0.3$ configuration improved discrimination across all folds, with mean PR-AUC of 0.193 (±0.017) and mean AUC-ROC of 0.633 (±0.019).

These results indicate that the reduced variable set captures limited predictive signal regardless of threshold and motivated the exploratory extension to the full variable set.

4.4. Full variable set analysis

To investigate whether the variable reduction from approximately 32 to 16 variables had been overly aggressive, an exploratory analysis was conducted using the complete set of original variables. The same bootstrap and cross-validation framework were applied (R = 200, Hill-Climbing, BDeu, stratified 5-fold), with three consensus thresholds: strength ≥ 0.7, ≥ 0.4, and ≥ 0.3.

4.4.1. Conservative graph (Strength ≥ 0.7): model collapse

With all variables included and a threshold of 0.7, no arc to the target node achieved sufficient strength. The node *alimentacao_saudavel_bin* was left without any parent in the consensus graph, effectively isolating it from the rest of the network. As a result, the model predicted the marginal probability (14%) for

every individual. AUC-ROC was exactly 0.500 (equivalent to random chance), PR-AUC equalled the prevalence baseline (0.143), and the Brier Skill Score was zero.

This result demonstrates that with a larger variable space and a sample of n = 847, the bootstrap cannot identify any dependency on the target with sufficient consistency to exceed the 0.7 threshold. Arc strength is distributed across multiple competing variables, none of which achieves the stability required by the conservative criterion. This collapse is illustrated in Figure 4.1.

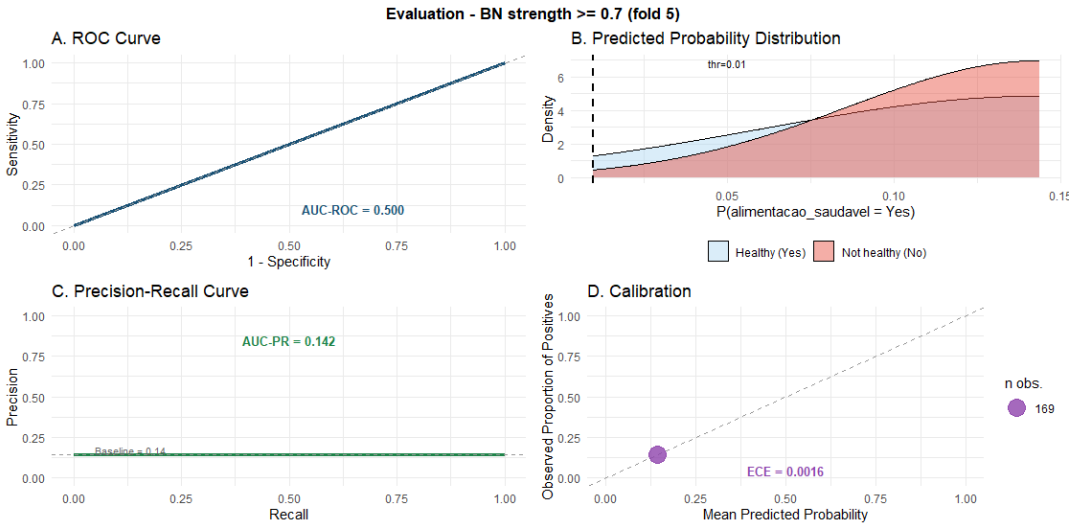

Figure 4.1- BN strength >= 0.7 (fold 5).

4.4.2. Permissive configurations (F≥0.4, F≥0.3, F≥0.2)

A consistent pattern of improvement was observed as the consensus threshold decreased from 0.7 to 0.3. The best overall performance was achieved with strength ≥ 0.3, yielding a PR-AUC of 0.241 (68% above the prevalence baseline), a Brier Skill Score of 0.040, and an F1 of 0.355. The precision-recall trade-off was also more balanced: precision increased to 0.266 with recall of 0.596, compared to the reduced model where precision was 0.173–0.183 with recall of 0.879–0.927.

The only metric that showed slight degradation was calibration (ECE increased from 0.009 in the reduced model to 0.030 in the full model with strength ≥ 0.3), likely reflecting the increased structural complexity. Table 4.5 reports predictive performance across the three consensus arc strength thresholds.

Table 4.5- Predictive performance of the Bayesian Network across three consensus arc strength thresholds. Values are mean ± SD across valid folds. F ≥ 0.7 and F ≥ 0.4: five valid folds (n = 847 evaluated); F ≥ 0.3: four valid folds (n = 678 evaluated, Fold 2 excluded). PR-AUC baseline = prevalence = 0.143.

Metric	Type	F ≥ 0.7	F ≥ 0.4 [a]	F ≥ 0.3	Δ (0.3 *vs* 0.7)
PR-AUC	Primary	0.143 ± 0.002	0.218	0.241 ± 0.050	+68%
Brier Skill Score (BSS)	Primary	0.000 ± 0.000	0.029	0.040 ± 0.038	+0.040 [c]
Brier score	Primary	0.122 ± 0.001	0.119	0.118 ± 0.005	−3%
AUC-ROC	Secondary	0.500 ± 0.000	0.640	0.673 ± 0.056	+35%
F1-score	Secondary	0.250 ± 0.003	0.326	0.355 ± 0.033	+42%
Precision	Secondary	0.143 ± 0.002	0.227	0.265 ± 0.065	+85%
Recall	Secondary	1.000 ± 0.000	0.685	0.596 ± 0.141	−40%
Specificity	Secondary	0.000 ± 0.000	— [b]	0.694 ± 0.152	+0.694 [c]
Log-Loss	Secondary	0.410 ± 0.003	— [b]	0.396 ± 0.020	−3%
ECE	Calibration	0.002 ± 0.001	0.025	0.030 ± 0.020	—
Classification threshold	—	0.010 ± 0.000	0.154	0.180 ± 0.034	+0.170

Exploratory evaluation at F ≥ 0.2 yielded a lower PR-AUC (0.196) compared to F ≥ 0.3 (0.241), confirming that further reductions in the arc strength threshold do not improve predictive performance and likely introduce spurious dependencies that degrade discrimination, as illustrated for Fold 5 in Figure 4.2. The F ≥ 0.3 threshold was therefore retained as the optimal balance between structural completeness and predictive utility.

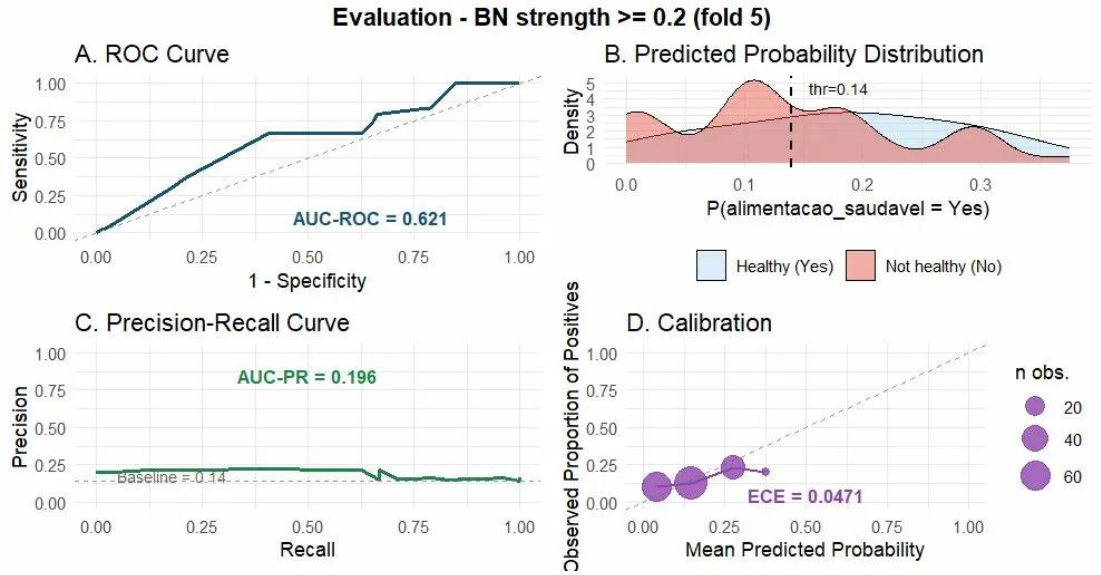

Figure 4.2- BN strength >= 0.2 (fold5).

4.4.3. Arc strengths to the target node

Table 4.6 presents arc strength estimates from each variable to the target node, using the full variable set (32 variables). Only two variables exceeded the 0.4 threshold: ***tabagismo_cat3*** (smoking status; strength =

0.655) and ***nuts2_cat5*** (NUTS II region of residence; strength = 0.555). These were the only direct parents of the target node in both the F ≥ 0.4 and F ≥ 0.3 consensus graphs.

Table 4.6 - Arc strength from each variable to the target node (full variable set, R = 200). Only variables with strength > 0.05 (n=847).

Variable	Strength	Direction	F ≥ 0.3
tabagismo_cat3	0.655	0.28	**Yes**
nuts2_cat5	0.555	0.40	**Yes**
ecras_interf_2s_cat5	0.175	0.84	No
sexo_cat	0.170	0.22	No
dores_osteo_musc_12m_bin	0.125	0.66	No
hipertensao_12m_bin	0.090	0.61	No
diabetes_12m_bin	0.070	0.07	No
ecras_risco_cat3	0.060	1.00	No
asma_dpoc_12m_bin	0.050	0.50	No

The target node had no children in either graph, meaning the Markov Blanket comprised these two variables exclusively. Despite the identical Markov Blanket, predictive performance differed substantially between the two thresholds (Table 4.8).

The consensus DAG (F ≥ 0.3, n=847) contains 43 arcs distributed across 32 variables. The target node *alimentacao_saudavel_bin* has exactly two direct parents, *tabagismo_cat3* (smoking status; arc strength = 0.655) and *nuts2_cat5* (NUTS II region of residence; arc strength = 0.555), which together constitute its complete Markov Blanket.

Beyond these two parents, *fumador_passivo_bin* (passive smoking) is the only additional variable connected to the target's cluster, sharing an arc with *nuts2_cat5*. The orientation of this *nuts2_cat5*–*fumador_passivo_bin* arc was not robustly resolved by the bootstrap (direction score ≈ 0.5) and is among the directionally ambiguous arc pairs identified in the Fold 2 analysis (Section 4.5).

The remaining 28 variables, including the entire cluster of chronic disease variables (*egs_cat3*, *hipertensao_12m_bin*, *diabetes_12m_bin*, *depressao_12m_bin*, and others), sleep, mental health, BMI, and physical activity, form internally coherent subgraphs that are entirely disconnected from the target node. No directed path exists between these clusters and *alimentacao_saudavel_bin*.

4.4.4. Full *vs* reduced variable set

The exploratory model with all variables outperformed the reduced model on every metric. These results suggest that the initial variable reduction to 16 aggregated variables may have been overly aggressive, removing information that, while individually weak, contributed in other ways. Table 4.7 sets out this comparison.

Table 4.7- Comparison between the best exploratory configuration (all variables, F ≥ 0.3 n=678) and the best reduced model configuration (16 variables, F ≥ 0.4 n=891).

Metrics	Reduced Variables F ≥ 0.4	All Variables F ≥ 0.3	Change
PR-AUC	0.182	0.241	**+32%**
Brier Skill	0.028	0.040	**+43%**
F1	0.300	0.355	**+18%**
Precision	0.183	0.266	**+45%**
AUC-ROC	0.625	0.673	**+8%**

4.4.5. Classification threshold

The F ≥ 0.7 configuration produced a model collapse: the target node was left without parents in the consensus graph, resulting in an AUC-ROC of exactly 0.500 and PR-AUC equal to the prevalence baseline. This configuration was retained for comparative purposes but is not considered a valid predictive model.

The F ≥ 0.3 configuration substantially outperformed the collapsed model on all primary and secondary metrics. The mean PR-AUC of 0.241 (±0.050) represents an improvement of 68% above the baseline prevalence of 0.143, confirming that the model captures meaningful signal for identifying healthy eaters. The BSS of 0.118 (±0.005) is below the baseline BSS of 0.122. The AUC-ROC of 0.673 (±0.056) confirms moderate discriminative ability. These results appear in Figure 4.3.

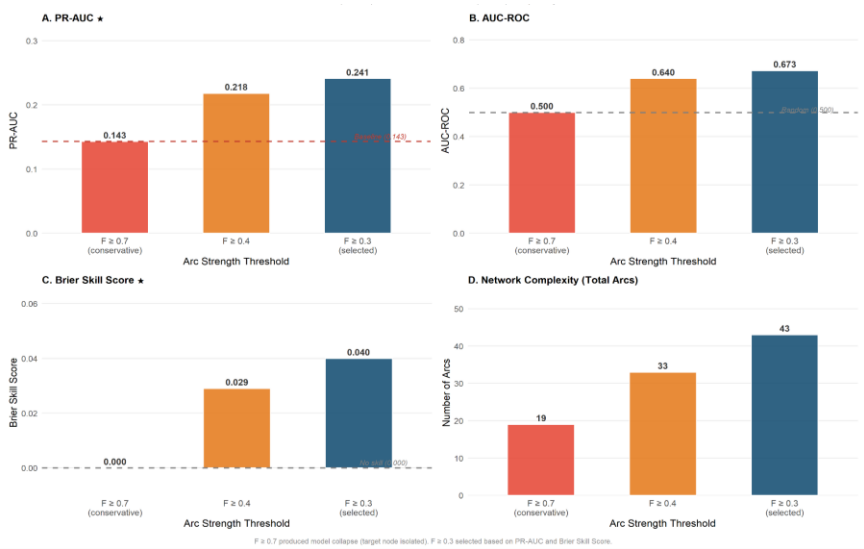

Figure 4.3 - Per-threshold results.

4.4.6. Final model selection

Following the exploratory comparison of the reduced and full variable sets across three arc strength thresholds (Sections 4.3–4.4), and before the detailed presentation of the conditional probability structure and predictive performance, the final model adopted for all subsequent analyses is declared here. This explicit declaration serves to clarify which configuration is being characterised in Sections 4.5 onwards and to consolidate the analytical decisions described in Chapter 3.

Selection criteria. The selection of the final configuration was guided by four criteria, applied in the following order of priority: (i) recovery of a non-empty consensus graph for the target node, ensuring that the model produces non-degenerate predictions; (ii) structural stability of the direct parents of the target across folds and across alternative configurations, ensuring that the principal substantive findings are not artefacts of methodological choices; (iii) predictive performance on held-out folds, evaluated through PR-AUC and Brier Skill Score as the primary metrics appropriate for imbalanced binary classification; and (iv) calibration of the predicted posterior probabilities, assessed through the Expected Calibration Error.

Final configuration. The configuration retained as the final model is the Hill-Climbing bootstrap Bayesian Network estimated on the full variable set (32 variables) under the arc strength threshold $F \geq 0.3$, with parameter estimation by Maximum A Posteriori using a Dirichlet prior (iss = 1), and a classification threshold optimised for the F1-score within each cross-validation fold. The stratified 5-fold cross-validation was performed under a pre-specified seed (seed = 42), with Fold 2 excluded from the performance summary due to the bootstrap procedure producing a partially directed acyclic graph (PDAG) at the $F \geq 0.3$ threshold that could not be coherently extended into a fully oriented DAG. Performance metrics are therefore reported as the mean across the four remaining valid folds (Folds 1, 3, 4, and 5).

Because Fold 2 was excluded from the final model, its aggregated results are pooled over 678 evaluated observations, whereas configurations that retained all five folds are pooled over the full analytical sample (n = 847). Comparisons between them should be read with this difference in mind.

Direct parents of the target under this configuration, two variables emerge as the direct parents of *alimentacao_saudavel_bin* in the consensus graph: *tabagismo_cat3* (smoking status, arc strength F = 0.655) and *nuts2_cat5* (NUTS II region of residence, arc strength F = 0.555). These two arcs are the only ones connecting any candidate predictor directly to the target node, and they remain stable across both the $F \geq 0.4$ and $F \geq 0.3$ thresholds, across all four valid folds, and across the sensitivity analyses reported in Section 4.9. All other predictors are positioned in the network as disconnected nodes.

Justification of the choice. The final configuration was preferred over each of the principal alternatives evaluated for the following reasons. The conservative threshold $F \geq 0.7$ produced an empty graph for the target node and a model that predicted the marginal prevalence ($\approx$ 14%) for every observation, yielding zero discriminative capacity. The intermediate threshold $F \geq 0.4$ recovered the same direct parents but produced lower PR-AUC and Brier Skill Score than $F \geq 0.3$. The further-permissive threshold $F \geq 0.2$ produced a lower

PR-AUC than F ≥ 0.3, confirming that further relaxation introduces dependencies that degrade rather than improve discrimination. The reduced variable set, while more parsimonious and clinically attractive, exhibited inferior predictive performance and less stable structural identification of the direct parents than the full variable set.

4.5. Final model: structural description

4.5.1. Consensus DAG

The consensus directed acyclic graph (DAG) was obtained by aggregating the network structures learned across R= 200 non-parametric bootstrap resamples of the data (n = 847, no missing values). For each resample, a structure was learned by hill-climbing optimisation of the BDeu score (imaginary sample size = 1). For every possible edge, two quantities were recorded: the arc strength F, defined as the proportion of resamples in which the arc appeared, and the probability of each direction conditional on the arc's presence. The average network retained only the arcs whose bootstrap support exceeded the selected threshold, orienting each in its most probable direction. The resulting graph therefore summarises the dependencies most robustly supported by the data, filtering out edges that arise only from sampling variability.

At the selected threshold of F ≥ 0.3, the consensus network comprised 43 arcs. The target node, `*alimentacao_saudavel_bin*` (Healthy Eating), had exactly two direct parents, `*nuts2_cat5*` (NUTS-II Region) and `*tabagismo_cat3*` (Smoking Status) — and no children. This indicates that, conditional on these two variables, healthy eating was rendered independent of the remaining behavioural, clinical-metabolic, mental-health and sociodemographic nodes in the network. Notably, this parent set was stable across thresholds: the same two arcs were recovered at F ≥ 0.3, F ≥ 0.4 and F ≥ 0.5, underscoring the robustness of the dependency. Substantively, a sociodemographic determinant (region of residence) and a behavioural risk factor (smoking status) constitute the immediate Markov-blanket parents through which the network channels predictive information towards the dietary outcome, while all other variables exert at most an indirect influence mediated by these two parents. The resulting consensus structure is shown in Figure 4.4.

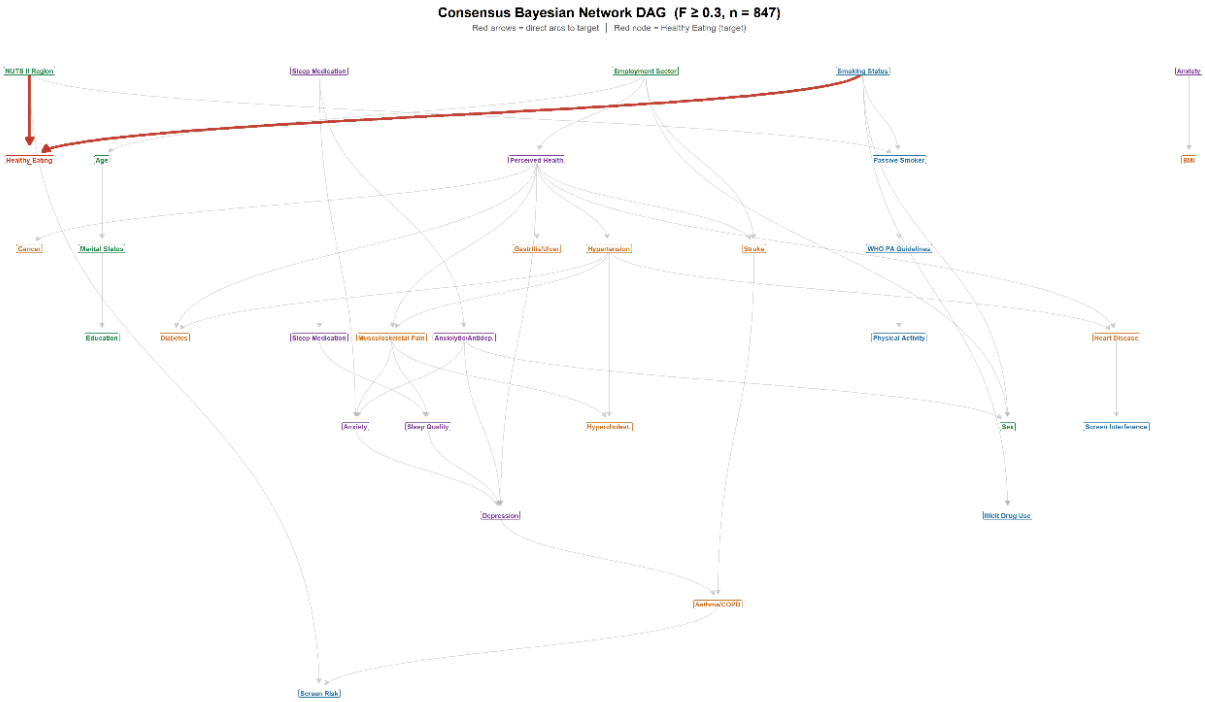

Figure 4.4 - Consensus DAG of the Bayesian Network (arc strength threshold F ≥ 0.3, n = 847). Network complexity across thresholds.

The total number of arcs in the consensus graph increased monotonically as the threshold decreased: 21 arcs at F ≥ 0.7, 31 at F ≥ 0.5, 38 at F ≥ 0.4, and 43 at F ≥ 0.3. While the conservative graph (F ≥ 0.7) was too sparse to support meaningful inference, the target node was isolated, retaining neither parents nor children; the more permissive thresholds recovered moderately stable dependencies that enriched the flow of information through the network. The trade-off between structural robustness and predictive utility therefore favoured the F ≥ 0.3 threshold in this exploratory analysis: it preserved a connected structure around the target without

admitting the unstable, low-support edges that proliferate below this level. All subsequent inference and evaluation were consequently based on the F ≥ 0.3 consensus network.

4.5.2. Conditional probability tables

This section presents the key Conditional Probability Tables (CPTs) estimated from the full dataset (n = 847) using MAP estimation with a Dirichlet prior (iss = 1) and the consensus DAG at arc strength threshold F ≥ 0.3. CPTs quantify the probabilistic relationships encoded in the Bayesian Network structure and are the basis for all predictive inference. CPTs for nodes without parents reflect marginal probability distributions estimated directly from the data.

4.5.3. Target node: *alimentacao_saudavel_bin*

The target node has two direct parents in the consensus DAG: *tabagismo_cat3* and *nuts2_cat5*.

Table 4.8 - P(*alimentacao_saudavel_bin* = Yes | *tabagismo_cat3*, *nuts2_cat5*) (n = 847).

tabagismo_cat3	**Alentejo**	**Algarve**	**AM Lisboa**	**Centro**	**Norte**
Ex-Smoker	0.002	0.007	0.120	0.334	0.200
Smoker	0.002	0.003	0.001	0.040	0.108
Never Smoked	0.168	0.113	0.038	0.299	0.178

The node *tabagismo_cat3* has no parents in the consensus graph. Marginal distribution: Ex-Smoker 17.6%, Smoker 25.9%, Never Smoked 56.5%.

The node *nuts2_cat5* has no parents in the consensus graph. Marginal distribution: North 38.1%, Lisbon Metropolitan Area 26.2%, Centre 23.3%, Alentejo 7.8%, Algarve 4.6%. Figures 4.5 and 4.6 present the bootstrap arc strength distributions obtained for the reduced and the full variable sets, respectively.

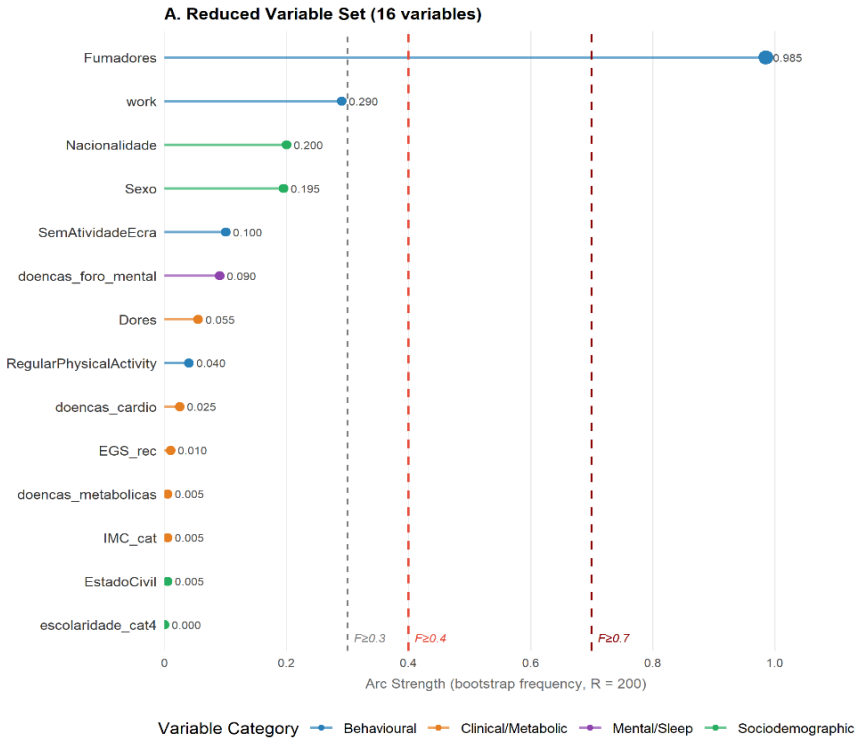

Figure 4.5 - Reduced Variable Set Bootstrap Results.

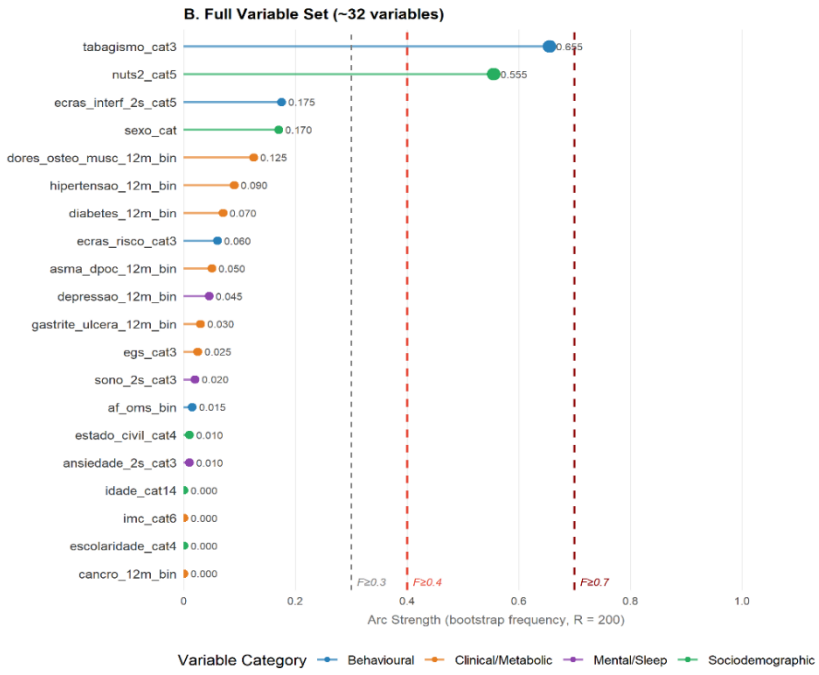

Figure 4.6 - Full Variable Set Bootstrap Result.

4.6. Final model: predictive performance

This section presents the predictive performance of the Bayesian Network evaluated through stratified 5-fold cross-validation. As already mentioned in Chapter 3, the primary evaluation metrics are PR-AUC and the BSS, selected due to class imbalance in the outcome variable (prevalence 14.3%). Calibration is assessed via the Expected Calibration Error (ECE). All results were computed under an optimised classification threshold identified by F1-score.

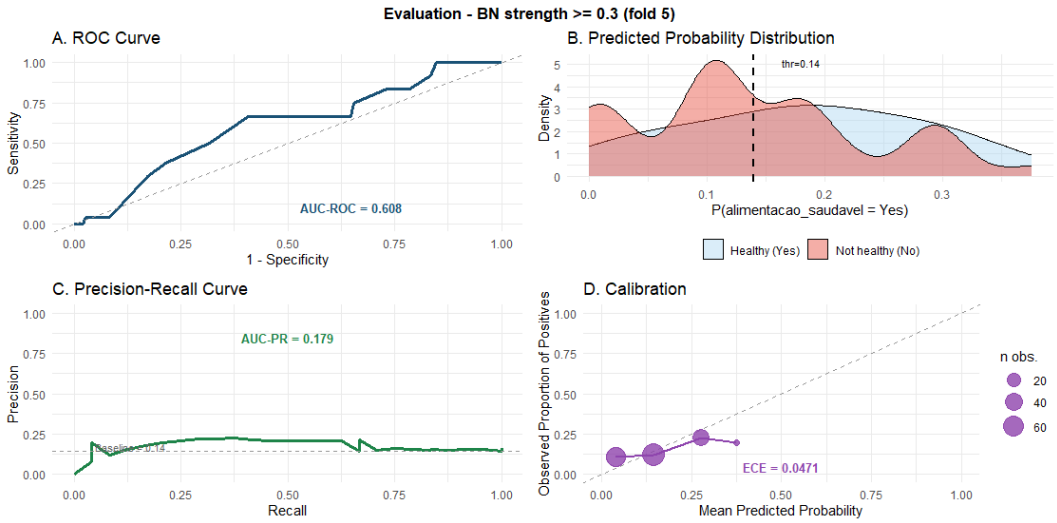

Figure 4.7 - Evaluation panel for the BN model (F ≥ 0.3, Fold 5).

4.6.1. Discrimination

Discriminative performance was summarised by two metrics: the area under the Precision–Recall curve (PR-AUC), the primary metric given the class imbalance, and the area under the ROC curve (ROC-AUC) as a secondary, threshold-independent measure (Figure 4.7, panels A and C). Averaged across the four valid folds, the model achieved a mean PR-AUC of 0.241 (±0.050) and a mean ROC-AUC of 0.673 (±0.056).

Because PR-AUC is bounded below by the positive-class prevalence (0.143), this value represents a 68% relative improvement over the no-information baseline, confirming that the network recovers genuine predictive signal for the healthy-eating class rather than reproducing the marginal rate. Discrimination was consistent in direction across partitions, with per-fold PR-AUC ranging from 0.179 to 0.296 and ROC-AUC from 0.608 to 0.723 (Table 4.11); every fold exceeded its prevalence baseline. The moderate ROC-AUC means that, for roughly two-thirds of randomly paired cases, the model assigns the higher posterior probability to the healthy-eating individual, while the comparatively low PR-AUC reflects the intrinsic difficulty of attaining high precision for a minority class with a 14% base rate.

4.6.2. Calibration

Calibration was assessed using the ECE, computed over 10 equally spaced probability bins. The F ≥ 0.3 model achieved a mean ECE of 0.030 (±0.020), indicating that predicted probabilities deviate from observed proportions by approximately 3 percentage points on average. The higher ECE relative to the reduced model (ECE = 0.009) reflects the increased structural complexity of the full-variable network — a known

degradation that does not negate the substantially improved discriminative performance. The corresponding calibration curve is shown in Figure 4.8.

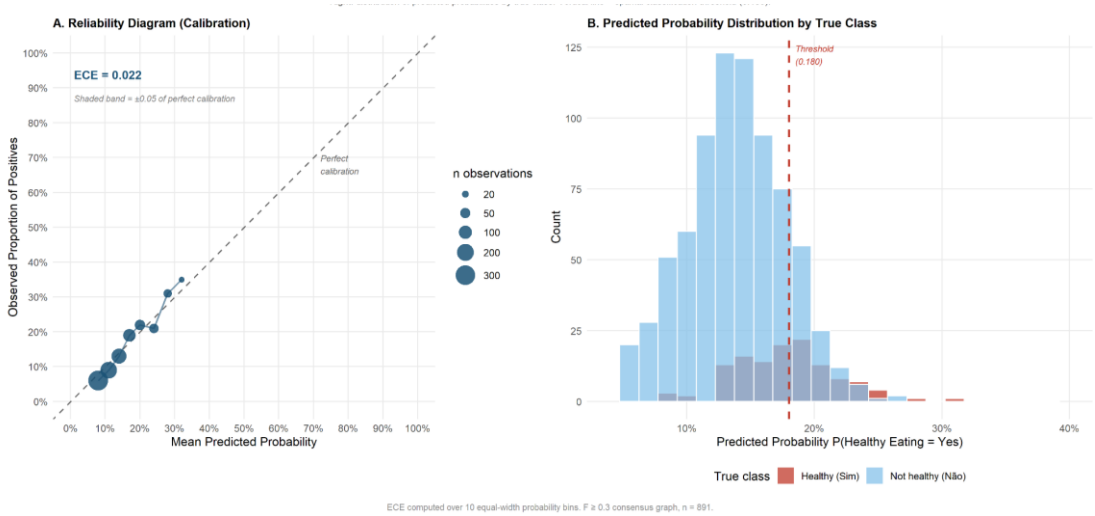

Figure 4.8 - Calibration assessment.

4.6.3. Confusion matrix

Classification performance at the F1-optimised threshold, pooled over the test partitions of the four valid folds (n = 678), is summarised in Table 4.9.

Table 4.9 - Confusion Matrix (n = 678).

	Predicted No	**Predicted Yes**
Real No	403	178
Real Yes	39	58

For the positive class (healthy eating), the pooled matrix yields a recall of 0.598, a precision of 0.246 and an F1 of 0.347 (per-fold means in Table 4.12). Reframed around the clinically relevant class (inadequate diet), the model correctly identified 403 of 581 inadequate-diet respondents (recall 69.4%) at a precision of 91.2%. The dominant residual error is the 178 inadequate-diet respondents classified as healthy eaters, a direct consequence of tuning the threshold to favour sensitivity over precision in the minority class.

4.6.4. Per-Fold results

Table 4.10 - Per-fold cross-validation results for the F ≥ 0.3 configuration. Fold 2 excluded due to bootstrap failure. Means computed over four valid folds (n = 678 evaluated, Fold 2 excluded).

Fold	PR-AUC	Brier	BSS	ROC-AUC	F1	Prec.	Recall	Thr.
1	0.225	0.123	0.017	0.645	0.342	0.221	0.760	0.17
2	—	—	—	—	—	—	—	—
3	0.296	0.112	0.083	0.723	0.400	0.355	0.458	0.22
4	0.262	0.115	0.059	0.716	0.353	0.273	0.500	0.19
5	0.179	0.120	0.001	0.608	0.323	0.213	0.667	0.14
Mean	**0.241**	**0.118**	**0.040**	**0.673**	**0.355**	**0.265**	**0.596**	**0.180**

Per-fold results, reported in Table 4.10, reveal moderate variability, particularly in PR-AUC (range: 0.179–0.296) and recall (range: 0.458–0.760). This variability is expected given the small number of positive cases per fold. However, all partitions produced a PR-AUC above the baseline prevalence (0.143), confirming that the model extracts predictive signal in a consistent manner.

During cross-validation, the bootstrap procedure for Fold 2 produced a partially directed acyclic graph (PDAG) at the $F \geq 0.3$ arc strength threshold that could not be consistently extended into a fully directed DAG. Inspection of the bootstrap results revealed five arc pairs with direction scores approximating 0.5 — namely *hipertensao_12m_bin* ↔ *doenca_cardiaca_12m_bin, idade_cat14* ↔ *setor_atividade_cat4, nuts2_cat5* ↔ *fumador_passivo_bin, ansiedade_12m_bin* ↔ *depressao_12m_bin, and imc_cat6* ↔ *ansiedade_2s_cat3* — indicating that the specific training partition of Fold 2 provided insufficient evidence to resolve the orientation of these arcs.

A technical workaround was available: the undirected arcs could be forcibly oriented (for instance, via a consistent extension of the CPDAG, as implemented in *cextend*), arbitrarily selecting one member of the Markov-equivalence class. This would have allowed maximum-a-posteriori parameter estimation to proceed and prevented the fold from being excluded. This solution was deliberately not adopted, as it is methodologically unsound: since the orientation would be chosen essentially at random, with no support from the data, imposing it would amount to fabricating a directional (and potentially causal) dependency that the analysis cannot justify. A methodologically correct resolution would require external domain knowledge to orient these arcs — specifically, expert input from a general and family medicine clinician, who could specify clinically plausible directions (e.g., as a whitelist/blacklist of arcs constraining the structure-learning step). In the absence of such input, excluding the fold was preferred over imposing an unsupported orientation, so that no edge direction reported in this work rests on an arbitrary choice. As a result, Fold 2 was excluded from the $F \geq 0.3$ performance summary, and mean metrics are reported over the four remaining valid folds.

Per-fold results for F ≥ 0.7 are not presented because the structural collapse of the model produces near-constant values across folds, as shown by the near-zero standard deviations reported in Table 4.5. Figure 4.9 compares predictive performance across folds.

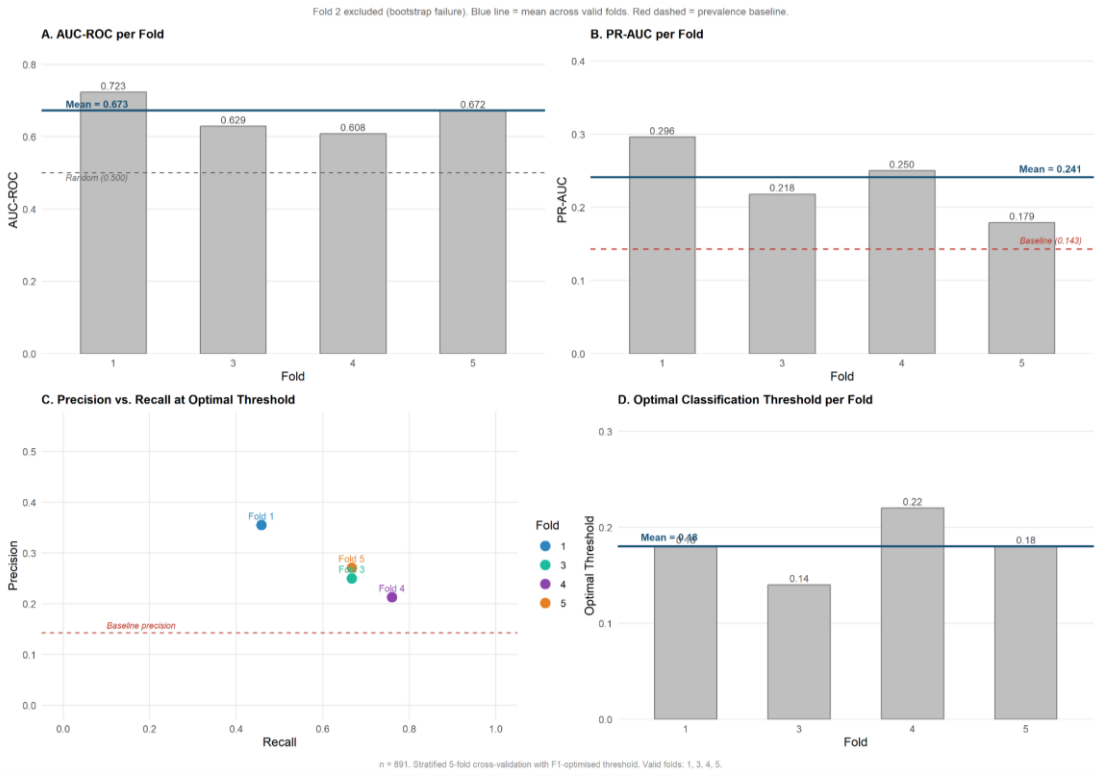

Figure 4.9 - Comparison between folds.

4.7. *What-if* Analysis

Given that the consensus graph identified smoking status (*tabagismo_cat3*) and NUTS II region of residence (*nuts2_cat5*) as the only direct parents of the target node, these two variables constitute the primary levers for *what-if* simulation. Interventions are expressed as conditional probability queries of the form P(*alimentacao_saudavel_bin* = Yes | evidence), where evidence specifies the state of one or both predictors. All posterior probabilities were computed via exact inference using the compiled *gRain* junction tree.

Table 4.11 - *What-if* scenario analysis: posterior probabilities of healthy eating under single and combined evidence conditions. Δ = change relative to marginal probability (0.142). Values derived from exact inference on the compiled *gRain* junction tree (F ≥ 0.3) (n = 847).

Scenario	P(Healthy Eating)	Δ vs. Marginal	Δ (%)	Direction
Marginal (no evidence)	0.142	—	—	—
Smoking: Never smoked	0.173	+0.031	+21.8%	Protective ↑
Smoking: Current smoker	0.045	−0.097	−68.4%	Adverse ↓
Smoking: Ex-smoker	0.188	+0.046	+32.3%	Protective ↑
Region: Norte	0.168	+0.026	+18.1%	Protective ↑
Region: Centro	0.225	+0.083	+58.3%	Protective ↑
Region: AM Lisboa	0.063	−0.079	−55.7%	Adverse ↓
Region: Alentejo	0.088	−0.054	−38.0%	Adverse ↓
Region: Algarve	0.054	−0.088	−62.1%	Adverse ↓
Best case: ex-smoker + Centro	**0.334**	**+0.192**	**+135.2%**	**Protective ↑**
Worst case: Current smoker + Algarve	**0.003**	**−0.139**	**−97.9%**	**Adverse ↓**

Queries involving the combination of smoking status (particularly active smokers) with small regional strata — namely Alentejo and Algarve — should be interpreted with caution. The corresponding cells in the conditional probability table for *alimentacao_saudavel_bin* contain zero or near-zero observed counts (n ≤ 1), meaning that the estimated probabilities reflect the Bayesian prior (iss = 1) rather than empirical data. Results for these combinations are therefore unreliable and are flagged accordingly.

Table 4.12 - Number of cases in each combination of the *what-if* class (n=847).

Healthy eating	Smoking status	Region	n
Yes	Ex-Smoker	Alentejo	**0**
Yes	Current smoker	Alentejo	**0**
Yes	Never smoked	Alentejo	4

Yes	Ex-Smoker	Algarve	**0**
Yes	Current smoker	Algarve	**0**
Yes	Never smoked	Algarve	2
Yes	Current smoker	AM Lisboa	**0**
Yes	Current smoker	Centro	1

4.7.1. Evaluated scenarios

Conditioning the network on smoking status alone produces the following posterior probabilities, computed via exact inference using the compiled *gRain* junction tree:

Never smoked: P(Healthy Eating) = 0.173, representing a +21.8% increase over the marginal.

Current smoker: P(Healthy Eating) = 0.045, representing a −68.4% reduction relative to the marginal.

Ex-smoker: P(Healthy Eating) = 0.188, representing a +32.3% increase over the marginal. Notably, ex-smokers exhibit the highest posterior probability of healthy eating among the three smoking categories — higher even than never-smokers (0.173).

Conditioning on the NUTS II region of residence reveals substantial geographic heterogeneity in the posterior probability of healthy eating, independent of smoking status:

Centro (+58.3%): The highest regional probability (0.225) suggests that individuals residing in the Centro region are substantially more likely to maintain healthy dietary patterns.

Norte (+18.1%): The Norte region (0.168) shows a moderate positive effect, consistent with its large share of the sample (38.1%) and the heterogeneous dietary practices.

Lisbon Metropolitan Area (−55.7%): This region (0.063) shows a markedly lower posterior probability.

Alentejo (−38.0%) and Algarve (−62.1%): These regions show substantially lower posterior probabilities (0.088 and 0.054, respectively). Given their small sample sizes (7.8% and 4.6% of the analytical sample, respectively), these estimates carry greater uncertainty and should be interpreted with caution.

Best case — ex-smoker in Centro approached one-third probability (P=0.334, +135.2%) increase over the marginal. This profile approaches approximately one-third probability of healthy eating, representing the most favourable combination of the two directly modelled determinants within this population. The representativeness of this specific cell is discussed below (see “Robust patterns despite limitations”), where the highest empirically grounded profile is identified.

Worst case — Current smoker + Algarve: P(Healthy Eating) = 0.003, a −97.9% reduction. This near-zero probability reflects the convergence of the two most adverse states across both direct predictors. While this result is mathematically coherent with the CPT structure, its extreme magnitude warrants caution given the very small cell count for this combination in the original dataset. Figure 4.10 summarises the impact of each intervention scenario on the probability of healthy eating.

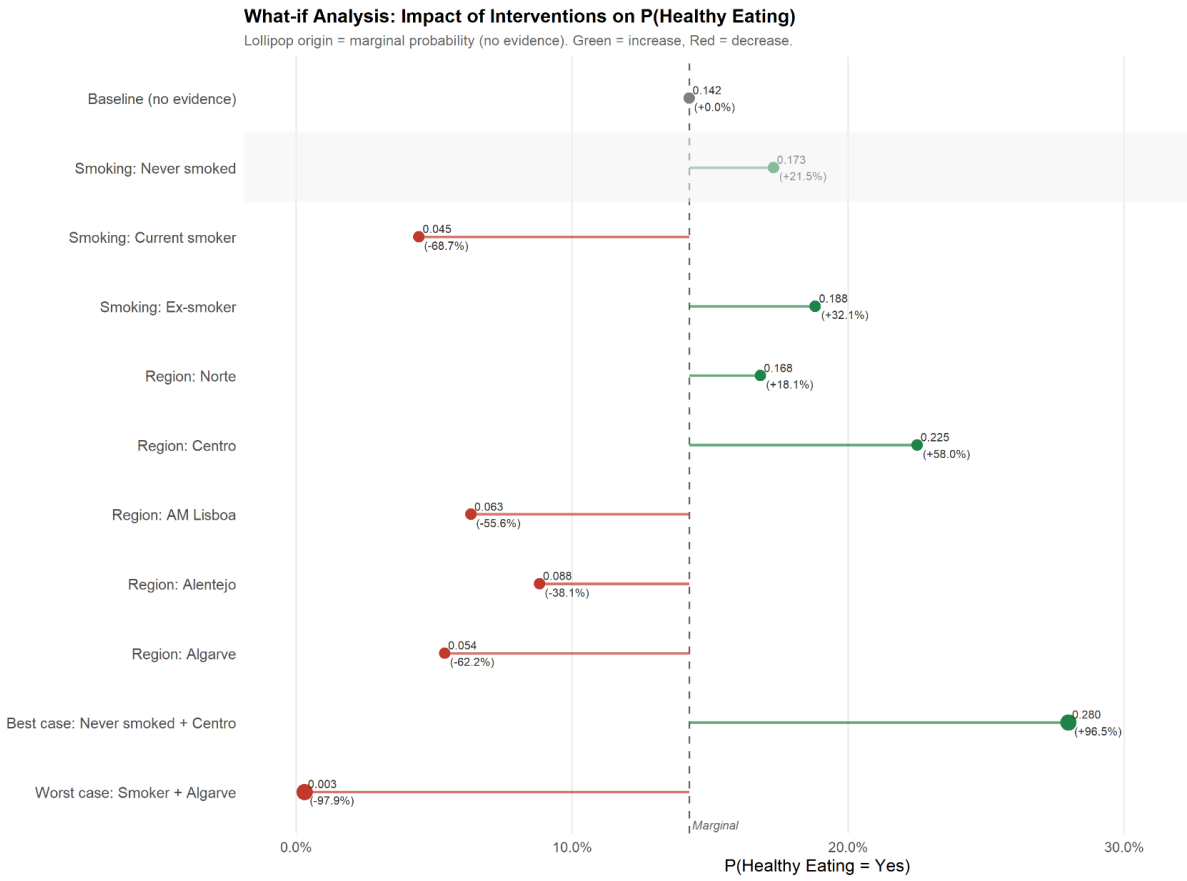

Figure 4.10 - *What-if* Analysis: Impact of interventions on P(Healthy Eating).

4.8. Comparison with baseline classifiers

To contextualise the predictive performance of the Bayesian Network within the landscape of standard classifiers used in nutritional epidemiology, the F ≥ 0.3 BN model was compared against two baseline models trained and evaluated under identical cross-validation conditions: Logistic Regression with L1 regularisation (LASSO) and Naïve Bayes with Laplace smoothing (λ = 1). Both baselines were fitted on the same stratified

5-fold splits, with threshold optimisation via F1-score maximisation applied consistently across all models. Table 4.13 reports this comparison.

Table 4.13 - Predictive performance comparison across models (BN F ≥ 0.3, Logistic Regression L1, Naïve Bayes). All models evaluated on the same stratified **4**-fold cross-validation splits with F1-optimised threshold. Primary metrics (PR-AUC, Brier Skill) in bold, n = 678, 4 valid folds).

Model	PR-AUC	AUC-ROC	BSS	Brier Skill	F1	Precision	Recall
Bayesian Network (F ≥ 0.3)	0.241	0.673	0.118	0.040	0.355	0.265	0.596
Logistic Regression (L1)	0.293	0.712	0.114	0.068	0.383	0.313	0.555
Naïve Bayes	0.304	0.699	0.120	0.018	0.371	0.410	0.522

4.8.1. Discriminative performance

On the primary discrimination metric (PR-AUC), the Logistic Regression (0.293) and Naïve Bayes (0.304) both outperformed the Bayesian Network (0.241). Similarly, both baselines achieved higher AUC-ROC values (LR: 0.712; NB: 0.699) compared to the BN (0.673). This result is noteworthy and requires careful interpretation: it does not imply that the BN is an inferior model for the purposes of this study but rather reflects the different objectives of the three modelling approaches.

Logistic Regression, as a discriminative model, is optimised directly for the classification boundary between positive and negative classes and benefits from the rich feature representation afforded by one-hot encoding of categorical variables. Naïve Bayes, despite its strong conditional independence assumption, performs competitively in high-dimensional discrete settings where the independence assumption is approximately satisfied. Both models are, by construction, optimised for predictive accuracy. The Bayesian Network, in contrast, is a generative model that learns the full joint distribution among all variables — a richer objective that comes at the cost of some discriminative efficiency.

4.8.2. Probabilistic calibration

On the Brier Skill Score, which evaluates the accuracy of predicted probabilities rather than binary classification, the Bayesian Network (0.040) substantially outperforms Naïve Bayes (0.018), while Logistic Regression achieves the highest score (0.068). The Naïve Bayes result is particularly striking: despite its higher PR-AUC, it produces substantially less well-calibrated probability estimates, likely due to the violation of the conditional independence assumption in the presence of correlated health and behavioural variables. BN's superior calibration relative to NB suggests that the learned dependency structure, even if imperfect, provides meaningful regularisation of probability estimates that the independence assumption cannot.

4.8.3. The value of the Bayesian Network beyond classification

The comparison reported here must be situated within the broader research objectives of this study. The research question is not *"which model best classifies individuals as healthy or unhealthy eaters?"* but rather *"how can Bayesian Networks contribute by describing dependencies and predicting healthy eating in the Portuguese population?"* (Section 1.1). In this context, the discriminative performance gap between the BN and the baselines is a secondary concern; the primary contribution of the BN lies in the learned dependency structure.

Three specific contributions are not replicable by logistic regression or Naïve Bayes:

First, the BN identifies the conditional independence structure of the data: BMI, age, and education have zero direct arc strength to healthy eating. Logistic regression cannot distinguish direct from indirect effects — it estimates marginal associations conditional on all covariates, which conflates these two types of influence.

Second, the BN supports probabilistic what-if reasoning (Section 4.7). The posterior probability of healthy eating can be computed for any combination of observed evidence, enabling scenario-based planning that is not available from regression coefficients or Naïve Bayes probability tables. Figure 4.11 presents this comparison graphically.



Figure 4.11 - Predictive performance: Bayesian Network *vs* baseline classifiers.

4.9. Sensitivity and robustness analyses

4.9.1. Age-stratified Bayesian network models

To examine whether the determinants of healthy eating behaviour differ across the life course, three separate Bayesian Network models were estimated for distinct age strata: young adults (20–35 years), middle-aged adults (35–65 years), and older adults (65 years and above). Age stratification allows each model to reflect the unique epidemiological and behavioural context of its target population, avoiding the masking of age-specific effects that would occur in a single pooled model.

Each stratum-specific model followed the same estimation pipeline described in Section 4.9.1: bootstrap hill-climbing structure learning (R = 200 resamples, BDe score, iss = 1), consensus graph construction (arc strength threshold = 0.50), MAP parameter estimation with Dirichlet priors, and performance evaluation via stratified 5-fold cross-validation with threshold optimisation. The target variable in all models was *alimentacao_saudavel_bin* (healthy eating: Yes/No).

4.9.1.1. Variable filtering by age stratum

An important methodological consequence of age stratification is the differential availability of variables across strata. Several binary health indicators exhibited near-zero or zero prevalence within specific

age groups, rendering them uninformative for structure learning. Table 4.14 identifies the variables excluded per stratum.

Table 4.14 - Variables excluded per age stratum due to insufficient variability (n= 847).

20–35	*avc_12m_bin*, *cancro_12m_bin*, *diabetes_12m_bin*, *doenca_cardiaca_12m_bin*, *asma_dpoc_12m_bin*, *gastrite_ulcera_12m_bin*
35–65	*cancro_12m_bin* (variability dependent on fold partition)
65+	*drogas_ilicitas_cat3*

This pattern is clinically coherent: chronic conditions such as diabetes, cardiovascular disease, and cancer are rare in the 20–35 age group, whereas illicit drug use is effectively absent in adults aged 65 and above. Far from constituting methodological limitations, these exclusions reflect the epidemiological reality of each stratum.

4.9.1.2. Predictive performance by age stratum

Table 4.15- Predictive performance of age-stratified Bayesian Network models (cross-validation).

Age Group	n	Prev. (%)	PR-AUC	Baseline	Gain (%)	AUC-ROC	F1	Recall	Precision	Brier	Threshold
20–35	161	11.2 %	0.181	0.112	+61.6%	0.615	0.264	0.800	0.185	0.098	0.094
35–65	450	14.0 %	0.215	0.140	+53.6%	0.567	0.291	0.822	0.182	0.125	0.110
65+	236	14.8 %	0.162	0.148	+9.5%	0.515	0.285	0.971	0.169	0.133	0.040

As reported in Table 4.15, the 35–65 stratum achieved the strongest discriminative performance, with a PR-AUC of 0.215 representing a 53.6% improvement over the baseline and an AUC-ROC of 0.567. This is also the largest stratum and the one with the most stable cross-validation results, consistent with the expectation that structure learning benefits from larger sample sizes.

The 20–35 stratum attained the highest AUC-ROC (0.615), suggesting reasonable rank-order discrimination among young adults. The PR-AUC of 0.181 corresponds to a 61.6% gain over the baseline prevalence of 11.2%. The recall of 0.800 confirms that the model successfully identifies 80% of young adults with healthy eating patterns.

The 65+ stratum presents the weakest overall performance, with a PR-AUC of 0.162 — only 9.5% above baseline — and an AUC-ROC of 0.515, approaching the performance of a random classifier. The

optimised classification threshold of 0.040 is notably low, and the corresponding recall of 0.971 suggests the model classifies nearly all observations as positive, reflecting limited discriminative ability.

4.9.1.3. Comparative interpretation across strata

The divergence in model performance across age strata constitutes a substantive finding. It demonstrates that the determinants of healthy eating are not uniform across the life course, and that a single pooled model would obscure the age-specific structure of these associations.

These findings should be interpreted with caution given the relatively small sample sizes in the 20–35 (n = 161) and 65+ (n = 236) strata, which constrain the stability of bootstrap structure learning and cross-validation estimates. Figure 4.12 summarises the comparative interpretation across strata.

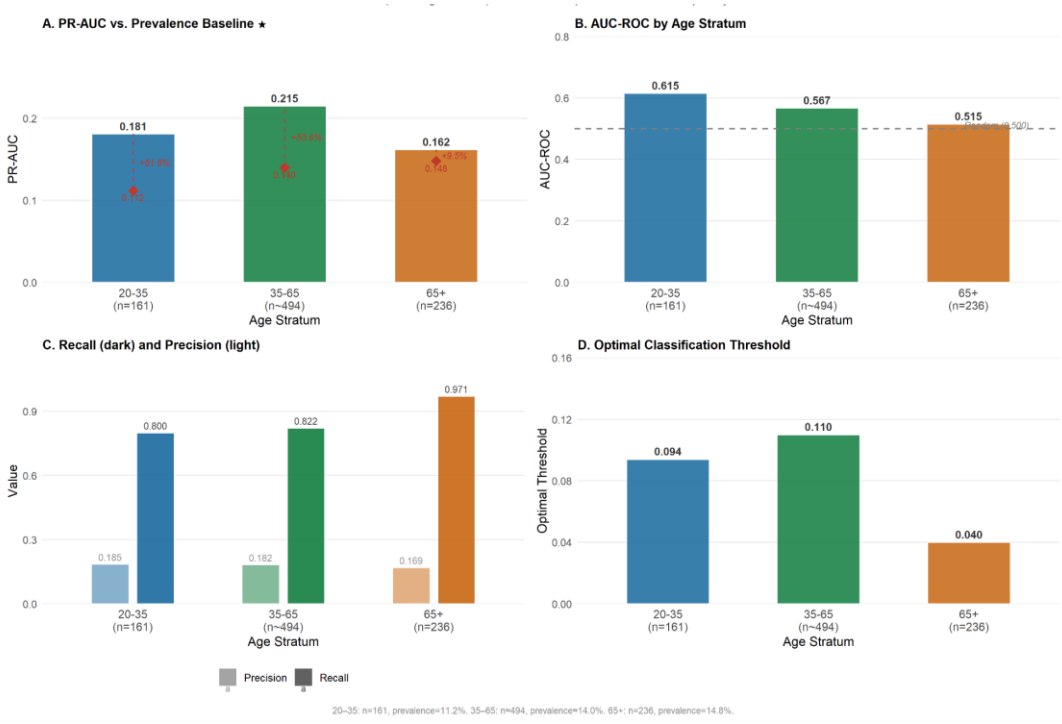

Figure 4.12- Comparative interpretation across strata.

4.9.2. Sensitivity to age stratification

As reported in Section 3.9.5, age-stratified models were estimated for three strata (20–35 [35 inclusive], 35–65 [65 inclusive], 65+ [older than 65]) to assess whether the pooled model masked age-specific dependency structures. The stratified analysis revealed substantial heterogeneity in predictive performance across age groups (Table 4.16 and 4.17), with PR-AUC ranging from 0.162 (65+, near-random) to 0.215 (35–65, strongest performance).

The near-random performance of the 65+ stratum (PR-AUC = 0.162, +9.5% above baseline; AUC-ROC = 0.515) is a robust finding rather than an artefact of small sample size alone. The optimised threshold of 0.040 and recall of 0.971 indicate that the model collapses to near-universally positive classification in this stratum, reflecting genuine difficulty in discriminating dietary behaviour among older adults using the available survey-based variables. This finding is consistent with the greater complexity of dietary determinants in older populations, where clinical comorbidities, polypharmacy, functional limitations, and social isolation interact in ways that cannot be captured by the present variable set.

Finding: Age stratification reveals that the pooled model's moderate performance conceals near-random discrimination in older adults (65+). A single pooled model would overestimate predictive capacity in this subgroup. Age-specific modelling is methodologically warranted and substantively informative.

4.9.2.1. Network structure and direct predictors by age stratum

Table 4.18 summarises the immediate neighbourhood of healthy eating (*alimentacao_saudavel_bin*) in each network at the consensus threshold of 0.30.

Table 4.16 - Network structure and direct predictors by age stratum.

Network	n	Direct predictors (parents)	Downstream nodes (children)	Markov blanket
Pooled	847	Smoking status (0.655); region of residence (0.555)	—	Region, smoking
20–35	161	—	Sex (0.49)	Hypercholesterolaemia, asthma/COPD, sex
35–65	450	Smoking status (0.66)	Screen-time interference (0.79)	Smoking, screen-time interference
65+	236	Region of residence (0.61)	Diabetes (0.43)	Diabetes, region

In the pooled network (n = 847), healthy eating had two direct predictors: region of residence (*nuts2_cat5*, strength = 0.72) and smoking status (*tabagismo_cat3*, strength = 0.60), both robustly oriented towards the target. It had no direct children, and all clinical comorbidities (hypertension, diabetes, cancer, cardiac disease, musculoskeletal pain) connected to healthy eating only indirectly, through a dense component centred on socioeconomic status (*egs_cat3*).

The stratified analyses revealed that these two pooled predictors do not operate uniformly across age, but rather split along age lines:

- In the 20–35 stratum (n = 161), healthy eating had no direct predictors at all; its only adjacency was an outgoing arc to sex (strength = 0.49), which was weakly supported and of unstable orientation (directional consistency = 58%). Neither region nor smoking entered the local structure of the target.

- In the 35–65 stratum (n = 450), smoking status was the sole direct predictor (strength = 0.66), while region dropped out. Healthy eating additionally pointed to screen-time interference (strength = 0.79).
- In the 65+ stratum (n = 236), the pattern inverted: region of residence became the sole direct predictor (strength = 0.61, directional consistency = 89%), smoking disappeared from the target's neighbourhood, and a weak arc to diabetes emerged (strength = 0.43).

Taken together, the two predictors recovered in the pooled model were each age-specific rather than population-wide: smoking was the dominant correlate of healthy eating in middle age, region in older age, and neither was retained in young adults.

These structural differences should nonetheless be interpreted with caution, for two reasons. First, the predictive performance of the networks was modest (pooled cross-validated AUC-ROC = 0.60, AUC-PR = 0.19, F1 = 0.29 at a baseline prevalence of 14.3%), indicating that the recovered dependencies reflect genuine but weak associations. Second, arc orientation in these cross-sectional, self-reported data is a statistical rather than a causal property: arcs directed *from* healthy eating to clinical outcomes (e.g. healthy eating → diabetes in the 65+ stratum) reflect conditional-dependence structure and should not be read as causal effect.

4.9.3. Sensitivity to arc strength threshold

The sensitivity analysis confirmed that the arc strength threshold materially affected predictive performance, while the target neighbourhood remained stable between the $F \geq 0.4$ and $F \geq 0.3$ configurations.

4.9.4. SMOTE

To improve model performance, particularly with respect to specificity, the hypothesis was considered that, in a dataset with such a low prevalence of healthy eaters (14.3%), the model may fail to learn discriminative patterns associated with the positive class due to the limited number of available cases. To address this, a synthetic oversampling approach was evaluated, despite not being a standard technique in Bayesian Network modelling.

Specifically, SMOTENC (Synthetic Minority Oversampling Technique for Nominal and Continuous features) was applied to the training partition of each cross-validation fold to generate synthetic observations of the positive class until a balanced class distribution of 50% was achieved. This experiment was conducted independently for both the 32-variable and the reduced 16-variable datasets. To assess not only the effect of balanced training but also the quality of the synthetic data itself, two evaluation strategies were employed: one in which the test set retained the original real-world class distribution (SMOTENC + real test), and one in which both training and test sets were balanced (SMOTENC + balanced test). For the SMOTENC configurations, Youden's J statistic was used as the classification threshold criterion, as it symmetrically optimises sensitivity and specificity. Table 4.17 reports the results of the SMOTENC sensitivity analysis.
because of that is therefore more appropriate than F1-score when the class distribution is artificially balanced.

Results for both configurations are presented in the following section.

Table 4.17 - SMOTENC sensitivity analysis: performance metrics by configuration and variable set. BSS: Brier Skill Score; ECE: Expected Calibration Error; FP: false positives. ‡ Balanced test set (≈50%) not directly comparable with original model (n=847).

Metric	Original 32v	SMOTE+real 32v	SMOTE+bal 32v ‡	Original 16v	SMOTE+real 16v	SMOTE+bal 16v
Threshold	F1 = 0.180	Youden = 0.310	Youden = 0.276	F1 = 0.100	Youden = 0.310	Youden = 0.396
PR-AUC	0.241	0.223	0.699	0.193	0.204	0.718
ROC-AUC	0.673	0.605	0.749	0.632	0.617	0.776
BSS	0.040	−0.612	−0.061	0.026	−0.850	0.160
ECE	0.030	0.205	0.249	0.019	0.243	0.143
Recall	0.596	0.663	0.825	0.807	0.663	0.825
Specificity	0.694	0.580	0.661	0.425	0.580	0.661
FP	178	322	260 ‡	441	322	260 ‡

When evaluated on the real test distribution, SMOTENC consistently produced negative Brier Skill Scores (−0.612 for 32v; −0.850 for 16v), indicating severe miscalibration attributable to the mismatch between the balanced training distribution (50%) and the real test prevalence (14.3%). The Expected Calibration Error increased from 0.030 in the original model to 0.205 and 0.243 in the SMOTENC + real test configurations for 32v and 16v, respectively.

Structural examination of the consensus DAG revealed that SMOTENC training produced highly unstable network structures. In the 32-variable dataset, the target node acquired 19 to 22 child nodes across folds. Parent nodes of the target were inconsistent across folds and included variables with no established relationship to dietary behaviour. These findings are consistent with the theoretical limitation of SMOTENC for Bayesian Network structure learning: synthetic interpolation between nominal variable categories generates combinations that may not exist in the original population, which are subsequently interpreted as genuine conditional dependencies by the bootstrap algorithm.

The SMOTENC + balanced test configurations produced markedly higher discrimination metrics (PR-AUC 0.699–0.718; ROC-AUC 0.749–0.776). However, these results are not directly comparable with the original model, as the test set prevalence of approximately 50% artificially inflates precision. When assessed

against the clinically relevant criterion, the inadequate diet incorrectly classified as healthy (false positives), no SMOTENC configuration improved upon the original model under real-world prevalence conditions.

4.9.5. Threshold comparison: F1 *vs* MCC *vs* Youden

Given the low positive class prevalence (14.3%), the predicted probabilities generated by the model are substantially compressed, rarely exceeding 0.30. Applying the conventional threshold of 0.5 results in no positive classifications (TP = 0), rendering the model clinically inoperable. Consequently, the optimal classification threshold was determined independently in each training fold by maximising the F1-score across a candidate range from 0.01 to 0.99.

Since the primary concern of this analysis was the number of individuals with inadequate dietary habits incorrectly classified as healthy (false positives, n = 178 in the final model), a threshold sensitivity analysis was conducted to determine whether an alternative optimisation criterion could improve specificity while maintaining clinically acceptable recall. The F1-score, used in the final model, was taken as the reference criterion. Three alternatives were evaluated: the Matthews Correlation Coefficient (MCC), which is considered more appropriate than F1 for imbalanced class distributions as it accounts for all four cells of the confusion matrix; Youden's J statistic, widely used in logistic regression contexts and included here as a potentially informative comparator; and a specificity-constrained maximization approach (Spec.max), which selects the threshold that maximizes specificity subject to a minimum recall of 40%, directly targeting the clinical objective of reducing false positives.

Table 4.18 - Classification threshold and performance metrics by optimisation criterion (mean ± SD, 4 valid folds, 32-variable model). Specificity maximisation subject to recall ≥ 40% (n = 678, 4 valid folds). The thresholds reported here (F1 = MCC = Youden = 0.208 ± 0.091) derive from the threshold-criteria comparison and are not the operating threshold of the final model; the final model applies the F1-optimised threshold of 0.180 reported in Table 4.10.

Criterion	Threshold	Recall	Specificity	F1-score	MCC
F1-score (selected)	0.208 ± 0.091	0.556 ± 0.170	0.730 ± 0.158	0.357 ± 0.051	0.233 ± 0.067
MCC	0.208 ± 0.091	0.556 ± 0.170	0.730 ± 0.158	0.357 ± 0.051	0.233 ± 0.067
Youden's J	0.208 ± 0.091	0.556 ± 0.170	0.730 ± 0.158	0.357 ± 0.051	0.233 ± 0.067
Specificity max*	0.210 ± 0.065	0.515 ± 0.095	0.733 ± 0.124	0.340 ± 0.062	0.203 ± 0.087

Within each of the four valid folds, F1-score, MCC, and Youden's J converged to the same fold-specific optimal threshold, yielding identical performance metrics (Recall = 0.556, Specificity = 0.730, F1 = 0.357, MCC = 0.233). This convergence confirms the robustness of the F1-score as the primary threshold

criterion and rules out threshold selection bias as a source of performance variability. Specificity maximisation produced a marginal gain of Δspecificity = +0.003 at the cost of Δrecall = −0.041, a trade-off without clinical significance given the objective of identifying individuals with inadequate dietary patterns. The F1-score criterion was therefore retained as the primary threshold optimisation method throughout all analyses.

4.9.6. Weighted bootstrap structure learning

Since the *bnlearn package* does not natively support observation weights in the *boot.strength()* function, a manual weighted bootstrap was implemented using *custom.strength*() to aggregate R = 200 independently learned DAGs per fold. Observation weights were set inversely proportional to class prevalence: w(Healthy) = 1/0.143 ≈ 7.0 and w(Not healthy) = 1/0.857 ≈ 1.17, yielding a Healthy/Not-healthy weight ratio of approximately 6:1. Parameter estimation was subsequently performed on unweighted real data in all cases, such that weighting affected only the structural learning phase. Tables 4.19 and 4.20 report the resulting performance metrics and the aggregated confusion matrix.

Table 4.19 - Performance comparison: original model (seed 42, 4 valid folds n=678) vs weighted bootstrap (5 valid folds n= 847), 32-variable dataset. FP: false positives (individuals with inadequate diet classified as healthy). Bold values indicate superior performance.

Metric	Original (4 folds)	Weighted Bootstrap (5 folds)	Δ
PR-AUC	0.241	0.229	−0.012
ROC-AUC	0.673	0.654	−0.019
BSS	0.040	−0.030	−0.070
Recall	0.596	0.761	+0.165
Specificity	0.694	0.563	−0.131
FP†	178	317	+139

Table 4.20 - Aggregated confusion matrix — weighted bootstrap (n = 847, 5 valid folds, F1-score threshold). Rows: real class. Columns: predicted class.

	Pred: Not healthy	Pred: Healthy
Real: Not healthy	TN = 409	FP = 317
Real: Healthy	FN = 29	TP = 92

The weighted bootstrap resolved the Fold 2 PDAG collapse, producing five valid DAGs across all folds. The variable *tabagismo_cat3* appeared as a parent of the target node in all five folds, consistent with the

original model structure. However, specificity decreased from 0.694 to 0.563, corresponding to an increase of 139 additional individuals with inadequate diet misclassified as healthy (FP = 317 vs 178). The Brier Skill Score became negative (BSS = −0.030), indicating deteriorated probabilistic calibration relative to a naïve prevalence-based classifier. An inverse weighting configuration — assigning greater weight to the majority class — was subsequently tested and produced severe model degradation (specificity = 0.085; FP = 664), confirming that bootstrap reweighting in either direction does not constitute a viable approach for improving specificity in this dataset.

4.9.7. Alternative fold stratification seed

As shown in Table 4.23, the Fold 2 PDAG collapse was investigated across multiple random seeds to determine whether it represented a structural property of the dataset or an artefact of the specific fold stratification produced by seed 42. Seeds 123 and 99 were tested and produced PDAG collapses in Fold 1 and Fold 5, respectively, confirming that the collapse is a structural phenomenon that shifts across folds with different seeds but does not disappear. Seed 44 was identified as the only tested configuration yielding five valid directed acyclic graphs with no PDAG collapse, and with an epidemiologically coherent structure: *tabagismo_cat3* appeared as a parent of the target in all five folds, and *nuts2_cat5* in four of five folds.

Table 4.21 compares the four seeds tested. Table 4.22 presents the aggregated confusion matrix for seed 44.

Table 4.21 - Comparison of random seeds tested for fold stratification. FP: false positives under F1-score threshold. Metrics for seeds 123 and 99 were not computed due to PDAG collapse in one-fold (seed 42, n= 678; seed 44, n=847).

Seed	Valid folds	PR-AUC	BSS	Specificity	FP
42 (original)	**4/5**	**0.241**	**0.040**	**0.694**	**178**
44	**5/5**	**0.219**	**0.023**	0.545	330
123	4/5 (Fold 1 collapses)	—	—	—	—
99	4/5 (Fold 5 collapses)	—	—	—	—

Table 4.22 - Aggregated confusion matrix — seed 44 (n = 847, 5 valid folds, F1-score threshold). Rows: real class. Columns: predicted class.

	Pred: Not healthy	Pred: Healthy
Real: Not healthy	**TN = 396**	**FP = 330**
Real: Healthy	FN = 29	TP = 92

Despite resolving the fold collapse, seed 44 produced inferior performance on all primary metrics (PR-AUC = 0.219 *vs* 0.241; BSS = 0.023 *vs* 0.040) and increased the number of misclassified individuals with

inadequate diet by 152 (FP = 330 *vs* 178), corresponding to a reduction in specificity from 0.694 to 0.545. Post-hoc selection of the random seed based on cross-validation performance would additionally constitute a form of data snooping, as it would involve using test-set information to influence a methodological decision that should be fixed a priori. The original seed 42 configuration was therefore retained as the final model.

4.9.8. Tree-Augmented Naïve Bayes

As a final sensitivity analysis, conducted at the recommendation of the dissertation supervisory committee, a Tree-Augmented Naïve Bayes (TAN) classifier was evaluated as an alternative Bayesian Network architecture. TAN extends the Naïve Bayes classifier by relaxing the assumption of full conditional independence between predictor variables: while the target node remains a parent of all explanatory variables (as in Naïve Bayes), each explanatory variable is additionally permitted to depend on one other explanatory variable, structured as a maximum-weight spanning tree derived from pairwise conditional mutual information. This architecture imposes a fixed, globally connected structure on the network, in contrast to the data-driven sparse structure produced by the Hill-Climbing bootstrap approach and was evaluated to determine whether enforcing broader inter-variable dependencies could improve predictive performance, particularly specificity.

Four TAN configurations were evaluated within the same methodological framework as the primary model: stratified 5-fold cross-validation with seed 42, F1-score threshold optimisation, and Dirichlet parameter estimation. Configurations varied by variable set (32-variable *vs* 16-variable) and imaginary sample size (ISS = 1 *vs* ISS = 5). Results are presented in the subsequent table.

Table 4.23 - Performance comparison: original HC bootstrap model vs. TAN configurations. FP: false positives (individuals with inadequate diet classified as healthy). BSS: Brier Skill Score.

Configuration	PR-AUC	AUC-ROC	Recall	Specificity	FP	BSS
Original HC 32v (seed 42)	0.241	0.673	0.596	0.694	178	0.040
TAN 32v ISS=1	0.200	0.611	0.472	0.745	185	-0.336
TAN 32v ISS=5	0.207	0.616	0.538	0.701	217	-0.326
TAN 16v ISS=1	0.204	0.638	0.564	0.704	227	-0.038
TAN 16v ISS=5	0.203	0.650	0.631	0.648	270	-0.033

Across all four TAN configurations, the original HC bootstrap model was superior in PR-AUC and the primary clinical criterion of false positives. The 32-variable TAN configurations produced strongly negative Brier Skill Scores (−0.336 and −0.326), indicating severe miscalibration attributable to the structural properties of the TAN architecture: with 31 predictor variables, each explanatory node has the target as a

parent, producing conditional probability tables with a number of cells that cannot be reliably estimated from the available sample size (n=847, ~121 positive cases). Increasing the imaginary sample size from ISS=1 to ISS=5 provided marginal improvement but did not resolve the miscalibration. The 16-variable configurations yielded substantially better calibration (BSS=−0.038 and −0.033), consistent with the reduced CPT dimensionality, but continued to produce more false positives than the original model (227 and 270 *vs* 178) with inferior discrimination (PR-AUC=0.203–0.204 *vs* 0.241).

Structurally, the TAN produced identical architecture across all five folds in each configuration — 61 arcs with 31 children of the target node for the 32-variable dataset, and 27 arcs with 14 children for the 16-variable dataset — reflecting the deterministic nature of the Chow-Liu spanning tree algorithm. This structural rigidity contrasts with the data-driven variability of the HC bootstrap, where arc inclusion is governed empirically by bootstrap frequency, and constitutes an inherent limitation of the TAN architecture for exploratory network analysis.

These results confirm that the original HC bootstrap model represents the optimal Bayesian Network configuration for this dataset. The TAN architecture, while computationally efficient and theoretically motivated, is not well-suited to datasets with high variable-to-sample ratios and pronounced class imbalance, where the conditional probability tables required by its fully connected structure cannot be reliably estimated.

5. Discussion

5.1. Main findings

The consensus DAG ($F \geq 0.3$, n=847) identified two direct determinants of healthy eating in the Portuguese adult population: smoking status (*tabagismo_cat3*, arc strength = 0.655) and region of residence (*nuts2_cat5*, arc strength = 0.555), which together constitute the complete Markov Blanket of the target node. These two parents are not, however, isolated from one another: passive smoking (*fumador_passivo_bin*) is the only additional variable connected to this cluster, linked to *nuts2_cat5* by an arc whose orientation the bootstrap procedure did not robustly resolve; the same *nuts2_cat5– fumador_passivo_bin* pair was flagged among the directionally ambiguous arcs (direction score ≈ 0.5) in the Fold 2 analysis (Section 4.5). All remaining variables in the dataset form subgraphs that are entirely disconnected from the target. The structure governing healthy eating is therefore a compact three-variable cluster: smoking status, region of residence, and passive smoking — within which smoking behaviour and geographic context act as the proximate (direct) determinants of dietary quality, while the broader set of behavioural, clinical, and sociodemographic variables shows no robust structural connection to the outcome at this sample size.

The direct determinants of healthy eating in this population, smoking status and region of residence, are not straightforwardly modifiable through individual-level behavioural interventions alone. Region of residence is a structural rather than an individual determinant, and its strong association with dietary behaviour suggests that contextual factors such as food availability, cultural dietary norms, socioeconomic infrastructure, and access to healthy food markets play a decisive role in shaping eating patterns across Portuguese territories. Notably, region of residence emerged as a direct predictor despite having been excluded from prior logistic regression models applied to the same dataset, where it did not reach conventional significance thresholds. This discrepancy illustrates a key advantage of the Bayesian Network framework: its capacity to detect dependencies that regression-based approaches may overlook, particularly when the relationship is non-linear or interacts with other variables.

The *what-if* analysis confirmed the substantive importance of both direct predictors. Active smokers showed a -68.4% reduction in the posterior probability of healthy eating relative to the marginal (P=0.045), consistent with the broader literature on behavioural co-occurrence and lifestyle clustering. Ex-smokers exhibited the highest posterior probability among all smoking categories (P=0.188, +32.3% above the marginal), higher even than never-smokers (P=0.173).

The observed association may reflect the clustering of health-promoting behaviours, whereby individuals who successfully quit smoking may also become more motivated to adopt healthier dietary habits and other positive lifestyle changes (5,21,26).

Region of residence revealed substantial geographic heterogeneity. Centro showed the highest posterior probability (+58.3% above the marginal, P=0.225), one possible explanation, not testable with the present data, is reflecting that as stronger adherence to traditional Mediterranean dietary practices, lower urbanisation, and greater availability of locally produced fresh foods. Norte showed a moderate positive effect (+18.1%, P=0.168). By contrast, the Lisbon Metropolitan Area (−55.7%, P=0.063) showed markedly lower probabilities, likely reflecting dietary patterns characteristic of dense urban environments, higher reliance on ultra-processed foods, fast food consumption, and time constraints on meal preparation. Alentejo (−38.0%, P=0.088) and Algarve (−62.1%, P=0.054) showed the lowest probabilities, though estimates for these regions carry greater uncertainty given their small sub-sample sizes (7.8% and 4.6%, respectively) and should be interpreted with caution. These findings call for regionally differentiated public health strategies rather than uniform national approaches.

The combined worst-case profile: current smoker residing in the Algarve yielded a near-zero posterior probability of healthy eating (P=0.003, −97.9%), while the best-case profile: ex-smoker in Centro approached one-third probability (P=0.334, +135.2%). These extreme profiles illustrate the magnitude of the joint effect of the two direct determinants and highlight the populations most in need of targeted intervention.

The remaining 28 variables, including chronic disease burden, sleep quality, mental health, BMI, physical activity, and sociodemographic characteristics, form internally coherent subgraphs that are structurally disconnected from the target node. This does not imply these variables are clinically irrelevant to diet; rather, it reflects that, given the sample size (n=847) and class prevalence (~14%), the bootstrap structure learning algorithm did not find sufficient and consistent evidence to retain directed paths from these clusters to the target across resamples. The absence of variables such as BMI, age, and education as direct predictors despite their well-documented associations with diet quality in the literature should therefore be interpreted as a statistical limitation rather than a substantive finding of independence.

The observed improvement in predictive performance between $F \geq 0.4$ and $F \geq 0.3$ (PR-AUC +10.6%) does not have a clear structural explanation: the Markov Blanket of the target is identical across both thresholds, and neither direct parent acquired new ancestors at the lower threshold. The 10 additional arcs in the $F \geq 0.3$ graph connect variables within the chronic disease and lifestyle subgraphs but create no new paths to the target. Given the small sample size and low prevalence, this performance difference may partly reflect sampling variability across cross-validation folds rather than genuine capture of additional structural dependencies. This constitutes a recognised limitation of the analysis.

Nevertheless, the network structure between non-target variables is itself of scientific interest. The learned associations like musculoskeletal pain and sleep quality, depression and asthma/COPD, sleep medication and anxiety, diabetes, hypertension and cardiovascular disease are well-supported in the clinical

literature and represent plausible comorbidity pathways. Their emergence from data-driven structure learning, without prior specification, lends additional credibility to the network as a representation of the underlying epidemiological structure of the Portuguese adult population, even where these pathways do not directly inform the prediction of healthy eating.

5.2. Clinical interpretation and public health

5.2.1. Smoking status and health behaviour clustering

The strong association between smoking status and healthy eating identified in this study is consistent with previous literature. Smoking has been associated with less favourable dietary behaviours, including a greater preference for energy-dense foods, higher consumption of unhealthy food items, and overall poorer diet quality (36,63). Consequently, smoking may be viewed as an important marker of unhealthy dietary patterns and lifestyle behaviours more broadly. This finding is in line with evidence suggesting that health-related behaviours tend to cluster, with smokers being more likely to engage in other risk behaviours, including higher alcohol consumption and less healthy dietary choices (5,8,26). This relationship may be particularly relevant in the present study, as alcohol intake contributed to the construction of the healthy eating outcome variable.

Former smokers showed the highest probability of healthy eating among all smoking categories. This finding may reflect the tendency for positive health behaviours to occur simultaneously, whereby individuals who successfully quit smoking often adopt additional lifestyle improvements, including healthier dietary habits (5,21,26). Although smoking cessation is frequently associated with short-term weight gain and increased caloric intake, these effects tend to be most pronounced during the initial period following cessation (64,65). Over the longer term, behavioural changes are often more holistic, resulting in healthier overall lifestyle patterns. The results observed in this study are consistent with this interpretation, suggesting that former smokers may represent a particularly health-conscious subgroup (66).

Never-smokers also demonstrated a relatively high probability of healthy eating, although lower than that observed among former smokers. One possible explanation is that individuals who have never smoked may be less prone to engaging in addictive or health-risk behaviours in general and may therefore be more likely to maintain healthier dietary habits throughout life. Together, these findings support the view that smoking status is not only a marker of tobacco exposure but also an indicator of broader lifestyle patterns that influence dietary quality (67). The regional distribution of current smokers is mapped in Figure 5.1.

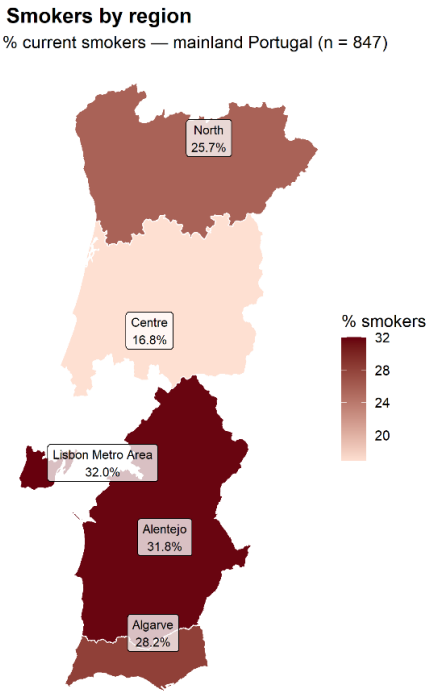

Figure 5.1 - Smokers by region.

5.2.2. Geographic gradient and the structural nature of region as a determinant

Region of residence emerged as a direct parent of healthy eating in the Bayesian Network, with substantial heterogeneity in posterior probabilities across the five mainland NUTS II regions. Centro showed the highest probability of healthy eating (P=0.225, +58.3% above marginal), followed by Norte (P=0.168, +18.1%). The remaining regions exhibited markedly lower probabilities: Alentejo (P=0.088, −38.0%), the Lisbon Metropolitan Area (P=0.063, −55.7%), and Algarve (P=0.054, −62.1%). This geographic gradient is consistent with previously documented regional disparities in Mediterranean dietary adherence in Portugal, where adherence has been shown to vary substantially across NUTS II regions despite the country's overall classification as Mediterranean (27,28,42).

The finding that Centro presented the highest probability of healthy eating aligns with regional patterns of Mediterranean diet adherence reported in previous Portuguese surveys. The Centro region encompasses areas with strong agricultural traditions, lower urbanisation density, and greater proximity to small-scale producers, all of which have been associated with higher consumption of fresh, locally produced foods and lower reliance on ultra-processed products (42,68). Norte showed an intermediate positive effect, consistent with its mixed rural-urban composition and the persistence of traditional dietary practices in inland subregions.

The lower probability of healthy eating in the Lisbon Metropolitan Area is consistent with multiple lines of evidence on the dietary impact of urbanisation in Portugal. Analyses of the IAN-AF cohort have shown that residents of the Lisbon Metropolitan Area have significantly higher ultra-processed food consumption compared to other mainland regions (42), a pattern attributable to the dietary characteristics of dense urban environments: greater reliance on convenience foods due to time constraints, higher exposure to fast-food retail, and reduced engagement with traditional meal preparation practices. The Lisbon Metropolitan Area has also been identified as a region where Mediterranean diet adherence has declined most markedly over recent decades (27,28,42).

The interpretation of results for Alentejo and Algarve requires caution. These regions account for 7.8% and 4.6% of the sample respectively ($n \approx 70$ and $n \approx 41$), and the wider uncertainty around these estimates reflects the limited sub-sample size. Nevertheless, the observed lower probability of healthy eating in these regions, particularly in the Algarve, is consistent with previously reported regional heterogeneity in dietary patterns in Portugal.

National evidence has shown that adherence to a Mediterranean dietary pattern varies across Portuguese regions and is influenced by socioeconomic and demographic factors, including age structure and educational attainment (28,42). In addition, studies conducted in the Algarve have identified relevant challenges related to food insecurity and inequalities in access to healthy food environments, which may contribute to differences in dietary quality at the regional level (35).

More broadly, dietary behaviour is shaped by structural determinants of food environments, including availability and accessibility of healthy food options, particularly in contexts where demographic ageing and socioeconomic constraints may limit dietary diversity (69,70). In this context, the results observed in the present study should be interpreted as reflecting broader structural and socioeconomic patterns rather than region-specific causal mechanisms.

A second methodological implication concerns the nature of region as a determinant. Unlike smoking status, region of residence is not a modifiable behaviour but a contextual characteristic that encodes a constellation of underlying factors: food availability, retail environment, dietary culture, socioeconomic infrastructure, and access to fresh produce. The strong association observed in the Bayesian Network should therefore not be interpreted as a causal effect of geography per se, but as evidence that the structural and environmental conditions captured by region of residence exert a substantial influence on dietary behaviour. This distinction has direct policy relevance: while individual-level interventions targeting smoking cessation may yield co-benefits for dietary quality (Section 5.2.1), interventions targeting regional dietary disparities will require structural rather than behavioural strategies, improving food retail diversity in urban areas, supporting local food systems in rural regions, and addressing the seasonal economic precarity that may affect dietary stability in tourism-dependent areas. The regional distribution of healthy eaters is mapped in Figure 5.2.

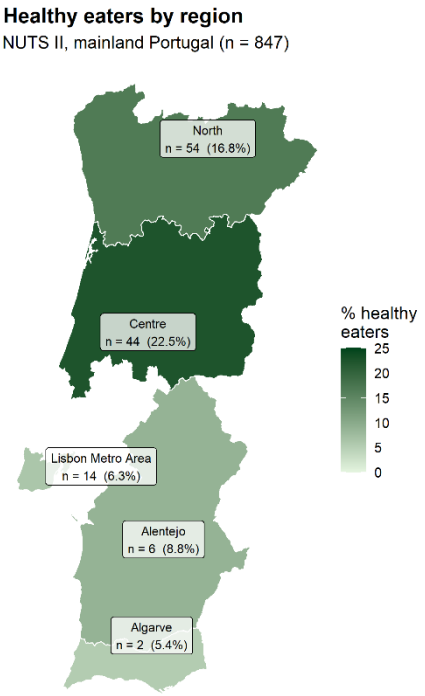

Figure 5.2 - Healthy eaters by region.

5.3. Joint effect of smoking status and region of residence

While Sections 5.2.1 and 5.2.2 examined smoking status and region of residence as separate direct determinants of healthy eating, the Bayesian Network framework allows the joint conditional probabilities of these two parents to be examined through the conditional probability table (CPT) of the target node. This joint perspective is conceptually informative: the probability of healthy eating in any given individual is determined not by either factor alone but by their combination yet must be interpreted with substantial caution given the structural limitations of the available data. Figure 5.3 presents the joint probabilities of healthy eating by region and smoking status.

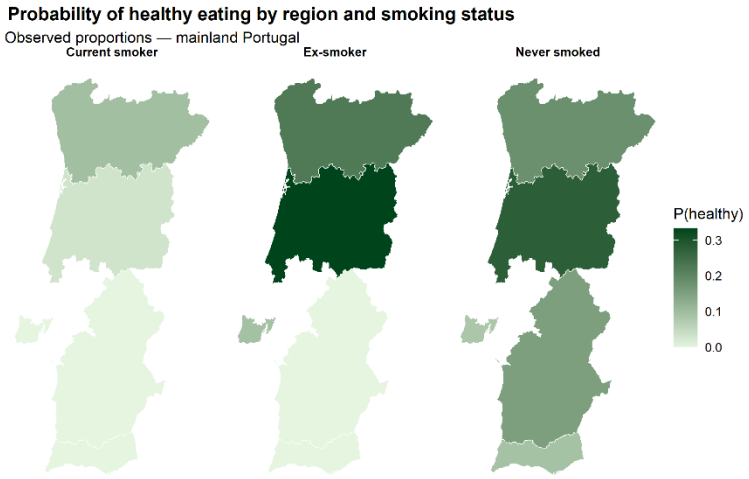

Figure 5.3 - Probability of healthy eating by region and smoking status.

5.3.1. Cautious interpretation: small cell counts and parameter smoothing

Before interpreting the values in this dissertation, it is essential to acknowledge that several cells of the CPT are populated by very few observations. With n=847, an overall positive class prevalence of approximately 14.3%, and three smoking categories distributed across five NUTS II regions, the joint cells defined by region × smoking status × outcome contain substantially fewer observations than the marginal counts might suggest. Algarve accounts for only 4.6% of the sample (n≈41) and Alentejo for 7.8% (n≈70). Within these regions, the number of active smokers reporting healthy eating is plausibly below five in absolute terms and may approach zero for some combinations. The parameter estimates for these cells are therefore strongly influenced by the Dirichlet smoothing prior (iss=1) applied during MAP estimation, rather than by direct empirical evidence.

This has two important implications. First, the apparent precision of values such as P=0.003 for active smokers in Algarve or P=0.001 for active smokers in the Lisbon Metropolitan Area is illusory; these probabilities reflect the absence of observed positive cases in those cells, not a robust empirical estimate of the underlying conditional probability. Second, the direction of the effect markedly lower probability of healthy eating in active smokers, particularly in regions with otherwise low baseline probabilities- is supported by the broader pattern across the CPT and by the marginal effects estimated from larger subsamples. The relative ordering of the cells is therefore more interpretable than their absolute numerical values.

5.3.2. Robust patterns despite limitations

Notwithstanding the limitations above, three patterns in the CPT are supported by sufficiently large subsamples to warrant interpretation.

First, the substantial positive effect of being an ex-smoker is observed consistently across the two largest regions in the sample, Norte (n≈340), where ex-smokers reach P=0.200 versus never-smokers at P=0.178, and Centro (n≈208), where ex-smokers reach P=0.334 versus never-smokers at P=0.299. These cells are populated by enough observations to support meaningful estimation, and the consistent direction of the effect, ex-smokers exhibiting higher probabilities than never-smokers, reproduces the marginal finding discussed in Section 5.2.1. This regional consistency supports the interpretation that smoking cessation is associated with broader lifestyle restructuring rather than being a sample artefact. A methodological caveat applies specifically to the best-case profile. Although ex-smokers in the Centro region attain the highest posterior probability in the entire CPT (P=0.334), the corresponding joint cell contains almost no observed positive cases (Table 4.12), so this particular estimate is dominated by the Dirichlet smoothing prior rather than by empirical data and its exact magnitude should be treated with caution. The most favourable profile supported by a sufficient number of observations is that of never-smokers in the Centro region (P=0.299, n≈118), which is reported here as the highest empirically grounded combination. The direction of the gradient — Centro yielding the highest probabilities and smoking cessation conferring an advantage — is preserved under both readings.

Second, the negative effect of active smoking on the probability of healthy eating is consistent across all five regions, even where absolute values cannot be precisely estimated. In every regional column of the CPT, active smokers exhibit lower probabilities than both ex-smokers and never-smokers. This monotonic pattern, while individually fragile in any single cell, is robust across the table as a whole and is consistent with extensive prior evidence on smoking and dietary quality (26,63,67).

Third, the regional gradient observed at the marginal level: Centro and Norte with higher probabilities of healthy eating, the Lisbon Metropolitan Area and the southern region with lower probabilities, is preserved within each smoking stratum. Never-smokers in Centro (P=0.299) exhibit substantially higher probabilities than never-smokers in the Lisbon Metropolitan Area (P=0.038), and the same pattern holds for ex-smokers and active smokers. The regional context appears to modulate dietary behaviour regardless of smoking status.

5.3.3. Combined effects: direction over magnitude

The CPT suggests that the combined effect of active smoking and unfavourable regional context is more pronounced than would be expected from the simple sum of marginal effects. The highest joint probability among ex-smokers in Centro (P=0.334) exceeds the marginal probability for either factor alone, while the lowest joint probabilities cluster among active smokers in southern and metropolitan regions. This pattern is consistent with broader evidence on the clustering of behavioural and contextual risk factors, where unfavourable behaviours and environments tend to co-occur and reinforce one another's effects on health outcomes (26,66,67,69). However, given the limitations discussed in Section 5.3.1, the magnitude of this interaction cannot be precisely quantified from the present sample. The direction of the effect is informative; its exact size is not.

5.3.4. Public health implications

Despite the limitations on precise estimation, the CPT identifies a coherent profile of populations at high risk of inadequate dietary patterns: active smokers residing in southern Portuguese regions or in the Lisbon Metropolitan Area. This identification is consistent with regional epidemiological evidence on smoking-attributable burden in Portugal and on broader regional disparities in health and dietary outcomes (2,28,42). The model therefore provides a probabilistic mechanism for identifying high-risk subgroups using only routinely collected behavioural and geographic information, even where the precise probability estimates for the smallest subgroups should be interpreted as approximate rather than definitive.

These findings underline two methodological recommendations for future work. First, larger or oversampled subsamples in less populated regions (Algarve, Alentejo) would substantially improve the precision of joint conditional probability estimates, particularly for the rarer combinations of risk factors, consistent with known limitations of sparse contingency tables in probabilistic graphical models (71). Second, the interpretive caution required for sparse cells should be made explicit in any practical application of the model, with predictions for individuals corresponding to small CPT cells being flagged as carrying greater uncertainty.

5.4. Variables disconnected from the target: statistical limitation or substantive finding?

The consensus DAG identified only two direct parents of the target node, smoking status and region of residence. The remaining 28 variables in the dataset, including BMI, age, education level, physical activity, sleep quality, mental health indicators, and chronic disease burden, formed structurally disconnected subgraphs with no directed paths to the outcome. This finding is, at first reading, counterintuitive: each of these variables has a well-established association with diet quality in the epidemiological and nutritional literature, and their absence from the Markov Blanket of the target requires careful interpretation.

The most parsimonious explanation is statistical. With n= 847 and a positive class prevalence of approximately 14.3%. Bayesian Network structure learning by Hill-Climbing with bootstrap aggregation evaluates the consistency of each candidate arc across 200 resamples; an arc is retained in the consensus graph only when its bootstrap frequency exceeds the predefined threshold ($F \geq 0.3$ in this analysis). For weak or moderately strong dependencies, the variability introduced by bootstrap resampling at this sample size is sufficient to push arc frequencies below the threshold, even where a true dependency exists in the underlying population. This phenomenon is well recognised in Bayesian Network structure learning, where the recovery of weaker dependencies becomes increasingly challenging as sample size decreases and signal strength weakens (72). The absence of variables such as BMI, age, and education as direct parents of the target should

therefore not be interpreted as evidence that these variables are independent of dietary behaviour, but rather as evidence that any such dependencies, if present, were not stable enough across bootstrap resamples to reach the consensus threshold. This is a conservative property of the methodology: the network reports only the dependencies for which it has robust empirical support.

A complementary methodological consideration concerns the choice of the arc strength threshold. The primary analysis used $F \geq 0.3$, a permissive threshold by Bayesian Network standards, chosen specifically to retain dependencies that, while not unequivocally robust, were supported by at least 30% of bootstrap resamples. Sensitivity analyses with $F \geq 0.4$ and $F \geq 0.7$ produced markedly sparser networks, with even fewer arcs surviving consensus. Importantly, however, the Markov Blanket of the target was identical across all three thresholds: at no threshold did any of the 28 disconnected variables gain a directed path to the target. This indicates that the absence of these variables from the predictive structure is not an artefact of the specific threshold chosen, but a stable property of the bootstrap distribution at this sample size. No threshold within the tested range produces a network in which BMI, age, or education directly inform the prediction of healthy eating.

While BMI is robustly associated with diet quality in observational studies (73), prospective evidence suggests that dietary quality acts primarily as an upstream determinant of obesity and BMI rather than the reverse (74). The temporal ordering of these variables therefore supports an interpretation in which dietary behaviour precedes changes in BMI, rather than BMI determining current dietary patterns. This interpretation is also consistent with Bayesian and causal modelling studies of obesity, which typically identify dietary and lifestyle factors as upstream drivers of BMI outcomes (52).

Were the network to have placed BMI as a parent of the target, this would have implied that current body composition determines current dietary choices and a relationship not well supported by the underlying causal mechanisms, even where the variables are cross-sectionally correlated. This argument extends, with varying degrees of strength, to other disconnected variables. Age and education are commonly modelled as predictors of diet quality in regression-based approaches, but their effects are typically substantially mediated by other variables including regional context, socioeconomic conditions, and behavioural lifestyle factors, some of which are not present in the dataset rather than acting as direct causes (8,20,22,52). In a Bayesian Network framework that explicitly models conditional dependencies, the influence of such variables may be absorbed by the direct parents of the target, leaving age and education without a direct arc but still indirectly informing the joint probability distribution.

A finding that supports the interpretation of methodological coherence rather than statistical failure is the structure of the disconnected subgraphs themselves. The 28 variables not connected to the target are not floating in isolation; they form internally coherent subgraphs that reproduce well-documented clinical and behavioural associations. The chronic disease subgraph links hypertension to cardiovascular disease, diabetes to hypertension, and metabolic conditions to musculoskeletal complaints. The mental health subgraph links anxiety to depression, sleep quality to anxiety, and screen time interference to sleep problems. The respiratory

health subgraph links asthma/COPD to depression, mediated by chronic disease burden. These internal structures are epidemiologically plausible and consistent with prior literature, indicating that the structure learning algorithm operates correctly and detects genuine dependencies where they exist. The fact that none of these subgraphs connects to the target node strengthens, rather than weakens, the interpretation that the target's true direct parents within the available variable set are indeed limited to smoking status and region of residence.

The combined reading of these considerations is that the absence of variables such as BMI, age, education, physical activity, and chronic disease burden from the Markov Blanket of the target reflects a methodologically conservative outcome rather than a substantive claim of independence. The absence of an arc to the target should not be read as a finding of no association since the Bayesian Network reports only dependencies for which there is robust evidence of conditional dependence; absence of an arc indicates absence of evidence rather than evidence of absence. Variables disconnected from the target may still be epidemiologically and clinically relevant to dietary behaviour at the population level; their absence from the predictive structure indicates only that they do not, given the constraints of the present data, contribute information beyond what is already captured by smoking status and region of residence. Where the literature suggests that a variable is downstream of the outcome of interest, BMI because of diet rather than as a cause, the network's failure to identify it as a parent is methodologically appropriate and aligned with the correct causal direction. The Bayesian Network framework, in this respect, performs a function that regression-based approaches typically cannot: it discriminates between variables that are statistical correlates of the outcome and variables that are direct conditional parents.

5.5. Model performance in context

The predictive performance of the Bayesian Network [PR-AUC = 0.241, ROC-AUC = 0.673, Brier Skill Score = 0.040] is modest in absolute terms but must be interpreted considering the dataset and the specific demands of evaluating models with severe class imbalance. A naïve reading of these values risks underestimating the model's utility; conversely, an uncritical comparison with benchmarks reported in dissimilar contexts risks the opposite distortion. This section situates the observed performance against three reference points: the theoretical baseline expected for a dataset of this prevalence, the published performance of comparable predictive models in nutritional and behavioural epidemiology, and the methodological limits imposed by sample size and outcome rarity.

The first reference point concerns the interpretation of PR-AUC itself in the context of severe class imbalance. Unlike ROC-AUC, which has a universal random-baseline value of 0.5 regardless of class distribution, the PR-AUC baseline is equal to the positive class prevalence (75). With a positive class prevalence of 14.3% in this analysis, the random baseline PR-AUC is 0.143. The observed PR-AUC of 0.241 therefore represents a 68% relative improvement over random classification. A meaningful gain in absolute terms, particularly given the conservative nature of the F1-optimised threshold and the structural conservatism

of the consensus DAG. However, it does not approach the discriminative performance reported in studies of more clinically well-defined outcomes, where predictive signals tend to be stronger and more consistently captured in routinely available variables (75,76).

The second reference point concerns comparable studies in the Bayesian Network literature. Applications of Bayesian Networks to healthcare prediction have been the subject of two recent comprehensive reviews (48,77), which together document a substantial concentration of work on cardiometabolic, oncological, and obstetric outcomes, with comparatively less emphasis on behavioural constructs and composite lifestyle indicators. The most frequently modelled outcomes in the available Bayesian Network literature are cardiometabolic conditions such as type 2 diabetes and coronary heart disease and obesity (52,53) outcomes that, while clinically important, are typically defined using standardised diagnostic criteria or physiological biomarkers. Healthy eating as defined in the present analysis is a composite behavioural variable derived from self-reported food consumption patterns assessed against the Portuguese National Food Wheel guidelines (42,78) and consequently aggregates substantial measurement variability and behavioural heterogeneity within a single binary label. The literature on Bayesian Network prediction of comparable behavioural composite outcomes remains limited, and direct performance benchmarks are therefore scarce.

In the broader literature on Mediterranean diet adherence in Mediterranean populations, comparable studies have consistently identified only a small number of robust predictors despite testing large variable sets. A systematic literature review covering 37 cross-sectional studies in Mediterranean countries (20) identified higher socioeconomic status, regular breakfast consumption, employment status, and demographic characteristics as the most consistent determinants of Mediterranean diet adherence at the individual level. In the Portuguese context specifically, analysis of the EpiDoC 3 cohort (n = 5647) identified age, educational level, income, and region of residence as the main factors associated with low adherence to the Mediterranean diet, with marked regional heterogeneity across the country (28). The convergence between the structural findings of the present analysis, which identified region of residence and smoking status as the two direct parents of healthy eating and the predictor sets identified in the broader literature on Mediterranean diet adherence lends external validity to the model, even where its absolute discriminative performance is modest.

The third reference point is the structural ceiling imposed by the data. With n = 847 and approximately 14% positive cases, the effective sample for learning conditional dependencies involving the target node is small by the standards of predictive modelling. Healthy eating itself is influenced by a wide and heterogeneous set of determinants including cost and household economic conditions, time availability, cooking practices, household composition, social context, and the food environment, many of which are not captured or only weakly captured by population health surveys (17,19). The Bayesian Network is, by construction, only able to identify dependencies between variables that are present in the dataset; the absence of richer behavioural and environmental measures necessarily imposes an upper bound on the achievable predictive performance. Bayesian Network methodology itself contributes additional conservatism, as bootstrap-based structure learning preferentially identifies the most robust dependencies and excludes those that, while potentially

genuine, cannot be confidently distinguished from sampling noise at moderate sample sizes. The observed performance therefore approximates the upper bound of what is achievable from this dataset with this methodology, rather than a deficit that could be substantially improved through alternative algorithmic choices, as confirmed by the sensitivity analyses presented in Section 4.9.

The Brier Skill Score of 0.040 indicates that the model's probabilistic predictions are better calibrated than a naïve classifier that always predicts the marginal prevalence, but only modestly so. The Expected Calibration Error of 0.030 supports this interpretation: the model's predicted probabilities are within 3 percentage points of the empirical frequencies on average, an acceptable level of calibration for a model of this type and complexity. In a clinical decision-support context, this means that the predicted probabilities can be interpreted as approximate but not precise estimates of individual-level risk adequate for triage and risk stratification, but not for direct quantitative comparisons of individual probability estimates without further validation. This interpretation is consistent with the role of Bayes' theorem in medical decision-making as a framework for updating prior beliefs in the presence of imperfect evidence, rather than as a source of definitive probabilistic claims (6).

Finally, the central clinical metric specificity of 0.694, corresponding to 403 individuals with inadequate dietary patterns correctly identified out of 581 in the test sample, should be understood as the relevant performance criterion for the intended use case. In a population where inadequate dietary patterns are a leading modifiable contributor to the national burden of non-communicable disease (2,3,7) and where almost a quarter of food consumption falls outside the recommendations of the Portuguese Food Wheel (42), the ability to correctly identify approximately 70% of individuals with poor dietary patterns using only smoking status and region of residence is a non-trivial outcome with potential value for primary care screening. This profile is consistent with the model's design priorities: a screening tool that prioritises the identification of the at-risk population over individual-level precision, and that achieves meaningful discrimination using only a small set of routinely collected behavioural and demographic variables. The performance is modest but, given the constraints of the data, defensible.

5.6. Class imbalance and methodological robustness

A central methodological concern throughout this work has been the substantial class imbalance in the target variable, with healthy eaters representing only 14.3% of the analytic sample. Class imbalance is a well-documented challenge for predictive modelling in health research, with both well-established consequences for model performance and an extensive though increasingly controversial toolkit of corrective techniques. The decisions made in this study regarding how to handle the imbalance, and the extensive sensitivity analyses conducted to test those decisions, constitute one of the central methodological contributions of this dissertation. This section discusses the implications of the imbalance, the rationale for the approach

taken, and the robustness of the conclusions considering the alternative approaches that were systematically evaluated and rejected.

The first issue raised by class imbalance in this dataset is the immediate inadequacy of the conventional classification threshold of 0.5. With a positive class prevalence of approximately 14%, the posterior probabilities estimated by the Bayesian Network exhibited limited separability between classes, rarely exceeding 0.30 even for the most extreme combinations of predictor values (79). Application of the default threshold therefore produced a degenerate classifier in which no individual was ever predicted to belong to the positive class, with the resulting sensitivity, F1-score, and Matthews Correlation Coefficient all collapsing to zero (80). This is a recognised consequence of applying default decision thresholds in imbalanced settings (59,75,81) and reflects the well-documented inadequacy of fixed 0.5 cut-offs under skewed class distributions.

This required the implementation of threshold optimisation as a non-discretionary methodological step rather than as an optional refinement. The choice of F1-score maximisation as the optimisation criterion (60,61), validated by the convergence of F1, MCC, and Youden's J (82) to similar thresholds across all valid folds, ensured that the model's discriminative capacity could be expressed in classification terms without introducing arbitrary post-hoc adjustment. This robust methodological approach directly addresses the known vulnerabilities of standard thresholding under class skew (83).

The second issue concerns whether the class imbalance itself should have been corrected through resampling at the training stage. Methods such as SMOTE and its variants are widely applied in machine learning pipelines and have become a default recommendation for imbalanced settings (59). However, methodological scholarship has increasingly challenged this approach. The distortion of posterior probabilities and the resulting poor model calibration introduced by data-level corrections had already been documented within the broader machine learning literature (79). This concern was recently reinforced by a large-scale simulation study (81), which demonstrated that all major imbalance correction methods, including random oversampling, random under-sampling, and SMOTE, produced models with poor calibration through systematic overestimation of the minority class probability, without improving discrimination as measured by ROC-AUC. Crucially, evidence shows that improvements in sensitivity and specificity traditionally attributed to imbalance correction can equally be obtained by shifting the probability threshold, without data resampling (81). The interpretation offered by this body of work is that class imbalance is not a problem to be corrected at the data level when the goal is to estimate calibrated probabilities; rather, the problem lies in the use of inappropriate classification thresholds applied to those probabilities.

The findings of the sensitivity analyses presented in Section 4.9 are entirely consistent with this position. SMOTENC oversampling at the training stage, evaluated across six configurations encompassing both variable sets and two test-distribution conditions, consistently produced Brier Skill Scores substantially below the original model when evaluated against the real-world prevalence distribution (BSS = −0.612 for the 32-variable dataset; BSS = −0.850 for the 16-variable dataset). The structural diagnostics were equally

revealing: SMOTENC training produced consensus DAGs in which the target node acquired between 19 and 22 child nodes, a hub-like topology incompatible with the epidemiological structure of the underlying data and with no precedent in the literature on dietary behaviour. These artefacts can be understood as the direct consequence of generating synthetic minority instances by interpolation between categorical features, a procedure that creates combinations of variable values that do not exist in the original population and that are subsequently interpreted by the bootstrap structure learning algorithm as genuine conditional dependencies. The theoretical incompatibility between SMOTE-based interpolation and primarily categorical data has been explicitly acknowledged in the methodological literature (84), and the present findings provide empirical confirmation of these concerns in a population health application using a Bayesian Network framework.

The weighted bootstrap analyses, in which observation weights inversely proportional to class prevalence were applied during structure learning, similarly failed to improve performance on the clinically relevant criteria. Although the weighted procedure successfully resolved the Fold 2 PDAG collapse and identified *tabagismo_cat3* as a parent of the target across all five folds, specificity decreased from 0.694 to 0.563, corresponding to an increase of 139 individuals with inadequate dietary patterns misclassified as healthy eaters. The Brier Skill Score also deteriorated from positive (BSS = 0.040) to negative (BSS = −0.030). The inverse weighting configuration, tested to determine whether increased weight on the majority class could improve specificity through alternative means, produced complete model degradation (specificity = 0.085; FP = 664), confirming that bootstrap reweighting in either direction was not a viable strategy in this dataset.

The Tree-Augmented Naïve Bayes architecture evaluated across four configurations produced strongly negative Brier Skill Scores in the 32-variable configurations and marginally negative scores in the 16-variable configurations, with increased false positives relative to the original model in all four cases. The structural rigidity of the TAN architecture, with all explanatory variables forced to have the target as a parent, inherently expands the parameter space. This constraint produced conditional probability tables of dimensionality incompatible with the sample size available (85). Under such severe data sparsity, the estimation process becomes unstable, regardless of the imaginary sample size used for Dirichlet smoothing, as over-parameterised tables tend to wash out real empirical patterns when smoothed (86).

Taken together, the sensitivity analyses converge on a single coherent conclusion: the performance characteristics of the original Hill-Climbing bootstrap model reflect the inherent signal available in the dataset given its sample size and prevalence, rather than methodological limitations correctable through resampling, reweighting, or architectural substitution. No alternative approach evaluated improved the model's specificity for identifying individuals with inadequate dietary patterns without simultaneously degrading either calibration, structural interpretability, or both. This finding extends the methodological position of van den Goorbergh (81) from logistic regression and standard machine learning algorithms to the Bayesian Network framework: in the present application, threshold optimisation alone proved sufficient to translate the model's probabilistic discrimination into operationally meaningful classification, and class imbalance corrections at the

data or structure-learning level offered no additional benefit while introducing substantial risks of calibration loss and structural distortion.

This conclusion carries implications beyond the present analysis. The increasing default application of resampling methods in health prediction modelling has been challenged on both conceptual and empirical grounds (81), and recent calls for greater methodological caution are strongly reinforced by the findings reported here. For Bayesian Network applications, where structural interpretability is a primary methodological asset in healthcare environments (77), the introduction of synthetic data through SMOTE-based methods carries the additional risk of distorting the true underlying data distribution and conditional independence structures (59). This directly compromises the structural validity that motivates the choice of this graphical framework in the first place. The methodological position adopted in this dissertation to preserve the original class distribution, address imbalance through threshold optimisation (61), and validate this approach through systematic sensitivity analyses is therefore not only defensible in light of the present results but strictly aligned with the emerging consensus in the broader literature on prediction modelling under class skew (60,81).

5.7. Comparison with previous studies

The structural findings of the present Bayesian Network namely, the identification of smoking status and region of residence as the two direct parents of healthy eating, the absence of direct effects for body mass index, age, and educational level, and the elevated probability of healthy eating among ex-smokers can be contextualized against three converging lines of evidence in the recent literature: (i) Portuguese population-based studies of dietary quality and adherence to the Mediterranean diet, (ii) cross-national European analyses of smoking and dietary behavior, and (iii) the small body of work applying Bayesian Networks to dietary outcomes. The convergence between the findings of the present analysis and this prior literature provides external validity to the model, while the divergences highlight the specific methodological contribution of the network approach.

5.7.1. Sociodemographic determinants of healthy eating in the portuguese population

Three recent population-based studies of the Portuguese adult population have addressed the determinants of dietary quality using regression-based methods, providing the closest empirical reference points for the present analysis. Analysing the third wave of the Epidemiology of Chronic Diseases cohort (EpiDoC 3, n = 5,647) (28), it was reported that low adherence to the Mediterranean diet was independently associated with younger age, lower educational attainment, lower household income, and region of residence, with marked heterogeneity across NUTS II regions. Using the IAN-AF 2015–2016 dataset for comparison (n

= 5,005, ages 3–84) (87), it was found that ultra-processed food consumption was significantly associated with age, educational level, civil status, urbanisation level, region of residence, and smoking status, with female current and former smokers exhibiting higher ultra-processed food intake than never-smokers. In another study also using the IAN-AF 2015–2016 (n = 3,852 adults), reported that adherence to the EAT-Lancet Planetary Health Diet was associated with sex, educational level, food insecurity, body mass index class, and physical activity level (88).

The set of variables identified as relevant determinants of dietary quality across these three Portuguese studies, like age, education, income, region, smoking status, food insecurity, and physical activity, coincides almost entirely with the candidate variable set considered in the present analysis. The fact that the Bayesian Network structure learning algorithm retained only smoking status and region of residence as direct parents of the target node, while relegating age, education, body mass index, and physical activity to indirect or disconnected positions in the consensus graph, does not contradict the prior literature: it reflects a substantively different statistical question. Regression-based analyses evaluate the marginal contribution of each variable conditional on a fixed adjustment set, whereas Bayesian Network structure learning evaluates conditional independence relationships across all candidate parent sets simultaneously, retaining only those variables whose direct dependence on the target is not screened off by other variables in the network.

Variables such as age and education, while strongly associated with dietary patterns at the marginal level within the Portuguese population (28,68), appear in the present network as either disconnected nodes. This specific network topology is mathematically consistent with the d-separation properties that govern conditional independence structures within the Bayesian Network framework (46).

The regional gradient observed in the present analysis with Centro exhibiting the highest probability of healthy eating (0.225) and the Lisbon Metropolitan Area (0.063), Alentejo (0.088), and Algarve (0.054) exhibiting the lowest is consistent with the pattern reported by other studies who identified Lisboa Metropolitan Area and Alentejo as regions with the highest odds of low Mediterranean diet adherence after adjustment for sociodemographic covariates (28). The convergence is particularly notable given the methodological differences between the two analyses: Mendonça *et al.* used an adjusted multivariable logistic regression on a sample more than six times larger than the present analytical sample, while the present analysis used unsupervised structure learning on a more compact variable set.

The geographic positioning of Centro as the region with the highest dietary quality is broadly consistent with national epidemiological data mapping the persistence of traditional dietary patterns outside major metropolitan centres (42). This regional trend reflects a higher adherence to traditional food guidelines characterised by an abundant intake of fresh fish, vegetables, and legumes, which historically define the country's rural and coastal food heritage (78). Conversely, the lower probability of healthy eating observed in the Lisbon Metropolitan Area aligns with multiple lines of evidence regarding the structural impact of urbanisation on the Portuguese population (42). This transition is heavily characterised by a significant rise in

ultra-processed food consumption and a greater reliance on out-of-home food preparation inherent to dense, industrialised urban environments (68).

5.7.2. Smoking status and dietary quality

The finding that smoking status emerged as one of only two direct parents of healthy eating in the consensus graph aligns with a substantial body of international literature documenting an inverse association between smoking and overall dietary quality. In a cross-sectional analysis of 80,296 United Kingdom adults from the Nuffield Health database, demonstrated that smokers exhibited consistently lower diet quality across multiple eating behaviours and food groups compared with non-smokers, with effects moderated by age, sex, and socioeconomic status (36). Evidence analysing the NHANES 2005–2018 database using the Healthy Eating Index-2020 has underscored the profound cross-sectional disparities in diet quality scores across different smoking statuses (51). These epidemiological evaluations demonstrate significantly lower diet quality among current tobacco users compared to former and never smokers. Crucially, the literature indicates a behavioural convergence over time, with former smokers gradually recovering dietary quality levels to match those of never smokers within a clear timeline post-cessation (63,64). Furthermore, large-scale population evidence has firmly established the inverse relationship between smoking habits and adherence to validated healthy dietary patterns, confirming that smoking intensity deeply modulates food choices and nutritional outcomes (36).

Two specific aspects of the present finding warrant particular attention against this background. First, the magnitude of the difference observed in the present analysis, a posterior probability of healthy eating of 0.045 for current smokers compared with 0.173 for never smokers, corresponding to a 68% relative reduction, is consistent in direction and broadly proportional to the differentials reported in the studies cited above, despite the substantial differences in outcome definition and sample size. Second, the unexpected finding that ex-smokers in the present analysis exhibited the highest posterior probability of healthy eating (0.188), exceeding even never-smokers (0.173), is consistent with established evidence on post-cessation behavioural trajectories, where former smokers often exhibit significant improvements in health behaviours that can lead to dietary patterns rivalling or slightly exceeding those of lifelong non-smokers (64,65). Two mechanisms are typically invoked to explain this pattern in the literature: (i) ex-smokers represent a behaviourally selected subgroup characterized by a successful change in a major health-risk behaviour, which frequently acts as a catalyst for a broader engagement with multiple health-promoting practices (66,67); and (ii) the post-cessation period is actively associated with deliberate dietary restructuring aimed at managing weight gain and mitigating cardiovascular risk (64). Both interpretations are highly compatible with the Bayesian Network framework, where the smoking status node operates as an effective proxy for an underlying latent behavioural cluster rather than as an isolated causal determinant of dietary choices in itself (63).

This latent behavioural shift finds strong empirical alignment in recent large-scale epidemiological research. Notably, an analysis of the NHANES 2005–2018 cohort using the Healthy Eating Index-2020

framework (89) reported significantly lower diet quality scores in current smokers (49.2) compared with both former (54.0) and never smokers (53.3). Crucially, this study demonstrated that former smokers not only bridge the nutritional gap but achieve dietary quality levels indistinguishable from and occasionally exceeding — those of never smokers within five to ten years of cessation (89) striking parallel between these international scores and the posterior probabilities generated by our Bayesian Network model underscores a robust cross-cultural phenomenon: tobacco cessation often triggers an over-compensatory health rebound, solidifying the role of smoking status as a critical marker for comprehensive lifestyle reorientation.

5.7.3. Bayesian network applications to dietary outcomes

Applications of Bayesian Networks to dietary outcomes remain comparatively rare in the literature, and the present study directly addresses this documented research gap (48,77)**.** Currently, the most directly comparable methodological precedent is a mixed-effect Bayesian Network developed to model the intricate relationships between dietary intake and continuous biological responses such as blood cholesterol, glucose, and insulin within a Finnish dietary intervention cohort (90). Although the primary outcome variable in that framework consisted of continuous biomarkers rather than the binary composite behavioural variables utilised in this dissertation, the core methodological framework remains highly analogous to our analysis (90). Specifically, both approaches rely on structure learning from observational data coupled with a rigorous bootstrap-based stability assessment. The empirical finding that personal and population-wide effects can be jointly captured within a single network strongly supports our conceptual choice of Bayesian Networks as a powerful and appropriate framework for representing the multivariate, interconnected structure of dietary behaviour (90).

Beyond the specific Bayesian Network literature, recent work in nutritional epidemiology has increasingly used Bayesian latent class models and other Bayesian approaches to identify dietary patterns from survey data (91). These approaches share with the present analysis a focus on representing dependence structures rather than estimating single-equation regression coefficients but differ in their focus on unsupervised pattern identification rather than supervised prediction of a defined target. The present study contributes to this small but growing literature by demonstrating that Bayesian Networks can be applied to a binary dietary quality target in a moderate-sized national survey sample, that the resulting network structure is stable under bootstrap resampling, and that the probabilistic predictions are calibrated with a Brier Skill Score above zero a level of methodological completeness that, to the best of our knowledge, has not previously been reported for Bayesian Network models of overall dietary quality in a European population.

5.7.4. Synthesis

Taken together, the convergence between the present findings and the prior literature is substantive on three counts. First, the variables identified by the Bayesian Network as direct parents of the target smoking

status and region of residence are both supported as relevant determinants of dietary quality by population-based Portuguese studies using independent samples and alternative methodologies (28,87,88). Second, the magnitude and direction of the smoking effect, including the elevated posterior probability among ex-smokers, are consistent with cross-national evidence using validated diet quality indices (36,89). Third, the regional gradient with Centro exhibiting the highest and the southern regions and the Lisbon Metropolitan Area exhibiting the lowest probabilities aligns with previously documented patterns of Mediterranean and Atlantic diet adherence in mainland Portugal. The divergence between the present analysis and the prior regression-based literature lies not in the variables identified but in their structural positioning: variables such as age, education, and body mass index, which appear as significant predictors in regression models, are represented in the network but did not connect with the target variable.

5.8. Specific contributions of this study

The present study makes three specific contributions to the literature, which can be organised along methodological, substantive, and applied dimensions. These contributions are framed in relation to the research question articulated in Section 1.1, namely, how Bayesian Networks can contribute to describing dependencies and predicting healthy eating in the Portuguese population, and should be understood as complementary rather than hierarchical.

5.8.1. Methodological contribution

The first contribution of this study is methodological. To the best of our knowledge, this is the first application of Bayesian Network structure learning to the prediction of overall dietary quality in the Portuguese adult population, and one of the few in the broader European literature. Prior applications of Bayesian Networks to dietary outcomes have predominantly focused on undernutrition and food insecurity in low- and middle-income countries (77), on personalised metabolic responses to dietary intervention (90), or on the identification of latent dietary patterns through Bayesian latent class models (91). The application of structure learning with bootstrap-based stability assessment, consensus graph construction, and explicit threshold optimisation to a composite binary measure of dietary quality in a Mediterranean population addresses a documented gap in the literature.

Three specific methodological elements of the present analysis warrant explicit recognition. First, the network framework permits the systematic separation of direct from indirect determinants of dietary behaviour, a distinction that is not accessible through standard regression-based approaches. The finding that body mass index, age, and educational level appear consistently as either disconnected or indirectly connected nodes in the consensus graph, rather than as direct parents of the target, is not a contradiction of the regression-based literature in which these variables emerge as significant predictors. It is a refinement of that literature: the

network identifies these variables as being conditionally independent of the target once smoking status and region of residence are considered. This d-separation property of the underlying conditional independence structure is invisible to logistic regression, which estimates marginal associations conditional on a fixed adjustment set without distinguishing between direct effects and effects mediated through other covariates.

Second, the network supports probabilistic what-if reasoning over arbitrary combinations of observed evidence. The posterior probability of healthy eating can be computed for any combination of values across the set of nodes via exact inference on the compiled junction tree, enabling scenario-based planning that is not available from regression coefficients or from Naïve Bayes probability tables. This capacity is conceptually relevant for public health risk stratification, where the question of interest is often not the marginal effect of a single covariate but the joint posterior probability of an outcome under a specific configuration of behavioural and demographic states.

Third, the present analysis demonstrates the feasibility and limits of Bayesian Network estimation in a sample of moderate size with severe class imbalance. The systematic sensitivity analyses across arc strength thresholds, alternative variable sets, age-stratified subsamples, synthetic minority oversampling (SMOTENC), weighted bootstrap resampling, and alternative fold stratification seeds provide an explicit account of how the principal structural and predictive findings of the model are affected by methodological choices. The stability of the two direct parents of the target node across all valid configurations, together with the documented destruction of interpretable network structure under SMOTENC, contributes to the methodological literature on the application of probabilistic graphical models to imbalanced categorical data.

5.8.2. Substantive contribution

The second contribution of this study is substantive. Beyond confirming the previously documented associations between sociodemographic and behavioural factors and dietary quality in the Portuguese population (28,87,88), the present analysis identifies the conditional independence structure relating these variables to one another and to healthy eating. Two substantive findings emerge from this structural analysis that are specific to the Bayesian Network framework and would not have been recoverable from regression-based approaches alone.

First, smoking status and region of residence are identified as the only two direct parents of healthy eating. Smoking status is a well-established correlate of dietary quality in international literature (36,89), and its emergence as a direct parent in the network is therefore consistent with the broader evidence base. Region of residence, by contrast, has typically been treated as a control variable rather than as a primary determinant in prior Portuguese analyses, and its emergence as a direct parent rather than as a mediated or confounding variable represents a substantive refinement of the existing literature. The interpretation of region as a contextual variable encodes a constellation of underlying factors (food availability, retail environment, dietary culture, socioeconomic infrastructure) that implies that interventions targeting regional dietary disparities will require structural rather than purely behavioural strategies.

Second, the conditional probability tables of the network reveal a coherent profile of populations at high risk of inadequate dietary patterns: active smokers residing in southern Portuguese regions or in the Lisbon Metropolitan Area, with a joint probability of healthy eating approaching zero in the most adverse configurations. The corresponding favourable profile for ex-smokers or never-smokers residing in the Centro region exhibits a probability of healthy eating approaching one third, more than two orders of magnitude higher than the worst-case combination. This kind of stratified probabilistic characterisation is a distinctive product of the network framework and provides an operational basis for population-level risk stratification that is not directly available from regression coefficients.

5.8.3. Applied contribution

The third contribution of this study is applied. The Bayesian Network estimated in this analysis is fully specified, validated through cross-validation, and accompanied by calibrated posterior probability estimates and a calibration error of 0.030. The model uses only routinely collected behavioural and geographic information, smoking status, and region of residence to identify approximately 70% of individuals with inadequate dietary patterns. In a population where inadequate dietary patterns are a leading modifiable contributor to the national burden of non-communicable disease, this level of discriminative capacity from minimal input information has potential value as a triage instrument in primary care settings.

Two specific applied implications follow from this. First, the network can be deployed as a probabilistic risk-stratification tool to identify subgroups of the Portuguese adult population most likely to benefit from targeted dietary interventions, based on information that is already routinely captured in family medicine and public health surveillance contexts. This is particularly relevant given the documented disparities between regional patterns of healthy eating and the geographic distribution of preventive nutritional services in mainland Portugal. Second, the explicit modelling of the joint conditional distribution of smoking status and region of residence rather than the marginal effect of each supports the design of integrated lifestyle interventions that address both behavioural and contextual determinants simultaneously, in line with the broader literature on the clustering of health-related behaviours.

Beyond these specific applied implications, the present study contributes to the wider methodological literature by demonstrating that Bayesian Networks can be applied productively to moderately sized national survey samples with severe class imbalance, provided that methodological choices are made transparent through systematic sensitivity analyses. The findings reported here, both positive and negative, are intended as a methodological reference for subsequent applications of probabilistic graphical models to dietary and other behavioural outcomes in the portuguese population.

5.9. Strengths and limitations

This section discusses the strengths and limitations of the present study. The limitations are organised in three groups according to the level at which they constrain the validity, the interpretability, or the practical applicability of the findings: limitations of the data source, limitations associated with the operationalisation of the target variable, and limitations regarding the translation of the model into clinical practice.

5.9.1. Strengths

The first strength of this study is methodological transparency. The structure learning procedure, the bootstrap stability assessment, the consensus graph construction, the threshold optimisation, the parameter estimation procedure, and the cross-validation framework are reported in sufficient detail to permit full replication of the analytical pipeline by an independent researcher with access to the underlying dataset. The systematic sensitivity analyses across arc strength thresholds, alternative variable sets, age-stratified subsamples, synthetic minority oversampling, weighted bootstrap resampling, and alternative fold stratification seeds provide an explicit map of the dependence between methodological choices and substantive findings and are uncommon in the comparable literature on Bayesian Network applications in nutritional epidemiology.

The second strength is the use of a fully probabilistic framework. The Bayesian Network produces calibrated posterior probabilities of healthy eating conditional on any combination of observed evidence, with a calibration error below the 0.05 threshold typically used in the clinical prediction literature. This calibration property is rare among machine learning approaches applied to imbalanced categorical outcomes and is a direct consequence of the explicit modelling of the joint distribution rather than of a discriminative decision boundary.

The third strength is interpretability. The estimated network represents the conditional independence structure of the data as an explicit directed acyclic graph, and the conditional probability tables of the target node and its parents are directly inspectable. This contrasts with black-box machine learning models — gradient-boosted trees, random forests, or neural networks — whose internal representations are typically not accessible without additional post-hoc explanation methods. In the context of public health applications, where transparency of the underlying mechanism is a substantive rather than a cosmetic requirement, this property of the Bayesian Network constitutes a meaningful methodological advantage.

The fourth strength is the use of a recent, population-based portuguese dataset collected through a study conducted by the Faculty of Medicine of the University of Porto in 2019 (5,8), with detailed information on dietary, behavioural, demographic, and health-related variables. While the analytical sample of 847 adults is moderate in size and not formally representative of the national population across all dimensions, it was representative in terms of sex, age group, and NUTS II regional distribution, and the dataset captures sufficient regional and behavioural variability to support meaningful structural inference.

5.9.2. Limitations of the data source

The first and most consequential limitation of this study concerns the data source itself. The dataset originates from a single study conducted by the Faculty of Medicine of the University of Porto in 2019 (5,8), and the analytical sample of n = 847 (originally n=891) adults is moderate by the standards of structural learning, which generally benefits from sample sizes in the thousands. While the bootstrap-based stability assessment partially mitigates the consequences of finite-sample noise on the consensus graph, it does not eliminate them: arcs with intermediate stability scores (in the 0.3–0.5 range) remain susceptible to sampling variability, and the choice of the arc strength threshold introduces a structural sensitivity that has been documented in the present analysis. A larger sample would be expected to reduce both the width of the stability distribution and the sensitivity of the consensus graph to the choice of threshold.

The second data limitation is the severe class imbalance of the target variable. Only approximately 14% of the analytical samples meet the composite criterion for healthy eating, which substantially complicates both the structure learning and the predictive evaluation phases of the analysis. Class imbalance reduces the effective information available for estimating the conditional probability distributions associated with the minority class, inflates the variance of cross-validation performance estimates, and creates the requirement for threshold optimisation that has been documented in the methodological chapter (Chapter 3). The use of stratified cross-validation, the application of F1-optimised thresholds, and the systematic evaluation of synthetic minority oversampling (SMOTENC) all address aspects of this limitation, but none of these procedures recovers the information that is missing in the underrepresentation of the positive class.

The third data limitation is the cross-sectional design. The dataset captures dietary, behavioural, demographic, and health-related variables at a single point in time, and no causal interpretation of the directed arcs in the consensus graph is warranted. The orientation of arcs in the Bayesian Network reflects the conditional independence structure of the joint distribution, not temporal precedence or causal mechanism. Two specific consequences follow. First, the relationship between smoking status and healthy eating cannot be interpreted as a causal effect of smoking on diet, nor of diet on smoking; both directions are consistent with the same observed dependence pattern. Second, the absence of follow-up data precludes any evaluation of whether the network would predict the future onset, persistence, or change of dietary behaviour, which is the prediction question that would be most relevant for preventive applications.

The fourth data limitation is the reliance on self-reported dietary information. The items used to construct the target variable number of main meals per day, frequency and portion size of fruit, vegetable, salad, and legume consumption, and alcohol intake are subject to the well-documented sources of measurement error that affect food frequency and short dietary recall instruments, including social desirability bias and recall bias. While these biases tend to attenuate associations rather than create spurious ones, their effect on a composite binary outcome that requires the simultaneous satisfaction of multiple criteria is non-linear and difficult to quantify; even small classification errors in any one component can shift the binary classification of an individual.

The fifth data limitation is the absence of variables that the international nutritional literature increasingly identifies as central to dietary quality, including ultra-processed food consumption, added sugar intake, sodium intake, dietary fibre, and the consumption of red and processed meat. The dataset does not contain detailed instruments for any of these components, and the target variable therefore captures a specific food-group-based notion of dietary quality that does not exhaust the contemporary definition of a healthy diet. This restriction is acknowledged in Chapter 3 and is consistent with the food-based operationalisation adopted from the World Health Organisation framing, but it does limit the comparability of the present findings with studies that use more comprehensive instruments such as the Healthy Eating Index, the Alternative Healthy Eating Index, or the Mediterranean Diet Score.

5.9.3. Limitations associated with the operationalisation of the target variable

Beyond the limitations of the data, the operationalisation of the target variable itself imposes a set of constraints on the interpretation of the findings that deserve explicit discussion. Healthy eating is, in its substantive sense, a continuous and multidimensional construct: it varies along multiple food groups, across nutrients, in relation to caloric intake, and over time. The reduction of this construct to a binary indicator is a methodological necessity dictated by the requirement that the target node of a Bayesian Network be categorical, but it introduces a substantial loss of information and a set of interpretative ambiguities that should be acknowledged.

First, the binary cut-off is asymmetric in its tolerance of departures from the criterion. The composite definition requires simultaneous satisfaction of all component conditions: appropriate meal structure, daily fruit intake within the recommended portion range, daily vegetable and legume intake within the recommended portion range, and non-excessive alcohol consumption for an individual to be classified as adopting a healthy diet. An individual who satisfies three of the four conditions but fails one is classified identically to an individual who fails all four. This all-or-nothing structure introduces a substantial collapse of the underlying variation: two individuals at very different positions on a continuous dietary quality scale may receive the same binary label, while two individuals at very close positions may receive opposite labels if one of them lies just above and the other just below a threshold for any of the component conditions.

Second, the binary outcome represents a proxy for healthy eating, not a direct measurement of it. The conditions defining the target capture food-based behaviours that are characteristic of healthy dietary patterns according to the World Health Organisation framing and the Mediterranean dietary pattern, but they do not exhaust the construct. An individual classified as having a healthy diet under the present operationalisation may nonetheless exhibit excessive intake of ultra-processed foods, added sugars, sodium, or saturated fats that are not captured by the available items; conversely, an individual classified as not healthy may fail the criterion only on the alcohol or meal-structure component while exhibiting an otherwise favourable nutrient profile. The findings of the network should therefore be interpreted as identifying predictors of a specific food-based proxy,

not as predictors of dietary quality in the comprehensive sense that the term carries in the broader nutritional literature.

Third, the prevalence of the positive class, approximately 14%, is a direct consequence of the conjunction of strict criteria, not necessarily a reflection of the actual prevalence of healthy eating in the Portuguese adult population under more permissive definitions. Alternative operationalisations of the same construct, using continuous adherence scores, partial satisfaction criteria, or component-specific outcomes, would yield substantially different prevalence distributions and would likely also yield different conditional independence structures. The robustness of the principal structural findings of the present analysis, particularly the identification of smoking status and region of residence as direct parents to alternative operationalisations of the target variable, was not formally evaluated and constitutes an area for further research.

Fourth, the requirement for simultaneous satisfaction of multiple conditions implies that the model effectively predicts a tail of the distribution of dietary behaviours rather than its centre. The substantive question being answered by the network is not what predicts the average level of dietary quality in the Portuguese population, but what predicts membership in the relatively small subgroup of individuals who satisfy a particularly demanding composite criterion. This distinction affects the interpretation of the conditional probability tables and the public health framing of the findings: the network identifies factors associated with the achievement of a stringent dietary standard, not factors associated with incremental dietary improvement.

5.9.4. Limitations for clinical and public health application

Beyond the data and the operationalisation, a third set of limitations concerns the translation of the findings of the present study into clinical practice. Although the network produces calibrated posterior probabilities and identifies a coherent profile of populations at high risk of inadequate dietary patterns, the model does not, on its own, resolve the underlying public health problem of inadequate diet in the Portuguese population, and several considerations need to be raised about its practical applicability.

First, the variables identified by the network as direct parents of healthy eating, smoking status and region of residence are not, in the immediate sense, clinically actionable in the context of a dietary intervention. Region of residence is a structural characteristic of an individual that cannot be modified through clinical practice; it acts as a marker of an underlying constellation of contextual factors (food availability, retail environment, dietary culture, socioeconomic infrastructure) whose modification requires structural rather than clinical interventions. Smoking status, while modifiable, is itself the target of a substantial body of clinical and public health practice in its own right, and its identification as a correlate of dietary quality does not generate a novel intervention pathway: the recommendation to support smoking cessation already exists on independent grounds, and the present analysis at best reinforces this recommendation rather than producing a distinctively dietary one.

Second, the model identifies predictors at a population level, not individual mechanisms or barriers. A primary care clinician seeking to support an individual patient in improving dietary quality requires information about the patient's specific motivations, food preferences, family context, financial constraints, time availability, and cooking skills, none of which is captured by the network. The probabilistic risk stratification that the network supports is informative for the design of population-level interventions and for the prioritisation of resources, but it does not substitute for the clinical assessment of individual determinants of dietary behaviour. The identification of an individual as belonging to a high-risk subgroup (active smoker, residing in a southern region or in the Lisbon Metropolitan Area) provides a prior probability, not a clinical pathway.

Third, the discriminative performance of the model, while improved over baseline classifiers and methodologically defensible given the class imbalance and the moderate sample size, remains modest in absolute terms. A PR-AUC of 0.241 and a recall of approximately 0.70 at the optimised threshold imply that approximately three in four individuals classified as having a healthy diet will not actually meet the composite criterion, and that approximately three in ten individuals with inadequate dietary patterns will be missed by the model. These performance levels are insufficient to support clinical decision-making about an individual patient and are best understood as supporting population-level risk stratification rather than individual triage.

Fourth, even if the predictive performance of the model were substantially improved, the underlying problem of inadequate diet in the portuguese population is not, in its substance, an information problem that can be solved by better identification of high-risk subgroups. The determinants of dietary behaviour identified in the present analysis region, smoking status, and the mediated effects of education, age, and body mass index point to a set of structural, behavioural, and contextual factors whose modification requires intervention strategies that operate at the population level (food policy, taxation, retail regulation, food labelling), at the community level (urban food environments, school meals, workplace catering), and at the individual level (motivational interviewing, dietary counselling, behavioural support). The Bayesian Network developed in this study can contribute to the design of such strategies by identifying populations to prioritise and contextual factors to address, but it cannot, by itself, deliver the changes in dietary behaviour that the underlying public health problem requires. The translation of structural insights into measurable improvements in population dietary quality remains a task for the broader public health system, of which this study is one input among many.

5.10. Implications for practice and public health policy

The findings of the present study have implications that operate at three distinct levels of public health intervention: at the level of population surveillance and risk stratification, at the level of primary care and clinical practice, and at the level of structural and policy interventions targeting the broader determinants of dietary behaviour. These three levels are complementary rather than alternative, and the implications described

in this section should be understood as elements of a coordinated public health response rather than as standalone recommendations. The discussion is framed conservatively, considering the limitations described in Section 5.9, and of the cross-sectional nature of the dataset and the proxy character of the target variable. The broader public health relevance of this framing is well established: inadequate dietary patterns are among the leading modifiable risk factors for the chronic non-communicable diseases that account for most of the premature mortality and disability in Portugal and globally (1–4).

5.10.1. Implications for population surveillance and risk stratification

The first implication of the findings concerns the use of probabilistic risk stratification at the population level. The Bayesian Network estimated in this analysis identifies smoking status and region of residence as the two direct parents of healthy eating, and the conditional probability tables of the model permit the joint computation of posterior probabilities of inadequate dietary patterns for any combination of these two variables. The most adverse active smokers residing in the Lisbon Metropolitan Area, Alentejo, or Algarve exhibit a posterior probability of healthy eating approaching zero, while the most favourable profile, ex-smokers or never-smokers residing in the Centro region, exhibits a probability close to one-third. This two-order-of-magnitude gradient, based on two variables that are routinely captured in population health surveillance, provides a parsimonious basis for the prioritisation of nutritional surveillance and intervention resources.

In practical terms, this implies that the regional dimension of nutritional inequality in the portuguese adult population deserves more sustained attention in public health surveillance than it has typically received. The convergence between the regional gradient identified in the present analysis and the pattern documented in independent Portuguese samples where the Lisbon Metropolitan Area and Alentejo are consistently identified as regions with lower dietary quality and Mediterranean diet adherence (28,87) provides additional support for the inclusion of NUTS II region as a primary stratified in dietary surveillance reporting, on the same footing as sex and age, and the disaggregation of nutritional indicators by region in periodic public health reports such as those produced under the National Programme for the Promotion of Healthy Eating (PNPAS) (13).

5.10.2. Implications for primary care and clinical practice

The second implication of the findings concerns the integration of dietary considerations into primary care practice. The model developed in the present analysis does not constitute a clinical decision support tool: the discriminative performance is insufficient for individual triage, the variables identified as direct parents are not directly modifiable through dietary counselling in isolation, and the cross-sectional design precludes any predictive interpretation of future dietary behaviour. The probabilistic framework underlying the Bayesian

Network is, in principle, well-suited to clinical reasoning under uncertainty (6), but the translation of population-level posterior probabilities into individual clinical decisions requires external validation in prospective cohorts and clinical settings that the present analysis cannot provide. Nevertheless, two specific implications for primary care can be drawn from the findings, framed conservatively and in awareness of these constraints.

First, the identification of smoking status as a direct correlate of dietary quality reinforces the case for the integration of dietary assessment into smoking cessation consultations. The evidence base for the association between smoking and dietary quality is consistent and substantial (36), and the period following smoking cessation during which dietary quality typically improves and converges with, or in some sub-scores slightly exceeds, that of never-smokers within five to ten years of cessation (89) represents a window of opportunity for the reinforcement of dietary change. The clinical implication is not that smoking cessation should be motivated by its dietary effects, but that dietary assessment and counselling should be systematically included in the post-cessation follow-up period, when the patient is already engaged in active behavioural change and is likely to be more receptive to the extension of that change to other health domains. This integration is consistent with the broader orientation of Portuguese family medicine towards lifestyle-based prevention of chronic disease (5).

Second, the regional gradient identified in the model has implications for the spatial allocation of preventive nutritional services in primary care. Regions with the lowest probability of healthy eating in the model, Lisbon Metropolitan Area, Alentejo, Algarve, coincide partially with regions where access to dedicated nutritional consultation within primary care settings has been documented as more limited (8). The structural inequality in the supply of preventive nutritional services reinforces the inequality in dietary quality observed in the data, and the present findings provide a quantitative basis for the prioritisation of investment in dietetic services in these regions. This recommendation is consistent with the broader literature on the role of primary care in the prevention of nutrition-related non-communicable disease (4,92) and with the objectives of the PNPAS framework (13).

5.10.3. Implications for structural and policy interventions

The third implication of the findings concerns the level at which the modification of dietary behaviour in the Portuguese adult population is most likely to be achieved. The identification of region of residence as a direct parent of healthy eating, which, points toward the primary importance of contextual rather than individual determinants of dietary quality (42,68). Region of residence is not an isolated causal determinant of diet; rather, it operates as a structural proxy for an underlying constellation of environmental factors. These include the geographic availability and affordability of fresh foods, the density of local food retailers, the prevalence of out-of-home food consumption, and the strong persistence of regional dietary traditions (42,78).

The modification of these structural factors requires policy interventions operating at the population level, of the kind documented in the broader international literature on food environment regulation and its effects on dietary behaviour (19,26).

Three specific policy domains follow from this framing. First, food environment regulation, including measures such as fiscal policy on sugar-sweetened beverages, restrictions on the marketing of ultra-processed foods, mandatory front-of-pack nutritional labelling, and incentives for the retail availability of fresh foods — has been documented in international comparative analyses as an effective lever for the modification of population dietary patterns (4,26). The Portuguese policy framework has implemented several of these measures under the PNPAS (13), and the findings of the present analysis are consistent with the continued strengthening of this regulatory approach, particularly in regions with the most adverse dietary profiles. The absence of variables capturing ultra-processed food consumption from the present dataset (45,51) further underscores the need for expanded nutritional surveillance instruments that capture the full range of dietary quality determinants in future population studies.

Second, school-based and workplace-based food environment interventions represent a policy domain where the regional gradient identified in the present analysis has direct relevance. Barriers to healthy food access at the household level, including income constraints, time availability, and cooking behaviours (17,19), are unevenly distributed across Portuguese regions and compound the structural food environment disadvantage already documented in the southern regions and in the Lisbon Metropolitan Area. School meal programs, workplace catering standards, and community food initiatives operating at the level of daily food exposure have the potential to shift dietary patterns at the population level in a way that individual clinical counselling cannot replicate (26).

Third, the integration of dietary policy with broader social and economic policy is supported by the findings of the present analysis. The mediated rather than direct position of education and income-related variables in the consensus graph implies that the dietary effects of socioeconomic disadvantage operate through behavioural and contextual mechanisms that are themselves subject to intervention. This finding is consistent with international evidence on the structural determinants of dietary quality, which identifies household income, food insecurity, and access to public services as upstream determinants of dietary behaviour whose effects are partially mediated through environmental and behavioural pathways (19,26,88).

5.10.4. The limits of public health policy and the need for coordinated action

A final consideration deserves explicit recognition. The implications described in this section operate at different levels of the public health system and require different actors, time horizons, and instruments. Surveillance and risk stratification operate within the existing administrative infrastructure and can be implemented in the short term. Primary care integration requires the coordination of family medicine, nutrition, and public health services within the National Health Service, and operates on a medium-term horizon.

Structural and policy interventions operate on the longest time horizon and require coordination across sectors of government that traditionally function independently of one another. The findings of the present study do not privilege any one of these levels over the others; they are intended to contribute evidence that supports the coordinated action of all three.

The substantive public health problem of inadequate diet in the Portuguese adult population with its documented contribution to the national burden of cardiovascular disease, type 2 diabetes, and obesity-related conditions (1–3) will not be resolved by any single intervention at any single level. Dietary behaviour is determined by a complex system of interacting individual, social, environmental, and structural factors (26), and interventions that address only one layer of this system are unlikely to produce sustained population-level changes. The Bayesian Network developed in this study can contribute to the design of coordinated multi-level strategies by identifying populations to prioritise and contextual factors to address, but the translation of structural insights into measurable improvements in population dietary quality remains a task for the broader public health system, of which this study is one input among many.

5.11. Future work

The limitations and findings of the present study point to several directions for future research, organised here from the most proximate extensions of the current analysis to broader methodological and applied developments.

The most immediate priority is replication in a substantially larger sample of the Portuguese adult population. The moderate sample size of n = 847 constrains the stability of arc strength estimates under bootstrap resampling and limits the granularity of regional and subgroup analyses, most notably the near-empty cells observed for smokers in the Alentejo and Algarve regions, which render the conditional probability estimates for those combinations unreliable. A nationally representative dietary survey with a sample in the thousands would permit tighter bootstrap stability distributions, finer NUTS II disaggregation, and more robust cross-validation estimates across all age strata.

The near-random discriminative performance observed in the 65 and older stratum (PR-AUC = 0.162, approaching baseline) constitutes a specific finding that warrants dedicated investigation. Dietary behaviour in older adults is shaped by a distinct constellation of factors: polypharmacy, functional limitations, social isolation, chewing and swallowing difficulties, and the cumulative effects of multi-morbidity that are not adequately captured by the current variable set. Future work should develop age-specific models for older adults using instruments designed to capture these dimensions, potentially integrating clinical and functional data from primary care records.

The cross-sectional design of the underlying dataset precludes any causal or longitudinal interpretation of the findings. A longitudinal extension of the study design tracking dietary behaviour, smoking status, and regional mobility over time would permit the estimation of dynamic Bayesian Networks, which

explicitly model temporal dependencies and allow inference about the direction and magnitude of change in dietary behaviour following events such as smoking cessation, regional relocation, or health status transitions (49). This is particularly relevant for the ex-smoker finding: the present analysis can only establish that ex-smokers exhibit higher posterior probabilities of healthy eating than never-smokers at a single time point; a longitudinal design would permit the modelling of dietary trajectory following cessation.

The target variable operationalisation is a further area for development. The present binary composite outcome captures a specific, stringent food-based proxy of healthy eating that necessarily excludes dimensions such as ultra-processed food consumption, added sugar and sodium intake, and dietary fibre components that are increasingly central to contemporary definitions of dietary quality (45,64). Future work should explore alternative operationalisations of the target, including continuous adherence scores such as the Healthy Eating Index or the Mediterranean Diet Score, or component-specific binary outcomes, and assess whether the structural findings of the present analysis, in particular the identification of smoking status and region as direct parents, are robust across these alternative definitions.

From a methodological standpoint, two extensions of the Bayesian Network framework merit investigation. First, the application of causal discovery algorithms such as the PC algorithm or FCI, which explicitly test for and represent confounding and latent variable structures, would permit a more rigorous separation of direct causal effects from statistical associations in the consensus graph, addressing a core limitation of the Hill-Climbing approach used here. Second, the integration of the estimated network into a clinical decision support prototype, with prospective evaluation of its utility as a risk stratification tool in primary care consultations, would provide the external validity evidence needed to assess whether the population-level probabilistic risk profiles identified in the present analysis translate into actionable information at the individual clinical level (6).

Finally, the finding that SMOTENC destroyed the interpretable structure of the Bayesian Network, producing a structurally implausible DAG while inflating ROC-AUC, raises a methodological question of broader relevance for the application of probabilistic graphical models to imbalanced categorical data. Systematic investigation of oversampling strategies compatible with the conditional independence assumptions of Bayesian Networks, including techniques that preserve the marginal and joint distributions of the minority class rather than interpolating between categorical observations, would constitute a methodological contribution to the field that extends beyond the specific application reported here.

6. Conclusion

6.1. Summary of the study

This dissertation developed a Bayesian Network model to describe the dependency structure among the determinants of healthy eating in the Portuguese adult population and to estimate individual posterior probabilities of adopting a healthy dietary pattern. The analysis was conducted on data from a population-based study carried out by the Faculty of Medicine of the University of Porto in 2019 (n = 847 adults), with healthy eating operationalised as a binary composite outcome combining regular meal structure, daily fruit and vegetable consumption within recommended portions, and non-excessive alcohol intake (positive class prevalence ≈ 14%).

The modelling pipeline applied bootstrap-based Hill-Climbing structure learning with BDeu scoring (R = 200 resamples) to construct a consensus directed acyclic graph at three arc strength thresholds ($F \geq 0.7$, $F \geq 0.4$, $F \geq 0.3$), followed by Maximum A Posteriori parameter estimation with a Dirichlet prior. Predictive performance was evaluated through stratified 5-fold cross-validation with F1-optimised classification thresholds, using PR-AUC and Brier Skill Score as primary metrics, and benchmarked against Logistic Regression with L1 regularisation and Naïve Bayes baseline classifiers. The methodological robustness of the principal findings was assessed through systematic sensitivity analyses across arc strength thresholds, alternative variable sets, age-stratified subsamples, synthetic minority oversampling (SMOTENC), weighted bootstrap resampling, and alternative fold stratification seeds.

6.2. Answer to the research question

The research question articulated in Section 1.1 — how can Bayesian Networks contribute by describing dependencies and predicting healthy eating in the Portuguese population? — is answered in three parts, corresponding to the three objectives of the study.

On the description of dependencies, the Bayesian Network identified smoking status (*tabagismo_cat3*) and region of residence (*nuts2_cat5*) as the only two direct parents of healthy eating in the consensus graph, a finding that proved structurally robust across all valid threshold and variable-set configurations and across all sensitivity analyses. Consistently reported as significant predictors of dietary quality in regression-based literature, body mass index, age, and educational level appeared in the network as disconnected. This d-separation property of the underlying conditional independence structure is a contribution that is not recoverable from standard regression approaches and represents the substantive methodological value of the network framework.

On the prediction of healthy eating, the final model ($F \geq 0.3$, 32 variables) achieved a PR-AUC of 0.241 — a 68% improvement over the prevalence baseline of 0.143, together with a Brier Skill Score of 0.040 and an Expected Calibration Error of 0.030, indicating that the predicted probabilities are both more accurate than the naive prevalence-based strategy and acceptably calibrated for practical use. The *what-if* analysis revealed a two-order-of-magnitude gradient in the posterior probability of healthy eating, ranging from values approaching zero in active smokers residing in the southern regions to 0.334 in ex-smokers residing in the Centro region. Ex-smokers exhibited the highest posterior probability of healthy eating among the three smoking categories (0.188), exceeding even never-smokers (0.173), a finding consistent across the two largest regions of the sample (Norte and Centro) and aligned with international evidence on dietary improvement following smoking cessation.

On the evaluation of performance and robustness, the comparison with baseline classifiers placed the Bayesian Network within the same performance range as Logistic Regression with L1 regularisation (PR-AUC = 0.293) and Naïve Bayes (PR-AUC = 0.304), with the baseline classifiers achieving marginally higher discriminative performance. This comparison should not be read as a deficit: the value of the Bayesian Network lies in the learned dependency structure and in the probabilistic what-if reasoning it supports, neither of which is replicable through regression or Naïve Bayes. The systematic sensitivity analyses confirmed the structural robustness of the two direct parents of the target across all valid configurations and documented that synthetic minority oversampling destroyed the interpretable network structure, a methodologically informative negative finding.

6.3. Final considerations

The findings of this study contribute to the limited but growing literature on the application of Bayesian Networks to dietary outcomes in well-developed food environments and address a documented gap in the Portuguese nutritional epidemiology literature. The identification of smoking status and region of residence as the two structural determinants of healthy eating has implications that extend beyond the methodological framework of the analysis itself. It supports the integration of dietary assessment into smoking cessation programs within primary care, the prioritisation of nutritional surveillance and preventive services in structurally disadvantaged regions, and the framing of dietary policy as a coordinated multi-level public health response rather than as an individual behavioural concern.

At the same time, the limitations of the analysis must be retained in the interpretation of its conclusions. The moderate sample size, the severe class imbalance, the cross-sectional design, the reliance on self-reported dietary data, and the proxy nature of the composite target variable all constrain the strength of the inferences that can be drawn. The discriminative performance of the model, while methodologically defensible given the constraints, is not sufficient to support individual clinical triage, and the variables identified as direct parents are not, in themselves, modifiable through dietary counselling alone. The Bayesian Network developed

in this study should therefore be understood as a methodological proof of concept and a population-level analytical instrument, not as a clinical decision support tool ready for direct deployment.

Beyond the specific findings, this dissertation contributes a methodological reference for the application of Bayesian Network structure learning to imbalanced categorical dietary outcomes in moderately sized national survey samples, with explicit documentation of methodological choices, sensitivity analyses, and limitations encountered. The reproducibility of the pipeline, the transparency of the structural and predictive findings, and the conservative framing of the implications are intended to support both replication in larger samples and the broader integration of probabilistic graphical models into the toolkit of nutritional epidemiology in Portugal.

The translation of the structural insights identified in this analysis into measurable improvements in the dietary quality of the Portuguese adult population and, ultimately, into a reduction of the burden of diet-related non-communicable disease remains a task for the broader public health system, of which this study is one contribution among many.

References

1. Noncommunicable diseases [Internet]. [cited 2026 Jan 20]. Available from: https://www.who.int/health-topics/noncommunicable-diseases

2. Henriques ADA. CITATION: direção-geral da saúde, institute for health metrics and evaluation. Portugal: the nation's health 1990–2016: an overview of the global burden of disease study 2016 results. Seattle, WA: IHME, 2018. 1–32.

3. Análise DD de S de I e. A saúde dos portugueses 2016. 2017 Apr 3;1–177.

4. Shang X, Liu J, Zhu Z, Zhang X, Huang Y, Liu S, et al. Healthy dietary patterns and the risk of individual chronic diseases in community-dwelling adults. Nat Commun. 2023 Oct 23;14(1):6704. doi:10.1038/s41467-023-42523-9

5. Páscoa R, Teixeira A, Gregório M, Carvalho R, Martins C. Patients' perspectives about lifestyle behaviors and health in the context of family medicine: a cross-sectional study in Portugal. Int J Environ Res Public Health. 2021 Mar 14;18(6):2981. doi:10.3390/ijerph18062981

6. Webb MPK, Sidebotham D. Bayes' formula: a powerful but counterintuitive tool for medical decision-making. BJA Educ. 2020 Jun 1;20(6):208–13. doi:10.1016/j.bjae.2020.03.002

7. Healthy diet [Internet]. [cited 2025 Feb 8]. Available from: https://www.who.int/news-room/fact-sheets/detail/healthy-diet

8. Páscoa R, Teixeira A, Monteiro H, Prazeres F, Martins C. Association of lifestyle and sociodemographic factors on multimorbidity: a cross-sectional study in Portugal. BMC Public Health. 2022 Dec 14;22(1):2341. doi:10.1186/s12889-022-14640-5

9. Trichopoulou A. Mediterranean diet: the past and the present. Nutr Metab Cardiovasc Dis. 2001 Aug 1;11(4 Suppl):1–4. PubMed PMID: 11894739.

10. Trichopoulou A, Vasilopoulou E. Mediterranean diet and longevity. 1–5.

11. Monteagudo C, Mariscal-Arcas M, Rivas A, Lorenzo-Tovar ML, Tur JA, Olea-Serrano F. Proposal of a Mediterranean diet serving score. Blachier F, editor. PLOS One. 2015 Jun 2;10(6):e0128594. doi:10.1371/journal.pone.0128594

12. Dernini S, Berry EM. Mediterranean diet: from a healthy diet to a sustainable dietary pattern. Front Nutr. 2015 May 7;2:1–7. doi:10.3389/fnut.2015.00015

13. alimentacaosaudavel.dgs.pt/wp-content/uploads/2023/07/Eat-Drink-and-Live-2023-PNPAS.pdf [Internet]. [cited 2026 Jan 21]. Available from: https://alimentacaosaudavel.dgs.pt/wp-content/uploads/2023/07/Eat-Drink-and-Live-2023-PNPAS.pdf

14. Trichopoulou A, Martínez-González MA, Tong TY, Forouhi NG, Khandelwal S, Prabhakaran D, et al. Definitions and potential health benefits of the Mediterranean diet: views from experts around the world. BMC Med. 2014 Jul 24;12:112–27. doi:10.1186/1741-7015-12-112 PubMed PMID: 25055810; PubMed Central PMCID: PMC4222885.

15. Trichopoulou A, Costacou T, Bamia C, Trichopoulos D. Adherence to a Mediterranean diet and survival in a greek population. N Engl J Med. 2003 Jun 26;348(26):2599–608. doi:10.1056/NEJMoa025039 PubMed PMID: 12826634.

16. Godos J, Castellano S, Marranzano M. Adherence to a Mediterranean dietary pattern is associated with higher quality of life in a cohort of italian adults. Nutrients. 2019 Apr 29;11(5):981–92. doi:10.3390/nu11050981 PubMed PMID: 31035736; PubMed Central PMCID: PMC6566890.

17. Kegler MC, Raskind IG, Bundy ŁT, Owolabi S, Veluswamy JK, Hernandez C, et al. Barriers to creating healthier home food environments: process evaluation results from two home food environment intervention studies. Fam Community Health. 2022 Apr 28;45(4):247. doi:10.1097/FCH.0000000000000335 PubMed PMID: 35536715.

18. Escoto K, Laska M, Larson N, Neumark-Sztainer D, Hannan P. Work hours and perceived time barriers to healthful eating among young adults. Am J Health Behav. 2012 Nov 1;36:786–96. doi:10.5993/AJHB.36.6.6

19. Wolfson JA, Ramsing R, Richardson CR, Palmer A. Barriers to healthy food access: associations with household income and cooking behavior. Prev Med Rep. 2019 Jan 31;13:298–305. doi:10.1016/j.pmedr.2019.01.023 PubMed PMID: 30792944; PubMed Central PMCID: PMC6363918.

20. Obeid C, Oenema A, Jaalouk D, Kremers SPJ, Gubbels JS. Determinants of adherence to the Mediterranean diet among adults in Mediterranean countries: a systematic literature review. Public Health Nutr. 28(1):e194. doi:10.1017/S1368980025101432 PubMed PMID: 41199361; PubMed Central PMCID: PMC12722083.

21. Orji R, Mandryk RL, Vassileva J. Towards a data-driven approach to intervention design: a predictive path model of healthy eating determinants. In: Bang M, Ragnemalm EL, editors. Persuasive Technology. Design for Health and Safety. Berlin, Heidelberg: Springer; 2012. p. 203–14. doi:10.1007/978-3-642-31037-9_18

22. Wongprawmas R, Mora C, Pellegrini N, Guiné RPF, Carini E, Sogari G, et al. Food choice determinants and perceptions of a healthy diet among italian consumers. Foods (Basel Switz). 2021 Feb 3;10(2):318–40. doi:10.3390/foods10020318 PubMed PMID: 33546323; PubMed Central PMCID: PMC7913531.

23. Aji AS, Rohana AJ, Geik OP, Rafdinal W, Mohammad WMZW, Yasin MAM, et al. Associations between food-related behaviours, nutrient intake and nutritional status through structural equation model (SEM) among clients undergoing community-based treatment and rehabilitation (CBTaR): a cross-sectional study in kelantan, malaysia. BMJ OPEN. 2026 Jan 1;16(1):e109777. doi:10.1136/bmjopen-2025-109777 PubMed PMID: 41513420.

24. Mollaei S, Roshani D, Salari-Moghddam A, Moradi Y, Moradpour F. Dietary and lifestyle predictors of obesity: a structural equation model approach [Internet]. Research Square; 2025 [cited 2026 Jan 24]. Available from: https://www.researchsquare.com/article/rs-7049033/v1 doi:10.21203/rs.3.rs-7049033/v1

25. Martins RC de O, Caldeira TCM, Soares MM, Mais LA, Claro RM. Leisure screen time and food consumption among brazilian adults. Int J Environ Res Public Health. 2024 Aug 26;21(9):1123. doi:10.3390/ijerph21091123 PubMed PMID: 39338006; PubMed Central PMCID: PMC11431508.

26. Zorbas C, Palermo C, Chung A, Iguacel I, Et. A. Factors perceived to influence healthy eating: a systematic review and meta-ethnographic synthesis of the literature. Nutr Rev. 2018;1–14. doi:10.1093/nutrit/nuy043

27. Andrade V, Jorge R, García-Conesa MT, Philippou E, Massaro M, Chervenkov M, et al. Mediterranean diet adherence and subjective well-being in a sample of portuguese adults. Nutrients. 2020 Dec 16;12(12):3837. doi:10.3390/nu12123837 PubMed PMID: 33339084; PubMed Central PMCID: PMC7765516.

28. Mendonça N, Gregório MJ, Salvador C, Henriques AR, Canhão H, Rodrigues AM. Low adherence to the Mediterranean diet is associated with poor socioeconomic status and younger age: a cross-sectional analysis of the EpiDoC cohort. Nutrients. 2022 Jan;14(6):1239. doi:10.3390/nu14061239

29. Ruddock HK, Brunstrom JM, Higgs S. The social facilitation of eating: why does the mere presence of others cause an increase in energy intake? Physiol Behav. 2021 Oct;240:113539. doi:10.1016/j.physbeh.2021.113539

30. Gahagan S. Development of eating behavior: biology and context. J Dev Behav Pediatr. 2012 Apr;33(3):261–71. doi:10.1097/DBP.0b013e31824a7baa

31. Woodside J, Young IS, McKinley MC. Culturally adapting the Mediterranean diet pattern – a way of promoting more 'sustainable' dietary change? Br J Nutr. 128(4):693–703. doi:10.1017/S0007114522001945 PubMed PMID: 35734884; PubMed Central PMCID: PMC9346614.

32. Phull S. The Mediterranean diet: socio-cultural relevance for contemporary health promotion. 1–6.

33. Louey J, He J, Partridge SR, Allman-Farinelli M. Facilitators and barriers to healthful eating among adolescents in high-income countries: a mixed-methods systematic review. Obes Rev. 2024;25(11):e13813. doi:10.1111/obr.13813

34. Higgs S, Thomas J. Social influences on eating. Curr Opin Behav Sci. 2016 Jun 1;Diet, behavior and brain function9:1–6. doi:10.1016/j.cobeha.2015.10.005

35. Pinto AM, Mercedes J. Determinants of healthy eating and physical activity in commuter college students: a qualitative study. J Am Coll Health. 2025 Dec 9;0(0):1–9. doi:10.1080/07448481.2025.2597910 PubMed PMID: 41367263.

36. Alruwaili A, King JA, Deighton K, Kelly BM, Liao Z, Innes A, et al. The association of smoking with different eating and dietary behaviours: a cross-sectional analysis of 80 296 united kingdom adults. Addiction. 2024;119(10):1737–50. doi:10.1111/add.16584

37. Dong HJ, Brain K, Olsson M, Dragioti E, Gerdle B, Ghafouri B. Eating habits and the desire to eat healthier among patients with chronic pain: a registry-based study. Sci Rep. 2024 Feb 27;14:4705. doi:10.1038/s41598-024-55449-z PubMed PMID: 38409442; PubMed Central PMCID: PMC10897138.

38. O'Loughlin I, Newton-John TRO. 'Dis-comfort eating': an investigation into the use of food as a coping strategy for the management of chronic pain. Appetite. 2019 Sep 1;140:288–97. doi:10.1016/j.appet.2019.05.027

39. Bai J, Zhu H, Ran X, Qu H, He L, Zhao Y. The time-stamped effects of screen exposure on food intake in adults: a meta-analysis of experimental studies. Appetite. 2025 Feb 1;206:107846. doi:10.1016/j.appet.2024.107846

40. Associação portuguesa de nutrição [Internet]. [cited 2026 Feb 2]. Available from: https://www.apn.org.pt/v0D0O0F/dieta-mediterranica

41. Costa RL da, Miguel I, Pereira L, Dias A, Gonçalves R. Hábitos alimentares saudáveis para a indústria alimentar em portugal. Rper. 2023 Sep 12;(66):66. doi:10.59072/rper.vi66.80

42. Lopes C, Torres D, Oliveira A, Severo M, Guiomar S, Alarcão V, et al. National food, nutrition and physical activity survey of the portuguese general population. EFSA Supporting Publ. 2017;14(12):1341E. doi:10.2903/sp.efsa.2017.EN-1341

43. Dn R. Diário de Notícias [Internet]. 2017 [cited 2026 Feb 2]. Quase 1/4 dos produtos consumidos estão fora da roda dos alimentos. Available from: https://www.dn.pt/arquivo/diario-de-noticias/quase-14-dos-produtos-consumidos-por-portugueses-estao-fora-da-roda-dos-alimentos-8672248.html

44. Retrato dos hábitos alimentares [Internet]. [cited 2026 Feb 2]. Available from: https://www.fipa.pt/consumidor/retrato-dos-habitos-alimentares?utm_source=chatgpt.com

45. Santos M, Nascimento AC, Santiago S, Gama AC, Calhau MA. _O sal na alimentação dos portugueses. 2016;(4).

46. Jaitha A. An introduction to the theory and applications of bayesian networks. 1–42.

47. McLachlan S, Dube K, Hitman GA, Fenton NE, Kyrimi E. Bayesian networks in healthcare: distribution by medical condition [Internet]. arXiv; 2020 [cited 2026 Feb 2]. Available from: http://arxiv.org/abs/2002.00224 doi:10.48550/arXiv.2002.00224

48. Polotskaya K, Muñoz-Valencia CS, Rabasa A, Quesada-Rico JA, Orozco-Beltrán D, Barber X. Bayesian networks for the diagnosis and prognosis of diseases: a scoping review. Mach Learn Knowl Extr. 2024 Jun;6(2):1243–62. doi:10.3390/make6020058

49. Begashaw GB, Zewotir T, Fenta HM. Dynamic bayesian network modeling for longitudinal data on child undernutrition in Ethiopia (2002–2016). Front Public Health. 2024 Dec 12;12:1–13. doi:10.3389/fpubh.2024.1399094

50. Research Explorer The University of Manchester [Internet]. [cited 2026 Feb 2]. Application of bayesian networks to problems within obesity epidemiology. Available from: https://research.manchester.ac.uk/en/studentTheses/application-of-bayesian-networks-to-problems-within-obesity-epide/

51. Wang W, Chang Y, Chen G. Association between healthy eating index-2020, alternative mediterranean diet scores, and gastrointestinal cancer risk in NHANES 2005-2018. Sci Rep. 2025 Feb 1;15(1):3983. doi:10.1038/s41598-025-88317-5 PubMed Central PMCID: PMC11787371.

52. Zhu W, Marchant R, Morris RW, Baur LA, Simpson SJ, Cripps S. Bayesian network modelling to identify on-ramps to childhood obesity. BMC Med. 2023 Mar 1;21(1):105–17. doi:10.1186/s12916-023-02789-8 PubMed PMID: 36944999; PubMed Central PMCID: PMC10031893.

53. Kong D, Chen R, Chen Y, Zhao L, Huang R, Luo L, et al. Bayesian network analysis of factors influencing type 2 diabetes, coronary heart disease, and their comorbidities. BMC Public Health. 2024 May 8;24(1):1267. doi:10.1186/s12889-024-18737-x PubMed Central PMCID: PMC11080276.

54. Guo Y, Bai G, Hu Y. Using bayes network for prediction of type-2 diabetes. In. 2012. p. 471–2.

55. Mohammadi M, Hosseini M, Tabatabaee H. Using bayesian network for the prediction and diagnosis of diabetes. 1–6.

56. Das S, Baffour B, Richardson A. Prevalence of child undernutrition measures and their spatio-demographic inequalities in Bangladesh: an application of multilevel bayesian modelling. BMC Public Health. 2022 May 18;22(1):1008. doi:10.1186/s12889-022-13170-4

57. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten Simple Rules for Reproducible Computational Research. PLOS Computational Biology. 2013 Oct 24;9(10):e1003285. doi:10.1371/journal.pcbi.1003285

58. Nunez Y, Gibson EA, Tanner EM, Gennings C, Coull BA, Goldsmith J, et al. Reflection on modern methods: good practices for applied statistical learning in epidemiology. Int J Epidemiol. 2021 May 17;50(2):685–93. doi:10.1093/ije/dyaa259 PubMed PMID: 34000733; PubMed Central PMCID: PMC8128480.

59. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009 Sep;21(9):1263–84. doi:10.1109/TKDE.2008.239

60. Luque A, Carrasco A, Martín A, de las Heras A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognit. 2019 Jul 1;91:216–31. doi:10.1016/j.patcog.2019.02.023

61. Lipton ZC, Elkan C, Naryanaswamy B. Optimal thresholding of classifiers to maximize F1 measure. In: Calders T, Esposito F, Hüllermeier E, Meo R, editors. Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg: Springer; 2014. p. 225–39. doi:10.1007/978-3-662-44851-9_15

62. Gregório M, Teixeira A, Páscoa R, Baptista S, Carvalho R, Martins C. The Problem-Solving Decision-Making scale—translation and validation for the Portuguese language: a cross-sectional study. BMJ Open. 2020 Jun 1;10(6):e033625. doi:10.1136/bmjopen-2019-033625 PubMed PMID: 32595147.

63. Chao AM, Wadden TA, Ashare RL, Loughead J, Schmidt HD. Tobacco smoking, eating behaviors, and body weight: a review. Curr Addctn Rep. 2019;6:191–9. doi:10.1007/s40429-019-00253-3 PubMed PMID: 33224710; PubMed Central PMCID: PMC7678013.

64. Wang X, Qin LQ, Arafa A, Eshak ES, Hu Y, Dong JY. Smoking cessation, weight gain, cardiovascular risk, and all-cause mortality: a meta-analysis. Nicotine Tob Res. 2021 Dec 1;23(12):1987–94. doi:10.1093/ntr/ntab076

65. Tian J, Gall S, Otahal P, Smith K, Dwyer T, Venn A. Smoking cessation and weight gain: are changing health behaviours explanatory?: jing tian. Eur J Public HEALTH. 2015 Oct 1;25(suppl_3):ckv174.040. doi:10.1093/eurpub/ckv174.040

66. Spring B, Stump T, Penedo F, Pfammatter AF, Robinson JK. Toward a health-promoting system for cancer survivors: patient and provider multiple behavior change. HEALTH Psychol : Off J Div HEALTH Psychol Am Psychol Assoc. 2019 Sep;38(9):840–50. doi:10.1037/hea0000760 PubMed PMID: 31436465; PubMed Central PMCID: PMC6709684.

67. Loef M, Walach H. The combined effects of healthy lifestyle behaviors on all cause mortality: a systematic review and meta-analysis. Prev Med. 2012 Sep;55(3):163–70. doi:10.1016/j.ypmed.2012.06.017 PubMed PMID: 22735042.

68. Moraes MM, Oliveira B, Afonso C, Santos C, Miranda RC, Rauber F, et al. Mediterranean diet, sociodemographic factors and ultra-processed food consumption in Portugal. Eur J Public HEALTH. 2020 Sep 1;30(Supplement_5):ckaa165.434. doi:10.1093/eurpub/ckaa165.434

69. Swinburn BA, Kraak VI, Allender S, Atkins VJ, Baker PI, Bogard JR, et al. The global syndemic of obesity, undernutrition, and climate change: the lancet commission report. Lancet. 2019 Feb 23;393(10173):791–846. doi:10.1016/S0140-6736(18)32822-8 PubMed PMID: 30700377.

70. Walker RE, Keane CR, Burke JG. Disparities and access to healthy food in the United States: a review of food deserts literature. HEALTH Place. 2010 Sep;16(5):876–84. doi:10.1016/j.healthplace.2010.04.013 PubMed PMID: 20462784.

71. Scutari M, Vitolo C, Tucker A. Learning Bayesian networks from big data with greedy search: computational complexity and efficient implementation. Stat Comput. 2019 Sep;29(5):1095–108. doi:10.1007/s11222-019-09857-1

72. Friedman, N., nachman, I. and peer, D. (1999) learning bayesian network structure from massive datasets the 'sparse candidate' algorithm. Uncertainty in artificial intelligence, 15(3), 206-215. - references - scientific research publishing [Internet]. [cited 2026 May 31]. Available from: https://www.scirp.org/reference/referencespapers?referenceid=34039

73. Schwingshackl L, Hoffmann G, Iqbal K, Schwedhelm C, Boeing H. Food groups and intermediate disease markers: a systematic review and network meta-analysis of randomized trials. Am J Clin Nutr. 2018 Sep 1;108(3):576–86. doi:10.1093/ajcn/nqy151 PubMed PMID: 30535089; PubMed Central PMCID: PMC6134288.

74. Dicken SJ, Batterham RL. The role of diet quality in mediating the association between ultra-processed food intake, obesity and health-related outcomes: a review of prospective cohort studies. Nutrients. 2021 Dec 22;14(1):23. doi:10.3390/nu14010023 PubMed PMID: 35010898; PubMed Central PMCID: PMC8747015.

75. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS One. 2015 Mar 4;10(3):e0118432. doi:10.1371/journal.pone.0118432

76. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22. doi:10.1016/j.jclinepi.2019.02.004

77. McLachlan S, Dube K, Hitman GA, Fenton NE, Kyrimi E. Bayesian networks in healthcare: distribution by medical condition. Artif Intell Med. 2020 Jul 1;107:101912. doi:10.1016/j.artmed.2020.101912

78. Rodrigues SSP, Franchini BM, Pinho ISM de, Graça APSR. The portuguese mediterranean diet wheel: development considerations. Br J Nutr. 2022 Oct;128(7):1315–21. doi:10.1017/S0007114521003743

79. Dal Pozzolo A, Caelen O, Johnson R, Bontempi G. Calibrating probability with undersampling for unbalanced classification. In. 2015. doi:10.1109/SSCI.2015.33

80. Chicco D, Jurman G. The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020 Jan 2;21(1):6–18. doi:10.1186/s12864-019-6413-7

81. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Inf Assoc. 2022 Sep 1;29(9):1525–34. doi:10.1093/jamia/ocac093

82. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32–5. doi:10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

83. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In. Pittsburgh, Pennsylvania: ACM Press; 2006 [cited 2026 Jun 1]. p. 233–40. Available from: http://portal.acm.org/citation.cfm?doid=1143844.1143874 doi:10.1145/1143844.1143874

84. Delgado R, Núñez-González JD. Bayesian network-based over-sampling method (BOSME) with application to indirect cost-sensitive learning. Sci Rep. 2022 May 24;12(1):8724. doi:10.1038/s41598-022-12682-8 PubMed PMID: 35610323; PubMed Central PMCID: PMC9130330.

85. Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers. Machine Learning. 1997 Nov 1;29(2):131–63. doi:10.1023/A:1007465528199

86. Koller D, Friedman N. Probabilistic graphical models: principles and techniques. Nachdr. Cambridge, Mass.: MIT Press; 2010. 1231 p. (Adaptive computation and machine learning).

87. Magalhães V, Severo M, Torres D, Ramos E, Lopes C, By Ian-Af Consortium. Characterizing energy intake misreporting and its effects on intake estimations, in the portuguese adult population. Public Health Nutr. 2020 Apr;23(6):1031–40. doi:10.1017/S1368980019002465 PubMed PMID: 31564257; PubMed Central PMCID: PMC10200469.

88. Carvalho C, Correia D, Lopes C, Torres D. Adherence to the EAT-lancet planetary health diet in Portugal and its associations with socioeconomic and lifestyle factors. Eur J Nutr. 2025 Apr 9;64(4):152–64. doi:10.1007/s00394-025-03661-6 PubMed PMID: 40205117; PubMed Central PMCID: PMC11982082.

89. Luo T, Tseng TS. Diet quality as assessed by the healthy eating index-2020 among different smoking status: an analysis of national health and nutrition examination survey (NHANES) data from 2005 to 2018. BMC Public Health. 2024 May 2;24(1):1212. doi:10.1186/s12889-024-18630-7 PubMed PMID: 38693488; PubMed Central PMCID: PMC11064397.

90. Turkia J, Mehtätalo L, Schwab U, Hautamäki V. Mixed-effect bayesian network reveals personal effects of nutrition. Sci Rep. 2021 Jun 8;11(1):12016. doi:10.1038/s41598-021-91437-3

91. Stephenson BJK, Wu SM, Dominici F. Identifying dietary consumption patterns from survey data: a Bayesian nonparametric latent class model. J R Stat Soc Ser A Stat Soc. 2024 Apr 12;187(2):496–512. doi:10.1093/jrsssa/qnad135

92. Neuhouser ML. The importance of healthy dietary patterns in chronic disease prevention. Nutr Res. 2019 Oct 1;70:3–6. doi:10.1016/j.nutres.2018.06.002

Annexes

Annex A

Faculty of Medicine, University of Porto

Questionnaire number |_____|

Introductory note

Dear Sir/Madam,

We would like to invite you to answer a questionnaire as part of a research project of the Faculty of Medicine of the University of Porto. The main goal is to understand your opinion about lifestyle.

The questionnaire is anonymous, meaning that your identity will never be recorded or revealed. It consists of 3 parts: the first on general health; the second on lifestyle and the third on sociodemographic data. The expected duration of the questionnaire is 20 to 25 minutes.

Your collaboration is voluntary. At any time, if you wish, you may give up answering the questions without any repercussions. Still, we would like to tell you that your collaboration is of utmost importance to the success of this study.

Want some further clarification?

Will you willingly answer this questionnaire?

Part I – General state of health

The following two questions ask your opinion about your health. Pay attention to each question and answer as honestly as possible. **Most important is your opinion.**

1. Overall, how do you consider your health?

 Very good Good ③ Reasonable Bad Very bad

2. Indicate if during the last 12 months you have had any of the following health problems (multi response)**:**

1	High blood pressure (hypertension)
2	Stroke
3	High blood cholesterol
4	Diabetes
5	Heart problems (infarction, angina pectoris, heart valve disease, heart failure)
6	Back, bone, joint or muscle pain
7	Chronic bronchitis, asthma, chronic obstructive pulmonary disease or emphysema
8	Gastritis or stomach ulcer/duodenum
9	Anxiety
10	Depression
11	Overweight
12	Obesity
13	Cancer
14	None of these

Part II – Lifestyle

In this part, we want to know what you think about lifestyle. When answering, do not worry about what is correct or not. Just answer according to what your personal opinion dictates.

Please indicate your agreement with each of the following sentences. Rate on a scale of 1 to 5, corresponding **1 to "Strongly disagree" and 5 to "Strongly agree"**.

		Strongly disagree	Partially disagree	Indifferent	Partially agree	Strongly agree
Diet						
1.	A healthy diet can prevent and help control some diseases.	1	2	3	4	5
2.	My diet is healthy.	1	2	3	4	5
3.	It is important that my family doctor asks/advises me about my diet.	1	2	3	4	5
4.	Usually, my family doctor asks/advises me about a healthy diet.	1	2	3	4	5
Physical activity						
1.	Regular physical activity can prevent and help control some diseases.	1	2	3	4	5
2.	My physical activity is regular.	1	2	3	4	5
3.	It is important that my family doctor asks/advises me about my physical activity.	1	2	3	4	5
4.	Usually, my family doctor asks/advises me about physical activity.	1	2	3	4	5
Alcohol intake						
1.	Excessive alcohol intake can cause and aggravate some diseases.	1	2	3	4	5
2.	My alcohol intake is excessive.	1	2	3	4	5
3.	It is important that my family doctor asks/advises me about my alcohol intake.	1	2	3	4	5
4.	Usually, my family doctor asks/advises me about alcohol intake.	1	2	3	4	5
Tobacco use						
1.	Smoking can cause and aggravate some diseases.	1	2	3	4	5
2.	I am an active smoker.	1	2	3	4	5
3.	It is important that my family doctor asks/advises me about smoking.	1	2	3	4	5
4.	Usually, my family doctor asks/advises me about smoking.	1	2	3	4	5
Illicit drugs						
1.	Illicit drugs can cause and aggravate some diseases.	1	2	3	4	5
2.	I do not use illicit drugs.	1	2	3	4	5
3.	It is important that my family doctor asks/advises me about illicit drugs.	1	2	3	4	5
4.	Usually, my family doctor asks/advises me about illicit drugs.	1	2	3	4	5
Sleep habits						
1.	Good quality sleep can prevent and help control some diseases.	1	2	3	4	5
2.	I have a good quality of sleep.	1	2	3	4	5
3.	It is important that my family doctor asks/advises me about sleep habits.	1	2	3	4	5
4.	Usually, my family doctor asks/advises me about sleep habits.	1	2	3	4	5
Screen activities (watching television; using the computer, mobile phone, game consoles, …)						
1.	Excessive screen activities can cause and aggravate some diseases.	1	2	3	4	5
2.	I have excessive screen activities.	1	2	3	4	5
3.	It is important that my family doctor asks/advises me about screen activities.	1	2	3	4	5
4.	Usually, my family doctor asks/advises me about screen activities.	1	2	3	4	5
Stress						
1.	A high level of stress can cause and aggravate some diseases.	1	2	3	4	5
2.	I have a high level of stress.	1	2	3	4	5
3.	It is important that my family doctor asks/advises me about my level of stress.	1	2	3	4	5
4.	Usually, my family doctor asks/advises me about managing stress.	1	2	3	4	5
Sedentarism (amount of time without physical activity)						
1.	Sedentarism can cause and aggravate some diseases.	1	2	3	4	5
2.	I am sedentary.	1	2	3	4	5
3.	It is important that my family doctor asks/advises me about sedentarism.	1	2	3	4	5
4.	Usually, my family doctor asks/advises me about sedentarism.	1	2	3	4	5

1. Please indicate:

1.1. **Weight** (without clothes and without shoes): _____(Kg)

1.2. **Height** (without shoes):_____(cm)

2. How many meals do you usually eat per day? Consider breakfast, lunch, and dinner as main meals and snacks and supper as intermediate meals.

Number of main meals: _____

Number os intermediate meals: _____

2.1. In the main meals you had yesterday, you consumed (multi response)**:**

1	Milk, yogurt or cheese
2	Vegetable soup
3	Bread
4	Beef
5	Fish
6	Potatoes, rice or pasta
7	Beans or chickpeas
8	Cakes, cookies, chocolates or desserts
9	Soft drinks, with or without gas
10	Natural juices, made from fresh fruit
11	Vegetables (salads, cooked vegetables)
12	Fast food (meal prepared and served quickly following a standardized and massified method, whether or not it can be eaten at the place of purchase)
13	Pre-cooked / industrially prepared meal (meal prepared according to an industrialized method that includes cooking partially or completely the same, keeping it according to refrigeration, freezing, vacuuming or canning processes, and before being consumed requires the completion of cooking heat treatment such as oven, microwave or short-term frying.)
14	None of these

2.2. How often do you eat fruit, including juices made from fresh fruit but excluding juices made from concentrates? (Do not consider canned or dried fruits.)

□ Once or more a day
□ 4 to 6 times a week
③ 1 to 3 times a week
□ 1 to 3 times a week
□ Never □ **Go to question 2.3.**

2.2.1. Usually, how many servings of fruit do you consume per day?

Number of servings of fruit:_____

2.3. How often, do you eat vegetables or salads, excluding potatoes and juices made from concentrates? (Also consider the juices made from fresh vegetables, canned vegetables, legumes (beans, lentils), soups (heat or cold) and vegetarian dishes.)

□ Once or more a day
□ 4 to 6 times a week
③ 1 to 3 times a week
□ 1 to 3 times a week
□ Never □ **Go to question 3.**

2.3.1. Usually, how many servings of vegetables or salads do you consume per day?

Number of servings of vegetables anda salads:_____

1. Think about the time you spend on activities such as working, caring for the house, caring for family members, studying, volunteering, and so on. Do not consider leisure or sports activities.
From the following options select the one that best describes the main way you perform the activities mentioned. (If you do several activities, consider the one that you spend the most time with.)

- Sitting or standing in activities involving slight physical exertion
- On the move or on tasks that require moderate physical exertion
- ③ In heavy or physically demanding jobs
- Do not perform any of the indicated activities

1.1. In a normal week during your travels, how many days are you walking for at least 10 minutes straight? Also, consider the weekend. (If you do not walk or do so for less than 10 minutes straight, record 0.)

Number of days:_____ **(If you answered 0, go to question 3.2.)**

1.1.1. And in a normal, in your travels, how long are you walking? (If your travel time varies greatly throughout the week, please state the average time per day.

- Less than 10 minutes
- 10 to 29 minutes
- ③ 30 to 59 minutes
- 1 hour to less than 2 hours
- 2 hours to less than 3 hours
- 3 hours or more

1.2. Now think of sports or leisure activities. In a normal week, how many days does physical activity last for less than 10 minutes in a row? Consider the weekend. (Examples: walking, playing ball, jogging, biking, or swimming. If you do not exercise or do less than 10 minutes in a row, record 0.)

Number of days:_____ **(If you answered 0, go to question 4.)**

1.2.1. On this (these) day(s), on average, how long do you practice physical activity?

- 10 to 29 minutes
- 30 to 59 minutes
- ③ 1 hour to less than 2 hours
- 2 horas to less than 3 hours
- 3 hours or more

1.3. In a normal week, how many days do you practice muscle-building activities, such as resistance training or bodybuilding? (If you do not engage in muscle-building activities, record 0. Consider all muscle-building activities, even if they have been considered before.)

Number of days:____

2. In the last 12 months, how often have you consumed alcoholic beverages of any kind (beer, wine, spirits, cocktails, liquors, mixtures of alcoholic beverages, among others)?

- Everyday or almost every day (Choose this option if you drank in the 365 days of the last 12 months or usually more than 6 days a week.)
- 5 to 6 days a week
- ③ 3 to 4 days a week
- 1 to 2 days a week
- 2 to 3 days a month **(Go to question 4.2.)**
- Once a month **(Go to question 4.2.)**
- Less than once a month **(Go to question 4.2.)**
- You have not consumed in the last 12 months because you have stopped drinking (**Choose this option if you have not consumed alcohol in the last 12 months, but have had at least one drink throughout your life without just proving it**.) **(Go to question 5.)**
- Never consumed, or only occasionally to taste (**Choose this option if you have never consumed a drink in your life, or, at most, sipped occasionally for the purpose of tasting the drink**.) **(Go to question 5.)**

2.1. Usually, between Monday and Thursday, how many days do you drink alcohol?

- 4 days
- 3 days

③ 2 days 1 day
 On neither day **(Go to question 4.2.)**

1.1.1. And on those days, on average, how many alcoholic drinks do you consume?

 16 or more drinks per day 10 to 15 drinks per day
③ 6 to 9 drinks per day 4 to 5 drinks per day
 3 drinks per day 2 drinks per day
 1 drink per day

1.2. And usually, between Friday and Sunday, how many days do you drink alcohol?

 3 days 2 days
③ 1 day On neither day **(Go to question 4.3.)**

1.2.1. And on one of those days, on average, how many alcoholic drinks do you consume?

 16 or more drinks per day 10 to 15 drinks per day
③ 6 to 9 drinks per day 4 to 5 drinks per day
 3 drinks per day 2 drinks per day
 1 drink per day

1.3. In the last 12 months, have you consumed 6 or more alcoholic beverages on a single occasion or event? For example, at a party, at a meal, going out with friends or alone at home. (Consider any type of beverage: beer, wine, spirits, cocktails, liqueurs, alcohol mixes, home-made alcoholic beverages, etc.)

 Every day or almost every day 5 to 6 days per week
③ 3 to 4 days per week 1 to 2 days per week
 2 to 3 days per month Once a month
 Less than once a month Not in the last 12 months
 Never in life

2. Do you smoke? (Consider any kind of tobacco, except the electronic cigarette. Answer regardless of regular consumption)

 Yes
 No **5.0.1. Have you ever smoked? (Consider any kind of tobacco, except the electronic cigarette. Answer regardless of regular consumption)**

 No **(Go to question 5.4.)**
 Yes **5.0.2. How long have you quit smoking?**
 Less than 1 year **(Go to question 5.4.)**
 More than 1 year **(Go to question 5.4.)**

2.1. You smoke:
 Daily **5.1.1. At what age did you start to smoke daily?** Age:_____
 Occasionally

2.2. On average how many cigarettes do you smoke per day? (Consider any kind of tobacco, except the electronic cigarette. Answer regardless of regular consumption)

Number of cigarettes: ___

2.3. Do you use an electronic cigarette? (Electronic mechanical device that produces inhalable vapor, with or without nicotine, also known as e-cigarette.)

 Yes No

1.1. How often are you indoors where other people smoke? (Consider only the smoke produced by other people. Examples of enclosed spaces: home, work, public spaces, restaurants, car, etc.)

 Daily
 Occasionally
③ Never **(Go to question 6.)**

1.2. And usually, how long are you exposed?
 Less than 1 hour per day
 1 hour or more

1.3. And this exposure to indoor tobacco smoke happens mostly where?

 At home In the car
③ At work In leisure places
 Other locations

2. Do you use illicit drugs? (Consider any type of drug: cannabis, cocaine (crack, ..), heroin, amphetamines (ecstasy, ..), hallucinogens, among others,… Answer regardless of regular use.)

 Yes
 No **6.0.1. Have you ever consumed? (Answer regardless of regular use.)**
 No **(Go to question 7.)**
 Yes **6.0.2. How long have you stopped consuming?**
 Less than 1 year **(Go to question 7.)**
 1 year or more **(Go to question 7.)**

2.1. You consume:
 Daily **6.1.1. At what age did you start to smoke daily?** Age: _____
 Occasionally

2.2. On average how many consumptions do you make per week? (Consider any type of drug. Answer regardless of regular use.)

Number of consumptions: _____

3. In the last 2 weeks, how often have you had sleep problems (difficulty falling asleep, little sleep or too much sleep)?

 Every day
 Most part of the days
③ Some days
 A few days
 Never

3.1. In the last 2 weeks, how often have you needed to take sleeping medication?

 Every day
 Most part of the days
③ Some days
 A few days
 Never

3.2. In the last 2 weeks, how often have you got restful sleep?
 Every day
 Most part of the days
③ Some days
 A few days
 Never

1. In a normal week, how many days do you perform screen activities (watching TV; using a computer, mobile phone, game consoles, etc.)? Also, consider the weekend.

Number of days: ____ **(If you answered 0, go to question 9.)**

1.1. And on a normal day, how much time do you use for screen activities? (If you do more than one activity, consider the one you do most often and tick only one answer. If your usage time varies a lot over the week, indicate the average time per day.)

- Less than 10 minutes
- 10 to 29 minutes
- ③ 30 to 59 minutes
- 1 hour to less than 2 hours
- 2 hours to less than 3 hours
- 3 hours or more

1.2. On which device do you usually do that screen activities? (If you use more than one type of device, consider what you use most often and tick only one answer.)

- Television
- Computer
- ③ Mobile phone
- Game consoles

1.3. In the last 2 weeks, how often did you do less than you wanted in your work or daily activities because you were doing screen activities?

- Every day
- Most part of the days
- ③ Some days
- A few days
- Never

2. In the last 2 weeks, how often have you done less than you wanted in your work or daily activities because you felt anxious/nervous?

- Every day
- Most part of the days
- ③ Some days
- A few days
- Never

2.1. In the last 2 weeks, how often have you felt calm and peaceful?

- Every day
- Most part of the days
- ③ Some days
- A few days
- Never

2.2. In the last 2 weeks, how often did you feel anxious/nervous?

- Every day
- Most part of the days
- ③ Some days
- A few days
- Never

2.3. In the last 2 weeks, have you had to take medication to control anxiety/nervousness?

 Every day
 Most part of the days
③ Some days
 A few days
 Never

Part III – Socio-demographic data

This part of the questionnaire aims to collect some data related to you, your education and your professional activity.

1. **What is your nationality?**
   Portuguese
   Foreign. Which one:_____
   ③ Stateless (without nationality)
   Does not know

2. **How old are you?** _____years

3. **Your gender:**
   Female Male

4. **What is your marital status?**
   Single
   Married
   ③ Married but legally separated
   Divorced
   Widowed
   Does not know

1. Of the following categories, which best describes your main occupation?

- Works on it owns
- Works for others
- ③ Student
- Doing military service
- Homemaker
- Retired
- Unemployed
- Does not know

2. What is your main profession?

Indicate:

_____

- Does not have a profession
- Does not Know

3. Which of the following is a health care recipient?

- ADSE (State Health Service Assistance)
- SSMJ (Justice Ministry Service)
- ③ IASFA (Institute of Social Action of the Armed Forces)
- SAD/PSP (PSP Disease Assistance Services)
- SAD/GNR (GNR Disease Assistance Services)
- SAMS (Banking Operations Department)
- Private Health Insurance
- SNS (National Health System)
- Others (indicate): _____

% Does not Know

4. What is your municipality of residence?

Indicate:

_____

Thank you so much for your cooperation!

Appendices

Appendix A - Distribution of Variables

Frequency (n) and relative frequency (%) of each variable. The full variable set corresponds to the final analytical sample (n = 847); the reduced variable set corresponds to the exploratory dataset prior to the exclusion of foreign nationals (n = 891).

Full variable set (n = 847)

Variable	Category	n	%
egs_cat3	Bad or very bad	78	9.2
	Fair	311	36.7
	Good or very good	458	54.1
hipertensao_12m_bin	No	676	20.2
	Yes	171	79.8
avc_12m_bin	No	831	98.1
	Yes	16	1.9
hipercol_12m_bin	No	707	83.5
	Yes	140	16.5
diabetes_12m_bin	No	750	88.5
	Yes	97	11.5
doenca_cardiaca_12m_bin	No	777	91.7
	Yes	70	8.3
dores_osteo_musc_12m_bin	No	513	60.6
	Yes	334	39.4
asma_dpoc_12m_bin	No	809	95.5
	Yes	38	4.5
gastrite_ulcera_12m_bin	No	801	94.6
	Yes	46	5.4
ansiedade_12m_bin	No	706	83.4
	Yes	141	16.6
depressao_12m_bin	No	793	93.6
	Yes	54	6.4

cancro_12m_bin	No	835	98.6
	Yes	12	1.4
Imc_cat6	Underweight	10	1.2
	Normal	389	45.9
	Overweight	339	40.0
	Obesity class 1	85	10.0
	Obesity class 2	19	2.2
	Obesity class 3	5	0.6
idade_cat14	20–24 years	54	6.4
	25–29 years	51	6.0
	30–34 years	56	6.6
	35–39 years	71	8.4
	40–44 years	83	9.8
	45–49 years	76	9.0
	50–54 years	74	8.7
	55–59 years	75	8.9
	60–64 years	71	8.4
	65–69 years	66	7.8
	70–74 years	53	6.3
	75–79 years	47	5.5
	80–84 years	38	4.5
	85 years and over	32	3.8
sexo_cat	Female	447	52.8
	Male	400	47.2
setor_atividade_cat4	Not working	319	37.7
	Primary sector	3	0.4
	Secondary sector	76	9.0
	Tertiary sector	449	53.0
nuts2_cat5	Norte	323	38.1
	Centro	197	23.3

	AM Lisboa	222	26.2
	Alentejo	66	7.8
	Algarve	39	4.6
alimentacao_saudavel_bin	No	726	85.7
	Yes	121	14.3
drogas_ilicitas_cat3	Never consumed	796	94.0
	Quit more than 1 year ago	39	4.6
	Consumes or quit less than 1 year ago	12	1.4
estado_civil_cat4	Single	170	20.1
	Married	479	56.6
	Divorced/Separated	93	11.0
	Widowed	105	12.4
escolaridade_cat4	None	24	2.8
	1st–3rd cycle	445	52.5
	Secondary education	274	32.3
	University education	104	12.3
tabagismo_cat3	Never smoked	479	56.6
	Ex-smoker	149	17.6
	Current smoker	219	25.9
nivel_af_cardio_cat4	Sedentário / Inativo (<10 min/sem)	156	18.4
	Low activity (10–149 min/week)	302	35.7
	Moderately active (150–299 min/week)	119	14.0
	Highly active (≥300 min/week)	270	31.9
af_oms_bin	Não cumpre (cardio<150)	458	54.1
	Cumpre (cardio>=150)	389	45.9
ecras_risco_cat3	Baixo (<1h/dia)	172	20.3
	Moderate (1–2h/day)	219	25.9
	Elevado (>2h/dia)	456	53.8
ecras_interf_2s_cat5	Minimal	713	84.2

	Little	94	11.1
	Some	28	3.3
	A lot	5	0.6
	Maximal	7	0.8
med_ansiedade_2s_bin	No	741	87.5
	Yes	106	12.5
fumador_passivo_bin	No	602	71.1
	Yes	245	28.9
ansiedade_2s_cat3	Low	191	22.6
	Moderate	321	37.9
	High	335	39.6
sono_2s_cat3	Good quality	497	58.7
	Moderate quality	226	26.7
	Poor quality	124	14.6
sono_med	Never	695	82.1
	A few days	42	5.0
	Some days	29	3.4
	Most days	12	1.4
	Every day	69	8.1
sono_med_2s_bin	No	695	82.1
	Yes	152	17.9

Reduced variable set (n = 891)

Variable	Category	n	%
EGS_rec	Bad or very bad	78	8.8
	Fair	322	36.1
	Good or very good	491	55.1
Dores	No	547	61.4
	Yes	344	38.6

IMC_cat	Underweight	13	1.5
	Normal	401	45.0
	Overweight	358	40.2
	Obesity class 1	93	10.4
	Obesity class 2	21	2.4
	Obesity class 3	5	0.6
Nacionalidade	Portuguese	850	95.4
	Foreign	41	4.6
Sexo	Female	478	53.6
	Male	413	46.4
EstadoCivil	Married	501	56.2
	Others	390	43.8
alimentacao_saudavel_bin	No	767	86.1
	Yes	124	13.9
RegularPhysicalActivity	No	635	71.3
	Yes	256	28.7
doencas_foro_mental	No	720	80.8
	Yes	171	19.2
escolaridade_cat4	1st cycle or none	224	25.1
	2nd and 3rd cycle	258	29.0
	Secondary	294	33.0
	University	115	12.9
Fumadores	No	665	74.6
	Yes	226	25.4
work	No	327	36.7
	Yes	564	63.3
SemAtividadeEcra	No	318	35.7
	Yes	573	64.3
doencas_cardio	No	615	69.0
	Yes	276	31.0

doencas_metabolicas1	No	678	76.1
	Yes	213	23.9