FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Classificação de Episódios Clínicos

Ricardo Manuel da Rocha Melo e Castro



Mestrado Integrado em Engenharia Informática e Computação

Orientador: Célia Talma Martins de Pinho Valente Oliveira Gonçalves

Orientador: Carlos Manuel Milheiro de Oliveira Pinto Soares

31 de Outubro de 2019

Classificação de Episódios Clínicos

Ricardo Manuel da Rocha Melo e Castro

Mestrado Integrado em Engenharia Informática e Computação

31 de Outubro de 2019

Resumo

Cada visita de um paciente a uma unidade de saúde é registada como um episódio clínico ao qual são associadas as diferentes observações do médico. Posteriormente e com base nessas notas clínicas, profissionais hospitalares atribuem códigos de diagnóstico ao episódio. Este processo de classificação é feito de forma manual consumindo demasiado tempo e estando sujeito a erros. Por um lado as notas clínicas são longas e contém muita informação não estruturada e irrelevante. Por outro, dentro dos milhares de códigos disponíveis, as diferenças entre alguns são mínimas tornando-os ainda mais difíceis de identificar para pessoas sem formação médica. O presente trabalho pretende auxiliar o pessoal hospitalar nessa decisão contribuindo para uma classificação mais rápida, uniforme e com uma taxa de erro menor.

Utilizando dados de 4199 episódios registados entre 2011 e 2018, criámos um modelo para a classificação de episódios clínicos. Abordámos o problema segundo uma estratégia Multirótulo pois um episódio pode ter diferentes diagnósticos. Na exploração do melhor modelo, procurámos igualmente aprofundar o conhecimento de diferentes dimensões associadas aos episódios clínicos. Elaborámos assim diferentes casos de forma a minimizar o impacto de limitações de dados e da própria natureza dos códigos e notas médicas.

Os resultados obtidos são positivos, havendo uma melhoria relativamente ao caso de referência. Os melhores resultados foram obtidos utilizando *Classifier Chains* com regressão logística, *Label Powersets* e *1 vs Rest* com regressão linear. Para a medida Correspondência Exacta tiveram valores médios de 0.557, 0.51 e 0.49. Relativamente à medida *Hamming Loss* estes 3 algoritmos tiveram valores de 0.204, 0.22 e 0.184, respetivamente.

A dimensão e balanceamento dos dados não nos permite tirar conclusões significativas quanto à distribuição temporal e tipologia dos diagnósticos. No entanto, os melhores resultados foram observados usando apenas consultas nos dados.

Palavras Chave: Aprendizagem Automática, Classificação de texto, Classificação Multirótulo, Episódios Clínicos, Diagnósticos Clínicos

Abstract

Each patient visit to a medical facility is registered as a clinical episode and all medical observations are associated to it. Based on those clinical notes, hospital staff later assigns diferent diagnostic codes to that episode. This manual classification process is time consuming and error prone. On one hand, clinical notes are long and contain too much unstructured and irrelevant information. On the other hand, amongst the thousands of codes available, the differences between some are minimal, making them even harder to identify, for staff with no medical training. The present study aims at helping hospital staff in that decision, contributing to a faster and more uniform classification, with less errors.

Using 4199 episodes, registered between 2011 and 2018, we created a model for the classification of clinical episodes. The underlying problem was approached using a Multi-Label strategy since an episode can have multiple diagnostics. In the search for the best model an effort was made to deepen our knowledge of the different dimensions of clinical episodes. This way, different cases were developed to minimize limitations related to data collection and the nature of codes and notes.

The results were positive with improvements over the baseline. The best ones were achieved with Classifier Chains with logistic regression, label powersets and linear regression wrapped in a 1 vs Rest classifier. For the Exact Match measure, these algorithms achieved 0.56, 0.51 and 0.49 scores, respectively. Considering the Hamming Loss measure, the same algorithms scored 0.204, 0.22 and 0.184. Due to dataset size and imbalance, no significative conclusions were made regarding the temporal distribution and typology of episodes. Still, the best output was registered using only appointments in the data.

keywords: Machine Learning, Document Classification, Multi-Label Classification, Clinical Episodes, Clinical Diagnostic

Agradecimentos

Gostaria de agradecer aos professores Carlos Soares e Célia Gonçalves pelo conhecimento, amizade e dedicação. Pela disponibilidade e esforço constantes, por garantirem que eu tinha o melhor acompanhamento e me orientarem nas dificuldades.

Às pessoas da Glintt que nos últimos 3 anos se tornaram uma segunda família. Em especial ao Pedro Rocha, João Gomes e Francisco Correia. Pelo contributo para esta dissertação e por me guiarem nestes anos.

À Carolina, que de diferentes formas me trouxe até aqui. Contigo sou melhor.

Aos meus pais, pelo amor e apoio incondicionais. Por garantirem que eu tivesse oportunidades e capacidade de as aproveitar. Ao meu irmão, pela consultoria gratuita durante estes anos, por estar perto sempre que preciso e me inspirar a ser melhor.

A todos os colegas que conheci no meu percurso académico, em especial ao Alcino e ao António. Por me incluírem e mostrarem que, quando se partilha os mesmos valores, a amizade não tem idade.

A todos os que me são próximos pela amizade e aprendizagem. Parte do que sou vem de vocês. Um abraço especial ao Pedro Silva e ao André Mota. Ao Pedro por me ajudar a focar no que é realmente importante. Ao André, que me acompanhou na faculdade e agora na Glintt, por ser o melhor exemplo do que é ser amigo.

Ao corpo docente do MIEIC. Pela vossa dedicação e foco constante nos alunos.

Um infinito obrigado,

Ricardo Melo e Castro

"It ain't what you don't know that gets you into trouble.
It's what you know for sure that just ain't so."

Unknown

Conteúdo

1 Introdução **1**
- 1.1 Motivação . . . 2
- 1.2 Objetivo . . . 2
- 1.3 Estrutura da Dissertação . . . 2

2 Revisão Bibliográfica **3**
- 2.1 Classificação de texto . . . 3
- 2.2 Classificação Multirótulo . . . 4
  - 2.2.1 Algoritmos Multirótulo . . . 4
  - 2.2.2 Métodos que Transformam o Problema . . . 4
  - 2.2.3 Métodos que Adaptam o Problema . . . 5
  - 2.2.4 Métricas de avaliação . . . 5
- 2.3 Trabalho Relacionado . . . 6

3 Caso de estudo **9**
- 3.1 Classificação Internacional de Doenças . . . 9
- 3.2 Natureza dos Dados Recolhidos . . . 10
  - 3.2.1 Episódios . . . 10
  - 3.2.2 Diagnósticos . . . 11
  - 3.2.3 Notas Clínicas . . . 11
- 3.3 Análise Exploratória dos Dados . . . 12
  - 3.3.1 Validação dos dados fornecidos . . . 12
  - 3.3.2 Descrição dos dados . . . 13
- 3.4 Conclusões . . . 15

4 Experiências e Resultados **17**
- 4.1 Configuração experimental . . . 17
  - 4.1.1 Dados utilizados . . . 17
  - 4.1.2 Método de estimação de desempenho . . . 20
  - 4.1.3 Medidas de avaliação . . . 20
  - 4.1.4 Preparação de dados . . . 21
  - 4.1.5 Software . . . 21
  - 4.1.6 Algoritmos . . . 22
- 4.2 Experiências . . . 22
  - 4.2.1 Conjunto de dados com rótulos selecionados . . . 22
  - 4.2.2 Rótulo adicional . . . 24
- 4.3 Resumo . . . 25

5 Conclusões e Trabalho Futuro **27**
5.1 Conclusões 27
5.2 Trabalho Futuro 28

Referências **29**

A Análise Exploratória de dados **33**
A.1 Distribuições de ocorrências códigos e notas por episódio 33

B Resultados **35**
B.1 Conjunto de dados com 4 rótulos 35
B.1.1 NaiveBayes 1vsAll 35
B.1.2 Multirótulo 1vsAll 36
B.1.3 Diferentes Algoritmos Multirótulo 38
B.2 Rótulo D 39

Lista de Figuras

3.1 Relação entre episódios, notas e diagnósticos 10
3.2 Distribuição dos episódios por ano 14
3.3 Número de diagnósticos por episódio 15
3.4 Distribuição de Consultas e Internamentos 16

4.1 Seleção de dados de treino e teste para o ano 2012 20
4.2 Resultados de Precisão para o algoritmo Naive Bayes em ambos os tipos de episódio 22
4.3 Resultados de Sensibilidade para o algoritmo Naive Bayes em consultas 22
4.4 Resultados de Precisão para Naive Bayes em internamentos 23
4.5 Resultados de Medida F1 para Naive Bayes em internamentos 23
4.6 Resultados da Média da Medida F1 para 1vsAll Linear SVN 23
4.7 Resultados da Média da Medida F1 para 1vsAll Logistic Regression 23
4.8 Resultados Hamming Loss para ambos os tipos de episódio 24
4.9 Resultados Correspondência Exacta para ambos os tipos de episódio 24
4.10 Resultados dos dois conjuntos de dados para Hamming Loss em ambos os tipos de episódio 24
4.11 Resultados dos dois conjuntos de dados para Correspondência Exacta em ambos os casos 24
4.12 Resultados dos dois conjuntos de dados para Hamming Loss em consultas 25
4.13 Resultados dos dois conjuntos de dados para Hamming Loss em internamentos . 25

A.1 Número de ocorrências de códigos 33
A.2 Número de notas por episódio 33

B.1 Resultados de Precisão para Naive Base em ambos os tipos de episódio 35
B.2 Resultados de Sensibilidade para Naive Base em ambos os tipos de episódio . . . 35
B.3 Resultados de Medida F1 para Naive Base em ambos os tipos de episódio 35
B.4 Resultados de Precisão para Naive Base em consultas 35
B.5 Resultados de Sensibilidade para Naive Base em consultas 36
B.6 Resultados de Medida F1 para Naive Base em consultas 36
B.7 Resultados de Precisão para Naive Base em internamentos 36
B.8 Resultados de Sensibilidade para Naive Base em internamentos 36
B.9 Resultados de Medida F1 para Naive Base em internamentos 36
B.10 Resultados da Média da Precisão para 1vsAll Naive Bayes 36
B.11 Resultados da Média da Sensibilidade para 1vsAll Naive Bayes 36
B.12 Resultados da Média da Medida F1 para 1vsAll Naive Bayes 37
B.13 Resultados da Média da Precisão para 1vsAll Linear SVN 37
B.14 Resultados da Média da Sensibilidade para 1vsAll Linear SVN 37
B.15 Resultados da Média da Medida F1 para 1vsAll Linear SVN 37

B.16 Resultados da Média da Precisão para 1vsAll Logistic Regression 37
B.17 Resultados da Média da Sensibilidade para 1vsAll Logistic Regression 37
B.18 Resultados da Média da Medida F1 para 1vsAll Logistic Regression 37
B.19 Resultados Hamming Loss para ambos os tipos de episódio 38
B.20 Resultados Hamming Loss para consultas 38
B.21 Resultados Hamming Loss parainternamentos 38
B.22 Resultados Correspondência Exacta para ambos os tipos de episódio 38
B.23 Resultados Correspondência Exacta para consultas 38
B.24 Resultados Correspondência Exacta para internamentos 38
B.25 Resultados dos dois conjuntos de dados para Hamming Loss em ambos os tipos de episódio 39
B.26 Resultados dos dois conjuntos de dados para Hamming Loss em consultas 39
B.27 Resultados dos dois conjuntos de dados para Hamming Loss em internamentos . 39
B.28 Resultados dos dois conjuntos de dados para Correspondência Exacta em ambos os tipos de episódios 39
B.29 Resultados dos dois conjuntos de dados para Correspondência Exacta em consultas 39
B.30 Resultados dos dois conjuntos de dados para Correspondência Exacta em internamentos 39

Lista de Tabelas

2.1 Matriz de pesquisa 7

3.1 Tabela de episódios 10
3.2 Tabela de diagnósticos 11
3.3 Tabela de Notas Clínicas 12
3.4 Diferença entre dados fornecidos e validados 13
3.5 Análise de Consultas e Internamentos 15

4.1 Tabela de conjuntos de dados selecionados 19

Abreviaturas e Símbolos

ICD9	Classificação Internacional de Doenças 9ª Revisão Modificação Clínica
MLkNN	Multi-Label k-Nearest Neighbor
LSVN	Linear Support Vector Classification
CC	Classification Chains
LR	Logistic Regression
GNB	Gaussian Naive Bayes
LB	Label Poweret
BR	Binary Relevance
NLTK	Natural Language Toolkit
SVM	Support Vector Machine
tf-idf	Term Frequency–Inverse Document Frequency
CNN	Convolutional Neural Networks
LSTM	Long Short-Term Memory
NN	Neural Network
RNN	Recurrent Neural Network
FFNN	Feed Forward Neural Network

Capítulo 1

Introdução

Cada vez que uma pessoa se dirige a uma unidade hospitalar, na qualidade de utente, é atendida por médicos, uma ou mais vezes. Toda a informação relevante destas interacções é registada em notas clínicas, quer sejam diagnósticos, procedimentos médicos efectuados bem como informação sobre o doente que auxilie na prestação de um melhor serviço médico. Esta visita é registada como um episódio clínico e a ele são associadas as diferentes doenças diagnosticadas. Esta atribuição é feita posteriormente por profissionais administrativos tendo como referência as notas médicas registadas.

A tarefa de classificar os diagnósticos é feita de forma manual consumindo bastante tempo e estando sujeita a erros. O responsável pela classificação tem de ler todas as notas e atribuir um ou mais códigos de diagnósticos apropriados. O texto das notas, além de ser tendencialmente longo, não é estruturado, contendo muita informação que, apesar de ser relevante para a prática médica, não é importante para a classificação em si. Por vezes, a pessoa responsável pela classificação não tem habilitações médicas para interpretar correctamente as notas. Este facto tem um peso ainda maior se considerarmos o vasto número de códigos disponíveis, cujas diferenças se podem prender em pormenores clínicos [LYNB08]. Dada a complexidade da informação disponível e a quantidade de códigos, não é incomum profissionais diferentes atribuírem classificações dispares ao mesmo episódio [PPN+13].

Todos estes factores tornam a tarefa de classificação de episódios clínicos ineficiente, demorando demasiado tempo, estando sujeita a erros e não garantindo uniformidade no processo de tomada de decisão. A nível administrativo, afecta não só a gestão hospitalar, que se baseia em grande parte nos diagnósticos e serviços prestados, assim como o respectivo financiamento. Do ponto de visto clínico, um incorrecto histórico do paciente poderá afectar futuramente o diagnóstico de um paciente.

1.1 Motivação

Este projeto foi realizado em colaboração com a Glintt - Global Intelligent Technologies. é uma empresa tecnológica com foco na área da saúde, nomeadamente nas áreas farmacêutica e hospitalar, oferecendo soluções para toda a cadeia de produção. Desta experiência acumulada é comum identificarem problemas aos quais os sistemas de informação existentes não conseguem responder. A classificação de episódios enquadra-se nesse grupo de problemas estando alinhado com a crescente necessidade de integração de sistemas de aprendizagem automática, nos sistemas atualmente implementados.

A exploração de aprendizagem computacional na área da saúde é antiga, devido à disponibilidade de dados. Contudo, em anos recentes, o estudo desta área tem ganho maior tracção, especificamente nas temáticas do processamento de informação não estruturada do paciente e da classificação de diagnósticos. Este interesse surge sob a promessa de um melhor apoio à decisão do médico e melhores cuidados prestados, assim como numa mais eficiente e menos custosa gestão hospitalar.

Enquanto que a maioria dos artigos aborda de forma transversal a classificação de diagnósticos [LYNB08][Nig16], alguns autores focam-se especificamente em certas especialidades ou doenças [SLD17] [PBM+07]. A definição do problema como classificação Multirótulo e classificação de documentos é comum, nos artigos identificados.

1.2 Objetivo

Este trabalho tem como objetivo o desenvolvimento de um modelo capaz de classificar com sucesso episódios clínicos com sucesso e aprofundar o conhecimento sobre as diferentes características inerentes à natureza dos episódios clínicos. Seguindo uma abordagem de classificação multirótulo espera-se poder contribuir para uma classificação que demore pouco tempo, com uma baixa percentagem de erros enquanto garante uniformidade nos seus resultados.

1.3 Estrutura da Dissertação

Para além do primeiro capítulo introdutório, esta dissertação contém mais 4 capítulos.

O capítulo 2 apresenta uma revisão de literatura, abordando trabalho relacionado com classificação de documentos, mais concretamente classificação de notas médicas. Nesse sentido abordou-se ainda a utilização de aprendizagem automática para diagnóstico médico, com incidência em algoritmos e métricas de avaliação para classificação Multirótulo. Seguidamente, no capítulo 3, é descrito o caso de estudo, o seu contexto, a natureza dos dados recolhidos e a análise exploratória dos mesmos. No capítulo 4 são descritas as experiências efetuadas e suas abordagens. São ainda apresentados os resultados obtidos. No último capítulo 5, são apresentadas as conclusões finais do trabalho desenvolvido assim como possível trabalho futuro.

Capítulo 2

Revisão Bibliográfica

A classificação de episódios clínicos é um tema amplamente estudado pela comunidade académica. Enquanto alguns se focam em doenças especificas [VD15][CXZ+15], outros abordam a temática de uma forma global [Nig16]. No segundo caso, onde se inclui o presente trabalho, o objecto de estudo pode ser encarado como um problema de classificação Multirótulo [PBM+07][BNKC+18].

Devido à sua natureza, a área de conhecimento da classificação de episódios clínicos encontra-se na intersecção de duas áreas muito estudadas em aprendizagem automática. Por um lado o tratamento de notas pode ser abordado como um problema de classificação de texto. por outro, dada particularidade de existirem vários diagnósticos por episódio, temos igualmente que estudar a classificação Multirótulo.

Por uma questão de consistência e uniformidade, a tradução para português de termos e conceitos de aprendizagem automática é feita a partir do livro *Extração de Conhecimento de Dados* [JG17].

2.1 Classificação de texto

A classificação automática de texto consiste na atribuição de um documento de texto a um conjunto de classes pré definidas, usando aprendizagem automática. É normalmente efectuada a partir de um conjunto de palavras significativas ou características extraídas do texto [DZ11]. A abordagem genérica a um problema de classificação de texto consiste em pré-processamento e classificação [TÇG+11].

Num texto a analisar nem toda a informação é relevante. Dada a quantidade de palavras que é preciso analisar, é assim necessário realizar diferentes operações que nos permitam melhorar a qualidade do texto. A remoção de palavras sem significado (*stop words*) é tradicionalmente usada. Estas palavras são normalmente as mais comuns numa língua como artigos, preposições ou pronomes, etc. A eliminação destas palavras permite diminuir o espaço de termos a analisar. Relativamente à representação do texto um método bastante utilizado é o *Term Frequency-Inverse*

Document Frequency (tf-idf), uma estatística numérica que classifica uma palavra (termo) quanto à sua importância num documento [VIN15].

Para classificação de documentos são identificados diferentes métodos como Naive Bayes, Naive Bayes Multinomial, Máquinas de vectores de suporte (*Support Vector Machines - SVMs*), K-vizinhos mais próximos[TÇG+11]. Contudo, dada a maior complexidade da classificação Multirótulo, continuamos o estudo deste tópico na secção seguinte.

2.2 Classificação Multirótulo

Nesta secção estudamos algoritmos e métricas de avaliação normalmente usadas num problema Multirótulo.

2.2.1 Algoritmos Multirótulo

Dependendo da sua abordagem, os algoritmos utilizados na classificação multirótulo podem ser categorizados como métodos que transformam o problema (*Problem Transformation Methods* e métodos que adaptam o algoritmo (*Algorithm Adaptation Methods)* [ZZ13],

2.2.2 Métodos que Transformam o Problema

Este tipo de métodos transforma o problema multirótulo num conjunto de problemas de classificação binária, nos quais podem posteriormente ser usados classificadores rótulo único. Alguns algoritmos representativos são *Classifier Chains* (CC) e *Binary Relevance* (BR).

- **Binary Relevance**

  Para q rótulos diferentes, este algoritmo decompõe o problema de Multirótulo num problema de $|q|$ classificações binárias independentes. Assim, um conjunto de dados D é decomposto em q conjuntos binários $D_{yj}, j = 1..q$ que são usados para construir q classificadores binários independentes. Para cada classificador binário, os exemplo associados com o respetivo rótulo são considerados como positivos e todos os restantes como negativos. Para classificar uma nova instância, este método associa a agregação dos rótulos previstos positivamente pelos q classificadores binários. O problema com BR é que os diferentes rótulos são tratados de forma independente. [SCML13]

- **Classifier Chains**

  Este método é em alguns aspectos equivalente a BR com a vantagem de considerar relações entre rótulos. O CC cria uma cadeia de classificadores binários ordenados, em que o resultado de um classificador é passado ao seguinte melhorando o seu conjunto de treino. [ZZ13]

- **Label Powerset**

  Esta abordagem transforma o problema de Multirótulo num problema Multiclasses. Para cada combinação possível de rótulos, este método cria um rótulo novo. Desta forma consegue captar a relação entre rótulos. Um dos problemas com este método é que quantos mais rótulos, maiores são as combinações possíveis aumentando muito o custo computacional. Outro problema é que, alguns rótulos gerados por combinações podem ter valores associados a poucas instâncias tornando o conjunto de dados desbalanceado [SCML13].

- **1vsAll** Esta abordagem é equivalente a BR tendo como base a classificação Multiclasses. Para cada q rótulos cria um novo conjunto de dados D_i, q onde os restantes rótulos são classificados como negativos. A cada novo conjunto de dados é aplicado um classificador binário [RK04].

2.2.3 Métodos que Adaptam o Problema

Esta abordagem passa por adaptar um algoritmo para efetuar classificação multirótulo. Em vez de adaptar o problema, este tipo de método tenta resolver o problema em toda a sua complexidade. Um exemplo é a adaptação multirótulo do algoritmo k-vizinhos mais próximos (*Multi-Label k-Nearest Neighbor - MLkNN*).

- **Multi Label k-Nearest Neighbors**

  Abordagem Multirótulo baseada em kNN. Para cada instância os k vizinhos mais próximos são identificados. Posteriormente, de acordo com o conjunto de rótulos destes vizinhos, o principio *maximum a posteriori (MAP)* é aplicado para determinar o conjunto de rótulos para essa instância [ZZ07].

2.2.4 Métricas de avaliação

Em problemas de classificação, os resultados dos modelos são avaliados segundo métricas convencionais como precisão, sensibilidade, Medida-F, etc. No contexto Multirótulo, esta avaliação é mais complicada uma vez que vários rótulos podem estar numa instância do conjunto de dados [ZZ13]. Sendo $D = \{(x_i, Y_i); 1 \leq i \leq n\}$ um conjunto de teste com n instâncias, Y_i rótulos verdadeiros e Z_i rótulos previstos para cada instância x_i.

- **Precisão**

  Percentagem de rótulos previstos correctamente na classe positiva sobre o total de rótulos previstos na classe positiva.

$$Precisão = \frac{1}{n}\sum_{i=1}^{n}\frac{|Z_i \bigcap Y_i|}{|Z_i|}$$

- **Sensibilidade**

  Taxa de acerto da classe positiva. Percentagem de rótulos previstos correctamente na classe positiva sobre o total de rótulos da classe positiva.

$$Sensibilidade = \frac{1}{n}\sum_{i=1}^{n}\frac{|Z_i \bigcap Y_i|}{|Y_i|}$$

- **Medida-F1** Média Harmónica da precisão e sensibilidade com peso igual.

$$Medida - F1 = \frac{1}{n} < \sum_{i=1}^{n}\frac{2|Z_i \bigcap Y_i|}{|Z_i| + |Y_i|}$$

- **Hamming loss**

  Representa a fração de rótulos previstos incorrectamente. Esta métrica considera os erros de previsão e omissão normalizados sobre o número total de classes e exemplos.

$$Hamming\,Loss = \frac{1}{n} * \sum_{i=1}^{n}\frac{1}{|L|} * |Z_i \triangle Y_i|$$

- **Correspondência Exacta**

  Percentagem de instâncias previstas correctamente. Para cada instância prevista é feita uma comparação rótulo a rótulo com a instância equivalente no conjunto de teste. É uma métrica rigorosa não distinguido entre previsões parcialmente corretas e totalmente incorrectas.

$$Correspondência\,Exata = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = Z_i)$$

2.3 Trabalho Relacionado

Durante o levantamento do estado de arte, toda a informação relevante recolhida nos diferentes artigos foi guardada na matriz da tabela 2.1. A matriz encontra-se organizada em diferentes dimensões - *Data, Volume de dados, Número de Rótulos Analisados, Algoritmos utilizados, Estimação de Desempenho, Métricas de Avaliação e Resultado*.

Referência	Data	vol. Dados	Rótulos	Algoritmos	Est. Desempenho	Métricas	Resultado
[Kar18]	2018	5000	17 ICD9	Convolutional Neural Networks (CNN)	70% Treino 15% Validation 15% Teste	Medida F1	F1=69
[Kar18]	2018	5000	17 ICD9	Long Short-Term Memory (LSTM)	70% Treino 15% Validation 15% Teste	Medida F1	F1=64.6
[Kar18]	2018	5000	17 ICD9	LSTM com attention	70% Treino 15% Validation 15% Teste	Medida F1	F1=67
[Kar18]	2018	5000	17 ICD9	CNN com Attention	70% Treino 15% Validation 15% Teste	Medida F1	F1=72.8
[PPN+13]	2013	22 815	5030 ICD9	SVM	90% Treino 10% Teste	Precisão Sensibilidade Medida F1	Precisão=86,7 Sensibilidade=16.4 Medida F1=27.6
[PPN+13]	2013	22 815	5030 ICD9	Hierarquia com SVM	90% Treino 10% Teste	Precisão Sensibilidade Medida F1	Precisão=57.7 Sensibilidade=30.0 Medida F1=39.5
[Nig16]	2016	31865	10 ICD9	Binary Relevance com Logistic Regression	50% Treino 25% Desenvolvimento 25% Teste	Precisão Sensibilidade Medida F1	Precisão=0.4381 Sensibilidade=0.2678 Medida F1=0.300
[Nig16]	2016	31865	10 ICD9	Feed Forward Neural Network (NN)	50% Treino 25% Desenvolvimento 25% Teste	Precisão Sensibilidade Medida F1	Precisão=0.6671 Sensibilidade=0.3062 Medida F1=0.3937
[Nig16]	2016	31865	10 ICD9	Recurrent NN	50% Treino 25% Desenvolvimento 25% Teste	Precisão Sensibilidade Medida F1	Precisão=0.5406 Sensibilidade=0.3610 Medida F1=0.4035
[Nig16]	2016	31865	10 ICD9	RNN com LSTM	50% Treino 25% Desenvolvimento 25% Teste	Precisão Sensibilidade Medida F1	Precisão=0.7488 Sensibilidade=0.3199 Medida F1=0.4168
[Nig16]	2016	31865	100 ICD9	Binary Relevance com Logistic Regression	50% Treino 25% Desenvolvimento 25% Teste	Precisão Sensibilidade Medida F1	Precisão=0.3322 Sensibilidade=0.1598 Medida F1=0.1815
[Nig16]	2016	31865	100 ICD9	FFNN	50% Treino 25% Desenvolvimento 25% Teste	Precisão Sensibilidade Medida F1	Precisão=0.3528 Sensibilidade=0.2022 Medida F1=0.234
[Nig16]	2016	31865	100 ICD9	RNN	50% Treino 25% Desenvolvimento 25% Teste	Precisão Sensibilidade Medida F1	Precisão=0.3410 Sensibilidade=0.2776 Medida F1=0.2439
[Nig16]	2016	31865	100 ICD9	RNN com LSTM	50% Treino 25% Desenvolvimento 25% Teste	Precisão Sensibilidade Medida F1	Precisão=0.1882 Sensibilidade=0.1956 Medida F1=0.1691

Tabela 2.1: Matriz de pesquisa

- **Volume de dados**

  O conjunto de dados maior tinha 31865 registos [Nig16] enquanto o mais pequeno consistia 1954 relatórios escritos de radiologia[Zha08].

- **Rótulos**

  Praticamente todos os artigos recolhidos usavam códigos da Classificação Internacional de Doenças 9ª Revisão Modificação Clínica (ICD9) [dS05]. Esta classificação atribui códigos a doenças ou lesões. O maior número de códigos distintos usados foi 5030 [Kar18] e o menor 10 [Nig16]. Este último autor também utilizou um conjunto de dados com 100 códigos. Outros autores usaram outro tipo de códigos definidos por si.

- **Algoritmos**

  Quatro autores usaram SVMs [PPN+13] [KTA16] [LYNB08] [Zha08]. Outro autor comparou diferentes versões de NN com um caso base de BR [KTA16].

- **Estimação de desempenho**

  Todos os autores usaram *Holdout*. O tamanho dos intervalos dos conjuntos de treino e teste variam entre autaores. Um autor utilizou 90% de treino e 10% de teste. Outro autor usou 70% de treino, 15% de teste e 15%[Kar18] dos dados para validação. Existe ainda um caso onde foram utilizado 50% para treino, 25% [Nig16] para desenvolvimento e 25% para teste.

- **Métricas**

  As medidas de avaliação mais comuns foram Precisão, Sensibilidade e Medida F1. Um autor recorreu ainda a *Micro Averaged F1* [Zha08] e outro *Ranking Loss* [Nig16].

- **Resultados**

  No geral os resultados foram positivos. Dois artigos obtiveram resultados acima de 80% para Precisão e Sensibilidade abaixo de 20%. Os restantes tiveram resultados piores com valores de F1 entre 40% e 80% e Precisão e Sensibilidade com valores piores que os referidos.

Capítulo 3

Caso de estudo

Este trabalho foi proposto pela Glintt (Global Technologies Solutions) e tem como objectivo a criação de um modelo de classificação de episódios clínicos. O modelo a criar deve auxiliar os profissionais hospitalares a associar diagnósticos a episódios clínicos, a partir das notas médicas geradas durante o acompanhamento do paciente. Este sistema permitirá aumentar a eficiência e uniformidade no processo de classificação, assim como diminuir o erro associado.

Na prática hospitalar corrente, os episódios só são classificados após o seu término, no fim de uma consulta ou, no caso de internamentos, após o paciente ter alta. Este ponto reflete-se na consideração acerca da data de um episódio, sendo apenas considerada a sua data de fim. Igualmente, consideramos que a análise e evolução de um modelo é feita no final de cada ano. Não é praticável gerar modelos mensalmente devido ao baixo volume de dados. Por outro lado, um modelo gerado a cada 5 ou 10 anos não conseguiria incluir a atualidade da atividade hospitalar. Consideramos assim o espaço temporal de um ano como o indicado para atualizar modelo, considerando os episódios que nele decorreram. Este período é igualmente ideal de forma a respeitar períodos administrativos. Assim, serão apenas considerados episódios pertencentes a anos para os quais existem episódios na sua totalidade.

3.1 Classificação Internacional de Doenças

Os diagnósticos presentes nos dados recolhidos são representados por códigos ICD9 [dS05]. Esta classificação resulta da adaptação da Classificação Internacional de Doenças da Organização Mundial de Saúde e atribui códigos a diferentes doenças ou lesões.

Existem aproximadamente 13 000 códigos diferentes. Cada código tem entre 3 e 5 caracteres sendo o primeiro um digito ou as letras V ou E e os restantes algarismos. Esta classificação apresenta uma estrutura hierarquia com dois ou mais níveis. O código 780.01 "Coma"pertence ao grupo "Sintomas, sinais e afecções mal definidas (780-799)"e consecutivos subgrupos, "Sintomas (780 - 789)","Sintomas Gerais (780)"e "Alteração do estado de consciência (780.0)".

Atributo	Descrição
DOENTE	*Identificador do paciente encriptado*
EPISODIO	*Identificador do episódio encriptado*
T_EPISODIO	*Tipo de episódio*
DT_INI	*Data de início do episódio*
DT_FIM	*Data de fim do episódio*

Tabela 3.1: Tabela de episódios

3.2 Natureza dos Dados Recolhidos

No âmbito do presente caso de estudo foram fornecidos dados relativos a episódios, diagnósticos e notas clínicas, registados entre Janeiro de 2011 e Abril de 2019. Os episódios dividem-se entre consultas e internamentos. A informação relativa aos episódios clínicos e paciente associado foi anonimizada por uma questão de confidencialidade.

Um episódio representa uma ida de um paciente a uma unidade médica podendo a sua duração ser de zero dias, no caso de uma consulta, como de vários dias, no caso de um internamento.

Um diagnóstico representa a identificação de um diagnóstico clínico associado ao paciente no âmbito de um episódio clínico. O mesmo episódio pode ter vários diagnósticos associados.

As notas clínicas contêm informação relativa às interações de um médico com um paciente, incluindo observações registadas pelo médico. Um episódio pode ter várias notas clínicas associadas.

A figura 3.1 representa a relação entre episódios, notas e diagnósticos.

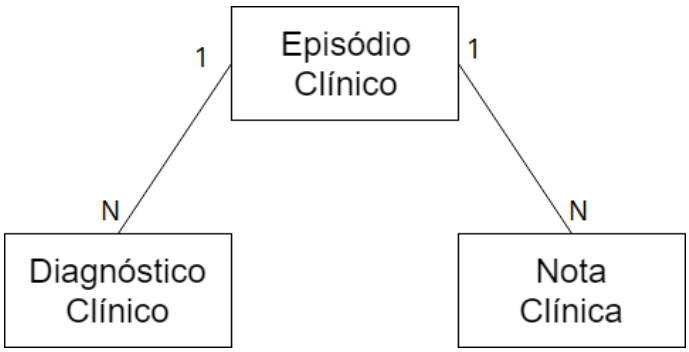


Figura 3.1: Relação entre episódios, notas e diagnósticos

3.2.1 Episódios

A tabela 3.1 descreve os atributos associados a um episódio clínico. Os atributos *DOENTE* e *EPISODIO* representam, respetivamente, os identificadores de um paciente e do episódio. *T_EPISODIO* diz respeito ao tipo de episódio podendo o seu valor ser *Consultas* ou *Internamentos*. As datas de início e fim de um episódio estão representadas nos campos *DT_INI* e *DT_FIM*.

Atributo	Descrição
EPISODIO	*Identificador do episódio encriptado*
INICIO_DIAG	*Data de início do diagnóstico*
FIM_DIAG	*Data de fim do diagnóstico*
CODIFICACAO	*Identificador da codificação do diagnóstico*
CODIGO	*Código do diagnóstico*
DESCR_DIAG	*Descrição do diagnóstico*

Tabela 3.2: Tabela de diagnósticos

3.2.2 Diagnósticos

Os atributos dos dados relativos aos diagnósticos encontram-se descritos na tabela 3.2. O atributo *EPISODIO* representa o identificador do episódio ao qual o diagnóstico se encontra associado.

O diagnóstico tem um código, *CODIGO*, segundo uma codificação *CODIFICACAO*. A codificação representa a codificação utilizada na classificação de um episódio. Nos nossos dados, a codificação é sempre ICD9. Os códigos que classificam o diagnóstico, são sempre códigos pertencentes à codificação utilizada. Os diagnósticos representam igualmente, através do atributo *DESCR_DIAG*, uma descrição do seu código. Por exemplo, para a codificação *ICD9*, e o código *5990*, temos a descrição *INFECCAO DO TRACTO URINARIO, LOCAL NAO ESPECIFICADO*.

As datas do diagnóstico, *INICIO_DIAG* e *FIM_DIAG* representam, respetivamente, as datas de início e fim do diagnóstico. Uma vez que um diagnóstico é uma classificação atribuída a um episódio, não possui duração própria, não sendo registada, na maioria das vezes a sua data de fim.

3.2.3 Notas Clínicas

Cada nota clínica é representada pela data do seu registo, *DT_REGISTO*, tipo de nota, *TIP_NOTA*, e conteúdo da nota, *OBSERVACAO*. Os atributos relativos às notas clínicas podem ser consultados na tabela 3.3. O conteúdo da nota representa as observações feitas pelo médico durante a sua interação com um utente. Este conteúdo é apresentado de forma não estruturada podendo conter informação tão diversa como: enquadramento sócio-económico, sintomas, medicação regular, medicação prescrita, perfil do utente, historial clínico do utente e/ou familiares,etc. Verifica-se de forma frequente, a utilização de siglas, diminutivos e presença de erros ortográficos. O exemplo seguinte descreve a caracterização apresentada:

> *PROCESSO NÃO DISPONIVEL*
>
> *pedido de colaboração em doente com DPOC, cardiopatia hipertrofica, DM e HTA, a fazer VNI, doente inernda por penumonia e insuf cardiaca.*
>
> *Analsiticaemtne hipnatremia, hiperkaliemia, PBNP- 7446, GSA com Fi O2 ?- pO2- 95, PCO2- 50 pH- normal. doetne com boa adaptação ao VNI, consciente colaborante oreintada, sat O2 com 4L/m- 96%. AP- sons pulmoanres prestnes roncus*

Atributo	Descrição
EPISODIO	*Identificador do episódio encriptado*
DT_REGISTO	*Data do registo da nota*
TIP_NOTA	*Tipo de nota*
OBSERVACAO	*Conteúdo da nota*

Tabela 3.3: Tabela de Notas Clínicas

dispersos, sem broncospasmo, discretas crepitações nas bases. Dada as varias co-morbilidades da doente sugere-se continuação de cuidados pela Medicina Interna. doente ja orientada para conslta de peumologia.

Nas notas está ainda presente o atributo *DOENTE* já descrito na secção anterior.

3.3 Análise Exploratória dos Dados

Os dados fornecidos contêm 66 898 diagnósticos e 131 120 notas clínicas. Nos diagnósticos estão presentes 3929 códigos ICD9 distintos. Estes dados estão associados a 23 176 episódios clínicos que se dividem em 13 403 consultas (42.17%) e 13 403 internamentos (57.83%).

3.3.1 Validação dos dados fornecidos

Os dados fornecidos inicialmente foram validados de forma a garantir que serviam o caso de estudo no seu âmbito. Esta validação teve como objetivo garantir que os episódios utilizados eram relativos a anos completos e que tinham notas e diagnósticos associados. Assegurou-se igualmente que os códigos dos diagnósticos eram ICD9. Foram realizadas as seguintes operações:

- **Eliminação de episódios de anos incompletos** — Para os episódios de 2019 só existe informação até Abril. 131 episódios e 263 diagnósticos associados foram eliminados.
- **Eliminação de episódios sem notas associadas** — De forma a garantir que todos os episódios têm notas clínicas associadas, foram removidos 1 038 episódios nestas condições. Foram igualmente descartados os 3 111 diagnósticos associados a este episódios.
- **Validação dos códigos ICD9** — Foram validados todos os diagnósticos para garantir que os seus códigos seguiam a formatação utilizada pela codificação ICD9. 10 códigos que não respeitavam essa formatação - *HIPERTEN*, *CABE2*, *ICT*, *ASC*, *OSTEOART*, *EPILEPSIA*, *BOC*, *ANEM*, *METAST* e *POLIPOS* - foram removidos.
- **Eliminação de episódios clínicos sem diagnósticos** — Por fim, verificou-se a existência de episódios que, em resultado do passo anterior, tivessem perdido todos os seus diagnósticos. Não foram encontrados episódios nessas condições.

Este processo de validação removeu um total de 1 169 episódios, 3 399 diagnósticos e 358 notas clínicas. Associados aos diagnósticos eliminados estão os seus códigos ICD9. O número total de códigos distintos após validação é de 68. O resultado do processo de validação encontra-se descrito na tabela 3.4 sendo a sua caracterização abordada na secção seguinte.

#	**Dados Originais**	**Dados Validados**	**Diferença (%)**
Nº Episódios	*23176*	*22007*	*1169 (-5.04%)*
Nº Consultas	*9773*	*9492*	*281 (-2.88%)*
Nº Internamentos	*13403*	*12515*	*888 (-6.63%)*
Nº Diagnósticos	*66898*	*63499*	*3399 (-5.08%)*
Nº Códigos Distintos	*3929*	*3861*	*68 (-1.73%)*
Nº Notas	*131120*	*130762*	*358 (-0.27%)*

Tabela 3.4: Diferença entre dados fornecidos e validados

3.3.2 Descrição dos dados

É importante termos um conhecimento aprofundado sobre o conjunto de dados fornecido, assim como dos diferentes elementos que o constituem. Conseguir quantificá-los sobre diferentes dimensões permite-nos antecipar erros futuros ou até mesmo a invalidação de casos de estudo.

- **Episódios por ano**

  A distribuição dos episódios por ano encontra-se representada na figura 3.2. Verifica-se que 2013 é o ano com mais episódios disponíveis com 5026 episódios enquanto que 2017 tem apenas 884 episódios (valor mínimo). Em média, cada ano tem cerca de 2750 episódios.

- **Duração dos episódios**

  A duração mínima de um episódio é de 0 dias, no caso das consultas, que se iniciam e terminam no mesmo dia. O episódio clínico mais longo foi um internamento de 356 dias. Em média, os episódios têm uma duração de 5,28 dias.

- **Diagnósticos por episódio** Na figura 3.3 é possível observar a distribuição do número de diagnósticos por episódio. Existem 9 913 episódios com apenas 1 diagnóstico (valor mínimo), enquanto que existem 2 episódios com o 24 diagnósticos associados (valor máximo). Em média, um episódio clínico tem 2,885 diagnósticos associados.

- **Ocorrências de códigos**

  Avaliando a ocorrência de códigos ICD9 no total dos diagnósticos verificamos que existem 1 393 códigos que ocorrem apenas 1 vez. Existe um código com um valor máximo de ocorrências de 2 551. Em média, cada código ocorre 16,45 vezes.

- **Notas por episódio**

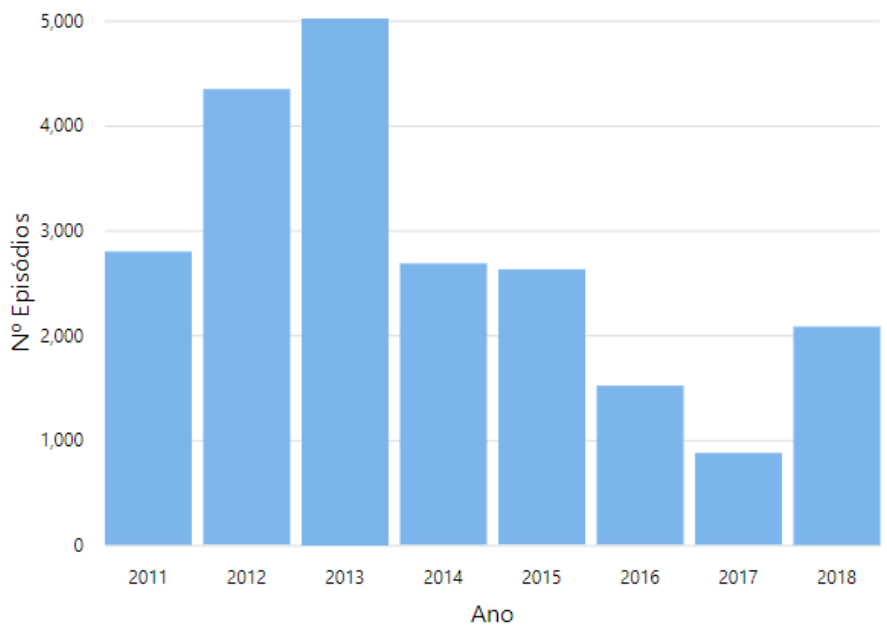


Figura 3.2: Distribuição dos episódios por ano

Relativamente às notas clínicas, 10 796 episódios têm apenas uma nota clínica, sendo a grande maioria consultas, enquanto que existe um episódio com 523 notas. Cada episódio tem em média 5.942 notas associadas.

- **Comprimento das notas**

  O comprimento máximo de uma nota clínica, em caracteres, foi de 607723 enquanto que o valor mínimo foi de 3 caracteres. Em média, uma nota clínica possui 5084.31 caracteres.

- **Consultas e Internamentos**

  Ao longo da análise exploratória de dados é possível observar diferenças entre consultas e internamentos. É assim importante explorar de forma mais detalhada estes dois tipos de episódios.

  A figura 3.4 representa a distribuição de episódios por ano fazendo a distinção entre consultas e internamentos. Nos quatro primeiros anos, entre 2011 e 2014, verificamos uma relação uniforme entre os dois tipos de episódios, com um maior número de internamentos disponíveis. Nos restantes anos esta proporção inverte-se. Apesar do número de consultas permanecer relativamente estável, o número de internamentos é significativamente menor.

  Relativamente aos atributos explorados previamente e aplicando ao contexto de consultas e internamentos, podem verificar-se diferenças. A duração média de um internamento é de 9,29 dias enquanto que, como verificado anteriormente, uma consulta decorre no espaço de minutos o que equivale a 0 dias. Esta diferença justifica que o número médio de notas

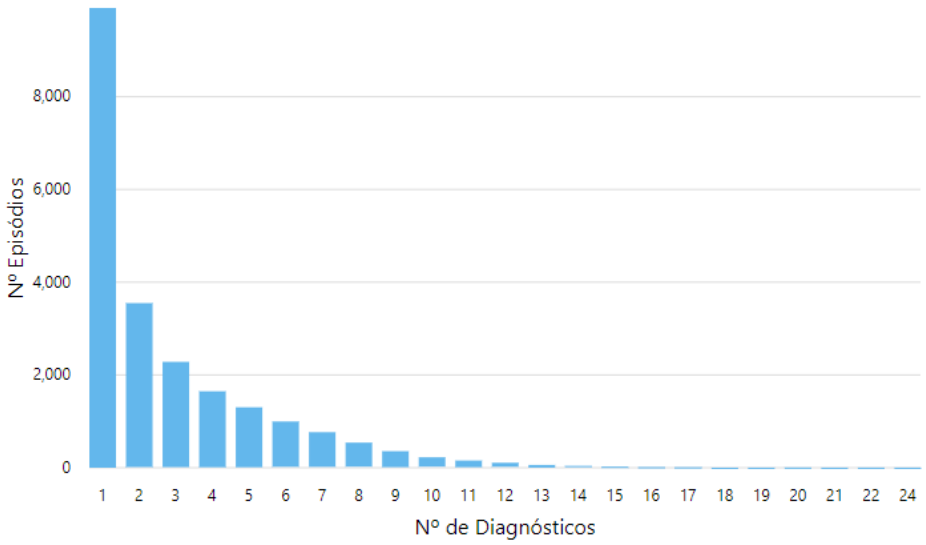


Figura 3.3: Número de diagnósticos por episódio

associadas a um internamento seja de 9,65 enquanto que o das consultas é de 1,05 notas. Igualmente, o tamanho médio das notas (texto a analisar), é proporcionalmente superior nos internamentos.

Existem cerca de 4,10 diagnósticos associados a um internamento contra os 1,29 diagnósticos associados a uma consulta. Relativamente aos códigos distintos presentes em todos os diagnósticos, encontramos 3327 códigos nas consultas e 1302 nos internamentos.

A tabela 3.5 descreve a análise entre os dois tipos de episódio.

3.4 Conclusões

Durante o processo de validação descrito na secção 3.3.1 foram eliminados episódios, diagnósticos e notas. A maior perda prendeu-se com os internamentos, sendo o valor final 6.63%

	Consultas	**Internamentos**
Nº Episódios	*9492*	*12515*
Nº Classificações	*12222*	*51277*
Nº Médio de Classificações por Episódio	*1,29*	*4,10*
Nº Códigos Distintos	*1302*	*3327*
Nº Notas Clínicas	*9978*	*120784*
Nº Médio de Notas por Episódio	*1,05*	*9,65*
Comprimento médio de notas por episódio	*659,36*	*8440,41*
Duração média de episódio	*0*	*9.29*

Tabela 3.5: Análise de Consultas e Internamentos

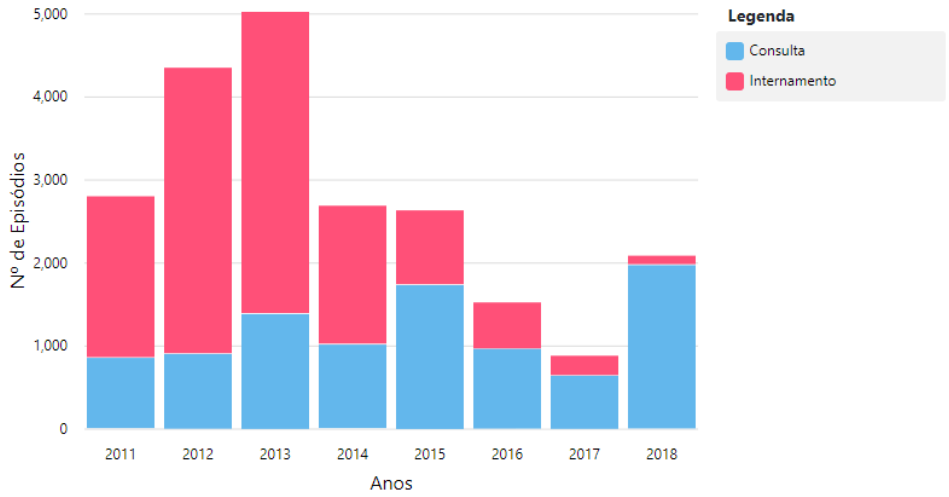


Figura 3.4: Distribuição de Consultas e Internamentos

inferior ao originalmente fornecido. Apesar do menor volume de dados e de haver igualmente menos 2,88% de consultas, a percentagem destas no total de episódios aumentou cerca de 1%. De uma forma geral podemos considerar que não há uma diferença significativa entre os dados recolhidos e os que efectivamente podemos considerar para experimentação.

A análise exploratória permitiu-nos compreender melhor os dados fornecidos e recolher um conjunto válido ao qual fosse possível aplicar o nosso caso de estudo. O comprimento médio das notas é bastante elevado. Adicionando a esse facto a diversidade de informação (muita dela irrelevante para o nosso caso) presente nos casos observados, não nos oferece garantias sobre a qualidade do texto a analisar. Relativamente aos diferentes códigos, estes ocorrem em média 16,45 vezes. Dada a amplitude e tamanho dessa amostra, muitos códigos não terão ocorrências suficientes para serem usados como treino e teste, garantindo qualidade do modelo. A distribuição das ocorrências agrupadas dos códigos e o número de notas por episódio seguem uma distribuição exponencial decrescente, equivalente à distribuição do número de diagnósticos por episódios representada na figura 3.3. As distribuições referidas podem ser consultadas no anexo A.1.

Durante o processo de validação do conjunto de dados a utilizar, foram removidos diagnósticos de alguns episódios. Apesar dessa perda, esses episódios mantiveram notas que contém informação relativa aos diagnósticos perdidos. Esta inconsistência é aceite em função dos dados fornecidos, mesmo sabendo que poderá introduzir algum ruído e/ou erro nos modelos preditivos.

Relativamente à análise de Consultas e Internamentos, existem diferenças a assinalar. O número médio de classificações por episódio, nos internamentos, é de 4,10 contra os 1,29 das consultas. O número médio de notas é igualmente superior nos internamentos, o que se reflecte no comprimento total do texto a analisar sendo este valor de cerca de 8440 caracteres para os internamentos e 659 para as consultas. Assim, dadas as diferenças verificadas, deverá ser positivo analisar o comportamento do mesmo modelo para cada um destes tipos de episódio.

Capítulo 4

Experiências e Resultados

Conforme referido na introdução do capitulo 3 é importante considerar a distribuição temporal dos episódios. Primeiramente reconhecendo que com os anos, o ambiente hospitalar sofre alterações podendo afetar o conteúdo das notas assim como o tipo de diagnósticos efetuados. Posteriormente, aproximando a modelação das condições reais de utilização do modelo. Na exploração de diferentes abordagens ao problema devem também ser considerados os dois tipos de episódios - consultas e internamentos.

4.1 Configuração experimental

Foi elaborada uma configuração experimental que permitisse não só classificar episódios clínicos mas igualmente elaborar diferentes experiências. A configuração experimental é idêntica para as diferentes experiências. Primeiramente as notas são pré-processadas removendo palavras irrelevantes e aplicando o tf-idf2.1.O conjunto de dados é dividido em dados de treino e teste de forma a treinar o modelo e avaliar a previsão obtida, respectivamente. Na avaliação do modelo são utilizadas diferentes métricas. Como resultado, além da avaliação dos algoritmos utilizados, deverá ser possível retirar conclusões sobre a própria condição dos episódios (dimensão temporal e tipologia).

4.1.1 Dados utilizados

Considerando a distribuição temporal do episódios, consideraram-se todos os códigos que permitissem treinar e testar o modelo para os diferentes anos para os quais existem episódios.

4.1.1.1 Conjunto de dados Base

Partindo do ano inicial de 2011 e usando o passado para prever o futuro, foi gerado um conjunto de dados que permitisse classificar episódios de 2012, 2013, 2014, 2015, 2016, 2017 e 2018.

Este processo consistiu em:

- **Selecionar códigos válidos para teste e treino**

  De forma a garantir que os dados selecionados eram válidos para treino e teste, organizaram-se os códigos por ano e mediu-se o número de ocorrências para cada um. A partir dessa distribuição, foram selecionados todos aqueles que tivessem pelo menos duas ocorrências no primeiro ano de treino e duas ocorrências em todos os anos de teste. Garantindo duas ocorrências no primeiro ano de treino, garantimos que esse código ocorre no treino de todos os anos, uma vez que no treino de cada novo ano utilizamos os dados dos anos anteriores. Para o teste, temos que garantir que um código tem duas ocorrências em cada ano de teste uma vez que os dados de teste não se repetem em anos posteriores.

- **Eliminação de códigos e episódios não aceites**

  Os restantes códigos foram eliminados dos episódios. Os episódios que perderam todos os seu códigos foram igualmente eliminados.

Este processo tem o mesmo efeito que o processo efetuado na secção 3.3.2. Contudo, o conjunto de dados final permite avaliar igualmente o comportamento do modelo para os diferentes tipos de episódios, sendo dividido em dois subconjuntos - Consultas e Internamento. Foram assim gerados os 3 conjuntos de dados descritos na tabela 4.1.

	Consultas	**Internamentos**	**Ambos**
Nº Episódios	868	3331	4199
Nº Códigos	1575	19085	20660
Nº Códigos Distintos	4	4	4
Nº Notas	904	26085	26989

Tabela 4.1: Tabela de conjuntos de dados selecionados

O conjunto de dados que agrega consultas e internamentos, possui 4199 episódios que se dividem em 868 consultas e 3331 internamentos. Contém igualmente 4 códigos distintos dispersos por 20660 diagnósticos e 26989 notas. Apesar da diversidade de códigos inicial, os rótulos selecionados são apenas 4 pois são os únicos que ocorrem em todos os anos pelos menos duas vezes tanto em internamentos como em consultas. Assim permitem-nos efetuar classificações em todos os anos das experiências utilizando apenas consultas ou internamentos e consequentemente os dois.

4.1.1.2 Conjunto de dados Base com rótulo adicional

Considerando o possível erro associado à eliminação dos diagnósticos de episódios, criámos um novo conjunto de dados. Para todos os episódios aos quais foram removidos diagnósticos, associou-se um novo diagnóstico com o código *"LabelD"*. Apesar de ter os mesmos episódios que o conjunto base, este novo conjunto tem 5 códigos distintos. Foram igualmente gerados dois subconjuntos exclusivamente com consultas e diagnósticos.

4.1.2 Método de estimação de desempenho

Para realizar experiências ao longo dos anos considerou-se a data de fim dos episódios pois estes apenas são classificados após o seu término. Assim, em cada experiência, considerámos para treino todos os episódios anteriores ao ano que queremos prever. Para teste, foram considerados todos os episódios terminados nesse ano, uma vez que geramos um novo modelo a cada ano.

A figura 4.1 exemplifica a seleção de diferentes episódios, para treino e teste, para a classificação de episódios relativos a 2012. Os episódios A, B e C são usados para treino pois terminam antes de 2012. Os episódios D,E e F são usados como teste pois terminam em 2012. Os episódios D E e F são descartados uma vez que terminam depois de 2012. Esta estimação de desempenho origina conjuntos de treino cada vez maiores com o decorrer dos anos.

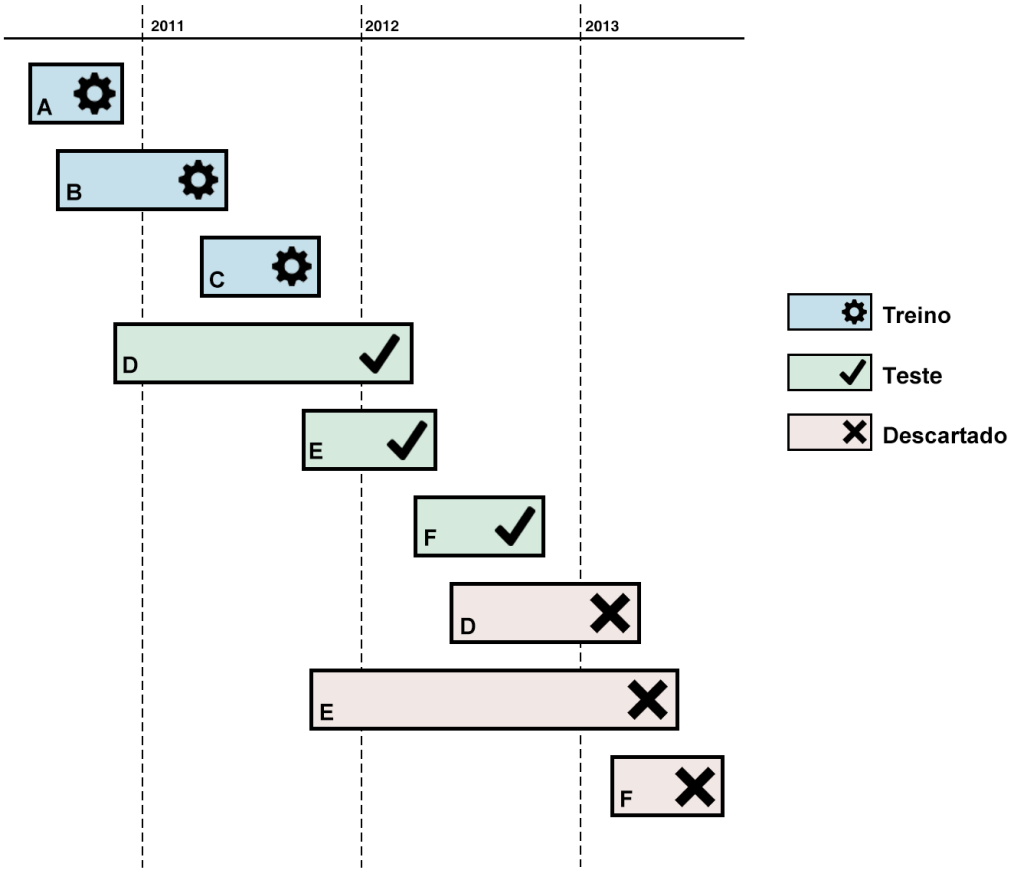


Figura 4.1: Seleção de dados de treino e teste para o ano 2012

4.1.3 Medidas de avaliação

De forma a avaliar as abordagens ao problema identificadas no capitulo 2 foram utilizadas diferentes métricas. Numa abordagem rótulo único foram utilizadas as métricas de avaliação Precisão e Sensibilidade de forma a avaliar as classificações positivas corretas. A medida F1, foi

utilizada como combinação das duas.

Em ambas as abordagens foram utilizadas Correspondência Exata e *Hamming Loss* como medidas de avaliação global do modelo. Apesar de ser uma medida rígida, a Correspondência Exacta foi utilizada para perceber o número de previsões cuja totalidade de rótulos foi prevista corretamente. Para perceber o número de códigos errados no total dos códigos utilizou-se a medida *Hamming Loss*.

4.1.4 Preparação de dados

De forma a podermos avaliar as notas clínicas, temos que transformar o texto inicial num formato próprio que possa ser analisado pelos classificadores. Inicialmente foram removidas palavras de paragem e posteriormente utilizámos o modelo *bag of words* ignorando gramática, ordem de palavras, ou semântica. Foi aplicada a medida tf-idf (frequência do termo–inverso da frequência nos documentos) primeiramente para atribuir um peso à ocorrência de cada palavra e posteriormente para diminuir o peso de palavras comuns que ocorrem demasiadamente. Aplicou-se também o método *CountVectorizer* para converter as notas numa matriz de contagem de *tokens*. Esta caracterização foi utilizada em todas as experiências realizadas neste estudo.

4.1.5 Software

No desenvolvimento do sistema de classificação de episódios clínicos foram utilizados o Natural Language Toolkit (NLTK) [nlt17] para processamento das notas médicas e o Scikit-Learn [SL] na criação dos modelos de classificação.

- **NLTK** - Plataforma desenvolvida em python que permite o acesso simples e intuitivo a diferentes recursos linguísticos. Oferece diferentes ferramentas que permitem, entre outros, analisar e categorizar texto, analisar estruturas linguísticas, converter texto em *tokens*, efetuar *stemming* e etiquetar classes de texto [nlt17].

- **Scikit-Learn** - Biblioteca de aprendizagem automática e modelação estatística implementada em python. Fornece diferentes algoritmos de classificação, regressão e agrupamento como, *SVM*s, *k*-nn, árvores de decisão e métodos baysianos, entre outros. Permite ainda a criação de *pipelines* que facilitam a elaboração de experiências e implementa diferentes métricas de avaliação [SL].

- **Scikit-Multilearn** - Biblioteca para classificação multirótulo construída sobre o ecosistema Scikit-Learn. Implementa por defeito, algoritmos e funcionalidades específicos de problemas multirótulo [SM].

A implementação destas bibliotecas e dos diferentes algoritmos foi baseada em diferentes artigos [Noo][Li][Sha].

4.1.6 Algoritmos

Conforme referido no capítulo anterior, no desenvolvimento do sistema de classificação foi utilizada a biblioteca Scikit-Learn. Nas experiências de Rótulo Único utilizámos o classificador *1vsRest* com diferentes algoritmos. Na primeira experiência, *Naive Bayes* e na seguinte *Logistic Regression* e *Linear Support Vector Classification*. Nas experiências seguintes introduzimos algoritmos Multirótulo. *Classifier Chains* e *Label Powerset* com *Logistic Regression* assim como *Binary Relevance* com *Gaussian Naive Bayes*. Utilizámos ainda o algoritmo Multirótulo adaptado do *K-nearest neighbours* (MLKnn) com um valor inicial de 10 vizinhos. Os restantes algoritmos usaram valores por defeito.

4.2 Experiências

Com a configuração experimental definida, elaborámos duas experiências, utilizando os dois conjuntos de dados definidos previamente.

4.2.1 Conjunto de dados com rótulos selecionados

Iniciámos esta experiência com uma abordagem Rótulo Único. Posteriormente aplicaram-se novos algoritmos para diferentes observações. Na primeira experiência utilizámos o conjunto de dados Base com os 4 rótulos selecionados.

4.2.1.1 NaiveBayes 1vsAll

Como primeira aproximação ao problema procurámos validar uma possível classificação Rótulo a Rótulo. Nesse sentido utilizámos unicamente o classificador *1vsRest* com o algoritmo *Naive Bayes*. Este algoritmo é um óptimo candidato pois obtém geralmente bons resultados com um custo computacional reduzido. Avaliámos o valor das precisão, sensibilidade e medida F1 ao longo dos 8 anos e para os diferentes tipos de episódio.

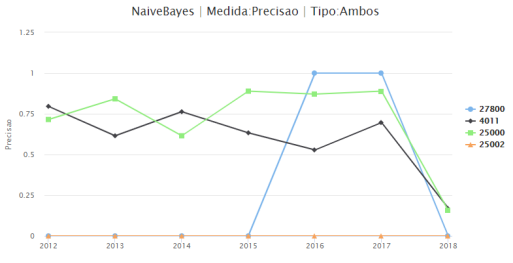


Figura 4.2: Resultados de Precisão para o algoritmo Naive Bayes em ambos os tipos de episódio

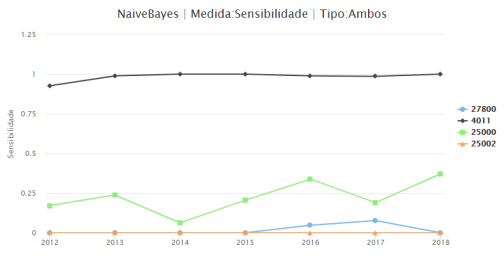


Figura 4.3: Resultados de Sensibilidade para o algoritmo Naive Bayes em consultas

Este modelo não consegue classificar correctamente o rótulo *25002*. Em nenhum momento atribuiu a classe positiva a esta etiqueta. Considerando a Precisão, o modelo é avaliado de forma

aceitável para os outros rótulos, dentro das consultas ou com ambos os episódios. Contudo não consegue ter o mesmo resultado para falsos positivos.

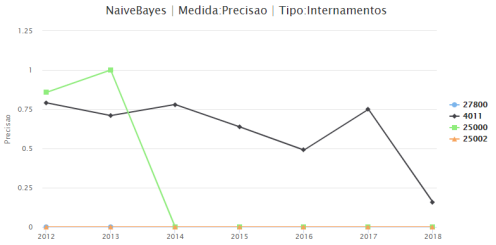


Figura 4.4: Resultados de Precisão para Naive Bayes em internamentos

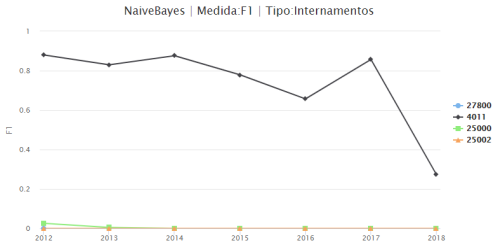


Figura 4.5: Resultados de Medida F1 para Naive Bayes em internamentos

Quando consideramos apenas os internamentos, o modelo tem um valor de próximo de 0,8 para a medida F1 para o rótulo *4011*, sofrendo uma quebra para 0.15 no último ano. Para os restantes rótulos a mesma medida tem um valor de 0.

4.2.1.2 Diferentes Algoritmos 1vsAll

Seguidamente introduzimos os restantes classificadores *1vsAll* com *Logistic Regression* e *Linear SVC*. Queremos avaliar o comportamento de mais algoritmos segundo a mesma abordagem Rótulo Único.

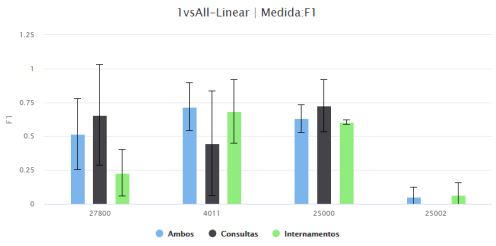


Figura 4.6: Resultados da Média da Medida F1 para 1vsAll Linear SVN

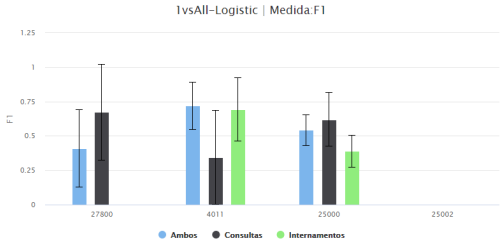


Figura 4.7: Resultados da Média da Medida F1 para 1vsAll Logistic Regression

O *Logistic Regression* é igualmente incapaz de prever a etiqueta *25002*, registando valores nulos para as 3 medidas. O Linear SVC tem um comportamento melhor e mais uniforme para os 3 tipos de episódios e é o único a identificar casos positivos para os 4 rótulos.

4.2.1.3 Diferentes algoritmos Multirótulo

Por fim, introduzimos os algoritmos Multi-Rótulo de forma a conseguirmos comparar os diferentes algoritmos e o seu comportamento ao longo dos anos. Avaliámos a correspondência exacta e as etiquetas mal previstas para todos os algoritmos implementados e para os diferentes tipos de episódio.

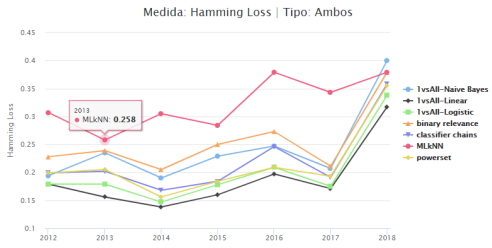


Figura 4.8: Resultados Hamming Loss para ambos os tipos de episódio

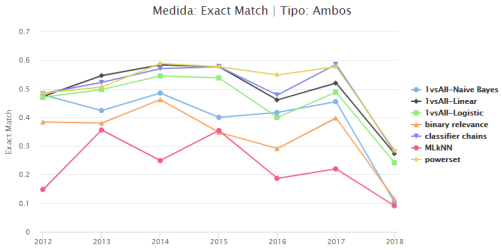


Figura 4.9: Resultados Correspondência Exacta para ambos os tipos de episódio

Para as diferentes medidas e tipos de episódio observamos que os 7 algoritmos seguem uma curva semelhante ao longo dos anos. Regra geral, de um ano para outro, os valores das medidas aumentam ou diminuem, da mesma forma para os diferentes algoritmos. Os valores para *Hamming Loss* aumentam ligeiramente ao longo do tempo registando um pico no último ano. A taxa de previsões exatas segue uma distribuição inversa, decrescendo ao longo do tempo com uma quebra no último ano. O algoritmo que classifica incorrectamente menos etiquetas é o *Linear Regression*. Para os três tipos de episódios obteve os valores mais baixos. Na previsão totalmente correta, apesar de ter um dos valores mais altos, *Powersets* e *Classifier Chains* têm os melhores resultados. Para ambas as métricas, *MLKnn* regista o pior comportamento, seguido de perto por *Binary Relevance* na Correspondência Exacta.

4.2.2 Rótulo adicional

Nesta experiência explorou-se a adição de um rótulo aos episódios que perderam diagnósticos, de forma a reduzir o erro associado. Para cada algoritmo, comparámos os dois conjuntos de dados nos diferentes tipos de episódio. Usámos as médias de *Hamming Loss* e Correspondência Exacta. Calculámos ainda o desvio padrão para perceber se possíveis diferenças seriam realmente significativas.

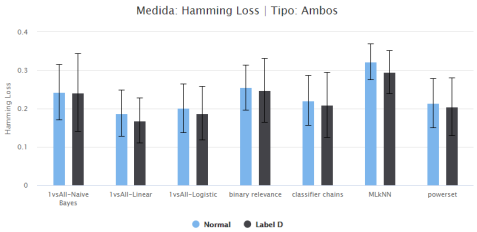


Figura 4.10: Resultados dos dois conjuntos de dados para Hamming Loss em ambos os tipos de episódio

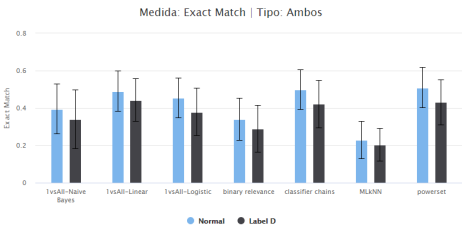


Figura 4.11: Resultados dos dois conjuntos de dados para Correspondência Exacta em ambos os casos

Observámos que, para os internamentos e com ambos os episódios, a medida *Hamming Loss* é menor no conjunto de dados com rótulo D. Contrariamente, a correspondência Exacta é maior no conjunto de dados original. Dos 7 algoritmos testados e para os diferentes episódios, *Linear SVN* é

o que tem melhores valores para *Hamming Loss*. O mesmo algoritmo tem os melhores resultados para Correspondência Exacta, a par de *powersets* e *classifier chains* identificados anteriormente.

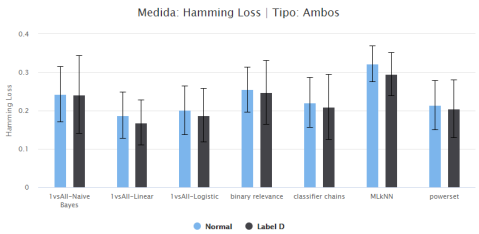


Figura 4.12: Resultados dos dois conjuntos de dados para Hamming Loss em consultas

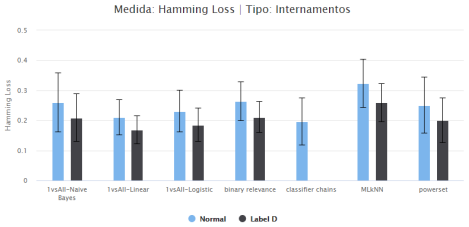


Figura 4.13: Resultados dos dois conjuntos de dados para Hamming Loss em internamentos

Comparando os diferentes tipos de episódio, os algoritmos obtiveram os melhores resultados quando utilizaram unicamente dados relativos a consultas.

No caso dos internamentos, o *Classifier Chain* foi incapaz de gerar um modelo preditivo com a etiqueta D. Este classificador necessita da presença de ambas as classes nos dados de teste o que não se verificava para o rótulo D, no último ano de teste.

4.3 Resumo

No decorrer das diferentes experiências, os algoritmos que de forma consistente tiveram os melhores resultados foram *Classifier Chain*, *powersets* e *1vsAll* com *Linear SVN*. Utilizando este último uma abordagem Rótulo Único, comprovámos a validade desta estratégia.

Os resultados observados na secção 4.2.1.3 permitem-nos concluir que o método de estimação utilizado não traz grandes vantagens. Com o decorrer dos anos e com conjuntos de treino maiores, os resultados não melhoram.

A introdução do rótulo D não trouxe melhorias. A diminuição da fração de códigos mal previstos é facilmente justificada pela elevada densidade dessa etiqueta e comprovada pela diminuição de correspondências exactas. Dada essa elevada densidade, os algoritmos passam a classificar positivamente o rótulo adicionado, em quase todas as instâncias de teste. Desta forma, classificam corretamente essa etiqueta aumentando a taxa de rótulos corretamente prevista. Por outro lado, considerando a totalidade de códigos de cada instância preveem de forma exacta menos instâncias.

Apesar dos resultados observados na segunda experiência, existe uma sobreposição dos intervalos de erro não se podendo concluir sobre os mesmos.

Pela mesma razão, não podemos tirar conclusões sobre os melhores resultados gerados pelos modelos nos diferentes momentos em que isolámos as consultas.

Os resultados completos encontram-se no anexo B

Capítulo 5

Conclusões e Trabalho Futuro

5.1 Conclusões

A classificação de notas clínicas é um processo lento, sujeito a erros e não uniforme quando efetuado por diferentes profissionais. A incorrecta atribuição de códigos de diagnóstico a episódios afeta a gestão hospitalar. Pode igualmente condicionar a elaboração do histórico clínico de um paciente. Este trabalho, desenvolvido em parceria com a *Glintt*, teve como objetivo a criação de um modelo de classificação de episódios clínicos que pudesse auxiliar os profissionais hospitalares nesta tarefa.

Durante a sua realização foram explorados métodos de pré-processamento de texto aplicados às notas clínicas assim como algoritmos de classificação multirótulo. Dada a complexidade e tamanho das notas, deveria ter havido um maior esforço no tratamento das mesmas. Apesar do elevado número de episódios e 3929 códigos distintos fornecidos inicialmente, acabámos por utilizar apenas 4. A seleção de dados, influenciada pela estimação de desempenho, acabou por limitar o estudo das técnicas de classificação no número rótulos utilizados. O problema subjacente a este trabalho representa um bom caso de estudo dado o elevado número de rótulos, acabando por ser demasiadamente reduzido.

Os resultados obtidos são positivos, havendo uma melhoria relativamente ao caso de referência. Os algoritmos com melhores resultados foram *Classifier Chains* com regressão logística, *Label Powersets* e *1 vs Rest* com regressão linear. A taxa de acerto exacto médio das instâncias previstas, com estes algoritmos, foi de 0.557, 0.51 e 0.49. Relativamente à medida *Hamming Loss*, que avalia a fração dos códigos previstos incorrectamente, estes 3 algoritmos tiveram valores médios de 0.204, 0.22 e 0.184, respetivamente.

Transversalmente, as medidas utilizadas tiveram melhores resultados com um conjunto de dados que constituído unicamente por consultas. Este facto está associado aos factores identificados na secção 3.3.2 - menor comprimento das notas, menor número de notas por episódio e menos diagnósticos por episódio. A utilização de modelos diferentes para consultas e internamentos poderá gerar melhores resultados. Relativamente à distribuição temporal dos episódios, observámos que a utilização de mais dados do passado não nos garante melhores resultados. Assim, poderia

ter sido usado outro método de estimação. Um método de estimação diferente poderia igualmente permitir um estudo com mais códigos de diagnóstico uma vez que o escolhido limitou a utilização dos dados.

Não foram obtidos melhores resultados com a utilização de um rótulo adicional. Apesar de uma maior taxa de acerto no total dos rótulos, o número de instâncias totalmente correctas é menor.

Apesar das limitações associadas, os resultados obtidos demonstram que será positivo aprofundar o estudo efetuado neste trabalho.

5.2 Trabalho Futuro

O melhoramento da qualidade dos dados poderá originar melhores resultados. Primeiramente utilizando um conjunto de dados com uma maior diversidade de códigos e um melhor balanceamento da sua distribuição. Depois através de uma melhor qualidade das notas recolhidas, com menos informação irrelevante para a classificação e menos erros. Apesar das limitações do contexto real em que as notas são produzidas, um melhoramento da qualidade das mesmas poderá gerar melhores resultados e viabilizar a introdução deste tipo de modelos em contextos reais. Nas experiências desenvolvidas, a maioria do esforço incidiu nos algoritmos de classificação. Um melhoramento do pré-processamento de texto poderá permitir melhores resultados. Adicionalmente, a experimentação com outro tipo de configurações dos algoritmos utilizados poderá melhorar os resultados obtidos.

O melhoramento da qualidade dos dados poderá igualmente permitir diferentes conclusões sobre a dimensão temporal e tipologia dos episódios clínicos. Por outro lado, a fase de experimentação deverá ser suportada por testes estatísticos.

A estimação de desempenho limitou a diversidade de códigos utilizados e consequentemente os número de episódios. Não tendo sido obtidos melhores resultados com o decorrer dos anos, as experiências realizadas podem ser simplificadas, com outra estimação de desempenho, sendo possível utilizar a maioria dos dados validados. Com mais códigos e uma distribuição mais equilibrada poderão explorar-se outras abordagens. Nesta situação, com menos códigos eliminados, a utilização do rótulo adicional poderá gerar outros resultados uma vez que irá ocorrer em menos episódios.

No âmbito da presente dissertação ficou por explorar a componente hierárquica dos episódios. Uma abordagem inicial a classes superiores, pode contribuir para a diminuição de códigos incorrectamente previstos. Da mesma forma, o modelo deverá beneficiar da utilização de dados do paciente associado ao episódio (sócio-demográficos, históricos clínico, medicação, entre outros).

Referências

[BNKC+18] Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad e Noemie Elhadad. Multi-label classification of patient notes: case study on icd code assignment. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[CXZ+15] Zhihua Cai, Dong Xu, Qing Zhang, Jiexia Zhang, Sai-Ming Ngai e Jianlin Shao. Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Molecular BioSystems*, 11(3):791–800, 2015.

[dS05] Serviço Nacional de Saúde. Classificação internacional de doenças 9ª revisão modificação clínica. Disponível em http://www2.acss.min-saude.pt/DepartamentoseUnidades/DepartamentoGest%C3%A3oeFinanciamentoPrestSa%C3%BAde/Codifica%C3%A7%C3%A3oCl%C3%ADnica/tabid/358/language/en-US/Default.aspx, Maio 2005.

[DZ11] Mita K Dalal e Mukesh A Zaveri. Automatic text classification: a technical review. *International Journal of Computer Applications*, 28(2):37–40, 2011.

[JG17] Katti Faceli Ana Lorena e Márcia Oliveira João Gama, André Carvalho. *Extração de conhecimento de dados*. Edições Sílabo, 2017.

[Kar18] Amitabha Karmakar. Classifying medical notes into standard disease codes using machine learning. *arXiv preprint arXiv:1802.00382*, 2018.

[KTA16] Manana Khachidze, Magda Tsintsadze e Maia Archuadze. Natural language processing based instrument for classification of free text medical records. *BioMed research international*, 2016, 2016.

[Li] Susan Li. Multi label text classification with scikit-learn. URL: https://towardsdatascience.com/multi-label-text-classification-with-scikit-learn-30714b7819c5.

[LYNB08] Lucian Vlad Lita, Shipeng Yu, Stefan Niculescu e Jinbo Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*, 2008.

[Nig16] Priyanka Nigam. Applying deep learning to icd-9 multi-label classification from medical records. Technical report, Technical report, Stanford University, 2016.

[nlt17] nltk. Usage statistics of content languages for websites, 2017. Last accessed 16 September 2017. URL: http://w3techs.com/technologies/overview/content_language/all.

REFERÊNCIAS

[Noo] Kartik Nooney. Deep dive into multi-label classification..! (with detailed case study). URL: https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff.

[PBM+07] John P Pestian, Christopher Brew, Paweł Matykiewicz, Dj J Hovermale, Neil Johnson, K Bretonnel Cohen e Włodzisław Duch. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 97–104. Association for Computational Linguistics, 2007.

[PPN+13] Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood e Noémie Elhadad. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2):231–237, 2013.

[RK04] Ryan Rifkin e Aldebaro Klautau. In defense of one-vs-all classification. *Journal of machine learning research*, 5(Jan):101–141, 2004.

[SCML13] Newton SpolaôR, Everton Alvares Cherman, Maria Carolina Monard e Huei Diana Lee. A comparison of multi-label feature selection methods using the problem transformation approach. *Electronic Notes in Theoretical Computer Science*, 292:135–151, 2013.

[Sha] Javed Shaikh. Machine learning, nlp: Text classification using scikit-learn, python and nltk. URL: https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nlt

[SL] Scikit-Learn. Scikit-learn. URL: http://https://scikit-learn.org.

[SLD17] Sumedh Sontakke, Jay Lohokare e Reshul Dani. Diagnosis of liver diseases using machine learning. In *2017 International Conference on Emerging Trends & Innovation in ICT (ICEI)*, pages 129–133. IEEE, 2017.

[SM] Scikit-Multilearn. Scikit-multilearn. URL: http://http://scikit.ml.

[TÇG+11] Dilara Torunoğlu, Erhan Çakirman, Murat Can Ganiz, Selim Akyokuş e M Zahid Gürbüz. Analysis of preprocessing methods on classification of turkish texts. In *2011 International Symposium on Innovations in Intelligent Systems and Applications*, pages 112–117. IEEE, 2011.

[VD15] S Vijayarani e S Dhayanand. Data mining classification algorithms for kidney disease prediction. *International Journal on Cybernetics & Informatics (IJCI)*, 4(4):13–25, 2015.

[VIN15] S Vijayarani, Ms J Ilamathi e Ms Nithya. Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1):7–16, 2015.

[Zha08] Yitao Zhang. A hierarchical approach to encoding medical concepts for clinical notes. In *Proceedings of the ACL-08: HLT Student Research Workshop*, pages 67–72, 2008.

[ZZ07] Min-Ling Zhang e Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.

REFERÊNCIAS

[ZZ13] Min-Ling Zhang e Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013.

REFERÊNCIAS

Anexo A

Análise Exploratória de dados

A.1 Distribuições de ocorrências códigos e notas por episódio

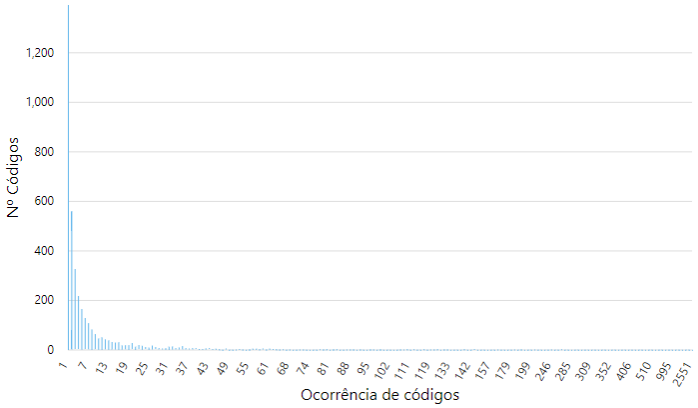


Figura A.1: Número de ocorrências de códigos

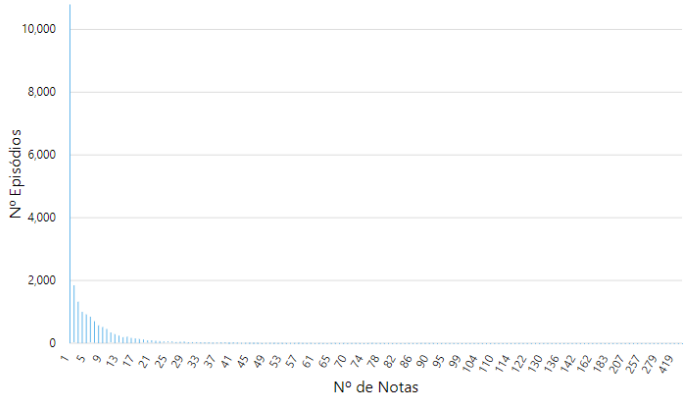


Figura A.2: Número de notas por episódio

Anexo B

Resultados

B.1 Conjunto de dados com 4 rótulos

B.1.1 NaiveBayes 1vsAll

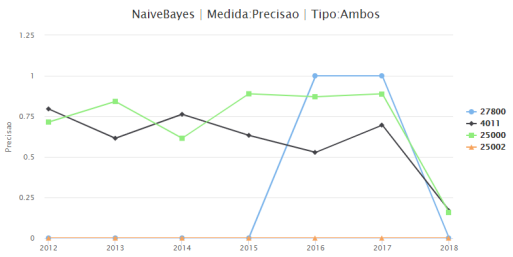


Figura B.1: Resultados de Precisão para Naive Base em ambos os tipos de episódio

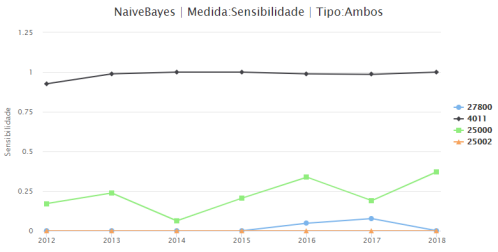


Figura B.2: Resultados de Sensibilidade para Naive Base em ambos os tipos de episódio

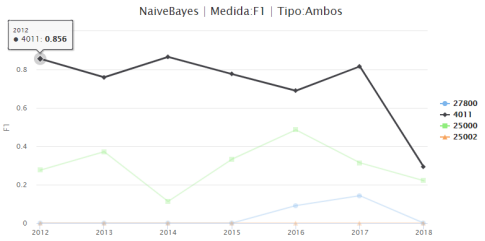


Figura B.3: Resultados de Medida F1 para Naive Base em ambos os tipos de episódio

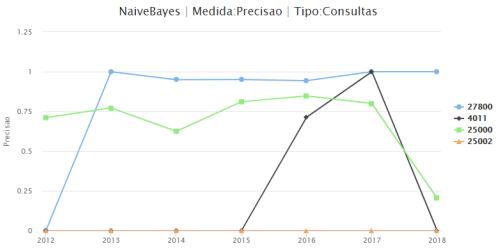


Figura B.4: Resultados de Precisão para Naive Base em consultas

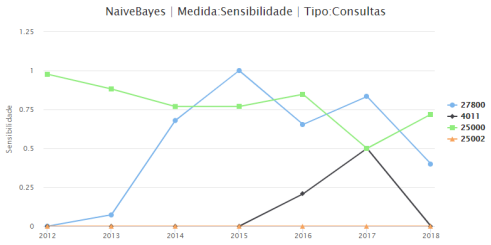


Figura B.5: Resultados de Sensibilidade para Naive Base em consultas

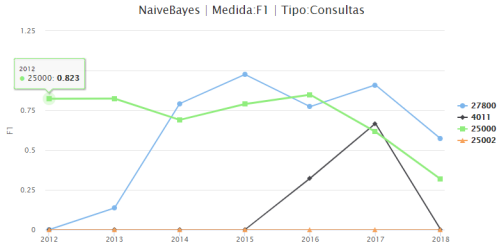


Figura B.6: Resultados de Medida F1 para Naive Base em consultas

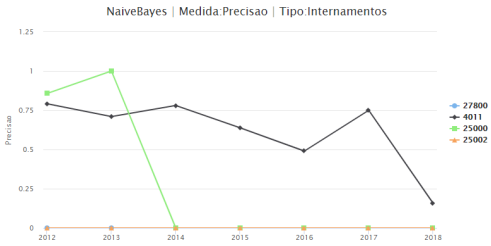


Figura B.7: Resultados de Precisão para Naive Base em internamentos

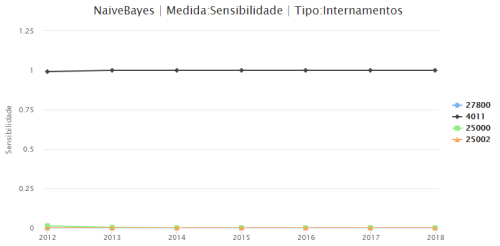


Figura B.8: Resultados de Sensibilidade para Naive Base em internamentos

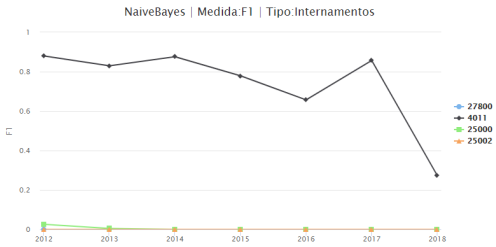


Figura B.9: Resultados de Medida F1 para Naive Base em internamentos

B.1.2 Multirótulo 1vsAll

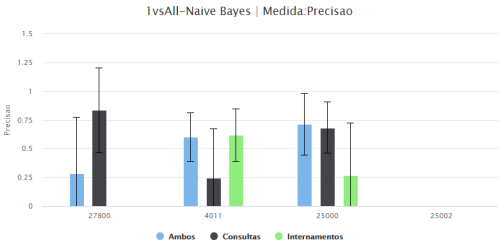


Figura B.10: Resultados da Média da Precisão para 1vsAll Naive Bayes

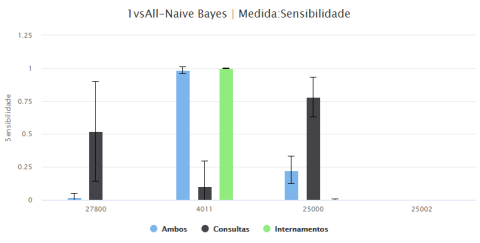


Figura B.11: Resultados da Média da Sensibilidade para 1vsAll Naive Bayes

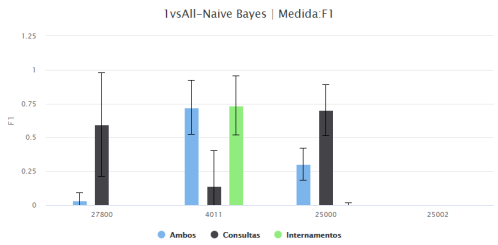


Figura B.12: Resultados da Média da Medida F1 para 1vsAll Naive Bayes

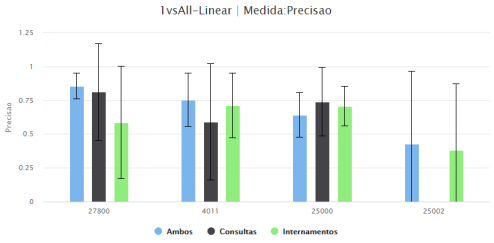


Figura B.13: Resultados da Média da Precisão para 1vsAll Linear SVN

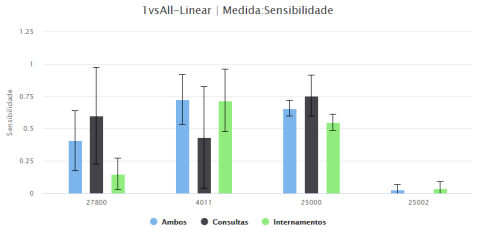


Figura B.14: Resultados da Média da Sensibilidade para 1vsAll Linear SVN

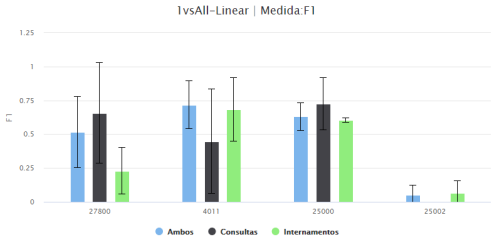


Figura B.15: Resultados da Média da Medida F1 para 1vsAll Linear SVN

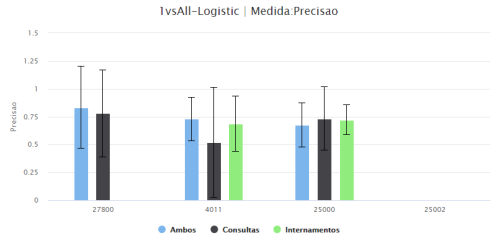


Figura B.16: Resultados da Média da Precisão para 1vsAll Logistic Regression

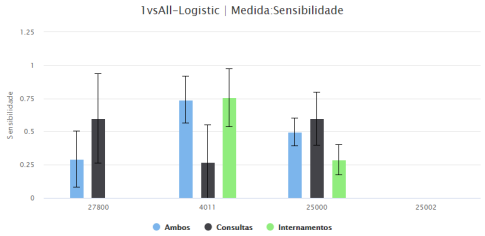


Figura B.17: Resultados da Média da Sensibilidade para 1vsAll Logistic Regression

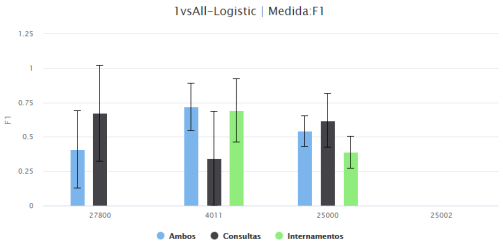


Figura B.18: Resultados da Média da Medida F1 para 1vsAll Logistic Regression

B.1.3 Diferentes Algoritmos Multirótulo

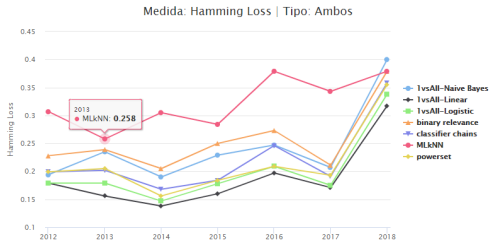


Figura B.19: Resultados Hamming Loss para ambos os tipos de episódio

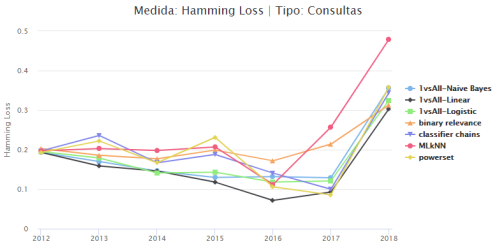


Figura B.20: Resultados Hamming Loss para consultas

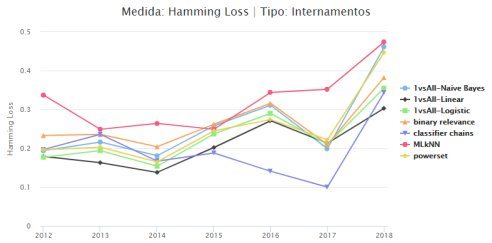


Figura B.21: Resultados Hamming Loss parainternamentos

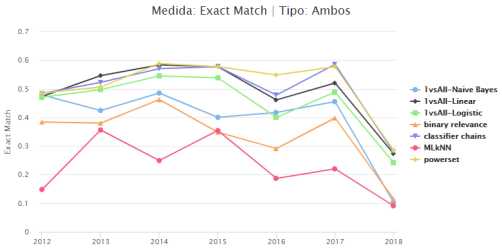


Figura B.22: Resultados Correspondência Exacta para ambos os tipos de episódio

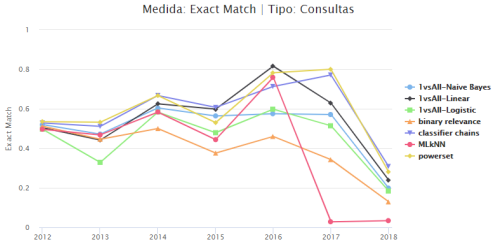


Figura B.23: Resultados Correspondência Exacta para consultas

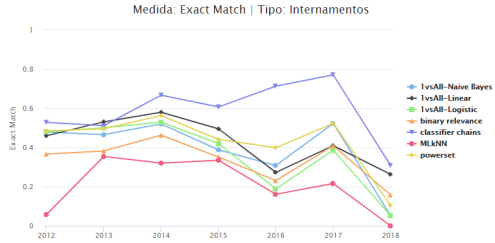


Figura B.24: Resultados Correspondência Exacta para internamentos

B.2 Rótulo D

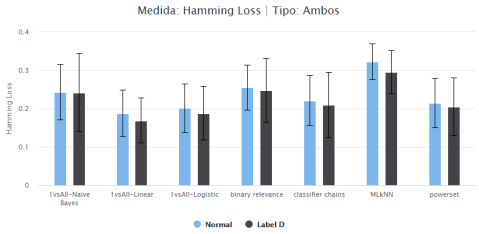


Figura B.25: Resultados dos dois conjuntos de dados para Hamming Loss em ambos os tipos de episódio

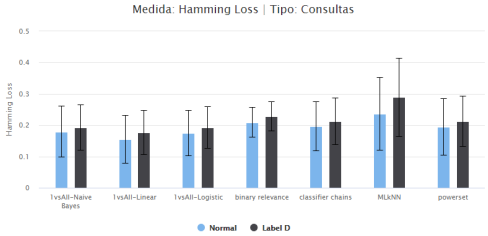


Figura B.26: Resultados dos dois conjuntos de dados para Hamming Loss em consultas

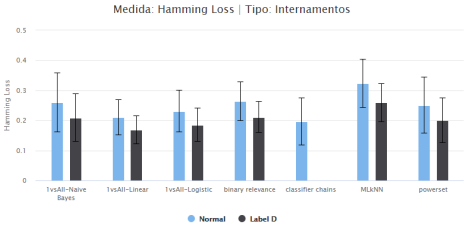


Figura B.27: Resultados dos dois conjuntos de dados para Hamming Loss em internamentos

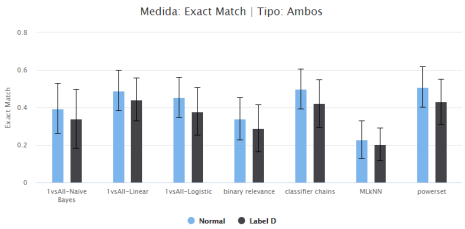


Figura B.28: Resultados dos dois conjuntos de dados para Correspondência Exacta em ambos os tipos de episódios

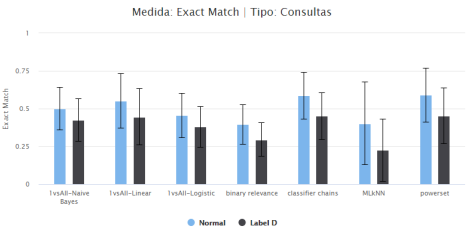


Figura B.29: Resultados dos dois conjuntos de dados para Correspondência Exacta em consultas

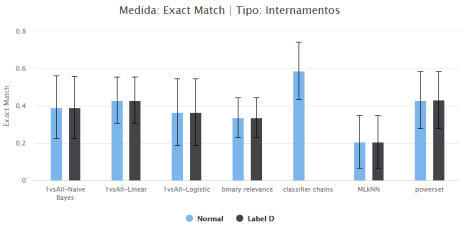


Figura B.30: Resultados dos dois conjuntos de dados para Correspondência Exacta em internamentos