
AGENDAMENTO DINÂMICO DE CIRCUITOS DE RECOLHA
SELETIVA DE RESÍDUOS PORTA-A-PORTA NÃO RESIDENCIAL

Nuno Miguel de Freitas Filipe

Dissertação

Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão

Orientado por

Professor Doutor João Manuel Portela da Gama, FEP

2018

Nota Biográfica

Nuno Miguel de Freitas Filipe nasceu em 1977, na cidade de Lisboa, Portugal.

Licenciado em Economia no ano 2000, pela Universidade Lusíada de Lisboa, inicia a sua carreira profissional na *Arthur Andersen*, onde participa em diversos projetos na área de contabilidade. Em 2001, ingressa na Sonae (Modelo Continente, S.A.), onde exerce funções de auditor no Departamento de Auditoria e Gestão de Risco da empresa, reportando diretamente ao Conselho de Administração.

Em 2005, inicia novo desafio profissional, na cidade do Porto, na LIPOR (Serviço Intermunicipalizado de Gestão de Resíduos do Grande Porto) como técnico superior no Gabinete de Controlo de Gestão, onde atualmente desempenha funções de Chefe de Divisão de Gestão e Sistemas de Informação.

Em representação da LIPOR, tem participado em diversos grupos de trabalho e fóruns a nível nacional e internacional, junto da Comissão Europeia, Ministério do Ambiente, Agência Portuguesa do Ambiente, Sociedade Ponto Verde, entre outras empresas e organizações, na discussão de diversas temáticas e desafios do setor dos resíduos.

A nível de formação profissional, tem realizado uma aposta contínua no desenvolvimento das suas competências, possuindo uma especialização em Auditoria Financeira pelo INDEG/ISCTE, um Mini-MBA pela EGE (*Atlantic Business School*), uma Pós-Graduação em Direção de Empresas pela PBS (*Porto Business School*), tendo ainda realizado nesta instituição diversos cursos sobre diferentes temáticas, com particular destaque para as áreas de Logística, Análise Financeira e Estratégia. Atualmente, frequenta o mestrado em Modelação e Análise de Dados e Sistemas de Apoio à Decisão, na FEP (Faculdade de Economia do Porto).

Agradecimentos

Gostaria de agradecer a todas as pessoas que direta ou indiretamente estiveram presentes comigo nesta caminhada.

Um especial agradecimento à LIPOR, pela sua visão e oportunidades que coloca ao desenvolvimento profissional e pessoal dos seus colaboradores.

Um grande obrigado ao Professor João Gama pela orientação prestada e pela sua dedicação ao projeto MADSAD.

Por último, uma palavra de amor e carinho à minha esposa Clara, pelo seu enorme apoio e compreensão por todas as horas que não estive presente, e um grande obrigado pela nossa maior conquista e desafio, o nosso filho Gil.

Resumo

A tendência de concentração populacional em grandes centros urbanos tem colocado elevadas pressões e desafios no planeamento, gestão e desenvolvimento das cidades. Têm sido realizadas diversas abordagens, rotuladas de *Smart Cities*, para estruturação de um modelo de desenvolvimento urbano sustentável. Um novo paradigma económico designado por “Economia Circular”, surge recentemente como uma prioridade da política europeia e nacional, para o uso sustentável dos recursos, encontrando-se definidas ambiciosas metas para a valorização de resíduos. Neste contexto, assume particular relevância a implementação e monitorização de circuitos de recolha seletiva porta-a-porta.

Com a presente tese, pretende-se contribuir para uma eficiente gestão de circuitos de recolha seletiva porta-a-porta não residencial (PAP-NR). É proposta uma solução pouco onerosa, face a soluções de sensorização e monitorização já existentes no mercado. Com base em dados históricos reais de recolha, assegura um agendamento dinâmico e flexível de circuitos de recolha, por via da previsão da colocação de resíduos à recolha pelos clientes e respetivo número total de contentores, e da identificação de regras que caracterizam a produção de resíduos pelos clientes. A solução proposta permitirá a obtenção de vantagens a nível de eficiências operacionais, com o consequente impacto na redução de custos da atividade de recolha e melhoria do nível de serviço prestado aos clientes.

O desenvolvimento dos modelos preditivos propostos foi efetuado com recurso à metodologia CRISP-DM, tendo sido utilizado um *data set* correspondente a um total de 26.177 registos de recolha, de janeiro a dezembro de 2017, num total de 141 clientes.

Com a adoção de árvores de decisão para abordagem aos problemas de previsão de colocação de resíduos à recolha pelos clientes (W-J48) e número de contentores produzidos (*Gradient Boosted Trees*), são obtidos os melhores desempenhos com taxas de acerto médias de 97,45% e 91,23%, respetivamente, para a amostra de clientes utilizada. Na geração das respetivas regras de decisão, foi adotado o algoritmo W-JRip, tendo sido apuradas 29 e 13 regras, com taxas de acerto de 73,46% e 84,78%, respetivamente.

Palavras-chave: *Smart Cities*, Recolha seletiva porta-a-porta, *RapidMiner*, CRISP-DM, Métodos preditivos, *Naive Bayes*, Regressão Logística, Árvores e regras de decisão.

Abstract

The global trend of increasing population concentration in large urban areas, presents many challenges for the planning, management and development of cities. There have been several approaches, commonly labelled of “Smart Cities”, to structure a sustainable urban development model. A new economic paradigm, called “Circular Economy” as recently emerged, as a priority of European and National policy, to address the sustainable use of natural resources, with ambitious targets for waste recovery. In such context, the implementation and monitoring of door-to-door waste collection schemes is particularly relevant.

This thesis aims to contribute for an efficient management of non-residential door-to-door selective waste collection schemes. It is proposed a solution, based on real waste collection data, for a dynamic and flexible scheduling, as an alternative for expensive container sensors and monitoring solutions already available in the market. The solution provides a prediction of the waste produced by the costumers and generates decision rules in order to provide insightful information on costumer’s waste production. Furthermore, provides operational efficiencies on waste collection operations, by reducing its costs and improving the service level provided to costumers.

The proposed predictive models were developed according to CRISP-DM standards, using a data set of 26,177 collection records regarding a total of 141 customers, from January to December 2017.

The best performances were achieved using the decision tree algorithms W-J48, for predicting the existence of waste to be collected in customer location, and Gradient Boosted Trees, to forecast the number of containers to collect, accounting an average accuracy of 97.45% and 91.23%, respectively, for a customer sample previously selected. To generate decision rules, W-JRip algorithm was adopted for both situations, with 29 and 13 rules being uncovered, with accuracy levels of 73.46% and 84.78%, respectively.

Keywords: Smart Cities, Door-to-door selective waste collection, RapidMiner, CRISP-DM, Predictive methods, Naive Bayes, Logistic Regression, Decision trees, Decision rules.

Índice

Nota Biográfica.....	ii
Agradecimentos	iii
Resumo	iv
Abstract	v
Índice	vi
Índice de Figuras	vii
Índice de tabelas	ix
1. Introdução e Motivação.....	1
2. Problema em análise.....	3
2.1. Enquadramento.....	3
2.2. Agendamento dinâmico de circuitos PAP-NR.....	3
3. Revisão da literatura	6
3.1. <i>Smart Cities</i> - Enquadramento.....	6
3.2. <i>Smart Cities</i> – Definição e evolução do conceito	7
3.3. <i>Smart Cities</i> – Modelos de desenvolvimento e desafios	10
3.4. Uma Gestão de Resíduos Urbanos <i>Smart</i>	13
3.5. <i>Data Mining</i> na otimização de circuitos de recolha seletiva de resíduos.....	15
4. Metodologia aplicada	17
4.1. Formulação e metodologia aplicada.....	17
4.2. Compreensão do negócio	18
4.3. Compreensão dos dados.....	21
4.3.1 Recolha dos dados	21
4.3.2 Análise do número de contentores recolhidos	22
4.3.3 Análise da eficiência dos circuitos de recolha	34
4.3.4 Análise da participação dos clientes aderentes.....	37
4.4. Preparação dos dados.....	42
4.5. Modelação	44
4.6. Avaliação	59
4.7. Implementação	69
5. Conclusões	70
6. Referências Bibliográficas	73
7. Anexos.....	75

Índice de Figuras

Figura 1: Características e fatores de uma <i>Smart City</i> (Fonte: adaptado de Giffinger et al., 2007)	8
Figura 2: <i>Urban Information Model</i> de Harrison (Fonte: adaptado de Harrison e Donnelly, 2011)	11
Figura 3: Etapas da metodologia CRISP-DM.....	18
Figura 4: Tipologia de circuitos de recolha de resíduos urbanos.....	19
Figura 5: Entradas de recicláveis por origem (Fonte: adaptado de Relatório e Contas LIPOR 2017).....	20
Figura 6: N.º de clientes e peso % por área de negócio, obtido via <i>Microsoft Excel</i>	22
Figura 7: Distribuição do n.º contentores recolhidos e peso % por área negócio e material, obtido via <i>Microsoft Excel</i>	23
Figura 8: Distribuição do n.º de contentores recolhidos por dia da semana, material e área de negócio, obtido via <i>Microsoft Excel</i>	24
Figura 9: Distribuição do n.º de contentores recolhidos por mês, material e área de negócio, obtida via <i>Microsoft Excel</i>	25
Figura 10: <i>Boxplots</i> do n.º total de contentores recolhidos por circuito de recolha, mês e material, obtidos via SPSS.....	26
Figura 11: Frequência do n.º total de contentores recolhidos por circuito de recolha, obtida via SPSS	28
Figura 12: Evolução diária do n.º de contentores recolhidos por circuito, obtida via <i>Microsoft Excel</i>	29
Figura 13: <i>Boxplots</i> do n.º de contentores recolhidos por ponto de recolha, mês e material, obtidas via SPSS	30
Figura 14: Frequência do n.º de contentores recolhidos e peso %, por ponto de recolha, obtido via SPSS .	32
Figura 15: Evolução mensal da frequência do n.º de contentores recolhidos e peso %, por ponto de recolha, obtido via SPSS.....	33
Figura 16: Cálculo da Taxa de Eficiência dos Circuitos de Recolha - TEC (%).....	34
Figura 17: Distribuição da Taxa de Eficiência da Recolha por mês e dia da semana, obtida via <i>Microsoft Excel</i>	35
Figura 18: Evolução do n.º de pontos recolha agendados <i>versus</i> recolhidos/ visitados, obtida via <i>Microsoft Excel</i>	35
Figura 19: Frequências de pontos de recolha recolhidos/ visitados, obtidas via SPSS.....	36
Figura 20: Cálculo da Taxa de Colocação de Resíduos à Recolha - TCR (%).....	37
Figura 21: Taxa de Colocação à Recolha total e por área de negócio, obtida via <i>Microsoft Excel</i>	37
Figura 22: Histograma da Taxa de Colocação à Recolha, obtido via SPSS.....	38
Figura 23: Taxa de Colocação à Recolha por dia da semana e área de negócio, obtida via <i>Microsoft Excel</i>	39
Figura 24: Taxa de Colocação à Recolha por mês e área de negócio, obtida via <i>Microsoft Excel</i>	40
Figura 25: Evolução do n.º de contentores recolhidos por visita/ cliente, obtida via <i>Microsoft Excel</i>	41
Figura 26: Evolução mensal do n.º de visitas por cliente, obtida via <i>Microsoft Excel</i>	41
Figura 27: Modelo preditivo implementado problema 1.1, <i>software RapidMiner</i>	49
Figura 28: Modelo preditivo problema 1.1 – detalhe do processo de Validação Cruzada, <i>software RapidMiner</i>	49
Figura 29: Árvore de decisão W-J48 obtida para um cliente – problema 1.1, obtida via <i>software RapidMiner</i> . 50	

Figura 30: Exemplos de regras de decisão geradas pelo algoritmo W-JRip – problema 1.2, obtido via <i>RapidMiner</i>	51
Figura 31: Modelo de geração de regras de decisão implementado – problema 1.2, <i>software RapidMiner</i>	52
Figura 32: Modelo de geração de regras de decisão – problema 1.2 – detalhe do processo de Validação Cruzada, <i>software RapidMiner</i>	52
Figura 33: Modelo preditivo implementado problema 2.1, <i>software RapidMiner</i>	54
Figura 34: Modelo preditivo problema 2.1 – detalhe do processo de Validação Cruzada, <i>software RapidMiner</i>	54
Figura 35: Árvore de decisão W-J48 obtida para um cliente – problema 2.1, via <i>software RapidMiner</i>	55
Figura 36: Exemplos de regras de decisão geradas pelo algoritmo W-JRip – problema 2.2, obtido via <i>RapidMiner</i>	57
Figura 37: Modelo de geração de regras de decisão implementado – problema 2.2, <i>software RapidMiner</i>	58
Figura 38: Modelo de geração de regras de decisão – problema 2.2 – detalhe do processo de Validação Cruzada, <i>software RapidMiner</i>	58

Índice de tabelas

Tabela 1: Metas PERSU 2020 e valores reais atingidos em 2016 (Fonte: adaptado de Ministério do Ambiente, 2014 e ERSAR, 2017)	2
Tabela 2: Encaminhamento de resíduos por destino final (Fonte: adaptado de Relatório e Contas LIPOR 2017)	20
Tabela 3: Evolução do n.º de clientes aderentes ao PAP-NR, obtido via <i>Microsoft Excel</i>	23
Tabela 4: Estatísticas descritivas do n.º total de contentores recolhidos por circuito de recolha, obtidas via SPSS	27
Tabela 5: Estatísticas descritivas do n.º de contentores recolhidos por ponto de recolha, obtidas via SPSS	31
Tabela 6: Estatísticas descritivas das variáveis pontos recolhidos e pontos agendados, obtidas via SPSS	36
Tabela 7: Descrição dos atributos do modelo preditivo	44
Tabela 8: Problemas em análise, abordagem metodológica e <i>outputs</i>	45
Tabela 9: Problemas em análise e respetivos <i>data sets</i>	45
Tabela 10: <i>Ranking</i> das medidas de avaliação Taxa de Acerto e <i>F-score</i> (valores médios) obtidas por algoritmos de classificação – problema 1.1, obtido via <i>RapidMiner</i> e trabalhado em <i>Microsoft Excel</i>	48
Tabela 11: N.º regras geradas e <i>ranking</i> das medidas de avaliação Taxa de Acerto e <i>F-score</i> obtidas por algoritmos geradores de regras de decisão – problema 1.2, obtido via <i>RapidMiner</i> e trabalhado em <i>Microsoft Excel</i>	51
Tabela 12: <i>Ranking</i> das medias de avaliação Taxa de Acerto e <i>F-score</i> (valores médios) obtidos por algoritmos de classificação – problema 2.1, obtido via <i>RapidMiner</i> e trabalhado em <i>Microsoft Excel</i>	53
Tabela 13: N.º regras geradas e <i>ranking</i> do nível de Taxa de Acerto e <i>F-score</i> obtidos por algoritmos geradores de regras de decisão – problema 2.2, obtido via <i>RapidMiner</i> e trabalhado em <i>Microsoft Excel</i>	56
Tabela 14: Matriz de Confusão	59
Tabela 15: Medidas de avaliação utilizadas	59
Tabela 16: <i>Ranking</i> dos algoritmos de classificação por medida de avaliação – problema 1.1, obtido via <i>RapidMiner</i> e trabalhado em <i>Microsoft Excel</i>	61
Tabela 17: <i>Ranking</i> da amplitude entre as medidas de Precisão e Sensibilidade (valores médios) – problema 1.1, obtido via <i>RapidMiner</i> e trabalhado em <i>Microsoft Excel</i>	63
Tabela 18: <i>Ranking</i> dos resultados obtidos pelos algoritmos de regras de decisão – problema 1.2, obtido via <i>RapidMiner</i> e trabalhado em <i>Microsoft Excel</i>	64
Tabela 19: <i>Ranking</i> dos algoritmos de classificação por medida de avaliação – problema 2.1, obtido via <i>RapidMiner</i> e trabalhado em <i>Microsoft Excel</i>	65
Tabela 20: <i>Ranking</i> da amplitude entre as medidas de Precisão e Sensibilidade (valores médios) – problema 2.1, obtido via <i>RapidMiner</i> e trabalhado em <i>Microsoft Excel</i>	66
Tabela 21: <i>Ranking</i> dos resultados obtidos pelos algoritmos de regras de decisão – problema 2.2, obtido via <i>RapidMiner</i> e trabalhado em <i>Microsoft Excel</i>	67
Tabela 22: <i>Ranking</i> dos resultados obtidos pelos algoritmos de regras de decisão com discretização do atributo classe – problema 2.2, obtido via <i>RapidMiner</i> e trabalhado em <i>Microsoft Excel</i>	68

1. Introdução e Motivação

Os mais recentes dados do Departamento de Assuntos Económicos e Sociais das Nações Unidas revelam que, atualmente, 54% da população mundial reside em grandes cidades, estimando-se atingir 66% no final da primeira metade do século XXI (ONU, 2014). Em Portugal, dados dos últimos censos realizados pelo Instituto Nacional de Estatística, demonstram que 72% da população portuguesa reside em áreas predominantemente urbanas, que representam apenas 18% do território nacional (INE, 2011).

A tendência de concentração populacional em grandes centros urbanos tem colocado elevadas pressões e desafios no planeamento, gestão e desenvolvimento das cidades, quer nos países desenvolvidos, quer nas economias emergentes. Estes desafios têm estimulado a reflexão num espectro alargado da sociedade, nomeadamente, junto das autoridades oficiais (governos e autoridades locais), organizações não governamentais, academia, empresas, classes de profissionais e indivíduos, em áreas tão diversas como o planeamento urbano, arquitetura, transportes, ambiente, ciências sociais, finanças, políticas públicas e, mais recentemente, a nível das próprias tecnologias de informação, por forma a colmatar os impactos nas condições económicas, sociais e ambientais das cidades. É assim imperativa a adoção de soluções inovadoras que permitam às cidades lidar com estes desafios.

Têm sido realizadas diversas abordagens, com maior ou menor nível de estruturação e âmbito de intervenção, a nível de planeamento urbano, *design*, finanças, construção, governança e gestão de serviços e infraestruturas urbanas, normalmente rotuladas de *Smart Cities* ou Cidades Inteligentes, existindo ainda múltiplas abordagens e *frameworks* para estruturação de um modelo de desenvolvimento urbano sustentável.

Estes desafios colocam-se igualmente ao setor dos resíduos, encontrando-se definidas pela União Europeia (UE) diretivas de política ambiental, transpostas para o quadro jurídico dos estados membros, com orientações e objetivos a cumprir, por forma a promover o uso sustentável de recursos e colmatar os impactos da produção de resíduos, provenientes das atividades socioeconómicas que caracterizam o quotidiano de pessoas e empresas. O uso sustentável dos recursos constitui, atualmente, uma prioridade da política europeia e nacional, traduzido num novo paradigma de desenvolvimento económico designado por “Economia Circular” em que, através da transformação dos resíduos num recurso, este novo modelo proporciona vantagens a nível de competitividade, crescimento e emprego, assim como para o ambiente. Dados da UE de 2010, revelam que são produzidos em 28 estados

membros, cerca de 2,7 mil milhões de toneladas de resíduos, o equivalente a uma capitação de 5,2 toneladas/ habitante/ano, tendo sido valorizados apenas 53%, e os restantes 47% eliminados, sobretudo por via de deposição em aterro. Em Portugal foram produzidos em 2012, 14,3 milhões de toneladas de resíduos, correspondendo a uma capitação de 1,35 toneladas/habitante/ano, tendo sido 69% encaminhados para operações de valorização e os restantes eliminados (Ministério do Ambiente, 2014).

Os resíduos urbanos (RU), resíduos provenientes de habitações (bem como outro resíduo que, pela sua natureza ou composição, seja semelhante ao resíduo proveniente de habitações), pela sua composição, origem e modelos de gestão, assumem especial relevância no contexto de gestão global de resíduos. As orientações comunitárias em matéria de gestão de resíduos urbanos encontram-se transpostas para o ordenamento jurídico nacional, no PERSU 2020 - Plano Estratégico para os Resíduos Urbanos, que estabelece a visão, os objetivos, as medidas, as metas nacionais e as metas específicas por Sistema de Gestão de RU, para o período de 2014 a 2020. Na tabela seguinte, são apresentadas as metas definidas para Portugal Continental, a atingir em 2020, e os valores reais registados no ano de 2016:

Indicador	Meta Nacional	
	Real 2016	Objetivo 2020
Preparação para reutilização e reciclagem	36%	50%
Reciclagem de resíduos embalagem	36 kg/(hab.ano)	47 kg/(hab.ano)
Deposição de RUB em Aterro	36%	35%

Nota: RUB - Resíduo Urbano Biodegradável

Tabela 1: Metas PERSU 2020 e valores reais atingidos em 2016 (Fonte: adaptado de Ministério do Ambiente, 2014 e ERSAR, 2017)

Refira-se ainda que, em fevereiro de 2018, foram aprovadas novas regras comunitárias para a gestão de resíduos, no âmbito do Pacote Economia Circular, tendo sido definidas metas mais ambiciosas, nomeadamente a nível da preparação para reutilização e reciclagem, com objetivos de 55% a alcançar em 2025, 60% em 2030 e 65% até 2035, e do desvio de RUB (Resíduos Urbanos Biodegradáveis) de Aterro, com uma meta de deposição máxima em aterro de 10% a alcançar em 2035 (Conselho Europeu, 2018).

Desta forma, é urgente a adoção de estratégias e iniciativas que promovam o incremento da separação dos resíduos pelos cidadãos, nomeadamente por via de uma aposta na implementação e monitorização de circuitos de recolha seletiva, com particular foco em circuitos de recolha porta-a-porta, que constitui um dos vetores estratégicos da estratégia nacional de gestão de resíduos consubstanciada no PERSU 2020.

2. Problema em análise

2.1. Enquadramento

A gestão de resíduos urbanos engloba todas as atividades de “recolha, o transporte, a valorização e a eliminação de resíduos, incluindo a supervisão destas operações, a manutenção dos locais de eliminação no pós-encerramento, bem como as medidas adotadas na qualidade de comerciante ou corretor” (Ministério do Ambiente, 2014). Estas atividades são geridas pelos municípios ou por entidades criadas por estes para o efeito, denominadas de SGRU – Sistemas de Gestão de Resíduos Urbanos.

Para a recolha dos resíduos, são agendados e realizados circuitos de recolha, nomeadamente, circuitos de recolha de proximidade (ecopontos), circuitos de recolha porta-a-porta residencial (PAP-R), circuitos de recolha porta-a-porta não residencial (PAP-NR), e circuitos de recolha a pedido, mediante solicitação de recolha pelo cidadão ou empresa.

Na presente tese, será abordada a atividade da recolha seletiva dos resíduos urbanos produzidos por atividades de comércio e outras (não residencial) que, pelo volume e tipologia de resíduos produzidos, são equiparáveis a resíduos urbanos, e a atividade de supervisão, nomeadamente, a avaliação das recolhas realizadas.

Nos circuitos de recolha porta-a-porta não residencial (PAP-NR), são por norma atribuídos a cada cliente aderente a este serviço, contentores descartáveis (sacos plásticos 120L) para acondicionamento de três diferentes tipologias de resíduos, identificadas por cores – Papel/ Cartão (saco azul), Embalagens (saco amarelo) e Vidro (saco verde). Caso a especificidade do negócio ou das instalações do estabelecimento do cliente assim o exija, excecionalmente são atribuídos contentores rígidos de capacidade adequada às suas necessidades. A recolha dos resíduos em circuitos de recolha PAP-NR é assegurada por viaturas de tipologia mista, sem equipamento de carga automática de contentores, destinadas a recolha multífluxo, isto é, recolhem simultaneamente as três tipologias de resíduos.

2.2. Agendamento dinâmico de circuitos PAP-NR

A execução dos circuitos de recolha PAP-NR, é realizada com base em rotas e datas pré-determinadas. As rotas estabelecem o percurso que a viatura deverá percorrer para recolha dos contentores presentes nos pontos de recolha definidos, tendo em conta o

número de pontos de recolha, total de contentores a recolher, capacidade dos contentores, tipologia e capacidade da viatura, número de recursos humanos disponíveis (motorista e ajudantes) e duração.

O agendamento destes circuitos ignora a produção de resíduos de cada estabelecimento comercial (se existem ou não resíduos para recolha, e o número de contentores colocados à recolha), e é efetuado de forma manual, dependendo da “sensibilidade” do gestor ou técnico responsável da área de gestão de resíduos dos municípios.

Propõe-se o desenvolvimento de um modelo que permita o agendamento dinâmico de circuitos de recolha seletiva porta-a-porta não residencial, por via da previsão da colocação de resíduos à recolha pelos clientes. Tendo assim em conta as previsões de entrega de resíduos por ponto de recolha, o modelo irá propor de forma dinâmica, os pontos de recolha que a viatura deverá visitar no circuito, e uma previsão do número total de contentores a recolher.

A não obrigatoriedade de paragem da viatura em todos os pontos de recolha, proporcionado pelo modelo de agendamento dinâmico a desenvolver, permitirá a obtenção de vantagens a nível de eficiências operacionais, com o consequente impacto na redução de custos da atividade de recolha e melhoria do serviço prestado ao cliente aderente à recolha seletiva PAP-NR, decorrentes de:

- i. realização de rotas de recolha otimizadas, aumentando a disponibilidade das viaturas para alocação a novos circuitos;
- ii. maior rentabilização dos ativos afetos à atividade de recolha (viaturas, equipamentos e recursos humanos);
- iii. evitar ocorrências de *overfilling* (esgotamento da capacidade de acondicionamento dos resíduos nos pontos de recolha) e respetivas consequências a nível de salubridade e necessidade de realização de recolhas urgentes.

Refira-se que, encontram-se disponíveis no mercado diversas soluções de sensorização de contentores que, monitorizando a sua taxa de enchimento, geram alertas para a sua recolha, proporcionado deste modo uma gestão dinâmica do agendamento e definição de circuitos de recolha de resíduos. Contudo, para além de onerosas, exigindo a realização de avultados investimentos para sensorização de equipamentos, estas soluções não são passíveis de aplicação em serviços de recolha PAP-NR, dada a utilização de sacos de plástico para o

acondicionamento de resíduos pelos clientes, e consequente impossibilidade de monitorização dos seus níveis de enchimento.

Com a presente tese, é proposta uma solução pouco onerosa que, com base em dados históricos reais de recolha, permite prever quer a colocação ou não de resíduos à recolha pelos clientes aderentes, quer a própria produção de resíduos, isto é, o número de contentores a recolher. Esta solução constituirá ainda um precioso auxílio no suporte às operações de angariação de novos clientes aderentes ao serviço PAP-NR, permitindo a previsão de uma rotina de recolha de um novo cliente, ultrapassando assim situações de *cold start* por inexistência de dados históricos.

3. Revisão da literatura

3.1. *Smart Cities* - Enquadramento

As mais recentes projeções de 2017 do Departamento de Assuntos Económicos e Sociais da ONU (Organização das Nações Unidas) relativas à população mundial, apontam para uma estimativa global de 7,6 mil milhões de habitantes em 2017, projetando um crescimento para 8,6 mil milhões em 2030, e para os 9,8 mil milhões de habitantes no final da primeira metade do século XXI (ONU, 2017). De acordo com a mesma organização, em 2014, 54% da população residia em áreas urbanas, estimando-se atingir os 66% em 2050 (ONU, 2014).

Esta tendência de concentração populacional decorrente da migração das populações de áreas rurais para grandes centros urbanos, em busca de melhores oportunidades, condições de vida, emprego e educação, tem colocado elevadas pressões e desafios no planeamento, gestão e desenvolvimento das cidades, quer nos países desenvolvidos, quer para as economias emergentes. Estes desafios têm estimulado a reflexão num espectro alargado da sociedade, nomeadamente, junto das autoridades oficiais (governos e autoridades locais), organizações não governamentais, instituições de ensino, empresas, classe de profissionais e indivíduos, em áreas tão diversas como por exemplo, planeamento urbano, arquitetura, transportes, ambiente, ciências sociais, finanças e políticas públicas e, mais recentemente, a nível das próprias tecnologias de informação.

Associado a este fenómeno e os decorrentes impactos nas condições económicas, sociais e ambientais das cidades, surgem desafios críticos a nível de fenómenos como a poluição, ruído, acidentes rodoviários, tráfico automóvel, produção de energia, gestão de recursos hidráulicos, gestão de resíduos, crime, ataques terroristas, fenómenos de exclusão de indivíduos e grupos étnicos, pobreza, etc., com um particular destaque para as alterações climáticas. As cidades são hoje os maiores contribuintes para o fenómeno das alterações climáticas. Cobrindo apenas cerca de 2% da superfície do globo terrestre, as cidades são responsáveis pelo consumo de 78% da energia mundial produzida, produzindo mais de 60% do total de emissões de CO₂ (UN-HABITAT, 2018).

A crescente consciencialização global para esta realidade encontra-se traduzida em variadíssimos esforços para o desenvolvimento de capacidades para a resolução destes problemas. Na Conferência das Nações Unidas para o Desenvolvimento Sustentável (Rio+20) realizada em 2012, os oito maiores bancos de desenvolvimento anunciaram a usa

intenção de investir 175 mil milhões de dólares num programa de incentivo ao desenvolvimento de sistemas de transporte sustentáveis¹. A IBM concedeu apoios em tecnologia e serviços no valor de 50 milhões de dólares, a 100 municípios em todo o mundo (ONU, 2012).

É assim imperativa a adoção de soluções inovadoras que permitam às cidades lidar com estes desafios. Têm sido realizadas diversas abordagens, com maior ou menor nível de estruturação e âmbito de intervenção, a nível de planeamento urbano, design, finanças, construção, *governance* e gestão de serviços e infraestruturas urbanas, normalmente rotuladas de *Smart Cities* ou Cidades Inteligentes.

3.2. *Smart Cities* – Definição e evolução do conceito

As abordagens aos desafios colocados por uma crescente população urbana encontram-se rotuladas em diversos conceitos, nomeadamente, *smart city*, *digital city*, *intelligent city*, *knowledge city*, *information city*, *creative city* e *urban innovation*, uns enfatizando aspetos tecnológicos, outros o desenvolvimento de capital humano e infraestruturas. No entanto, o termo *Smart City* tem registado uma utilização crescente pela sua maior amplitude conceptual, abrangendo os aspetos presentes nos diversos rótulos que têm sido utilizados para definir os variadíssimos esforços realizados para tornar as cidades mais inteligentes (Gil-Garcia, Pardo e Nam, 2015). De acordo com Bollier (1998, in Harrison e Donnelly, 2011), o conceito de *Smart Cities* poderá ter tido a sua origem no movimento *Smart Growth* do final da década de 90, uma coligação de organizações ambientais, profissionais e políticas ligados à gestão urbana, que defendiam a adoção de novas políticas de planeamento urbano. Este movimento criticava os efeitos sociais, económicos e ambientais decorrentes do modelo de desenvolvimento suburbano, com a deslocação massiva dos cidadãos, das cidades para os seus subúrbios, defendendo a adoção de um planeamento central pelos governos sobre onde e como os cidadãos vivem e trabalham, com o objetivo de forçar o retorno destes às cidades (Dilorenzo, 1999). Contudo, o conceito de *Smart Cities* tem um alcance maior que uma política de urbanização inteligente. Apesar de ambos os conceitos se encontrarem focados no bem-estar e qualidade de vida dos cidadãos, é omissa no urbanismo inteligente uma perspetiva de gestão escalável de recursos no longo prazo (Khatoun e Zeadally, 2016). A partir de 2005, o

¹ African Development Bank, Asian Development Bank, CAF–Development Bank of Latin America, European Bank for Reconstruction and Development, European Investment Bank, Inter-American Development Bank, Islamic Development Bank, World Bank.

vocabulo *Smart Growth*, passa a ser adotado por diversas empresas tecnológicas (Cisco, 2005, IBM, 2009, Siemens, 2004), na aplicação de complexas soluções informáticas para gestão de infraestruturas e serviços urbanos, nomeadamente, construção, transportes, distribuição de água e eletricidade, e segurança pública. Desde então, o conceito evoluiu para uma perceção mais lata, traduzindo quase toda e qualquer forma de solução inovadora no planeamento, desenvolvimento e gestão de cidades (Harrison e Donnelly, 2011).

Giffinger et al. (2007) define *Smart City* como uma cidade que apresenta um bom desempenho de forma prospetiva, isto é, em termos de visão de futuro, em seis características, assentes na economia (*Smart Economy*), pessoas (*Smart People*), governança (*Smart Governance*), mobilidade (*Smart Mobility*), ambiente (*Smart Environment*) e no bem-estar (*Smart Living*).

<i>Smart Economy</i> (Competitividade) - Espírito inovador; - Empreendedorismo; - Imagem de marca da economia; - Produtividade; - Flexibilidade laboral; - Internacionalização; - Capacidade de transformação.	<i>Smart People</i> (Capital Social e Humano) - Nível de qualificações; - Aprendizagem contínua; - Pluralidade social e étnica; - Flexibilidade; - Criatividade; - Mentalidade "aberta"; - Cidadania ativa.	<i>Smart Governance</i> (Participação) - Participação na tomada de decisão; - Serviços públicos e sociais; - Governação transparente; - Estratégias políticas e perspetivas.
<i>Smart Mobility</i> (Transportes e TIC) - Acessibilidade local; - Acessibilidade nacional e internacional; - Disponibilidade de infraestruturas tecnológicas de informação e comunicação (TIC); - Sistema de transportes seguro, inovador e sustentável.	<i>Smart Environment</i> (Recursos naturais) - Atratividade de condições naturais; - Poluição; - Proteção ambiental; - Gestão de recursos sustentável.	<i>Smart Living</i> (Qualidade de vida) - Infraestruturas culturais; - Condições de saúde; - Segurança individual; - Qualidade habitacional; - Infraestruturas de educação; - Atratividade turística; - Coesão social.

Figura 1: Características e fatores de uma *Smart City* (Fonte: adaptado de Giffinger et al., 2007)

No âmbito do projeto CITYkey da Comissão Europeia, desenvolvido ao abrigo do programa Horizonte 2020, para a definição de um *framework* de monitorização e avaliação do desempenho de soluções *Smart City*, são definidos os próprios conceitos de *Smart City* e *Smart Project*, numa visão igualmente holística a nível de âmbitos de intervenção. Deste modo, “uma cidade inteligente é uma cidade que mobiliza e usa os recursos disponíveis de forma eficiente (incluindo, mas não limitado ao capital social e cultural, capital financeiro, recursos naturais, informações e tecnologia)”, integrando no seu propósito os eixos da melhoria da qualidade de vida dos cidadãos, a utilização eficiente e sustentável dos recursos, o desenvolvimento de uma economia verde e *governance* (Bosch et al., 2017). De igual modo, projetos com um

impacto no desenvolvimento das capacidades de intervenção das cidades nestes eixos, que ativamente mobilizem cidadãos e *stakeholders*, adotam abordagens inovadoras e que integrem múltiplos sectores, são considerados *Smart Projects* (Bosch et al., 2017). Já o Parlamento Europeu, introduz, em 2014, um conceito que coloca em destaque o papel das tecnologias de informação e comunicação (TIC), definindo *Smart City* como uma cidade que procura lidar com as questões públicas por via de soluções baseadas em TIC, numa base de cooperação intermunicipal e *multi-stakeholder* (Manville et al., 2014). Outra definição de *Smart City* que coloca em destaque o papel das TIC é avançada por Anthony Townsend, como “lugares em que a tecnologia da informação é combinada com infraestrutura, arquitetura (...), para lidar com problemas sociais, económicos e ambientais” (2014, in Khatoun, 2016). A *Japan Smart Community Alliance* (JSCA) estabelece ainda que *Smart City* é “uma comunidade onde várias tecnologias de próxima geração e sistemas sociais avançados são efetivamente integrados e utilizados, incluindo o uso eficiente da energia, a utilização de calor e fontes de energia não utilizadas, a melhoria dos sistemas de transporte locais e a transformação da vida quotidiana dos cidadãos” (JSCA, 2018).

Anthopoulos (2017), aborda de forma bastante abrangente o conceito de *Smart City*, destacando o papel desempenhado pelas TIC e pela inovação, na forma como as cidades abordam de forma sustentável (a nível económico, social e ambiental), os desafios colocados em seis dimensões (pessoas, economia, *governance*, mobilidade, ambiente e modo de vida), idênticas às referenciadas por Giffinger et al. (2007). Anthopoulos refere ainda que, estas abordagens serão tão diferentes, quanto as diferenças a nível de *performance* das TIC e inovação, bem como das prioridades das diferentes cidades, produzindo assim abordagens distintas na forma como abordam os seus desafios e, deste modo, produzindo diferentes formas de *Smart City*.

Inúmeras definições de *Smart City* têm sido propostas ao longo dos últimos anos, com maior ou menor abrangência e âmbitos de intervenção, e com diferentes focos, desde o envolvimento dos cidadãos, adoção de métodos colaborativos, até ao papel desempenhado pelas TIC. No entanto, em todas estas definições existe uma tónica comum, de que soluções inovadoras, eficiência de processos, transformação digital e tecnologia, são os grandes facilitadores do desenvolvimento de um modelo urbano mais sustentável, com particular destaque para o papel central desempenhado pelas TIC, na capacitação das cidades, para satisfação das necessidades dos seus cidadãos e otimização dos seus recursos.

3.3. *Smart Cities* – Modelos de desenvolvimento e desafios

Harrison (2011), defende que as motivações das cidades investirem em abordagens *Smart City* são essencialmente de natureza económica e que, no decorrer da crise económico-financeira de 2008-2009, as cidades começaram a perceber que se encontram em competição com outras cidades, em formatos que não haviam até então experienciado. Potenciada pela internet e redes de distribuição internacionais, a natureza desta competição transcende as fronteiras nacionais, encontrando-se as cidades em direta competição com outras cidades do outro lado do globo, não competindo apenas pela atração de investimento e criação de emprego, mas também pela atração de jovens das gerações Y e Z, os arquitetos de um novo e fortalecido desenvolvimento económico. O interesse das cidades num modelo de desenvolvimento urbano sustentável, passa então a estar ancorado numa gestão de imagem, da marca da cidade. Deste modo, existe atualmente um grande foco dos decisores políticos no *branding* das cidades para atração de populações jovens e criativas para as suas cidades, com a publicitação e desenvolvimento de diversas iniciativas *smart* algo difusas e com reduzida perceção da sua causa-efeito (rede pública de wi-fi, ciclovias, postos de abastecimento elétricos de viaturas, digitalização interfaces das instituições públicas com os cidadãos, etc.). Tal, traduz as abordagens *ad hoc* para o desenvolvimento de *Smart Cities*, na medida em que existe uma perceção que produzem o bem, mas existe ainda um reduzido entendimento do porquê, à semelhança do verificado na medicina pré-científica.

É assim necessária uma arquitetura de modelos e abordagens *Smart City* com forte fundamentação teórica. Com inspiração na biologia, surgem estudos que demonstram capacidade de explicar a escalabilidade de muitos fenómenos e aspetos das cidades (West, 1999, in Harrison, 2011), a escalabilidade de sistemas urbanos regidas por estruturas de rede, numa analogia a sistemas biológicos, nomeadamente o sistema nervoso e sistema circulatório (Batty, 2008, in Harrison, 2011). Outra linha de estudo, numa perspetiva mais macro, considera as cidades como sistemas complexos, auto organizativos (Allen, 1997 e Portugali, 2000, in Harrison, 2011), proporcionando uma compreensão do crescimento, operação, declínio e evolução das cidades. Esta noção de sistema complexo encontra-se também patente no trabalho desenvolvido por Forrester no seu livro "*Urban Dynamics*", em que as interações entre sistemas abstratos, como por exemplo "tráfego", são decompostas em complexos modelos matemáticos (Forrester, 1971, in Harrison, 2011). Harrison (2011), refere ainda que, o bom funcionamento de uma cidade requer um apropriado nível de

complexidade, quer nas suas infraestruturas físicas (estradas, edifícios e comunicações), quer a nível económico e cultural.

A complexidade na gestão das cidades é igualmente abordada por Visnjic (Visnjic et al., 2016), referenciando as cidades como o “ecossistema dos ecossistemas”, estruturadas e geridas dinamicamente como empresas, num ambiente de constante mudança, onde inputs de organizações especializadas são coordenados e integrados num serviço final, e como plataformas de mercado, em que os líderes da cidade intermedeiam as interações entre os diferentes prestadores de serviço e os cidadãos.

Harrison (2011), define sistemas urbanos como entidades elementares ou entidades complexas compostas por entidades mais simples e propõe um modelo – *Urban System Model*, para estruturar e classificar os diferentes tipos de informações contidas ou que fluem nessas redes. Este modelo propõe uma representação destes sistemas urbanos, num grupo de cinco camadas, contendo diferentes tipos de informação relacionados.



Figura 2: *Urban Information Model* de Harrison (Fonte: adaptado de Harrison e Donnelly, 2011)

Não existe no modelo um sentido de hierarquização das camadas e o próprio agrupamento e número de camadas presentes em cada grupo é de alguma forma arbitrária. Importante, é o entendimento da natureza multicamada e multigrupo que o conceito de serviço assume neste modelo. Os serviços constituem operações com que os cidadãos interagem na cidade, e que podem ser despoletados pelos próprios cidadãos ou por outros serviços. A título de exemplo, o serviço de recolha de resíduos urbanos prestado aos cidadãos, e que encontrar-se-á representado no grupo dos serviços, encontra-se igualmente identificado em outros grupos e camadas, na medida em que as viaturas de recolha circulam

em estradas (grupo das infraestruturas); na construção das estradas são ocupadas parcelas de terreno (grupo das infraestruturas) e impactam na sua topografia (grupo meio ambiente); o combustível consumido pelas viaturas é originário de recursos de jazidas de petróleo (grupo meio-ambiente), que é extraído e refinado por refinarias (grupo serviços) e transportado em pipelines (grupo infraestruturas); a disponibilidade dos cidadãos em separar os seus resíduos para reciclagem (grupo sistemas sociais – camada pessoas); a implementação de políticas governamentais e instrumentos económicos de incentivo à reciclagem (grupo sistemas sociais – camada políticas), etc.

O acesso em tempo real à informação ao nível individual dos cidadãos, constitui um dos fundamentos basilares nas abordagens *Smart City*, “tornando visível, o invisível” (Harrison, 2011). A atual proliferação e utilização de sistemas de informação nos ambientes urbanos, potenciada pelo IoT (Internet of Things), gera uma diversidade de informação sobre os cidadãos, a nível de volume e formato, que permite a análise e deteção de padrões de comportamento, quer a nível do indivíduo, quer de uma forma agregada. A capacidade das cidades colherem os benefícios desta informação encontrar-se-á traduzida num bem sucedido desenvolvimento de um modelo urbano sustentável, sem esquecer, contudo, as questões inerentes à privacidade e proteção de dados pessoais.

Adicionalmente à privacidade dos dados e à proteção dos mesmos contra ciberataques, constituem ainda desafios à implementação e gestão de modelos urbanos sustentáveis *Smart City*, o grau de envolvimento dos cidadãos, o custo e acesso a fontes de energia, e o elevado volume de investimento exigível (Khatoun and Zeadally, 2016). O investimento neste tipo de projetos inovadores é percecionado pelos investidores como de elevado risco, uma vez que se encontram associados a elevados períodos de retorno, elevados montantes de investimento requeridos, e a uma limitada capacidade de financiamento do setor público. Contudo, tem-se verificado um crescente investimento neste tipo de projetos, quer por entidades públicas (governamentais e municipais), quer por entidades privadas, estimando-se atingir um volume de investimento de 20 mil milhões € em 2020 (Navigant Research, in Khatoun e Zeadally, 2016).

O sucesso de projetos *Smart City*, encontra-se intrinsecamente ligado a uma dimensão social, nomeadamente, do grau de conexão e participação dos cidadãos nas mais diversas plataformas de comunicação (redes sociais, *smartphones*, portais municipais), interagindo com diversos *smart services*, quer em casa, quer em locais públicos. Questões como a privacidade e

proteção contra ciberataques desempenham assim um papel fundamental na implementação de qualquer estratégia *Smart City* (Khatoun e Zeadally, 2016; Yigitcanlar, 2015).

Yigitcanlar (2015), ao abordar os desafios tecnológicos e técnicos das *Smart Cities*, destaca desafios a nível de infraestruturas e interfaces *cloud*, escalabilidade de redes, interoperabilidade entre sistemas. O autor evoca igualmente desafios a nível económico, decorrentes do elevado volume de investimento requerido por este tipo de projetos, desafios a nível da sociedade, pelas alterações que este novo paradigma preconiza a nível do modo de vida dos cidadãos, acesso e grau de maturação tecnológica dos cidadãos, desafios a nível ambiental (gestão de recursos hídricos, gestão de resíduos), desafios a nível de *governance* (transparência e participação e colaboração dos cidadãos na tomada de decisão) e, por último, desafios decorrentes da própria implementação alargada de um modelo de *Smart City*, nomeadamente, a nível da convergência das novas tecnologias com as existentes infraestruturas urbanas físicas (estradas, fornecimento de água e energia elétrica, esgotos, sistemas de segurança, etc.).

3.4. Uma Gestão de Resíduos Urbanos *Smart*

Os desafios de estruturação de um modelo de desenvolvimento urbano sustentável, colocam-se igualmente ao nível da gestão dos resíduos urbanos (RU), produzidos por uma crescente população urbana. Este fator, aliado a uma crescente consciencialização dos cidadãos para as questões ambientais, coloca uma pressão acrescida aos gestores e decisores políticos na implementação de medidas sustentáveis na recolha e deposição de resíduos (Poser e Awad, 2006, Xue et al., 2015, in Akhtar et al., 2017).

A extensão das atividades e soluções adotadas para a gestão dos resíduos, dependerá das estratégias definidas pelos municípios e/ ou sistemas de gestão de resíduos urbanos. O processo de gestão de resíduos urbanos envolve diversas atividades, desde a produção do resíduo, a sua separação e deposição pelos cidadãos em contentores, recolha e transporte, separação multimaterial para recuperação de materiais para reciclagem, valorização orgânica de resíduos urbanos biodegradáveis (RUB) (com produção de composta orgânico para aplicação em agricultura ou digestão anaeróbia de RUB com aproveitamento para produção de energia elétrica), até ao seu destino final, com deposição em aterro ou para incineração, onde é passível de gerar a produção de energia elétrica.

As atividades ligadas à recolha e transporte dos resíduos até ao seu destino, por via do tipo e número de meios empregues (viaturas, pessoal, necessidade de percorrer um elevado número de quilómetros em rotas pré-determinadas para paragem em diversos pontos de recolha, elevados consumos de combustível das viaturas, etc.), representam uma fatia importante dos orçamentos municipais, podendo ascender a mais de 70% do seu orçamento total, sendo grande parte destinado a combustível (Tavares et al., 2009), com repercussão nas taxas cobradas aos cidadãos a título de gestão de resíduos.

O uso sustentável dos recursos constitui, atualmente, uma prioridade da política europeia e nacional, traduzido num novo paradigma de desenvolvimento económico designado por “Economia Circular” em que, através da transformação dos resíduos num recurso, este novo modelo traz vantagens a nível de competitividade, crescimento e emprego, assim como para o ambiente.

Num recente inquérito global promovido pelo *International Solid Waste Association* (ISWA), junto de diversos atores da indústria, sobre os impactos da 4ª Revolução Industrial na gestão de resíduos, demonstram que o setor já se encontra consciente das mudanças em curso, estando convicto que, as novas tecnologias emergentes, irão criar as soluções necessárias para lidar com os desafios da gestão de resíduos, desde o eco design, à prevenção da produção de resíduos e economia circular dos plásticos. Contudo, a profundidade e velocidade desta revolução, encontra-se subestimada pelos agentes do sector. É ainda destacada a urgência deste processo de mudança, prevendo que, num período de 15 anos, organizações que não abracem este novo modelo de desenvolvimento, se encontrarão obsoletas (ISWA, 2017). De acordo com os resultados deste inquérito, os maiores impactos no sector dos resíduos serão produzidos pela introdução de novos materiais, sensores avançados e IoT, sendo ainda expectável que o *social media* e aplicações telemóveis produzam maiores impactos que a robotização, inteligência artificial e condução autónoma. Para 2030, os atores da indústria consideram que já será possível a total robotização da triagem de resíduos e de centrais de reciclagem, e *chatbots* como plataforma de comunicação com os cidadãos em matérias de sensibilização ambiental, como a prevenção da produção de resíduos e a reciclagem. Antes de 2030, será atingível o *eco design* de produtos de consumo, economia circular para os plásticos, reduções significativas de emissões CO₂ produzidas pelo setor. O inquérito demonstra ainda que as áreas que as organizações do setor deverão investir mais serão *mobile apps*, novas soluções de sensorização, *social media* e *Big Data*.

Encontram-se atualmente em curso, diversas iniciativas e projetos na área de *smart waste management*, estando ainda a ser comercializadas no mercado diversas plataformas informáticas, nomeadamente a nível da gestão da recolha de resíduos urbanos, com o objetivo de aumentar a eficiência da recolha de resíduos, redução de custos e de emissões de CO₂. Estas plataformas centram-se na monitorização dos níveis de enchimento dos contentores (ecopontos, compactadores, etc.) por via de diferentes tecnologias de sensorização e monitorização, e com base nesta informação, otimizar o agendamento e a execução de circuitos e rotas de recolha de resíduos. O âmbito de intervenção destas soluções varia desde a disponibilização de diferentes soluções de sensorização *wireless* (ultrassónicos, geolocalização e monitorização de níveis de temperatura e movimento/ inclinação), otimização de rotas, plataformas de monitorização dos circuitos de recolha (*war view room of operations*), soluções de acompanhamento das equipas de recolha na execução dos circuitos, soluções de previsão de níveis de enchimento, gestão de ocorrências e manutenção de equipamentos, monitorização da participação dos cidadãos na separação dos resíduos, até propostas de digitalização de toda a cadeia de atividades do processo, desde o agendamento e/ ou pedido de recolha de resíduos, até à disponibilização de informação e faturação aos cidadãos e clientes. Contudo, todas estas plataformas, carecem de um *framework* mais abrangente e estruturado, de suporte à tomada de decisão, quer a nível da natureza dos indicadores, integrando a informação operacional, com informação económica e ambiental, quer a nível da consideração do papel desempenhado pelos diferentes atores a nível de gestão da recolha de resíduos, desde o operacional do terreno, até ao nível da administração.

Existem já em curso experiências de utilização deste tipo de soluções informáticas na gestão de recolha de seletivas de resíduos urbanos, referindo-se a título de exemplo as cidades de Bordéus, França (plataforma *Syren*), Groningen, Holanda (plataforma *Mavis*), e em Portugal, as cidades da Maia (plataforma *Compta*) e Cascais, Portugal (plataforma *Smart Bin*).

3.5. *Data Mining* na otimização de circuitos de recolha seletiva de resíduos

Existe diversa literatura na área de otimização de circuitos de recolha seletiva de resíduos urbanos, nomeadamente a nível de definição de rotas, tendo em conta restrições a nível de diferentes tipologias e capacidades de contentores e viaturas, dimensão dos circuitos de recolha, locais de deposição dos resíduos (aterro, centros de transferência, centrais de

triagem, etc.), entre outras, com recurso a metodologias do tipo *Arc Routing Problems* (ARP) e *Capacity Arc Routing Problems* (CARP) (Maria Rodrigues e Ferreira, 2015).

Este tipo de problema de otimização, designado em diversos estudos como *Vehicle Routing Problem* (VRP) (Bautista et al., 2008, in Akhtar et al., 2017) ou como *Capacitated Vehicle Routing Problem* (CVRP), quando introduzidas restrições de capacidade (Dantzig e Ramser, 1959, in Akhtar et al., 2017), é computacionalmente bastante exigente, impossibilitando a adoção de métodos exatos de programação linear, (Kulcar, 1996, in Akhtar et al., 2017), programação inteira mista (Tung and Pinnoi, 2000, Badran and El-Haggar, 2006, Agha, 2006, in Akhtar et al., 2017), em realidades de sistemas de gestão de resíduos de maior dimensão, isto é, em problemas de escala alargada.

Deste modo, foram adotadas diversas abordagens por via de heurísticas, com particular destaque para *Nearest Neighbourhood Search Algorithm* (Faccio et al., 2011, in Akhtar et al., 2017) e *Greedy Algorithm* (Bautista e Pereira, 2006, Sahoo et al., 2005, in Akhtar et al., 2017). No entanto, estas técnicas apresentam ainda tempos de computação elevados e baixa precisão para os problemas de recolha de resíduos (Viotti et al., 2003, in Akhtar et al., 2017). De acordo com o mesmo autor, a abordagem por via de meta heurísticas têm-se revelado, recentemente, bastante populares, destacando a utilização de técnicas como *Ant Colony Optimization* (ACO) (Islam e Rahman, 2012, Liu e He, 2012, in Akhtar et al., 2017), *Genetic Algorithms* (GAs) (Karadimas et al., 2007, Viotti et al., 2003, in Akhtar et al., 2017), e *Particle Swarm Optimization* (PSO) (Son, 2014, Kuo et al., 2012, in Akhtar et al., 2017).

Contudo, em todas estas abordagens é omissa a avaliação em tempo real do nível de enchimento dos contentores, e que Akhtar et al. (2017) aborda, desenvolvendo uma nova meta heurística que integra o algoritmo *Backtracking Search Optimization* (BSA), com os dados produzidos por sensores de medição do nível de enchimento dos contentores, num problema de *Capacitated Vehicle Routing Problem* (CVRP). Outra abordagem a este nível, é proposta por Vu e Kaddoum (2017) que, através de técnicas de regressão, classificação e teoria dos grafos, assegura uma previsão dos níveis de enchimento dos contentores, classificação da localização dos contentores e, por via da aplicação de técnica de *Genetic Algorithm*, procede à otimização de rotas.

4. Metodologia aplicada

4.1. Formulação e metodologia aplicada

As equipas de recolha de resíduos executam circuitos de recolha PAP-NR com base em rotas pré-estabelecidas. Estas rotas estabelecem o percurso a percorrer pela viatura para a recolha dos contentores presentes nos pontos de recolha definidos, tendo em conta o número de pontos de recolha, total de contentores a recolher, capacidade dos contentores, tipologia e capacidade da viatura, número de recursos humanos disponíveis (motorista e ajudantes), e duração.

O agendamento destes circuitos ignora a produção de resíduos de cada estabelecimento comercial (se o cliente possui ou não resíduos para recolha e respetivo número de contentores) e é efetuado de forma manual, dependendo da “sensibilidade” do gestor ou técnico responsável da área de gestão de resíduos dos municípios.

Deste modo, propõe-se um modelo para previsão da colocação de resíduos à recolha pelos clientes, que possibilite o agendamento dinâmico e flexível de circuitos de recolha, adequados e dimensionados às necessidades do município e/ou empresa responsável pela recolha, sem prejuízo do cumprimento dos níveis de serviço aos clientes. O modelo terá como *outputs*:

1. Previsão da colocação de resíduos à recolha pelos clientes, propondo assim, de forma dinâmica, os pontos de recolha a considerar na definição da rota que a viatura deverá percorrer;
2. Previsão do número total de contentores apresentados à recolha a ser considerado no dimensionamento e definição da rota;
3. Identificação de regras que caracterizam a produção de resíduos pelos clientes.

Na abordagem ao desenvolvimento deste modelo será adotada a metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*), geralmente aceite e consensual entre profissionais para tarefas de *data mining*, ilustrada na figura seguinte:

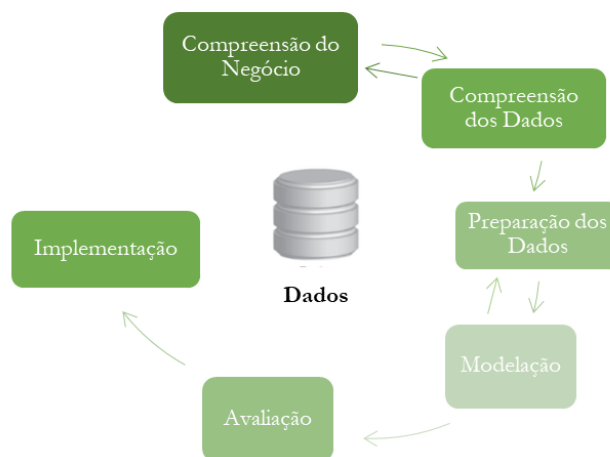


Figura 3: Etapas da metodologia CRISP-DM

Numa fase inicial de preparação, será efetuado um enquadramento referente à compreensão do negócio e o tipo de problema a abordar para uma melhor definição dos requisitos e objetivos no desenvolvimento do modelo. Serão ainda abordados o processo de recolha, familiarização e análise primária de dados com vista à identificação de problemas de qualidade, existência de relações e conjuntos interessantes – compreensão dos dados, bem como a sua preparação para utilização na construção do modelo preditivo, através da transformação de dados, resolução de questões de qualidade e seleção dos atributos relevantes - preparação dos dados. Numa fase intermédia será abordada a construção do modelo, através da utilização de diferentes algoritmos e afinação dos seus parâmetros – modelação. Por último, será efetuada a avaliação do modelo desenvolvido e respetiva implementação.

4.2. Compreensão do negócio

A LIPOR – Serviço Intermunicipalizado de Gestão de Resíduos do Grande Porto é a entidade responsável pela gestão, valorização e tratamento dos Resíduos Urbanos produzidos pelos oito municípios que a integram: Espinho, Gondomar, Maia, Matosinhos, Porto, Póvoa de Varzim, Valongo e Vila do Conde, tratando anualmente, cerca de 500 mil toneladas de resíduos urbanos (RU) produzidos por cerca de 1 milhão de habitantes.

Sustentada nos modernos conceitos de gestão de RU, que preconizam a adoção de sistemas integrados e a minimização da deposição de resíduos em aterro, a LIPOR

desenvolveu uma estratégia integrada de valorização, tratamento e confinamento dos RU, baseada em três componentes principais: Valorização Multimaterial, Valorização Orgânica e Valorização Energética, completadas por um Aterro Sanitário, para receção dos rejeitados e de resíduos previamente preparados.

A recolha dos resíduos urbanos é assegurada diretamente pelas Câmaras Municipais associadas da LIPOR, ou por entidades por elas subcontratadas. Para a recolha dos resíduos são agendados e realizados circuitos de recolha de acordo com o ponto de recolha, nomeadamente, circuitos de recolha de proximidade (ecopontos), circuitos de recolha porta-a-porta residencial (PAP-R), circuitos de recolha porta-a-porta não residencial (PAP-NR) e circuitos de recolha a pedido, mediante solicitação de recolha pelo cidadão ou empresa.

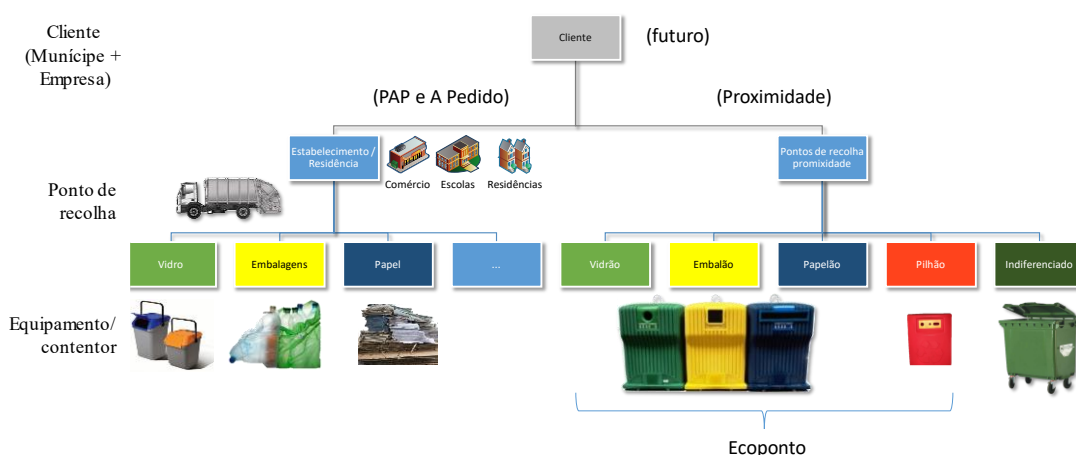


Figura 4: Tipologia de circuitos de recolha de resíduos urbanos

Os resíduos recolhidos são encaminhados para valorização nas diferentes unidades de tratamento da LIPOR, nomeadamente:

- Valorização Multimaterial: triagem e encaminhamento para a indústria recicladora de resíduos recicláveis (Vidro, Papel/ Cartão e Embalagens de plástico e metálicas);
- Valorização Orgânica: compostagem de resíduos urbanos biodegradáveis (RUB) recolhidos seletivamente, para produção de um corretivo orgânico de elevada qualidade comercializado sob a marca Nutrimais®;
- Valorização Energética: produção de energia elétrica a partir da incineração de resíduos indiferenciados, não separados pelos cidadãos e empresas;
- Confinamento Técnico: deposição em aterro sanitário de resíduos não passíveis das formas de valorização descritas anteriormente, e que constitui a última solução de tratamento no sistema integrado de gestão de resíduos da LIPOR.

Destino Final	2014	2015	2016	2017	Δ% 17/16	% Peso 2017
Reciclagem Multimaterial	44 754	45 296	46 791	47 304	1,1%	9,4%
Valorização Orgânica	47 169	48 649	47 944	50 911	6,2%	10,1%
Valorização Energética	385 972	402 786	402 058	395 643	-1,6%	78,3%
Confinamento Técnico	21 497	24	4 759	11 649	144,8%	2,3%
Total	499 392	496 756	501 552	505 507	0,8%	100,0%

Tabela 2: Encaminhamento de resíduos por destino final (Fonte: adaptado de Relatório e Contas LIPOR 2017)

A LIPOR, em articulação com as Câmaras Municipais associadas, assegura ainda, de forma permanente e dinâmica, a monitorização das diferentes tipologias de circuitos de recolha de resíduos.

Na presente tese, será especificamente abordada a atividade da recolha seletiva dos resíduos urbanos produzidas por atividades de comércio e outras (não residencial), nomeadamente, o agendamento de circuitos de recolha porta-a-porta não residencial (PAP-NR), sendo ainda realizada uma análise e avaliação das recolhas realizadas com proposta de indicadores.

Os circuitos PAP-NR foram responsáveis pela recolha de um total de 5.170 toneladas de resíduos recicláveis em 2017, representado cerca de 11% dos quantitativos totais encaminhados para valorização multimaterial, e apresenta uma importância estratégica crescente para o cumprimento das metas nacionais e comunitárias em matéria de gestão de resíduos urbanos.

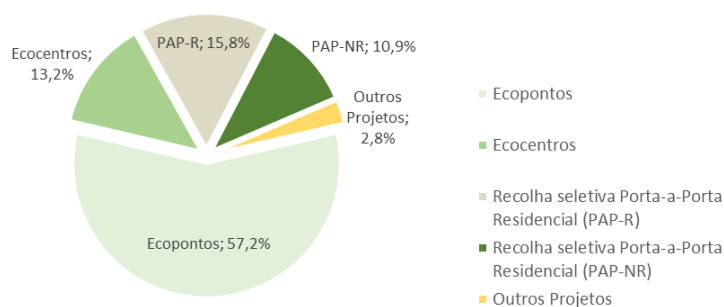


Figura 5: Entradas de recicláveis por origem (Fonte: adaptado de Relatório e Contas LIPOR 2017)

4.3. Compreensão dos dados

4.3.1 Recolha dos dados

O registo dos contentores recolhidos em cada cliente é assegurado manualmente pelos operadores de recolha, numa listagem extraída do *software* desenvolvido internamente pela LIPOR, para monitorização e gestão de circuitos de recolha seletiva de resíduos, e que acompanha a equipa de recolha na realização do circuito. A listagem possui identificação do município, do circuito, data e turno, o tipo de recolha, a sequência de recolha do ponto de recolha na rota ou circuito estabelecido, identificação do cliente (nº LIPOR, nome e morada), e o número de contentores recolhidos por fluxo de resíduo (Papel/ Cartão, Embalagens e Vidro).

Após a conclusão do circuito, estes dados são introduzidos manualmente no *software* de monitorização e gestão de circuitos de recolha seletiva de resíduos. Refira-se que, à data da elaboração da presente tese, encontravam-se em curso diversos projetos pilotos para automatização da recolha de informação pelos operadores durante a operação de recolha, bem como o desenvolvimento de ligação do *software* ao sistema de pesagens da LIPOR, para associação do registo de pesagem da viatura à entrada das instalações, por forma a aferir os quantitativos totais de resíduos recolhidos por circuito e fluxo.

Para abordagem ao desenvolvimento de um agendamento dinâmico de circuitos, foi utilizada a informação de um circuito de recolha PAP-NR, de um município associado da LIPOR, que apresentava o maior volume de dados disponível, nomeadamente o histórico completo de um ano de atividade, de janeiro a dezembro de 2017, num total de 258 dias, que se traduzem em 258 circuitos realizados. Este *data set* contém informação relativa a um total de 26.177 exemplos e 9 atributos:

- Data do circuito – data de execução do circuito estabelecido;
- Nº sequencial – sequência do ponto de recolha no circuito de recolha;
- Cliente – nome do cliente;
- Morada – morada do cliente, correspondente ao ponto de recolha;
- Atividade – descrição do tipo de atividade comercial exercida pelo cliente;
- Geolocalização – latitude e longitude da morada do estabelecimento do cliente;
- Papel – número de contentores recolhidos do fluxo de Papel/ Cartão;

- Embalagens - número de contentores recolhidos do fluxo de Embalagens de plástico e metal;
- Vidro - número de contentores recolhidos do fluxo de Vidro.

Para análise do *data set* e desenvolvimento do modelo pretendido, para previsão da colocação e produção de resíduos à recolha pelos clientes, foram adicionados novos atributos, descritos em maior detalhe no capítulo “4.4 Preparação dos dados”, nomeadamente:

- a) Horário de funcionamento do estabelecimento (número de horas);
- b) Clima, referente às condições meteorológicas verificadas no dia anterior à recolha;
- c) Feriado_Festividades_Fds, identificando a ocorrência da recolha após fim-de-semana, feriados e outras datas festivas;
- d) Centralidade, que define o nível de centralidade do estabelecimento do cliente com base na sua geolocalização, numa escala de 1 a 3, correspondendo 1 ao nível máximo de centralidade e 3, ao nível mínimo de centralidade.

4.3.2 Análise do número de contentores recolhidos

Numa primeira análise dos dados, verificou-se a existência de um total de 141 clientes aderentes ao projeto PAP-NR em 2017, sendo a Restauração a área de negócio predominante, seguida do Comércio, com um peso de 58% e 33%, respetivamente:

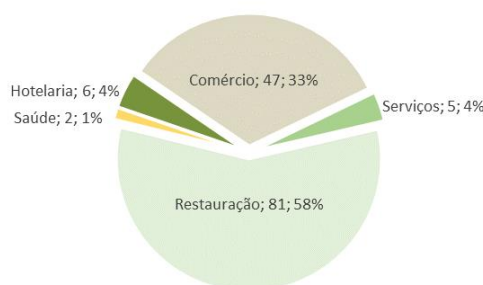


Figura 6: N.º de clientes e peso % por área de negócio, obtido via *Microsoft Excel*

O projeto PAP-NR arrancou em janeiro de 2017, com um total de 115 aderentes, tendo-se verificado um número máximo de 140 aderentes em junho. No mês de junho verificaram-se alterações na base de clientes do PAP-NR, registando-se saídas e entradas de novos aderentes. A partir do mês de julho e até ao final do ano, o número de aderentes

estabilizou num total de 131 clientes. Os maiores crescimentos da base de clientes registaram-se nas áreas de negócio de Comércio e Restauração. Globalmente, durante o ano de 2017, o serviço de recolha PAP-NR abrangeu um total de 141 clientes.

Área de Negócio	janeiro	fevereiro	março	abril	maio	junho	julho	agosto	setembro	outubro	novembro	dezembro	Global Ano
Comércio	38	38	38	38	39	46	43	43	43	43	43	43	47
Hotelaria	4	4	4	4	4	6	6	6	6	6	6	6	6
Restauração	69	69	69	69	69	81	75	75	75	75	75	75	81
Saúde	0	0	0	0	0	2	2	2	2	2	2	2	2
Serviços	4	4	4	4	4	5	5	5	5	5	5	5	5
Total Geral	115	115	115	115	116	140	131	131	131	131	131	131	141

Tabela 3: Evolução do n.º de clientes aderentes ao PAP-NR, obtido via *Microsoft Excel*

Analisando o número de contentores recolhidos, verifica-se que durante o ano de 2017 foram recolhidos um total de 52.364 contentores, apresentando o material Papel/ Cartão o maior número de recolhas, com um peso de 44,5% do total de recolhas, seguido dos materiais Embalagens e Vidro, com um peso de 31% e 24,5%, respetivamente.

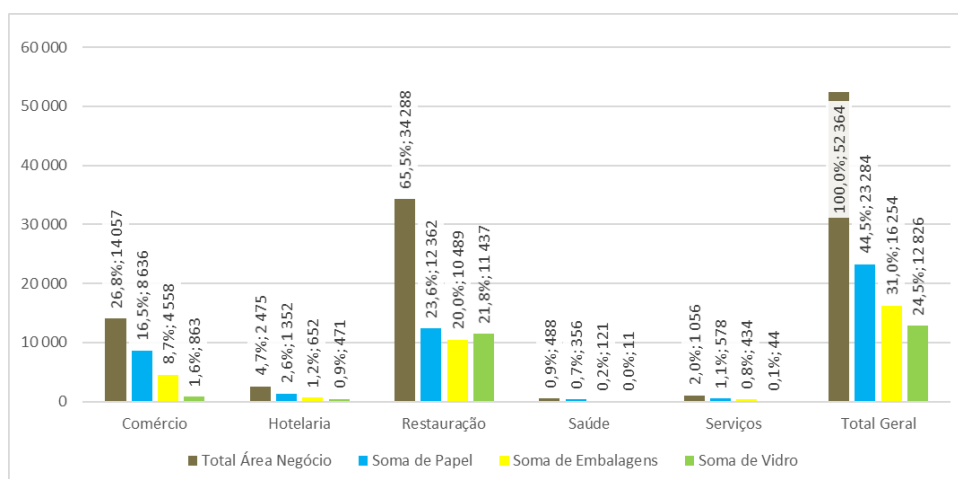


Figura 7: Distribuição do n.º contentores recolhidos e peso % por área negócio e material, obtido via *Microsoft Excel*

Transversalmente a todas as áreas de negócio, o material Papel/ Cartão apresenta o maior número de contentores recolhidos. Na área de negócio da Restauração, o material Vidro apresenta o segundo maior número de contentores recolhidos, enquanto nas restantes áreas de negócio é o material com o menor número de contentores recolhidos. As áreas de negócio Restauração e Comércio são as mais representativas no número de contentores recolhidos durante o ano de 2017, contabilizando 92,3% do total das recolhas. A área de negócio da Restauração é responsável pelo maior número de contentores recolhidos com um peso de 65,5% do total de recolhas realizadas, seguida do Comércio com um peso de 26,8%, Hotelaria com 4,7%, Serviços com 2% e, por último, a Saúde com um peso de 0,9%.

Os circuitos de recolha são realizados diariamente, de segunda a sexta-feira, com exceção de alguns feriados e outras datas festivas. Na figura seguinte, é apresentada a distribuição do número de contentores recolhidos por tipologia de material e dia da semana, sendo possível verificar que sexta-feira é o dia com maior número de recolhas, seguido da segunda-feira. Analisando por área de negócio, na Restauração, Saúde e Comércio, o maior número de recolhas verifica-se à sexta-feira, enquanto nas áreas de Hotelaria e Serviços, ocorre à quinta-feira.

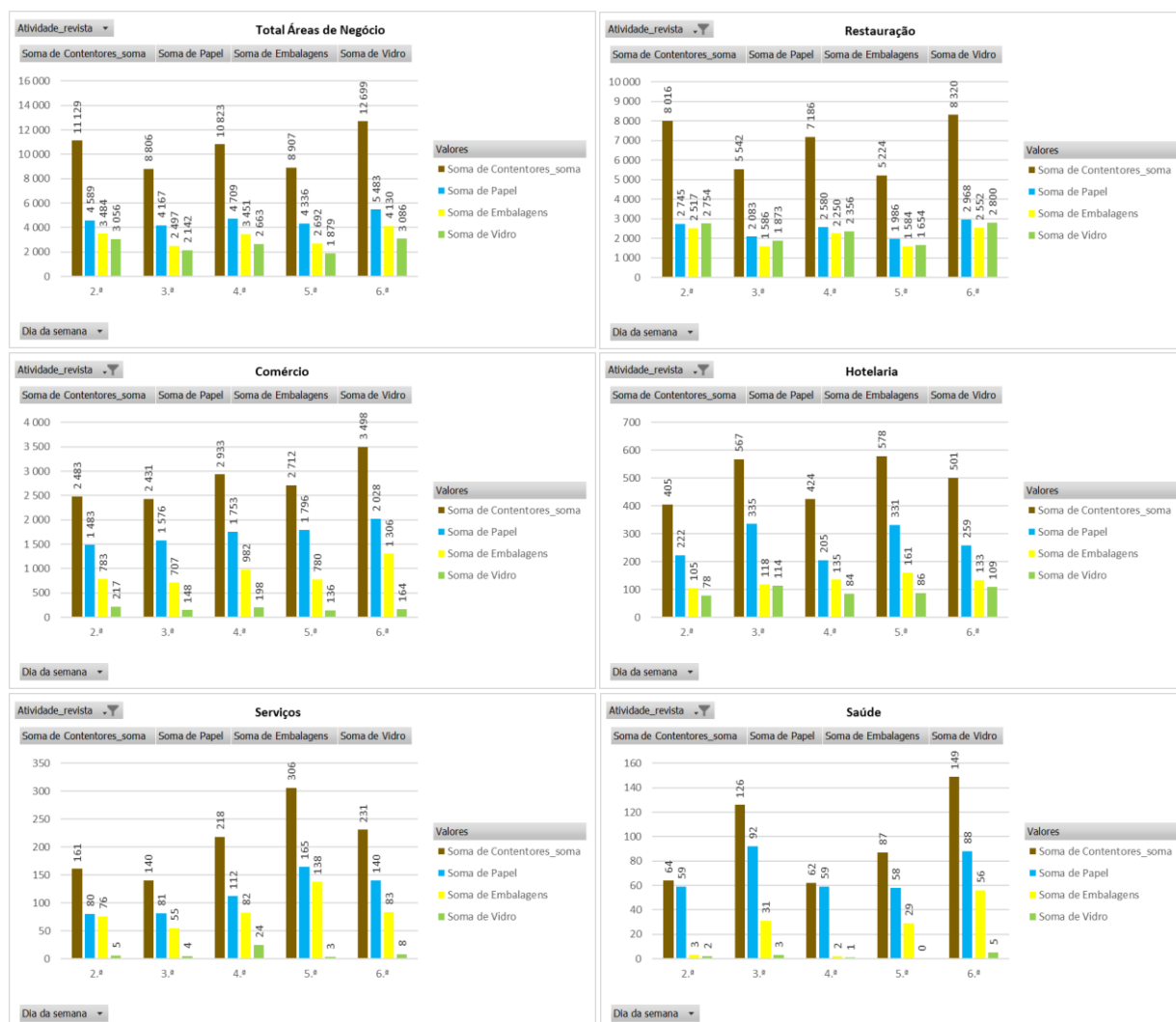


Figura 8: Distribuição do n.º de contentores recolhidos por dia da semana, material e área de negócio, obtido via *Microsoft Excel*

Analisando a evolução do número de contentores recolhidos por mês, verifica-se um crescimento da produção de resíduos a partir do mês de abril, com um pico no mês de agosto, mantendo-se, no segundo semestre do ano, os níveis de produção de resíduos superiores aos apresentados no primeiro semestre de 2017. Este crescimento encontra-se associado ao

alargamento da base de clientes/ aderentes ao projeto PAP-NR, em particular no Comércio e Restauração, conforme já referido anteriormente, bem como de uma maior participação dos clientes na separação e apresentação de resíduos à recolha, que é posteriormente analisada em detalhe na avaliação da Taxa de Colocação à Recolha dos clientes. Refira-se ainda que, o circuito PAP-NR em análise, pertence a um município com um elevado fluxo populacional nas épocas de verão e fim de ano. Este efeito de sazonalidade, é particularmente sentida nas áreas de negócio da Restauração, Hotelaria e Comércio, enquanto nas restantes áreas, a produção mantém-se estável. Este efeito é sentido com maior dimensão no material Papel/ Cartão e nas áreas de negócio da Hotelaria e Comércio, verificando-se um elevado crescimento do número de contentores recolhidos deste material a partir do mês de abril.

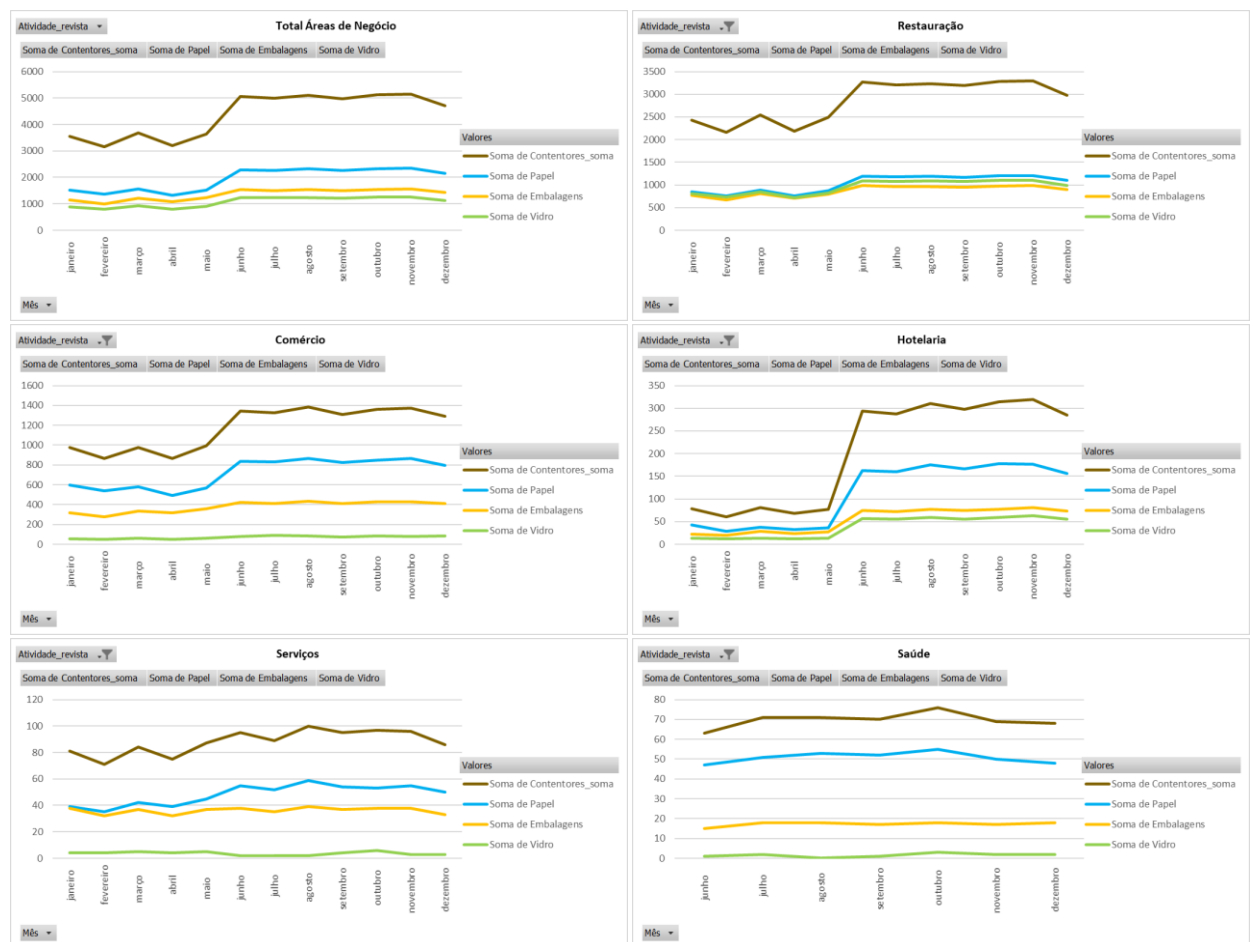


Figura 9: Distribuição do n.º de contentores recolhidos por mês, material e área de negócio, obtida via *Microsoft Excel*

A nível de tipologia de material, durante todo o ano, o Papel/ Cartão apresenta o maior número de contentores recolhidos, seguido das Embalagens e, por último, pelo Vidro.

O crescimento verificado na produção de resíduos no segundo semestre de 2017, encontra-se igualmente refletido no número total de contentores recolhidos por dia, atendendo que, diariamente, é assegurado apenas um único circuito de recolha. Em média, foram recolhidos aproximadamente 203 contentores por circuito de recolha. O valor máximo atingido de contentores recolhidos por circuito é de 286, registado no mês de dezembro (08.12.2017), e o valor mínimo de 134, foi registado nos meses de março, abril e maio.

No mês de junho, verifica-se a maior amplitude entre máximo e mínimo de contentores recolhidos. Conforme já referido anteriormente, no decorrer do mês de junho verificou-se um alargamento da base de clientes aderentes ao serviço PAP-NR, que não se refletiu logo no início do mês.

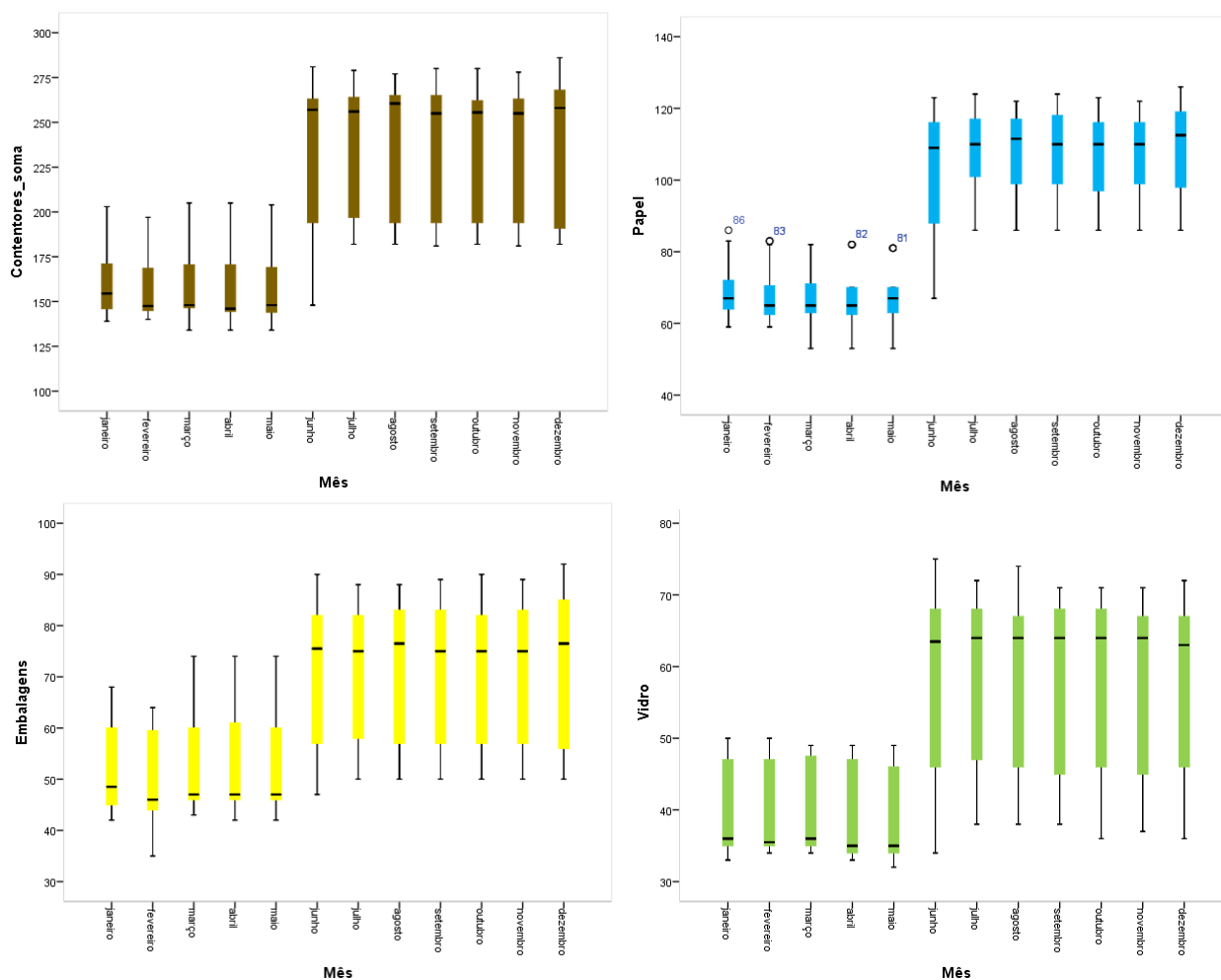


Figura 10: *Boxplots* do n.º total de contentores recolhidos por circuito de recolha, mês e material, obtidos via SPSS

Apenas no material Papel/ Cartão são verificadas ocorrências de *outliers* moderados, nos meses de janeiro, fevereiro, abril e maio. Nos restantes materiais, não se verifica a ocorrência de *outliers*. Nos materiais Papel/ Cartão e Embalagens, o valor máximo de contentores recolhidos por circuito é verificado em dezembro, enquanto no material Vidro, ocorre no mês de junho.

	N Válido	Média	Mediana	Moda	Desvio Padrão	Variância	Assimetria	Curtose	Amplitude	Mínimo	Máximo	Soma	Percentis		
Papel	258	90,25	86,00	122	22,607	511,098	-,004	-1,418	73	53	126	23284	68,50	86,00	112,25
Embalagens	258	63,00	59,00	46	15,735	247,595	,302	-1,357	57	35	92	16254	48,00	59,00	77,00
Vidro	258	49,71	47,00	46	13,633	185,871	,354	-1,423	43	32	75	12826	36,00	47,00	64,00
Contentores_soma	258	203,0	194,00	194	49,482	2448,47	,272	-1,404	152	134	286	52364	154,75	194,00	260,00

Tabela 4: Estatísticas descritivas do n.º total de contentores recolhidos por circuito de recolha, obtidas via SPSS

O número total de contentores de Papel/ Cartão recolhidos por circuito, varia entre um mínimo de 53 e um máximo de 126 contentores, com uma mediana de 86 contentores e uma média de 90,25 contentores recolhidos por circuito de recolha. Conforme já referido anteriormente, o aumento da produção de resíduos verificada no segundo semestre de 2017, é refletido com maior intensidade no Papel/ Cartão, em que *outliers* moderados de valores máximos de recolha de 81 a 86 contentores observados nos meses de janeiro, fevereiro, abril e maio, passam a constituir valores abaixo dos mínimos verificados a partir de julho. Os dados do número de contentores de Papel/ Cartão recolhidos por circuito de recolha, são aproximadamente simétricos, com uma assimetria negativa de -0,004.

Em média foram recolhidos 63 contentores de Embalagens por circuito durante o ano de 2017, tendo-se observado um mínimo de 35 (registado em fevereiro), um máximo de 92 contentores (registado em dezembro), e uma mediana de 59 contentores. Os dados do número de contentores recolhidos por circuito apresentam ainda uma assimetria positiva de 0,302.

O menor número de contentores recolhidos por circuito é verificado no material Vidro, com um mínimo de 32 contentores, verificado no mês de maio. O valor máximo de contentores de Vidro recolhidos, é observado no mês de junho, com um total de 75. A média do número contentores recolhidos é aproximadamente 50, apresentando ainda uma mediana de 47 e uma assimetria positiva de 0,354.

Durante o ano de 2017, foi registado um máximo de recolha de 286 contentores por circuito de recolha, uma média de 203 contentores, registando-se com maior frequência a ocorrência de recolha de um total de 194 contentores.

Nas figuras seguintes, são apresentados os histogramas do número total de contentores recolhidos por circuito de recolha, total e por tipo de material:

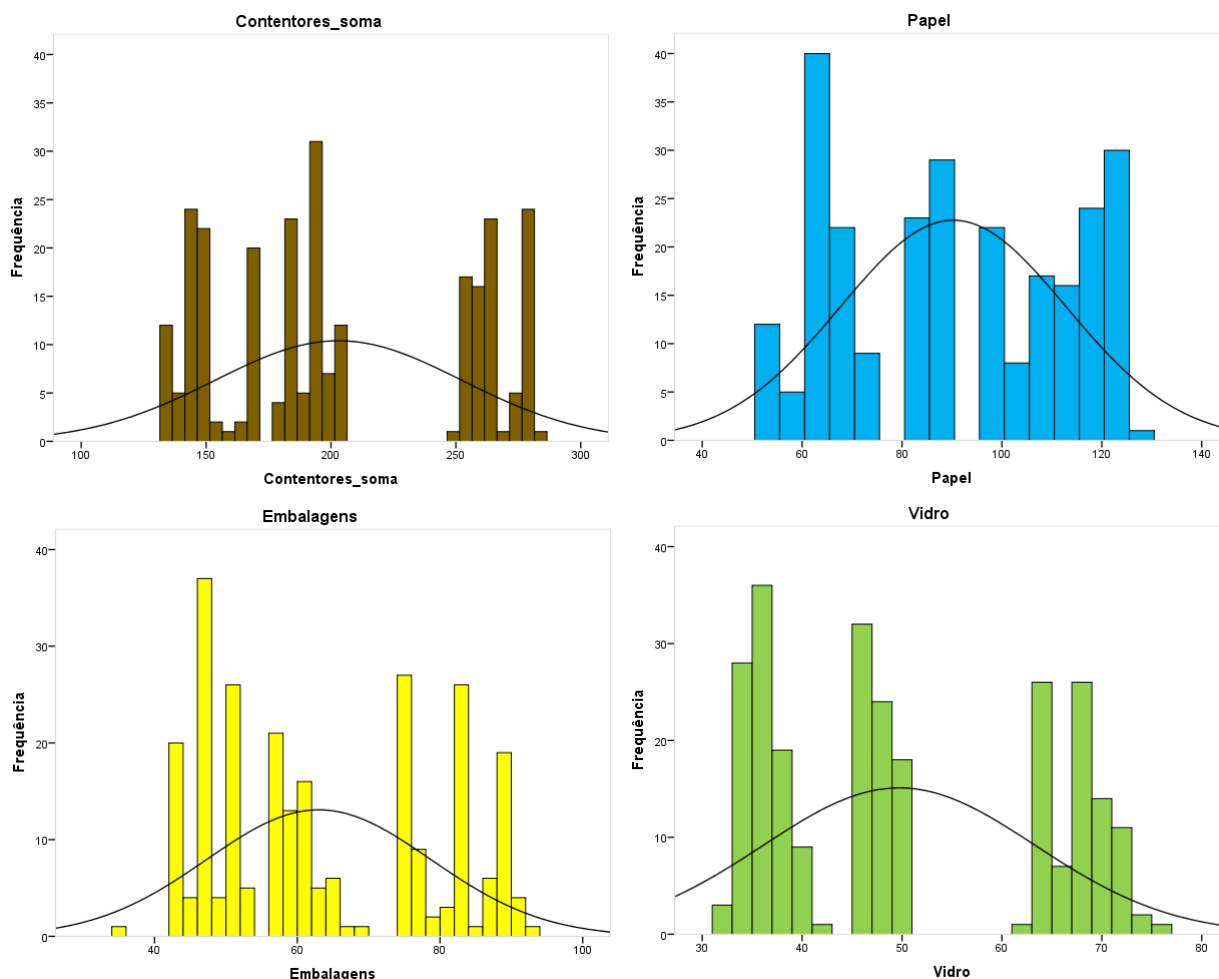


Figura 11: Frequência do n.º total de contentores recolhidos por circuito de recolha, obtida via SPSS

No gráfico seguinte é apresentada a evolução diária do número de contentores recolhidos por circuito, sendo destacados os feriados e outros dias festivos, procurando-se avaliar a ocorrência de picos no número de contentores recolhidos, nestes ou após estes dias.

A sexta-feira constitui, consistentemente ao longo do ano de 2017, o dia da semana com o maior número de contentores de recolhidos. Com o alargamento da base de clientes aderentes ao serviço PAP-NR em junho, a quarta-feira-feira que durante os primeiros 5 meses do ano constituía o dia da semana com menor número de recolhas, passa a posicionar-se como o segundo dia da semana com maior número de recolhas, ocupando o lugar da segunda-feira, que passa a ser o terceiro dia da semana com maior número de recolhas.

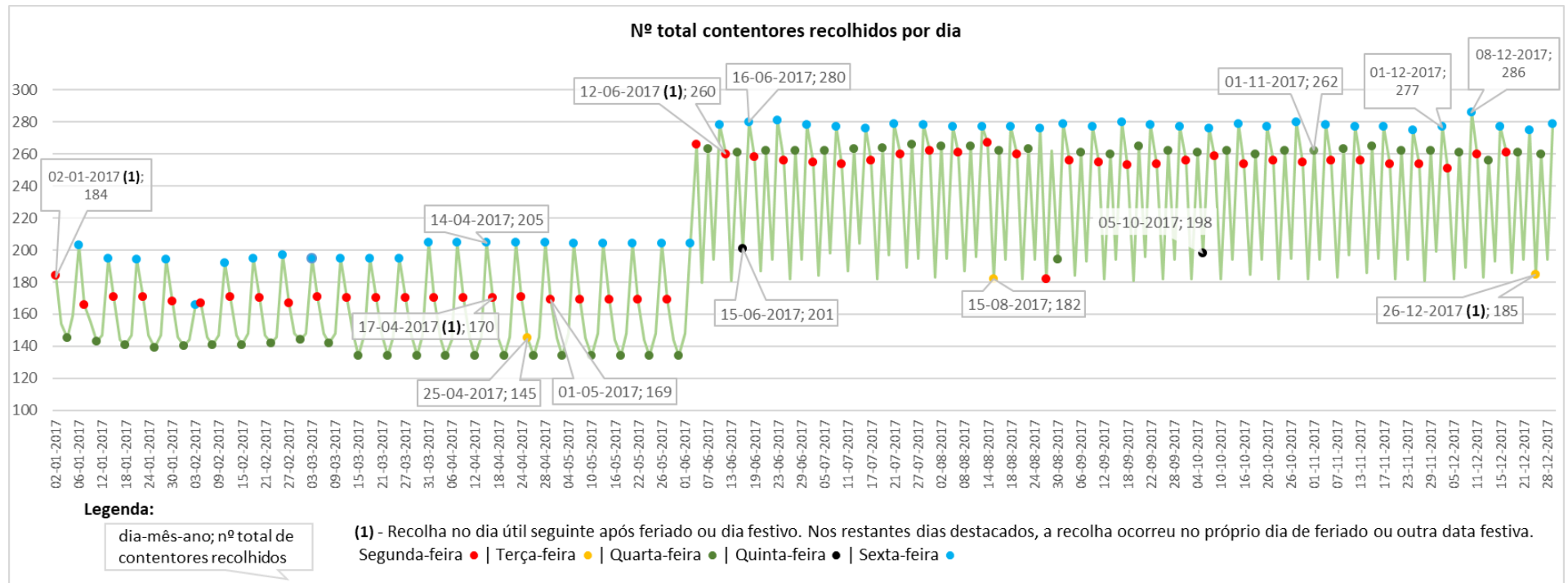


Figura 12: Evolução diária do nº de contentores recolhidos por circuito, obtida via *Microsoft Excel*

Verifica-se ainda uma suavização da diferença dos quantitativos de contentores recolhidos entre a sexta-feira e a segunda e quarta-feira, com particular expressão no mês de agosto (exceção dos dias 28 e 30 de agosto, que regista um valor de contentores recolhidos anormalmente baixo). Para além dos normais picos registados nas recolhas realizadas às segundas, quartas e sextas-feiras, não se verifica a ocorrência de picos extraordinários decorrentes de feriados e dias festivos.

A. A designação dos feridos e outros dias festivos, e respetiva ocorrência de realização de recolha nesses dias, podem ser consultados no Anexo

Nas figuras seguintes, encontra-se representada a evolução mensal do número de contentores recolhidos por ponto de recolha, total e por material:

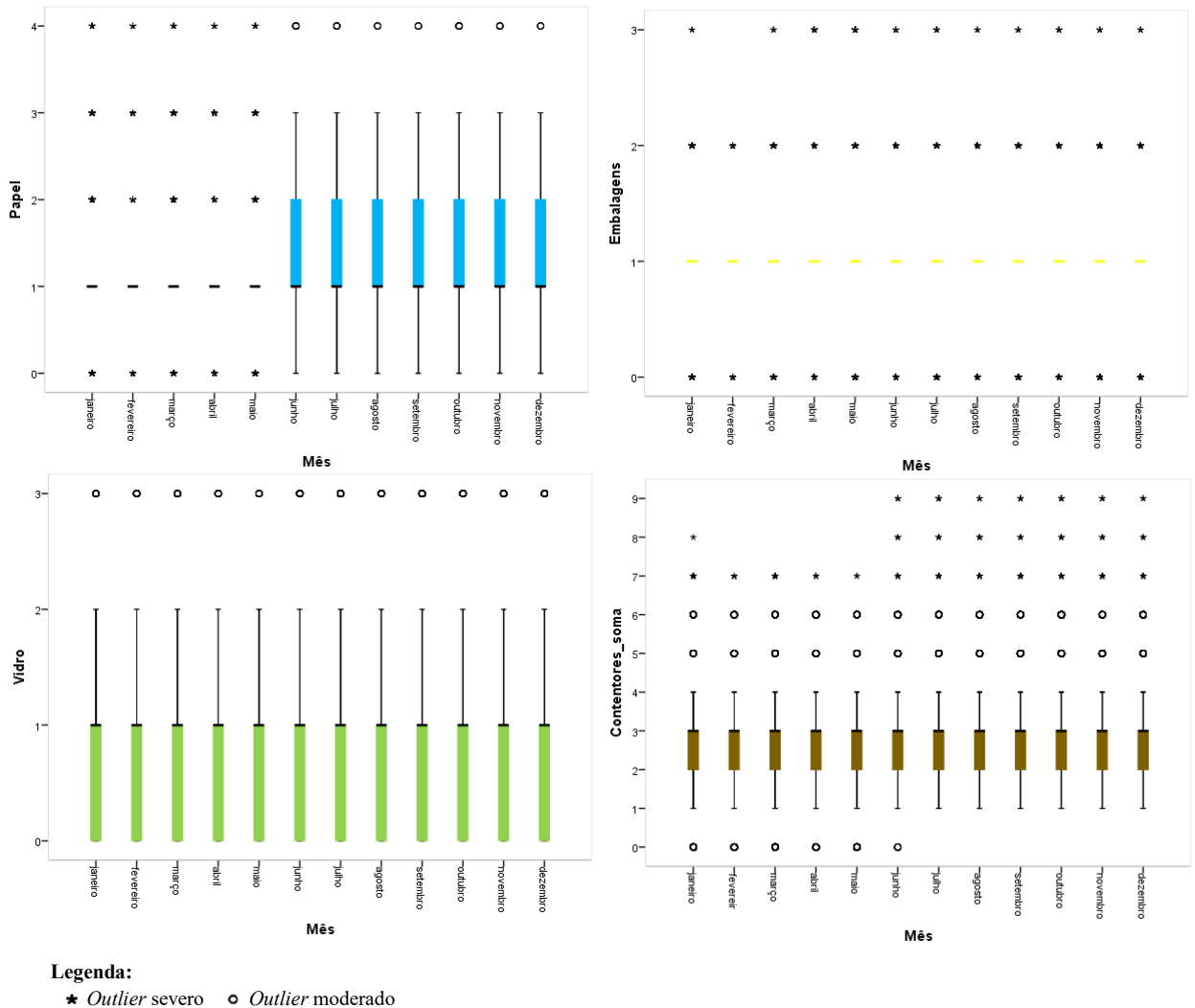


Figura 13: *Boxplots* do n.º de contentores recolhidos por ponto de recolha, mês e material, obtidas via SPSS

Relativamente ao material Papel/ Cartão, o número de contentores recolhidos varia entre 0 e 4, com uma mediana de 1 contentor e uma média de 1,26 contentores recolhidos por ponto de recolha. Nos meses de janeiro a maio, constata-se que 0, 2, 3 e 4 contentores recolhidos, constituem *outliers* severos; nos restantes meses do ano, as ocorrências de recolhas de 4 contentores constituem *outliers* moderados, verificando-se ainda uma amplitude interquartis de 1 contentor. Os dados do número de contentores de Papel/ Cartão recolhidos por ponto de recolha, apresentam uma assimetria positiva de 0,763.

O material Embalagens apresenta uma mediana de 1 contentor por recolha, com uma média de 0,88 contentores, constituindo *outliers* severos, para todos os meses do ano, a ocorrência de recolhas de 0, 2 e 3 contentores, com exceção do mês de fevereiro em que não se verificam ocorrências de recolha de 3 contentores. Os dados do número de contentores recolhidos apresentam ainda uma assimetria negativa de -0,164.

O Vidro apresenta ao longo do ano uma mediana de 1 contentor, com uma amplitude interquartis de 1 contentor, constituindo o máximo de 3 contentores um *outlier* moderado. A média do número contentores recolhidos é de 0,7, apresentando ainda uma assimetria positiva de 0,768.

Analisando o somatório dos contentores recolhidos para os três materiais, é possível inferir que o número de contentores recolhidos, por ponto de recolha, varia entre 0 e 9, apresentando uma mediana de 3 contentores, uma média de 2,84 contentores e uma assimetria positiva de 0,755. O valor máximo de número de contentores recolhidos por ponto de recolha, é registado nos meses de junho a dezembro, com a recolha de 9 contentores, em consonância com a evolução verificada no material Papel/ Cartão, que constitui o material com o maior número de recolhas, conforme já descrito anteriormente. Em janeiro é atingido um valor máximo de 8 contentores, enquanto que, nos meses de fevereiro a maio, é registado um máximo de 7 contentores. A recolha de 5 e 6 contentores constituem *outliers* moderados ao longo do ano. No mês de janeiro, a recolha de 7 e 8 contentores constituem *outliers* severos. Nos meses de fevereiro a maio, verifica-se a ocorrência de um *outlier* severo com a recolha de 7 contentores, enquanto nos restantes meses do ano, a recolha de 7, 8 e 9 contentores constituem *outliers* severos. Refira-se ainda que, apenas nos meses de janeiro a junho são registadas ocorrências de visitas ao cliente sem contentores recolhidos, constituindo estas situações *outliers* moderados.

	N Estatística	Amplitude Estatística	Mínimo Estatística	Máximo Estatística	Soma Estatística	Média Estatística	Desvio Padrão Estatística	Variância Estatística	Assimetria Estatística	Erro Padrão
Papel	18409	4	0	4	23284	1,26	,685	,470	,763	,018
Embalagens	18409	3	0	3	16254	,88	,471	,222	-,164	,018
Vidro	18409	3	0	3	12826	,70	,689	,474	,768	,018
Contentores_soma	18409	9	0	9	52364	2,84	1,255	1,575	,755	,018

Tabela 5: Estatísticas descritivas do n.º de contentores recolhidos por ponto de recolha, obtidas via SPSS

Outra forma de análise da distribuição do número de contentores recolhidos total e por material, poderá ser visualizada com recurso a gráficos de barras, sendo então possível analisar a frequência do número de contentores recolhidos por ponto de recolha.

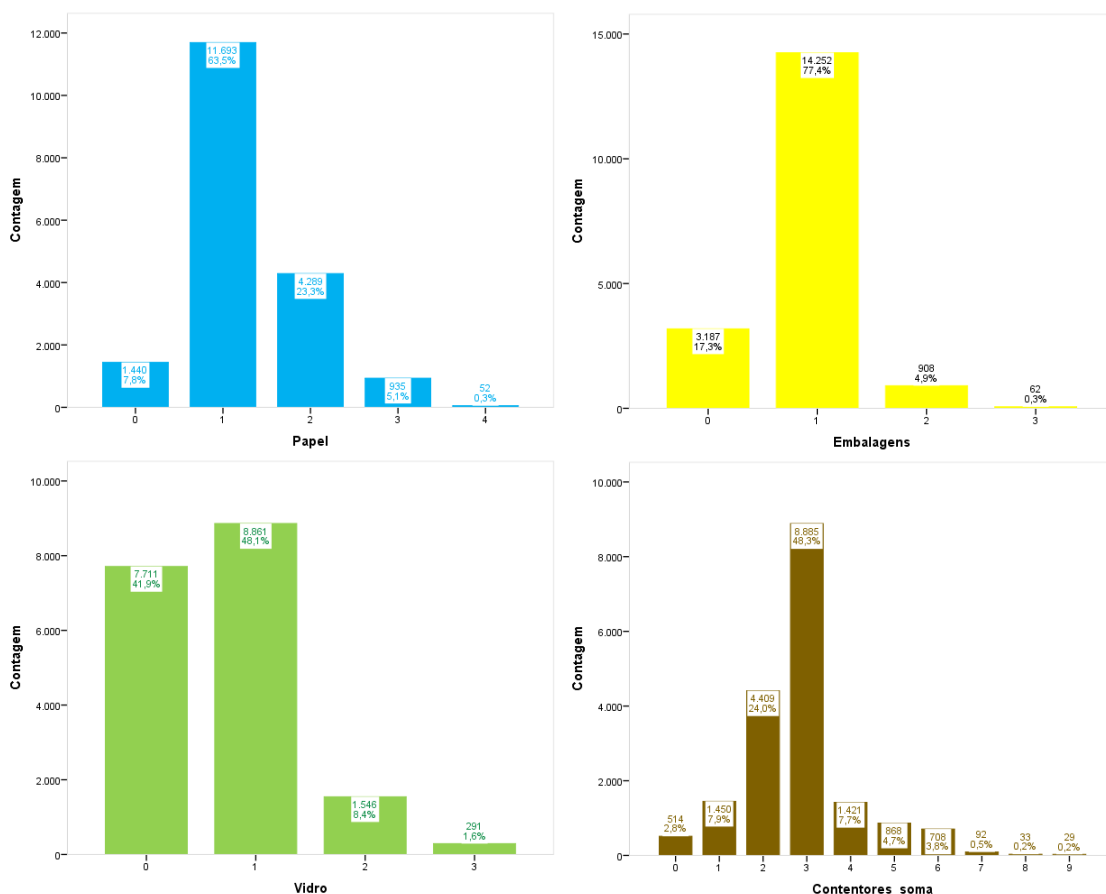


Figura 14: Frequência do n.º de contentores recolhidos e peso %, por ponto de recolha, obtido via SPSS

Em todos os materiais, a ocorrência da recolha de 1 contentor por recolha assume o maior peso, com maior expressão nas Embalagens com 77,4% dos casos, seguido do Papel com 63,5% e, por último, pelo Vidro, com 48,1% dos casos. No material Vidro, destaque-se ainda a elevada ocorrência de visitas a clientes com 0 contentores recolhidos, que representa 41,9% das recolhas. Recorde-se que o Vidro é o material com menor número total de contentores recolhidos durante o ano de 2017.

Analisando o número total de contentores recolhidos por ponto de recolha, a recolha de 3 contentores constituiu o cenário mais representativo, ocorrendo em 48,3% dos casos. Na figura seguinte é possível verificar a evolução mensal do número de contentores recolhidos por ponto de recolha.

Agendamento Dinâmico de Circuitos de Recolha Seletiva de Resíduos Porta-a-Porta Não Residencial

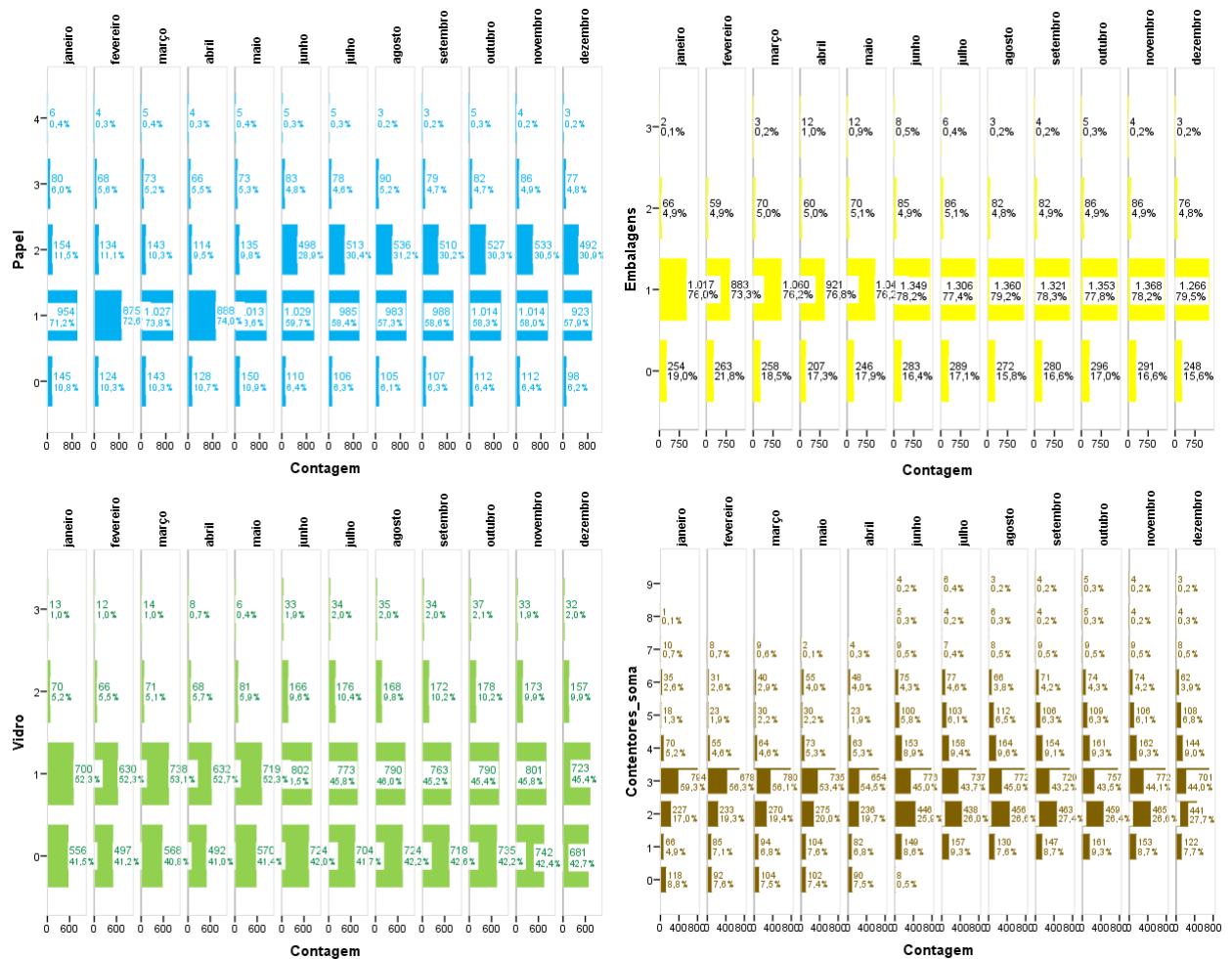


Figura 15: Evolução mensal da frequência do n.º de contentores recolhidos e peso %, por ponto de recolha, obtido via SPSS

Relativamente ao material Papel, apesar da ocorrência da recolha de 1 contentor por recolha ser a mais representativa ao longo do ano, é possível observar, a partir de junho, um ligeiro decréscimo para valores aproximados de 58%, face a um peso medio de 73% verificado nos primeiros 5 meses do ano. Em oposição, a ocorrência da recolha de 2 contentores passa a representar a partir de junho valores médios de 30%, face aos 11% verificado nos primeiros meses do ano. As Embalagens apresentam ao longo do ano valores médios de 78% de ocorrências de recolhas de 1 contentor.

Conforme já referido anteriormente, Vidro é o material em que se verifica o maior número de ocorrências de recolha de 0 contentores, registando valores bastante próximos das ocorrências de recolha de 1 contentor. Refira-se ainda que, a partir do mês de junho,

verifica-se um crescimento das ocorrências da recolha de 2 contentores, com um valor médio de 10%, face aos 6% registados nos primeiros meses do ano.

Analisando o número total de contentores recolhidos por ponto de recolha, a ocorrência de recolha de 3 contentores é a mais representativa ao longo do ano. A recolha de 2 contentores apresenta um crescimento ao longo do ano, evoluindo de 17% em janeiro, para um peso de 27,7% das recolhas no final do ano. A partir do mês de junho, é possível verificar um crescimento do peso das ocorrências de recolha 4 e 5 contentores, destacando-se ainda, com uma reduzida expressão, a existência de ocorrências de recolhas de 8 e 9 contentores por ponto de recolha.

4.3.3 Análise da eficiência dos circuitos de recolha

Para uma análise mais aprofundada do desempenho dos circuitos de recolha, foi definido o seguinte indicador - **Taxa de eficiência dos circuitos (TEC)**. A eficiência dos circuitos de recolha, traduz o grau de cumprimento dos circuitos de recolha agendados pela respetiva equipa de recolha, tendo em consideração o número de pontos de recolha visitados pela equipa (com ou sem resíduos recolhidos) e o número total de pontos de recolha agendados para o circuito de recolha.

$$\text{Taxa eficiência circuitos (\%)} = \frac{N^{\circ} \text{ pontos recolha visitados}}{N^{\circ} \text{ pontos recolha agendados}}$$

Figura 16: Cálculo da Taxa de Eficiência dos Circuitos de Recolha - TEC (%)

Analisando a taxa de eficiência dos circuitos de recolha, é possível verificar que nos primeiros quatro meses do ano, são registadas taxas próximas de 100%, o que significa que, quase todos os clientes agendados para recolha, foram visitados pela equipa de recolha. De julho a dezembro, a eficiência dos circuitos de recolha decresce para valores próximos de 60%. Quarta e sexta-feira são os dias da semana que registam os maiores valores médios de taxa de eficiência da recolha, com aproximadamente 83%, em oposição à terça-feira, que regista o valor mais baixo com cerca de 67%.

Agendamento Dinâmico de Circuitos de Recolha Seletiva de Resíduos Porta-a-Porta Não Residencial



Figura 17: Distribuição da Taxa de Eficiência da Recolha por mês e dia da semana, obtida via *Microsoft Excel*

Analisando ainda a evolução comparativa do número de pontos de recolha agendados *versus* o número de pontos de recolha recolhidos, isto é, visitados pela equipa de recolha, é possível verificar que, a partir do início do mês de junho e até ao final do ano, para todos os circuitos de recolha é agendada visita à totalidade dos pontos de recolha, correspondente a 131 clientes.

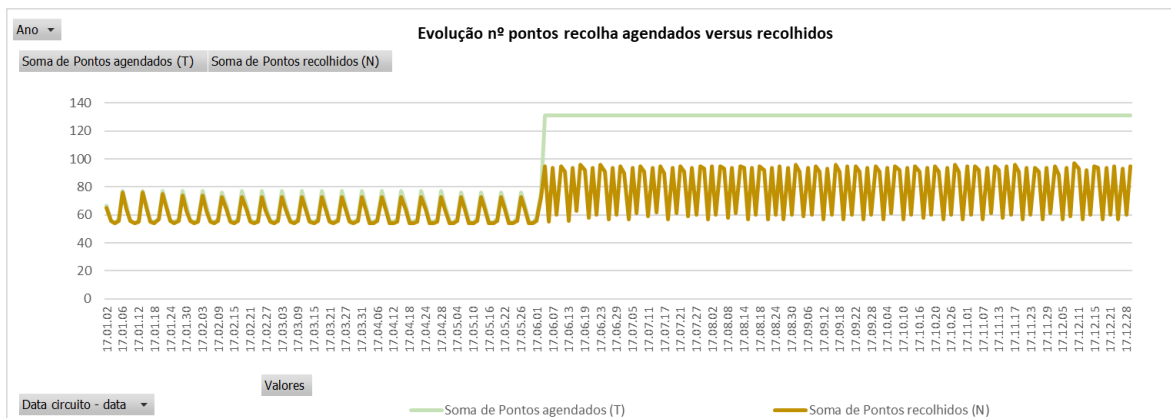


Figura 18: Evolução do n.º de pontos recolha agendados *versus* recolhidos/ visitados, obtida via *Microsoft Excel*

Na figura seguinte, é ainda apresentado histograma com a frequência das observações da variável pontos recolhidos, sendo possível verificar que durante o ano de 2017, foi observado um máximo de 97 clientes visitados por circuito de recolha e um mínimo de 54, sendo ainda 95 pontos de recolha visitados o segundo valor mais frequente, valores distantes dos 131 pontos de recolha agendados, o que explica o decréscimo verificado nas taxas de eficiência dos circuitos.

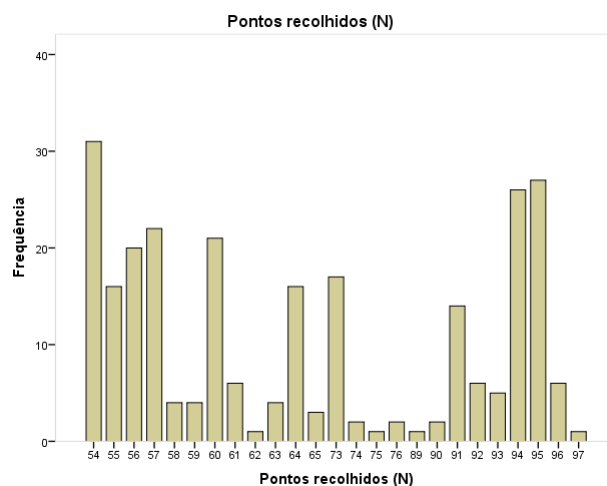


Figura 19: Frequências de pontos de recolha recolhidos/ visitados, obtidas via SPSS

	Estatísticas														
	N		Média	Mediana	Moda	Desvio Padrão	Variância	Assimetria	Curtose	Amplitude	Mínimo	Máximo	Percentis		
	Válido	Omisso											25	50	75
Pontos recolhidos (N)	258	0	71,37	63,50	54	16,859	284,241	,456	-1,595	43	54	97	56,00	63,50	93,00
Pontos agendados (T)	258	0	101,70	131,00	131	34,499	1190,188	-,372	-1,795	77	54	131	58,00	131,0	131

Tabela 6: Estatísticas descritivas das variáveis pontos recolhidos e pontos agendados, obtidas via SPSS

Pela análise das figuras anteriores e das estatísticas descritivas dos pontos de recolha recolhidos e pontos de recolha agendados, é ainda possível verificar uma elevada variabilidade destas variáveis, bem patente nos elevados desvios padrões e amplitudes estatísticas. Esta variabilidade assume maior expressão a partir do mês de junho, verificando-se uma alternância quase diária entre circuitos com valores elevados de pontos de recolha recolhidos e circuitos com baixos valores de pontos de recolha recolhidos. Esta dicotomia poderá ainda ser constatada pela análise da distribuição de frequências dos pontos recolhidos representada no respetivo histograma, onde se destacam dois grupos opostos de valores frequentes, nomeadamente [54,57] e [94,95].

4.3.4 Análise da participação dos clientes aderentes

Para uma análise mais aprofundada da participação dos clientes aderente ao projeto PAP-NR, foi definido o indicador **Taxa de colocação de resíduos à recolha (TCR)**, que pretende avaliar o número de vezes que o serviço de recolha foi prestado/ disponibilizado ao cliente (número de vezes que o cliente foi visitado pelas equipas de recolha), e o número de entregas de resíduos pelo cliente (número de vezes que o cliente apresentou à recolha 1 ou mais contentores, independentemente do tipo de resíduo). Este conceito é traduzido pela seguinte expressão:

$$\text{Taxa colocação de resíduos à recolha (\%)} = \frac{N^{\circ} \text{ entregas cliente}}{N^{\circ} \text{ visitas ao cliente}}$$

Figura 20: Cálculo da Taxa de Colocação de Resíduos à Recolha - TCR (%)

Em 2017 foi registada uma taxa de colocação à recolha global de 97,2%, o que traduz uma elevada participação dos clientes. Todas as áreas de negócio apresentam taxas de colocação à recolha próximas de 100%, apresentando a área de negócio da Restauração, área com o maior número de clientes, uma taxa de 98,1%.

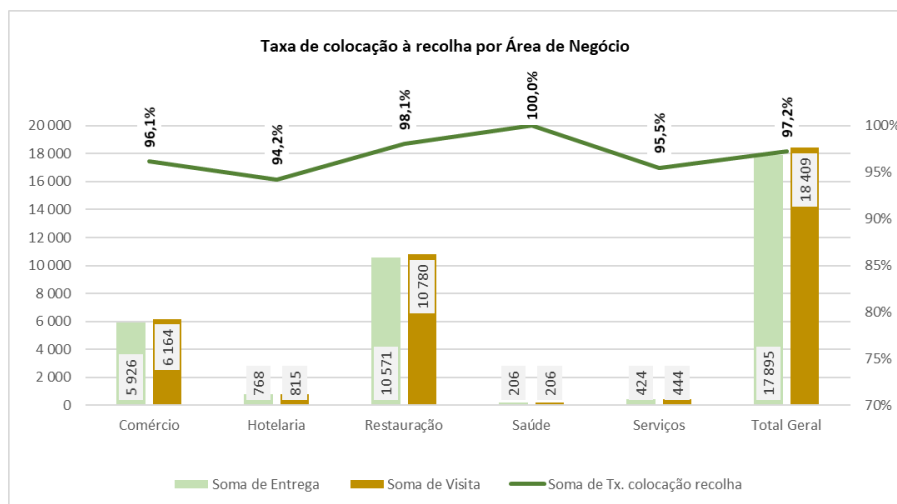


Figura 21: Taxa de Colocação à Recolha total e por área de negócio, obtida via *Microsoft Excel*

Na figura seguinte, é ainda apresentado histograma com a frequência das observações da taxa de colocação à recolha pelos clientes aderentes, sendo a taxa de 100% o valor mais frequente observado para um total de 117 clientes.

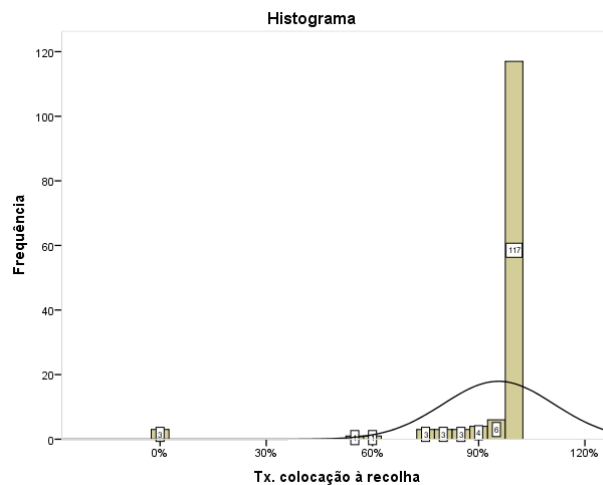


Figura 22: Histograma da Taxa de Colocação à Recolha, obtido via SPSS

A maior taxa de colocação à recolha verifica-se à segunda e sexta-feira com valores de aproximadamente 98%, e a menor à terça-feira com 96,5%, à semelhança do comportamento anteriormente identificado a nível do número de contentores recolhidos por dia da semana.

As áreas de negócio menos representativas em número de clientes aderentes, nomeadamente Saúde e Serviços, apresentam taxas de 100% na maioria dos dias da semana (exceção à sexta-feira, na área da Saúde, em que é registada taxa de 84,1%). Na Hotelaria verifica-se a maior variabilidade das taxas de colocação à recolha com um mínimo de 85%, à segunda-feira, e uma taxa máxima de 100% registada à terça e sexta-feira.

Nas áreas de negócio da Restauração e Comércio, que apresentam o maior número de aderentes, verifica-se uma maior estabilidade nas taxas de colocação à recolha ao longo da semana; contudo, não chegam a atingir a taxa máxima de 100%.

Na Restauração, verificam-se valores próximos de 100%, com um mínimo de 97,4% à quinta-feira e um máximo de 98,8% registado à terça e sexta-feira.

No Comércio regista-se a taxa máxima de 98% à segunda-feira, e a mínima de 91,5% à terça-feira.

Agendamento Dinâmico de Circuitos de Recolha Seletiva de Resíduos Porta-a-Porta Não Residencial

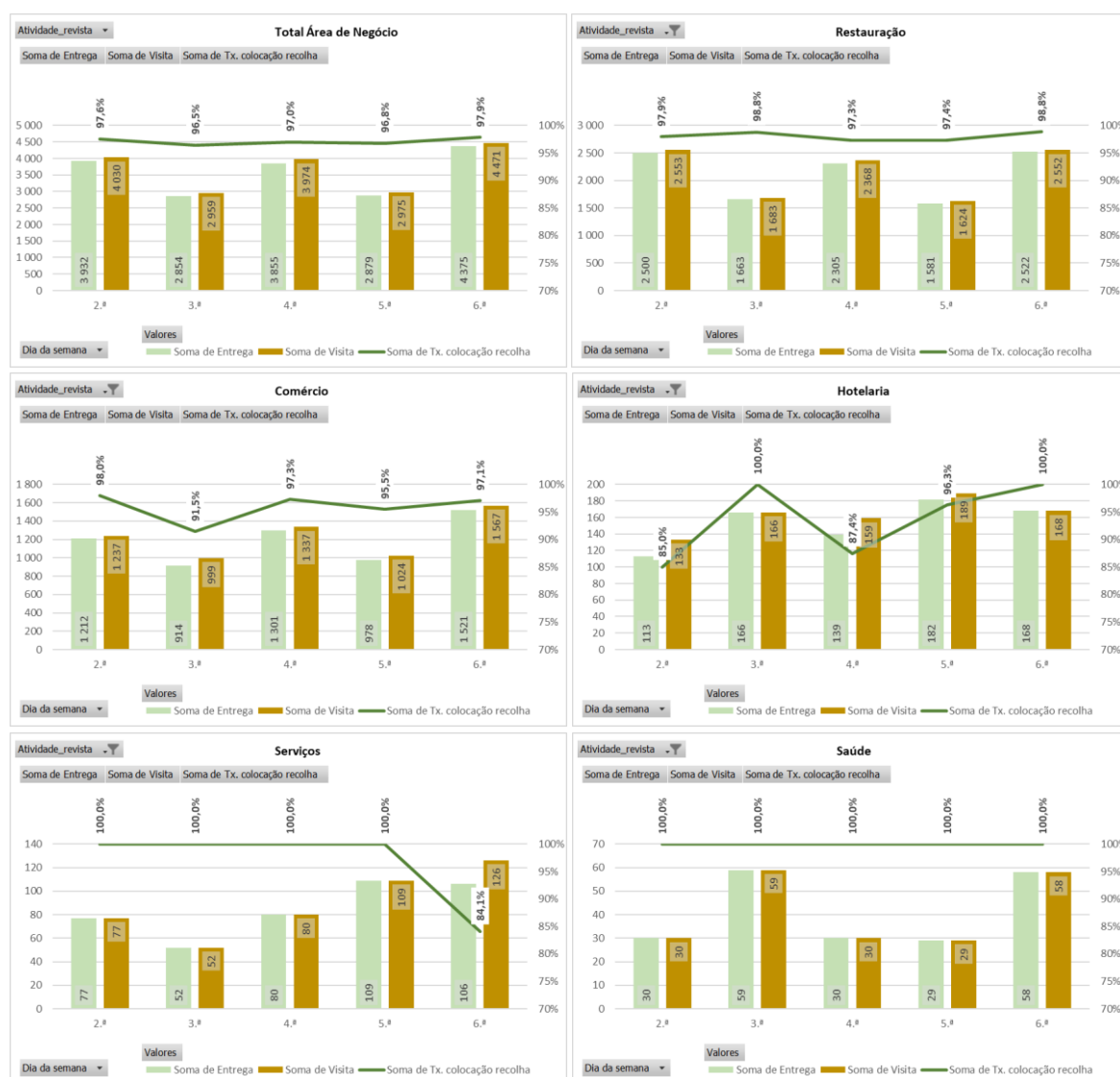


Figura 23: Taxa de Colocação à Recolha por dia da semana e área de negócio, obtida via *Microsoft Excel*

Analisando a evolução mensal da taxa de colocação à recolha, é possível verificar que a taxa mais baixa é obtida no arranque do projeto PAP-NR em janeiro de 2017, com uma taxa de 91,2%, sendo registadas a partir do mês de julho, taxas de 100%, período coincidente com o aumento do número de contentores recolhidos, conforme já referido anteriormente.

Em todas as áreas de negócio a taxa de colocação à recolha máxima de 100% é obtida a partir do mês de julho, com exceção da Hotelaria e Saúde, em que a taxa máxima é registada a partir do mês de junho. Na área de negócio de Hotelaria, é ainda observada a menor taxa de colocação à recolha de todas as áreas de negócio durante o ano de 2017, com uma taxa de 72,5% em fevereiro.

Agendamento Dinâmico de Circuitos de Recolha Seletiva de Resíduos Porta-a-Porta Não Residencial

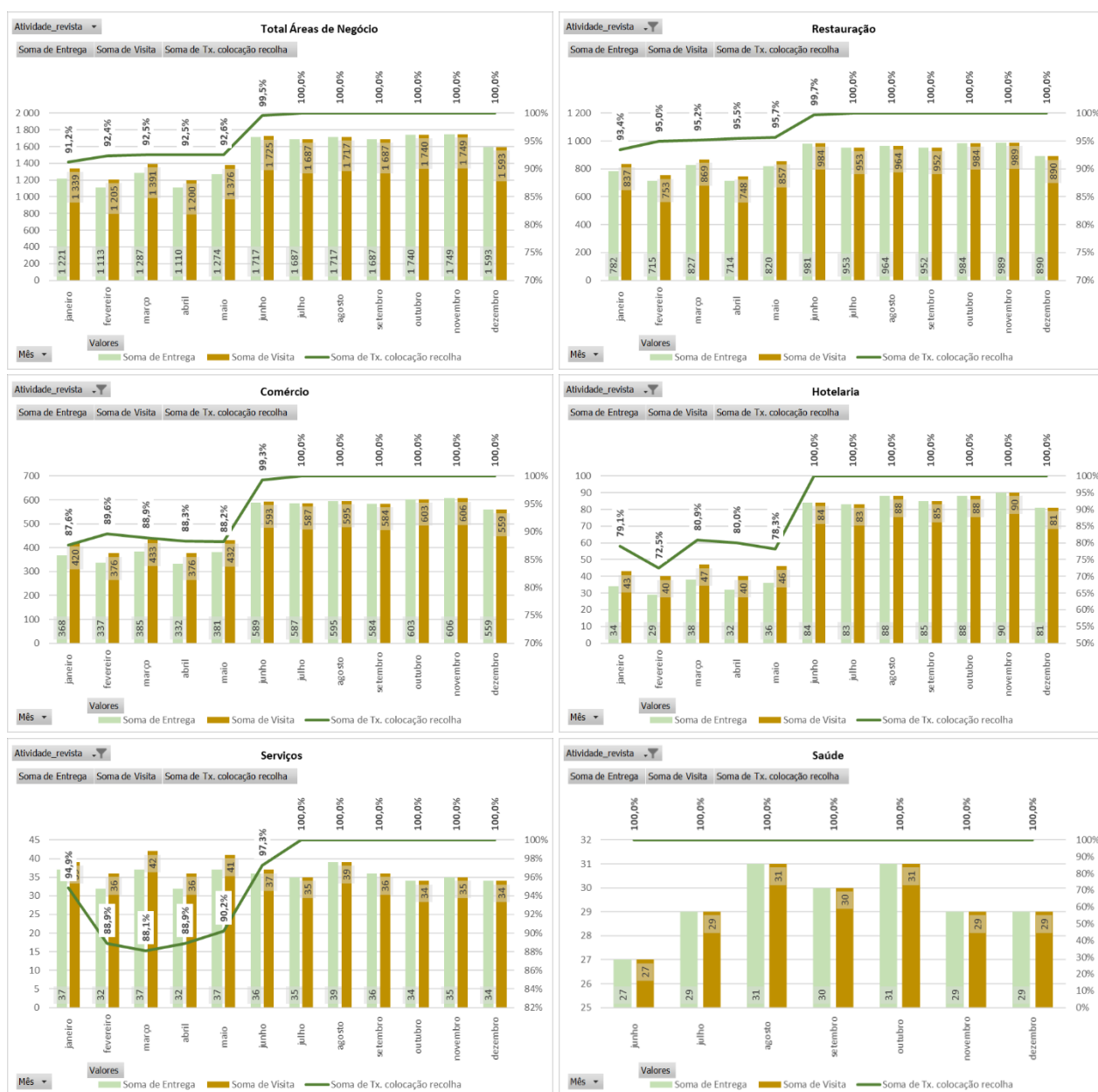


Figura 24: Taxa de Colocação à Recolha por mês e área de negócio, obtida via *Microsoft Excel*

Na área de negócio de Serviços, que apresenta um número reduzido de clientes, verifica-se uma diminuição do número de visitas e entregas a partir do mês de junho, continuando a registar-se uma taxa de colocação à recolha de 100%, tendo-se mantido o mesmo número de clientes. Nas restantes áreas de negócio, verifica-se um crescimento do número de visitas e entregas a partir do mês de junho, sendo registadas consistentemente taxas de 100%, de julho até ao final do ano.

Analisando a evolução do número de contentores recolhidos por visita, é possível observar um crescimento do número médio de contentores recolhidos, a partir do mês de junho, atingindo valores máximos próximos de 3,3 contentores recolhidos por visita, em oposição a valores máximos registados de aproximadamente 2,8 contentores no primeiro semestre. Este crescimento resulta, não só, do alargamento da base de clientes aderentes ao projeto PAP-NR, em particular no Comércio e Restauração, conforme já referido anteriormente, bem como de uma maior participação dos clientes na separação e apresentação de resíduos à recolha. Verifica-se ainda, um crescimento da frequência da recolha, como pode ser observado na figura abaixo relativa à evolução do número de visitas por cliente. O valor máximo do número de visitas por cliente verifica-se em dezembro com uma média de 13,4 visitas, em que cada cliente é visitado em média 3 a 4 vezes por semana. Em fevereiro e abril registam-se valores mínimos de 10,5 e 10,4 visitas por cliente, respetivamente, que se traduzem numa média de 2 a 3 visitas semanais.

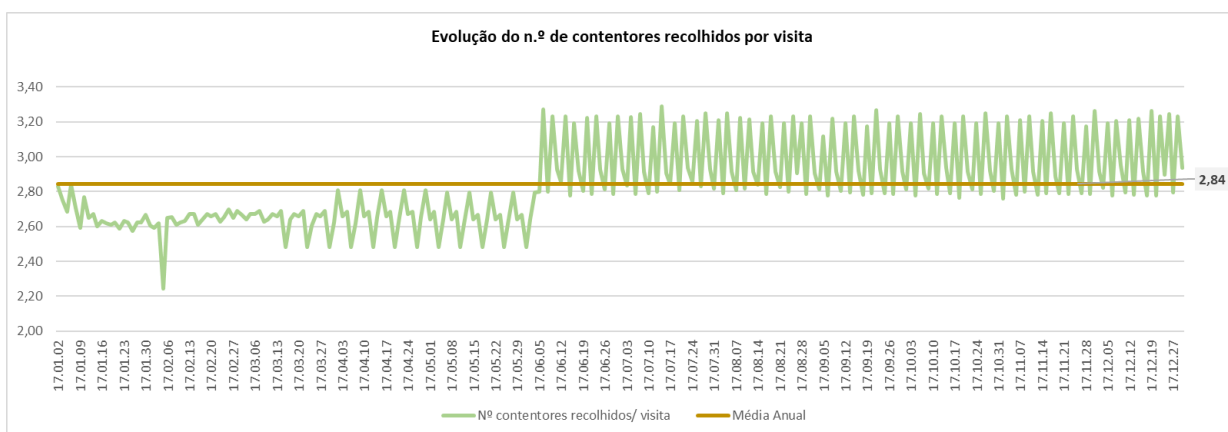


Figura 25: Evolução do n.º de contentores recolhidos por visita/ cliente, obtida via *Microsoft Excel*

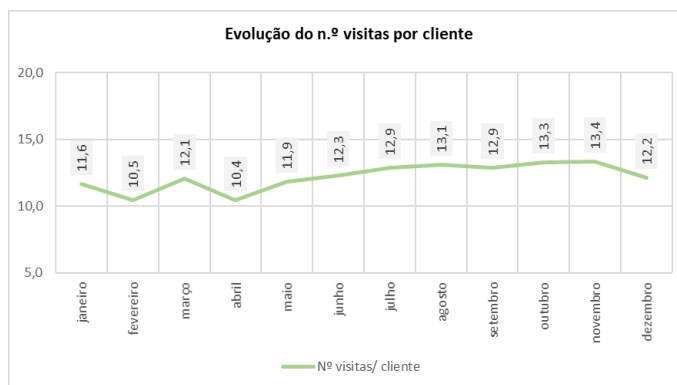


Figura 26: Evolução mensal do n.º de visitas por cliente, obtida via *Microsoft Excel*

4.4. Preparação dos dados

Esta etapa do processo de *data mining* é essencial para a construção de um modelo preditivo eficiente em termos exigência computacional, simplicidade, velocidade do processo de indução, bem como da sua capacidade de predição. Procura-se assim, através da limpeza e transformação dos dados, não enviesar a construção e avaliação do modelo preditivo. Trata-se de um processo iterativo, nomeadamente na seleção dos atributos a utilizar e a *performance* obtida pelo modelo, tendo sido testados diferentes conjuntos de atributos.

Após extração dos dados da base de dados procedeu-se a uma análise preliminar dos mesmos, procurando identificar-se e corrigir ocorrências de valores em falta, dados redundantes e inconsistentes, *outliers* e existência de atributos não relevantes, tendo sido assegurado o seguinte:

- a) Valores em falta: Foram apenas identificados valores em falta no atributo “Atividade”, tendo sido garantido o devido preenchimento da informação. Esta ocorrência identificou-se num número reduzido de exemplos da tabela de clientes aderentes (1,3%);
- b) *Outliers*: Os outliers identificados foram avaliados como fidedignos, não correspondendo assim a ruído nos dados, resultantes de erros de registo da informação, nomeadamente, no número de contentores recolhidos por fluxo e cliente.
- c) Seleção de atributos: Procedeu-se à eliminação de atributos não relevantes para a análise e construção do modelo pretendido, nomeadamente, “Nº sequencial”, “Morada” e “Geolocalização”, dados relevantes para o desenvolvimento de um modelo de otimização de rotas, mas que não se enquadra no âmbito e objetivo da presente tese;
- d) Transformação e criação de atributos: O atributo “Data do circuito” com exemplos no formato dd:mm:yy, foi desdobrado em dois novos atributos – dia da semana e mês, num formato, mais adequado ao tipo de modelo desenvolvido. Para conveniência da análise e desenvolvimento do modelo, foram criados os seguintes atributos:
 - ✓ “Contentores_soma” - referente ao somatório do número de contentores recolhidos por cliente;
 - ✓ “Entrega_RU(1/0)” - identifica a ocorrência de entrega de contentores à recolha pelo cliente aderente ao circuito PAP-NR (variável binária 0 e 1, correspondendo 0 à não entrega de contentores à recolha e 1, o oposto).

Procurando-se identificar e analisar as variáveis que influenciam a produção de resíduos neste tipo de circuitos, foram introduzidas no *data set*, quatro novas variáveis:

- ✓ Horário – identifica o número de horas de funcionamento do estabelecimento comercial do cliente. Os dados foram discretizados em três classes nominais de horários: reduzido (horário compreendido no intervalo de horas de $[1,7]$), normal (intervalo $]7,12]$) e alargado (intervalo $]12,24]$);
 - ✓ Feriado_Festividades_Fds - identifica as recolhas realizadas após fins de semana, feriados e outras datas festivas específicas do município, por forma a analisar o impacto nos quantitativos de resíduos recolhidos (variável binária 0 e 1, correspondendo 0 à não ocorrência de recolha após fim-de-semana, feriado e dias festivos, e 1 o oposto);
 - ✓ Centralidade – variável que, tendo em conta a geolocalização do estabelecimento comercial do cliente, define o seu nível de centralidade numa escala de 1 a 3. O nível 1 corresponde a uma zona central, com maior concentração de infraestruturas (estabelecimentos comerciais e serviços, equipamentos e infraestruturas sociais, ambientais e culturais), e elevada circularização de pessoas decorrente de um elevado número de residentes e/ ou população flutuante. O nível mínimo 3 corresponderá a zonas de subúrbios, com menor disponibilidade de infraestruturas e circularização de pessoas.
 - ✓ Clima – identifica as condições meteorológicas verificadas no dia anterior à recolha, nomeadamente a nível de temperatura. Esta variável foi construída com base em dados diários da temperatura máxima do ar ($^{\circ}\text{C}$), da estação meteorológica do Porto - Pedras Rubras, fornecidos pelo IPMA – Instituto Português do Mar e da Atmosfera. Para uma melhor representação dos dados, quer para a sua análise, quer para a construção e desempenho do modelo preditivo, os dados numéricos da temperatura em graus centígrados, foram discretizados em três classes nominais, de acordo com as escalas de classificação climática de Köppen-Geiger e Trewartha (Köppen-Geiger, 1954, e Trewartha & Horn, 1980, in Belda et al., 2014): frio (temperaturas no intervalo $]-\infty, 18]$), ameno (temperaturas compreendidas no intervalo $]18, 22,2]$) e quente (temperaturas no intervalo de $]22,2, +\infty[$).
- e) Dados redundantes e inconsistentes: Verificou-se que a descrição da atividade comercial exercida pelo cliente apresentava alguma redundâncias e inconsistências na sua nomenclatura, nomeadamente, uma excessiva proliferação de tipos de atividade, a existência de descrições alternativas para um mesmo tipo de atividade, descrições genéricas sem identificação do tipo de atividade (ex: “Outros”). Deste modo, procedeu-

se a uma revisão, uniformização e simplificação da nomenclatura de descrição do tipo de atividade, passando-se de 51 diferentes tipos para um total de 5, nomeadamente Comércio, Restauração, Saúde, Hotelaria e Serviços. O agrupamento das atividades na nomenclatura revista pode ser consultado no Anexo B.

Após a realização das operações atrás descritas e tendo em conta a análise efetuada a nível da compreensão dos dados, o *data set* final a utilizar na construção do modelo é composto pelos seguintes atributos:

Atributos	Formato	Intervalo de valores
Dia_semana	polinomial	segunda, terça, quarta, quinta, sexta
Mes	polinomial	janeiro, fevereiro, março, abril, maio, junho, julho, agosto, setembro, outubro, novembro, dezembro
Nome_cliente	polinomial	-
Atividade_revista	polinomial	Comércio, Restauração, Saúde, Hotelaria, Serviços
Horario	real	reduzido, normal, alargado
Feriado_Festividades_Fds(1/0)	binomial	[0:1]
Clima	polinomial	frio, ameno, quente
Papel	polinomial	[0:4]
Embalagens	polinomial	[0:3]
Vidro	polinomial	[0:3]
Entrega_RU(1/0)	binomial	[0:1]

Tabela 7: Descrição dos atributos do modelo preditivo

4.5. Modelação

Conforme já referido no capítulo 4.1, em que é formulado o problema em análise, na presente tese propõe-se o desenvolvimento de um modelo preditivo para a previsão da colocação ou não de resíduos à recolha pelos clientes, previsão do número de contentores colocados à recolha e identificação de regras que caracterizam a produção de resíduos pelos clientes. Tendo em conta os problemas em análise e *outputs* propostos, foram adotadas abordagens metodológicas distintas, enquadradas em problemas de classificação em *supervised learning*.

Nos quadros seguintes, são apresentados os problemas em análise, abordagem metodológica adotada e respetivo *output*, bem como as respetivas variáveis do *data set* utilizadas:

Problemas			Abordagem metodológica	Output	
1	Cliente possui resíduos para recolha?	1.1	Previsão da colocação de resíduos à recolha pelos clientes.	<i>Naive Bayes</i> Árvores de decisão Regressão Logística	1 (=Sim) 0 (=Não)
		1.2	Identificação de regras caracterizadoras da entrega de resíduos pelos clientes.	Árvores de decisão Regras de decisão	Regras de decisão
2	Quantos contentores o cliente possui para recolha?	2.1	Previsão do n.º de contentores colocados à recolha (somatório dos três fluxos - Papel/ Cartão, Embalagens e Vidro) pelos clientes.	<i>Naive Bayes</i> Árvores de decisão Regressão Logística	Nº total de contentores
		2.2	Identificação de regras caracterizadoras do n.º total de contentores que os clientes colocam à recolha.	Árvores de decisão Regras de decisão	Regras de decisão

Tabela 8: Problemas em análise, abordagem metodológica e *outputs*

Atributos	Tipo	Problema 1		Problema 2	
		1.1	1.2	2.1	2.2
Dia_semana	Atributo	X	X	X	X
Mes	Atributo	X	X	X	X
Nome_cliente	Atributo				
Atividade_revista	Atributo		X		X
Horario	Atributo		X		X
Feriado_Festividades_Fds(1/0)	Atributo	X	X	X	X
Clima	Atributo	X	X	X	X
Contentores_soma	Atributo classe			X	X
Entrega_RU(1/0)	Atributo classe	X	X		

Tabela 9: Problemas em análise e respetivos *data sets*

Na abordagem metodológica aos problemas em análise, foram testados métodos probabilísticos, com recurso à utilização do classificador *Naive Bayes*, métodos de regressão, com utilização do conceito de regressão logística, e métodos baseados em procura, com utilização de árvores e regras de decisão. Para além da previsão da apresentação ou não de resíduos à recolha (problema 1.1, com *output* binário Sim/Não) e do número total de contentores colocados à recolha (problema 2.1 – problema *multitarget* com *output* número total de contentores), procurou-se identificar as características que definem quando os clientes apresentam contentores à recolha (problema 1.2) e o número de contentores apresentados à recolha (problema 2.2.), apurando-se regras com recurso a algoritmos de regras de decisão.

A regressão constitui um dos métodos mais utilizados em análise preditiva, tendo como conceito base a predição de um valor, ou classe, de uma variável dependente, com base num conjunto de observações. Em problemas de classificação de variáveis categóricas, não é possível a utilização de técnicas de regressão linear, apropriadas para variáveis contínuas, sendo então necessário o recurso à regressão logística, com a aplicação de uma função discreta para a predição das categorias.

A utilização de métodos probabilísticos em problemas de classificação, baseados na aplicação do Teorema de *Bayes*, que pressupõe uma forte independência entre as variáveis, constituem métodos de fácil implementação e de modelação bastante eficiente (Gama et al., 2015). Estes métodos, para além de revelarem um bom desempenho numa variedade alargada de domínios, incluindo situações de clara dependência entre os atributos (Domingos e Pazzani, 1997, em Gama et al., 2015), são robustos à presença de ruídos e atributos irrelevantes (Kononenko, 1991, em Gama et al., 2015).

Segundo Gama (2015), os métodos de aprendizagem baseados na procura, procuram representar um espaço de possíveis soluções, com recurso à generalização de exemplos e a uma função de avaliação das hipóteses formuladas, construindo modelos baseados em árvores (árvores de decisão para problemas de classificação e árvores de regressão para problemas de regressão). Através de uma estratégia de dividir para conquistar, problemas complexos são divididos, de forma recursiva, em problemas mais simples. A avaliação da capacidade de um atributo discriminar as classes, é efetuada com recurso a diversas medidas, como por exemplo, o ganho de informação (baseado no conceito de entropia que traduz a dificuldade de predição do atributo alvo) ou o erro quadrático (em problemas de regressão). Gama (2015) refere a utilização alargada de algoritmos de árvores de decisão no meio académico e empresarial, destacando ainda como vantagens a sua eficiência e fácil interpretabilidade.

A interpretabilidade constitui um dos fatores críticos de sucesso na proposta de soluções para os problemas abordados na presente tese, procurando-se extrair informação relevante para suporte ao processo de gestão de circuitos de recolha e tomada de decisão. A utilização de algoritmos de classificação, se bem que capazes de gerar previsões de colocação de resíduos à recolha pelos clientes, podem constituir uma “caixa-negra”, uma vez que nem sempre proporcionam informação entendível dos critérios de classificação utilizados. Deste modo, recorreu-se à utilização de métodos de classificação baseados em regras para superação desta limitação. As regras esclarecem o motivo da decisão de classificação, evitando a ocorrência de conclusões imprevisíveis, ou de difícil compreensão (Dubitzky, 2013, in Włodzisław, 2013, in Tayel et al., 2013). Estes métodos consistem na geração de um modelo através de um conjunto de dados de treino, composto por um conjunto de regras, cada regra com antecedentes e uma consequência. Se os antecedentes forem satisfeitos, um exemplo será rotulado de acordo com o definido na consequência.

Para identificação das regras caracterizadoras da entrega de resíduos pelos clientes, foi desenvolvido um modelo geral, com o objetivo de fornecer ao técnico municipal responsável pelo agendamento dos circuitos de recolha, informação de fácil compreensão que traduza as situações em que o cliente apresenta resíduos para recolha e respetivo número de contentores. Esta informação, é particularmente relevante para suporte às operações de prospeção de novos clientes aderentes ao circuito de recolha PAP-NR, ou mesmo para o estabelecimento de novos circuitos, permitindo a previsão de uma rotina de recolha adequada ao novo cliente, ultrapassando assim situações de *cold start* por inexistência de dados históricos. Deste modo, foram testados os seguintes algoritmos para o apuramento de regras de decisão:

- W-JRip: Algoritmo baseado numa Redução Incremental Repetida para Produzir Redução de Erros. Utiliza uma estratégia de dividir para conquistar em que o conjunto de exemplos é dividido em subconjuntos, gera uma regra para um subconjunto, remove esse subconjunto e conquista os subconjuntos restantes (Cohen, 1995, e Furnkranz, 1999, in Tayel et al., 2013);
- W-Part: Este algoritmo é uma combinação do algoritmo C4.5 utilizado para a criação de árvores de decisão e o algoritmo base do JRip (Redução Incremental Repetida para Produzir Redução de Erros). Este algoritmo utiliza igualmente uma estratégia de dividir para conquistar em que, a partir de um conjunto de exemplos, e de forma iterativa, são criadas e removidas sub-árvores, que contém ramos inexplorados, sendo os nós da sub-árvore expandidos com base na entropia mínima (Tayel et al., 2013);
- W-Ridor: O algoritmo Ridor (*Ripple Down Rules*), tem como conceito base a geração de uma regra geral e, em seguida, a criação de exceções para essa regra. Se um exemplo satisfizer a premissa da regra geral, será classificado na classe principal, a menos que corresponda a uma das exceções, sendo então classificado nas classes associados às regras de exceção (Gaines e Compton, 1995, in Tayel et al., 2013);
- *Subgroup Discovery*: Operador do *software RapidMiner* que implementa uma procura exhaustiva de subgrupos, apurando regras que descrevam subconjuntos da população que sejam suficientemente grandes e estatisticamente incomuns;
- *Rule Induction*: Operador do *software RapidMiner*, com uma lógica de funcionamento semelhante ao algoritmo W-JRip, em que, partindo das classes menos prevalentes, o

algoritmo iterativamente cresce e remove as regras até que não existam exemplos positivos ou a taxa de erro seja maior que 50%;

- *Tree to Rules (Decision Tree)*: Operador do *software RapidMiner* que determina o conjunto de regras a partir de uma árvore de decisão.

Os modelos foram desenvolvidos com recurso ao *software RapidMiner Studio Educational* (versão 9.0.001), tendo sido utilizadas as extensões disponíveis para acesso à livreria do *software Weka*, nomeadamente, para utilização dos algoritmos W-JRip, W-Part, W-Ridor, W-OneR e W-J48.

Problema 1.1: Previsão da colocação de resíduos à recolha pelos clientes

Na abordagem ao problema da previsão da colocação ou não de resíduos à recolha pelos clientes, desenvolveu-se um modelo preditivo para cada cliente. Deste modo, as variáveis do *data set* “Nome_cliente”, “Horario”, “Centralidade” e “Atividade_revista”, foram excluídas do modelo uma vez que caracterizam inequivocamente o cliente. Procedeu-se ainda a uma filtragem dos exemplos do *data set*, tendo sido seleccionados uma amostra de 5 clientes, 1 por cada área de negócio, com maior número de contentores apresentados à recolha, de um total de 141 clientes da base de dados de clientes aderentes ao serviço PAP-NR durante o ano de 2017. O atributo “Entrega_RU(1/0)” foi definido como atributo classe, com dois *outputs* (1 = Sim e 0 = Não).

Na tabela seguinte, é apresentado o *ranking* dos resultados obtidos para cada classificador, tendo como base de avaliação do desempenho preditivo o recurso a validação cruzada com 10 iterações, e as medidas de Taxa de Acerto e *F-score*:

Algoritmos	Taxa Acerto		<i>F-score</i>
	Valor	Desvio Padrão	
W-J48 (Weka)	97,45%	2,49%	98,60%
Logistic Regression (RapidMiner Studio Core)	96,94%	2,77%	98,35%
Decision Tree (RapidMiner Studio Core)	96,93%	2,66%	98,34%
Naive Bayes (RapidMiner Studio Core)	96,82%	2,58%	98,27%
Random Forest (RapidMiner Studio Core)	95,59%	3,44%	97,48%
W-JRip (Weka)	94,96%	2,96%	96,90%
W-OneR (Weka)	94,12%	3,29%	96,10%
Gradient Boosted Trees (RapidMiner Studio Core)	93,69%	5,25%	96,58%

Tabela 10: *Ranking* das medidas de avaliação Taxa de Acerto e *F-score* (valores médios) obtidas por algoritmos de classificação – problema 1.1, obtido via *RapidMiner* e trabalhado em *Microsoft Excel*

Foram obtidos níveis de desempenho bastante elevados com todos os algoritmos, tendo sido obtido o melhor desempenho preditivo com a utilização do algoritmo W-J48, que regista os maiores valores médios de Taxa de Acerto e *F-score*, com 97,45% e 98,6%, respetivamente. Os algoritmos *Logistic Regression*, *Decision Tree* e *Naive Bayes*, seguem-se ao W-J48, apresentando Taxas de Acerto e níveis de *F-score* bastante similares na ordem dos 97% e 98%, respetivamente. No capítulo “4.6 Avaliação”, é realizada em maior detalhe a avaliação dos classificadores utilizados, e justificada a opção pela adoção do algoritmo W-J48 na construção do modelo preditivo.

Nas figuras seguintes, é apresentado o modelo desenvolvido com recurso ao *software RapidMiner Studio*:

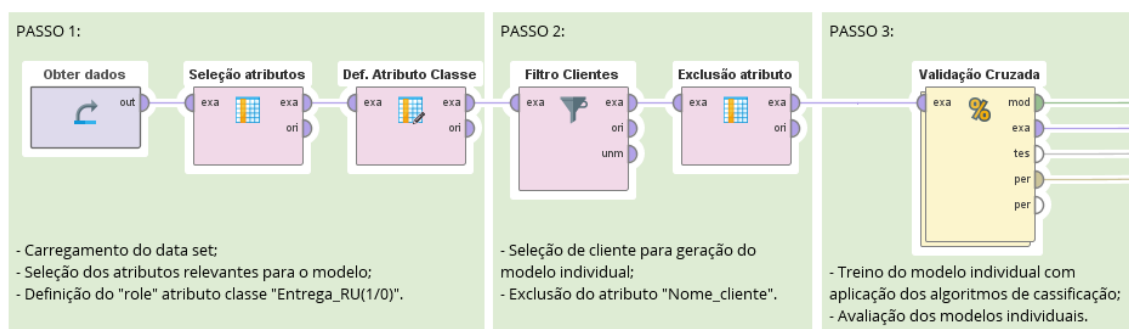


Figura 27: Modelo preditivo implementado problema 1.1, *software RapidMiner*

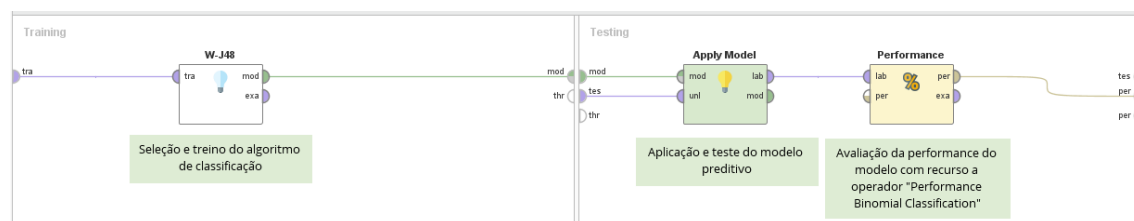


Figura 28: Modelo preditivo problema 1.1 – detalhe do processo de Validação Cruzada, *software RapidMiner*

O algoritmo W-J48, recorre à utilização algoritmo de C4.5 para criação da árvore decisão, a partir dos exemplos do *data set*, selecionando em cada nó, o atributo que melhor separa os exemplos, enquadrados nas classes em análise. A avaliação da divisão, que indica a medida em que um dado atributo discrimina as classes, é efetuado com recurso ao ganho de informação, que possui como conceito fundamental subjacente, o conceito de entropia. A entropia mede a aleatoriedade de uma variável aleatória, isto é, a dificuldade de predição do atributo alvo (Gama et al., 2015). O atributo com o maior ganho de informação (partição que proporciona a maior redução na entropia) é escolhido para tomar a decisão de partição de um conjunto de exemplos.

Na figura seguinte, é apresentada a árvore de decisão obtida com o algoritmo W-J48, para um dos clientes selecionados na amostra:

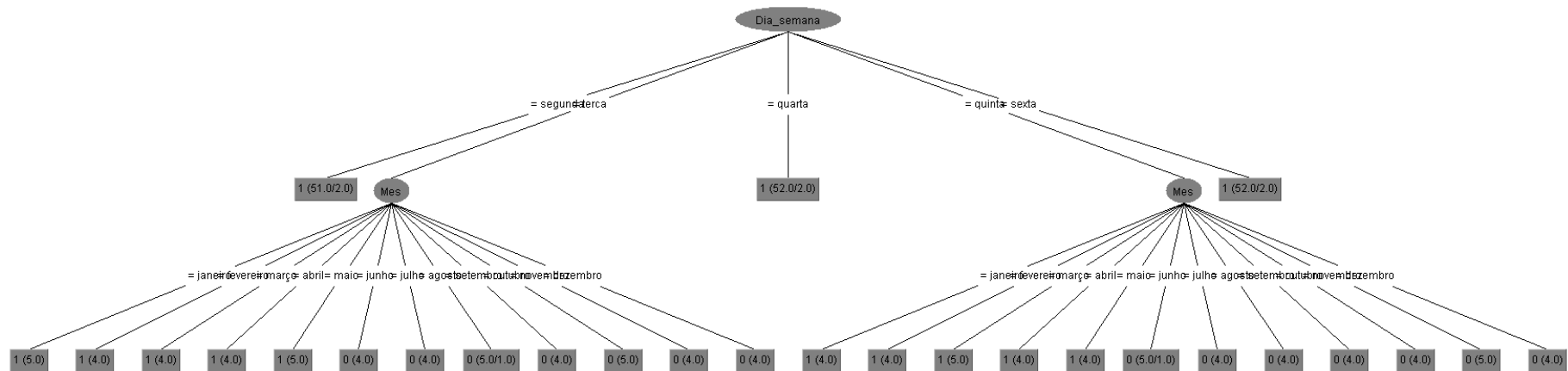


Figura 29: Árvore de decisão W-J48 obtida para um cliente – problema 1.1, obtida via *software RapidMiner*

Na construção da árvore de decisão para o cliente selecionado, é possível verificar que para partição dos exemplos do *data set*, o algoritmo considerou as variáveis “Dia_semana” e “Mes”.

Problema 1.2: Identificação de regras caracterizadoras da entrega de resíduos pelos clientes

Para identificação das regras caracterizadoras da entrega de resíduos pelos clientes e, conforme já referido anteriormente, foi desenvolvido um modelo geral, pelo que foi excluída apenas a variável “Nome_cliente”.

Na tabela seguinte são apresentados o número de regras geradas por cada algoritmo e o *ranking* dos resultados obtidos, tendo como referencial de avaliação as medidas de Taxa de Acerto e *F-score*:

Algoritmos	Nº regras geradas	Taxa Acerto		<i>F-score</i>
		Valor	Desvio Padrão	
W-Part (Weka)	83	76,59%	0,79%	83,58%
Tree to Rules (Decision Tree) (RapidMiner Studio Core)	418	76,02%	0,78%	83,18%
W-Ridor (Weka)	145	74,03%	1,60%	81,66%
W-JRip (Weka)	29	73,46%	1,00%	81,53%
Rule Induction (RapidMiner Studio Core)	42	72,03%	3,07%	81,70%
Subgroup Discovery (RapidMiner Studio Core)	20	44,13%	1,27%	36,00%

Tabela 11: N.º regras geradas e *ranking* das medidas de avaliação Taxa de Acerto e *F-score* obtidas por algoritmos geradores de regras de decisão – problema 1.2, obtido via *RapidMiner* e trabalhado em *Microsoft Excel*

Para a geração e interpretação de regras de decisão, recorreu-se à utilização do algoritmo W-JRip, uma vez que apresenta níveis de *performance* bastante próximos dos algoritmos melhores classificados, gerando um número de regras adequado a uma efetiva utilização e compreensão dos utilizadores, em oposição ao elevado número de regras gerados em outros algoritmos testados. No capítulo “4.6 Avaliação”, é realizada em maior detalhe a avaliação dos algoritmos utilizados.

Na figura seguinte, são apresentadas 3 das 29 regras geradas pelo algoritmo W-JRip, retratando as situações em que o cliente não apresente resíduos à recolha (a totalidade das regras geradas podem ser consultadas no Anexo C):

```
(Dia_semana = terca) and (Mes = novembro ) and (Horario = Normal) and (Centralidade = 1) => Entrega_ RU(1/0)=0
(Dia_semana = terca) and (Mes = dezembro ) and (Horario = Normal) and (Centralidade = 1) => Entrega_ RU(1/0)=0
(Dia_semana = quinta) and (Mes = setembro ) and (Atividade_revista = Restauração) => Entrega_ RU(1/0)=0
```

Figura 30: Exemplos de regras de decisão geradas pelo algoritmo W-JRip – problema 1.2, obtido via *RapidMiner*

Pela análise das duas primeiras regras apresentadas, é possível concluir que em estabelecimentos comerciais com nível de centralidade 1, e regime de horário normal (7 a 12 horas de funcionamento), à terça-feira e nos meses de novembro e dezembro, o cliente não

possui resíduos para a recolha, pelo que este ponto de recolha não deverá ser considerado no agendamento do circuito de recolha. Interpretando a terceira regra apresentada, é possível verificar que um estabelecimento comercial da área de atividade da Restauração, no mês de setembro e à quinta-feira, não possui resíduos para recolha.

Refira-se ainda que, com a aplicação de filtros de atributos no modelo desenvolvido, será possível o apuramento de análises complementares, nomeadamente a geração de regras por tipo de atividade, nível de centralidade do estabelecimento comercial do cliente, bem como gerar regras cliente a cliente. No Anexo D, são disponibilizados exemplos de regras extraídas para estas situações.

Nas figuras seguintes, é apresentado o modelo desenvolvido com utilização do algoritmo W-JRip e recurso ao *software RapidMiner Studio*:

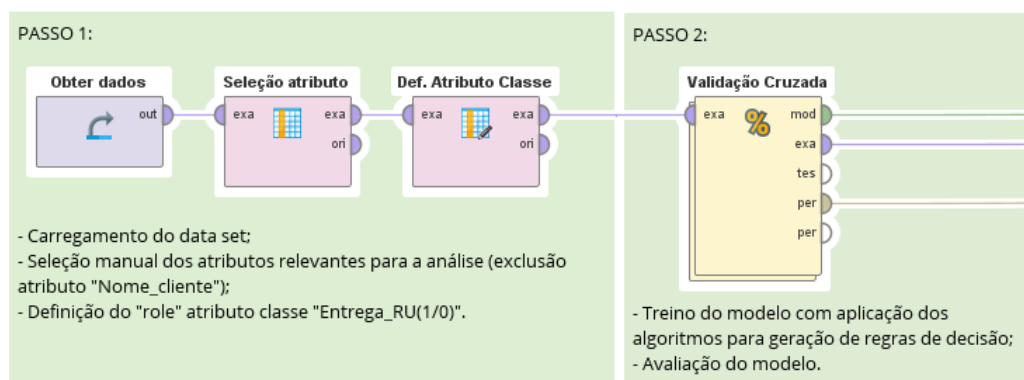


Figura 31: Modelo de geração de regras de decisão implementado – problema 1.2, *software RapidMiner*

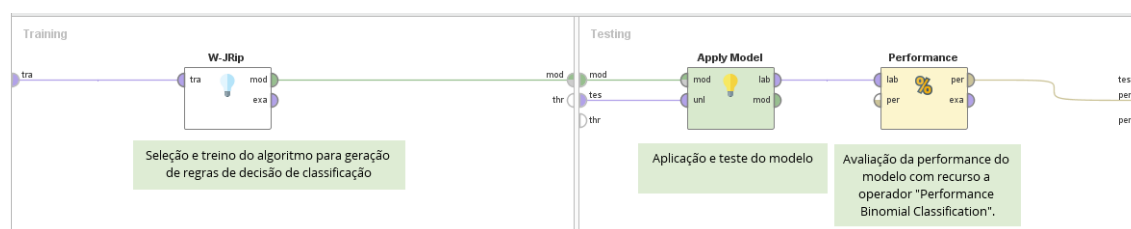


Figura 32: Modelo de geração de regras de decisão – problema 1.2 – detalhe do processo de Validação Cruzada, *software RapidMiner*

Problema 2.1: Previsão do número total de contentores colocados à recolha pelos clientes

À semelhança do assegurado no problema 1.1, na abordagem ao problema da previsão do número total de contentores colocados à recolha (somatório dos 3 fluxos de material – Papel/ Cartão, Embalagens e Vidro), desenvolveu-se um modelo preditivo para cada cliente, com exclusão das variáveis do *data set* que caracterizam inequivocamente o cliente (“Nome_cliente”, “Horario”, “Centralidade” e “Atividade_revista”), com amostragem de 5 clientes, 1 por cada área de negócio, com o maior número de contentores apresentados à recolha. Foram ainda utilizados os algoritmos testados no problema 1.1, com idêntica metodologia de validação, com recurso a validação cruzada com 10 iterações. Na tabela seguinte, é apresentado o *ranking* dos resultados obtidos para cada classificador, tendo como base de avaliação do desempenho preditivo do modelo, as medidas de Taxa de Acerto e *F-score*:

Algoritmos	Taxa Acerto		<i>F-score</i>
	Valor	Desvio Padrão	
Gradient Boosted Trees (RapidMiner Studio Core)	91,23%	5,22%	93,79%
W-J48 (Weka)	89,10%	5,90%	92,07%
Decision Tree (RapidMiner Studio Core)	88,08%	6,73%	91,98%
Random Forest (RapidMiner Studio Core)	83,89%	7,29%	87,27%
Logistic Regression (RapidMiner Studio Core)	83,03%	5,40%	83,39%
W-JRip (Weka)	81,88%	8,15%	83,81%
Naive Bayes (RapidMiner Studio Core)	81,22%	5,93%	81,12%
W-OneR (Weka)	73,50%	6,87%	82,46%

Tabela 12: *Ranking* das medias de avaliação Taxa de Acerto e *F-score* (valores médios) obtidos por algoritmos de classificação – problema 2.1, obtido via *RapidMiner* e trabalhado em *Microsoft Excel*

Foram obtidos níveis de Taxa de Acerto e *F-score* elevados, mas inferiores aos registados no problema de 1.1, referente à previsão de colocação de resíduos à recolha pelo cliente. O algoritmo *Gradient Boosted Tree*, que apresentava a pior *performance* do conjunto de algoritmos testados no problema 1.1, com um nível médio de Taxa de Acerto de 93,69%, apresenta neste problema de classificação *multitarget*, o melhor desempenho preditivo, registando os maiores valores médios de Taxa de Acerto e *F-score*, com 91,23% e 93,79%, respetivamente. Segue-se o algoritmo W-J48, que apresenta níveis de desempenho similares, com níveis médios de Taxa de Acerto e *F-score*, de 89,1% e 92,07%, respetivamente. No capítulo “4.6 Avaliação”, é realizada em maior detalhe a avaliação dos classificadores utilizados.

Para a construção do modelo no *software RapidMiner Studio*, recorreu-se à utilização do algoritmo *Gradient Boosted Tree*, conforme ilustrado nas figuras seguintes:

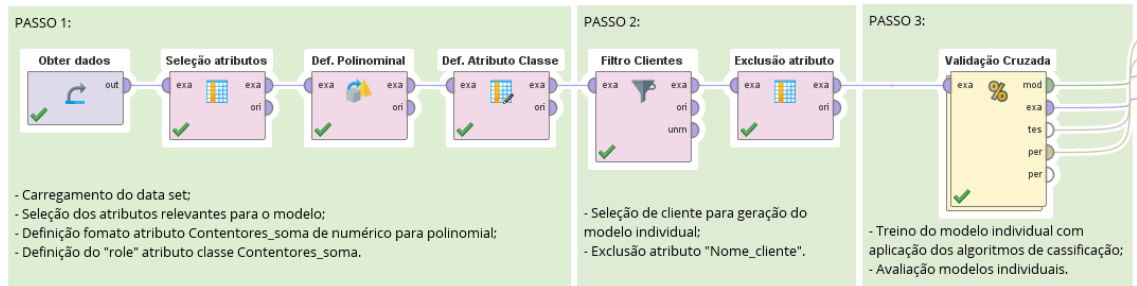


Figura 33: Modelo preditivo implementado problema 2.1, *software RapidMiner*

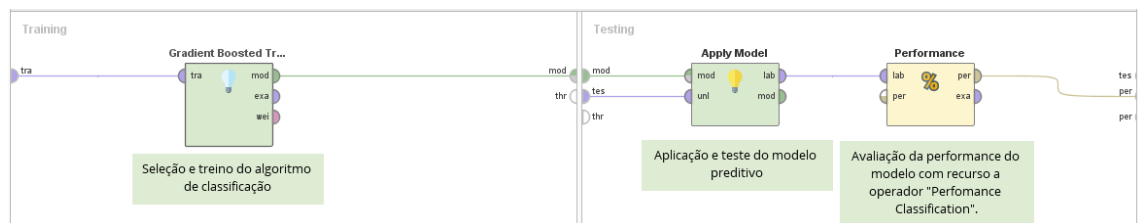


Figura 34: Modelo preditivo problema 2.1 – detalhe do processo de Validação Cruzada, *software RapidMiner*

Apesar do algoritmo *Gradient Boosted Tree* ter sido selecionado para a construção do modelo preditivo, para representação da árvore de decisão obtida para um dos clientes selecionados na amostra, a título exemplificativo, optou-se pela utilização do do W-J48. Pela análise da árvore obtida é possível constatar a normal recolha de 3 contentores que, em dias úteis após fim-de-semana, feriados ou outras datas festivas, poderá ascender a 6 contentores., dependendo do mês do ano.

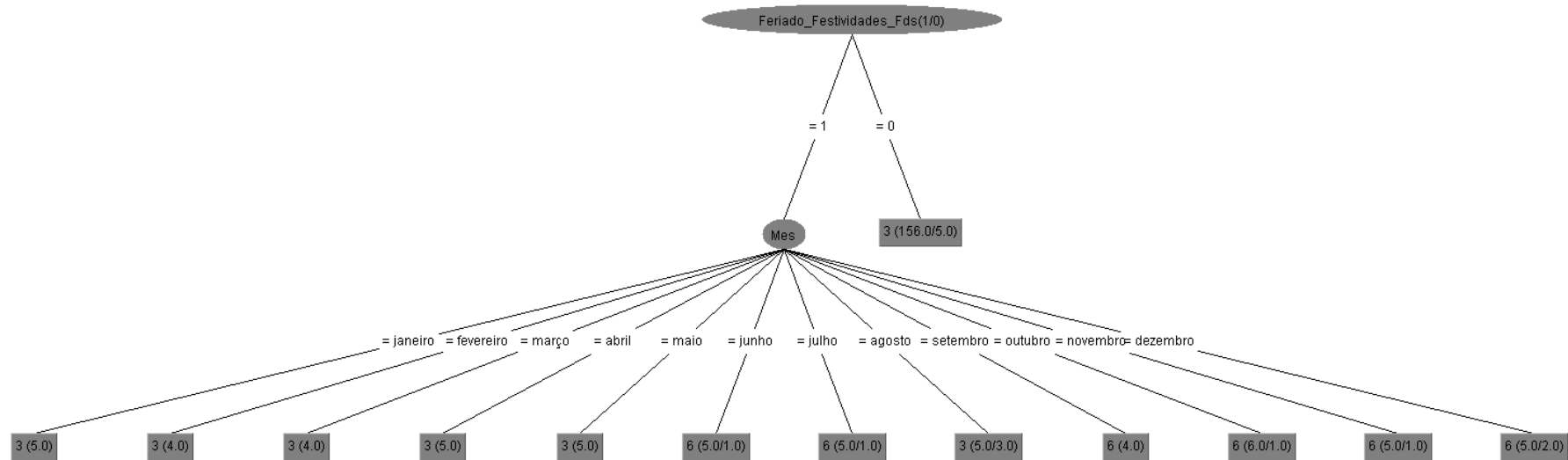


Figura 35: Árvore de decisão W-J48 obtida para um cliente – problema 2.1, via *software RapidMiner*

É possível verificar que para partição dos exemplos do *data set*, o algoritmo considerou as variáveis “Feriado_Festividades_Fds(1/0)” e “Mes”.

Problema 2.2: Identificação de regras caracterizadoras do número total de contentores que os clientes colocam à recolha

Após identificação das regras que caracterizam a ocorrência da entrega de resíduos pelos clientes, complementarmente, é importante a compreensão dos quantitativos que os clientes apresentam à recolha. A identificação de regras que caracterizam as situações que os clientes apresentam um determinado número de contentores à recolha, revelam-se igualmente de particular importância para a prospeção de novos clientes aderentes ou no estabelecimento de um novo circuito de recolha PAP-NR, nomeadamente, no dimensionamento e afetação de recursos (viaturas, contentores e recursos humanos). Deste modo, foi desenvolvido um modelo geral, com exclusão da variável “Nome_cliente”, tendo sido testados os 6 algoritmos utilizados no problema 1.2 para o apuramento das regras de decisão. Manteve-se a metodologia de avaliação do modelo desenvolvido, com recurso a validação cruzada com 10 iterações, com processo de amostragem estratificada, por forma a garantir a proporção de exemplos de cada classe, idêntica à observada no *data set*.

Na tabela seguinte, são apresentados o número de regras geradas por cada algoritmo e o *ranking* dos resultados obtidos, com recurso a validação cruzada com 10 iterações, tendo como referencial de avaliação as medidas de Taxa de Acerto e *F-score*:

Algoritmos	Nº regras geradas	Taxa Acerto		<i>F-score</i>
		Valor	Desvio Padrão	
W-Part (Weka)	110	85,23%	0,55%	91,60%
W-JRip (Weka)	13	84,78%	0,44%	91,42%
Tree to Rules (Decision Tree) (RapidMiner Studio Core)	218	84,87%	0,64%	91,39%
W-Ridor (Weka)	490	84,52%	0,62%	91,28%
Rule Induction (RapidMiner Studio Core)	29	84,30%	0,34%	91,17%
Subgroup Discovery (RapidMiner Studio Core)	60	82,39%	0,04%	90,35%

Tabela 13: N.º regras geradas e *ranking* do nível de Taxa de Acerto e *F-score* obtidos por algoritmos geradores de regras de decisão – problema 2.2, obtido via *RapidMiner* e trabalhado em *Microsoft Excel*

Refira-se que, simulações iniciais do modelo desenvolvido se traduziram em taxas máximas de acerto de aproximadamente 66%. Deste modo, foram efetuadas e testadas novas abordagens ao modelo, tendo sido atingidas as melhores *performances* com a discretização do atributo classe “Contentores_soma” em três classes, nomeadamente:

- [1,3] – ocorrência de recolha de 1 a 3 contentores no estabelecimento comercial do cliente, situação mais frequente que ocorre em 82,4% das recolhas realizadas;
- [4,6] – ocorrência de recolha de 4 a 6 contentores, verificada em 16,7% das recolhas;

- $[7, +\infty]$ – ocorrências pontuais e extremas de ocorrência da colocação à recolha de 7 contentores ou superior, correspondendo a apenas 0,9% das recolhas realizadas.

Os algoritmos testados apresentam níveis de *performance* bastante similares entre si. Para a geração e interpretação das regras de decisão, e à semelhança do problema 1.2, optou-se pela utilização do algoritmo W-JRip, tendo em conta o número adequado de regras a uma efetiva utilização e compreensão, gerando assim valor para o gestor responsável pela definição e agendamento de circuitos de recolha (foram gerados pelo W-Rip 13 regras, face às 110 regras geradas pelo W-Part). No capítulo “4.6 Avaliação”, é realizada em maior detalhe a avaliação dos algoritmos utilizados.

Na figura seguinte, são apresentadas 3 das 13 regras geradas pelo algoritmo W-JRip, caracterizadoras das situações em que o cliente apresenta um número de contentores à recolha, diferente da primeira classe definida, referente à ocorrência entre 1 a 3 contentores (a totalidade das regras geradas podem ser consultadas no Anexo E):

```
(Centralidade = 3) and (Horario = Alargado) and (Atividade_revista = Restauração) and (Dia_semana = terca) => Contentores_soma=[7;+inf.]  
(Atividade_revista = Restauração) and (Centralidade = 2) and (Dia_semana = quarta) and (Horario = Normal) => Contentores_soma=[4-6]  
(Atividade_revista = Restauração) and (Centralidade = 2) and (Horario = Alargado) and (Mes = agosto ) => Contentores_soma=[4-6]
```

Figura 36: Exemplos de regras de decisão geradas pelo algoritmo W-JRip – problema 2.2, obtido via *RapidMiner*

Na primeira regra é possível constatar a ocorrência de apresentação à recolha de um número de contentores igual ou superior a 7, às terças-feiras, em estabelecimentos comerciais da área de atividade de Restauração, com nível de centralidade 3 e horário de funcionamento alargado. A segunda regra define que às quartas-feiras, são recolhidos 4 a 6 contentores em estabelecimentos da área de Restauração, com nível de centralidade 2 e horário de funcionamento normal. Por último, na terceira regra, é possível verificar que em estabelecimentos da área de Restauração, situados em zona enquadrada no nível de centralidade 2, com horário largado e durante o mês de agosto, são apresentados à recolha pelos clientes, 4 a 6 contentores.

À semelhança do referido no modelo desenvolvido para abordagem ao problema 1.2, com a aplicação de filtros de atributos, será possível o apuramento de análises complementares para uma compreensão mais alargada das regras pelo gestor responsável pela gestão dos circuitos de recolha, nomeadamente a geração de regras por tipo de atividade, nível de centralidade do estabelecimento comercial do cliente, bem como a geração de regras

cliente a cliente. No Anexo F, são disponibilizados exemplos de regras extraídas para estas situações.

Nas figuras seguintes, é apresentado o modelo desenvolvido com utilização do algoritmo W-JRip e recurso ao *software RapidMiner Studio*:

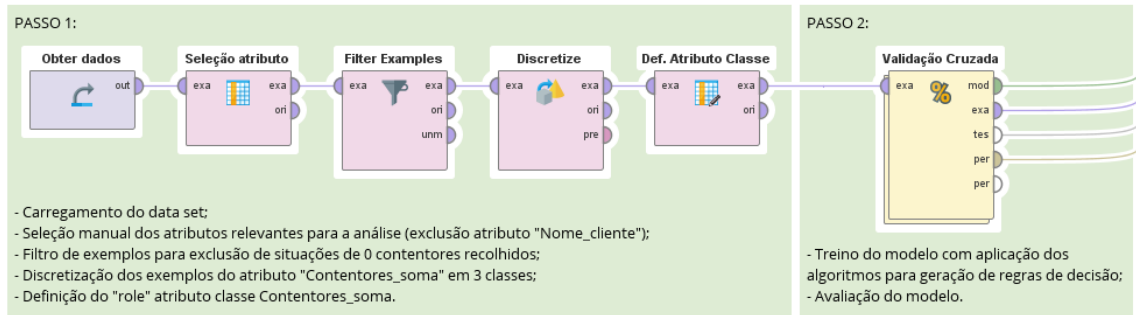


Figura 37: Modelo de geração de regras de decisão implementado – problema 2.2, *software RapidMiner*

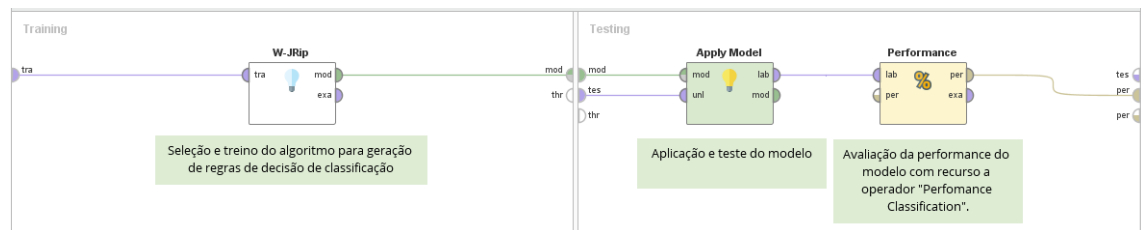


Figura 38: Modelo de geração de regras de decisão – problema 2.2 – detalhe do processo de Validação Cruzada, *software RapidMiner*

4.6. Avaliação

Na avaliação dos modelos preditivos desenvolvidos para abordagem aos problemas 1.1 e 2.1, e nos modelos de geração de regras de decisão desenvolvidos para os problemas 1.2 e 2.2, foi testada e avaliada a utilização de diferentes algoritmos de classificação, com base nos conjuntos de atributos selecionados e com recurso ao conceito de validação cruzada, que recorre à utilização de subconjuntos disjuntos do *data set* para treino do modelo e avaliação da predição, por forma a evitar a produção de estimativas otimistas (Monard e Baranaus, 2003, in Gama et al., 2015). Para o efeito, foi utilizado o operador *Cross-Validation* do *software RapidMiner*, tendo sido definida uma validação cruzada com 10 iterações, com processo de amostragem estratificada, por forma a garantir a proporção de exemplos de cada classe, idêntica à observada no *data set*.

Na avaliação do desempenho dos modelos recorreu-se ainda à utilização de Matriz de Confusão, a partir da qual é possível apurar diversas medidas de desempenho (Gama et al., 2015) tendo sido utilizadas as medidas Taxa de Acerto, Precisão, Sensibilidade e *F-score*, conforme ilustrado nas tabelas seguintes:

Classe Predita	Classe Verdadeira	
	Entrega RU=1 (Sim)	Entrega RU=0 (Não)
Entrega RU=1 (Sim)	VP	FP
Entrega RU=0 (Não)	FN	VN

Tabela 14: Matriz de Confusão

Medidas	Definição	Cálculo
Taxa Acerto	Performance agregada de todo o modelo de classificação (<i>accuracy</i>).	$(VP+VN)/(VP+VN+FP+FN)$
Precisão	Proporção de exemplos positivos classificados corretamente sobre o total de predições positivas. -> Dada a predição de uma dada classe pelo algoritmo, qual a probabilidade de se encontrar correta?	$VP/(VP+FP)$
Sensibilidade	Taxa de acerto na classe positiva - Proporção de exemplos positivos classificados corretamente sobre o total de exemplos verdadeiros. -> Dada uma classe, conseguirá o algoritmo detectá-la?	$VP/(VP+FN)$
<i>F-score</i>	Média harmónica da precisão e sensibilidade, em que uma pontuação de F1 atinge seu melhor valor em 1 (precisão e sensibilidade perfeitas) e a pior em 0.	$F1 = 2 \text{ (precisão} * \text{sensibilidade)} / \text{(precisão} + \text{sensibilidade)} = 2VP / (2VP + FP + FN)$

Tabela 15: Medidas de avaliação utilizadas

Nos casos particulares de problemas multiclasse, nomeadamente no problema 2, em que se procura prever o número de contentores colocados à recolha e apurar as regras que caracterizam estas ocorrências, as medidas de Precisão e Sensibilidade “podem ser facilmente generalizadas a problemas com mais de duas classes pela consideração de cada classe como positiva em relação ao conjunto das demais classes” (Gama, 2015, p.192).

Dado que o *data set* utilizado para a construção dos modelos propostos não é balanceado, a simples utilização da Taxa de Acerto como medida de avaliação poderá conduzir a resultados enganadores. Deste modo, para além de adoção de medidas complementares de análise como a Precisão e Sensibilidade, assume particular relevância a utilização do conceito de *F-score*, que combina ambas as medidas, ultrapassando ainda a limitação da sua interpretação individual.

A título de exemplo, um desempenho máximo de 100% de Precisão, não fornece informação do número de exemplos não classificados corretamente numa dada classe, traduzindo apenas todos os exemplos classificados como pertencentes à classe, realmente pertencem a essa classe. Por outro lado, um nível de Sensibilidade máximo, fornece apenas informação de que cada exemplo de uma dada classe, foi classificado como pertencente a essa classe, não fornecendo, contudo, informação do número de exemplos classificados incorretamente nessa classe (Gama et al., 2015). Na presente tese, foi adotada a medida *F-score* “*F1*”, que traduz a atribuição de idêntica ponderação (1) às medidas de Precisão e Sensibilidade.

Problema 1.1: Previsão da colocação de resíduos à recolha pelos clientes

Na tabela seguinte é apresentado o *ranking* da classificação do desempenho da capacidade preditiva dos 8 algoritmos testados, para a amostra de 5 clientes selecionada, agrupado por tipo de medida de avaliação:

Medida de avaliação: Taxa de Acerto e respetivo desvio padrão

Algoritmos	Cliente					Máx.	Média	Min.	Amplitude p-p.
	1	2	3	4	5				
W-J48 (Weka)	96,89%	98,20%	98,14%	97,39%	96,62%	98,20%	97,45%	96,62%	1,6 p.p.
	2,37%	1,80%	2,28%	2,61%	3,38%	3,38%	2,49%	1,80%	1,6 p.p.
Logistic Regression (RapidMiner Studio Core)	96,51%	97,10%	98,14%	96,34%	96,62%	98,14%	96,94%	96,34%	1,8 p.p.
	2,74%	2,20%	2,28%	3,36%	3,28%	3,36%	2,77%	2,20%	1,2 p.p.
Decision Tree (RapidMiner Studio Core)	96,51%	97,84%	97,66%	97,39%	95,24%	97,84%	96,93%	95,24%	2,6 p.p.
	2,14%	1,76%	2,34%	2,61%	4,47%	4,47%	2,66%	1,76%	2,7 p.p.
Naive Bayes (RapidMiner Studio Core)	96,12%	98,20%	95,78%	97,39%	96,62%	98,20%	96,82%	95,78%	2,4 p.p.
	1,79%	1,80%	3,33%	2,61%	3,38%	3,38%	2,58%	1,79%	1,6 p.p.
Random Forest (RapidMiner Studio Core)	87,58%	98,20%	98,14%	97,39%	96,62%	98,20%	95,59%	87,58%	10,6 p.p.
	7,11%	1,80%	2,28%	2,61%	3,38%	7,11%	3,44%	1,80%	5,3 p.p.
W-JRip (Weka)	84,91%	98,20%	97,66%	97,39%	96,62%	98,20%	94,96%	84,91%	13,3 p.p.
	4,66%	1,80%	2,34%	2,61%	3,38%	4,66%	2,96%	1,80%	2,9 p.p.
W-OneR (Weka)	80,23%	98,20%	98,14%	97,39%	96,62%	98,20%	94,12%	80,23%	18 p.p.
	6,38%	1,80%	2,28%	2,61%	3,38%	6,38%	3,29%	1,80%	4,6 p.p.
Gradient Boosted Trees (RapidMiner Studio Core)	93,35%	94,25%	95,32%	93,74%	91,81%	95,32%	93,69%	91,81%	3,5 p.p.
	7,41%	3,28%	3,64%	5,12%	6,79%	7,41%	5,25%	3,28%	4,1 p.p.

Medida de avaliação: Precisão

Algoritmos	Cliente					Máx.	Média	Min.	Amplitude p-p.
	1	2	3	4	5				
W-J48 (Weka)	96,97%	98,20%	98,13%	97,40%	96,62%	98,20%	97,46%	96,62%	1,6 p.p.
Decision Tree (RapidMiner Studio Core)	96,95%	98,19%	98,12%	97,40%	96,58%	98,19%	97,45%	96,58%	1,6 p.p.
Logistic Regression (RapidMiner Studio Core)	96,48%	98,18%	98,13%	97,37%	96,62%	98,18%	97,36%	96,48%	1,7 p.p.
W-OneR (Weka)	96,13%	98,20%	98,13%	97,40%	96,62%	98,20%	97,30%	96,13%	2,1 p.p.
Naive Bayes (RapidMiner Studio Core)	96,00%	98,20%	98,09%	97,40%	96,62%	98,20%	97,26%	96,00%	2,2 p.p.
Gradient Boosted Trees (RapidMiner Studio Core)	93,17%	98,13%	99,02%	97,30%	96,45%	99,02%	96,81%	93,17%	5,9 p.p.
W-JRip (Weka)	92,35%	98,20%	98,12%	97,40%	96,62%	98,20%	96,54%	92,35%	5,9 p.p.
Random Forest (RapidMiner Studio Core)	86,49%	98,20%	98,13%	97,40%	96,62%	98,20%	95,37%	86,49%	11,7 p.p.

Medida de avaliação: Sensibilidade

Algoritmos	Cliente					Máx.	Média	Min.	Amplitude p-p.
	1	2	3	4	5				
Random Forest (RapidMiner Studio Core)	98,97%	100,00%	100,00%	100,00%	100,00%	100,00%	99,79%	98,97%	1 p.p.
W-J48 (Weka)	98,79%	100,00%	100,00%	100,00%	100,00%	100,00%	99,76%	98,79%	1,2 p.p.
Logistic Regression (RapidMiner Studio Core)	98,97%	98,90%	100,00%	98,93%	100,00%	100,00%	99,36%	98,90%	1,1 p.p.
Naive Bayes (RapidMiner Studio Core)	98,97%	100,00%	97,62%	100,00%	100,00%	100,00%	99,32%	97,62%	2,4 p.p.
Decision Tree (RapidMiner Studio Core)	98,45%	99,63%	99,52%	100,00%	98,60%	100,00%	99,24%	98,45%	1,6 p.p.
W-JRip (Weka)	87,11%	100,00%	99,52%	100,00%	100,00%	100,00%	97,33%	87,11%	12,9 p.p.
Gradient Boosted Trees (RapidMiner Studio Core)	98,45%	95,97%	96,19%	96,26%	95,10%	98,45%	96,39%	95,10%	3,4 p.p.
W-OneR (Weka)	76,80%	100,00%	100,00%	100,00%	100,00%	100,00%	95,36%	76,80%	23,2 p.p.

Medida de avaliação: F-Score

Algoritmos	Cliente					Máx.	Média	Min.	Amplitude p-p.
	1	2	3	4	5				
W-J48 (Weka)	97,87%	99,09%	99,06%	98,68%	98,28%	99,09%	98,60%	97,87%	1,2 p.p.
Logistic Regression (RapidMiner Studio Core)	97,71%	98,54%	99,06%	98,14%	98,28%	99,06%	98,35%	97,71%	1,3 p.p.
Decision Tree (RapidMiner Studio Core)	97,69%	98,90%	98,82%	98,68%	97,58%	98,90%	98,34%	97,58%	1,3 p.p.
Naive Bayes (RapidMiner Studio Core)	97,46%	99,09%	97,85%	98,68%	98,28%	99,09%	98,27%	97,46%	1,6 p.p.
Random Forest (RapidMiner Studio Core)	92,31%	99,09%	99,06%	98,68%	98,28%	99,09%	97,48%	92,31%	6,8 p.p.
W-JRip (Weka)	89,65%	99,09%	98,82%	98,68%	98,28%	99,09%	96,90%	89,65%	9,4 p.p.
Gradient Boosted Trees (RapidMiner Studio Core)	95,74%	97,04%	97,58%	96,78%	95,77%	97,58%	96,58%	95,74%	1,8 p.p.
W-OneR (Weka)	85,38%	99,09%	99,06%	98,68%	98,28%	99,09%	96,10%	85,38%	13,7 p.p.

Tabela 16: *Ranking* dos algoritmos de classificação por medida de avaliação – problema 1.1, obtido via *RapidMiner* e trabalhado em *Microsoft Excel*

Foram obtidos níveis de Taxa de Acerto bastante elevados, com um máximo de 98,2%, atingido simultaneamente por 5 algoritmos, nomeadamente, W-J48, *Naive Bayes*, W-JRip, W-OneR e *Random Forest*, e um mínimo de 80,23%, verificado com a utilização do algoritmo W-OneR. Os algoritmos *Gradient Boosted Tree* e *Random Forest* apresentam uma elevada variabilidade nos valores de Taxa de Acerto obtidos, face aos demais algoritmos, registando desvios padrão máximos de 7,41% e 7,11%, respetivamente. A menor variabilidade é registada no algoritmo W-J48, seguida do *Naive Bayes* e *Decision Tree*, com valores médios de desvio-padrão de 2,49%, 2,58% e 2,66%, respetivamente. O classificador W-J48 apresenta a melhor *performance*, apresentando uma taxa de acerto média de 97,45%, e o mais baixo desvio-padrão do conjunto de algoritmos avaliados. Os algoritmos *Logistic Regression*, *Decision Tree* e *Naive Bayes*, seguem-se ao W-J48, apresentando taxas de acerto bastante similares na ordem dos 96,9%. Contudo, destaca-se o algoritmo *Naive Bayes* que apresenta o menor desvio padrão nos valores de taxa de acerto obtidos.

Complementarmente à avaliação da Taxa de Acerto dos algoritmos, é importante analisar a sua capacidade de não classificar aleatoriamente os exemplos para uma classe. Esta capacidade é avaliada com a obtenção de valores próximos de Precisão e Sensibilidade (Tayel, Reif e Dengel, 2013). Deste modo, um algoritmo com um valor de Precisão muito maior que a Sensibilidade, classificará a maioria dos exemplos na classe negativa, em que o cliente não apresenta resíduos à recolha ($\text{Entrega_RU}(1/0)=0$, isto é, Não). Apenas alguns exemplos serão corretamente classificados como ocorrência de colocação de resíduos à recolha pelos clientes ($\text{Entrega_RU}(1/0)=1$, isto é, Sim).

Os valores médios de Precisão obtidos pelos 8 algoritmos são bastante elevados e similares, apresentando o algoritmo *Gradient Boosted Trees* o valor máximo obtido de 99,02%, e o *Random Forest* o valor mínimo registado de 86,49%. Os algoritmos W-J48 e *Decision Tree* apresentam os maiores valores médios de Precisão, com valores de aproximadamente 97,5%. A medida de Sensibilidade, apresenta igualmente valores elevados e similares entre os algoritmos avaliados. Apenas no algoritmo *Gradient Boosted Trees* não é obtida o valor máximo de 100% de Sensibilidade. O valor mínimo de 76,8% é verificado com a aplicação do algoritmo W-OneR. Os algoritmos *Random Forest* e W-J48 apresentam os maiores valores médios de Sensibilidade, com valores de aproximadamente 99,8%.

Algoritmos	Ranking Taxa de Acerto	Taxa Acerto		F-score	Precisão	Sensibilidade	Amplitude p.p. (1) - (2)
		Valor	Desvio Padrão				
Gradient Boosted Trees (RapidMiner Studio Core)	8º	93,69%	5,25%	96,58%	96,81%	96,39%	0,4 p.p.
W-JRip (Weka)	6º	94,96%	2,96%	96,90%	96,54%	97,33%	0,8 p.p.
Decision Tree (RapidMiner Studio Core)	3º	96,93%	2,66%	98,34%	97,45%	99,24%	1,8 p.p.
W-OneR (Weka)	7º	94,12%	3,29%	96,10%	97,30%	95,36%	1,9 p.p.
Logistic Regression (RapidMiner Studio Core)	2º	96,94%	2,77%	98,35%	97,36%	99,36%	2 p.p.
Naive Bayes (RapidMiner Studio Core)	4º	96,82%	2,58%	98,27%	97,26%	99,32%	2,1 p.p.
W-J48 (Weka)	1º	97,45%	2,49%	98,60%	97,46%	99,76%	2,3 p.p.
Random Forest (RapidMiner Studio Core)	5º	95,59%	3,44%	97,48%	95,37%	99,79%	4,4 p.p.

Tabela 17: *Ranking* da amplitude entre as medidas de Precisão e Sensibilidade (valores médios) – problema 1.1, obtido via *RapidMiner* e trabalhado em *Microsoft Excel*

Analisando comparativamente os valores médios obtidos pelas medidas de Precisão e Sensibilidade para os 8 algoritmos, é possível ainda verificar que os algoritmos apresentam valores bastante próximos, traduzindo a sua capacidade de não classificar aleatoriamente um exemplo em uma determinada classe. Os algoritmos *Gradient Boosted Trees* e WJ-Rip com desempenhos inferiores a nível de Taxa de Acerto, apresentam as menores diferenças com 0,4 e 0,8 pontos percentuais, respetivamente.

Avaliando o desempenho dos algoritmos com recurso à medida *F-score*, são obtidos valores bastante elevados e similares entre o universo de algoritmos testados, verificando-se uma reduzida amplitude entre o valor médio máximo de 98,6% obtido com o W-J48, e o valor médio mínimo de 96,1%, registado com a utilização do algoritmo W-OneR. Os algoritmos *Logistic Regression*, *Decision Tree* e *Naive Bayes*, apresentam valores bastante próximos aos verificados no W-J48, com *F-scores* de aproximadamente 98%. Os elevados níveis de *F-score* verificados, traduzem a capacidade de o modelo classificar adequadamente os exemplos nas várias classes em análise.

Face ao exposto, optou-se pela utilização do algoritmo W-J48 na construção do modelo de classificação para previsão da colocação de resíduos à recolha pelos clientes. Apesar de todos os algoritmos apresentarem valores de Taxa de Acerto e *F-score* bastante elevados e similares, o algoritmo W-J48 apresenta os maiores valores médios de Taxa de Acerto (com o menor desvio padrão) e *F-score*, registando ainda uma reduzida amplitude entre as medidas de Precisão e Sensibilidade.

Problema 1.2: Identificação de regras caracterizadoras da entrega de resíduos pelos clientes

No modelo de geração de regras relativas à entrega de resíduos pelos clientes, o melhor desempenho é obtido com a utilização do algoritmo W-Part, registrando a maior Taxa de Acerto de 76,59%, com uma baixa variabilidade, registrando um desvio padrão de apenas 0,79%. Com este algoritmo é ainda obtido o maior nível de *F-score* com 83,58% e uma reduzida amplitude de 6,9 p.p., entre as medidas de Precisão e Sensibilidade.

O operador do *RapidMiner*, *Subgroup Discovery*, apresenta a pior *performance* com uma Taxa de Acerto de apenas 44,13%, registrando ainda níveis bastante baixos de Sensibilidade de 23,09%, traduzindo a sua menor capacidade para previsão e identificação de ocorrências de entrega dos contentores à recolha pelos clientes ($\text{Entrega_RU}(1/0)=1$, isto é, Sim). O operador do *Rapid Miner*, *Rule Induction*, com o maior nível de Sensibilidade de 91,32%, revela uma maior propensão para classificação exemplos como pertencentes à classe positiva.

Na tabela seguinte são apresentados os resultados gerados pelos algoritmos de geração de regras de decisão:

Algoritmos	Nº regras geradas	Taxa Acerto		<i>F-score</i>	Precisão (1)	Sensibilidade (2)	Amplitude p.p. (1)-(2)
		Valor	Desvio Padrão				
W-Part (Weka)	83	76,59%	0,79%	83,58%	80,30%	87,15%	6,9 p.p.
Tree to Rules (Decision Tree) (RapidMiner Studio Core)	418	76,02%	0,78%	83,18%	79,92%	86,71%	6,8 p.p.
W-Ridor (Weka)	145	74,03%	1,60%	81,66%	78,96%	84,55%	5,6 p.p.
W-JRip (Weka)	29	73,46%	1,00%	81,53%	77,74%	85,71%	8 p.p.
Rule Induction (RapidMiner Studio Core)	42	72,03%	3,07%	81,70%	73,91%	91,32%	17,4 p.p.
Subgroup Discovery (RapidMiner Studio Core)	20	44,13%	1,27%	36,00%	81,62%	23,09%	58,5 p.p.

Tabela 18: *Ranking* dos resultados obtidos pelos algoritmos de regras de decisão – problema 1.2, obtido via *RapidMiner* e trabalhado em *Microsoft Excel*

Os algoritmos W-Part, *Tree to Rules* e W-Ridor, apesar de apresentarem os melhores níveis de *performance* do universo de algoritmos testados, geram um elevado número de regras. Tendo em conta o propósito da criação de regras de decisão, em gerar informação facilmente interpretável e útil ao processo de gestão de circuitos de recolha e suporte à tomada de decisão, optou-se pela utilização do algoritmo W-JRip, uma vez que apresenta níveis de *performance* próximos dos algoritmos melhores classificados, gerando um número de regras adequado a uma efetiva utilização e compreensão dos utilizadores, em oposição ao elevado número de regras gerados em outros algoritmos testados.

Problema 2.1: Previsão do número total de contentores colocados à recolha pelos clientes

Na tabela seguinte é apresentado o *ranking* de classificação do desempenho da capacidade preditiva dos algoritmos testados, para a amostra de 5 clientes selecionada, agrupado por tipo de medida de avaliação:

Medida de avaliação: Taxa de Acerto e respetivo desvio padrão

Algoritmos	Cliente					Máx.	Média	Min.	Amplitude p.p.
	1	2	3	4	5				
Gradient Boosted Trees (RapidMiner Studio Core)	93,80%	84,15%	95,37%	89,58%	93,24%	95,37%	91,23%	84,15%	11,2 p.p.
	5,54%	5,86%	2,88%	6,66%	5,17%	6,66%	5,22%	2,88%	3,8 p.p.
W-J48 (Weka)	91,86%	82,33%	91,65%	86,42%	93,24%	93,24%	89,10%	82,33%	10,9 p.p.
	4,72%	7,57%	5,67%	6,37%	5,17%	7,57%	5,90%	4,72%	2,9 p.p.
Decision Tree (RapidMiner Studio Core)	89,12%	79,81%	91,60%	88,00%	91,86%	91,86%	88,08%	79,81%	12,1 p.p.
	5,49%	9,18%	4,07%	9,03%	5,90%	9,18%	6,73%	4,07%	5,1 p.p.
Random Forest (RapidMiner Studio Core)	80,62%	71,23%	92,12%	82,26%	93,24%	93,24%	83,89%	71,23%	22 p.p.
	7,37%	9,27%	5,75%	8,88%	5,17%	9,27%	7,29%	5,17%	4,1 p.p.
Logistic Regression (RapidMiner Studio Core)	87,60%	57,58%	95,37%	85,42%	89,19%	95,37%	83,03%	57,58%	37,8 p.p.
	5,12%	5,96%	2,88%	4,50%	8,55%	8,55%	5,40%	2,88%	5,7 p.p.
W-JRip (Weka)	77,45%	62,17%	90,69%	85,87%	93,24%	93,24%	81,88%	62,17%	31,1 p.p.
	9,16%	12,42%	5,86%	8,16%	5,17%	12,42%	8,15%	5,17%	7,3 p.p.
Naive Bayes (RapidMiner Studio Core)	88,34%	48,53%	94,42%	82,24%	92,57%	94,42%	81,22%	48,53%	45,9 p.p.
	6,53%	5,79%	2,76%	8,29%	6,28%	8,29%	5,93%	2,76%	5,5 p.p.
W-OneR (Weka)	62,77%	50,32%	83,68%	77,50%	93,24%	93,24%	73,50%	50,32%	42,9 p.p.
	7,64%	8,22%	4,68%	8,64%	5,17%	8,64%	6,87%	4,68%	4 p.p.

Medida de avaliação: Precisão

Algoritmos	Cliente					Máx.	Média	Min.	Amplitude p.p.
	1	2	3	4	5				
Gradient Boosted Trees (RapidMiner Studio Core)	95,69%	86,42%	96,26%	92,13%	93,26%	96,26%	92,75%	86,42%	9,8 p.p.
Decision Tree (RapidMiner Studio Core)	89,43%	81,40%	96,09%	91,11%	93,26%	96,09%	90,26%	81,40%	14,7 p.p.
W-J48 (Weka)	91,74%	81,82%	94,57%	84,85%	93,26%	94,57%	89,25%	81,82%	12,7 p.p.
Random Forest (RapidMiner Studio Core)	85,71%	69,88%	93,68%	81,82%	93,26%	93,68%	84,87%	69,88%	23,8 p.p.
Logistic Regression (RapidMiner Studio Core)	86,61%	54,24%	97,30%	85,11%	93,02%	97,30%	83,26%	54,24%	43,1 p.p.
W-OneR (Weka)	94,12%	46,88%	94,08%	83,95%	93,26%	94,12%	82,46%	46,88%	47,2 p.p.
W-JRip (Weka)	83,04%	55,17%	92,23%	86,96%	93,26%	93,26%	82,13%	55,17%	38,1 p.p.
Naive Bayes (RapidMiner Studio Core)	89,57%	44,64%	97,27%	79,61%	93,26%	97,27%	80,87%	44,64%	52,6 p.p.

Medida de avaliação: Sensibilidade

Algoritmos	Cliente					Máx.	Média	Min.	Amplitude p.p.
	1	2	3	4	5				
W-J48 (Weka)	96,52%	88,89%	96,13%	96,55%	97,65%	97,65%	95,15%	88,89%	8,8 p.p.
Gradient Boosted Trees (RapidMiner Studio Core)	96,52%	86,42%	99,45%	94,25%	97,65%	99,45%	94,86%	86,42%	13 p.p.
Decision Tree (RapidMiner Studio Core)	95,65%	86,42%	95,03%	94,25%	97,65%	97,65%	93,80%	86,42%	11,2 p.p.
Random Forest (RapidMiner Studio Core)	88,70%	71,60%	98,34%	93,10%	97,65%	98,34%	89,88%	71,60%	26,7 p.p.
Logistic Regression (RapidMiner Studio Core)	95,65%	39,51%	99,45%	91,95%	94,12%	99,45%	84,14%	39,51%	59,9 p.p.
W-JRip (Weka)	80,87%	59,26%	98,34%	91,95%	97,65%	98,34%	85,61%	59,26%	39,1 p.p.
Naive Bayes (RapidMiner Studio Core)	89,57%	30,86%	98,34%	94,25%	97,65%	98,34%	82,13%	30,86%	67,5 p.p.
W-OneR (Weka)	41,74%	37,04%	87,85%	78,16%	97,65%	97,65%	68,49%	37,04%	60,6 p.p.

Medida de avaliação: F-score

Algoritmos	Cliente					Máx.	Média	Min.	Amplitude p.p.
	1	2	3	4	5				
Gradient Boosted Trees (RapidMiner Studio Core)	96,10%	86,42%	97,83%	93,18%	95,40%	97,83%	93,79%	86,42%	11,4 p.p.
W-J48 (Weka)	94,07%	85,21%	95,34%	90,32%	95,40%	95,40%	92,07%	85,21%	10,2 p.p.
Decision Tree (RapidMiner Studio Core)	92,44%	83,83%	95,56%	92,66%	95,40%	95,56%	91,98%	83,83%	11,7 p.p.
Random Forest (RapidMiner Studio Core)	87,18%	70,73%	95,96%	87,10%	95,40%	95,96%	87,27%	70,73%	25,2 p.p.
W-JRip (Weka)	81,94%	57,14%	95,19%	89,39%	95,40%	95,40%	83,81%	57,14%	38,3 p.p.
Logistic Regression (RapidMiner Studio Core)	90,91%	45,71%	98,36%	88,40%	93,57%	98,36%	83,39%	45,71%	52,6 p.p.
Naive Bayes (RapidMiner Studio Core)	89,57%	36,50%	97,80%	86,32%	95,40%	97,80%	81,12%	36,50%	61,3 p.p.
W-OneR (Weka)	57,83%	41,38%	90,86%	80,95%	95,40%	95,40%	73,28%	41,38%	54 p.p.

Tabela 19: *Ranking* dos algoritmos de classificação por medida de avaliação – problema 2.1, obtido via *RapidMiner* e trabalhado em *Microsoft Excel*

Face ao modelo desenvolvido para previsão da colocação de resíduos à recolha pelos clientes – problema 1.1, verifica-se uma ligeira degradação dos níveis de *performance*, resultante da sua natureza *multitarget*, com maior número de classes alvo. Contudo, continuam a verificar-se Taxas de Acerto elevadas, com um máximo de 95,37% obtido com os algoritmos *Gradient Boosted Tree* e *Logistic Regression*, e um mínimo de 48,53%, atingido com o classificador *Naive Bayes*. Recorde-se que, no problema 1.1, mesmo o pior valor de Taxa de Acerto de 80,23%, obtido com o algoritmo W-OneR, representa um bom nível de *performance*. Verifica-se ainda uma maior variabilidade dos resultados de Taxa de Acerto obtidos. Em termos médios, o melhor desempenho preditivo é verificado com a utilização do algoritmo *Gradient Boosted Trees*, com uma Taxa de Acerto média de 91,23%, e o mais baixo desvio padrão verificado de 5,22%. O algoritmo W-J48 apresenta um nível de *performance* bastante próximo, de 89,10%. Refira-se que, o algoritmo *Gradient Boosted Tree*, apresentava no problema 1.1 a pior *performance* do conjunto de algoritmos testados, com um nível médio de Taxa de Acerto de 93,69%. O algoritmo de regressão testado, *Logistic Regression*, que no problema 1.1 apresenta níveis de *performance* superiores, apenas superado pelo W-48, regista uma perda da sua capacidade preditiva neste problema *multitarget*, passando a apresentar desempenhos medianos.

Algoritmos	Ranking Taxa de Acerto	Taxa Acerto		F-score	Precisão (1)	Sensibilidade (2)	Amplitude p.p. (1) - (2)
		Valor	Desvio Padrão				
Logistic Regression (RapidMiner Studio Core)	5º	83,03%	5,40%	83,39%	83,26%	84,14%	0,9 p.p.
Naive Bayes (RapidMiner Studio Core)	7º	81,22%	5,93%	81,12%	80,87%	82,13%	1,3 p.p.
Gradient Boosted Trees (RapidMiner Studio Core)	1º	91,23%	5,22%	93,79%	92,75%	94,86%	2,1 p.p.
Decision Tree (RapidMiner Studio Core)	3º	88,08%	6,73%	91,98%	90,26%	93,80%	3,5 p.p.
W-JRip (Weka)	6º	81,88%	8,15%	83,81%	82,13%	85,61%	3,5 p.p.
Random Forest (RapidMiner Studio Core)	4º	83,89%	7,29%	87,27%	84,87%	89,88%	5 p.p.
W-J48 (Weka)	2º	89,10%	5,90%	92,07%	89,25%	95,15%	5,9 p.p.
W-OneR (Weka)	8º	73,50%	6,87%	82,46%	82,46%	68,49%	14 p.p.

Tabela 20: *Ranking* da amplitude entre as medidas de Precisão e Sensibilidade (valores médios) – problema 2.1, obtido via *RapidMiner* e trabalhado em *Microsoft Excel*

Analisando comparativamente os valores médios obtidos pelas medidas de Precisão e Sensibilidade, com exceção do W-OneR, os algoritmos continuam a apresentar valores próximos entre estas medidas. Apesar de as amplitudes serem ligeiramente superiores às verificadas no problema 1.1, a capacidade dos algoritmos em classificar aleatoriamente um exemplo em uma determinada classe, não é afetada numa abordagem a um problema *multitarget*. No algoritmo *Logistic Regression*, que apresenta um desempenho mediano a nível de Taxa de Acerto, é obtida a menor amplitude com 0,9 p.p. O algoritmo *Gradient Boosted Trees*, continua a apresentar bons desempenhos, registando a terceira menor amplitude, com 2,1 p.p. Avaliando o desempenho dos algoritmos com recurso ao *F-score*, verifica-se uma

degradação da sua *performance* face ao problema 1.1, tendo os algoritmos *Gradient Boosted Trees* e W-J48, apresentado os maiores valores médios, com 93,79% e 92,07%, respetivamente, inferiores ao mínimo de 96,1%, obtido com o algoritmo W-OneR no problema 1.1. Globalmente, os algoritmos *Gradient Boosted Trees* e W-J48 apresentam os melhores níveis de desempenho na construção de um modelo de previsão do número de contentores colocados à recolha pelos clientes. Apesar do algoritmo *Gradient Boosted Tree* apresentar uma *performance* ligeiramente superior à do W-J48, optou-se pela utilização deste último na construção do modelo preditivo, dada a maior interpretabilidade da árvore de decisão obtida para os utilizadores da informação.

Problema 2.2: Identificação de regras caracterizadoras do número total de contentores que os clientes colocam à recolha

Na tabela seguinte são apresentados os resultados gerados pelos algoritmos de geração de regras de decisão, bem como o respetivo número de regras geradas:

Algoritmos	Nº regras geradas	Tempo execução	Taxa Acerto		Taxa Erro	<i>F-score</i>	Precisão (1)	Sensibilidade (2)	Amplitude p.p. (1)-(2)
			Valor	Desvio Padrão					
W-Part (Weka)	243	3 s	65,79%	0,91%	34,21%	76,45%	67,89%	87,47%	19,6 p.p.
Rule Induction (RapidMiner Studio Core)	164	11 s	64,50%	0,74%	35,50%	75,93%	67,19%	87,29%	20,1 p.p.
Tree to Rules (Decision Tree) (RapidMiner Studio Core)	515	1 s	64,20%	0,79%	35,80%	66,77%	85,84%	66,76%	19,1 p.p.
W-JRip (Weka)	28	12 s	62,48%	0,79%	37,52%	74,04%	61,41%	93,20%	31,8 p.p.
W-Ridor (Weka)	3443	8 m 49 s	60,13%	1,48%	39,87%	72,61%	67,78%	78,19%	10,4 p.p.
Subgroup Discovery (RapidMiner Studio Core)	180	12 m 3 s	49,65%	0,03%	50,35%	66,36%	49,65%	100,00%	50,3 p.p.

Tabela 21: *Ranking* dos resultados obtidos pelos algoritmos de regras de decisão – problema 2.2, obtido via *RapidMiner* e trabalhado em *Microsoft Excel*

Verifica-se uma degradação da *performance* dos resultados obtidos pelos diversos algoritmos testados, face à sua utilização já abordada no problema 1.2, referente à identificação das regras de colocação de resíduos à recolha ou não pelos clientes. A maior Taxa de Acerto verificada é de apenas de 65,79%, obtida com a utilização do algoritmo W-Part, seguida do operador *Rule Induction* do *RapidMiner*, com um resultado próximo, de 64,5%. O algoritmo W-Part apresenta ainda o maior *F-score* com 76,45%. O operador do *RapidMiner*, *Subgroup Discovery*, apresenta a pior *performance*, obtendo uma Taxa de Acerto de apenas 49,65%. Registam-se igualmente valores reduzidos de Precisão e Sensibilidade, com uma amplitude elevada entre estas medidas, com particular incidência no operador do *RapidMiner*, *Subgroup Discovery* e W-JRip, com uma amplitude de 50,3 p.p e 31,8 p.p., respetivamente. O algoritmo W-JRip, que no problema 1.2 apresentava níveis de *performance* próximos dos melhores desempenhos, tendo sido inclusive selecionado para a geração de regras de decisão, revela-se inadequado na abordagem ao problema 2.2., para geração de regras caracterizadoras

do número de contentores colocados à recolha. A necessidade de classificação de um maior número de classes do atributo classe pelo modelo, reflete-se quer no número de regras geradas, quer no tempo de execução do modelo, substancialmente superiores ao verificado na abordagem ao problema 1.2. Face aos resultados obtidos, foram efetuadas e testadas novas abordagens à modelação do modelo. Com a introdução da discretização do atributo classe, com o objetivo de reduzir o número de exemplos a classificar, conforme já abordado anteriormente no capítulo “4.5 Modelação”, foram atingidas melhores *performances*:

Algoritmos	Nº regras geradas	Tempo execução	Taxa Acerto		<i>F-score</i>	Precisão (1)	Sensibilidade (2)	Amplitude p.p. (1)-(2)
			Valor	Desvio Padrão				
W-Part (Weka)	110	1 s	85,23%	0,55%	91,60%	86,42%	97,44%	11 p.p.
W-JRip (Weka)	13	2 s	84,78%	0,44%	91,42%	85,65%	98,01%	12,4 p.p.
Tree to Rules (Decision Tree) (RapidMiner Studio Core)	218	1 s	84,87%	0,64%	91,39%	86,25%	97,18%	10,9 p.p.
W-Ridor (Weka)	490	1 m 17 s	84,52%	0,62%	91,28%	85,27%	98,20%	12,9 p.p.
Rule Induction (RapidMiner Studio Core)	29	2 s	84,30%	0,34%	91,17%	85,04%	98,26%	13,2 p.p.
Subgroup Discovery (RapidMiner Studio Core)	60	3 m 3 s	82,39%	0,04%	90,35%	82,39%	100,00%	17,6 p.p.

Tabela 22: *Ranking* dos resultados obtidos pelos algoritmos de regras de decisão com discretização do atributo classe – problema 2.2, obtido via *RapidMiner* e trabalhado em *Microsoft Excel*

Nesta nova abordagem ao problema, foram obtidos níveis de Taxa de Acerto bastante superiores e similares entre o universo de algoritmos testados, com um máximo de 85,23% obtido com o algoritmo W-Part, e um mínimo de 82,39% atingido pelo operador do *RapidMiner*, *Subgroup Discovery*. Verifica-se ainda um crescimento nos níveis de Precisão e Sensibilidade, uma diminuição do número de regras geradas e dos tempos de execução do modelo. Relativamente aos níveis de *F-score* obtidos, o algoritmo W-Part apresenta igualmente o maior valor com 91,6%, verificando-se igualmente níveis bastante elevados nos restantes algoritmos, com um mínimo bastante próximo de 90,35%, obtido com a utilização do operador do *RapidMiner*, *Subgroup Discovery*. O algoritmo W-JRip, que na anterior simulação de modelação não se encontrava a nível das melhores *performances*, passa a apresentar a segunda melhor *performance* nos algoritmos testados, com um *F-score* de 91,42%. O novo modelo desenvolvido apresenta níveis de *performance* superiores aos obtidos no problema 1.2 abordado anteriormente, para previsão de colocação ou não de resíduos à recolha, em que foi registada uma Taxa de Acerto máxima de apenas 76,59%.

Na abordagem ao problema de identificação de regras que caracterizam o número de contentores entregues pelos clientes, considerou-se mais apropriada a adoção do algoritmo W-Rip, uma vez que apresenta níveis de *F-score* e Taxa de Acerto obtidos bastante próximos do W-Part, e um número adequado de regras a uma efetiva utilização e compreensão dos utilizadores da informação (foram gerados pelo W-Rip 13 regras, face às 110 regras geradas pelo W-Part).

4.7. Implementação

Encontra-se em fase de desenvolvimento interno pela LIPOR, um *software* para monitorização e avaliação dos circuitos de recolha seletiva de resíduos, denominado de *Data Centre* LIPOR. A implementação dos modelos propostos na presente tese, passarão pela sua articulação com o referido *software*, disponibilizando um módulo para o agendamento dinâmico dos circuitos de recolha.

Os modelos desenvolvidos em *RapidMiner* deverão ser atualizados periodicamente, e suportarão, numa lógica diária ou semanal, a definição dos circuitos de recolha, nomeadamente, os clientes que a equipa de recolha deverá visitar para recolha dos seus resíduos, e uma estimativa do número de contentores a recolher.

Encontra-se em fase final de implementação pela LIPOR, a automatização do registo das recolhas pelos operadores durante a operação de recolha, por via da utilização de terminais móveis, bem como o desenvolvimento de ligação ao *software* de pesagens da LIPOR, para associação do registo de pesagem da viatura de recolha, aos quantitativos de resíduos recolhidos por circuito, tipo de material e por cliente.

No módulo a disponibilizar no *Data Centre* LIPOR, será possível a monitorização e avaliação da execução dos circuitos de recolha e da participação dos clientes, sendo ainda disponibilizadas as regras caracterizadoras da entrega de resíduos pelos clientes e respetivo número de contentores. Esta informação, para além de suportar a compreensão dos agendamentos propostos pelo *software*, é particularmente relevante no suporte às operações de prospeção de novos clientes e estabelecimento de novos circuitos de recolha seletiva porta-a-porta não residencial.

Os modelos desenvolvidos em *RapidMiner* caracterizam-se pela sua simplicidade de construção e utilização, bem como pela sua escalabilidade. O *software RapidMiner*, disponibiliza ainda conectores com bases de dados, pelo que será de fácil implementação a integração da informação produzida pelos modelos com a base de dados da LIPOR, para articulação com o *Data Centre* LIPOR.

A solução proposta exige um reduzido nível de investimento, uma vez que carece apenas de licenciamento do *RapidMiner*, disponível gratuitamente, da aquisição de dados climatológicos históricos junto do IPMA (Instituto Português do Mar e da Atmosfera), sendo a integração da informação produzida pelos modelos propostos com o *Data Centre* LIPOR, assegurada internamente pela equipa de TI (Tecnologias de Informação) da LIPOR.

5. Conclusões

A tendência de concentração populacional em grandes centros urbanos tem colocado elevadas pressões e desafios no planeamento, gestão e desenvolvimento das cidades, por forma a colmatar os consequentes impactos nas suas condições económicas, sociais e ambientais. Têm sido realizadas diversas abordagens, normalmente rotuladas de *Smart Cities*, para estruturação de um modelo de desenvolvimento urbano sustentável, em diversos domínios, nomeadamente na área de gestão de resíduos.

O uso sustentável dos recursos constitui atualmente, uma prioridade da política europeia e nacional, traduzida num novo paradigma de desenvolvimento económico designado por “Economia Circular” em que, através da transformação dos resíduos num recurso, este novo modelo proporciona vantagens a nível de competitividade, crescimento e emprego, assim como para o ambiente. No caso particular da gestão de resíduos urbanos, foram definidas em fevereiro de 2018, novas e ambiciosas metas para a valorização de resíduos, pelo que é urgente a adoção de estratégias e iniciativas que promovam o incremento da separação dos resíduos pelos cidadãos, nomeadamente por via de uma aposta na implementação e monitorização de circuitos de recolha seletiva, com particular foco em circuitos de recolha porta-a-porta, que constitui um dos vetores estratégicos da estratégia nacional de gestão de resíduos.

Na presente tese, foi abordada a monitorização e recolha seletiva porta-a-porta não residencial (PAP-NR) dos resíduos urbanos produzidos por atividades de comércio e outras que, pelo volume e tipologia de resíduos produzidos, são equiparáveis a resíduos urbanos. A execução dos circuitos de recolha PAP-NR é realizada com base em rotas e datas pré-determinadas, sendo o seu agendamento assegurado de forma manual, dependendo da “sensibilidade” do gestor ou técnico responsável da área de gestão de resíduos dos municípios, ignorando a produção de resíduos de cada estabelecimento comercial.

Face a soluções de sensorização de contentores já existentes no mercado, é proposta uma solução pouco onerosa que, com base em dados históricos reais de recolha, possibilita o agendamento dinâmico e flexível de circuitos de recolha, por via da previsão da colocação de resíduos à recolha pelos clientes e respetivo número total de contentores, e da identificação de regras que caracterizam a produção de resíduos pelos clientes. Esta solução constituirá ainda um precioso auxílio no suporte às operações de angariação de novos clientes

ou estabelecimento de um novo circuito de recolha, ultrapassando assim situações de *cold start* por inexistência de dados históricos, permitindo ainda a obtenção de vantagens a nível de eficiências operacionais, com o consequente impacto na redução de custos da atividade de recolha e melhoria do nível de serviço prestado aos clientes.

Foi realizada uma análise exaustiva das variáveis do *data set*, referente ao histórico de recolhas de um circuito PAP-NR da área de influência da LIPOR, no período de janeiro a dezembro de 2017. Decorrente desta análise, foram propostos indicadores para avaliação do desempenho da execução dos circuitos de recolha pelas equipas de recolha - Taxa de eficiência dos circuitos % (TEC), e da participação dos clientes aderentes ao serviço PAP-NR – Taxa de colocação de resíduos à recolha % (TCR). Foram ainda introduzidas novas variáveis no *data set*, com o objetivo de melhor caracterizar a produção de resíduos.

Para resposta aos problemas em análise e *outputs* propostos, foram adotadas abordagens metodológicas distintas, enquadradas em problemas de classificação em *supervised learning*. Deste modo, no desenvolvimento dos modelos propostos, foram testados métodos probabilísticos, com recurso à utilização do classificador *Naive Bayes*, métodos de regressão, com utilização do conceito de regressão logística, e métodos baseados em procura, com utilização de algoritmos de árvores e regras de decisão. Constituindo a interpretabilidade, um dos fatores críticos de sucesso na proposta de soluções para os problemas abordados na presente tese, o recurso a algoritmos de regras de decisão revelou-se de especial importância, para a geração de informação entendível e de suporte à tomada de decisão para os gestores dos circuitos de recolha.

A avaliação dos modelos foi realizada com recurso a validação cruzada com 10 iterações, tendo sido adotadas como medidas de avaliação a Taxa de Acerto, Precisão, Sensibilidade e *F-score*.

Com o modelo desenvolvido para previsão da colocação de resíduos à recolha pelos clientes, foram obtidos níveis de *performance* bastante elevados no universo de algoritmos testados, tendo sido obtido o melhor desempenho preditivo com o algoritmo de árvore de decisão W-J48, com uma Taxa de Acerto de 97,45% e um *F-score* de 98,6%. Na abordagem ao problema *multitarget* de previsão do número de contentores colocados à recolha, verificou-se uma redução da *performance* preditiva do modelo, tendo sido obtido o melhor resultado com outro algoritmo de árvore de decisão - *Gradient Boosted Tree*, com uma Taxa de Acerto de 91,23% e um *F-score* de 93,79%.

Para a geração de regras de decisão, a opção pelo algoritmo a utilizar nos modelos, foi baseada na interpretabilidade e adequabilidade do número de regras geradas, para além das medidas de avaliação anteriormente descritas. Deste modo, para a geração de regras caracterizadoras da colocação de resíduos à recolha, foi adotado o algoritmo W-JRip, tendo sido geradas 29 regras, com uma Taxa de Acerto de 73,46% e um *F-score* de 81,53%, cujo *output* descreve as situações em que os clientes não colocam resíduos à recolha. Na geração de regras caracterizadoras do número total de contentores produzidos pelos clientes, o algoritmo W-JRip revelou-se igualmente a melhor opção, tendo sido obtidas 13 regras, com uma Taxa de Acerto de 84,78% e um *F-score* de 91,42%.

Por último, destaque-se a facilidade de implementação dos modelos desenvolvidos, pela sua simplicidade, escalabilidade e facilidade de integração da informação produzida com o *software* desenvolvido internamente pela LIPOR para gestão e monitorização dos circuitos de recolha seletiva.

6. Referências Bibliográficas

- Akhtar, M. et al. (2017) 'Backtracking search algorithm in CVRP models for efficient solid waste collection and route optimization', *Waste Management*, 61, pp. 117–128.
- Anthopoulos, L. G. (2017) *Understanding Smart Cities: A Tool for Smart Government or an Industrial Trick?* Larissa, Greece: Springer International Publishing.
- Belda, M. et al. (2014) 'Climate classification revisited: from Koppen to Trewartha', *Climate Research*, 59, pp. 1–13.
- Bosch, P. et al. (2017) CITYkeys Indicators for smart city projects and smart cities. Disponível em: <http://nws.eurocities.eu/MediaShell/media/CITYkeysD14Indicatorsforsmartcityprojectsandsmartcities.pdf>.
- Dilorenzo, T. J. (1999) *Suburban Legends: Why Smart Growth Is Not So Smart*. Washington, USA: Center for the Study of American Business Washington University. Disponível em: https://wc.wustl.edu/files/wc/imce/suburban_legend.pdf.
- ERSAR (2017) *PERSU 2020 Plano Estratégico para os Resíduos Urbanos Relatório de Monitorização 2016*. Disponível em: <http://www.ersar.pt/pt/site-comunicacao/site-noticias/Paginas/relatorio-de-monitorizacao-do-persu-20201023-7489.aspx>.
- European Council (2018) EU ambassadors approve new rules on waste management and recycling - Consilium, 2018-02-23. Disponível em: <http://www.consilium.europa.eu/en/press/press-releases/2018/02/23/eu-ambassadors-approve-new-rules-on-waste-management-and-recycling/>
- Gama, J. et al. (2015) *Extração de Conhecimento de Dados - Data Mining*. 2ª Edição. M. Robalo. Lisboa: Edições Sílabo.
- Giffinger, R. et al. (2007) *Smart cities Ranking of European medium-sized cities*. Vienna. Disponível em: http://www.smart-cities.eu/download/smart_cities_final_report.pdf.
- Gil-Garcia, J. R., Pardo, T. A. and Nam, T. (2015) 'What makes a city smart? Identifying core components and proposing an integrative and comprehensive conceptualization', *Information Polity*, 20(1), pp. 61–87.
- Harrison, C. and Donnelly, I. a. (2011) 'A Theory of Smart Cities', in *Proceedings of the 55th Annual Meeting of the ISSS - 2011*, Hull, UK, pp. 1–15. doi: 10.1017/CBO9781107415324.004.
- ISWA - International Solid Waste Association (2017) 4th Industrial Revolution on the Waste Management Sector. Disponível em: <https://www.iswa.org/home/news/news-detail/browse/12/article/press-release-waste-management-will-be-transformed-by-new-technologies/109/>.
- JSCA (2018) Smart Community Development. Disponível em: <https://www.smart-japan.org/english/index.html>.
- Khatoun, R. and Zeadally, S. (2016) 'Smart Cities: Concepts, Architectures, Research Opportunities.', *Communications of the Association for Computing Machinery*, 59(8), pp. 46–57.
- LIPOR (2017) *Relatório e Contas LIPOR 2017*. Baguim do Monte, Portugal. Disponível em: <https://www.lipor.pt/pt/bibliotecas/>.
- Manville, C. et al. (2014) *Mapping Smart Cities in the EU*. Bruxelas, Bélgica. Disponível em: [http://www.europarl.europa.eu/RegData/etudes/etudes/join/2014/507480/IPOL-ITRE_ET\(2014\)507480_EN.pdf](http://www.europarl.europa.eu/RegData/etudes/etudes/join/2014/507480/IPOL-ITRE_ET(2014)507480_EN.pdf).

- Maria Rodrigues, A. and Ferreira, J. S. (2015) 'Waste collection routing - Limited multiple landfills and heterogeneous fleet', *Networks*, 65(2), pp. 155–165.
- Ministério do Ambiente, O. do T. e E. (2014) PERSU 2020 - Plano Estratégico para os Resíduos Sólidos Urbanos, *Diário da República*. Lisboa, Portugal. Disponível em: http://www.apambiente.pt/_zdata/DESTAQUES/2014/Portaria_PlanoEstrategico_PERSU2020_final.pdf.
- Tavares, G. et al. (2009) 'Optimisation of MSW collection routes for minimum fuel consumption using 3D GIS modelling', *Waste Management*. Elsevier Ltd, 29(3), pp. 1176–1185.
- Tayel, S., Reif, M. and Dengel, A. (2013) 'Rule-based Complaint Detection using RapidMiner', in *RCOMM 2013. RapidMiner Community Meeting and Conference (RCOMM-2013)*. Porto, Portugal, pp. 141–149. Disponível em: http://www.dfki.de/web/forschung/publikationen/renamerFileForDownload?filename=paper.pdf&file_id=uploads_2041.
- Tayel, S., Reif, M. and Dengel, A. (2013) 'Rule-based Complaint Detection using RapidMiner', in *RCOMM 2013. RapidMiner Community Meeting and Conference (RCOMM-2013)*. Porto, Portugal, pp. 141–149.
- UN-HABITAT (2018) Climate Change. Available at: <https://unhabitat.org/urban-themes/climate-change/>.
- United Nations, Department of Economic and Social Affairs, P. D. (2014) World Urbanization Prospects, The 2014 revision. New York, USA. Disponível em: <http://esa.un.org/unpd/wup/Highlights/WUP2014-Highlights.pdf>.
- United Nations, Department of Economic and Social Affairs, P. D. (2017) World Population Prospects The 2017 Revision Key Findings and Advance Tables. New York, USA. Disponível em: https://esa.un.org/unpd/wpp/Publications/Files/WPP2017_KeyFindings.pdf.
- United Nations News Centre (2012) Rio + 20: Development banks to invest \$ 175 billion in sustainable transport Secretary-General Spokesperson 's Office. Disponível em: http://www.un.org/apps/news/story.asp?NewsID=42287#.Wk_HB99l-Uk.
- Visnjic, I. et al. (2016) 'Governing the City', *California Management Review*, 59(1), pp. 109–140.
- Vu, D. D. and Kaddoum, G. (2017) 'A Waste City Management System for Smart Cities Applications', in *Advances in Wireless and Optical Communications*, pp. 225–229.
- Yigitcanlar, T. (2015) 'Smart cities: an effective urban development and management model?', *Australian Planner*. Taylor & Francis, 52(1), pp. 27–34.

7. Anexos

Anexo A – Feriados e outros dias festivos em 2017, e ocorrência de recolha de resíduos

Dia	Designação	Dia semana	Ocorrência de Recolha (Sim/Não)
01-01-2017	Dia de Ano Novo	Dom.	Não
14-04-2017	Sexta-Feira Santa	6ª	Sim
16-04-2017	Páscoa	Dom.	Não
25-04-2017	Dia da Liberdade	3ª	Sim
01-05-2017	Dia do Trabalhador	2ª	Sim
10-06-2017	Dia de Portugal	Sab.	Não
15-06-2017	Corpo de Deus	5ª	Sim
16-06-2017	Feriado Municipal de Espinho	6ª	Sim
15-08-2017	Assunção de Nossa Senhora	3ª	Sim
05-10-2017	Implantação da República	5ª	Sim
01-11-2017	Dia de Todos os Santos	4ª	Sim
01-12-2017	Restauração da Independência	6ª	Sim
08-12-2017	Dia da Imaculada Conceição	6ª	Sim
25-12-2017	Natal	2ª	Não
26-12-2017	Tolerância de ponto	3ª	Sim

Anexo B – Revisão da nomenclatura da atividade dos clientes – agrupamento e descrição

Atividade: Comércio			
Descrição: Estabelecimentos comerciais dedicados a retalho generalista e especializado			
Designação:			
Artigos de confeitaria	Artigos de marroquinaria	Artigos religiosos e ceras	Artigos têxteis lar
Eletrodomésticos	Eletrodomésticos	Produtos regionais	Pronto-a-vestir
Congelados	Drogaria	Estação de serviço	Florista
Frutaria	Peixaria	Garrafeira	Mercearia/ minimercado
Supermercado	Comércio generalista	Perfumaria e cosméticos	Telecomunicações
Decoração	Comércio jornais e revistas	Sapataria	Ótica
Loja de animais	Retrosaria	Mobiliário	-
Atividade: Restauração			
Descrição: Preparação e fornecimento de refeições em restaurantes e estabelecimentos similares.			
Designação:			
Café	Café/ snack bar	Cafetaria	Padaria
Confeitaria	Salão de chá	Restaurante	Bar
Atividade: Saúde			
Descrição: Estabelecimentos tratamento ambulatorio, internamento e venda de fármacos			
Designação:			
Farmácia	Lar 3ª idade	-	-
Atividade: Hotelaria			
Descrição: Estabelecimentos de alojamento e alimentação a hóspedes.			
Designação:			
Hotel	Pousada	Campismo	-
Atividade: Serviços			
Descrição: Estabelecimentos prestadores de serviços especializados			
Designação:			
Agência de viagens	Publicidade	Lavandaria	Imprensa

Anexo C – Regras de decisão geradas pelo algoritmo W-JRip, caracterizadoras da entrega de resíduos pelos clientes, obtido via *RapidMiner*

JRIP rules:

=====

```
(Dia_semana = terca) and (Horario = Normal) and (Clima = Quente) and (Centralidade = 1) => Entrega_RU(1/0)=0 (1213.0/456.0)
(Dia_semana = quinta) and (Horario = Normal) and (Atividade_revista = Restauração) and (Centralidade = 1) and (Clima = Quente) and (Mes = setembro ) => Entrega_RU(1/0)=0 (78.0/24.0)
(Dia_semana = quinta) and (Horario = Normal) and (Centralidade = 1) and (Mes = novembro ) => Entrega_RU(1/0)=0 (285.0/102.0)
(Dia_semana = quinta) and (Horario = Normal) and (Atividade_revista = Restauração) and (Centralidade = 1) and (Clima = Quente) and (Mes = julho ) => Entrega_RU(1/0)=0 (104.0/35.0)
(Dia_semana = quinta) and (Horario = Normal) and (Centralidade = 1) and (Mes = dezembro ) => Entrega_RU(1/0)=0 (228.0/80.0)
(Dia_semana = quinta) and (Clima = Quente) and (Mes = agosto ) => Entrega_RU(1/0)=0 (524.0/241.0)
(Dia_semana = quinta) and (Mes = outubro ) and (Atividade_revista = Restauração) => Entrega_RU(1/0)=0 (300.0/125.0)
(Dia_semana = terca) and (Centralidade = 3) and (Horario = Normal) => Entrega_RU(1/0)=0 (396.0/111.0)
(Horario = Normal) and (Atividade_revista = Serviços) and (Clima = Quente) => Entrega_RU(1/0)=0 (376.0/155.0)
(Dia_semana = terca) and (Mes = novembro ) and (Horario = Normal) and (Centralidade = 1) => Entrega_RU(1/0)=0 (228.0/73.0)
(Dia_semana = terca) and (Mes = dezembro ) and (Horario = Normal) and (Centralidade = 1) => Entrega_RU(1/0)=0 (228.0/73.0)
(Dia_semana = quinta) and (Mes = setembro ) and (Atividade_revista = Restauração) => Entrega_RU(1/0)=0 (222.0/101.0)
(Centralidade = 3) and (Atividade_revista = Hotelaria) => Entrega_RU(1/0)=0 (310.0/104.0)
(Dia_semana = quinta) and (Mes = junho ) and (Horario = Normal) and (Centralidade = 1) => Entrega_RU(1/0)=0 (244.0/99.0)
(Horario = Normal) and (Atividade_revista = Serviços) and (Mes = novembro ) => Entrega_RU(1/0)=0 (60.0/15.0)
(Dia_semana = quinta) and (Mes = setembro ) and (Horario = Normal) and (Centralidade = 1) => Entrega_RU(1/0)=0 (109.0/39.0)
(Dia_semana = terca) and (Centralidade = 3) and (Clima = Quente) and (Atividade_revista = Restauração) and (Horario = Alargado) => Entrega_RU(1/0)=0 (143.0/30.0)
(Centralidade = 2) and (Atividade_revista = Restauração) and (Horario = Normal) and (Dia_semana = quarta) => Entrega_RU(1/0)=0 (299.0/136.0)
(Horario = Normal) and (Centralidade = 3) and (Feriado_Festividades_Fds(1/0) = 0) and (Atividade_revista = Restauração) and (Dia_semana = sexta) => Entrega_RU(1/0)=0 (162.0/74.0)
(Dia_semana = quinta) and (Mes = julho ) and (Centralidade = 1) => Entrega_RU(1/0)=0 (220.0/101.0)
(Centralidade = 2) and (Clima = Quente) and (Atividade_revista = Restauração) and (Mes = julho ) and (Dia_semana = quinta) => Entrega_RU(1/0)=0 (52.0/16.0)
(Centralidade = 2) and (Mes = novembro ) and (Atividade_revista = Restauração) and (Dia_semana = quinta) => Entrega_RU(1/0)=0 (65.0/20.0)
(Centralidade = 2) and (Clima = Quente) and (Atividade_revista = Restauração) and (Mes = setembro ) and (Dia_semana = sexta) and (Horario = Normal) => Entrega_RU(1/0)=0 (40.0/15.0)
(Centralidade = 2) and (Clima = Quente) and (Atividade_revista = Restauração) and (Horario = Normal) and (Dia_semana = sexta) and (Mes = julho ) => Entrega_RU(1/0)=0 (32.0/12.0)
(Mes = outubro ) and (Dia_semana = quinta) and (Centralidade = 1) and (Atividade_revista = Comércio) => Entrega_RU(1/0)=0 (112.0/40.0)
(Centralidade = 2) and (Clima = Quente) and (Atividade_revista = Restauração) and (Mes = agosto ) and (Dia_semana = sexta) and (Horario = Normal) => Entrega_RU(1/0)=0 (32.0/12.0)
(Dia_semana = terca) and (Mes = setembro ) and (Clima = Ameno) and (Horario = Normal) and (Centralidade = 1) and (Atividade_revista = Comércio) => Entrega_RU(1/0)=0 (25.0/6.0)
(Mes = dezembro ) and (Dia_semana = quinta) and (Atividade_revista = Restauração) and (Centralidade = 2) and (Horario = Alargado) => Entrega_RU(1/0)=0 (20.0/4.0)
=> Entrega_RU(1/0)=1 (20070.0/4474.0)
```

Number of Rules : 29

Anexo D – Identificação de regras caracterizadoras da entrega de resíduos pelos clientes, geradas pelo algoritmo W-JRip – exemplos de filtros de regras por atributos, obtido via *RapidMiner*

- Exemplo de regras geradas pelo algoritmo W-JRip, para um cliente

```
JRIP rules:
=====

(Mes = julho ) and (Dia_semana = terca) => Entrega_RU(1/0)=0 (4.0/3.0)
(Clima = Ameno) and (Dia_semana = quarta) and (Mes = novembro ) => Entrega_RU(1/0)=0 (3.0/2.0)
(Mes = julho ) and (Dia_semana = segunda) => Entrega_RU(1/0)=0 (5.0/4.0)
(Mes = dezembro ) and (Dia_semana = sexta) => Entrega_RU(1/0)=0 (5.0/4.0)
(Mes = outubro ) and (Dia_semana = terca) and (Clima = Ameno) => Entrega_RU(1/0)=0 (1.0/0.0)
=> Entrega_RU(1/0)=1 (174.0/0.0)

Number of Rules : 6
```

- Exemplo de regras geradas pelo algoritmo W-JRip, para a atividade de Restauração

```
JRIP rules:
=====

(Horario = Normal) and (Dia_semana = quinta) and (Centralidade = 1) and (Mes = novembro ) => Entrega_RU(1/0)=0 (130.0/42.0)
(Horario = Normal) and (Dia_semana = terca) and (Centralidade = 3) => Entrega_RU(1/0)=0 (150.0/5.0)
(Horario = Normal) and (Dia_semana = quinta) and (Centralidade = 1) and (Mes = agosto ) => Entrega_RU(1/0)=0 (104.0/32.0)
(Horario = Normal) and (Dia_semana = quinta) and (Centralidade = 1) and (Mes = dezembro ) => Entrega_RU(1/0)=0 (104.0/32.0)
(Horario = Normal) and (Dia_semana = terca) and (Clima = Quente) and (Centralidade = 1) => Entrega_RU(1/0)=0 (539.0/233.0)
(Dia_semana = quinta) and (Mes = outubro ) => Entrega_RU(1/0)=0 (300.0/125.0)
(Dia_semana = quinta) and (Clima = Quente) and (Mes = julho ) => Entrega_RU(1/0)=0 (300.0/127.0)
(Horario = Normal) and (Centralidade = 2) and (Clima = Quente) and (Feriado_Festividades_Fds(1/0) = 0) and (Mes = agosto ) => Entrega_RU(1/0)=0 (120.0/49.0)
(Dia_semana = quinta) and (Mes = setembro ) => Entrega_RU(1/0)=0 (300.0/125.0)
(Horario = Normal) and (Centralidade = 2) and (Dia_semana = quarta) => Entrega_RU(1/0)=0 (275.0/127.0)
(Dia_semana = terca) and (Centralidade = 3) and (Horario = Alargado) => Entrega_RU(1/0)=0 (254.0/78.0)
(Horario = Normal) and (Mes = junho ) and (Dia_semana = quinta) => Entrega_RU(1/0)=0 (170.0/72.0)
(Horario = Normal) and (Centralidade = 2) and (Feriado_Festividades_Fds(1/0) = 0) and (Mes = julho ) => Entrega_RU(1/0)=0 (64.0/29.0)
(Dia_semana = quinta) and (Horario = Alargado) and (Mes = dezembro ) => Entrega_RU(1/0)=0 (128.0/57.0)
(Horario = Normal) and (Centralidade = 2) and (Mes = novembro ) => Entrega_RU(1/0)=0 (136.0/65.0)
(Horario = Normal) and (Mes = dezembro ) and (Dia_semana = terca) => Entrega_RU(1/0)=0 (136.0/57.0)
(Dia_semana = quinta) and (Horario = Alargado) and (Clima = Quente) and (Mes = agosto ) => Entrega_RU(1/0)=0 (128.0/56.0)
(Horario = Normal) and (Centralidade = 3) and (Feriado_Festividades_Fds(1/0) = 0) and (Clima = Quente) and (Dia_semana = quarta) => Entrega_RU(1/0)=0 (92.0/40.0)
(Dia_semana = quinta) and (Horario = Alargado) and (Mes = novembro ) => Entrega_RU(1/0)=0 (160.0/70.0)
(Horario = Normal) and (Mes = setembro ) and (Dia_semana = terca) => Entrega_RU(1/0)=0 (58.0/26.0)
(Horario = Normal) and (Mes = dezembro ) and (Centralidade = 3) and (Dia_semana = sexta) => Entrega_RU(1/0)=0 (25.0/9.0)
(Horario = Normal) and (Mes = novembro ) and (Dia_semana = terca) => Entrega_RU(1/0)=0 (104.0/42.0)
(Horario = Normal) and (Clima = Quente) and (Centralidade = 3) and (Dia_semana = sexta) => Entrega_RU(1/0)=0 (91.0/38.0)
(Horario = Normal) and (Mes = dezembro ) and (Centralidade = 2) and (Dia_semana = quinta) => Entrega_RU(1/0)=0 (32.0/12.0)
=> Entrega_RU(1/0)=1 (11454.0/2431.0)

Number of Rules : 25
```

- Exemplo de regras geradas pelo algoritmo W-JRip, para o nível de centralidade 1

```
JRIP rules:
=====

(Dia_semana = quinta) and (Horario = Normal) and (Atividade_revista = Restauração) and (Clima = Quente) and (Mes = agosto ) => Entrega_RU(1/0)=0 (104.0/32.0)
(Dia_semana = terca) and (Horario = Normal) and (Clima = Quente) and (Atividade_revista = Comércio) and (Mes = junho ) => Entrega_RU(1/0)=0 (100.0/23.0)
(Dia_semana = quinta) and (Horario = Normal) and (Atividade_revista = Restauração) and (Mes = setembro ) => Entrega_RU(1/0)=0 (104.0/33.0)
(Dia_semana = terca) and (Horario = Normal) and (Clima = Quente) and (Atividade_revista = Comércio) and (Mes = agosto ) => Entrega_RU(1/0)=0 (100.0/24.0)
(Dia_semana = quinta) and (Horario = Normal) and (Mes = novembro ) => Entrega_RU(1/0)=0 (285.0/102.0)
(Dia_semana = terca) and (Horario = Normal) and (Mes = dezembro ) => Entrega_RU(1/0)=0 (228.0/73.0)
(Dia_semana = quinta) and (Horario = Normal) and (Mes = julho ) => Entrega_RU(1/0)=0 (228.0/84.0)
(Horario = Normal) and (Dia_semana = terca) and (Clima = Quente) and (Mes = julho ) and (Atividade_revista = Comércio) => Entrega_RU(1/0)=0 (100.0/26.0)
(Dia_semana = terca) and (Horario = Normal) and (Mes = novembro ) => Entrega_RU(1/0)=0 (228.0/73.0)
(Dia_semana = quinta) and (Mes = dezembro ) and (Horario = Normal) => Entrega_RU(1/0)=0 (228.0/80.0)
(Dia_semana = quinta) and (Mes = outubro ) and (Horario = Normal) and (Atividade_revista = Restauração) => Entrega_RU(1/0)=0 (104.0/33.0)
(Dia_semana = terca) and (Clima = Quente) and (Horario = Normal) and (Atividade_revista = Restauração) => Entrega_RU(1/0)=0 (539.0/233.0)
(Dia_semana = quinta) and (Mes = junho ) and (Horario = Normal) => Entrega_RU(1/0)=0 (254.0/105.0)
(Dia_semana = quinta) and (Atividade_revista = Comércio) and (Mes = outubro ) => Entrega_RU(1/0)=0 (112.0/40.0)
(Atividade_revista = Serviços) and (Clima = Quente) => Entrega_RU(1/0)=0 (466.0/173.0)
(Dia_semana = quinta) and (Mes = agosto ) and (Atividade_revista = Comércio) => Entrega_RU(1/0)=0 (112.0/41.0)
(Atividade_revista = Serviços) and (Mes = novembro ) => Entrega_RU(1/0)=0 (60.0/15.0)
(Dia_semana = quinta) and (Mes = setembro ) and (Atividade_revista = Comércio) => Entrega_RU(1/0)=0 (112.0/39.0)
(Dia_semana = terca) and (Mes = outubro ) and (Atividade_revista = Comércio) => Entrega_RU(1/0)=0 (140.0/36.0)
(Dia_semana = terca) and (Mes = setembro ) and (Horario = Normal) => Entrega_RU(1/0)=0 (135.0/39.0)
(Atividade_revista = Serviços) and (Mes = dezembro ) => Entrega_RU(1/0)=0 (60.0/18.0)
=> Entrega_RU(1/0)=1 (13795.0/2790.0)

Number of Rules : 22
```

Anexo E – Regras de decisão geradas pelo algoritmo W-JRip, caracterizadoras do número total de contentores que os clientes colocam à recolha, obtido via *RapidMiner*

JRIP rules:

=====

```
(Centralidade = 3) and (Horario = Alargado) and (Atividade_revista = Restauração) and (Dia_semana = terca) => Contentores_soma=[7;+inf.] (78.0/32.0)
(Horario = Alargado) and (Atividade_revista = Comércio) => Contentores_soma=[4-6] (252.0/3.0)
(Horario = Alargado) and (Centralidade = 2) and (Atividade_revista = Hotelaria) => Contentores_soma=[4-6] (60.0/6.0)
(Atividade_revista = Restauração) and (Centralidade = 2) and (Clima = Quente) and (Horario = Alargado) and (Mes = julho ) => Contentores_soma=[4-6] (57.0/19.0)
(Atividade_revista = Restauração) and (Centralidade = 2) and (Dia_semana = quarta) and (Horario = Normal) => Contentores_soma=[4-6] (136.0/56.0)
(Atividade_revista = Restauração) and (Centralidade = 2) and (Horario = Alargado) and (Mes = agosto ) => Contentores_soma=[4-6] (56.0/18.0)
(Atividade_revista = Restauração) and (Horario = Reduzido) and (Centralidade = 1) and (Dia_semana = sexta) => Contentores_soma=[4-6] (134.0/56.0)
(Atividade_revista = Restauração) and (Centralidade = 2) and (Mes = novembro ) and (Horario = Alargado) => Contentores_soma=[4-6] (57.0/17.0)
(Atividade_revista = Restauração) and (Clima = Quente) and (Centralidade = 2) and (Horario = Alargado) and (Mes = setembro ) => Contentores_soma=[4-6] (42.0/14.0)
(Atividade_revista = Restauração) and (Mes = outubro ) and (Centralidade = 2) and (Horario = Alargado) => Contentores_soma=[4-6] (59.0/20.0)
(Atividade_revista = Restauração) and (Horario = Reduzido) and (Centralidade = 1) and (Dia_semana = quinta) => Contentores_soma=[4-6] (80.0/37.0)
(Atividade_revista = Restauração) and (Mes = dezembro ) and (Centralidade = 2) and (Horario = Alargado) => Contentores_soma=[4-6] (54.0/19.0)
=> Contentores_soma=[1-3] (16830.0/2374.0)
```

Number of Rules : 13

Anexo F – Identificação de regras caracterizadoras do número total de contentores que os clientes colocam à recolha, geradas pelo algoritmo W-JRip – exemplos de filtros de regras por atributos, obtido via *RapidMiner*

- Exemplo de regras geradas pelo algoritmo W-JRip, para um cliente

```
JRIP rules:
=====

(Feriado_Festividades_Fds(1/0) = 1) and (Clima = Quente) => Contentores_soma=[4-6] (27.0/8.0)
(Dia_semana = sexta) and (Clima = Quente) => Contentores_soma=[4-6] (23.0/6.0)
(Dia_semana = sexta) and (Mes = dezembro ) => Contentores_soma=[4-6] (5.0/0.0)
(Dia_semana = sexta) and (Mes = novembro ) => Contentores_soma=[4-6] (4.0/0.0)
(Dia_semana = segunda) and (Mes = novembro ) => Contentores_soma=[4-6] (4.0/0.0)
(Feriado_Festividades_Fds(1/0) = 1) and (Mes = dezembro ) and (Dia_semana = segunda) => Contentores_soma=[4-6] (3.0/0.0)
(Feriado_Festividades_Fds(1/0) = 1) and (Dia_semana = sexta) => Contentores_soma=[4-6] (2.0/0.0)
=> Contentores_soma=[1-3] (167.0/7.0)

Number of Rules : 8
```

- Exemplo de regras geradas pelo algoritmo W-JRip, para a atividade Restauração

```
JRIP rules:
=====

(Centralidade = 3) and (Horario = Alargado) and (Dia_semana = terca) => Contentores_soma=[7;+inf.] (78.0/32.0)
(Centralidade = 2) and (Dia_semana = quarta) and (Horario = Normal) => Contentores_soma=[4-6] (136.0/56.0)
(Horario = Reduzido) and (Centralidade = 1) and (Dia_semana = sexta) => Contentores_soma=[4-6] (134.0/56.0)
(Centralidade = 2) and (Horario = Alargado) and (Clima = Quente) and (Mes = julho ) => Contentores_soma=[4-6] (57.0/19.0)
(Centralidade = 2) and (Horario = Alargado) and (Mes = novembro ) => Contentores_soma=[4-6] (57.0/17.0)
(Clima = Quente) and (Centralidade = 2) and (Horario = Alargado) and (Mes = agosto ) => Contentores_soma=[4-6] (52.0/17.0)
(Clima = Quente) and (Mes = outubro ) and (Centralidade = 2) and (Horario = Alargado) => Contentores_soma=[4-6] (40.0/14.0)
(Dia_semana = terca) and (Mes = dezembro ) and (Centralidade = 2) and (Horario = Alargado) => Contentores_soma=[4-6] (12.0/4.0)
=> Contentores_soma=[1-3] (10005.0/1739.0)

Number of Rules : 9
```

- Exemplo de regras geradas pelo algoritmo W-JRip, para o nível de centralidade 3

```
JRIP rules:
=====

(Horario = Alargado) and (Atividade_revista = Restauração) and (Dia_semana = terca) => Contentores_soma=[7;+inf.] (78.0/32.0)
(Horario = Alargado) and (Atividade_revista = Comércio) => Contentores_soma=[4-6] (252.0/3.0)
(Atividade_revista = Hotelaria) and (Dia_semana = terca) => Contentores_soma=[4-6] (29.0/0.0)
=> Contentores_soma=[1-3] (2175.0/329.0)

Number of Rules : 4
```