Departamento de Ciências e Tecnologias da Informação

Triagem de pedidos de assistência médica

Ália Gerardo

Dissertação submetida como requisito parcial para obtenção do grau de

Mestre em Informática e Gestão

Orientador

Doutor Ricardo Daniel Santos Faro Marques Ribeiro, Professor Auxiliar,

ISCTE-IUL

Outubro, 2017

Agradecimentos

Passado mais de um ano após ter surgido a ideia do objecto de estudo deste trabalho, o número de pessoas a quem é devido um agradecimento é relativamente elevado.

Ao orientador, o Prof. Doutor Ricardo Ribeiro, que aceitou o meu convite e que por diversas vezes se mostrou mais optimista do que eu. Agradeço a sua disponibilidade, motivação e orientação durante todo este trabalho.

Aos colaboradores do INEM, com os quais tive oportunidade de falar. Agradeço por permitirem que fosse possível a realização deste estudo, toda a vossa colaboração e saliento a vossa impressionante simpatia.

Aos meus colegas, que me desculparam sempre o mau-humor.

E, finalmente, à minha família e amigos, que me apoiaram, desculparam, incentivaram e permitiram que o trabalho fosse concluído.

Resumo

Nesta dissertação foi avaliada a capacidade de efectuar a triagem de pedidos de assistência médica recorrendo a técnicas de Data Mining.

Com base na revisão da literatura decidiu-se seguir a metodologia de Cios et. al (2000), tendo-se explorado diversas abordagens. Uma das principais razões para a escolha desta metodologia foi o facto de se verificar que é a mais utilizada em estudos na área da saúde.

Os dados utilizados consistem em 2.070.227 pedidos de assistência médica com as variáveis Ano, Mês, Dia, Dia da Semana, Hora, Distrito, Concelho, Prioridade, Tipo de Ocorrência, Faixa Etária e Sexo, sendo a variável Prioridade o nível de triagem atribuído, podendo este assumir um de quatro valores Emergentes, Urgente, Pouco-urgente e Não-urgente.

O tratamento de dados médicos exige cuidados que vão além dos requisitos habituais neste tipo de trabalhos. Para além da dificuldade na obtenção de dados por questões de confidencialidade, é importante que o resultado seja transparente e perceptível e cuidadosamente avaliado.

Nesse sentido, foram aplicados os algoritmos árvores de decisão (J48), o Naïve Bayes e Máquinas de Vectores de Suporte (SMO e LibSVM) considerando a escala real de quatro níveis (Emergente, Urgente, Pouco-urgente e Não-urgente). Foi igualmente considerada uma escala de dois níveis, derivada a partir da escala real. As medidas de avaliação utilizadas foram a taxa de acerto, sensibilidade e especificidade.

Os resultados mostram que as técnicas de Data Mining são mais eficazes a efectuar a triagem considerando apenas dois níveis. Igualmente se demonstrou nas diferentes abordagens que as Máquinas de Vectores de Suporte são mais eficazes que as restantes técnicas utilizadas.

Palavras-chave: *Knowledge Discovery in Databases*, Triagem médica, *Data Mining*, classificação.

Abstract

In this dissertation was evaluated the ability to perform the screening of medical assistance requests using Data Mining techniques.

Based on the literature review it was decided to follow the methodology of Cios et. al (2000), and several approaches have been explored. One of the main reasons for choosing this methodology was the fact that it is used most frequently in healthcare studies.

The data consists of 2,070,227 requests of medical assistance and it features the following variables: Year, Month, Day, Day of the Week, Hour, District, County, Priority, Type of Occurrence, Age Group and Gender. The variable for Priority is the level of triage attributed, which may assume one of four values: Emergent, Urgent, Less Urgent and Nonurgent.

The processing of medical data demands a supplementary degree of caution when comparing to other kinds of data. In addition to the difficulties of obtaining sensitive and confidential information, it is important that the results are transparent, perceptible and carefully evaluated.

In this regard, the following algorithms are applied: Decision Tree (J48), the Naïve Bayes and Support Vector Machines (SMO and LibSVM), considering the four-levels of the real scale: Emergent, Urgent, Less Urgent and Nonurgent. A two-level scale was also derived from the original scale. The evaluation measures used were: Accuracy, Sensitivity and Specificity.

The results show that Data Mining techniques are more effective performing triage considering only two levels. It has also been demonstrated in the different approaches investigated that the Support Vector Machines are more effective than the other techniques analyzed.

Keywords: *Knowledge Discovery in Databases*, medical triage, *Data Mining*, classification.

Índice

Índice de Tabelas

Índice de Figuras

Lista de Abreviações

AD – Árvore de Decisão

ANFIS – Adaptive Neuro-Fuzzy Inference System

CHAID – Chi-square Automatic Interaction Detection

CODU – Centro de Orientação de Doentes Urgentes

CRISP-DM – Cross Industry Standard Process for Data Mining

DM – Data Mining

INEM – Instituto Nacional de Emergência Médica

KDD – Knowledge Discovery in Databases

LERS – Learning from Examples using Rough Sets

NB – Naïve Bayes

RL – Regressão Logística

RNA – Rede Neuronal Artificial

SEMMA – Sample, Explore, Modify, Model, and Assess

SVM – Support Vector Machine – Máquina de Vetores de Suporte

TETRICOSY – Telephonic Triage and Couseling System

VMER – Viatura Médica de Emergência e Reanimação

WEKA – Waikato Environment for Knowledge Analysis

Capítulo 1

Introdução

O objectivo inicial deste trabalho consistia na identificação das chamadas falsas recebidas no INEM, através de técnicas de *Data Mining*, consideradas técnicas avançadas de análise de dados. Para tal, foi efectuado um pedido de dados a esse instituto onde se identificou o objectivo do pedido. Contudo, ao longo do formal processo para obtenção de dados, concluiu-se que a maioria das chamadas falsas são filtradas na Central 112, responsabilidade da Protecção Civil. Uma vez que se chegou a esta conclusão numa altura em que não se dispunha de tempo para um novo processo de obtenção de dados, após uma visita e reunião com colaboradores do INEM foi-nos disponibilizada uma base de dados que permite efectuar um estudo com outro objectivo – identificar a prioridade de um pedido de assistência médica.

1.1. Definição do Problema e Objectivos

Em Portugal, são recorrentes temas como o aumento do défice no Serviço Nacional de Saúde, a falta de médicos de família, os elevados tempos de espera nos hospitais e centros de saúde, resultantes do grande número de pacientes que recorrem a esses serviços. Quaisquer medidas que resultem numa maior eficácia e eficiência destes serviços serão assim importantes, por forma a que estes possam dar resposta, principalmente, em casos de urgência.

Numa situação de emergência é apropriado recorrer-se ao Número Europeu de Emergência, 112, a partir do qual os pedidos de socorro que digam respeito a situações de urgência ou emergência médica são transferidos para o Instituto Nacional de Emergência Médica, I. P. (INEM), mais concretamente para os Centros de Orientação de Doentes Urgentes (CODU). Os profissionais destes centros, onde se incluem

médicos e técnicos, atendem e avaliam os pedidos de socorro, efetuando uma triagem, coordenando vários serviços e meios para que possam ser disponibilizados os recursos ajustados a cada caso (Serviço Nacional de Saúde, 2017; INEM, 2017).

Esta triagem é efectuada com o apoio de um sistema informático de fluxos de triagem, implementado em 2012, o TETRICOSY® (Telephonic Triage and Couseling System) que foi desenvolvido pelo próprio INEM. Contudo, numa análise efectuada pelo Centro Hospitalar de Lisboa Ocidental aos accionamentos da respetiva Viatura Médica de Emergência e Reanimação (VMER), constata-se que existem alguns problemas na triagem efectuada pelo sistema, algo que é demonstrado pelo aumento acentuado do número de accionamentos injustificados da VMER, e igualmente de outros meios, conforme relatório do Tribunal de Contas de 2016 (Tribunal de Contas, 2017; VMER SFX, 2017).

Estatísticas públicas disponibilizadas no site do INEM (http://www.inem.pt) mostram o número elevadíssimo de pedidos de socorro recebidos anualmente por estes centros. Em 2015, por exemplo, foram rececionadas mais de 1,3 milhões de chamadas, que resulta numa média superior a 3,5 mil chamadas por dia. Estão disponíveis igualmente estatísticas por prioridade de ocorrência, podendo verificar-se que, por exemplo, em novembro de 2016, 13% dos contactos recebidos poderiam ter sido evitados, pois ou não foram acionados quaisquer meios ou foram encaminhados para a linha Saúde 24 (INEM, 2017).

Posto isto, será importante efectuar uma triagem o mais correta possível – perceber quais os pedidos de socorro que poderão ser evitados; e dentro dos justificados identificar os que carecem de assistência privilegiada e altamente capacitada – para que a instituição que nos socorre em caso de efetiva urgência se torne mais eficaz e eficiente. Para tal, pretende-se, com recurso a um modelo informático de extracção de conhecimento de grande quantidade de dados, empregando técnicas de *Data Mining*, analisar dados fornecidos pelo INEM para que se possa eventualmente perceber que factores são determinantes na atribuição de um nível de prioridade e predizer qual a prioridade atribuída a um novo pedido de socorro.

Nesse sentido, e para que os objectivos sejam atingidos, a investigação deverá conseguir responder ao seguinte:

- Existirão variáveis determinantes na classificação dos pedidos de socorro?

- Será possível predizer a prioridade de um pedido de socorro?

1.2. Estrutura da Dissertação

Para dar resposta a estas questões, o trabalho foi organizado em seis capítulos, conforme o seguinte:

Capítulo 2 onde são apresentados os conceitos básicos desta área de estudo utilizados ao longo do trabalho de modo.

Capítulo 3 que corresponde à revisão de trabalho relacionado de forma que se possa conhecer quais os resultados habitualmente obtidos e quais as metodologias, técnicas, métricas e variáveis frequentemente utilizadas.

Capítulo 4 onde de acordo com a literatura são identificados os cuidados na análise de dados em saúde, de forma a perceber quais as particularidades e aspectos que devem ser tidos em conta.

Capítulo 5 apresenta a metodologia utilizada para efectuar a triagem de pedidos de assistência.

E finalmente, o Capítulo 6 no qual são apresentadas as conclusões, limitações desta investigação e sugestões para futuros trabalhos.

Capítulo 2

Conceitos Básicos

Neste capitulo será efectuada uma explicação de alguns termos que serão utilizados ao longo do trabalho, de forma a possibilitar ao leitor uma completa compreensão da investigação.

2.1. *Knowledge Discovery in Databases*, *Data Mining* e *Machine Learning*

Todos os passos na análise dos dados pertencem a uma das fases do processo frequentemente designado por *Knowledge Discovery in Databases* (KDD), em português "Descoberta de Conhecimento em Bases de Dados". Esta é a designação mais corrente do processo de extracção de conhecimento a partir de bases de dados e deve-se a Gregory Piatetsky-Shapiro, remontando ao ano de 1989. Este processo é definido como sendo não trivial, interactivo e iterativo, por envolver várias etapas que conduzem à identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados (Fayyad et al., 1996a; Piatetsky-Shapiro, 2000).

Data Mining (DM) e *Machine Learning* são termos amplamente utilizados na literatura, sendo difícil distingui-los. Podem encontrar-se, por exemplo, as seguintes definições: *Data Mining* consiste na aplicação de algoritmos para analisar dados ou extrair padrões em categorias específicas a partir dos dados (Klosgen & Zytkow, 1996). *Machine Learning* é uma área de investigação que estuda como os computadores podem aprender, ou melhorar a sua performance baseando-se nos dados, tendo especial foco no

desenvolvimento de programas que automaticamente aprendam e reconheçam padrões complexos (Han, Pei & Kamber, 2001).

As técnicas de *Data Mining* e *Machine Learning* podem ser classificadas, essencialmente, em duas categorias, supervisionadas e não supervisionadas. Técnicas supervisionadas aplicam algoritmos que aprendem com base em dados anotados, organizados em classes. Neste caso, estes algoritmos designam-se de classificadores. Caso se pretendam predizer valores contínuos, os algoritmos designam-se regressores. Técnicas não-supervisionadas, aplicadas na análise de *clusters* (conjuntos de dados), recorrem a algoritmos que aprendem com base em dados não anotados, sendo o próprio algoritmo responsável por identificar grupos nos dados (Han, Pei & Kamber, 2001).

A aplicação destas técnicas tem um de dois objectivos principais – descrição ou predição. Refere-se que se pretende descrição quando o objectivo consiste em encontrar características ou propriedades nos dados e predição quando se utilizam os dados para prever acontecimentos futuros (Han, Pei & Kamber, 2001).

Antecedendo a aplicação de qualquer técnica, existe uma fase na qual são analisados e preparados os dados a utilizar, devendo ter-se especial cuidado com valores desconhecidos. Os valores desconhecidos são elementos que não seguem o comportamento da generalidade dos dados (Jiawei & Kamber, 2001).

2.2. Avaliação dos Resultados das Técnicas de DM

Em geral, para validar os algoritmos, o conjunto original de dados deverá ser dividido em dois conjuntos, um de treino e outro de teste. O conjunto de dados de treino é utilizado para a fase de aprendizagem – desenvolvimento do modelo, o conjunto de teste servirá para verificar a eficácia do modelo contruído. (Olson & Delen, 2008). Esta divisão pode ser efectuada através do método k-fold cross-validation (Kohavi, 1995) muitas vezes utilizando-se k=10 (10-fold cross validation). Através deste método o conjunto de dados é aleatoriamente dividido em k conjuntos, mutuamente exclusivos, de aproximadamente igual dimensão. O modelo é treinado e testado k vezes, sendo que de cada vez apenas um dos conjuntos é utilizado como teste, e os restantes para treino. A avaliação do modelo será o resultado da média aritmética dos vários testes (Olson & Delen, 2008).

O resultado do modelo é regularmente espelhado sobe a forma de matriz, designada de matriz de confusão, cujas colunas representam a predição efectuada pelo modelo e as linhas os valores reais. A métrica mais frequentemente utilizada para avaliação é a taxa de acerto (Equação 1). Esta métrica consiste no rácio do total das correctas predições sobre o total de instâncias (Hossin & Sulaiman, 2015).

Apesar de um modo geral a comparação de resultados ser efectuada pela análise da taxa de acerto, existe um conjunto de outras métricas de avaliação dos resultados em DM, nomeadamente a sensibilidade (Equação 2) e especificidade (Equação 3), definidas no trabalho de Lavrač et. al (1999) e cuja formulação é a seguinte:

$$Taxa\ de\ acerto = \frac{número\ de\ instâncias\ correctamente\ classificadas}{número\ total\ de\ instâncias} \quad (1)$$

$$Sensibilidade = \frac{verdadeiros\ positivos}{verdadeiros\ positivos + falsos\ negativos} \quad (2)$$

$$Especificidade = \frac{verdadeiros\ negativos}{verdadeiros\ negativos + falsos\ positivos} \quad (3)$$

Importa referir que a taxa de acerto é aplicável tanto a classificações com apenas duas classes (binárias) como em classificações com três ou mais classes. Por outro lado, a sensibilidade e especificidade, são métricas frequentemente utilizadas para classificações binárias (Cios & Moore, 2002, Sokolova & Lapalme, 2009). Sensibilidade – também conhecida como *recall* ou rácio dos verdadeiros positivos – é a métrica que avalia a predição dos positivos, isto é, a identificação de valores como sendo de determinada classe, entre todas as instâncias dessa mesma classe. A especificidade – igualmente conhecida como rácio dos verdadeiros negativos – avalia a predição dos negativos, a identificação dos valores como não pertencendo a determinada classe, entre todas as instâncias que de facto não pertencem a essa classe (Hossin & Sulaiman, 2015). Apesar destas métricas serem essencialmente usadas em classificações binárias, nos restantes tipos de classificações são importantes para distinguir os resultados da predição para cada classe individualmente e para se efectuarem comparações entre estas (Guler & Ubeyli, 2007, Baldi et al., 2000, Lavrač et. al, 1999).

2.3.Metodologias de KDD

Para a execução deste tipo de análises a grandes conjuntos de dados, a utilização de uma metodologia ajuda a perceber o processo de KDD e serve como guia para o planeamento e execução do mesmo, o que permite para além de reduzir tempo e custos, que os projectos sejam mais facilmente compreendidos e aceites. As metodologias não são mais do que um conjunto de passos a seguir, em sequência, que podem envolver ciclos e iterações. (Kurgan & Musilek, 2006). Das diversas metodologias existentes as principais são de Fayyad et al. (1996b), CRISP-DM (Shearer, 2000; Wirth & Hipp, 2000), Cios et al. (2000) e SEMMA (SAS Institute Inc., 1997), e cujos passos são consultáveis no final deste capítulo, na Tabela 1.

A metodologia de Fayyad et al., (1996b) serve como estrutura básica para todas as metodologias, envolve 9 fases e classifica-se como especialmente voltada para a investigação. Aponta-se como principal desvantagem desta metodologia o facto de a fase de processamento dos dados ser anterior à escolha da técnica de DM, o que não permite que os dados estejam completamente ajustados implicando eventuais repetições de etapas (Cios & Kurgan, 2005).

O CRISP-DM, desenvolvido por um consórcio de 4 empresas (SPSS, NCR, Daimler Chrysler e OHRA), disponibilizado no ano 2000, é a mais frequentemente utilizada, envolve 6 fases, e é caracterizada por ser uma metodologia orientada para o mundo industrial (Kurgan & Musilek, 2006).

A metodologia de 6 fases de Cios et. al, que surge igualmente no ano 2000, é referenciada como sendo uma adaptação da CRISP-DM às necessidades dos investigadores (Marbán et al., 2009, Kurgan & Musilek, 2006). Esta metodologia distingue-se pelo facto de requerer conhecimento da terminologia e dos conceitos de DM e é destacada por ser a única a fornecer indicações detalhadas sobre as possíveis iterações e possuir uma descrição modificada da última fase que enfatiza que o conhecimento descoberto para uma área em particular poderá ser aplicado numa área diferente (Kurgan & Musilek, 2006). Kurgan & Musilek (2006) salientam igualmente que é frequentemente utilizada no âmbito da saúde.

Por último, a metodologia SEMMA, desenvolvida em 1997 pela SAS Institute, sendo SEMMA um acrónimo para Sample, Explore, Modify, Model e Assess – os 5 passos do

processo. É desenhada para a utilização do software SAS Enterprise Miner. Esta metodologia é, à primeira vista, menos completa que a CRISP-DM, mas analisado aprofundadamente, engloba todos os mesmos passos do processo de KDD (Azevedo & Santos, 2008).

Tabela 1 – Principais modelos de KDD

	Fayyad et al. (1996b)		CRISP-DM (2000)		Cios et. al (2000)		SEMMA (1997)
1.	Desenvolvimento e compreensão do domínio de aplicação	**1.**	Compreensão do negócio	**1.**	Compreensão do domínio do problema		
2.	Criação da base de dados alvo	**2.**	Compreensão dos dados	**2.**	Compreensão dos dados	**1.**	Sample
3.	Limpeza e pré-processamento dos dados	**3.**	Preparação dos dados	**3.**	Preparação dos dados	**2.**	Explore
4.	Redução e visualização dos dados					**3.**	Modify
5.	Escolha da tarefa de DM						
6.	Escolha da técnica de DM						
7.	*Data Mining*	**4.**	Criação do Modelo	**4.**	*Data Mining*	**4.**	Model
8.	Interpretação dos padrões	**5.**	Avaliação	**5.**	Avaliação do conhecimento descoberto	**5.**	Assess
9.	Consolidação do conhecimento descoberto	**6.**	Entrega	**6.**	Utilização do conhecimento descoberto		

Capítulo 3

Trabalho Relacionado

No início de um processo de investigação é importante a análise de trabalhos já realizados sobre o mesmo tema ou abordagens similares à que se pretende usar ainda que o tema não seja tão próximo, na medida em que permite perceber quais as particularidades e aspectos importantes para o sucesso do trabalho bem como as metodologias e técnicas mais adequadas.

Nesse sentido, e tendo em conta que pretendemos perceber, com recurso a técnicas de DM, quais as variáveis importantes na atribuição de um determinado nível de prioridade a um pedido de socorro recebido no INEM procurámos, essencialmente, estudos nos quais fossem aplicadas técnicas de DM e onde o objecto fosse triagem na área da saúde.

3.1. Árvores de Decisão

Zhang & Szolovits (2008) consideram que as Árvores de Decisão são adequadas para problemas na área da saúde, nos casos em que o resultado é um de um número finito de opções e os dados podem conter erros e valores desconhecidos. Utilizam esta técnica com o objectivo de classificar condições dos pacientes como de alarme ou não-alarme. Nesta análise incluem diversos atributos numéricos, tais como frequência cardíaca, pulso, frequência respiratória, pressão sanguínea e níveis de oxigénio. Incluem ainda classificações efectuadas pela interpretação desses valores pelo sistema de monitorização existente no serviço, com indicação do estado de alarme e severidade e ainda, a classificação pelo médico ou enfermeiro. Recorrendo ao algoritmo C5.0 (Quinlan, 2004), e também a Rede Neuronal Artificial (Bishop, 1995), tentam classificar os eventos, considerando como correcta a avaliação do médico ou

enfermeiro. Numa amostra de 196 horas de monitorização a Árvore de Decisão apresentou taxa de acerto de 98% e a Rede Neuronal Artificial 99%.

Li, Guo & Handly (2009) num estudo com o propósito de predizer as admissões num serviço de urgência, com uma base de dados composta por 2.784 registos de pacientes admitidos, utilizaram quatro tipo de técnicas de DM: Árvores de Decisão (Rokach & Maimon, 2014), Naïve Bayes (NB), Regressão Logística (RL) (Hosmer et al. 2013) e Máquinas de Vectores de Suporte (SVM) (Cortes & Vapnik, 1995). Nesta análise utilizam atributos com informação demográfica (como o idade e sexo) medidas clínicas (como frequência cardíaca e respiratória) e ainda a queixa do paciente. Este último atributo foi alvo de tratamento para que existisse alguma standardização, tendo o resultado sido alvo das quatro técnicas. Nesta investigação, a técnica de AD foi a que piores resultados apresentou, com taxa de acerto de 76,21% (NB: 77,38%, RL: 77,34% e SVM: 78,21%).

Zmiri, Shahar & Taieb-Maimon (2012) num trabalho levado a cabo para melhorar a triagem num serviço de urgência, verificam igualmente se as técnicas de DM são mais eficazes caso as variáveis de output forem em menor número, recorrendo ao algoritmo C4.5 (Quinlan, 2014) de Árvores de Decisão e ao modelo Naïve Bayes. Estas técnicas foram aplicadas numa base de dados com 402 registos, para os quais foram consideradas diversos atributos como a idade, sexo, temperatura, pulso, pressão sanguínea, queixa do paciente, histórico do paciente e nota do enfermeiro. Cada registo foi alvo de classificação por um médico especialista, sendo essa classificação considerada como a efetiva. Na análise considerando uma escala de triagem de 4 níveis a AD apresenta taxa de acerto de 49,75% e NB de 56,72%. Posteriormente, considerando uma escala de apenas 2 níveis (agregando os níveis iniciais 2, 3 e 4) a taxa de acerto da AD subiu para os 70,65% e de NB para 72,64%.

Izad Shenas et. al. (2014) numa investigação com o propósito de criar um modelo preditivo que consiga identificar os pacientes com elevado custo aplicam dois algoritmos de AD – o C5.0 (Quinlan, 2004) e o CHAID (CHi-squared Automatic Interaction Detection) (Magidson, 1994) este último conhecido por requerer bases de dados de maior dimensão – e também uma Rede Neuronal Artificial (RNA). Como amostra utilizaram uma extensa base de dados de saúde americana, com 98.175 registos e perto de 1800 variáveis. Essa base foi alvo de redução, quer do número de registos quer do número de atributos, com fundamento na revisão de literatura efectuada pelos

investigadores e análise estatística, que não iremos detalhar tendo em conta que não será aplicável na nossa investigação. Terminam com apenas 39 atributos – que incluem informação sobre o actual estado de saúde, informação histórica do paciente e dados demográficos – e 31.704 registos sobre os quais, uma vez aplicado o modelo, a taxa de acerto das técnicas de AD foram superiores à da RNA, tendo CHAID uma taxa de 86,3% e C5.0 de 93,7% enquanto que a RNA atingiu 76,2%.

Verifica-se, igualmente, que a aplicação de AD, para além de classificação e predição, poderá ser utilizada para interpretação, sendo exemplo disso a investigação Lin et al. (2010). Nesta investigação, realizada em Taiwan, analisam-se os casos de diagnósticos anormais num serviço de urgência e a sua relação com o nível atribuído pela triagem. Para 501 registos foram escolhidos 10 atributos – temperatura, pressão arterial sistólica e diastólica, pulso, frequência respiratória, nível de oxigénio, nível de triagem do enfermeiro, nível de triagem do médico, tempo de decisão e decisão final. Após implementado um algoritmo de classificação – o K-means (MacQueen, 1967) – para análise de clusters aplicam o algoritmo C4.5 para análise dos resultados, referindo escolherem uma AD pela sua facilidade de interpretação.

3.2. Redes Neuronais Artificiais

No já referido estudo de Zhang & Szolovits (2008) a técnica de Redes Neuronais Artificiais (RNA) foi distinguida como sendo útil em investigações na área da saúde, por permitir extrair padrões não lineares e trabalhar com dados que contenham erros. Nessa investigação, através da implementação de uma RNA da classe *feedforward*, com apenas um nó de output e uma camada escondida considerando um número de neurónios igual ao número de nós de entrada, que neste caso foram todos os atributos. O algoritmo de treino considerado foi o de retropropagação. Verificou-se que esta técnica foi mais eficaz em classificação do que a AD, tendo a RNA apresentado taxa de acerto de 99% enquanto que a AD 98%.

Azeez et al. (2013) demonstram o sucesso da aplicação deste tipo de técnicas no seu estudo efectuado na Malásia, que com base numa amostra de 2.223 registos, tenta-se predizer o nível de triagem (numa escala de 3). Foi aplicada a RNA com uma camada escondida, 12 neurónios e 20 nós (atributos) de entrada. Esta técnica foi escolhida por ser considerada como sendo uma técnica simples. Utilizaram, ainda, um modelo hibrido

Adaptive Neuro-fuzzy Inference (ANFIS) (Jang, 1993.). Este último, apesar de se distinguir pela sua capacidade de perceber a partir dos dados as características que efectivamente influenciam a predição, apresentou valores inferiores de taxa de acerto – ANFIS: 94%, RNA: 96,7%.

Por outro lado, no estudo de Izad Shenas et al. (2014), também já mencionado por aplicar dois algoritmos de AD. Em oposição aos mencionados anteriormente, a RNA apresentou valores de taxa de acerto significativamente inferiores aos registados pelas técnicas de AD – RNA: 76%, CHAID: 86% e C5.0: 94%.

3.3. Máquina de Vectores de Suporte

No já referido estudo de Li, Guo & Handly (2009) em que foram aplicadas os 4 tipo de técnicas de DM (AD, NB, RL e SVM) a SVM destacou-se, com uma taxa de acerto de 78,21%. Os autores propõem ainda duas abordagens relativamente às variáveis para melhorar a predição, aplicadas na técnica SVM. Efectuaram duas abordagens distintas: uma analisando a semântica do atributo queixa do paciente, e outra introduzindo uma função kernel para integrar vários atributos. Estas alternativas de análise conduziram de facto a melhores resultados com taxas de acerto de 81,21% e 79,32% respectivamente, sendo a da SVM original de 78,21%.

3.4. Naïve Bayes

O classificador Naïve Bayes (NB), como referimos acima, de acordo com Li, Guo & Handly (2009), mostrou-se menos eficaz que a técnica SVM (taxa de acerto NB: 77,38%, SVM: 78,21%). Quando comparada com Árvores de Decisão e Regressão Logística, provou tratar-se de uma técnica com boa capacidade preditora (taxa de acerto NB: 77,38%, AD: 76,21%, RL:77,34%).

Este resultado verifica-se também no estudo Zmiri, Shahar & Taieb-Maimon (2012) com a aplicação dos algoritmos C4.5 e NB. O classificador NB apresentou igualmente uma taxa de acerto superior, considerando uma escala de 4 níveis de triagem de alcançou 56,72% (contra os 49,75% da AD) e com a escala de 2 níveis atingiu uma taxa de 72,64% (contra os 70,65% da AD).

Peck, et al. (2011) referem que esta técnica é uma ferramenta simples motivo pelo qual a escolheram aplicar no seu estudo que pretendia predizer a admissão de pacientes num Serviço de Urgência. Com base numa amostra de 621 casos, utilizando como atributos a admissão, idade, meio de chegada ao serviço, nível de urgência, departamento na urgência e queixa do paciente. O classificador NB, testando a ferramenta comparativamente com a análise de um especialista, apesar de não ter apresentado melhores resultados, tendo uma sensibilidade de 55,66% contra 53,48% obtida pelo especialista, foi considerado como tendo valor preditor e, por se tratar de uma técnica de simples utilização, referem que a sua aplicação poderia ser importante num fluxo de triagem devidamente ajustado.

Em termos de especificidade, também no já por diversas vezes comentado estudo de Li, Guo & Handly (2009), apesar de ter mostrado valores de taxa de acerto inferiores aos da técnica SVM, a NB provou ser melhor em termos de especificidade (NB: 83,03%, SVM: 82,86%, AD:81,93%, RL:82,24%).

3.5. Redes Bayesianas

Sadeghi et al. (2006) aplicam, num processo de triagem, a técnica Rede Bayesiana por a considerarem superior a Árvores de Decisão por permitir incluir conhecimento do domínio como input. Nessa investigação foi utilizada amostra com 90 registos, tendo como atributos a idade, sexo, histórico do paciente, os seus sintomas e a sua principal queixa. Estes registos poderiam ser classificados em 4 níveis de triagem, em que consideraram como classificação efectiva a definida posteriormente pelos médicos do serviço. Comparam o resultado da triagem da técnica de DM com a indicada por um Especialista de Emergência e verificou-se que ambos tiveram uma taxa de acerto de 56%. Em termos de sensibilidade a triagem efectuada pela Rede Bayesiana apresentou um valor de 90% contra os 64% do Especialista e em termos de especificidade, a Rede obteve um valor de 25%, contra 48% do Especialista.

Abad-Grau et al. (2008) destacam esta técnica numa revisão da literatura onde é efectuada uma análise comparativa entre sistemas de triagem com regras baseadas em técnicas de DM –Árvores de Decisão e Redes Bayesianas – e regras estipuladas por Especialistas. Concluem que a taxa de acerto dos sistemas desenhados por Especialistas é baixa, apontando como principal problema a existência de um elevado número de

variáveis que podem influenciar a decisão. Consideram como promissora a utilização de sistemas com base em técnicas de DM, destacando precisamente as Redes Bayesianas por permitirem interpretação do resultado, modificações efectuadas por especialistas e não terem em conta a ordem com que os factores surgem.

3.6. Outras Técnicas

Para além das técnicas abordadas acima existem outras que, apesar de serem usadas com menos frequência em estudos de triagem, poderiam igualmente ser aplicadas no nosso estudo.

Foram já abordadas técnicas menos comuns como a ANFIS que, na pesquisa de Azeez, D., et al. (2013), mostrou menos eficácia que a RNA e a Regressão Logística (RL) mencionadas por Li, Guo & Handly (2009).

Também o algoritmo LEM2 (Grzymala-Busse, 1992) utilizado para criação de regras que, de acordo com Lavrac (2005), se trata de um componente dos sistemas LERS (Learning from Examples using Rough Sets) que tem por base a Teoria *Rough Set* (Pawlak, 1982), foi considerado útil devido ao facto de que todo o conhecimento extraído ser baseado unicamente nos dados, não sendo necessário conhecimento prévio para definição de hipóteses. Esta técnica foi utilizada na área da saúde por Lin et al. (2011) para criação de regras de modo a perceber o que conduz a que um determinado caso tenha elevados custos num serviço de urgência. Com uma amostra de 22.990 casos, depois de efectuada uma análise de clusters, aplicaram o algoritmo LEM2 tendo este uma taxa de acerto de 93,7%.

Em Portugal, Alves (2015) com uma base de dados de uma urgência hospitalar do norte do país, efectua uma análise que pretende extrair conhecimento que permita apoiar e melhorar a gestão de sistemas de saúde. Numa base com 118.262 registos utilizaram-se como atributos a data, motivo de urgência, cor da pulseira, alta médica, serviço de destino, informação geográfica, sexo e idade. Foi efectuada uma análise de clusters com o objectivo de verificar se foram encontrados padrões relevantes para a gestão de sistemas de saúde. Numa fase seguinte, Alves (2015) aplica o algoritmo *apriori* (Agrawal & Srikant, 2008), com o objectivo de analisar regras de associação. Nessa ultima análise, considera que as regras originadas estão de acordo com os clusters criados, tirando-se as mesmas conclusões.

Constatamos assim que não existe uma técnica que apresente sempre melhores resultados em predição. Contudo verifica-se pela Tabela 2, que sintetiza o que abordámos neste capítulo, que os elevados e baixos valores de taxa de acerto para a mesma técnica podem ser influenciados pelo número de classes do problema.

Observámos a eficácia dos algoritmos de DM, apesar disso, como referem Patel et al. (2008), verifica-se existir uma resistência ao uso de sistemas automáticos na área da saúde e que a sua aceitação depende, entre outros factores (que serão abordados com maior detalhe no capítulo seguinte), essencialmente, do grau com que o sistema ajuda os profissionais a atingir os seus objectivos.

Tabela 2 – Apresentação dos valores de taxa de acerto, por técnica

		Taxa de acerto	
Árvores de Decisão			
Zhang & Szolovits, (2008)	C5.0	98,00%	2 classes
Li, Guo & Handly, (2009)		76,21%	2 classes
Zmiri, Shahar & Taieb-Maimon, (2012)	C4.5	49,75%	4 classes
		70,65%	2 classes
Izad Shenas et al., (2014)	CHAID	86,00%	2 classes
	C5.0	94,00%	2 classes
Bayes Networks			
Sadeghi et al., (2006)		56,00%	4 classes
Redes Neuronais Artificiais			
Zhang & Szolovits, (2008)		99,00%	2 classes
Azeez et al., (2013)		96,70%	3 classes
Izad Shenas et al., (2014)		76,00%	2 classes
Máquinas de Vectores de Suporte			
Li, Guo & Handly, (2009)		78,21%	2 classes
Naïve Bayes			
Li, Guo & Handly, (2009)		77,38%	2 classes
Zmiri, Shahar & Taieb-Maimon, (2012)		56,72%	4 classes
		72,64%	2 classes
Outras Técnicas			
Azeez et al., (2013)	ANFIS	94,00%	3 classes
Lin et al., (2011)	LEM2	93,70%	2 classes
Li, Guo & Handly, (2009)	RL	77,34%	2 classes

Capítulo 4

Cuidados na Análise de Dados em Saúde

No capítulo anterior foi analisada a literatura em se aplica KDD essencialmente para efeito de triagem em saúde dado ser esse o objecto deste trabalho.

4.1. Aplicações de KDD na Área da Saúde

KDD é aplicado em diversas áreas, tais como astronomia, marketing, investimento, detecção de fraudes, entre outros. A sua aplicação na área da saúde tem sido igualmente crescente, como referem Fayyad et al. (1996a).

Piatetsky-Shapiro (2000), num estudo em que é efectuada uma revisão de uma década de aplicação de KDD, refere que os principais profissionais interessados em KDD/DM são do mundo empresarial, quando inicialmente era maioritário o interesse de investigadores.

Apesar disso, Wasan, Bhatnagar & Kaur (2006) referem que com o aumento do volume de dados nos países desenvolvidos, os investigadores têm adoptado DM para extrair conhecimento na área da saúde. Foram desenvolvidos diversos trabalhos de investigação, com maior foco em doenças em particular, essencialmente as mais sensíveis à população, tais como cancro, diabetes, obesidade infantil, doenças cardiovasculares, como mostram os trabalhos de Hariz et al. (2012) e Esfandiari et al. (2014).

4.2. Singularidades do Domínio

KDD na área da saúde tem presentes particularidades que não surgem noutras áreas. Esfandiari et al. (2014) indicam que, ao contrário do que acontece em outras áreas, na área da saúde a aplicação de KDD não procura encontrar padrões, mas sim encontrar e, acima de tudo, explicar as minorias que fogem a esses padrões. Nesse mesmo trabalho, agrupam as aplicações deste processo de extracção de conhecimento como tendo os seguintes principais objectivos: aumento de eficiência e redução de erro (aplicável a doença específica); redução de tempo e custos; automatização (sistemas de suporte) e extracção de conhecimento (relações, riscos, novo conhecimento). No que diz respeito à automatização, Patel et al. (2008) referiram que os sistemas desenvolvidos deverão ser criados para ajudar os profissionais a evitar erros, assegurar qualidade e eficiência em cuidados de saúde.

Diversos estudos analisados no capítulo anterior fazem menção ao trabalho de Cios & Moore (2002) onde são elencados uma série de constrangimentos e dificuldades da aplicação de KDD na área da saúde. Nesse trabalho destacam questões relacionadas com o acesso aos dados, fazendo a particularmente interessante observação de que no geral as pessoas apreciam que sejam disponibilizados dados para avanços científicos, mas quanto a fornecer os seus próprios dados já se mostram relutantes. Percebemos assim, que existem dificuldades quanto à propriedade e privacidade dos dados – questões éticas e legais.

Com a digitalização, em saúde existe uma enorme quantidade de dados que se encontram registados em diversas fontes como hospitais, clinicas, farmácias, centros de saúde; e sob diversas formas, tais como texto, imagens, sons, etc. Até mesmo dados registados na mesma fonte e sob a mesma forma poderão ser difíceis de analisar. Pensemos num tão simples exemplo de dois médicos que analisam dois gémeos com gripe, um poderá registar gripe e o outro constipação, tal acontece porque o registo não é standardizado e muitas vezes é feito em texto livre. Todos estes factores levam a que os dados sejam muito heterogéneos o que dificulta igualmente a investigação (Cios & Moore, 2002).

Por outro lado, existe a questão relacionada com a singularidade da área da medicina – em que o resultado pode distinguir-se entre a vida e a morte – salientada por Esfandiari et al. (2014), o que torna o processo de investigação extremamente exigente. Koh & Tan

(2005) salientam o facto de que este tipo de investigações requer um intensivo planeamento e preparação e que, muitas vezes, os resultados poderão ser padrões e relações sem significado. Assim, Shillabeer & Roddick (2007) concluem que a área da saúde tem uma forte, estabelecida e aceite metodologia tradicional de investigação, confiando os médicos mais facilmente na opinião de um especialista do que no resultado da aplicação de técnicas de DM.

Relativamente ao constrangimento relativo ao acesso aos dados, será interessante mencionar alguns acontecimentos recentes. Existiu uma iniciativa desenhada pelo Serviço Nacional de Saúde de Inglaterra para extrair dados sociais e de saúde, de fontes como hospitais e clínicas, para fins que incluíam a investigação, o *care.data.* Esta iniciativa acabou por fracassar devido às questões da privacidade e, segundo a opinião de van Staa, et al. (2016), o insucesso desta iniciativa deve-se ao facto de não ter sido adquirida a confiança dos pacientes, cidadãos e profissionais de saúde. Referem que, se estes estivessem sido melhor informados, teriam aceite disponibilizar os seus dados. Schadt & Chilukuri (2015) também defendem este ponto de vista, indicando que a utilização de KDD será de forma continua, evolutiva e que os indivíduos disponibilizarão os seus dados porque terão percepção do benefício.

Já Piatetsky-Shapiro, (2000), na revisão que efectuou de uma década de KDD, referencia precisamente que áreas relacionadas com questões de privacidade têm um crescimento lento. Dixon-Woods & Ashcroft (2008) tinham igualmente referido que existe um longo trabalho a desenvolver na aquisição de confiança na investigação do público e na investigação em saúde.

4.3. Benefícios do KDD

Torna-se assim necessário insistir e comprovar a potencialidade de KDD na área da saúde. Shillabeer & Roddick (2007) referem que algum do conhecimento aceite hoje tem por base o acaso e que, para aquisição de conhecimento, é necessário pôr de parte metodologias tradicionais. Bellazzi & Zupan (2008) defendem este tipo de processo por se tratar de uma boa ferramenta, na medida em que permite uma análise de dados de forma abrangente e com integração de conhecimento anterior, envolvendo assim abordagens de várias áreas científicas e possuindo uma capacidade explicativa nos seus modelos.

A área da saúde, segundo Wasan, Bhatnagar & Kaur (2006), é uma área rica em informação, mas pobre em conhecimento. Com a enorme quantidade de dados, armazenada em formato digital, sem sistemas de análise automática não será possível beneficiar de todo o potencial destes dados. KDD permite extrair conhecimento de forma mais rápida e assim tornar igualmente mais célere a tomada de decisão.

4.4. Princípios a Garantir

A aplicação de KDD em saúde como referimos, engloba alguns cuidados. À parte dos problemas em relação ao acesso aos dados, Bellazzi & Zupan (2008) referem que, numa fase inicial para escolha das técnicas a empregar, deverá dar-se resposta a um conjunto de questões: (1) O resultado deve ser transparente e perceptível? (2) O modelo deve ser utilizado na tomada de decisão? (3) O modelo deverá indicar a probabilidade de cada resultado? Mencionam ainda um segundo conjunto de questões, às quais deverá ser dada resposta no decorrer do processo, importantes para a validação dos resultados: (1) Serão os dados e características das variáveis suficientes? (2) Quais as variáveis mais preditivas? Quais deverão ser consideras? (3) Qual a relação entre as variáveis e o resultado? (4) Existirão relações entre as variáveis que interessam? Existirão factores derivados importantes?

No âmbito da triagem em saúde, Abad-Grau et al. (2008) indicam ser fundamental ter em conta na escolha das técnicas, a sua interpretabilidade e ainda a robustez ao lidar com redundância, inconsistência, valores desconhecidos e com a incerteza. Jothi, Rashid, & Husain (2015), numa análise à aplicação de DM na área da saúde, identificam que os principais trabalhos têm como objectivo a predição e identificam como causa de resultados erróneos, precisamente, questões relacionadas com conjuntos de dados desequilibrados e com presença de valores desconhecidos.

Relembrando que nesta área muitas vezes apenas poderemos ter um de dois resultados – a vida ou morte – mais do que em qualquer outra, quando aplicadas técnicas de DM, deverá ter-se especial cuidado na análise da sensibilidade e especificidade. Por exemplo, ao analisarem-se pedidos de socorro, classificar como não-urgente um pedido urgente poderá levar à morte de um ser humano. Assim, é fundamental uma cuidadosa avaliação dos resultados.

Relativamente às medidas de avaliação de resultados, verificámos no capitulo anterior que as mais comuns são a taxa de acerto, especificidade e sensibilidade, facto confirmado por Bellazzi & Zupan (2008).

Independentemente da metodologia, técnicas e métricas adoptadas no processo de KDD, estará sempre presente a fase de divulgação dos resultados. Também nesta fase, na área da saúde, será necessário um especial cuidado, a falta de detalhe na publicação poderá conduzir a interpretações erradas. Shillabeer & Roddick (2007) comentam, a título de exemplo, a existência de uma errada generalização de um estudo que indicou que fumar não tinha influencia directa no cancro de pele, tendo sido afirmado em certas publicações que não tinha influencia no cancro, em geral, e não apenas no de pele. Em Smith & Ebrahim (2002) podem ver-se outros exemplos de que muitas vezes os media, e não só, tendem a apresentar a informação de uma forma mais atractiva, mas menos especifica e cuidada, que poderá conduzir a falsas interpretações. Identificam-se assim três passos essenciais na divulgação dos resultados: os resultados não devem ser publicados apenas com base na correlação; os resultados devem ser explicados e, finalmente, deverão ser replicados, confirmados e documentados.

Capítulo 5

Triagem de Pedidos de Assistência

A partir deste ponto será efectuada a aplicação de uma metodologia de KDD optando-se pela de Cios et al. (2000) uma vez que, como referimos no Capítulo 2, tem sido frequentemente utilizada em análises semelhantes, isto é, na área da investigação e em estudos na área da saúde (Kurgan & Musilek, 2006).

Esta metodologia envolve 6 passos – cuja ilustração aplicada a esta investigação é apresentada na Figura 1 – e onde é feita uma menção específica a 7 possíveis iterações que poderão ser necessárias para obtenção de conhecimento válido, novo e útil. Sobre estas possíveis iterações faremos referência no final de cada uma das fases.

Figura 1 – Modelo de 6 passos de Cios et al. (2000) aplicado na investigação

Para realizar esta investigação foi utilizado o software WEKA. O WEKA é software de código aberto, que disponibiliza vários algoritmos de *Data Mining* que permitem responder aos três tipos de problemas mais frequentes – classificação, regressão e *clustering* (Witten et al. 2016).

5.1. Compreensão do Domínio do Problema

Nesta primeira fase inclui-se a definição do problema, objectivos do projecto e revisão das acuais soluções para o problema. Conforme indicado anteriormente, nesta investigação propomo-nos a analisar os pedidos de socorro recebidos pelo INEM, estudando eventuais correlações entre variáveis e, recorrendo a técnicas de DM, efectuar predição do nível de prioridade desses pedidos de socorro. Actualmente, o INEM utiliza o sistema TETRICOSY® por si desenvolvido, que mediante o input de informação do operador que atende a chamada atribui um nível de prioridade à ocorrência. Está identificado que este sistema efectua accionamentos injustificados de VMER e outros meios, tornando assim importantes todas as formas de melhorar a priorização dos pedidos de socorro recebidos, sendo esse o principal objectivo desta investigação.

5.2. Compreensão dos Dados

De acordo com Cios & Kurgan (2005), na fase de compreensão dos dados inclui-se a recolha e tomada a decisão de que dados serão necessários e sob que formato. Os dados utilizados nesta investigação foram cedidos pelo próprio INEM, enviados em formato Excel, e correspondem a todos os pedidos de socorro recebidos pelo INEM nos anos de 2014 e 2015. Esta informação foi disponibilizada em resposta a um pedido de dados relativo aos pedidos de socorro recebidos pelo INEM. Após um processo de aprovação, foi enviado o conjunto de dados cujas variáveis foram seleccionadas de forma a respeitar os níveis de confidencialidade existente nesse instituto. Os dados incluem a data da ocorrência, hora, distrito, concelho, prioridade, tipo de ocorrência, faixa etária e sexo do individuo a socorrer, sendo que nesta investigação a variável que se pretende predizer será a prioridade.

Na totalidade foi-nos disponibilizada informação sobre 2.070.227 ocorrências – correspondendo 1.049.928 ao ano 2014 e 1.020.299 ao ano 2015.

Na primeira análise aos dados, verifica-se a não existência de valores nulos, isto é, valores estranhos que seriam considerados erros na base de dados. Foram retirados espaços, acentos e cedilhas, derivado o atributo dia da semana através da data, desdobrada a data em ano, mês, dia e derivado o dia da semana. Constatou-se a existência de valores desconhecidos nos atributos sexo, concelho e distrito. Com este tratamento dos dados, os atributos e seus formatos são os que se podem observar na Tabela 3.

Tabela 3 – Atributos e respectivos formatos

Atributo	Formato
Ano	Numérico
Mês	Numérico
Dia	Numérico
Dia_Semana	Nominal
Hora	Numérico
Distrito	Nominal
Concelho	Nominal
Prioridade	Nominal
Tipo de Ocorrência	Nominal
Faixa Etária	Nominal
Sexo	Nominal

Analisando detalhadamente cada um dos atributos e começando pela variável a predizer observa-se que estamos perante uma escala de quatro níveis, visto que o atributo Prioridade assume os seguintes valores: Emergente, Não-urgente, Pouco_urgente e Urgente. No capítulo Trabalho Relacionado, evidenciou-se que é frequente a existência de escalas de dois ou quatro níveis. Este facto leva a que possamos realizar, tal como Zmiri, Shahar & Taieb-Maimon, (2012), um estudo com dois tipos de escala. Partindo de uma escala inicial de quatro níveis, agrupando os dois níveis inferiores e os dois níveis superiores resultará uma escala de dois níveis. Na análise com dois níveis, será importante a análise das métricas sensibilidade e especificidade. Poderemos considerar como negativos os casos Pouco-urgentes e os Não-urgentes e em positivos os casos Urgentes e Emergentes. A sensibilidade medirá a capacidade do modelo em identificar os casos Urgentes e Emergentes e a especificidade medirá a capacidade do modelo em identificar os casos Não-urgentes e Pouco-urgentes.

Seguindo com a análise da variável a predizer, verifica-se que a grande maioria das ocorrências são classificadas como Urgentes (75%) e apenas uma muito reduzida percentagem – 1% – é classificada como Não Urgente, como se mostra na Tabela 4 e correspondendo o Gráfico 1.

Este facto leva a que tenhamos que considerar que os dados são desequilibrados, e de forma a evitar que a performance das técnicas seja prejudicada devido a este facto, como alerta Japkowicz & Stephen (2002), será adoptado um processo na fase de preparação dos dados para corrigir o desequilíbrio.

Tabela 4 – Nº de ocorrência por prioridade

Prioridade	Nº Ocorrências	%
Emergente	278.242	13%
Urgente	1.561.326	75%
Pouco_Urgente	204.344	10%
Nao_Urgente	26.315	1%
Total	**2.070.227**	**100%**

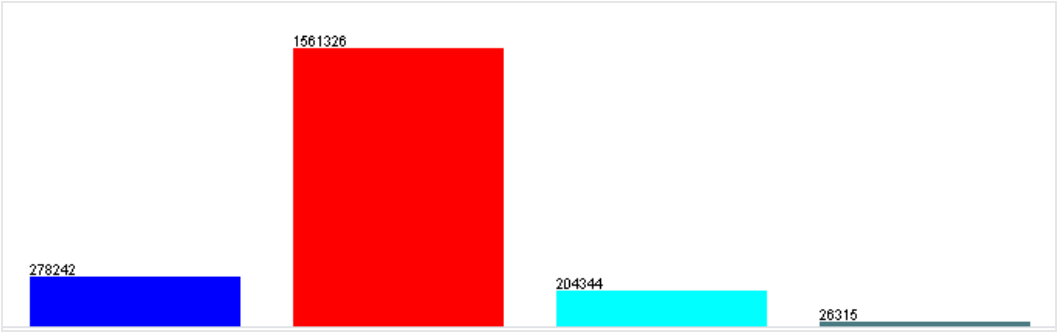

Gráfico 1 – Atributo: prioridade (classe = prioridade)

Uma vez que os dados enviados correspondem ao registo de pedidos de socorro entre Janeiro de 2014 e Dezembro de 2015, no que diz respeito ao atributo ano, os únicos valores possíveis são 2014 e 2015. Existem ocorrências em todos os meses, dias, dias de semana e horas, uma vez que os valores do atributo Mês são valores entre 1 e 12, do atributo Dia valores entre 1 e 31, para o atributo Dia da Semana, segunda a domingo e finalmente para o atributo hora valores entre 0 e 23.

Analisando os Gráfico 2 e 3 é evidente o facto de que o dia 31 tem um número inferior de dados e que das 0 às 7 horas o nº de ocorrências é claramente inferior ao das restantes.

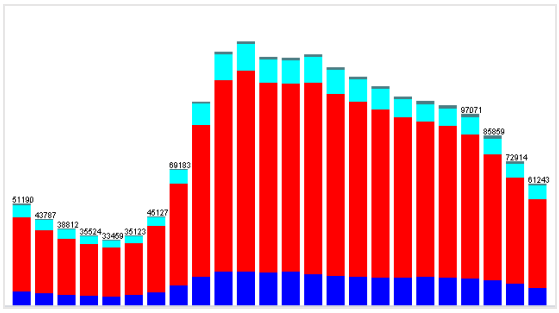

Gráfico 2 – Atributo: hora (classe = prioridade)

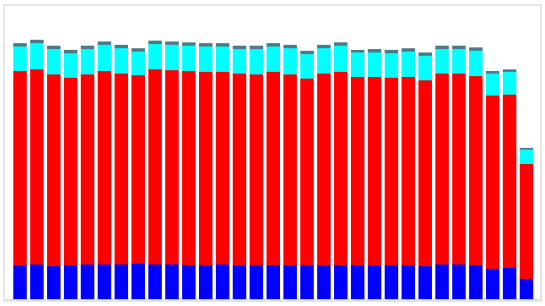

Gráfico 3 – Atributo: dia (classe = prioridade)

Relativamente aos dados é ainda de destacar que o número de registos do atributo mês com valor igual a 9 é substancialmente inferior ao dos restantes apesar da distribuição por prioridade – atributo classe – ser restante ao dos semelhantes, como se mostra no Gráfico 4. Analisando na perspectiva em que a classe é o atributo Ano, rapidamente se percebe que não estão disponíveis dados de Setembro de 2015, como comprova o Gráfico 5.

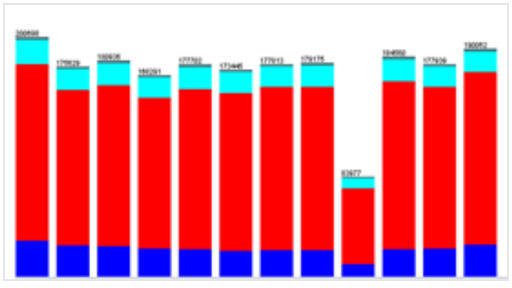

Gráfico 4 – Atributo: mês (classe = prioridade)

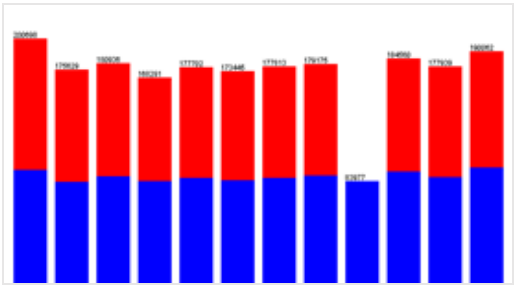

Gráfico 5 – Atributo: mês (classe = ano)

Igualmente se verifica que em todos os concelhos e consequentemente em todos os distritos de Portugal Continental foram efectuados pedidos de socorro, uma vez que surgem os 278 concelhos e 20 distritos tendo Lisboa e o Porto um número de ocorrências evidentemente superior.

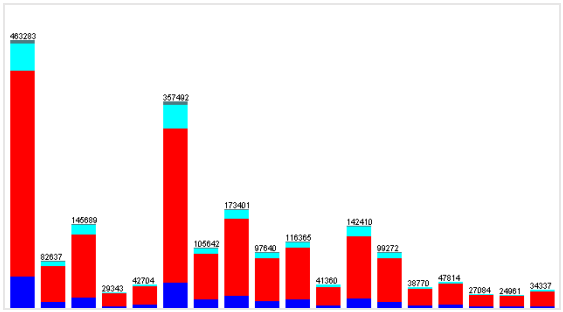

Gráfico 6 - Atributo: distrito (classe = Prioridade)

Constata-se ainda, relativamente a estes dois atributos, que existem 23 instâncias sem Concelho e Distrito.

Quanto ao atributo Sexo verifica-se uma distribuição uniforme e que o número de instâncias sem esta informação representa 0% do conjunto total de dados (Tabela 5).

Tabela 5 – Atributo: sexo

Sexo	Número ocorrências	%
Feminino	1.096.195	53%
Masculino	965.192	47%
Valores desconhecidos	8.840	0%
Total Geral	**2.070.227**	100%

A idade da pessoa a socorrer é nos informada através do atributo Faixa Etária, na Tabela 6demonstra-se como é claramente superior o número de ocorrências com pessoas idosas (57%) e mínimo o de lactentes (1%).

Tabela 6 – Atributo: faixa etária

Faixa Etária	Número ocorrências	%
Lactente_[0_-_1_ano]	16.927	1%
Crianca_[2_-_17_anos]	138.911	7%
Jovem_[18_-_30_anos]	195.989	9%
Adulto_[31_-_59_anos]	536.450	26%
Idoso_(>=_60_anos)	1.181.950	57%
Total Geral	**2.070.227**	**100%**

Finalmente, o último atributo a analisar é o tipo de ocorrência onde existe igualmente uma distribuição gravemente desequilibrada. A grande maioria dos pedidos de socorro (68%) é classificado com o tipo Doença súbita, seguindo-se o tipo Queda (16%). Os restantes dez tipos são igualmente pouco frequentes.

Tabela 7 – Atributo: tipo de ocorrência

Tipo de Ocorrência	Número ocorrências	%
Pedido_Apoio_Diferenciado	42.382	2%
Queimadura	3.741	0%
Doenca_Subita	1.401.230	68%
Agressao	29.900	1%
Queda	323.745	16%
Paragem_Cardiorrespiratoria	31.253	2%
Gravidez/Parto	21.027	1%
Intoxicacao	66.168	3%
Acidente_de_Viacao	52.320	3%
Psiquatria	55.171	3%
Pediatria	42.648	2%
Afogamento	642	0%
Total Geral	**2.070.227**	**100%**

Na metodologia adoptada, é identificada uma possível iteração para a fase anterior – compreensão do domínio do problema (Cios & Kurgan, 2005). De facto, e uma vez que tínhamos conhecimento de que era possível, solicitámos uma visita ao INEM. Considerámos que a mesma seria importante para obter mais informação da forma como são recepcionados os pedidos de assistência e como os operadores do INEM, perante cada situação, lhes atribuem determinado nível de prioridade, resumidamente, como é efectuada actualmente a triagem.

5.3. Preparação dos Dados

A preparação dos dados é uma fase crucial no processo de descoberta de conhecimento. É nesta fase que deverá ser efectuada a escolha dos dados a utilizar como input, avaliação dos atributos mediante análise de eventuais correlações, limpeza dos dados, remoção de ruido, amostragem, etc. (Cios & Kurgan, 2005).

Nesta investigação, no que diz respeito à limpeza dos dados e considerando-a relativa aos valores desconhecidos, tendo em conta que estes representam 0% dos dados optou-se por mantê-los, pois não influenciarão os resultados. Foram ainda convertidos todos os atributos, que se encontravam em formato numérico, em atributos nominais.

A avaliação dos atributos, essencialmente utilizada na selecção de atributos quando existe elevada dimensionalidade, foi efectuada com recurso ao método *CorrelationAttributeEval* – que permite criar um ranking dos atributos mediante as suas correlações (Pearson's) com o atributo definido como classe – e o *GainRatioAttributeEval* – que permite igualmente criar um ranking, mas com base no *Gain Ratio* medindo o valor de cada atributo relativamente ao atributo definido como classe.

Foram seleccionados estes dois métodos devido ao facto do coeficiente de Correlação de Pearson (Pearson, 1895) ser uma das mais frequentemente métricas de análise e o *Gain Racio* ser utilizado na construção de árvores de decisão (Witten et al. 2016). A correlação detectará apenas dependências lineares entre a variável e a classe (Guyon & Elisseeff, 2003) e os resultados poderão variar entre -1 e 1, sendo resultado 0 obtido quando não existe qualquer correlação linear. O *Gain Ratio* ajusta o método de Ganho de Informação de forma a suprimir a tendência deste de seleccionar atributos com maior número de valores, pois esse tipo de atributos poderá ter maior número de situações em que um valor tenha apenas uma classe. Este é o método utilizado pelo algoritmo C4.5 para escolha dos nós na criação das Árvores de Decisão (Hand et al. 2001).

Para efectuar esta análise, tendo em conta os limites computacionais do equipamento utilizado e o presente volume de dados – 2.070.227 com 11 atributos – foi necessário criar uma amostra com 20% dos dados. Para gerar esta amostra, foi criado um novo atributo que resulta simplesmente da concatenação de Ano com Mês, deste modo

possibilitando que fosse então definida uma distribuição uniforme da amostra com base neste novo atributo.

Tabela 8 – Avaliação dos atributos: ganho de informação e correlação

	Ganho de Informação		Correlação	
1	0,078226	Tipo_Ocorrencia	0,03925	Tipo_Ocorrencia
2	0,00682	Faixa_Etaria	0,02727	Ano
3	0,001226	Concelho	0,01735	Sexo
4	0,000895	Distrito	0,01684	Faixa_Etaria
5	0,000738	Ano	0,00896	Distrito
6	0,000682	Sexo	0,00683	AnoMes
7	0,000382	AnoMes	0,00621	Mes
8	0,000335	Hora	0,0046	Hora
9	0,000203	Mes	0,00444	Concelho
10	4,62E-05	Dia	0,00189	Dia_Semana
11	4,53E-05	Dia_Semana	0,00185	Dia

O resultado da aplicação dos dois métodos encontra-se acima na Tabela 8, onde são visíveis as baixas relações entre os atributos com a classe a predizer, a Prioridade da ocorrência. O ranking gerado, quer pela correlação quer pelo *Gain Ratio,* classifica a variável Tipo de Ocorrência como a melhor variável preditora. No que diz respeito às restantes variáveis, os métodos não fornecem a mesma classificação, dado que as variáveis não surgem nas mesmas posições.

Relativamente ao atributo Ano, de acordo com um dos métodos seria considerado um atributo importante, visto surgir como o segundo atributo com maior correlação com a classe, não sendo visível esta importância de acordo com o método G*ain Ratio.* Este facto que será estudado na aplicação das técnicas e discutido de acordo com os resultados.

Verifica-se ainda que na análise do *Gain Ratio* em segundo e terceiro lugar são apresentadas as variáveis Faixa-etária e Concelho, respectivamente, enquanto que na análise da correlação estas posições são atribuídas à variável Ano e Sexo.

Relativamente aos atributos Dia, Dia-da-Semana, Mês e Hora ambos os métodos indicam que a sua relação com a classe prioridade é muito reduzida.

Recordando os estudos abordados na secção anterior, nestes verificámos que os atributos utilizados possuíam um caracter essencialmente clínico, o que poderia indicar

desde logo que a informação que temos disponível nesta investigação não será eventualmente a mais relevante para o objectivo de predição da prioridade de um pedido de socorro efectuado ao INEM.

Apesar destes resultados, e tendo em conta que não seria possível obter outro tipo de informação por questões de confidencialidade – dificuldade frequente neste tipo de investigações, conforme abordámos em secção anterior – todos os atributos foram utilizados na seguinte fase do processo de extracção de conhecimento, a fase de *Data Mining*.

Nesta fase de *Data Mining* foram efectuadas quatro abordagens sendo desta forma possível verificar se determinados factores influenciam os resultados e garantir a validade dos mesmos. Existem três aspectos comuns às quatro abordagens: a partir do conjunto inicial de dados foram criados três subconjuntos: de treino, desenvolvimento e teste que, como referido anteriormente, permite validar os algoritmos. O segundo aspecto comum é o facto de que se garante que o conjunto de teste possui informação cronologicamente posterior à dos restantes conjuntos. E, finalmente, o conjunto de treino é dividido em 5 amostras de forma a verificar se existe influencia da selecção dos dados de treino no desempenho das técnicas e deste modo podendo efectuar-se uma correcta validação dos resultados.

Na Abordagem 1, como se resume na Tabela 9, o conjunto total de dados foi ordenado por data de ocorrência. As 1.049.928 do ano 2014 foram utilizadas para treino. As restantes 1.020.299, que se encontravam ordenadas, foram separadas uma a uma de modo a garantir que os conjuntos, de desenvolvimento e teste, fossem abrangentes e semelhantes. Após esta divisão, foi retirado o atributo Ano para que não haja relação temporal, uma vez que se pretende que a classificação possa ser aplicada no futuro, não podendo então estar o atributo Ano em causa e a influenciar a classificação de determinado pedido de assistência médica.

Tabela 9 – Divisão dos dados na Abordagem 1

Conjuntos	Abordagem 1	
treino	1.049.928	Ocorrências de 2014 sem atributo ano
desenvolvimento	510.150	Primeira metade das ocorrências de 2015 sem atributo ano
teste	510.149	Segunda metade das ocorrências de 2015 sem atributo ano
Total	**2.070.227**	

Na abordagem 2, utilizando os dados não ordenados reservaram-se os relativos a Maio e Dezembro de 2015 para teste, de modo a assegurar que se encontravam presentes as 4 estações do ano e que eram empregues na fase de treino um grande volume de dados. Os restantes dados, que não se encontravam ordenados, foram separados em dois conjuntos de igual proporção entre treino e desenvolvimento, conforme Tabela 10, do mesmo modo que na abordagem anterior, de forma a garantir que os conjuntos fossem uniformes.

A ordenação dos dados foi desconsiderada nesta abordagem devido ao facto de termos verificado na análise da relação entre os atributos Dia, Dia-da-Semana, Mês e Hora e a classe se encontravam nas mais baixas posições do ranking. No entanto, e relembrando que o atributo Ano surgiu em segundo lugar na análise da correlação, nesta segunda abordagem este facto será alvo de análise, sendo as técnicas aplicadas aos conjuntos de dados de duas formas – uma com o atributo Ano e outra sem este atributo.

Tabela 10 – Divisão dos dados na Abordagem 2

Conjuntos		Abordagem 2
treino	714.125	50% das ocorrências entre Janeiro de 2014 a Abril de 2015
desenvolvimento	714.126	50% das ocorrências entre Janeiro de 2014 a Abril de 2015
teste	641.976	Ocorrências de Maio a Dezembro de 2015
Total	**2.070.227**	

Na Abordagem 3 pretende-se efectuar uma análise semelhante à realizada por Zmiri, Shahar & Taieb-Maimon, (2012), isto é, partindo de um problema com quatro classes efectua-se uma análise considerando apenas duas classes. Para tal a escala de prioridade inicial – Emergente, Urgente, Pouco Urgente e Não Urgente – foi dividida em dois níveis, agrupando-se Emergente com Urgente e Pouco Urgente com Não Urgente, através de uma simples concatenação dos valores da variável Prioridade. Com esta abordagem a variável Prioridade passa a ter a distribuição demonstrada na Tabela 11.

Tabela 11 – Variável prioridade na Abordagem 3

Prioridade	Nº Ocorrências	%
Emergente + Urgente	1.839.568	89%
Pouco_Urgente + Nao_Urgente	230.659	11%
Total	**2.070.227**	**100%**

Finalmente, a Abordagem 4 consiste na aplicação de uma matriz de custos aos resultados obtidos na Abordagem 1, recalculando a taxa de acerto para cada técnica considerando penalizações diferentes aos erros de classificação efectuados pelos algoritmos.

Matriz de Custos					
a	**b**	**c**	**d**	**<-- classificação WEKA**	
-	0,5	2	2	**a**	**Emergente**
0,5	-	1	2	**b**	**Urgente**
1	0,5	-	0,5	**c**	**Pouco_Urgente**
1	1	0,5	-	**d**	**Nao_Urgente**

Figura 2 – Matriz de Custos - Abordagem 4

Esta matriz, apresentada na Figura 2, atribui penalizações às classificações incorrectas, o que conduzirá a que se mantenha o número de instancias correctamente classificadas, mas fruto das ponderações, o número de instâncias consideradas como incorrectamente classificadas alterar-se-á, conduzindo a uma taxa de acerto diferente. Nestas penalizações considera-se mais gravoso a sub-triagem do que a sobre-triagem, dado que na sub-triagem estão em causa vidas humanas enquanto que na sobre-triagem estão apenas em causa o desperdício de recursos. Foi solicitada a validação do INEM à matriz apresentada, no entanto não foi possível concluir se é a certa, tendo sido referido que a penalização superior na sub-triagem será de facto uma consideração correcta, mas a ordem de grandeza das diferentes penalizações será difícil obter consenso.

A matriz será aplicada nas matrizes confusão das quatro técnicas obtidas na Abordagem 1, sendo posteriormente calculada uma nova taxa de acerto.

Com estas quatro abordagens poderemos analisar se alguns aspectos influenciam os resultados, nomeadamente na Abordagem 2, em que se analisará a influência do atributo Ano, na Abordagem 3 em que se verifica a influencia do número de classes a predizer e, ainda, na Abordagem 4 onde será efectuada a análise do custo real da predição efectuada pelos algoritmos.

Para realizar estas quatro abordagens, e ainda na fase de preparação dos dados, será necessário corrigir o desequilíbrio identificado na fase de compreensão dos dados. Conforme demonstrado anteriormente na Tabela 4, 75% das ocorrências foram classificadas como Urgente, 13% como Emergentes, 10% como Pouco-Urgente e

apenas 1% como Não-Urgente. De entre as formas para lidar com conjuntos de dados desequilibrados, as mais comuns são a sub-amostragem (*undersampling*) e sobre-amostragem (*Oversampling*) (Grzymala-Busse, J. W., 2005). Nesta investigação cujo volume inicial de dados é superior a 2 milhões de instâncias e tendo a classe a predizer – Prioridade – como valor menos frequente os casos Não-Urgentes com 26.315 ocorrências, a forma escolhida foi a sub-amostragem, evitando a utilização de dados não reais.

Para assegurar que a amostra utilizada representa com precisão o conjunto total de dados foram, como referido anteriormente, criadas cinco amostras para cada abordagem, através da função sub-amostragem. Essa função foi parametrizada de modo a garantir uma distribuição uniforme da classe a predizer (Prioridade) e para tal o conjunto de dados obtido terá 6% do volume inicial.

Tabela 12 – Volume de dados após sub-amostragem

	Abordagem 1	Abordagem 2	Abordagem 3
Treino (por amostra)	61.254	42.844	61.254
Desenvolvimento	510.150	714.126	510.150
Teste	510.149	714.126	510.149

Conforme se resume na Tabela 12, na Abordagem 1 e 3 ao invés de na fase de treino serem utilizadas 1.049.928 instâncias, teremos 5 amostras com 61.254 instâncias cada, e na Abordagem 2 ao invés de 714.125 teremos 42.844 em cada uma das 5 amostras, ficando garantido em qualquer uma delas que a classe a predizer se encontra equilibrada. Não é apresentado o volume de dados da Abordagem 4 uma vez que consiste na aplicação de uma matriz de custos aos resultados obtidos na Abordagem 1.

Na metodologia escolhida, é identificada uma possível iteração para a fase anterior – compreensão dos dados (Cios & Kurgan, 2005). Nessa fase inclui-se a etapa de recolha dos dados, e de facto quando efectuada a análise da relação entre os atributos e a classe a predizer, tendo em conta que os resultados apontavam para uma fraca relação entre os mesmos, caso fosse possível, teríamos optado pela recolha de novos dados onde se incluíssem outros atributos cuja relação com a classe fosse significativa. Uma vez que a cedência por parte do INEM de outro conjunto de dados não seria possível, as próximas etapas serão efectuadas com os dados disponíveis.

5.4. *Data Mining*

A fase de *Data Mining* é referenciada por Cios & Kurgan (2005) como sendo uma fase chave do processo de KDD, sendo nesta que de facto o novo conhecimento é descoberto. Cios & Kurgan (2005) salientam que, apesar da importância desta fase, geralmente exige menos tempo que a fase anterior, a de preparação dos dados, apontando como sendo uma das maiores dificuldades a falta de escalabilidade das técnicas de DM factor importante tendo presente o volume de dados em causa.

Na análise aos trabalhos relacionados em secção anterior, concluímos que as principais técnicas empregues em investigações similares são Árvores de Decisão, Redes Neuronais Artificiais, SVMs, Naïve Bayes e Redes Bayesianas.

Recordando que, de acordo com Abad-Grau et al. (2008), factores essenciais na escolha da técnica são a sua robustez ao lidar com redundância e valores desconhecidos, mas essencialmente a sua interpretabilidade, e por esse motivo uma das técnicas a aplicar será o J48, a implementação do algoritmo C4.5 de Árvores de Decisão no WEKA.

Com a mesma justificação indicada por Peck, et al. (2011), iremos igualmente utilizar a técnica Naïve Bayes (John & Langley, 1995) pela sua simplicidade.

No que diz respeito à selecção de uma técnica considerada como mais eficiente, optámos por recorrer a SVM em detrimento das RNA por apresentar maior escalabilidade (Bolchini & Cassano, 2014) e por ser identificada como popular em KDD na área da saúde (Esfandiari et al., 2014). No WEKA foram empregues os algoritmos LibSVM (Chang & Lin, 2011) e SMO (Platt, 1998) que se distinguem essencialmente no método de resolução do problema quadrático, SMO implementa a versão de Platt enquanto que a LibSVM considera a SMO-type proposta por Fan et al. (2005).

Na aplicação das quatro técnicas – Naïve Bayes, J48, LibSVM e SMO – foram mantidos os parâmetros por omissão da versão 3.9.1 do WEKA, e cujas imagens estão disponíveis nos Anexos (Figura 3, 6 e 7). Importa destacar que nestas parametrizações o algoritmo J48 tem definido um número mínimo de instancias por folha 2 e confiança de 0,25. Quanto aos algoritmos SVM, o LibSVM tem parametrizado por omissão o tipo S-SCV, C=1,0, e kernel do tipo rbf *(radial basis function)*, enquanto que o SMO kernel do tipo polinomial e igualmente C=1,0.

Numa primeira fase, utilizámos as quatro técnicas em cada uma das cinco amostras de treino, considerando o respectivo conjunto de desenvolvimento de acordo com cada abordagem e apuram-se os resultados apresentados na tabela seguinte.

Tabela 13 – Taxas de acerto e tempos médios de execução fase de desenvolvimento

	Técnica	Amostra 1	Amostra 2	Amostra 3	Amostra 4	Amostra 5	Tempo médio
ABORDAGEM 1	**J48**	31,86	36,32	34,19	32,88	35,87	00:00:14
	Naïve Bayes	34,25	35,91	35,28	33,96	36,21	00:00:22
	SMO	36,59	38,93	37,57	35,69	37,30	01:48:07
	LibSVM	52,70	52,60	52,58	52,55	51,69	01:12:28
ABORDAGEM 2	**Com atributo Ano**						
	J48	35,54	35,73	36,38	35,99	35,96	00:02:17
	Naïve Bayes	36,05	35,30	35,46	35,34	35,73	00:00:12
	SMO	40,26	40,94	40,11	39,91	39,94	03:30:29
	LibSVM	50,65	52,22	50,62	50,62	50,66	01:25:41
	Sem atributo Ano						
	J48	35,62	35,95	36,38	35,93	36,17	00:00:03
	Naïve Bayes	35,93	35,35	35,36	35,33	35,67	00:00:05
	SMO	40,26	40,92	39,67	39,65	40,09	03:00:16
	LibSVM	50,65	52,22	50,62	50,62	50,65	02:55:55
ABORDAGEM 3	**J48**	54,00	54,01	53,50	54,39	54,16	00:00:01
	Naïve Bayes	59,58	59,79	59,44	59,56	59,18	00:00:02
	SMO	61,79	62,08	61,12	61,00	62,05	02:47:04
	LibSVM	60,36	60,36	60,36	60,36	60,36	01:16:01

Apesar do volume de dados de cada amostra no treino ser substancialmente inferior ao inicial, para desenvolvimento foi utilizado um grande volume de dados. Nesta primeira fase foram necessárias 89 Horas, 56 minutos e 33 segundos para correr todas as técnicas para as cinco amostras, sendo evidente que a escalabilidade das técnicas de SVM é inferior à das restantes, como se mostra na Tabela 13 – Taxas de acerto e tempos médios de execução fase de desenvolvimento.

Considerou-se importante mencionar os tempos de execução porque, tal como referem Cios & Kurgan (2005), uma das dificuldades da utilização de DM é a falta de escalabilidade de algumas técnicas e numa utilização real a quantidade de dados a tratar é de grande dimensão.

Conforme se observa na Tabela 13, as taxas de acerto de cada algoritmo não sofrem grandes oscilações consoante a amostra utilizada no treino. Por este motivo, na fase

seguinte, em que se utilizam os conjuntos de teste, foram utilizadas para treino as primeiras amostras de cada abordagem da fase de desenvolvimento. Os resultados (taxas de acerto) de cada técnica na fase de teste de cada abordagem são apresentados na Tabela 14.

Tabela 14 – Taxas de acerto na fase de teste

	ABORDAGEM 1	ABORDAGEM 2		ABORDAGEM 3
		Com Ano	Sem Ano	
J48	31,81	35,35	35,17	53,87
Naïve Bayes	34,26	41,15	36,44	59,54
SMO	36,62	51,31	39,61	61,68
LibSVM	52,66	49,45	49,45	60,28

De acordo com a metodologia seguida nesta investigação, nesta fase poderão ocorrer três iterações (Cios & Kurgan, 2005). No entanto, tendo em conta que a análise dos resultados será aprofundada na fase seguinte, será nela que faremos referência à necessidade de existirem essas iterações.

5.5. Avaliação do Conhecimento Descoberto

Cios & Kurgan (2005) referem que nesta fase é efectuada a avaliação se a informação obtida é nova e interessante, interpretação pelos especialistas da área dos resultados para confirmação se existe novo conhecimento descoberto.

Contudo, como não seria possível avaliar a informação conjuntamente com especialistas da área, nesta secção será incluída uma análise relacionada com a fase anterior de DM, que consiste na apresentação, para cada uma das abordagens, dos resultados de cada uma das técnicas utilizadas, efectuando uma análise comparativa entre as diferentes abordagens e os Trabalho Relacionados.

Serão apresentadas as métricas anteriormente referidas – taxa de acerto, sensibilidade e especificidade – sendo as últimas importantes na Abordagem 3 por se tratar de uma classificação com apenas 2 classes, porém através da Abordagem 4 mostrar-se-á a sua relevância em classificações de 4 classes. Será também apresentada, e porque estamos perante uma investigação em que o tema é a triagem, uma análise às classificações incorrectas de modo a serem avaliados os casos de sub-triagem e sobre-triagem.

5.5.1. Abordagem 1

Os dados utilizados nesta abordagem foram preparados considerando o ano 2014 para treino e dividindo o ano 2015 para desenvolvimento e teste.

Tabela 15 – Taxas de acerto Abordagem 1

Técnica	Desenvolvimento					Teste
	Amostra 1	Amostra 2	Amostra 3	Amostra 4	Amostra 5	
J48	31,86	36,32	34,19	32,88	35,87	31,81
Naïve Bayes	34,25	35,91	35,28	33,96	36,21	34,26
SMO	36,59	38,93	37,57	35,69	37,30	36,62
LibSVM	52,70	52,60	52,58	52,55	51,69	52,66

Na Tabela 15 estão apresentadas as taxas de acerto obtidas nesta abordagem, podendo verificar-se que as variações não são significativas em função dos conjuntos de dados utilizados, por esse motivo a análise será aprofundada aos resultados obtidos no teste.

Tabela 16 – Resultados Abordagem 1 - Teste

J48

a	b	c	d		←	classificação WEKA	Especificidade	Sensibilidade
34.031	8.891	21.944	1.771	66.637	a	Emergente	77,20%	51,07%
91.539	99.286	164.366	35.685	390.876	b	Urgente	85,75%	25,40%
8.411	7.334	26.694	4.086	46.525	c	Pouco_Urgente	59,41%	57,38%
1.192	772	1.863	2.284	6.111	d	Nao_Urgente	91,76%	37,38%
135.173	116.283	214.867	43.826	**510.149**			**Taxa de Acerto**	**31,81%**

Naïve Bayes

a	b	c	d		←	classificação WEKA	Especificidade	Sensibilidade
35.046	11.943	15.410	4.238	66.637	a	Emergente	74,84%	52,59%
99.996	116.155	126.159	48.566	390.876	b	Urgente	81,39%	29,72%
10.353	9.284	21.082	5.806	46.525	c	Pouco_Urgente	69,17%	45,31%
1.257	971	1.374	2.509	6.111	d	Nao_Urgente	88,37%	41,06%
146.652	138.353	164.025	61.119	**510.149**			**Taxa de Acerto**	**34,26%**

SMO

a	b	c	d		←	classificação WEKA	Especificidade	Sensibilidade
33.768	13.420	17.634	1.815	66.637	a	Emergente	76,79%	50,67%
92.639	127.439	138.327	32.471	390.876	b	Urgente	79,07%	32,60%
9.053	10.488	23.326	3.658	46.525	c	Pouco_Urgente	66,03%	50,14%
1.247	1.061	1.542	2.261	6.111	d	Nao_Urgente	92,47%	37,00%
136.707	152.408	180.829	40.205	**510.149**			**Taxa de Acerto**	**36,62%**

LibSVM

a	b	c	d		←	classificação WEKA	Especificidade	Sensibilidade
17.795	32.408	14.859	1.575	66.637	a	Emergente	99,78%	26,70%
522	227.992	127.671	34.691	390.876	b	Urgente	52,99%	58,33%
33	21.846	20.571	4.075	46.525	c	Pouco_Urgente	68,91%	44,21%
419	1.815	1.608	2.269	6.111	d	Nao_Urgente	92,00%	37,13%
18.769	284.061	164.709	42.610	**510.149**			**Taxa de Acerto**	**52,66%**

A representação da Árvore de Decisão do algoritmo J48, exposta na Figura 7 do Anexo A, é pouco interpretável (número de folhas 4.496 e tamanho da árvore 4.809) facto que deverá ser considerado como um aspecto negativo pois, conforme referido no Capítulo 4, é importante que o modelo seja transparente e perceptível. Confirma-se que, conforme análise do *Gain Ratio* efectuada na Preparação dos Dados, o primeiro nó tem como atributo o Tipo de Ocorrência.

No que diz respeito às métricas, qualquer uma das técnicas apresentou baixas taxas de acerto. Ainda que os resultados não possam ser directamente comparáveis com os estudos em Trabalho Relacionado, visto estarmos perante conjuntos de dados e variáveis diferentes, nenhum dos classificadores atingiu valores semelhantes, mesmo que apenas se compare com estudos de 4 classes.

Pode observar-se que o algoritmo J48 efectuou menos classificações correctas que os restantes, com uma taxa de acerto de 31,81%, seguido pelo Naïve Bayes com 34,26%. As técnicas de SVM foram de facto mais eficientes, nos entanto os resultados entre os dois algoritmos são substancialmente diferentes, tendo o SMO atingido uma taxa de acerto de 36,62%, relativamente próxima à das técnicas anteriores enquanto que o LibSVM atinge uma taxa de 52,66%.

Quanto à análise das classificações incorrectas, na Tabela 17 foram divididas as classificações incorrectas de cada algoritmo como sendo casos de sub-triagem ou sobre-triagem podendo verificar-se que em todas as técnicas o número de casos de sub-triagem é superior ao de sobre-triagem.

Analisando algoritmo a algoritmo, os resultados não são iguais aos relativos às taxas de acerto. Verifica-se assim que apesar de no total de instâncias o algoritmo LibSVM classificar correctamente 52,66%, nos restantes 47,34% uma grande maioria (89%) são casos de sub-triagem, e por isso devem ser considerados erros de maior gravidade.

Tabela 17 – Classificações incorrectas na Abordagem 1

Técnica	Instancias classificadas incorrectamente				
	sub-triagem	%	sobre-triagem	%	total
J48	236.743	68%	111.111	32%	**347.854**
Naïve Bayes	212.122	63%	123.235	37%	**335.357**
SMO	207.325	64%	116.030	36%	**323.355**
LibSVM	215.279	89%	26.243	11%	**241.522**

5.5.2. Abordagem 2

Nesta abordagem foi avaliada a influência do atributo Ano e a divisão dos dados consistiu na separação de Janeiro de 2014 a Abril 2015 em dois conjuntos para treino e desenvolvimento e os restantes dados – Maio 2015 a Dezembro 2015 – para teste.

Tabela 18 – Taxas de acerto Abordagem 2

Com atributo Ano						
	Desenvolvimento					
Técnica	Amostra 1	Amostra 2	Amostra 3	Amostra 4	Amostra 5	Teste
J48	35,54	35,73	36,38	35,99	35,96	35,35
Naïve Bayes	36,05	35,3	35,46	35,34	35,73	41,15
SMO	40,26	40,94	40,11	39,91	39,94	51,31
LibSVM	50,65	52,22	50,62	50,62	50,66	49,45
Sem atributo Ano						
	Desenvolvimento					
Técnica	Amostra 1	Amostra 2	Amostra 3	Amostra 4	Amostra 5	Teste
J48	35,62	35,95	36,38	35,93	36,17	35,17
Naïve Bayes	35,93	35,35	35,36	35,33	35,67	36,44
SMO	40,26	40,92	39,67	39,65	40,09	39,61
LibSVM	50,65	52,22	50,62	50,62	50,65	49,45

Na Tabela 18 são apresentadas as taxas de acerto obtidas, e verifica-se que o atributo Ano não tem influência significativa nos resultados, inclusivamente as taxas de acerto do algoritmo LibSVM são aproximadamente iguais. Este facto poderá considerar-se positivo, uma vez que que se pretende aplicar um sistema a ocorrências futuras não podendo o Ano distinguir o nível de prioridade atribuído a um pedido de assistência médica.

Com esta abordagem, uma vez mais se verifica que as oscilações das taxas de acerto não são relevantes na fase de desenvolvimento e comparativamente com o teste, sendo novamente por este motivo que prosseguiremos a análise aos resultados do teste e considerando a versão em que não se considera o atributo Ano.

Tabela 19 – Resultados Abordagem 2 – Teste sem atributo Ano

Modelo 1 - J48 - Teste								
a	**b**	**c**	**d**		←	**classificação WEKA**	**Especificidade**	**Sensibilidade**
40.110	16.679	17.111	6.895	80.795	**a**	**Emergente**	76,91%	49,64%
115.086	159.754	141.474	78.623	494.937	**b**	**Urgente**	78,50%	32,28%
13.040	13.557	22.613	9.282	58.492	**c**	**Pouco_Urgente**	72,55%	38,66%
1.474	1.375	1.594	3.309	7.752	**d**	**Nao_Urgente**	85,05%	42,69%
169.710	191.365	182.792	98.109	**641.976**			**Taxa de Acerto**	**35,17%**

Modelo 1 - Naïve Bayes - Teste								
a	**b**	**c**	**d**		←	**classificação WEKA**	**Especificidade**	**Sensibilidade**
36.332	17.967	18.727	7.769	80.795	**a**	**Emergente**	81,45%	44,97%
93.254	168.649	153.358	79.676	494.937	**b**	**Urgente**	77,20%	34,07%
9.693	14.176	25.506	9.117	58.492	**c**	**Pouco_Urgente**	70,21%	43,61%
1.175	1.375	1.732	3.470	7.752	**d**	**Nao_Urgente**	84,77%	44,76%
140.454	202.167	199.323	100.032	**641.976**			**Taxa de Acerto**	**36,44%**

Modelo 1 - SMO - Teste								
a	**b**	**c**	**d**		←	**classificação WEKA**	**Especificidade**	**Sensibilidade**
35.918	21.001	20.615	3.261	80.795	**a**	**Emergente**	82,22%	44,46%
89.592	187.679	169.528	48.138	494.937	**b**	**Urgente**	73,56%	37,92%
8.954	16.365	27.668	5.505	58.492	**c**	**Pouco_Urgente**	67,07%	47,30%
1.214	1.510	2.010	3.018	7.752	**d**	**Nao_Urgente**	91,03%	38,93%
135.678	226.555	219.821	59.922	**641.976**			**Taxa de Acerto**	**39,61%**

Modelo 1 - LibSVM - Teste								
a	**b**	**c**	**d**		←	**classificação WEKA**	**Especificidade**	**Sensibilidade**
23.987	35.642	18.912	2.254	80.795	**a**	**Emergente**	97,15%	29,69%
15.393	264.778	168.018	46.748	494.937	**b**	**Urgente**	55,92%	53,50%
15	27.123	25.748	5.606	58.492	**c**	**Pouco_Urgente**	67,59%	44,02%
607	2.052	2.174	2.919	7.752	**d**	**Nao_Urgente**	91,39%	37,65%
40.002	329.595	214.852	57.527	**641.976**			**Taxa de Acerto**	**49,45%**

De forma análoga à abordagem anterior, a representação da Árvore de Decisão não é de fácil interpretação, conforme Figura 8 dos Anexos, sendo inclusivamente de maior dimensão, com um número de folhas 17.887 e tamanho da árvore 18.902.

Relativamente às taxas de acerto, os resultados são semelhantes aos verificados anteriormente, no entanto ligeiramente superiores. Os classificadores J48 e Naïve Bayes continuam a ter as taxas mais baixas, mas já com 35,17% e 36,44% respectivamente. Quanto aos algoritmos de SVM, o classificador SMO apresenta uma taxa consideravelmente mais elevada, tendo 49,45% já o LibSVM foi o único a piorar o desempenho com esta abordagem, dado ter uma taxa de acerto de 49,45% enquanto que na abordagem anterior apresentou 52,66%.

No que diz respeito à sensibilidade, destaca-se que J48 foi o algoritmo mais sensível na prioridade Emergente, isto é, de entre todos os casos que são na realidade Emergentes, J48 consegue identificar um maior número.

Tabela 20 – Classificações incorrectas na Abordagem 2

Técnica	Instancias classificadas incorrectamente				
	sub-triagem	%	sobre-triagem	%	total
J48	270.064	65%	146.126	35%	**416.190**
Naïve Bayes	286.614	70%	121.405	30%	**408.019**
SMO	268.048	69%	86.353	31%	**387.693**
LibSVM	277.180	85%	47.364	15%	**324.544**

Com esta abordagem aos dados, uma vez mais os erros de classificação são predominantemente casos de sub-triagem, sendo apenas importante fazer notar que o classificador com taxa de acerto mais baixa, J48, nesta análise é o que apresenta melhor prestação.

5.5.3. Abordagem 3

Esta abordagem tem por base um problema de classificação com apenas duas classes Emergente_Urgente e Pouco_Urgente_Nao_Urgente, de modo a efectuar uma análise semelhante à efectuada por Zmiri & Taieb-Maimon, (2012).

Tabela 21 – Taxas de acerto Abordagem 3

Técnica	Amostra 1	Amostra 2	Amostra 3	Amostra 4	Amostra 5	Teste
J48	53,9959	54,0114	53,4952	54,3889	54,1552	53,8672
Naïve Bayes	59,5778	59,7856	59,4386	59,5607	59,1828	59,5357
SMO	61,7918	62,0808	61,1240	61,0046	62,0533	61.6757
LibSVM	60,3599	60,3599	60,3599	60,3599	60,3599	60,2773

Nesta nova classificação, não se verificam alterações relevante nos resultados consoante as amostras e assim, tal como anteriormente, vamos incidir uma vez mais a avaliação apenas no conjunto de teste e cujos resultados foram os seguintes:

Tabela 22 – Resultados Abordagem 3 - Teste

J48						
a	b		←	classificação WEKA	Especificidade	Sensibilidade
240.145	217.368	457.513	a	Emergente_Urgente	65,84%	52,49%
17.978	34.658	52.636	b	Pouco_Urgente_Nao_Urgente	52,49%	65,84%
258.123	252.026	**510.149**			**Taxa de Acerto**	**53,87%**

Naïve Bayes						
a	b		←	classificação WEKA	Especificidade	Sensibilidade
273.643	183.870	457.513	a	Emergente_Urgente	57,14%	59,81%
22.558	30.078	52.636	b	Pouco_Urgente_Nao_Urgente	59,81%	57,14%
296.201	213.948	**510.149**			**Taxa de Acerto**	**59,54%**

SMO						
a	b		←	classificação WEKA	Especificidade	Sensibilidade
287.067	170.446	457.513	a	Emergente_Urgente	52,38%	62,75%
25.065	27.571	52.636	b	Pouco_Urgente_Nao_Urgente	62,75%	52,38%
312.132	198.017	**510.149**			**Taxa de Acerto**	**61,68%**

LibSVM						
a	b		←	classificação WEKA	Especificidade	Sensibilidade
279.671	177.842	457.513	a	Emergente_Urgente	52,88%	61,13%
24.803	27.833	52.636	b	Pouco_Urgente_Nao_Urgente	61,13%	52,88%
304.474	205.675	**510.149**			**Taxa de Acerto**	**60,28%**

Na Tabela 23 é efectuada a comparação entre a esta abordagem e a abordagem 1, em que se comparam a aplicação de técnicas de DM para triagem das ocorrências considerando níveis de triagem com duas e quatro classes.

Tabela 23 – Comparação taxa de acerto Abordagens 2 e 3

Técnica	Abordagem 3	Abordagem 1	Variação (p.p.)
J48	**57,87%**	31,81%	26,06%
Naïve Bayes	**59,54%**	34,26%	25,28%
SMO	**61,68%**	36,62%	23,66%
LibSVM	**60,28%**	52,66%	9,02%

Tal como no trabalho de Zmiri & Taieb-Maimon, (2012) verifica-se que considerando apenas duas classes todas as técnicas são mais eficazes. O algoritmo J48 aumentou significativamente a sua taxa de acerto – 26,06% p.p. – em Zmiri, Shahar & Taieb-Maimon, (2012) observou-se um aumento 20,90%.

As restantes técnicas, apesar de incrementos menos expressivos, apresentaram taxas a rondar os 60%, atingindo o algoritmo SMO a taxa de acerto mais elevada até agora

apresentada nesta investigação – 61,68%. O ranking das técnicas nesta abordagem mantém-se inalterado face ao das anteriores, mas é de salientar que as diferenças entre as técnicas são menos significativas.

Em termos de sensibilidade e especificidade, os valores também melhoraram com esta abordagem, mas são uma vez mais inferiores aos obtidos em trabalhos semelhantes. A técnica SMO mostrou maior sensibilidade na prioridade mais importante (Emergente_Urgente) o que significa que entre as ocorrências desta prioridade, foram classificadas correctamente um maior número de instâncias do que com as restantes técnicas. Este facto é espelhado na Tabela 24, onde se verifica mais uma vez e com grande expressão que estamos perante, de forma geral, de erros de sub-triagem. Contudo, verifica-se que com esta abordagem os classificadores SVM, que mantém taxas de acerto superiores, não são os que efectuam maior número de erros de sub-triagem, ao contrário do verificado nas abordagens anteriores.

Tabela 24 – Classificações incorrectas na Abordagem 3

Técnica	Instancias classificadas incorrectamente				
	sub-triagem	%	sobre-triagem	%	total
J48	217.368	92%	17.978	8%	**235.346**
Naïve Bayes	183.870	89%	22.558	11%	**206.428**
SMO	170.446	87%	25.065	13%	**195.511**
LibSVM	177.842	88%	24.803	12%	**202.645**

5.5.4. Abordagem 4

Nesta última abordagem, aos resultados obtidos na Abordagem 1 foi aplicada a matriz de custos apresentada na Figura 2, na qual é atribuído um custo superior aos erros de sub-triagem de forma a efectuarmos uma avaliação aos resultados mais ajustados ao tema em causa. A matriz foi aplicada nas matrizes confusão das quatro técnicas, de forma a ser possível calcular uma nova taxa de acerto, e os resultados são os apresentados na Tabela 25, onde foram incluídos igualmente os resultados da Abordagem 1.

Tabela 25 – Resultado Abordagem 4

Técnica	Abordagem 4	Abordagem 1
J48	**31,66%**	31,81%
Naïve Bayes	**34,00%**	34,26%
SMO	**37,27%**	36,62%
LibSVM	**50,58%**	52,66%

Verifica-se que apenas o algoritmo SMO melhorou o resultado, tendo sido relativamente elevado o decréscimo do algoritmo LibSVM. Na análise efectuada anteriormente às classificações incorrectas efectuadas pelos algoritmos (Tabela 25) observou-se que a na sua maioria dizem respeito a erros de sub-triagem, e este facto conduziu aos resultados apresentados, em que à excepção do algoritmo SMO para todas as técnicas a aplicação da matriz de custos resulta em taxas de acerto inferiores. Na Tabela 22 percebemos o motivo para esta excepção, uma vez que o SMO possui maior sensibilidade na classe Emergente_Urgente igualmente demonstrado na Tabela 24, conduzindo a que nesta abordagem, tendo erros de menor custo, apresente uma taxa de acerto superior.

Finalizada a análise dos resultados obtidos por cada uma das técnicas de DM nas diferentes abordagens efectuadas, análise essa que de acordo com a metodologia de Cios et al. (2000) se inclui na fase anterior – *Data Mining* – faremos agora a apresentação das três iterações que poderiam ocorrer nessa mesma fase. Posteriormente serão observadas as duas iterações relativas a esta fase de avaliação do conhecimento descoberto.

A primeira iteração referida na fase de *Data Mining* consiste em recuar à primeira fase – compreensão do domínio do problema – quando o resultado das técnicas de DM não é o esperado e um melhor conhecimento do domínio poderá ser necessário para alteração dos objectivos (Cios & Kurgan, 2005).

Verificámos que os valores das taxas de acerto obtidas pelas várias técnicas nas diferentes abordagens não são os valores óptimos, tais como valores acima dos 90%. Igualmente concluímos que os erros na classificação consistem essencialmente em erros de sub-triagem, sendo precisamente esse tipo de erros o menos desejado na área em questão. Estes resultados poderiam conduzir a que, caso existisse maior tempo

disponível e conjuntamente com os especialistas, fosse alterado os objectivos da investigação de forma a obter melhores resultados.

A segunda iteração após a fase de *Data Mining* consiste em recuar à fase de compreensão dos dados, que poderá ocorrer quando a falha das técnicas resulta de um fraco conhecimento dos dados, nomeadamente quando as ferramentas escolhidas não se ajustam aos dados (Cios & Kurgan, 2005).

Nesta investigação utilizámos as técnicas Árvore de Decisão, Naïve Bayes e dois algoritmos de SVM que não se mostraram incompatíveis com os dados.

A terceira e última iteração que poderá ocorrer na fase de *Data Mining* consiste em retroceder à fase de preparação dos dados quando se identifica que é necessário um tratamento específico aos dados para a utilização de determinada técnica e que se desconhecia inicialmente (Cios & Kurgan, 2005). Esta necessidade não foi sentida no decorrer desta investigação.

Concluída a observação das iterações da fase de *Data Mining*, analisaremos seguidamente as duas relativas a esta fase, relembrando que nesta fase de avaliação do conhecimento deverá ser efectuar a avaliação por parte dos especialistas se a informação obtida é nova e interessante.

A primeira iteração consiste em recuar à fase de compreensão do domínio do problema e poderá ocorrer quando se chegam a conclusões inválidas ou impossíveis (Cios & Kurgan, 2005). Apesar de não existir uma avaliação dos resultados por parte de especialistas da área, consideramos que não se chega nesta investigação a conclusões impossíveis ou inválidas.

A segunda iteração consiste em retroceder à fase anterior de *Data Mining*, quando mediante a avaliação dos especialistas se conclui que não existe descoberta de novo conhecimento e se acredita que a utilizam de diferentes técnicas poderá conduzir a resultados diferentes (Cios & Kurgan, 2005). Dado que não temos a avaliação dos especialistas da área, consideramos que nesta investigação existe descoberta de novo conhecimento.

5.6. Utilização do Conhecimento Descoberto

Esta fase do processo é exclusivamente reservada aos proprietários dos dados (neste caso ao INEM), que consistiria no planeamento de como e quando seria utilizado o conhecimento descoberto.

Capítulo 6

Conclusões

Este estudo tem como objectivos analisar a existência de factores determinantes na atribuição de um nível de prioridade num pedido de socorro recebido pelo INEM e ainda analisar se será possível, com recurso a técnicas de DM predizer o nível de prioridade de um pedido de socorro. Esta análise considerava-se relevante uma vez que está identificado que o sistema actualmente utilizado como suporte à triagem efectuada no CODU efectua sobre-triagem.

Os dados disponibilizados pelo INEM foram 2.070.227 ocorrências, entre Janeiro de 2014 e Dezembro de 2015 (excepto o mês de Setembro de 2015) com as seguintes variáveis: Ano, Mês, Dia, Dia da Semana, Hora, Distrito, Concelho, Prioridade, Tipo de Ocorrência, Faixa Etária e Sexo.

Através da correlação de Pearson e damétrica *Gain Ratio*, observou-se que as variáveis presentes não eram relevantes na atribuição da prioridade de uma ocorrência – e assim pouco preditoras –, sendo a variável de maior relevância o Tipo de Ocorrência (Ganho de informação: 0,078; Correlação Pearson: 0,039). Nesta análise destacou-se o facto de aparentemente existir correlação entre o Ano e a Prioridade, facto esse avaliado na aplicação das técnicas de DM no Capítulo 5 onde se concluiu que o atributo Ano não tem influência nos resultados.

Com base numa justificação, foi seguida a metodologia de KDD de 6 passos de Cios et al. (2000) e feita referência às possíveis iterações em cada fase do processo. Esta metodologia funcionou como mapa de um caminho a percorrer que consideramos ser uma mais valia na realização deste tipo de investigações.

Para efectuar a triagem de pedidos de assistência foram efectuadas quatro abordagens. Nas primeiras duas abordagens foram considerados os níveis reais de triagem, variando apenas o período dos dados utilizados para treino, desenvolvimento e teste. Importa referir que na segunda abordagem foi avaliada a relação do atributo Ano da triagem de um pedido de assistência.

Na terceira abordagem foram considerados apenas dois níveis de triagem (agrupando os dois de prioridade mais elevada e os dois de prioridade mais baixa). Finalmente, na ultima abordagem foi aplicada uma matriz de custos de decisão.

As taxas de acerto obtidas neste estudo em cada abordagem estão apresentadas na Tabela 26, bem como as de dois trabalhos relacionados que considerámos que efectuaram uma abordagem semelhante – Zmiri, Shahar & Taieb-Maimon, (2012) e Li, Guo & Handly, (2009), ainda que não possam ser directamente comparáveis.

Tabela 26 – Taxas de acerto obtidos nesta investigação e de trabalhos relacionados

Técnica	Abordagem 1 (4 classes)	Abordagem 2 sem Ano (4 classes)	Zmiri, Shahar & Taieb-Maimon, (2012) (4 classes)	Abordagem 3 (2 classes)	Zmiri, Shahar & Taieb-Maimon, (2012) (2 classes)	Li, Guo & Handly, (2009) (2 classes)	Abordagem 4 (matriz)
AD	31,81%	35,17%	49,75%	57,87%	70,65%	76,21%	31,66%
NB	34,26%	36,44%	56,72%	59,54%	72,64%	77,38%	34,00%
SVM (SMO)	36,62%	39,61%	-	60,28%	-	78,21%	37,27%
SVM (LibSVM)	52,66%	49,45%	-	61,68%	-		50,58%

Verifica-se pelas três primeiras abordagens que as máquinas de vectores de suporte são mais eficazes que as restantes técnicas, tal como tínhamos verificado na revisão da literatura.

A representação gráfica das Árvores de Decisão não é interpretável (Figura 7, 8 e 9 dos Anexos), o que conduz a que não possamos dar resposta a um dos requisitos indicados por Bellazzi & Zupan (2008).

Observa-se que entre as duas primeiras abordagens as taxas de acerto não sofrem alterações significativas indicando que a ordenação dos dados, neste caso e com as variáveis em presença, não afecta os resultados da predição.

Para além disso, foi constatado na abordagem 2 que, ao contrário do indicado na análise da correlação dos atributos com a classe, que o atributo Ano não tem influência nos resultados das várias técnicas, significando assim que o Ano não influência a prioridade de um pedido de socorro.

Na abordagem 3 verifica-se um aumento significativo das taxas de acerto comparativamente com a abordagem 1, sendo maior o número de classificações correctas efectuado pelos algoritmos considerando 2 classes do que considerando 4 classes.

Conforme abordado no ponto 5 do Capítulo 5, predominantemente as classificações incorrectas dizem respeito a erros de sub-triagem, facto esse igualmente verificável pela métrica sensibilidade. Uma vez que estamos perante triagem na área da saúde, os erros de sub-triagem são considerados mais gravosos, resultando a que as penalizações consideradas na matriz aplicada na abordagem 4 superiores nesse tipo de erros. O referido anteriormente conduziu a que os resultados nesta ultima abordagem fossem inferiores aos obtidos na abordagem 1.

Na analise da tabela acima, verificamos que os resultados por nós obtidos são inferiores, tanto analisando classificações com 4 classes como nas de 2 classes. No entanto, deverá observar-se que as condições são diferentes, nomeadamente no que diz respeito às variáveis consideradas. Li, Guo & Handly, (2009) consideraram idade, sexo, frequência cardíaca e respiratória e ainda a queixa do paciente, enquanto que Zmiri, Shahar & Taieb-Maimon, (2012) idade, sexo, temperatura, pulso, pressão sanguínea, queixa do paciente, histórico do paciente e nota do enfermeiro. Nestes dois estudos foram consideradas variáveis do âmbito clínico, e que por motivo de confidencialidade não nos foram disponibilizadas.

Bellazzi & Zupan (2008) fizeram notar que a escolha das variáveis a considerar no estudo é um passo importante para a obtenção de bons e fiáveis resultados. Assim, os resultados menos positivos obtidos neste trabalho poderão estar relacionados com as variáveis, factor no qual a nossa influência não é significativa.

Importa relembrar que este tipo de investigações, mediante condições devidamente adequadas, obtém bons resultados, tendo sido observado que as técnicas de *Data Mining* poderão servir de suporte nas instituições de saúde.

6.1. Sugestões para Futura Investigação

Tendo em conta a importância para a população em geral desta temática, será importante desenvolverem-se novas investigações neste âmbito. Sugerimos que, com recurso a técnicas de *Data Mining* ou de outras técnicas de análise de dados, sejam efectuados trabalhos que permitam identificar as chamadas falsas recebidas na Central 112 (objectivo inicial deste trabalho).

Desenvolver novos estudos no âmbito da triagem, com outras variáveis, que possam ser utilizados de suporte na tomada de decisão nos serviços de saúde.

Sugerimos, igualmente que seja efectuada uma análise aos erros identificados da triagem efectuada pelo TETRICOSY® que motiva os accionamentos injustificados da VMER, que mencionámos na introdução.

Finalmente, poderia ainda, uma outra linha de investigação levar em conta o conteúdo da chamada usando técnicas de Processamento Computacional da Língua.

Bibliografia

Abad-Grau, M. M., Ierache, J., Cervino, C., & Sebastiani, P. (2008). Evolution and challenges in the design of computational systems for triage assistance. *Journal of biomedical informatics*, 41(3), 432-441.

Alves, D. D. S. (2016). Saúde em Portugal: estudo das urgências hospitalares através do Data Mining (Dissertação de Mestrado não editada, Mestrado em Gestão de Informação). Universidade Nova de Lisboa, Lisboa.

Azeez, D., Ali, M. A. M., Gan, K. B., & Saiboon, I. (2013). Comparison of adaptive neuro-fuzzy inference system and artificial neutral networks model to categorize patients in the emergency department. *SpringerPlus*, *2*, 416.

Azevedo, A. I. R. L., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. IADS-DM.

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, *77*(2), 81-97.

Bellazzi, R., Ferrazzi, F., & Sacchi, L. (2011). Predictive data mining in clinical medicine: a focus on selected methods and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(5), 416-430.

Bishop, C. M. (1995). Neural networks for pattern recognition. Oxford university press.

Blum, R. L., & Wiederhold, G. C. (1982). Studying hypotheses on a time-oriented clinical database: an overview of the RX project. *Proceedings of the Symposium on Computer Applications in Medical Care*, 245-253.

Bolchini, C., & Cassano, L. (2014, October). Machine learning-based techniques for incremental functional diagnosis: A comparative analysis. In Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT), *2014 IEEE International Symposium on* (pp. 246-251). IEEE.

Canlas, R. D. (2009). *Data mining in healthcare: Current Applications and Issues.* (Dissertação de Mestrado não editada, Master of Science in Information Technology). Carnegie Mellon University, Australia.

Carter, P., Laurie, G. T., & Dixon-Woods, M. (2015). The social licence for research: Why care.data ran into trouble. *Journal of Medical Ethics*, 41, 404-409.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3), 27.

Cios, K. J., & Kurgan, L. A. (2005). Trends in data mining and knowledge discovery. In *Advanced techniques in knowledge discovery and data mining* (pp. 1-26). Springer London.

Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. *Artificial intelligence in medicine*, *26*(1), 1-24.

Cios, K. J., Teresinska, A., Konieczna, S., Potocka, J., & Sharma, S. (2000). A knowledge discovery approach to diagnosing myocardial perfusion. *IEEE Engineering in Medicine and Biology Magazine*, 19(4), 17-25.

Cios, Krzysztof J., and Lukasz A. Kurgan. "Trends in data mining and knowledge discovery." *Advanced techniques in knowledge discovery and data mining*. Springer London, 2005. 1-26.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.

Dixon-Woods, M., & Ashcroft, R. E. (2008). Regulation and the social licence for medical research. *Medicine, Health Care and Philosophy*, *11*(4), 381–391.

Esfandiari, N., Babavalian, M. R., Moghadam, A. M. E., & Tabar, V. K. (2014). Knowledge discovery in medicine: Current issue and future trend. *Expert Systems with Applications*, *41*(9), 4434–4463.

Fan, R. E., Chen, P. H., & Lin, C. J. (2005). Working set selection using second order information for training support vector machines. *Journal of machine learning research*, 6(Dec), 1889-1918.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996b). Knowledge discovery and data mining: towards a unifying framework. In KDD (Vol. 96, pp. 82-88).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From data mining to knowledge discovery in databases. *AI Magazine*, 37–54.

Grzymala-Busse, J. W. (1992). LERS-a system for learning from examples based on rough sets. In *Intelligent decision support* (pp. 3-18). Springer Netherlands.

Grzymala-Busse, J. W. (2005). Rule induction. In *Data Mining and Knowledge Discovery Handbook* (pp. 277-294). Springer US.

Guler, I., & Ubeyli, E. D. (2007). Multiclass support vector machines for EEG-signals classification. *IEEE Transactions on Information Technology in Biomedicine*, 11(2), 117-126.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hand, D. J., Mannila, H., & Smyth, P. (2001). Principles of data mining. MIT press.

Hariz, M., Adnan, M., Husain, W., & Rashid, N. A. (2012). Data mining for medical systems: a review. In *Proceedings of the international conference on advances in computer and information technology* (pp. 17-22).

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 1.

INEM, *Carteira de Serviços, CODU*. Disponível em: http://www.inem.pt/PageGen.aspx?WMCM_PaginaId=27856

INEM, *Institucional, Estatísticas, Indicadores de Desempenho do INEM*. Disponível em: http://www.inem.pt/stats/stats.asp

Izad Shenas, S. A., Raahemi, B., Hossein Tekieh, M., & Kuziemsky, C. (2014). Identifying high-cost patients using data mining techniques and a small set of non-trivial attributes. *Computers in Biology and Medicine*, *53*, 9–18.

Jang, J. S. (1993). ANFIS: adaptive-network-based fuzzy inference system. *IEEE transactions on systems, man, and cybernetics*, 23(3), 665-685.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent data analysis, 6(5), 429-449.

John, G. H., & Langley, P. (1995, August). Estimating continuous distributions in Bayesian classifiers. *In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* (pp. 338-345). Morgan Kaufmann Publishers Inc.

Jothi, N., Rashid, N. A., & Husain, W. (2015). Data Mining in Healthcare - A Review. *Procedia Computer Science*, *72*, 306–313.

Koh, H. C., & Tan, G. (2005). Data mining applications in healthcare. *Journal of Healthcare Information Management*, *19*(2), 64–72.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence*, *14*(12), 1137–1143.

Kurgan, L. A., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review*, *21*(01), 1-24.

Lavrač, N., Flach, P., & Zupan, B. (1999, June). Rule evaluation measures: A unifying view. *In International Conference on Inductive Logic Programming* (pp. 174-185).

Li, J., Guo, L., & Handly, N. (2009, November). Hospital admission prediction using pre-hospital variables. In *Bioinformatics and Biomedicine, 2009. BIBM'09. IEEE International Conference on* (pp. 283-286). IEEE.

Lin, W. T., Wang, S. T., Chiang, T. C., Shi, Y. X., Chen, W. Y., & Chen, H. M. (2010). Abnormal diagnosis of Emergency Department triage explored with data mining

technology: An Emergency Department at a Medical Center in Taiwan taken as an example. *Expert Systems with Applications*, *37*(4), 2733-2741.

Lin, W. T., Wu, Y. C., Zheng, J. S., & Chen, M. Y. (2011). Analysis by data mining in the emergency medicine triage database at a Taiwanese regional hospital. *Expert Systems with Applications*, *38*(9), 11078-11084.

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. *In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

Magidson, J. (1994). The CHAID approach to segmentation modeling: Chi-squared automatic interaction detection. *Advanced methods of marketing research*, 118-159.

Marbán, Ó., Mariscal, G., & Segovia, J. (2009). A data mining & knowledge discovery process model. *In Data Mining and Knowledge Discovery in Real Life Applications.* InTech.

Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*. Springer Science & Business Media.

Patel, V. L., Zhang, J., Yoskowitz, N. A., Green, R., & Sayan, O. R. (2008). Translational cognition for decision support in critical care environments: a review. *Journal of biomedical informatics*, *41*(3), 413-431.

Pawlak, Z. (1982). International Journal of Computer and Information Sciences, 11, 341–356.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*, 240-242.

Peck, J., Gaehde, S., Benneyan, J., Graves, S., & Nightingale, D. (2011). Using Prediction To Improve Patient Flow in a Health Care. *Proceedings of the 2011 Society of Health Systems Conference*, 1–6.

Piatetsky-Shapiro, G. (2000). Knowledge discovery in databases: 10 years after. *ACM SIGKDD Explorations Newsletter*, *1*(2), 59-61.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.

Quinlan, J. R. (2014). C4.5: programs for machine learning. Elsevier.

Quinlan, R. (2004). Data mining tools See5 and C5.0.

Rokach, L., & Maimon, O. (2014). Data mining with decision trees: theory and applications. World scientific.

Sadeghi, S., Barzi, A., Sadeghi, N., & King, B. (2006). A Bayesian model for triage decision support. *International Journal of Medical Informatics*, *75*(5), 403-411.

SAS, 1997, SAS Institute Inc, From Data to Business Advantage: Data Mining, SEMMA Methodology and the SAS System (White Paper).

Schadt, E., & Chilukuri, S. (2015). The role of big data in medicine. *McKinsey & Company*.

Serviço Nacional de Saúde, *Instituto Nacional de Emergência Médica (INEM), I.P*. Disponível em: https://www.sns.gov.pt/entidades-de-saude/instituto-nacional-de-emergencia-medica-ip/

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, 5(4), 13-22.

Shillabeer, A., & Roddick, J. F. (2007, December). Establishing a lineage for medical knowledge discovery. In *Proceedings of the sixth Australasian conference on Data mining and analytics-Volume 70* (pp. 29-37). Australian Computer Society, Inc..

Smith, G. D., & Ebrahim, S. (2002). Data dredging, bias, or confounding. *Bmj*, *325(7378)*, 1437-1438.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427-437.

Tribunal de Contas, Atos do Tribunal, Relatórios de Apuramento de Responsabilidades Financeiras, Relatório nº 12/2016 - 1ª Secção. Disponível em: http://www.tcontas.pt/pt/actos/rel_arf/2016/arf-dgtc-rel012-2016-1s.pdf

van Staa, T. P., Goldacre, B., Buchan, I., & Smeeth, L. (2016). Big health data: the need to earn public trust. *BMJ: British Medical Journal (Online)*, *354*.

VMER SFX, *Notícias, Impacto do TETRICOSY na actividade da VMER*. Disponível em: http://www.vmersfxavier.com/docs/index.php?news=20&idNoticia=1613

Wasan, S. K., Bhatnagar, V., & Kaur, H. (2006). The impact of data mining techniques on medical diagnostics. *Data Science Journal*, 5, 119-126.

Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. *In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29-39).

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Zhang, Y., & Szolovits, P. (2008). Patient-specific learning in real time for adaptive monitoring in critical care. *Journal of biomedical informatics*, *41*(3), 452-460.

Zmiri, D., Shahar, Y., & Taieb-Maimon, M. (2012). Classification of patients by severity grades during triage in the emergency department using data mining methods. *Journal of evaluation in clinical practice*, 18(2), 378-388.

Anexo A – Parametrizações dos Algoritmos

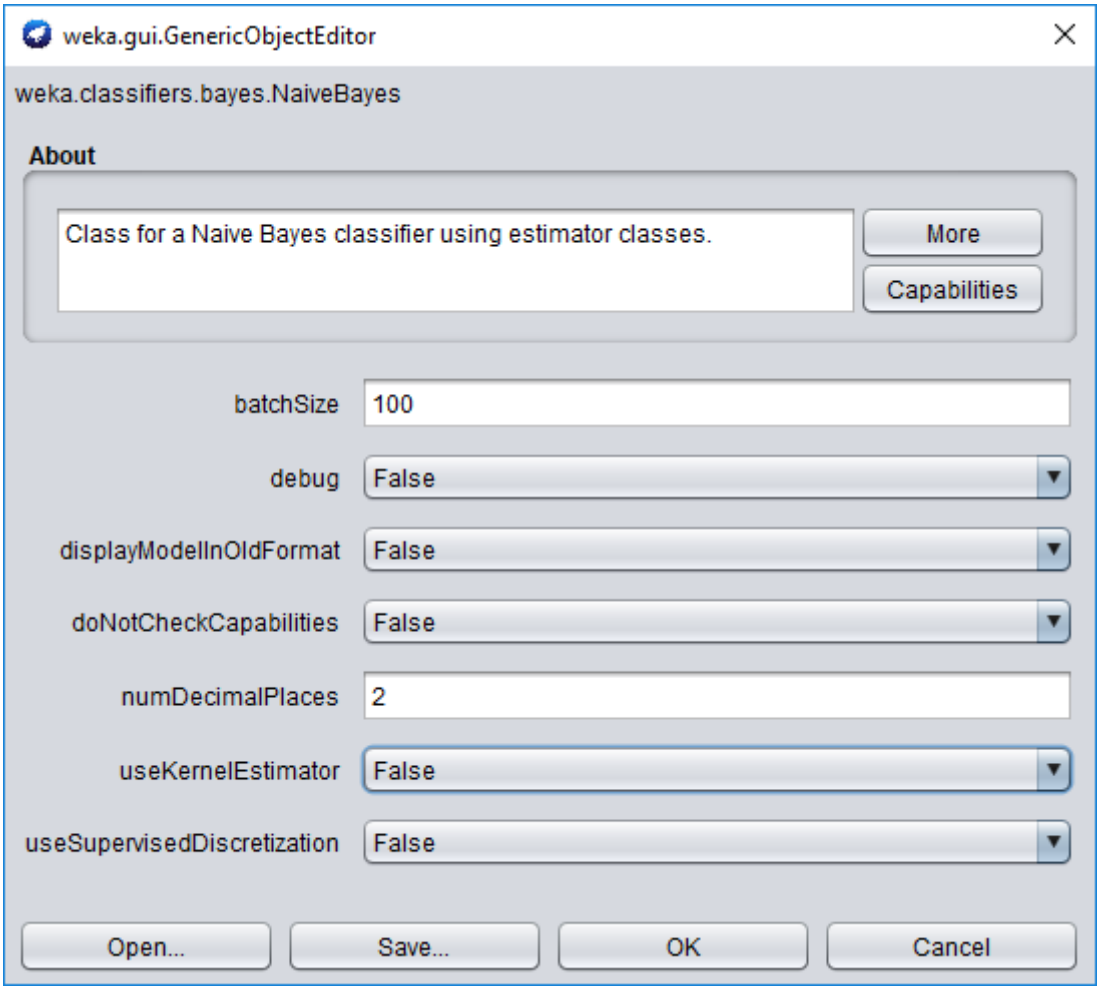

Figura 3 – Parametrizações Naïve Bayes

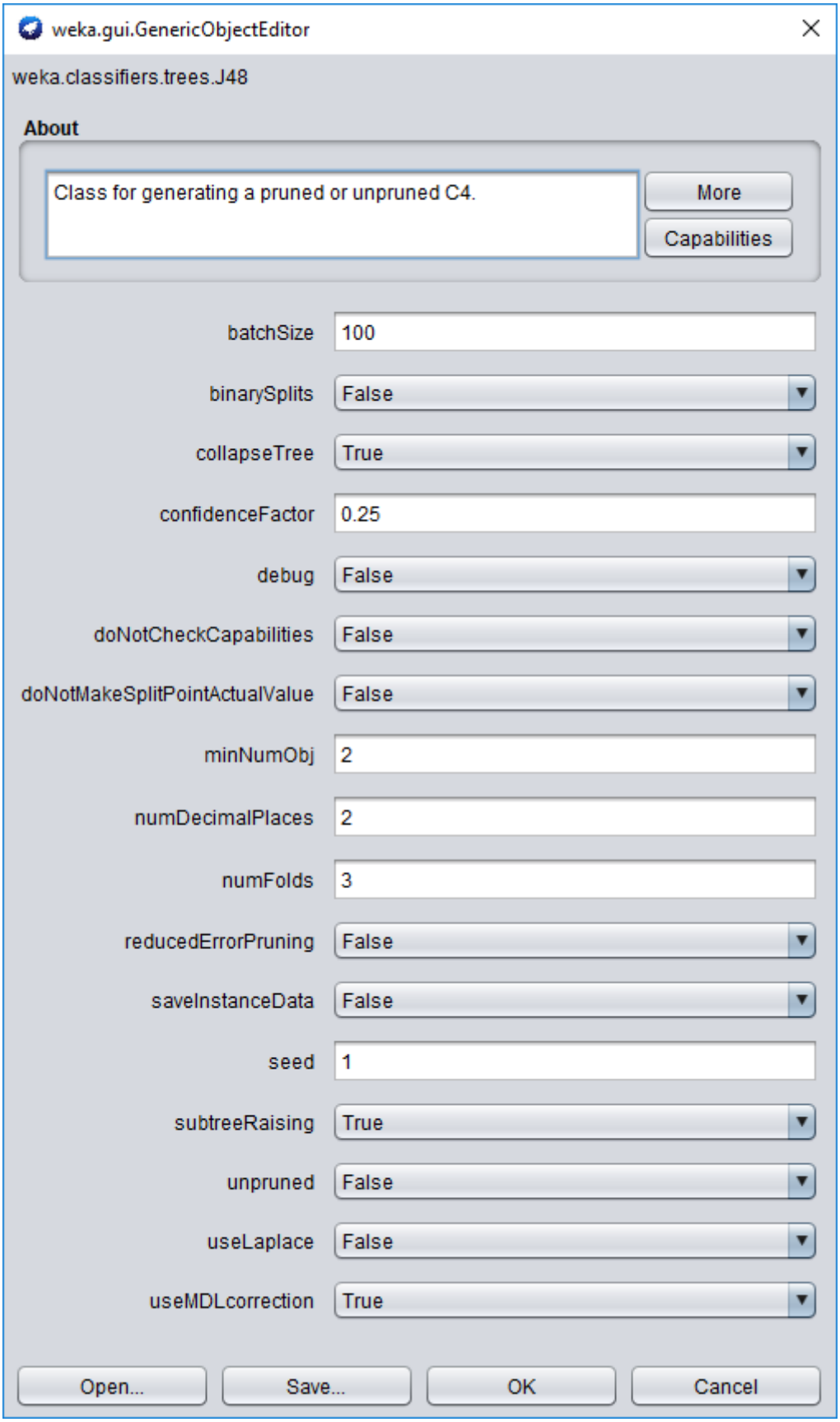

Figura 4 – Parametrizações J48

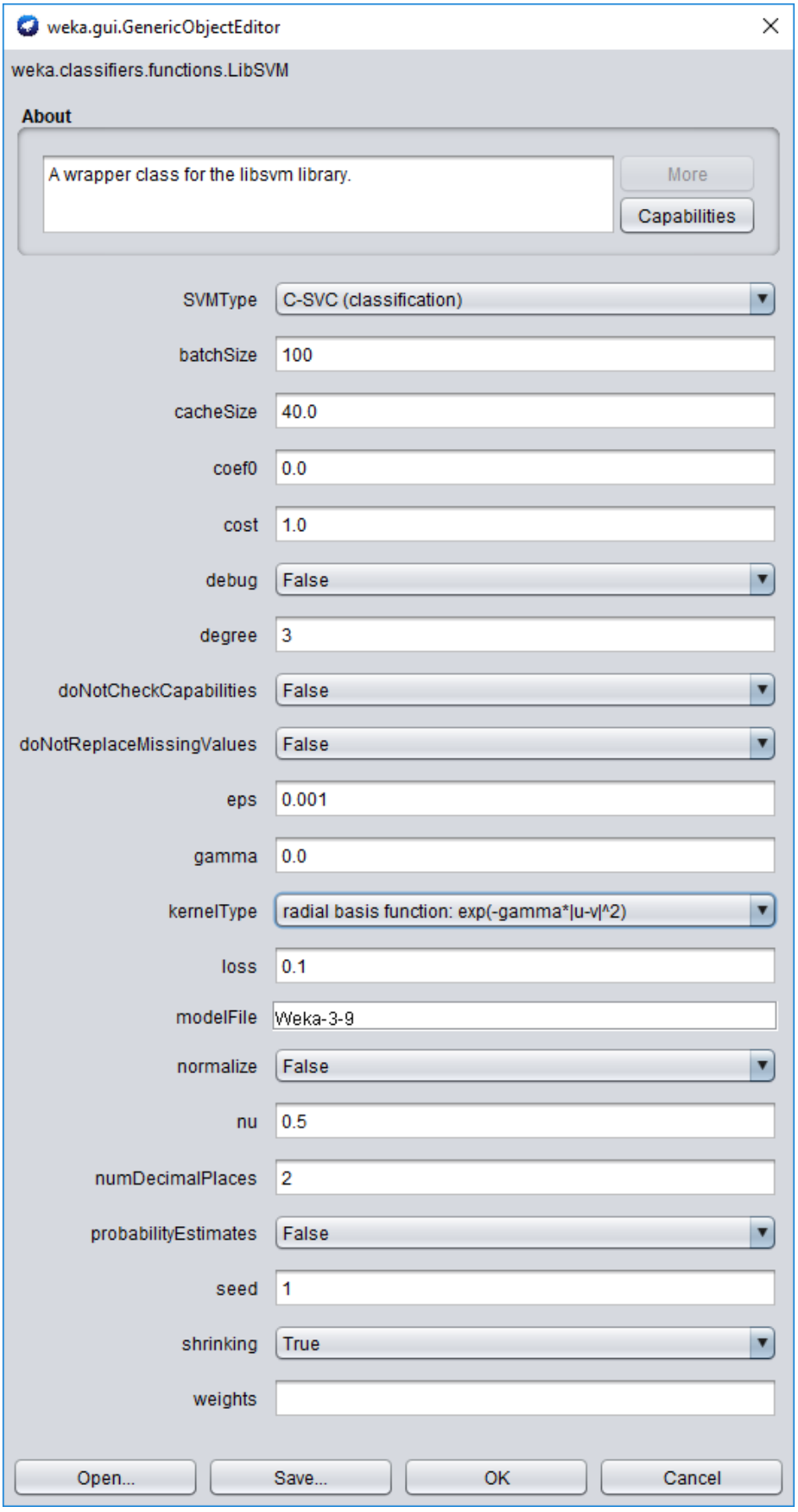

Figura 5 – Parametrizações LibSVM

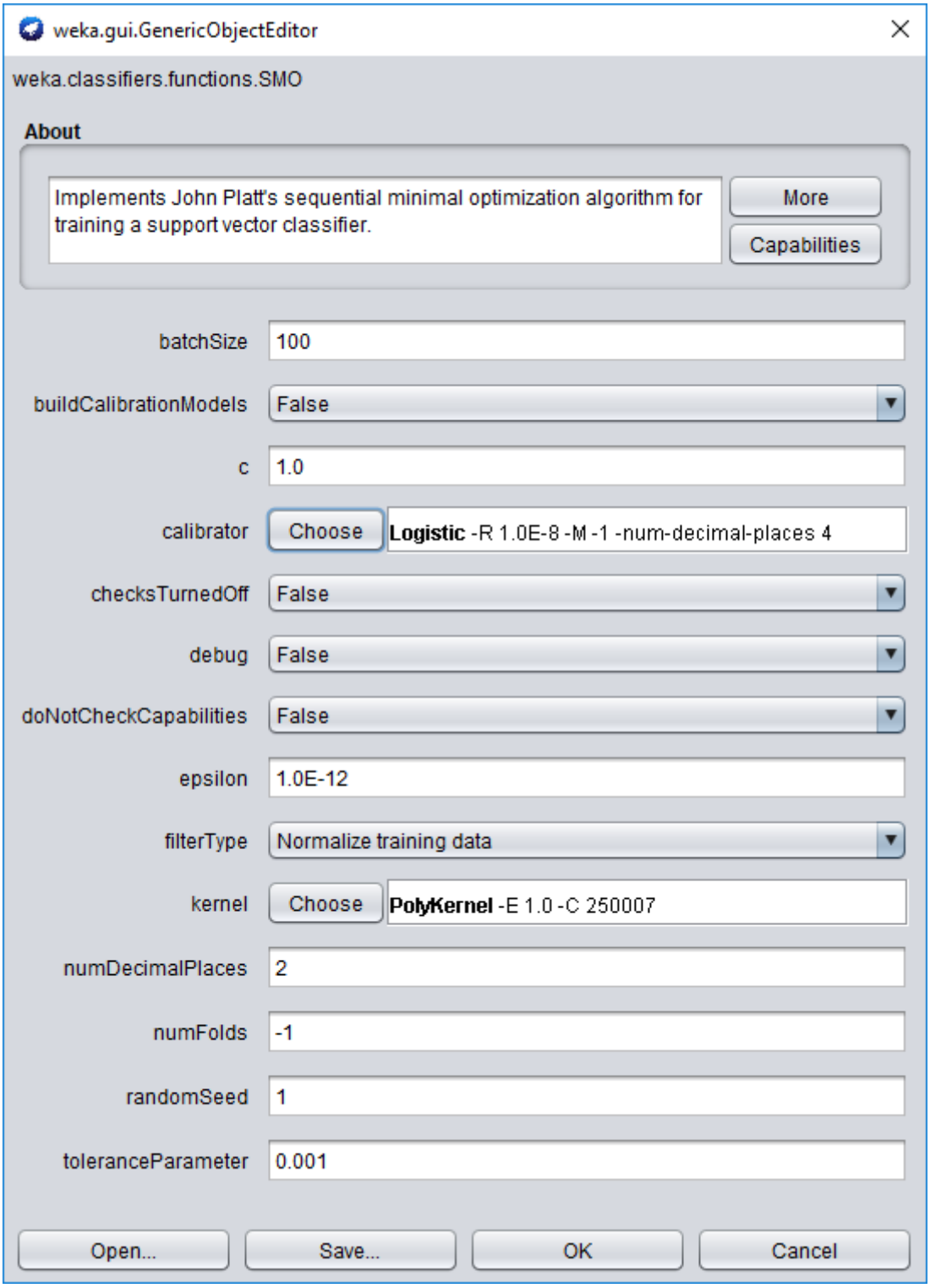

Figura 6 – Parametrizações SMO

Anexo B – Representação das Árvores de Decisão

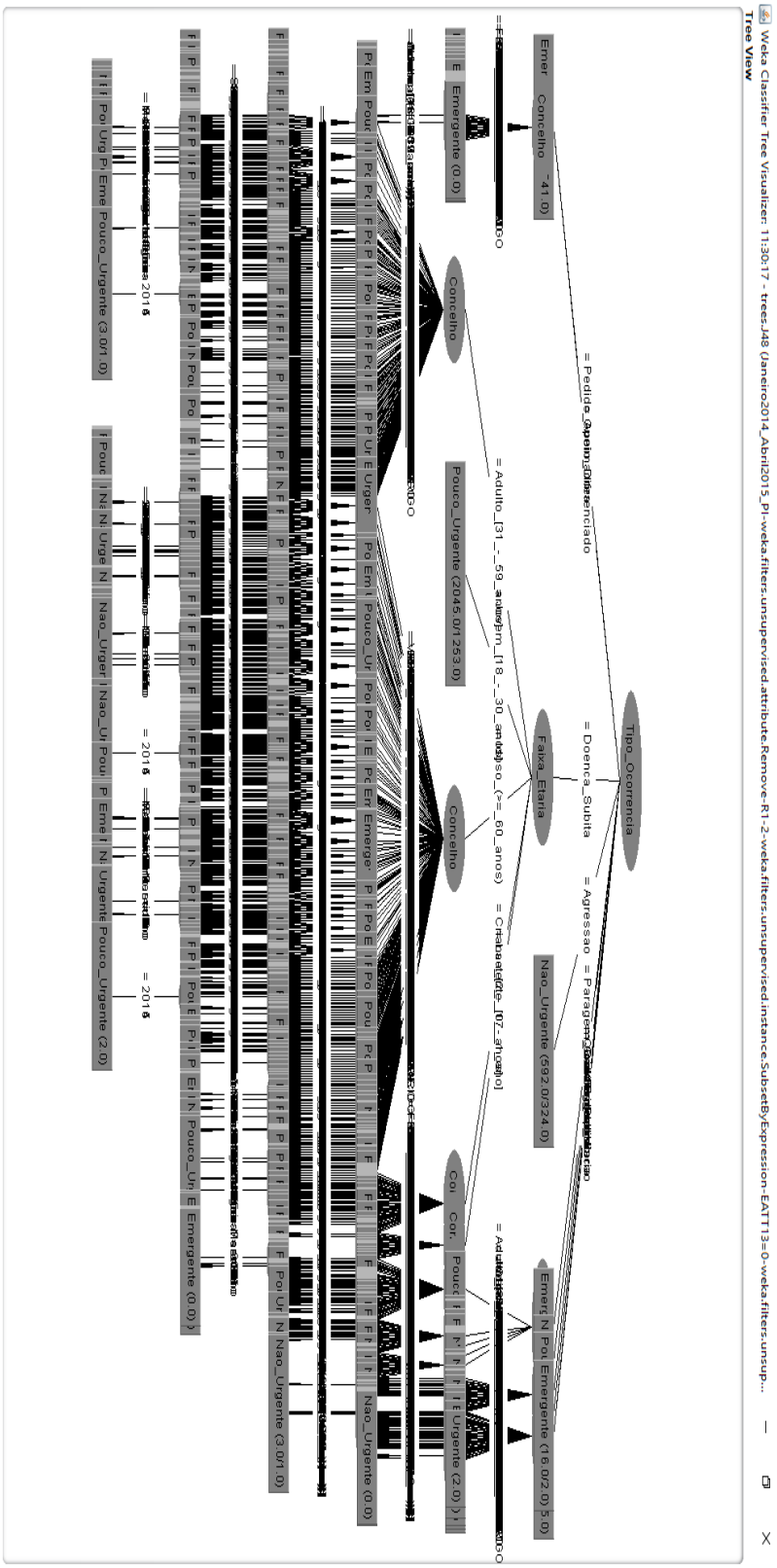

Figura 7 – Representação árvore decisão Abordagem 1 – Teste

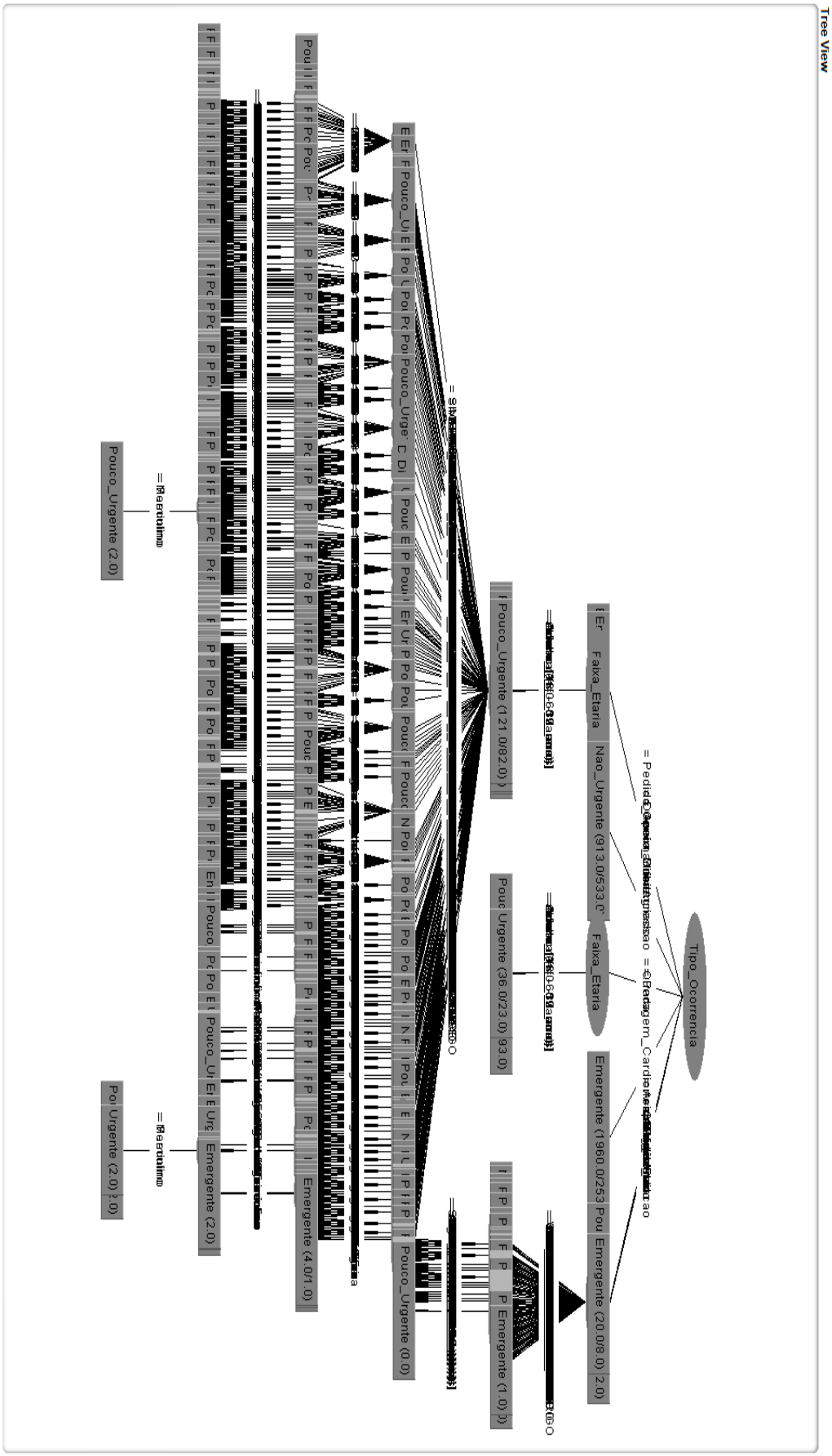

Figura 8 – Representação árvore decisão Abordagem 2 com atributo Ano – Teste

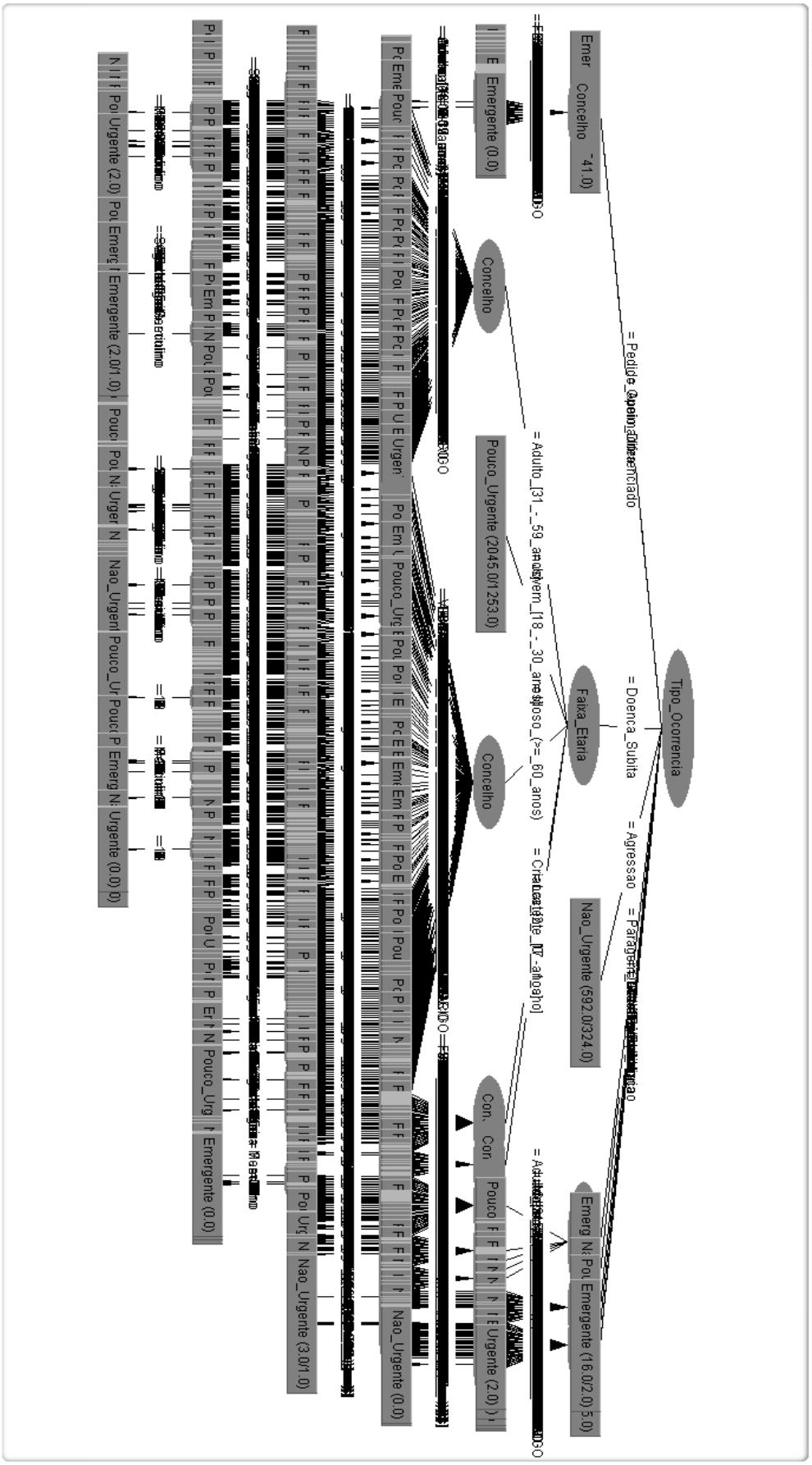

Figura 9 – Representação árvore decisão Abordagem 2 sem atributo Ano – Teste

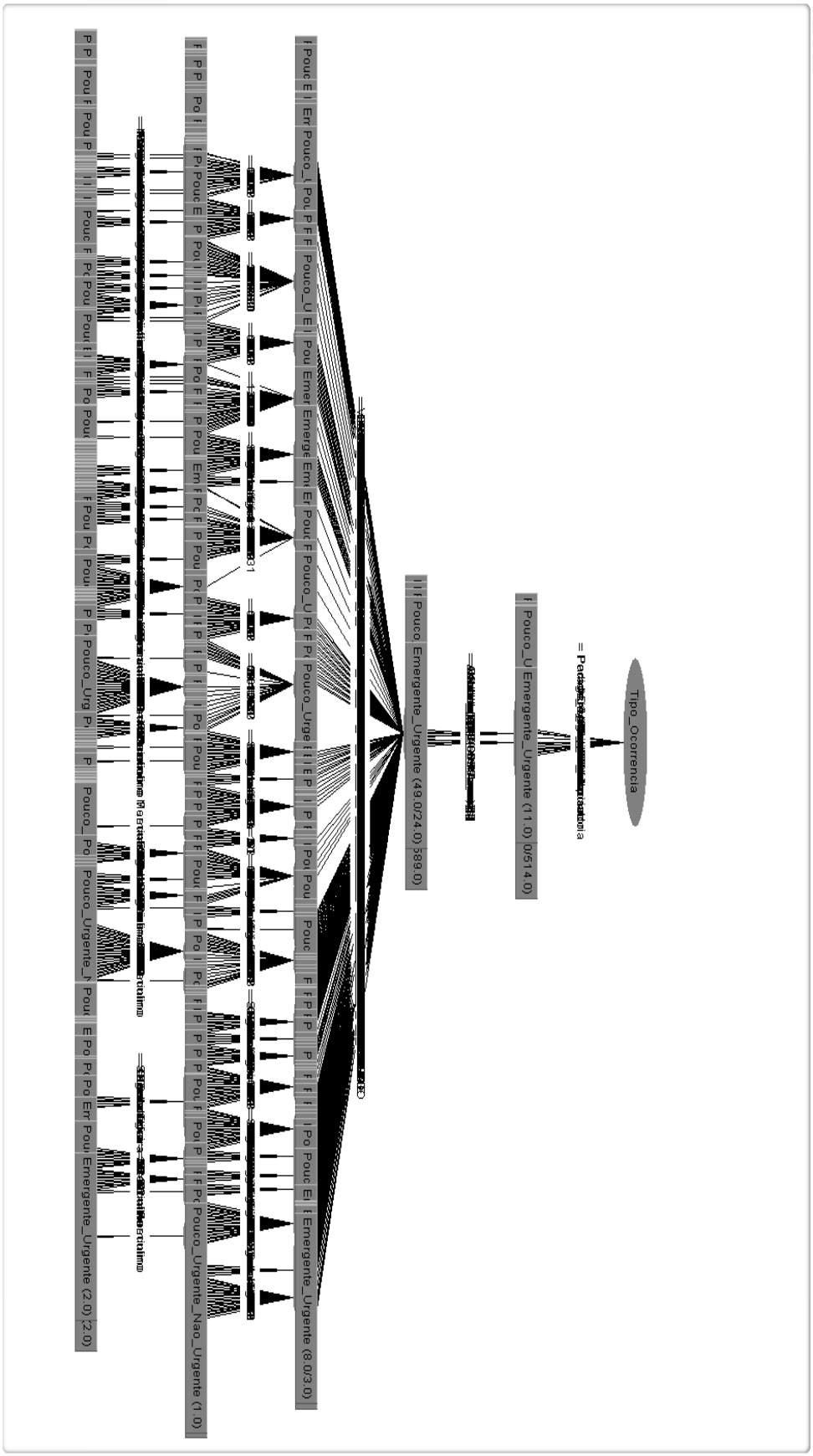

Figura 10 – Representação árvore decisão Abordagem 3 – Teste