Pedro Miguel da Silva Ferreira

Aplicação de Algoritmos de Aprendizagem Automática para a Previsão de Cancro de Mama

Departamento de Ciência de Computadores
Faculdade de Ciências da Universidade do Porto
Porto, Outubro de 2010

Pedro Miguel da Silva Ferreira

Aplicação de Algoritmos de Aprendizagem Automática para a Previsão de Cancro de Mama

Dissertação submetida à Faculdade de Ciências da Universidade do Porto como parte dos requisitos para a obtenção do grau de Mestre em Engenharia de Redes e Sistemas Informáticos

Orientadora: Prof. Doutora Inês Dutra

Co-Orientador: Doutor Nuno Fonseca

Departamento de Ciência de Computadores
Faculdade de Ciências da Universidade do Porto
Porto, Outubro de 2010

Ao meu avô, à minha mãe, restante família e amigos

Este documento foi preparado com o processador de texto MS Word 2007. O sistema de citações de referências bibliográficas utiliza a norma ISO 690 de acordo com a Organização Internacional de Normalização – ISO.

Alguns termos presentes nesta dissertação não foram sujeitos a tradução da língua inglesa para a portuguesa pelo simples facto de estarem amplamente aceites, difundidos e até mesmo enraizados na comunidade académica que estuda o processo de mineração de dados e as técnicas de aprendizagem automática.

Todos os endereços de Internet referenciados na bibliografia foram acedidos pela última vez durante o mês de Outubro de 2010.

Agradecimentos

Gostaria de apresentar os meus agradecimentos, acima de tudo, à Prof. Doutora Inês Dutra e ao Doutor Nuno Fonseca por todo o apoio, disponibilidade e óptimas sugestões ao longo deste trabalho. Sem as suas orientações, o sucesso desta investigação não teria sido possível.

Deixo também uma palavra de agradecimento à Dra. Elizabeth Burnside e ao Dr. Ryan Woods pela assistência prestada na aplicação dos dados fornecidos.

Pretendo igualmente agradecer aos restantes professores e colaboradores da Faculdade de Ciências da Universidade do Porto por me terem proporcionado um ambiente de aprendizagem único ao longo do meu percurso académico.

Uma nota de agradecimento a André Rodrigues, Andress Teixeira, Bruna Pereira, Bruno Luz, Bruno Pinho, Carlos Elói, Carlos Oliveira, Carlos Soares, César Ferreira, Diana Almeida, Diogo Pacheco, Duarte Azevedo, Duarte Silva, Eduardo Burnay, Engerbeth Vivas, Filipe Cunha, Filipe Sousa, Hélder Lourenço, Helena Lagoa, Hugo Figueiredo, Hugo Vieira, Jason Araújo, João Campos, João Melhorado, João Raimundo, José Amador, Juliana Gonçalves, Luís Correia, Miguel Silva, Nuno Vidal, Odair Tavares, Pedro Azevedo, Pedro Borges, Pedro Freitas, Pedro Gomes, Pedro Martins, Pedro Vilaça, Ricardo Castro, Ricardo Luís, Rui Marques, Rui Pedrosa, Tiago Bastos, Tiago Caçador e Vânia Rodrigues pelo apoio proporcionado ao longo do curso, com especial destaque para Alexandra Ferreira, Ana Areal, Bernardo Pina, Bruno Lopes, Bruno Silva, Cristiana Costa, Filipe Azevedo, João Barros, Luís Valente, Margarida Franco, Miguel Barros, Nuno Marques, Pedro Duarte, Ricardo Costa, Sílvia João e Tiago Silva pela paciência demonstrada e conselhos sábios.

Finalmente, agradeço aos meus pais por me terem proporcionado todas as condições para a minha formação académica. Em especial, um muito obrigado à minha mãe, por todo o apoio, tolerância e afecto revelados ao longo dos anos, mas também por ser a força motivadora deste trabalho.

Esta dissertação é dedicada à memória do meu avô e amigo Joaquim Lopes da Silva, falecido no decorrer do presente ano.

O projecto em questão foi parcialmente suportado pelos programas HORUS (PTDC/EIA-EIA/100897/2008) e DigiScope (PTDC/EIA-CCO/100844/2008) e também pela Fundação para a Ciência e a Tecnologia (FCT/Portugal). Foi ainda financiado, através de uma Bolsa de Iniciação Científica (BIC), pelo *Center for Research in Advanced Computing Systems* (CRACS), grupo autónomo do Instituto de Engenharia de Sistemas e Computadores do Porto (INESC Porto LA).

Resumo

O rastreio de cancro de mama consiste na examinação periódica da mama de uma mulher com o principal objectivo de detectar indícios de cancro numa fase inicial. O exame mais utilizado para este fim é a mamografia que, apesar da existência de técnicas mais avançadas, é considerado o método mais económico e eficiente para a detecção de cancro de mama num estado precoce.

Investigamos, recorrendo a técnicas de aprendizagem automática, como os atributos obtidos a partir de mamografias se relacionam com malignidade. Em particular, o foco deste estudo é o modo como a densidade de massa dos nódulos poderá influenciar esse conceito. Para este fim, aplicamos diferentes algoritmos de aprendizagem ao conjunto de dados, fazendo uso das ferramentas do sistema WEKA, assim como efectuamos testes de significância aos resultados. Validamos igualmente estes resultados através da apresentação dos mesmos a especialistas na área médica em questão.

São três as conclusões a que chegamos:

a) A classificação automática de uma mamografia poderá alcançar resultados semelhantes ou mesmo superiores aos obtidos pelos próprios especialistas, o que permitirá aos médicos concentrarem-se mais rapidamente num determinado exame que necessite de um estudo mais aprofundado;

b) A densidade de massa parece ser efectivamente um bom indicador de malignidade, tal como estudos anteriores sugeriam;

c) Conseguimos obter classificadores capazes de preverem densidade de massa dos nódulos com um nível qualitativo tão bom como o de um especialista sem qualquer tipo de informação relativa a biópsias.

Abstract

Breast screening is the regular examination of a woman's breasts to find breast cancer in an initial stage. A widely used exam to this end is mammography that, despite the existence of more advanced technologies, is considered the cheapest and most efficient method to detect cancer in a preclinical stage.

We investigate, using machine learning techniques, how attributes obtained from mammographies can relate to malignancy. In particular, this study focus is on how mass density can influence malignancy from a data set of 348 patients containing, among other information, results of biopsies. To this end, we applied different learning algorithms on the data set using the WEKA tools, and performed significance tests on the results. We also validated our results presenting them to specialists in mammographies.

The conclusions are threefold:

a) Automatic classification of a mammography can reach equal or better results than the ones annotated by specialists, which can help doctors to quickly concentrate on some specific mammogram for a more thorough study;

b) Mass density seems to be a good indicator of malignancy, as previous studies suggested;

c) We can obtain classifiers that can predict mass density with a quality as good as the specialist blind to biopsy.

Índice

Índice de Tabelas

Índice de Figuras

Abreviaturas e Acrónimos

ACR	*American College of Radiology*
arff	*attribute-relation file format*
BIC	Bolsa de Iniciação Científica
BI-RADS®	*Breast Imaging Reporting and Data System*
CDI	Carcinoma ductal invasor
CDIS	Carcinoma ductal *in situ*
CLI	Carcinoma lobular invasor
CLIS	Carcinoma lobular *in situ*
CRACS	*Center for Research in Advanced Computing Systems*
csv	*comma separated values*
DAG	*Directed Acyclic Graph*
DigiScope	*DIGItally enhanced stethosCOPE for clinical usage*
EMA	Erro Médio Absoluto
EUA	Estados Unidos da América
FCT	Fundação para a Ciência e a Tecnologia
FN	*False Negative*

FNR	*False Negative Rate*
FP	*False Positive*
FPR	*False Positive Rate*
GUI	*Graphical User Interface*
HORUS	*Horn Representations of Uncertain Systems*
HTML	*Hyperlink Text Markup Language*
ICC	Instâncias Correctamente Classificadas
IIC	Instâncias Incorrectamente Classificadas
ILP	*Inductive Logic Programming*
INESC	Instituto de Engenharia de Sistemas e Computadores
JDBC	*Java Database Connectivity*
KDD	*Knowledge Discovery in Databases*
KKT	*Karush-Kuhn-Tucker*
LA	Laboratório Associado
MySQL	*My Structured Query Language*
NMD	*National Mammography Database*
PLI	Programação Lógica Indutiva
PR	*Precision-Recall*
ROC	*Receiver Operating Curve*
SMO	*Sequential Minimal Optimization*
SVM	*Support Vector Machine*
TFN	Taxa de Falsos Negativos
TFP	Taxa de Falsos Positivos
TN	*True Negative*

TNR	*True Negative Rate*
TP	*True Positive*
TPR	*True Positive Rate*
TVN	Taxa de Verdadeiros Negativos
TVP	Taxa de Verdadeiros Positivos
UCI	*University of California, Irvine*
USA	*United States of America*
WEKA	*Waikato Environment for Knowledge Analysis*

Capítulo 1

Introdução

O cancro de mama[1], apesar de se tratar de um dos maiores flagelos da sociedade actual, pode ser combatido através da aplicação de programas de rastreio, que têm como principal função detectar indícios de cancro numa fase inicial. O exame mais utilizado para este fim é a mamografia[2] – considerado o método mais económico e eficiente para a detecção de cancro de mama num estado precoce.

Habitualmente, os nódulos[3] encontrados são classificados de acordo com o sistema BI-RADS® (*Breast Imaging Reporting and Data System*) criado pelo *American College of Radiology* (ACR). Este sistema introduziu na área médica um léxico padrão que é utilizado por radiologistas na classificação de nódulos.

[1] Tumor maligno que se desenvolve nas células do tecido mamário. Apresenta-se diversas vezes como uma massa dura e irregular que, quando palpada, se diferencia do resto da mama pela sua consistência. Localiza-se habitualmente no quadrante supero-externo da mama.

[2] Exame radiológico específico para examinação da mama.

[3] Lesões sólidas, elevadas, com mais de 1 cm de diâmetro e geralmente bem delimitadas. Também conhecidas como tumores. De notar que o termo tumor não é sinónimo de cancro. Um tumor pode ser benigno ou maligno.

Vários estudos têm sido desenvolvidos na aplicação de métodos de aprendizagem automática para o estudo do cancro de mama – um dos tipos de cancro mais comuns em todo o mundo. A maioria dos trabalhos presentes na literatura aplica redes neuronais artificiais como forma de diagnosticar este tipo de cancro. Outros trabalhos, por sua vez, focam-se no prognóstico da doença, recorrendo a métodos de aprendizagem indutiva.

O nosso estudo incide, essencialmente, na influência da densidade de massa dos nódulos na previsão de malignidade, no entanto, também abordamos outras questões potencialmente interessantes.

Apesar de alguns estudos no passado terem defendido que densidade de massa seria um indicador pouco fiável de malignidade [JDB+91, CL93, Sic91], investigações recentes [DBD+05, WOS+09, WB10] revelam que a densidade de massa dos nódulos poderá efectivamente ter uma maior importância do que alguns trabalhos anteriores sugeriram.

Nesta dissertação, fazemos uso de um universo de dados fornecido pelos investigadores norte-americanos Woods e Burnside, tendo-lhes aplicado métodos de aprendizagem automática na tentativa de resposta a várias questões. Mesmo aplicando uma metodologia diferente daquela utilizada anteriormente por estes cientistas [WB10], confirmamos que densidade de massa e malignidade estão de facto relacionados. Além do mais, demonstramos que os classificadores gerados neste trabalho são capazes de prever densidade de massa e malignidade com um nível qualitativo semelhante à previsão efectuada por um especialista, assumindo-se como óptimas plataformas de apoio a médicos e radiologistas.

1.1 Motivação

O cancro de mama é o tipo de cancro mais comum entre as mulheres (excluindo o cancro de pele), correspondendo à segunda causa de morte por cancro no sexo feminino.

Trata-se de uma das doenças com maior impacto na sociedade, não só por ser muito frequente, e associada a uma imagem de extrema gravidade, mas também porque

agride um órgão carregado de simbolismo. Apresenta, portanto, repercussões aos mais variados níveis: físico, psicológico, familiar e social.

Actualmente, existem vários tipos de procedimentos aplicados ao tratamento do cancro de mama. No entanto, a melhor forma de prevenir este tipo de cancro é através da realização de programas de rastreio, sendo a mamografia o método mais utilizado, acima de tudo, pelo seu carácter económico e eficiente na detecção de cancro de mama num estado precoce. Apesar da realização de mamografias como forma de prevenção, o recurso a biópsias será sempre uma hipótese a ser equacionada nos momentos em que surjam dúvidas quanto à natureza dos nódulos observados.

Nesta dissertação investigamos essencialmente a influência da densidade de massa dos nódulos na previsão de malignidade. Embora alguns trabalhos no passado tenham defendido que densidade de massa seria um indicador pouco fiável de malignidade, investigações recentes revelam que a densidade de massa dos nódulos poderá efectivamente ter uma maior importância do que alguns estudos anteriores tentaram sugerir.

Como tal, fazendo uso de um universo de dados de 348 pacientes, tentamos provar que a densidade de massa é, de facto, um factor preponderante no diagnóstico de cancro de mama. Além do mais, através da tentativa de construção de classificadores capazes de preverem densidade de massa e malignidade com altos níveis de rigor, poder-se-á evitar no futuro o recurso a biópsias em casos que poderão suscitar dúvidas. Pela simples aplicação destes classificadores a esses mesmos casos, vários milhares de euros poderão ser economizados e posteriormente encaminhados para áreas de pesquisa mais necessitadas.

Sendo assim, a principal causa que nos move ao longo desta investigação, passa, acima de tudo, pela consciência de que poderemos dar um contributo, por mínimo que seja, na descoberta da cura para um dos flagelos mais mortais da nossa sociedade – o cancro de mama.

1.2 Objectivos

Neste documento pretende-se apresentar os resultados de um trabalho de investigação sobre o modo como atributos obtidos a partir de mamografias se relacionam com malignidade. Em particular, o foco deste estudo é a forma como a densidade de massa dos nódulos poderá influenciar a malignidade de um conjunto de dados de 348 pacientes.

A finalidade deste trabalho é, portanto:

i. Encontrar relações entre os atributos através da aplicação de técnicas de aprendizagem automática aos dados;

ii. "Aprender" modelos capazes de auxiliarem os médicos na avaliação imediata de mamografias.

1.3 Estrutura do Documento

Este documento está organizado em seis capítulos:

Capítulo 1 – Este capítulo introduz o tema da dissertação, assim como revela a motivação e objectivos inerentes a este trabalho. É apresentada também a estrutura que segue o documento, além de uma breve nota bibliográfica.

Capítulo 2 – Neste capítulo são abordadas as técnicas e ferramentas a que recorremos na elaboração desta dissertação. É focado o processo de Descoberta de Conhecimento, com especial destaque para o conceito de Mineração de Dados, assim como é efectuada uma contextualização das técnicas de Aprendizagem Automática aplicadas ao problema do cancro de mama. Por último, é introduzido o software utilizado para a execução das experiências.

Capítulo 3 – Neste capítulo é efectuado um levantamento do estado da arte relacionada com o conceito de cancro de mama e respectivos estudos ao longo dos últimos anos.

Capítulo 4 – Este capítulo introduz, inicialmente, os dados fornecidos para a execução das experiências em que são aplicados métodos de aprendizagem automática em tarefas de classificação. Em seguida, é descrita a forma como esses mesmos dados foram seleccionados. Por último, é efectuada uma explicação do modo como foi aplicada a aprendizagem *10-fold cross-validation* ao longo dos diferentes ensaios, assim como a forma de aplicação dos modelos gerados a conjuntos de dados desconhecidos.

Capítulo 5 – Neste capítulo são apresentados os resultados obtidos após a execução das experiências. Posteriormente, é efectuada a análise a esses mesmos resultados através da tentativa de resposta a três questões essenciais:

1. Será densidade de massa um factor relevante no diagnóstico de cancro de mama?

2. Será possível obter classificadores capazes de preverem densidade de massa com um nível qualitativo semelhante ao de um radiologista?

3. Qual o comportamento dos classificadores gerados num conjunto de dados desconhecidos?

Capítulo 6 – Finalmente, este capítulo apresenta as considerações finais, onde é efectuado um balanço sobre todo o trabalho realizado, com especial destaque para os objectivos propostos. O capítulo termina com uma abordagem ao trabalho futuro.

1.4 Nota Bibliográfica

Algumas partes desta dissertação estão presentes no artigo *Studying the relevance of Breast Imaging Features*[4] [FDF+11], o qual foi aceite na conferência: *International Conference on Health Informatics* (*HealthInf*, 2011) que terá lugar na cidade de Roma, em Itália, entre os dias 26 e 29 de Janeiro de 2011.

[4] Ver *Apêndice A*.

Capítulo 2

Background

Neste capítulo iremos abordar as técnicas e ferramentas a que recorremos na elaboração desta dissertação. Deste modo, será focado o processo de Descoberta de Conhecimento, destacando o conceito de Mineração de Dados, assim como será efectuada uma contextualização das técnicas de Aprendizagem Automática aplicadas ao problema do cancro de mama. Apresentaremos também diferentes formas de validação destes métodos de Aprendizagem Automática. Por último, será introduzido o software que utilizamos (WEKA) para a realização das experiências**.**

2.1 Descoberta de Conhecimento

"A quantidade de dados recolhidos e armazenados ao longo do tempo tem crescido de forma considerável em praticamente todas as áreas da sociedade" [Fon06]. Um exemplo disso mesmo é o aumento exponencial de dados relativos à biotecnologia [BKML+05, BWF+00], onde o volume de dados tem vindo a duplicar em cada 3 a 6 meses. Nesta situação particular, tal como em muitas outras, o processamento de todos os dados é uma tarefa extremamente dispendiosa e em alguns casos até impossível, quer humanamente quer computacionalmente.

Estes problemas justificam assim o crescente interesse na descoberta automática de conhecimento em universos de dados extensos.

A Descoberta de Conhecimento em Base de Dados, do inglês *Knowledge Discovery in Databases* (KDD) visa alcançar esse objectivo. Trata-se de um processo de identificação de dados potencialmente úteis e válidos, que por sua vez levará à extracção de padrões que sejam devidamente compreensíveis e representativos do universo em questão [FPSS96]. Neste contexto, os dados são um conjunto de factos, enquanto os padrões dizem respeito a "pedaços" de conhecimento extraídos de um determinado universo e que têm a particularidade de descrever um subconjunto desses mesmos dados. De notar que os padrões podem ser considerados conhecimento: "um padrão que se revele interessante e suficientemente preciso (de acordo com os critérios do utilizador) é designado conhecimento" [FPSS96]. Por outro lado, um modelo pode ser visto como um conjunto de padrões que caracteriza todo o universo de dados.

De seguida, iremos abordar as três fases que constituem o processo de descoberta de conhecimento:

- **Pré-processamento de dados;**
- **Mineração de dados;**
- **Pós-processamento de conhecimento;**

e que se encontram representadas na Figura 1.

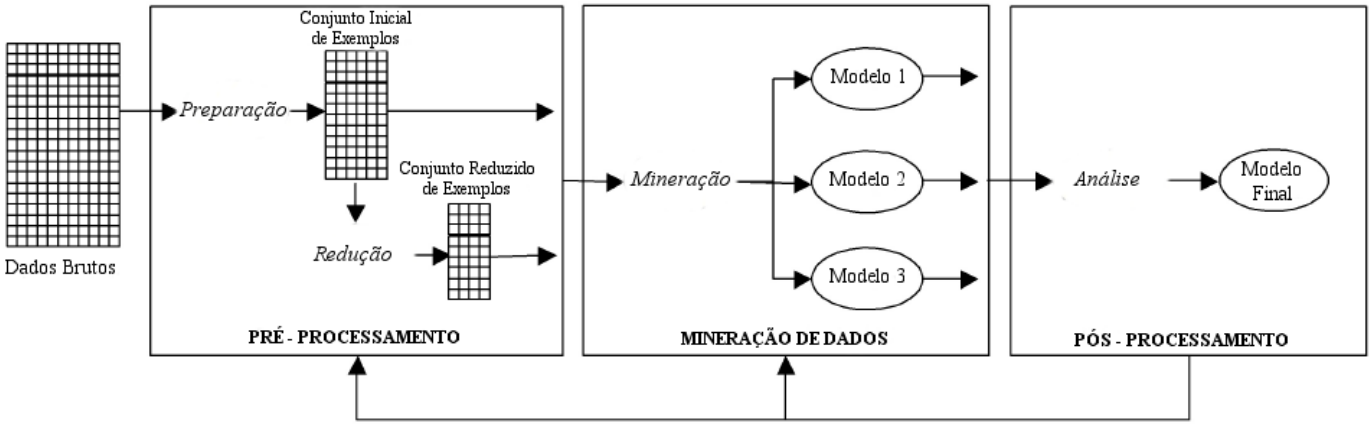


Figura 1 - Fases no processo de Descoberta de Conhecimento (adaptado de [Lee05])

2.1.1 Pré-processamento de dados

A fase de pré-processamento visa essencialmente "conhecer" os dados de forma a prepará-los para a fase seguinte. Ao longo desta etapa, as seguintes tarefas são realizadas [HK06]:

- **Integração dos dados**

A integração dos dados consiste em reunir dados provenientes de múltiplas fontes numa base de dados coerente. Uma integração cuidada a partir de múltiplas fontes poderá ajudar a reduzir e evitar redundâncias/inconsistências no conjunto de dados final, o que, por sua vez, permitirá uma melhoria quer na precisão como na velocidade de todo o processo de mineração.

Existe um vasto número de questões a serem consideradas ao longo da tarefa de integração, como o esquema de integração, redundâncias nos dados que eventualmente possam surgir, assim como a detecção de eventuais conflitos entre dados.

- **Limpeza dos dados**

A limpeza dos dados pode ser aplicada para a remoção de inconsistências e/ou para a correcção de erros nos dados. Pode igualmente efectuar o preenchimento de valores em falta, assim como identificar ou remover dados que não se enquadrem no universo que está a ser alvo de estudo.

Caso os utilizadores não considerem os dados que estão a utilizar como fiáveis, será pouco provável que confiem nos resultados de qualquer processo de mineração aplicado ao universo em questão. Além do mais, dados que não sejam considerados fiáveis poderão causar confusão aquando do processo de mineração, o que poderá conduzir a resultados pouco precisos.

- **Selecção dos dados**

Ao longo da tarefa de selecção dos dados, algumas técnicas (uma vez aplicadas) permitem uma representação reduzida do universo de dados em estudo. Apesar do conjunto de dados poder sofrer uma redução considerável ao nível do volume, a integridade dos dados originais mantém-se, ou seja, a mineração de um grupo de dados reduzido deverá ser mais eficiente e simultaneamente deverá produzir os mesmos resultados analíticos.

Uma das inúmeras estratégias utilizadas na selecção dos dados é a **selecção de atributos**. A selecção de atributos permite reduzir o tamanho do universo de dados em questão através da remoção de atributos redundantes ou irrelevantes. O objectivo deste tipo de selecção é encontrar o menor número de atributos, tal que a probabilidade resultante da distribuição das classes se aproxime o mais possível da distribuição original obtida aquando da utilização de todos os atributos. O simples facto do número de atributos presentes nos padrões ser menor, torna mais fácil a compreensão desses mesmos padrões.

- **Transformação dos dados**

Fase em que os dados são transformados em formatos apropriados para o processo de mineração através de operações de agregação, generalização, normalização ou discretização. Algumas destas operações, tais como a normalização e agregação são procedimentos adicionais de pré-processamento que podem contribuir para o sucesso do processo de mineração.

É importante referir que a maioria dos erros é corrigida ao longo desta etapa de transformação de dados, nomeadamente erros que têm como base erro humano, sendo exemplo disso mesmo os erros originados por um processamento de dados incorrecto. Nos casos em que são encontradas discrepâncias, é necessário definir e aplicar uma série de transformações para as rectificar.

Também nesta etapa, os dados são modificados ou consolidados em formatos devidamente apropriados para o processo de mineração.

A transformação dos dados poderá envolver uma série de operações, tais como:

- ***Smoothing*:** Remoção de ruído dos dados.
- **Agregação:** Aplicação de operações de agregação aos dados que permitem resumir um conjunto de valores num único, através de operações aritméticas (média; máximo; mínimo; soma; entre outros).
- **Generalização:** Generalização dos dados, onde dados primitivos são substituídos por conceitos de nível superior através da aplicação de hierarquias de conceito.
- **Normalização:** Dados são dimensionados de forma a serem inseridos em intervalos de referência relativamente curtos.
- **Construção de Atributos:** Novos atributos são construídos e adicionados a partir do conjunto de atributos dado, com o objectivo de melhorarem o processo de mineração.

A Figura 2 resume os passos do pré-processamento de dados descritos acima.

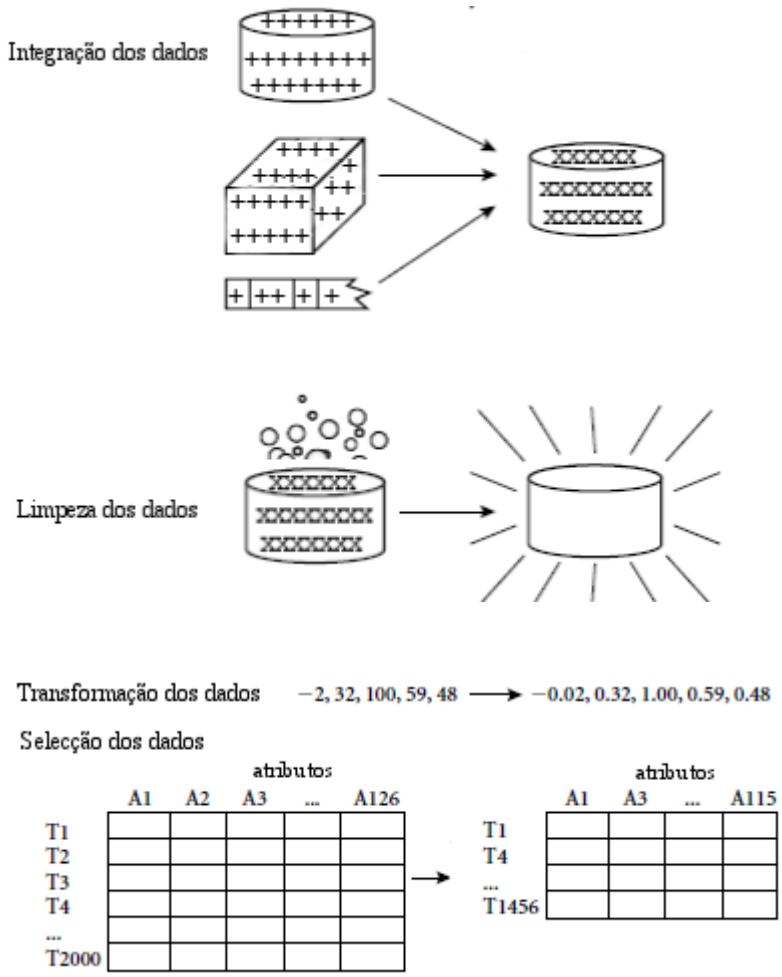


Figura 2 - Etapas no pré-processamento de dados (adaptado de [HK06])

Na fase de pré-processamento, os métodos de visualização de dados, assim como a utilização de estatísticas descritivas (médias, desvios-padrão) [MBK98], assumem um papel fundamental no conhecimento prévio dos dados, podendo mesmo auxiliar na selecção dos algoritmos mais adequados para a fase de mineração [RPMP03, WF00]. Além do mais, o pré-processamento é a actividade que requer um esforço acrescido ao longo de todo o processo de descoberta de conhecimento. Aliás, estima-se que cerca de 80% do tempo despendido em todo o processo seja utilizado para o pré-processamento de dados [Pyl99].

Em suma, os dados reais tendem a ser incompletos, inconsistentes, e em alguns casos, pouco fiáveis. No entanto, as técnicas de pré-processamento garantem a qualidade dos dados envolvidos, desde logo, auxiliando quer na melhoria da precisão como na melhoria da eficiência de processos de mineração decorrentes [HK06]. O pré-processamento de dados é, desta forma, um passo extremamente importante ao longo de todo o processo de descoberta de conhecimento, uma vez que possibilita que decisões de qualidade sejam baseadas em dados, também eles, de qualidade.

2.1.2 Mineração de dados

Em termos históricos, o conceito relativo à procura de padrões consistentes em universos de dados extensos tem sido apelidado de diversas formas, entre as quais: extracção de conhecimento, descoberta de informação, processamento de padrões de dados, entre outros [FPSS96]. Em alguns casos, o próprio termo mineração de dados, do inglês *Data Mining*, confunde-se na literatura como descoberta de conhecimento.

Sendo assim, torna-se essencial referir que sob o nosso ponto de vista e de acordo com alguns autores [FPSS96], o processo de descoberta de conhecimento procura extrair informação relevante a partir de um determinado conjunto de dados.

A mineração, por sua vez, refere-se a uma etapa de todo este processo, e muito provavelmente a mais importante. Trata-se da aplicação de algoritmos específicos na extracção de padrões dos dados [FPSS96].

Os dois principais objectivos do processo de mineração de dados são, na prática, a previsão e a descrição. Quer os modelos de previsão como os de descrição são construídos a partir de observações. Não existe uma separação total entre estas duas categorias de modelos, podendo um modelo de previsão servir também como descrição e vice-versa. A função do modelo, descritivo ou de previsão, vai depender da forma de representação do próprio modelo e do seu foco.

De seguida, apresentamos as principais diferenças entre previsão e descrição no contexto de mineração de dados:

- No que diz respeito à construção de **modelos de previsão**, o objectivo principal é prever o valor de alguma variável num determinado universo de dados, sendo que essa previsão é baseada no modelo construído a partir de valores de outras variáveis já previamente conhecidas. Caso o valor da variável que está a ser alvo de previsão (classe) assuma um valor numérico (contínuo), trata-se de um problema de regressão. Se a variável for categórica então estamos perante um problema de classificação, em que cada categoria é designada como *valor de classe* [Fon06]. Os modelos de previsão, tal como acima mencionado, podem oferecer uma descrição dos dados, no entanto existem muitos que não são de fácil interpretação (redes neuronais ou *support vector machines*[5], por exemplo) e, portanto, são considerados apenas como de previsão. Os modelos de previsão que utilizam representação em árvore ou em forma de regras, também podem ser considerados modelos descritivos.

- No **modelo descritivo**, o objectivo fundamental é, tal como o próprio nome indica, descrever padrões interessantes relativos ao universo de dados em causa. *Clustering*, por exemplo, é uma das tarefas do modelo descritivo e consiste em agrupar todos os dados semelhantes entre si em subconjuntos [Fon06]. Este tipo de modelo pode igualmente ser utilizado para previsão nos casos em que apresentamos uma nova instância e este decide a qual grupo esta nova instância irá pertencer. Modelos baseados em regras também são considerados modelos descritivos.

É importante sublinhar que a diferença principal entre uma tarefa de previsão e uma tarefa de descrição está directamente relacionada com a existência ou não de classes pré-definidas para os dados. Os modelos descritivos estão habitualmente associados à modelação de relações entre dados que não são previamente rotulados (aprendizagem não supervisionada) enquanto os modelos de previsão estão geralmente relacionados de forma

[5] Ver subsecção 2.2.4.

directa à modelação de dados que pertencem a uma determinada classe previamente conhecida.

Torna-se relevante distinguir aprendizagem supervisionada de aprendizagem não supervisionada. Deste modo, na **aprendizagem supervisionada**, cada exemplo é associado a uma classe (rótulo), que, tal como acima referido, poderá ser discreta, sendo neste caso designada por classificação, ou no entanto poderá ser contínua, denominada de regressão [Lee05]. Na **aprendizagem não supervisionada**, por sua vez, não existe informação sobre a classe associada a cada exemplo [Lee05]. A aprendizagem é efectuada descobrindo similaridades nos dados, ou seja, pretende-se encontrar agrupamentos de dados com características semelhantes [Cru07]. A tarefa de *clustering* é um tipo de aprendizagem não supervisionada.

A importância quer da previsão como da descrição para determinadas aplicações da mineração de dados poderá variar consideravelmente dependendo da natureza dos dados e dos objectivos do utilizador.

São várias as tarefas de mineração de dados que poderão ser aplicadas quer para previsão como para descrição. De seguida passamos a descrever algumas delas:

- **Classificação**

  A classificação consiste no processo de encontrar um modelo (ou função) que descreva e distinga classes de dados ou conceitos. Depois de encontrado esse modelo, é possível aplicá-lo de forma a prever a classe de um novo objecto. O modelo gerado é baseado na análise de um conjunto de dados, designado por conjunto de treino (objectos cuja classe é previamente conhecida) [HK06].

  Para a execução da tarefa de classificação é possível aplicar uma série de métodos de aprendizagem automática (Figura 3), nomeadamente: **árvores de decisão**, **regras de classificação (regras *if-then*)**, **programação lógica indutiva**, ***support***

vector machines, **redes bayesianas**, ***ensemble***, entre outros[6]. Na Figura 3, por exemplo, são representados diferentes modelos de classificação para um mesmo problema. Neste caso particular é relacionada a idade de um indivíduo X e o seu rendimento, inserindo-o numa determinada categoria.

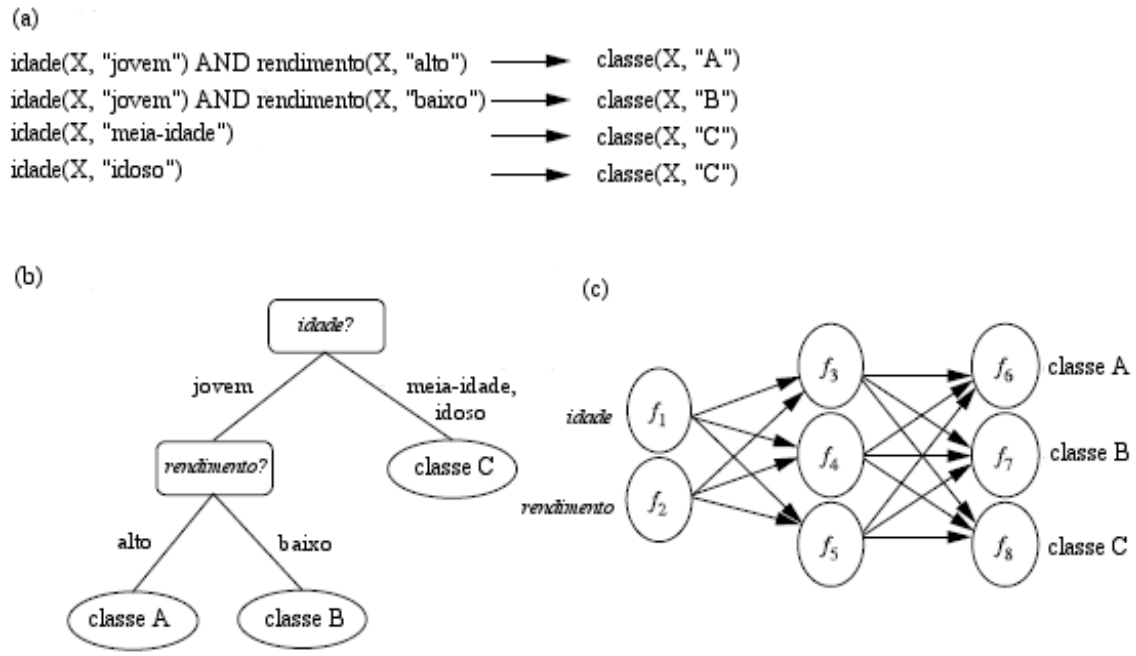


Figura 3 - Um modelo de classificação pode ser representado de várias formas, tais como: (a) regras de classificação, (b) árvores de decisão, ou (c) redes neuronais (adaptado de [HK06])

- **Regressão**

A regressão é habitualmente utilizada para a previsão de valores de variáveis dependentes (variáveis que se pretende prever) a partir de uma ou mais variáveis independentes (atributos conhecidos) e nos casos em que essas mesmas variáveis são contínuas. Trata-se de uma tarefa utilizada na aproximação dos dados recebidos.

[6] Os métodos de aprendizagem automática serão descritos em detalhe na secção 2.2.

Existem inúmeras formas de regressão, tais como **linear, linear múltipla, polinomial, robusta,** entre outras. Dois dos tipos de regressão mais populares são a **regressão linear** e a **regressão linear múltipla**. A **regressão linear** visa encontrar a melhor forma de relacionar dois atributos (ou variáveis), de modo a que um dos atributos possa ser utilizado na previsão do outro.

Tomemos como exemplo a situação ilustrada na Figura 4:

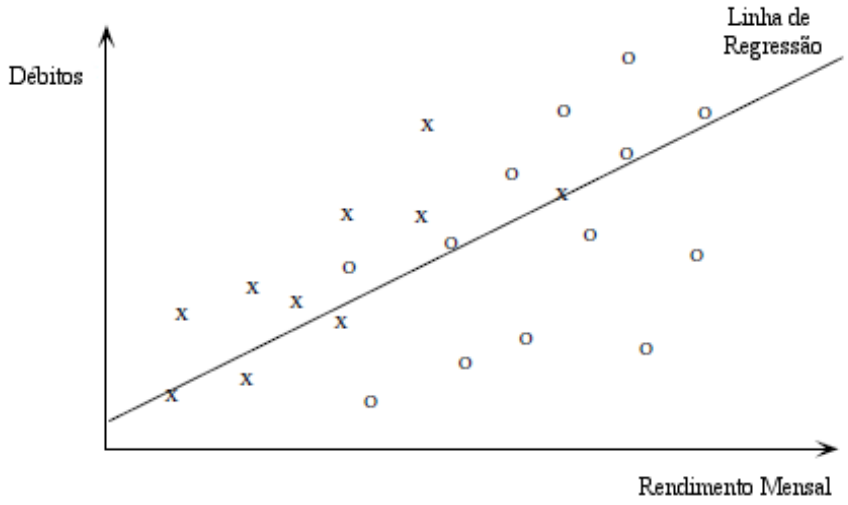


Figura 4 - Exemplo de regressão linear entre total de débitos de um conjunto de indivíduos e o valor dos seus rendimentos (adaptado de [FPSS96])

A Figura 4 representa um conjunto de dados bidimensional composto por 23 casos. Cada ponto no gráfico refere-se a uma pessoa a quem lhe foi facultado um empréstimo por um determinado banco. O eixo horizontal representa os rendimentos dessas pessoas, enquanto o eixo vertical refere-se ao total de débitos desses mesmos indivíduos (hipoteca, prestação do automóvel, etc.). Os dados foram divididos em duas classes distintas: os *x's* representam pessoas que não têm cumprido com os seus pagamentos ao banco; os *o's*, por sua vez, referem-se a pessoas que têm pago as suas prestações de acordo com os prazos estipulados pelo banco em questão.

A Figura 4 ilustra, portanto, o resultado de uma simples regressão linear onde os débitos das pessoas são apresentados como uma função linear dos seus rendimentos. Nesta situação particular o ajuste é baixo, uma vez que existe apenas uma correlação extremamente fraca entre as duas variáveis. Deste modo, é possível referir que uma variável aleatória *y* (variável dependente), pode ser modelada como uma função linear de uma outra variável aleatória *x* (variável independente), de acordo com a equação:

$$y = ax + b$$

em que a variância de *y* é assumida como sendo constante. No contexto da mineração de dados, *x* e *y* são atributos numéricos enquanto *a* e *b* são coeficientes de regressão.

A **regressão linear múltipla**, por sua vez, é uma extensão da regressão linear, onde mais do que dois atributos estão relacionados e os dados são adaptados a uma plataforma multidimensional. Permite que uma variável dependente *y* seja modelada como uma função linear de duas ou mais variáveis independentes [HK06].

- **Regras de Associação**

As regras de associação são um tipo de regras geradas a partir de padrões frequentes. Este tipo de mineração poderá gerar um vasto número de regras, no entanto, muitas dessas mesmas regras acabam por se revelar redundantes ou até pouco esclarecedoras quanto à existência de correlações entre atributos. Sendo assim, as regras geradas poderão ser alvo de um processo de análise de modo a encontrar correlações estatísticas, podendo mesmo conduzir *a posteriori* a regras de correlação.

O principal objectivo desta tarefa de mineração é encontrar associações interessantes ou relações de correlação dentro de um universo de dados extenso. A descoberta de relações de associação dentro de universos de dados extensos assume-se cada vez mais como um factor extremamente importante em diversas áreas de negócio.

Um exemplo típico da aplicação de regras de associação é a análise dos hábitos de compra dos consumidores. Este processo caracteriza-se por efectuar um estudo dos hábitos de consumo através da descoberta de associações entre diferentes itens adquiridos. Este tipo de conhecimento permite aos vendedores desenvolver uma série de estratégias de marketing. Por exemplo, se os consumidores compram leite, qual a probabilidade de comprarem também pão numa mesma visita ao hipermercado? Esta informação poderá conduzir a um aumento substancial no volume de vendas, uma vez que permitirá aos operadores logísticos efectuarem uma selecção/planeamento dos seus produtos. No exemplo em questão, a simples colocação do pão próximo do leite poderá levar a um aumento significativo nas vendas em conjunto destes dois consumíveis [HK06].

Do universo de algoritmos existentes, o algoritmo *Apriori* [AS94] é o mais referenciado na descoberta de regras de associação. Trata-se de um algoritmo utilizado para encontrar associações relevantes entre atributos. Além do mais, com o *Apriori* são definidos uma série de parâmetros que determinam quais associações são ou não interessantes para o utilizador.

Este algoritmo gera regras do tipo $X \rightarrow Y$, onde $X = \{x1, x2,\ldots, x_n\}$ e $Y = \{y1, y2,\ldots, y_m\}$ são conjuntos de itens [CS04]. Esta associação afirma que se a característica X está presente, à partida a característica Y também estará. Por exemplo, uma imagem de raios-X contendo características *a* e *b* provavelmente irá exibir a característica *c*. O algoritmo Apriori caracteriza-se, portanto, por efectuar uma série de buscas sucessivas num determinado universo de dados, mantendo um óptimo desempenho no que respeita ao tempo de processamento [AS94].

- ***Clustering***

A tarefa de *clustering*, em português agrupamento, visa identificar um conjunto finito de categorias ou grupos que descrevam um conjunto de dados [JD88, TSM85]. É um método de aprendizagem não supervisionada e uma técnica comum na análise estatística de dados utilizada em inúmeras áreas, desde aprendizagem automática[7] até mineração de dados, reconhecimento de padrões, análise de imagens, bioinformática, entre outras.

Voltando ao exemplo dos empréstimos bancários (introduzido para explicar a tarefa de regressão), a Figura 5 ilustra um possível agrupamento em que o conjunto de dados é dividido em três grupos.

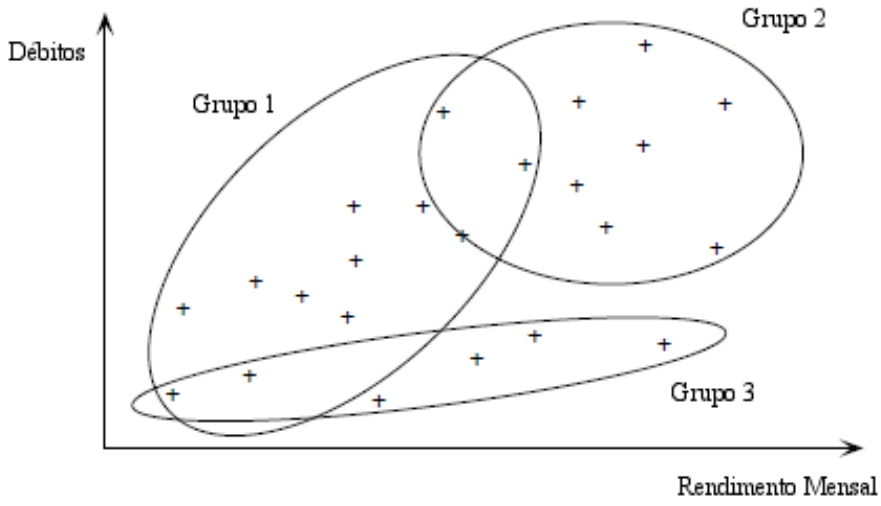


Figura 5 - Tarefa de *clustering* em que um conjunto de dados é dividido em três grupos (adaptado de [FPSS96])

[7] Ver secção 2.2.

É notória a sobreposição de grupos, permitindo que alguns pontos do conjunto de dados pertençam a mais do que um aglomerado. Relevante também é o facto de não serem conhecidas as classes a que cada ponto pertence [FPSS96]. Aliás, ao contrário da tarefa de classificação, a tarefa de *clustering* analisa objectos sem recorrer a uma classe específica. Na generalidade dos casos, as classes não estão presentes nos dados de treino porque não são conhecidas à partida. Este tipo de tarefa poderá ser utilizado para esse propósito, ou seja, para gerar classes de objectos [HK06].

Normalmente, os algoritmos utilizados neste tipo de tarefa são aqueles que utilizam alguma medida de distância entre pontos. O objectivo desses algoritmos é maximizar a distância entre grupos e simultaneamente minimizar a distância entre indivíduos do mesmo grupo.

De notar uma vez mais que a mineração de dados é apenas um passo em todo o processo de descoberta de conhecimento, no entanto põe a descoberto uma série de padrões para avaliação até então desconhecidos [HK06].

2.1.3 Pós-processamento de conhecimento

O objectivo principal da fase de pós-processamento é avaliar, validar e consolidar o conhecimento extraído [Lee05]. Interpretando os resultados recorrendo, por exemplo, à visualização dos padrões obtidos ou à tradução de padrões considerados úteis para formas que sejam de fácil compreensão, são modos de efectuar a avaliação do conhecimento a que se chegou. Devem igualmente ser avaliados de forma a garantir que os resultados são fiáveis e estatisticamente significativos[8].

[8] A significância estatística trata-se de uma ferramenta matemática utilizada para determinar se o resultado de uma experiência se deve a uma relação entre factores específicos ou se resulta apenas de um simples acaso.

A validação é também uma etapa a ter em conta. Efectuando-se uma comparação entre o conhecimento adquirido e o conhecimento prévio, eventuais conflitos serão eliminados.

A consolidação do conhecimento extraído, por sua vez, é executada a partir do momento em que esse mesmo conhecimento é associado a sistemas de apoio à decisão, ou então nas situações em que é disponibilizado ao utilizador através de documentação própria.

Em suma, as três fases (pré-processamento de dados, mineração de dados e pós-processamento do conhecimento) são fundamentais para que o processo de descoberta de conhecimento[9] seja bem sucedido.

Na secção seguinte iremos abordar em detalhe, a aprendizagem automática e métodos associados – parte integrante das tarefas de mineração de dados.

2.2 Métodos de Aprendizagem Automática

"Desde a invenção dos computadores que o homem se tem questionado se estes foram concebidos para a aprendizagem. Se fosse possível compreender como programá-los para "aprenderem" (i.e. para melhorarem de forma automática com a experiência) o impacto seria enorme. Imaginemos, por exemplo, na área da saúde, os computadores a "aprenderem" a partir de registos médicos quais os tratamentos mais eficazes para novas doenças. Uma compreensão bem sucedida do modo como tornar os computadores capazes de "aprender" permitiria uma abertura muito maior no que respeita a novas formas de utilização destas máquinas, assim como conduziria a novos níveis de competência e personalização. Além do mais, uma compreensão detalhada dos algoritmos de processamento de informação aquando da utilização de métodos de aprendizagem automática poderia contribuir para um melhor entendimento quer das capacidades como das limitações da aprendizagem humana." [Mit99].

[9] Uma descrição mais detalhada do processo de descoberta de conhecimento poderá ser encontrada em [BA96].

A aprendizagem automática visa compreender o modo como criar programas que permitam melhorar o desempenho das máquinas em determinadas tarefas, nomeadamente através do conceito "experiência". Sendo assim, os algoritmos de aprendizagem automática têm-se revelado extremamente úteis em diversos domínios, desde logo têm sido especialmente importantes na resolução de problemas de mineração de dados, onde universos de dados extensos poderão conter implicitamente informação considerada de valor e que poderá ser descoberta automaticamente.

Estes algoritmos são igualmente essenciais em domínios de compreensão difícil, em que os próprios humanos não possuem capacidades para o desenvolvimento de algoritmos eficazes.

Os diferentes métodos de aprendizagem automática utilizados nesta dissertação serão brevemente introduzidos em seguida.

2.2.1 Árvores de Decisão

As árvores de decisão caracterizam-se por utilizarem a estratégia de divisão e conquista. Sendo assim, focam-se num problema considerado complexo, dividindo-o em problemas mais simples e recursivamente aplicando a mesma estratégia a sub-problemas. No final, as soluções dos sub-problemas podem ser combinadas para gerar a solução do problema inicial [Gam99].

As árvores de decisão classificam instâncias ordenando-as desde a raíz até um determinado nó-folha, o qual designa a classificação da instância em causa. Cada nó na árvore especifica um determinado atributo da instância, enquanto cada ramo descendente corresponde a um dos possíveis valores para o atributo em questão. Uma instância é classificada começando pela raíz da árvore, testando o atributo definido pelo nó e posteriormente descendo o ramo correspondente ao valor do atributo dado. Todo este processo é depois repetido para a sub-árvore cuja raíz é um novo nó.

A Figura 6 ilustra uma árvore de decisão típica. Neste caso particular, as manhãs de Sábado são classificadas consoante se são adequadas ou não para a prática de ténis.

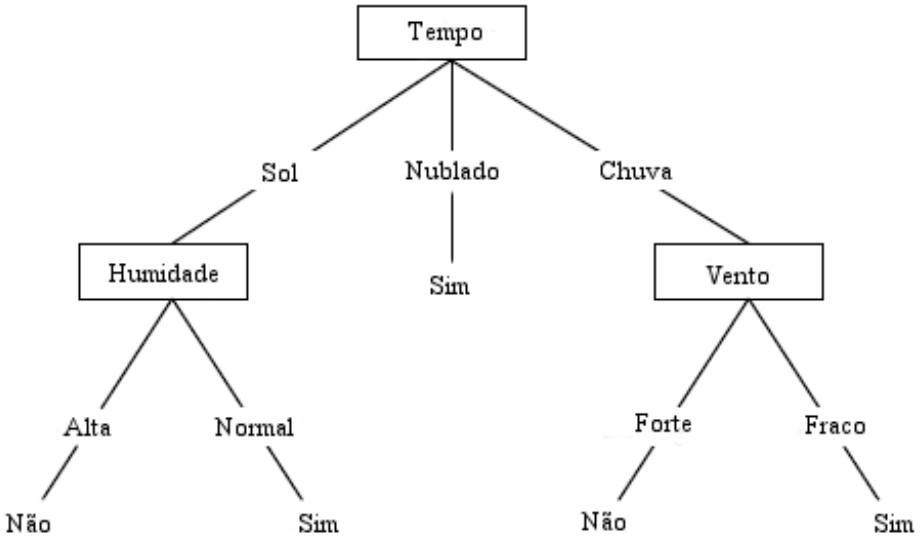


Figura 6 - Árvore de decisão que representa o conceito *JogarTenis*. Um exemplo é classificado ordenando-o ao longo da árvore até ao nó-folha apropriado, retornando em seguida a classificação associada a essa folha (neste caso, Sim ou Não) (adaptado de [Mit99])

A árvore presente na Figura 6 pode ser representada pela seguinte expressão:

Se (Tempo = Sol Λ Humidade = Normal)

V (Tempo = Nublado)

V (Tempo = Chuva Λ Vento = Fraco)

Então pode-se jogar ténis

Por outro lado, esta árvore pode também expressar as condições quando não é desejável a prática de ténis, nomeadamente quando o dia é de sol e a humidade está alta, ou então nos casos em que está a chover e o vento é forte.

De um modo geral, as árvores de decisão representam uma disjunção de conjunções, isto é, cada caminho desde a raíz da árvore até uma determinada folha diz

respeito à conjunção de atributos, e a árvore propriamente dita corresponde à disjunção dessas mesmas conjunções.

2.2.2 Regras de Classificação

Uma das representações possíveis de modelos de aprendizagem é a representação recorrendo a regras de classificação, do inglês *if-then rules*. Várias abordagens ao nível da aprendizagem produzem este tipo de regras, como por exemplo a aprendizagem de regras proposicionais e a aprendizagem de árvores [Fon06]. Aliás, tal como referenciado na subsecção 2.2.1, uma forma de "aprender" conjuntos de regras é, inicialmente, "aprendendo" uma árvore de decisão, traduzindo posteriormente essa mesma árvore num conjunto de regras equivalentes – uma regra para cada nó-folha da árvore.

Existe uma variedade de algoritmos que "aprende" conjuntos de regras. Estes algoritmos, na maioria dos casos, apresentam uma série de particularidades interessantes, nomeadamente estão aptos para a aprendizagem de regras de 1ª ordem que contêm variáveis. Este facto é significativo, uma vez que este tipo de regras é bastante mais expressivo do que as regras proposicionais. Além do mais, estes algoritmos recorrem a algoritmos sequenciais que, por seu lado, "aprendem" uma regra de cada vez até chegarem ao conjunto de regras final.

2.2.3 Programação Lógica Indutiva

A Programação Lógica Indutiva (PLI), do inglês *Inductive Logic Programming* (ILP) é um outro exemplo de uma abordagem de aprendizagem capaz de produzir regras de classificação – regras *if-then*.

Os modelos descobertos pela PLI são habitualmente representados como programas lógicos – subconjuntos de lógica de 1ª ordem, enquanto os padrões surgem como cláusulas. Um modelo é, desta forma, um conjunto de regras. Os sistemas de PLI criam modelos a partir de dados de *input* que, por sua vez, são obtidos após um processo

de treino de um determinado conjunto de exemplos. Os modelos são também frequentemente gerados a partir de conhecimento prévio, do inglês *background knowledge*. Quer os exemplos como o conhecimento prévio são representados na maioria das vezes como programas lógicos.

Vários sistemas de PLI utilizam habitualmente uma abordagem que busca (através da aprendizagem) a descoberta de padrões. Aliás, essa mesma abordagem recorre à procura de um único padrão que apresente as propriedades desejadas. O espaço de procura de padrões poderá ser extremamente vasto ou até mesmo infinito. Por isso mesmo, os sistemas de PLI frequentemente empregam estratégias de procura, tais como: a procura *greedy*, *randomized* ou mesmo a procura *branch-and-bound*. Independentemente da estratégia utilizada, cada padrão gerado é avaliado de modo a determinar a sua qualidade. Os padrões que se revelem desadequados são imediatamente descartados, enquanto os padrões potencialmente interessantes são posteriormente expandidos em etapas do processo de procura. A procura termina quando um padrão que preencha todos os requisitos é encontrado.

A avaliação de um determinado padrão visa testar se esse mesmo padrão, juntamente com a informação relativa ao conhecimento prévio permite perceber os exemplos de treino. É importante referir, no entanto, que o processo relativo à avaliação de um padrão, mesmo para pequenos conjuntos de exemplos de treino, é extremamente demorado.

De seguida, passamos a enumerar as principais vantagens da PLI [Fon06]:

- **Expressividade:** A lógica de 1ª ordem permite representar uma série de conceitos mais complexos do que as tradicionais linguagens atributo-valor.
- **Facilidade de Leitura:** É discutível o facto de que as fórmulas lógicas são de leitura mais acessível do que as árvores de decisão ou mesmo do que um conjunto de equações lineares. No entanto, são potencialmente legíveis. Se o conhecimento se encontra estruturado, uma representação de 1ª ordem é provavelmente mais fácil de ler do que uma representação de ordem zero.

- **Uso de conhecimento prévio:** O conhecimento envolvente pode ser codificado e facultado como conhecimento prévio. A fonte desse mesmo conhecimento poderá ser um "perito" ou um sistema de descoberta. Em alguns casos, o conhecimento prévio poderá crescer ao longo do próprio tempo de descoberta.

A expressividade da lógica de 1ª ordem fornece aos modelos gerados flexibilidade e compreensão. No entanto, os sistemas de PLI são afectados com limitações significativas que reduzem a sua aplicabilidade em tarefas de mineração de dados. A maioria dos sistemas de PLI executa os seus processos na memória principal, limitando a capacidade de processamento de bases de dados extensas. Além do mais, estes sistemas são computacionalmente dispendiosos - a avaliação individual de regras poderá demorar períodos de tempo consideráveis. No caso de aplicações complexas, os sistemas de PLI poderão mesmo demorar várias horas até retornarem um modelo.

Assim sendo, os baixos níveis de eficiência são, sem sombra de dúvidas, os maiores obstáculos com que os sistemas de PLI se deparam.

2.2.4 Support Vector Machines

As *Support Vector Machines* (SVM's) são um conjunto de métodos supervisionados utilizados quer para classificação como para regressão.

Em tarefas que requerem a aprendizagem de duas classes, o objectivo de uma SVM é encontrar a melhor função de classificação que permita a distinção entre membros de duas classes num conjunto de treino. Para um conjunto de dados linearmente separados, uma função de classificação linear corresponde a um hiperplano $f(\boldsymbol{x})$ que atravessa as duas classes, dividindo-as. No momento em que esta função é determinada, a nova instância $\boldsymbol{x}_n$ é classificada de acordo com o sinal da função $f(\boldsymbol{x}_n)$; $\boldsymbol{x}_n$ pertence à classe positiva se $f(\boldsymbol{x}_n) > 0$ [WKQ+07].

Uma vez que existe um número extremamente vasto de hiperplanos, o recurso a uma SVM garante que a melhor função é encontrada depois de maximizada a margem entre as duas classes. A margem em questão é a quantidade de espaço ou separação existente entre essas duas classes. Em termos geométricos, a margem corresponde à distância mais curta entre um conjunto de pontos mais próximos entre si e um determinado ponto no hiperplano. Tendo esta definição geométrica, é possível maximizar a margem, sendo que apesar de existir um número infinito de hiperplanos (Figura 7), apenas um é solução para a SVM em causa.

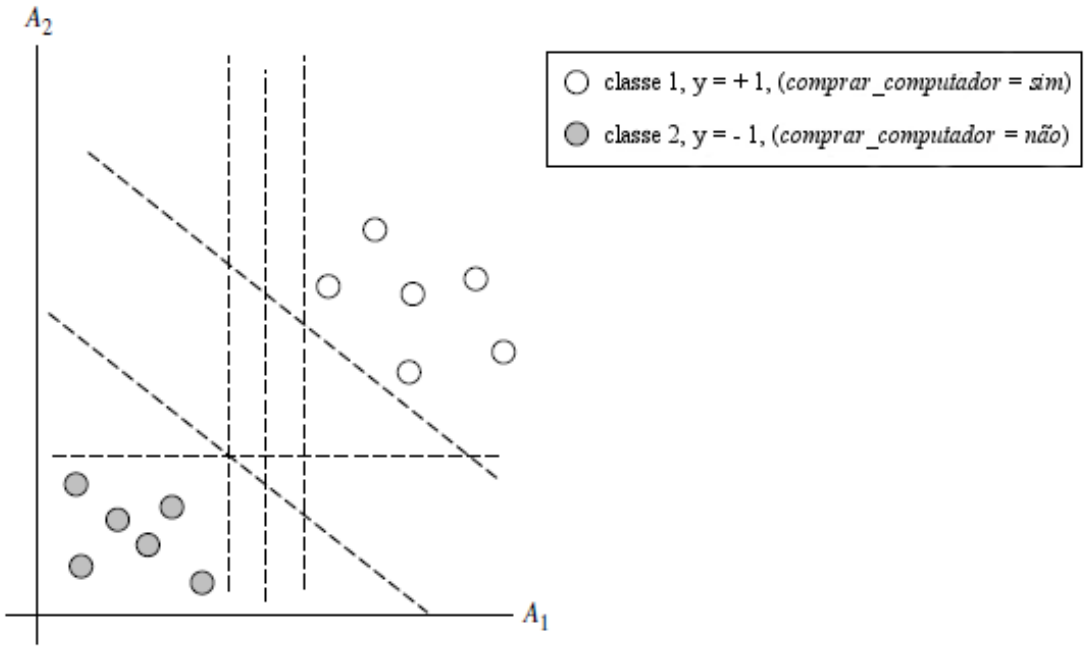


Figura 7 - Existe um número infinito de hiperplanos possíveis.

Nota: A figura em questão representa o conceito *comprar_computador*, o qual prevê se um determinado cliente de uma loja de electrónica é capaz de adquirir ou não um computador (adaptado de [HK06])

A razão pela qual uma SVM procura encontrar uma margem máxima num hiperplano, prende-se com o facto de oferecer uma melhor capacidade de generalização (Figura 8). Permite não só uma melhor performance em termos de classificação nos dados de treino, como fornece bons indicadores para uma correcta classificação de dados futuros.

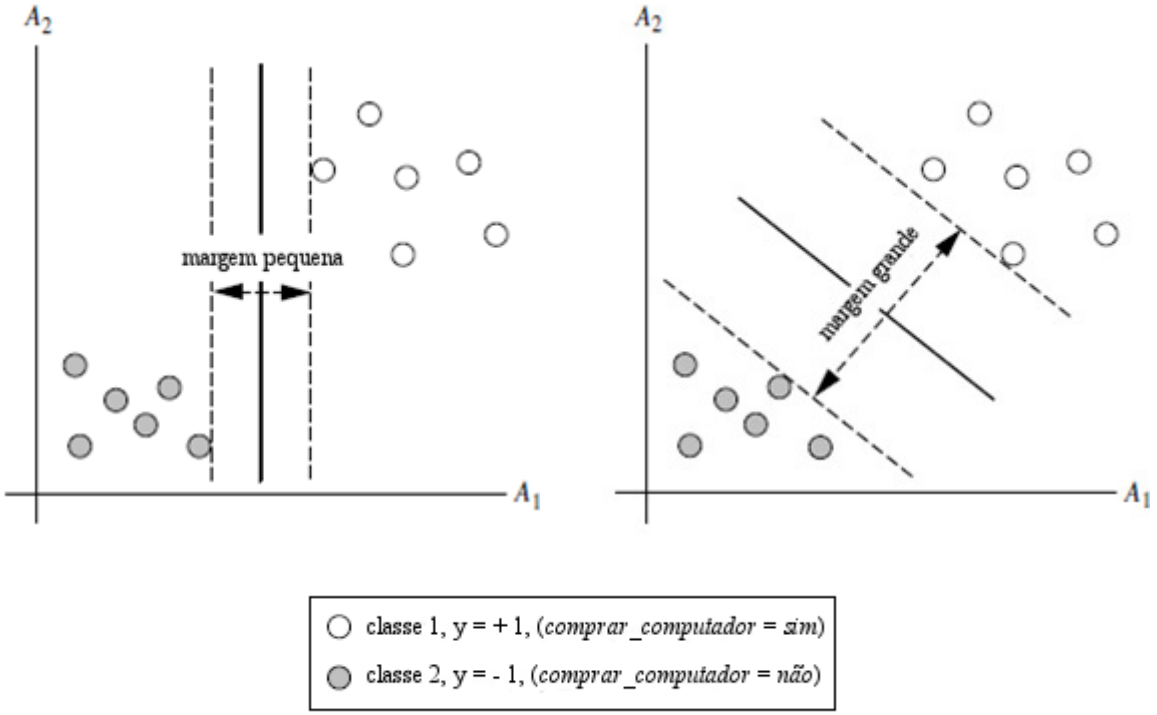


Figura 8 - Nesta figura estão presentes dois hiperplanos possíveis e respectivas margens. A margem maior, à partida, revelará uma capacidade de generalização também superior.

Nota: A figura em questão representa o conceito *comprar_computador*, o qual prevê se um determinado cliente de uma loja de electrónica é capaz de adquirir ou não um computador (adaptado de [HK06])

De modo a assegurar que é efectivamente encontrado um hiperplano com margem máxima, um classificador SVM maximiza a função seguinte em ordem a $\vec{\mathbf{w}}$ e b:

$$L_P = \frac{1}{2} \parallel \vec{\mathbf{w}} \parallel - \sum_{i=1}^{t} \alpha_i y_i (\vec{\mathbf{w}} \cdot \vec{\mathbf{x}}_\mathbf{i} + b) + \sum_{i=1}^{t} \alpha_i$$

onde t é o número de exemplos de treino, e $\alpha_i, i = 1, \ldots, t,$ são números não negativos, tal que as derivações de L_P em ordem a α_i são zero. α_i diz respeito aos multiplicadores de *Lagrange* enquanto a sigla L_P é designada por *Lagrangian*. Nesta equação, os vectores $\vec{\mathbf{w}}$ e a constante b definem o hiperplano.

Por último, é importante referir que apesar dos classificadores SVM serem extremamente precisos, acabam por se revelar relativamente lentos aquando do processamento de extensos conjuntos de dados.

2.2.5 Métodos Bayesianos

Michie *et al*. [MST94] fornecem um estudo detalhado em que comparam o classificador *naive Bayes* (um dos algoritmos de aprendizagem bayesiana) a uma série de outros algoritmos de aprendizagem, nomeadamente algoritmos relacionados com árvores de decisão e redes neuronais. Este estudo revela que o classificador *naive Bayes* é extremamente competitivo com vários destes algoritmos em inúmeras situações, e em alguns casos supera mesmo estes métodos [Mit99].

Para certas tarefas de aprendizagem, o classificador *naive Bayes* está entre os classificadores conhecidos mais eficazes. Trata-se de um algoritmo extremamente fácil de construir assim como de interpretar. A simplicidade e robustez do classificador *naive Bayes* fazem dele um bom candidato para a combinação de regras aprendidas [PK95]. Poderá até não ser o melhor classificador possível numa determinada situação, no entanto, na maioria dos casos é extremamente robusto, revelando altos níveis de performance [WKQ+07].

Sendo assim, os métodos de aprendizagem assentes em redes bayesianas são relevantes para o estudo da aprendizagem automática por duas razões essenciais. A primeira prende-se com o facto dos algoritmos de aprendizagem bayesiana que calculam probabilidades para determinadas hipóteses estarem, tal como acima mencionado, entre as abordagens mais utilizadas para a resolução de vários tipos de problemas.

A segunda razão pela qual os métodos bayesianos são importantes no estudo da aprendizagem automática diz respeito ao facto de providenciarem uma perspectiva útil na compreensão de diversos algoritmos de aprendizagem que não manipulam explicitamente probabilidades. Aliás, uma das dificuldades inerentes à aplicação de métodos bayesianos é o facto de habitualmente exigirem o conhecimento de uma série de probabilidades. Nos

casos em que estas probabilidades não são conhecidas, são frequentemente alvo de estimativa baseada em conhecimento prévio (dados disponíveis anteriormente ou suposições sobre a forma de distribuições subjacentes).

Os métodos bayesianos caracterizam-se portanto por associarem uma probabilidade a cada previsão, o que representa o nível de confiança do classificador na classificação final [DCO+04]. Outra das dificuldades que os métodos bayesianos apresentam é o custo computacional significativo necessário para determinar a hipótese de *Bayes* óptima para o caso geral [Mit99].

A seguinte equação é conhecida como regra de *Bayes*:

$$P(Z \mid Y) = \frac{P(Y \mid Z)P(Z)}{P(Y)}$$

Esta equação está subjacente a todos os sistemas actuais de inteligência artificial para inferência probabilística.

À primeira vista, a regra de *Bayes* poderá não parecer propriamente muito útil, uma vez que exige três termos (uma probabilidade condicional e duas probabilidades não condicionais) apenas para calcular uma probabilidade condicional. No entanto, a regra de *Bayes* é de facto relevante, acima de tudo, nos casos em que existem boas estimativas de probabilidades para os três termos e é necessário calcular um quarto termo. Exemplos disso mesmo são os diagnósticos médicos, em que frequentemente existem probabilidades condicionais sobre relações causais, sendo que a partir daí se pretende obter um determinado diagnóstico.

Um classificador de *Bayes* é portanto uma regra que prevê a classe mais provável para um dado exemplo, baseado na distribuição (assumida como sendo conhecida) do conjunto de dados considerado [Lee05].

Relativamente à topologia de uma rede bayesiana, esta é composta por um grafo dirigido em que cada nó representa uma variável aleatória.

Especificamente uma rede bayesiana apresenta-se do seguinte modo:

- Um conjunto de variáveis aleatórias compõe os nós da rede. As variáveis podem ser discretas ou contínuas;
- Um conjunto de arcos conecta pares de nós. Se existe um arco do nó X ao nó Y; X é designado "pai" de Y;
- Cada nó Xi apresenta uma distribuição de probabilidades condicional P (Xi | Pais (Xi)) que quantifica o efeito dos "pais" em cada nó.
- O grafo é um grafo dirigido acíclico, do inglês *Directed Acyclic Graph* (DAG).

Consideremos agora um exemplo que ilustra uma aprendizagem que recorre a uma rede bayesiana. Uma determinada pessoa (António) instalou um alarme anti-roubo na sua habitação. Trata-se de um alarme relativamente fiável na detecção de um assalto, no entanto responde ocasionalmente a pequenos tremores de terra. O António tem dois vizinhos, o João e a Maria, os quais se comprometeram a telefonar-lhe para o emprego nos momentos em que o alarme tocasse. O João telefona sempre que ouve o alarme tocar, no entanto, algumas vezes confunde o toque do telefone com o alarme e portanto liga ao António também nessas situações. A Maria, por outro lado, costuma ouvir música com o volume muito alto, logo em alguns casos não ouve o toque do alarme.

De acordo com estes dados, vamos estimar a probabilidade de ocorrer um assalto. Uma possível rede bayesiana que ilustra este problema encontra-se representada na Figura 9.

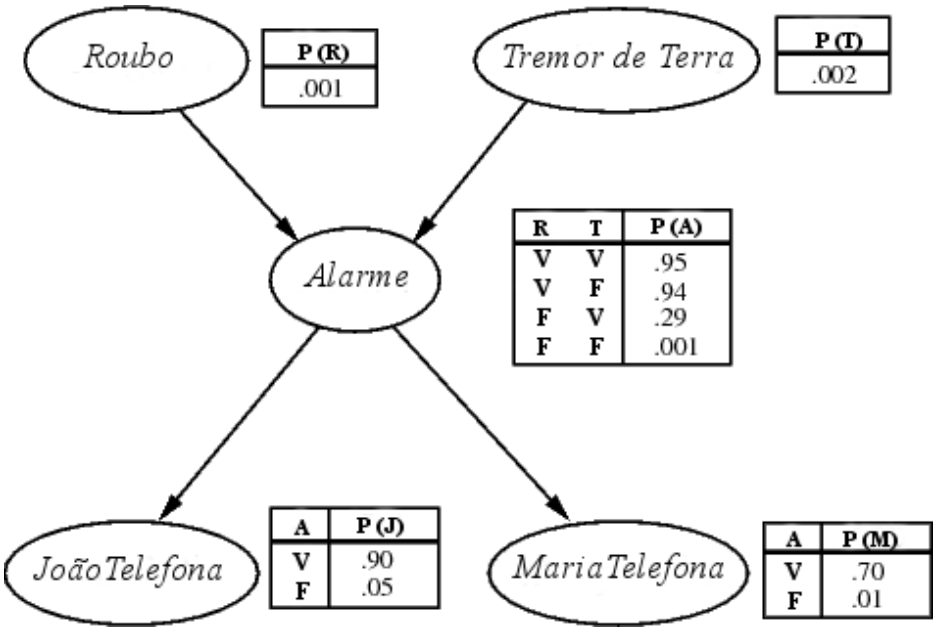


Figura 9 - Rede bayesiana onde estão presentes quer a topologia da rede como as tabelas de probabilidades condicionais.

Nota: Nas tabelas, as letras R, T, A, J, M referem-se respectivamente aos termos *Roubo* (*Assalto*), *Tremor de Terra*, *Alarme*, *JoãoTelefona* e *MariaTelefona*, enquanto as letras V e F representam os termos *Verdadeiro* e *Falso* (adaptado de [RN03])

Neste momento, ignoremos as distribuições condicionais presentes na Figura 9 e concentremo-nos na topologia da rede. No caso da rede relativa ao assalto propriamente dito, a topologia mostra que quer o assalto como o tremor de terra directamente afectam a probabilidade do alarme disparar, mas o facto do João ou a Maria telefonarem apenas depende do alarme.

É importante chamar a atenção de que a rede não apresenta nós que correspondem respectivamente às acções da Maria ouvir música com o volume alto ou ao facto do telefone tocar e confundir o João. Estes factores estão resumidos nas incertezas associadas com as ligações do *Alarme* ao *JoãoTelefona* e *MariaTelefona*. De facto, as probabilidades sumarizam um conjunto infinito de circunstâncias nas quais o alarme poderia não disparar (humidade elevada, falha de energia, etc.) ou o facto do João e da Maria falharem na sua missão de alertar o António (jantar fora, de férias, etc.).

Foquemo-nos agora nas distribuições condicionais presentes na Figura 9. Nesta figura, cada distribuição é exibida como uma tabela de probabilidades condicionais. Por

sua vez, cada linha nas tabelas contém a probabilidade condicional do valor de cada nó para um caso condicionado. Um caso condicionado trata-se de uma combinação possível de valores para os nós-pai.

No caso de variáveis booleanas, a partir do momento que é conhecida que a probabilidade de ocorrência de um valor *Verdade* é *p*, a probabilidade de ocorrência de um valor *Falso* terá que ser obrigatoriamente 1-*p*, daí a omissão do segundo número na Figura 9. Geralmente, a tabela para uma variável booleana com *k* pais booleanos contém 2^k probabilidades independentes. Um nó sem "pais" apresenta apenas uma linha, representando as probabilidades de cada valor possível da variável [RN03].

Existem duas formas de compreender a semântica de uma rede bayesiana. A primeira sugere a visualização da rede como uma representação da distribuição de probabilidades conjuntas. A segunda remete-nos para uma espécie de codificação de uma colecção de declarações condicionais independentes. As duas formas são equivalentes, apesar de que a primeira acaba por ser útil na compreensão do modo como as redes são construídas, enquanto a segunda visualização torna-se útil no desenvolvimento de procedimentos de inferência.

Uma rede bayesiana fornece uma descrição completa do domínio que representa. Qualquer entrada na distribuição de probabilidades conjuntas pode ser calculada através da informação presente na rede. Uma entrada genérica na distribuição conjunta é a probabilidade de um conjunto de tarefas específicas para cada variável, segundo a notação $P(X_1 \ldots X_n)$. O valor desta entrada é dado pela fórmula:

$$P(X_1 \ldots X_n) = \prod_{i=1}^{n} P(X_i \mid \quad pais\ (X_i))$$

onde $pais(X_i)$ revela os valores específicos das variáveis em $Pais(X_i)$. Sendo assim, cada entrada na distribuição conjunta é representada pelo produto dos elementos adequados das tabelas de probabilidades condicionais na rede bayesiana. Estas tabelas fornecem uma representação decomposta da distribuição conjunta.

É possível ilustrar todo este processo através do cálculo, por exemplo, da probabilidade do alarme tocar, mas nem um assalto nem um tremor de terra terem ocorrido, e no entanto quer a Maria como o João terem telefonado.

Assim sendo, tendo em conta a Figura 9 e de acordo com a regra de *Bayes*, temos:

$$P(j \Lambda m \Lambda a \Lambda \neg r \Lambda \neg t)$$

$$= P(j \mid a)\, P(m \mid a)\, P(a \mid \neg r \Lambda \neg t)\, P(\neg r)\, P(\neg t)$$

$$= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 = 0.00063$$

(**Nota:** As letras j, m, a, r, t, representam respectivamente as palavras *JoãoTelefona*, *MariaTelefona*, *Alarme*, *Roubo* (*Assalto*), *Tremor de Terra*).

Uma distribuição conjunta pode ser utilizada para responder a qualquer questão sobre o domínio em causa. Deste modo, se uma rede bayesiana é a representação de uma distribuição conjunta, então também poderá ser utilizada para responder a qualquer questão, nomeadamente através da soma de todas as entradas conjuntas consideradas relevantes.

2.3 Validação dos Métodos de Aprendizagem Automática

A aplicação dos diferentes métodos de aprendizagem automática requer um processo que permita garantir que os resultados obtidos sejam fiáveis e estatisticamente significativos. Existem inúmeras abordagens que asseguram a avaliação da qualidade e características de um modelo e que incluem nomeadamente a utilização de métricas de validade estatística, que têm como principal objectivo detectar possíveis anomalias nos dados ou no próprio modelo.

É extremamente importante analisar a exactidão e a confiança de um determinado modelo. Sendo assim, a exactidão trata-se de uma medida que revela se o modelo em causa está de acordo com os resultados obtidos, fazendo uso das características extraídas dos dados fornecidos. A confiança, por sua vez, avalia o modo como um modelo se

comporta em conjuntos de dados diferentes. Caso o modelo gere o mesmo tipo de previsões ou então localize padrões semelhantes (independentemente dos dados de teste fornecidos), poderemos dizer que estamos perante um modelo fiável.

Existem diversos tipos de validação, nomeadamente:

- **Validação de sub-amostras aleatórias repetidas:** Este método divide aleatoriamente o conjunto de dados para treino e para validação. A desvantagem associada à utilização deste método recai no facto de que uma determinada amostra de elementos poderá nunca ser alvo de selecção enquanto, por exemplo, uma outra amostra poderá ser escolhida várias vezes.

- ***N fold Cross-Validation*:** Método de validação em que os dados são divididos em *N* subconjuntos (blocos de dimensão semelhante – *folds*) para uma aprendizagem de *N* iterações. Ao longo do processo de treino são utilizados *N* – 1 blocos, e apenas um para teste, sendo este diferente a cada iteração [Cru07]. Este processo é repetido para as *N* amostras. A performance do classificador é definida de acordo com a média dos *N* testes. A vantagem da aplicação deste método prende-se, acima de tudo, com o facto de todos os dados serem utilizados.

- **Validação *Leave-One-Out*:** Método semelhante ao *N fold cross-validation*, diferindo apenas no tamanho da amostra que, neste caso particular, é de apenas um elemento no conjunto de teste. Deste modo, sendo *l* o tamanho do *dataset*, o treino é efectuado com *l* – 1 elementos, sendo o teste posteriormente realizado com o elemento reservado (elemento de teste) [DKG00].

- **Validação *Hold-Out Percentage Split*:** O conjunto de teste é escolhido de modo aleatório, habitualmente cerca de 20 a 30% dos elementos. Os restantes dados são alvo de treino e em seguida validados no conjunto reservado (conjunto de teste) [DKG00].

Este conjunto de formas de validação e avaliação de classificadores procura, tal como referido anteriormente, garantir que os resultados são fiáveis e estatisticamente significativos. No entanto, o processo de avaliação de um determinado modelo necessita igualmente levar em consideração dois factores extremamente comuns aquando da utilização de universos de dados extensos: as **classes desbalanceadas** e o problema de ***overfitting***; frequentemente responsáveis pela baixa qualidade dos resultados obtidos na classificação de dados, e que agora passamos a explicar:

- **Classes desbalanceadas:** [LR06] Vários algoritmos de aprendizagem automática consideram que os valores que uma determinada classe poderá assumir, apresentarão, à partida, probabilidades iguais. Esse facto nem sempre ocorre, tal como é exemplo o conjunto de dados alvo de estudo nesta dissertação. Sendo assim, nesta situação particular, tal como em muitas outras relacionadas com sistemas de detecção de células cancerígenas, o número de casos anormais (malignos) que estão disponíveis para treino é consideravelmente inferior ao número de casos ditos normais (benignos). Este desbalanceamento poderá afectar a taxa de acertos para a classe de menor ocorrência.

  Algumas abordagens com vista ao balanceamento do conjunto de dados poderão envolver desde a remoção de tuplos da classe dominante até à replicação aleatória de tuplos da classe de menor ocorrência. No entanto, estas duas perspectivas acarretam igualmente alguns senãos. Na primeira abordagem existe o problema de dados potencialmente úteis serem eliminados. Na segunda abordagem, a partir do momento em que o conjunto de treino é aumentado, consequentemente, o tempo de aprendizagem também será maior. Segundo Hoste [Hos05] esta segunda abordagem poderá igualmente conduzir ao problema de *overfitting* (perda de capacidade de generalização) quando utilizada com árvores de decisão. Apesar de tudo, estas duas perspectivas aumentam significativamente o desempenho dos classificadores em algumas situações [Hos05].

- ***Overfitting*:** Situação em que o modelo gerado adapta-se bastante bem aos casos utilizados na aprendizagem, no entanto apresenta fracos resultados nos casos de teste [Cru07]. Quando um algoritmo procura pelos melhores parâmetros para um determinado modelo, utilizando um conjunto de dados limitado, poderá modelar não apenas os padrões gerais, mas também ruídos específicos do próprio conjunto de dados, resultando numa fraca performance do modelo nos dados de teste. Uma das possíveis soluções para este problema é a aplicação de *cross-validation* aos dados [FPSS96].

2.3.1 Métricas de Desempenho

As métricas de desempenho garantem igualmente a fiabilidade dos resultados. Tratam-se de medidas numéricas que quantificam a performance de um determinado classificador [Rae08].

Sendo assim, em seguida apresentamos as diferentes métricas utilizadas, de modo a certificar a qualidade dos resultados obtidos.

(**Nota:** Ao longo desta subsecção iremos concentrar-nos em problemas com apenas duas classes, no entanto, é importante referir que as noções aqui expostas poderão ser estendidas a várias classes. Deste modo, como tratamos duas classes apenas, serão utilizadas as siglas: TP, TN, FP e FN que representam respectivamente os termos *True Positive* (Verdadeiros Positivos), *True Negative* (Verdadeiros Negativos), *False Positive* (Falsos Positivos) e *False Negative* (Falsos Negativos)).

- **Matriz de Confusão**

Utilizada em classificação, a matriz de confusão, do inglês *confusion* (ou *contingency*) *matrix*, possibilita uma visualização inequívoca dos resultados de um determinado modelo [KP98]. Os resultados são apresentados sob a forma de uma tabela de duas entradas (considerando problemas de apenas duas classes): uma das entradas é constituída pelas classes desejadas, a outra pelas classes previstas pelo modelo. As células, por sua vez, são preenchidas com o número de instâncias que correspondem ao cruzamento das entradas.

Uma matriz de confusão é definida do seguinte modo:

	positivos previstos	negativos previstos
positivos originais	*TP*	*FN*
negativos originais	*FP*	*TN*

em que:

TP é o número de **previsões correctas** para uma instância que é **positiva**;
FN é o número de **previsões incorrectas** para uma instância que é **positiva**;
FP é o número de **previsões incorrectas** para uma instância que é **negativa**;
TN é o número de **previsões correctas** para uma instância que é **negativa**;

A Figura 10 ilustra um exemplo de uma matriz de confusão, em que a entrada vertical são as classificações obtidas por um modelo, e a entrada horizontal são as classificações originais dos dados. É possível constatar que no caso da classe *high*, de um universo de 81 instâncias, foram classificadas correctamente 49, e incorrectamente 32. Já no caso da classe *iso*, de um conjunto de 99 instâncias, 71 foram classificadas correctamente, sendo que 28 instâncias foram classificadas incorrectamente.

```
 a  b   <-- classified as
49 32 |  a = high
28 71 |  b = iso
```

Figura 10 - Exemplo de uma matriz de confusão

- **Instâncias Correctamente Classificadas (*Accuracy*)**

A percentagem de instâncias correctamente classificadas (ICC), também conhecida como *accuracy*, é a percentagem de instâncias que o classificador previu correctamente. Corresponde à taxa de exemplos positivos e negativos correctamente classificados. Esta métrica é calculada de acordo com a seguinte fórmula:

$$\text{ICC} = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

- **Instâncias Incorrectamente Classificadas**

A percentagem de instâncias incorrectamente classificadas (IIC) é o número de instâncias que o classificador previu incorrectamente. Corresponde à taxa de exemplos positivos e negativos incorrectamente classificados. Esta métrica é calculada de acordo com a seguinte fórmula:

$$\text{IIC} = \frac{(FP + FN)}{(TP + TN + FP + FN)}$$

- **Precisão**

Precisão, do inglês *precision*, é uma medida que originalmente foi introduzida com o objectivo de medir a eficácia de um motor de busca ao retornar informação considerada relevante. Nesse caso concreto, a precisão é a fracção de documentos recuperados por um motor de busca e que são igualmente relevantes.

Por sua vez, na avaliação de classificadores, a precisão é definida como [Rae08]:

$$\text{Precisão} = \frac{TP}{(TP + FP)}$$

- ***Recall* (Sensibilidade, Taxa de Verdadeiros Positivos)**

Na avaliação de classificadores, *recall*, sensibilidade e taxa de verdadeiros positivos (TVP), do inglês *True Positive Rate* (TPR), têm em comum o facto de serem definidos segundo a mesma fórmula [Rae08]:

$$Recall = \text{Sensibilidade} = \text{TVP} = \frac{TP}{(TP + FN)}$$

- **Taxa de Verdadeiros Negativos (Especificidade)**

Taxa de Verdadeiros Negativos (TVN), do inglês *True Negative Rate* (TNR) quantifica a proporção de casos negativos que foram correctamente classificados.

Quer a taxa de verdadeiros negativos como a especificidade são ambas definidas de acordo com a seguinte fórmula:

$$\text{Especificidade} = \text{TVN} = \frac{TN}{(TN + FP)}$$

- **Taxa de Falsos Positivos**

Taxa de Falsos Positivos (TFP), do inglês *False Positive Rate* (FPR) quantifica a proporção de casos negativos que foram incorrectamente classificados como positivos. É definida como:

$$\text{TFP} = \frac{FP}{(FP + TN)} = 1 - \text{Especificidade}$$

- **Taxa de Falsos Negativos**

Taxa de Falsos Negativos (TFN), do inglês *False Negative Rate* (FNR) quantifica a proporção de casos positivos que foram incorrectamente classificados como negativos. É definida como:

$$\text{TFN} = \frac{FN}{(TP + FN)} = 1 \text{ - Sensibilidade}$$

- ***F-Measure***

F-Measure mede a eficácia de um classificador, nomeadamente em termos de precisão e *recall*. É possível definir uma medida *F-Measure* que atribua peso arbitrário quer para precisão como para *recall*. Essa medida é conhecida como F_1-*Measure*, uma vez que atribui igual importância a essas duas métricas [Rae08]. A fórmula para F_1-*Measure* (média harmónica entre precisão e *recall*) é a seguinte:

$$F = \frac{2 \cdot \text{Precisão} \cdot Recall}{\text{Precisão} + Recall}$$

- **Estatística *Kappa***

Inúmeras pessoas, cujo foco de trabalho é a observação e interpretação de exames médicos, como por exemplo, a interpretação de mamografias, assim como uma série de outros exames de diagnóstico, habitualmente se deparam com situações em que existem diversas opiniões para um mesmo caso. Os estudos que medem a concordância entre dois ou mais observadores devem incluir uma estatística que tome em consideração o facto de que em certas ocasiões os observadores poderão concordar ou discordar apenas por acaso.

A estatística *Kappa* (ou coeficiente *Kappa*) é a estatística mais utilizada para abordar este tipo de problema.

A equação que traduz esta estatística é definida do seguinte modo:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

em que Pr(a) é a concordância observada relativa entre os avaliadores e Pr(e) é a probabilidade hipotética de ocorrer concordância por simples acaso, fazendo uso dos dados observados para calcular as probabilidades de cada observação.

Um *kappa* igual a 1 indica concordância perfeita, enquanto um *kappa* igual a 0 indica concordância equivalente a um simples acaso [VG05].

- **Área ROC**

A curva ROC (*Receiver Operating Curve*) representa a taxa de verdadeiros positivos (TVP) em função da taxa de falsos positivos (TFP).

A área sob a curva ROC varia entre 0 e 1, sendo que 1 representa o classificador perfeito, e 0 um classificador que está sempre errado. Uma área ROC de 0.5 indica um classificador que é aproximadamente aleatório.

A área ROC é habitualmente escolhida em detrimento da *accuracy* aquando da utilização de conjuntos de dados que se apresentam balanceados, uma vez que captura mais eficazmente o equilíbrio entre verdadeiros positivos e verdadeiros negativos [Rae08].

Além do mais, no âmbito da saúde por exemplo, a área da curva ROC permite estabelecer uma relação entre a sensibilidade de um teste diagnóstico e a especificidade, como limiar para indicação da variação positiva de um teste. É

frequentemente utilizada para escolha de diferentes testes de diagnóstico, apesar de não ter em conta a prevalência da patologia testada [MMC09].

- **Curvas *Precision-Recall***

As curvas ROC poderão apresentar perspectivas demasiado optimistas quanto à performance de um determinado algoritmo nos casos em que possa existir um desbalanceamento grande na distribuição de classes.

As curvas *Precision-Recall* (PR) apresentam-se como uma alternativa às curvas ROC para tarefas que envolvam conjuntos de dados desbalanceados. Uma diferença bastante grande entre um espaço ROC e um espaço PR é a própria representação visual das curvas. As curvas PR poderão conduzir à detecção de diferenças entre algoritmos que, apenas pela análise de uma área ROC, eram imperceptíveis à partida. Exemplos de curvas ROC e PR são apresentados na Figura 11. Estas curvas, construídas a partir dos mesmos modelos de aprendizagem referentes a um conjunto de dados altamente desbalanceado, permitem tornar mais evidentes as diferenças entre estes espaços.

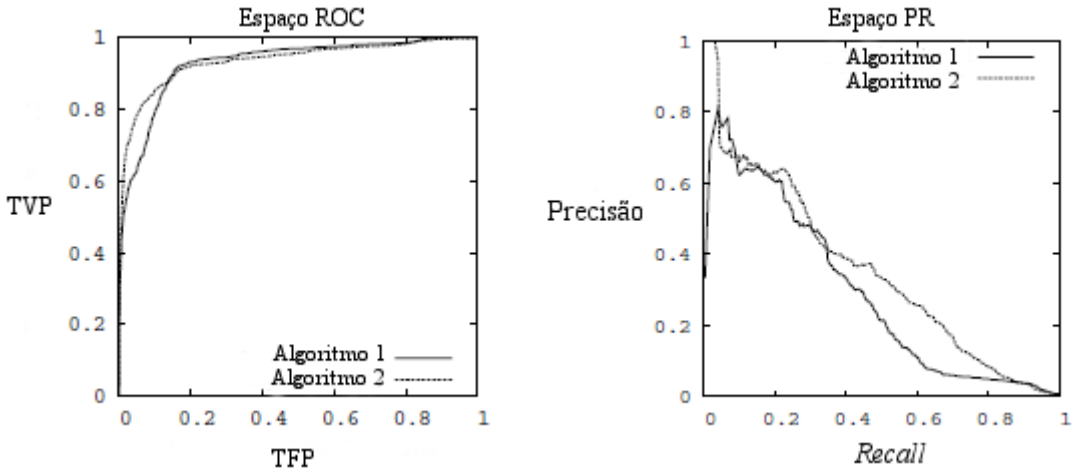


Figura 11 - Diferenças entre comparar algoritmos num espaço ROC e num espaço PR (adaptado de [DG06]

Como o objectivo de um espaço ROC é que este se situe no canto superior esquerdo, quando observamos as curvas presentes no gráfico da esquerda, ficamos com uma ideia de que se trata de um espaço ROC que se aproxima bastante desse cenário ideal.

Num espaço PR, por sua vez, o objectivo principal é que este se situe no canto superior direito. Deste modo, analisando as curvas do gráfico da direita é possível constatar que existe ainda uma enorme margem de progressão para aperfeiçoamentos [DG06].

Assim sendo, podemos concluir que a área ROC é preferencialmente escolhida para análise de resultados quando o conjunto de dados que está a ser alvo de estudo é balanceado. No sentido oposto, isto é, em casos de desbalanceamento na distribuição de classes, as curvas *Precision-Recall* são a melhor forma de compreender a qualidade dos resultados obtidos.

2.4 WEKA

Não sendo o objectivo desta dissertação constituir um manual de utilização de ferramentas de mineração de dados é, no entanto, oportuno efectuar uma breve descrição das ferramentas utilizadas. Deste modo, em seguida iremos descrever as funcionalidades essenciais apresentadas pelo sistema WEKA, uma vez que se tratou da ferramenta a que recorremos para a realização de todas as experiências inerentes a este trabalho.

Criado pela Universidade de Waikato na Nova Zelândia, o software WEKA (*Waikato Environment for Knowledge Analysis*) foi desenvolvido na linguagem de programação Java (linguagem orientada a objectos), sendo que implementa uma grande variedade de técnicas [WF05]. Uma vez que é escrito em Java, o código encontra-se apto para ser executado em diferentes plataformas, conferindo um certo grau de portabilidade ao sistema.

Disponibiliza igualmente diversos algoritmos de pré-processamento de dados, bem como de análise de resultados. O conjunto de técnicas que implementa permite a utilização da ferramenta em diversos problemas, desde classificação até regressão, por exemplo.

Grande parte dos recursos do software WEKA encontra-se acessível através da sua interface gráfica, que passamos a descrever.

2.4.1 Interface Gráfica

A interface gráfica da ferramenta WEKA, do inglês *Graphical User Interface* (GUI), possui uma janela – WEKA GUI *Chooser* – (Figura 12) que permite aos utilizadores escolherem quais as aplicações que pretendem utilizar de modo a extraírem informação dos seus dados.

Figura 12 - Janela inicial do WEKA (GUI Chooser)

Este menu é composto por quatro botões, cada um deles para cada uma das quatro principais funcionalidades que compõem o sistema WEKA. Sendo assim, estes botões poderão ser utilizados para despoletarem as seguintes aplicações [Cru07]:

- **Explorer**

  Proporciona um ambiente gráfico de manipulação de dados pela utilização de diversos algoritmos. Trata-se da interface mais fácil de usar, conduzindo o utilizador através de menus e formulários, impedindo-o de fazer escolhas não aplicáveis e simultaneamente apresentando *pop-ups* de informação relativos ao preenchimento de vários campos. Embora seja intuitivo, torna-se necessário evidenciar alguns elementos estratégicos desta aplicação *Explorer*. Sendo assim, a Figura 13 apresenta elementos de pré-processamento [Sil04]:

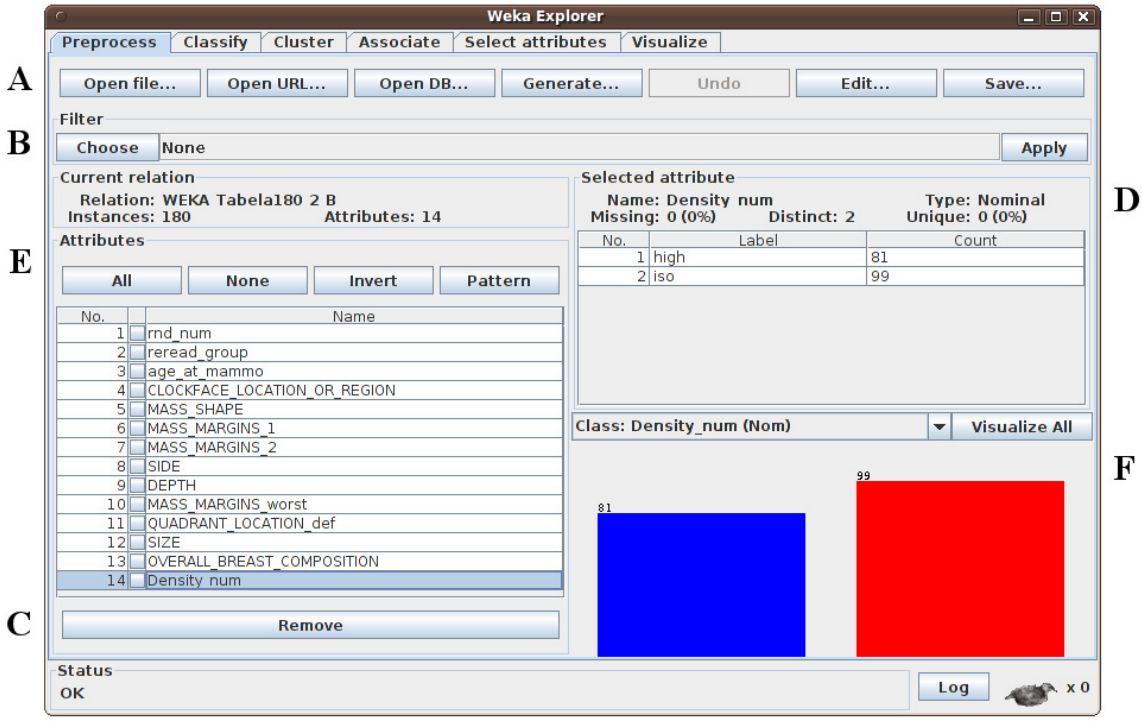


Figura 13 - Pré-processamento no WEKA Explorer (*Preprocess*)

- **(A) –** *Open File*, *Open URL*, *Open DB*: através destes botões é possível seleccionar, respectivamente, bases de dados a partir de ficheiros locais (formato *arff*), bases de dados remotas (Web) ou apenas diferentes tipos de bases de dados (via JDBC[10]);

- **(B)** – No botão *Filter* é possível efectuar sucessivas filtragens de atributos e instâncias na base de dados previamente carregada (aplicação de operações de selecção, discretização, entre outras);

- **(C)** – Uma vez escolhidos os principais atributos que serão alvo de estudo, todos os outros poderão ser removidos através do botão *Remove*, que se encontra no final da lista de atributos;

[10] *Java Database Connectivity* – Conjunto de classes e interfaces escritas em Java que fazem o envio de instruções SQL para qualquer base de dados relacional.

- **(D)** – Navegando interactivamente pelos atributos (quadro *Attributes* **(E)**) é possível obter informações quantitativas e estatísticas sobre os mesmos (quadro *Selected attribute* **(D)**). Por exemplo, o atributo seleccionado na lista de atributos da Figura 13 - *Density_num* - permite-nos constatar que a distribuição de valores *high* e *iso* na base de dados é relativamente homogénea, tal como ilustrado pelos histogramas coloridos presentes no canto inferior direito da imagem **(F)**. Sendo assim, neste caso concreto, temos 81 instâncias do tipo *high* (rectângulo azul) e 99 instâncias do tipo *iso* (rectângulo vermelho).

No WEKA *Explorer* é também possível desenvolver tarefas de classificação, tal como ilustrado na Figura 14:

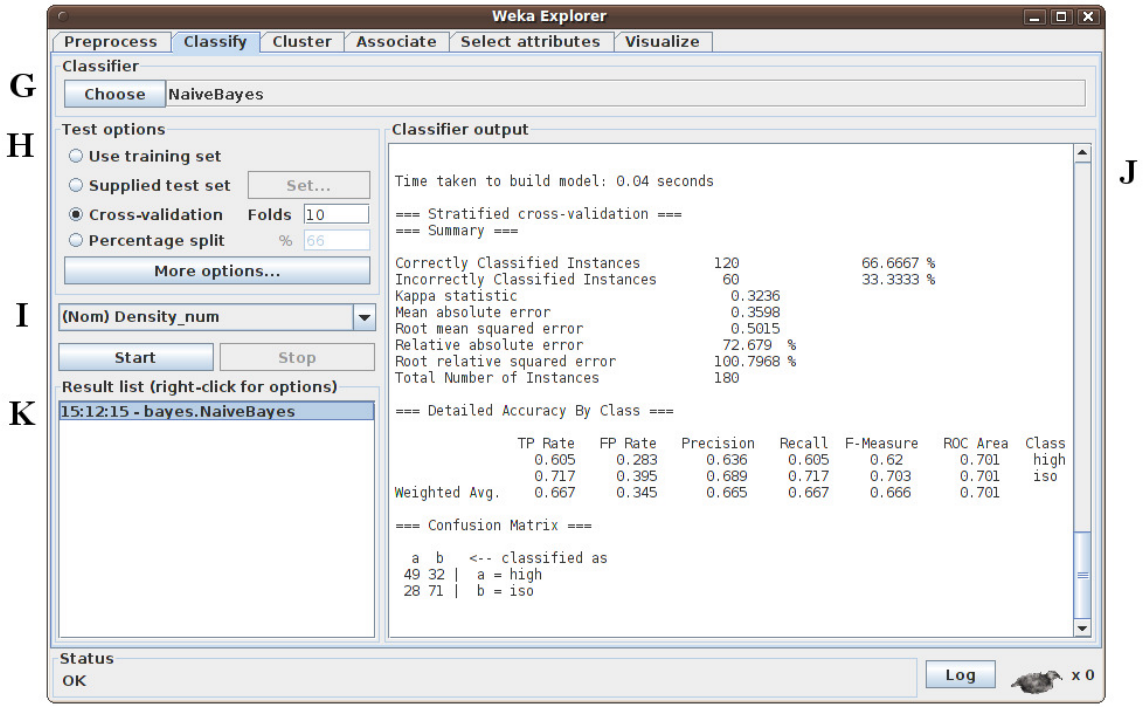


Figura 14 - Classificação no WEKA Explorer (*Classify*)

- **(G)** – Selecção e parametrização do algoritmo a ser utilizado (J48, *RandomForest*, SMO, *naive Bayes*, *BayesNet*, etc.);

- **(H)** – Permite seleccionar a opção de teste e validação do modelo gerado. Entre as opções de teste temos: a utilização do próprio conjunto de dados de treino, o uso de um outro conjunto apenas para testes, a aplicação de *cross-validation* aos dados, assim como a separação de parte do conjunto de treino para teste (*Hold-Out Percentage Split*);
- **(I)** – Selecção do atributo classe para a tarefa de classificação;
- **(J)** – Resumo da tarefa efectuada com dados estatísticos, nomeadamente métricas de desempenho, matrizes de confusão, entre outros;
- **(K)** – Pressionando o botão direito do rato em cima de uma entrada que se encontre na *Result list*, é possível aceder a um menu que permite, entre outras coisas, a visualização de uma representação gráfica da estrutura do classificador gerado (i.e. árvores de decisão, redes bayesianas), assim como possibilita a visualização das respectivas curvas ROC e PR.

As opções *Cluster*, *Associate* e *Select attributes* possuem interfaces semelhantes, fornecendo determinadas opções a estas tarefas. No caso de tarefas de *clustering* a interface disponibiliza a opção de ignorar atributos, uma vez que é extremamente comum que neste tipo de tarefa, um ou mais atributos gerem "ruído" ao longo deste processo. Já na fase de selecção (*Selected attributes*) é possível escolher o algoritmo avaliador de atributos, assim como o método de busca para a tarefa em causa [Sil04].

- **Experimenter**

Permite testar técnicas diferentes em classificação ou regressão, de modo a compará-las. Apesar destas operações serem igualmente possíveis quer no *Explorer* como no *KnowledgeFlow*, no *Experimenter*, no entanto, é possível escolher desde diversos conjuntos de dados a serem utilizados numa só experiência, como várias técnicas a serem experimentadas, e até o número de repetições (*runs*) do teste em questão, entre outras escolhas.

Posteriormente, a experiência em causa é executada sem ser necessária a supervisão do utilizador. Os resultados são depois guardados num ficheiro para análise. É fundamental referir que esta interface revelou-se a interface ideal para a execução de experiências, pelo que foi utilizada para a realização dos diversos ensaios relativos a esta dissertação.

- **KnowledgeFlow**

Permite o desenvolvimento de projectos de mineração de dados num ambiente gráfico com fluxos de informação. Por outro lado, de entre as várias vantagens que possui, é de destacar o *layout* intuitivo, assim como o facto de permitir o processamento de dados em *batch*[11] ou de modo incremental, que por sua vez permitem a sua aplicação a conjuntos de dados de elevada dimensão. Além do mais, possibilita o processamento paralelo, em que cada fluxo de dados distinto é processado na respectiva *thread*[12].

- **SimpleCLI (*Command Line Interface*)**

Proporciona uma interface que permite a execução directa de comandos do WEKA. Embora disponibilize todas as funcionalidades, requer um elevado grau de conhecimento dos comandos que poderão ser utilizados.

[11] Termo referente a um processamento de dados que ocorre através de um conjunto de tarefas que se encontram enfileiradas, sendo que o sistema operativo apenas processa a próxima tarefa após o término completo da tarefa anterior.

[12] Forma de um determinado processo se dividir em duas ou mais tarefas que possam ser executadas simultaneamente.

O formato de ficheiros utilizado no decorrer das experiências foi o formato *arff*. No próprio cabeçalho deste tipo de ficheiros são descritos os atributos, tal como ilustrado no seguinte exemplo:

```
@relation WEKA_Tabela180_2_B

@attribute rnd_num numeric
@attribute reread_group {salkowski,burnside,sisney}
@attribute age_at_mammo numeric
@attribute CLOCKFACE_LOCATION_OR_REGION {11.0,12.0,C,7.0,1.0,10.0,3.0,6.0,2.0,4.0,5.0,8.0,9.0}
@attribute MASS_SHAPE {X,R,O,L}
@attribute MASS_MARGINS_1 {I,S,D,M,U}
@attribute MASS_MARGINS_2 {I,S,D,M,U}
@attribute SIDE {R,L}
@attribute DEPTH {P,M,A}
@attribute MASS_MARGINS_worst {Indistinct,Spiculated,Circumscribed,Microlobulated,Obscured}
@attribute QUADRANT_LOCATION_def {'Upper Outer','Upper Inner','Lower Inner','Lower Outer'}
@attribute SIZE numeric
@attribute OVERALL_BREAST_COMPOSITION {'scattered fibroglandular densities','almost entirely fat',
@attribute Density_num {high,iso}
```

Figura 15 - Exemplo do conteúdo de um ficheiro do tipo *arff*

Para a criação de ficheiros deste tipo foi utilizado o próprio conversor disponibilizado pela ferramenta WEKA, sendo que os dados foram carregados em formato *csv*[13], ou seja, separados por vírgulas.

Os algoritmos que implementam SVM's utilizam o método SMO (para tarefas de classificação) enquanto, por exemplo, no que diz respeito às árvores de decisão um dos algoritmos utilizados para classificação é o J48, que mais não é do que uma simples implementação para o WEKA do famoso algoritmo C4.5 (criado por J. Quinlan) [WKQ+07]. Sendo assim, tal como vimos, é possível referir que a ferramenta WEKA permite aplicar as quatro tarefas principais de aprendizagem automática apresentadas na subsecção 2.1.2 relativa à mineração de dados, ou seja, podemos dizer que esta ferramenta permite a aplicação das tarefas de classificação, regressão, regras de associação e *clustering* a inúmeros conjuntos de dados.

[13] *comma separated values*.

Como nota de conclusão, o WEKA disponibiliza portanto uma variedade bastante grande de algoritmos de mineração de dados, desde algoritmos relativos a redes neuronais até *support vector machines* (SVM's), árvores de decisão, entre outros. Como tal, na subsecção seguinte iremos concentrar-nos nos algoritmos/classificadores a que recorremos para a realização das experiências.

2.4.2 Classificadores

Tal como mencionado anteriormente, o sistema WEKA reúne um vasto conjunto de algoritmos de aprendizagem automática para a resolução de tarefas de mineração de dados.

Deste modo, durante a utilização desta ferramenta, doze desses algoritmos foram aplicados ao universo de dados alvo de estudo. A escolha destes doze algoritmos prende-se com o facto de estarem subjacentes a praticamente todos os métodos de aprendizagem automática, desde árvores de decisão até regras de classificação, *support vector machines*, redes bayesianas, etc. Com esta abordagem multidisciplinar pretende-se, acima de tudo, estudar o comportamento dos diferentes algoritmos na classificação dos dados em causa e por conseguinte extrair os classificadores que se revelem mais exactos.

Sendo assim, a Tabela 1 apresenta uma síntese das principais características dos doze algoritmos a que recorremos para a execução das diversas experiências.

Métodos de Aprendizagem Automática	Algoritmos	Características
Redes Bayesianas	**BayesNet**	Rede de aprendizagem *Bayes* que utiliza vários algoritmos de procura e métricas de qualidade. Algoritmo de base para um classificador que segue a estrutura de uma rede bayesiana.
	NaiveBayes	Classificador probabilístico baseado na aplicação do teorema de *Bayes* (estatística bayesiana). É designado *naive* uma vez que os valores dos atributos são condicionalmente independentes.
Support Vector Machines	**SMO**	Algoritmo eficiente para a implementação da técnica SVM. Substitui todos os valores em falta e transforma atributos nominais em binários. Normaliza, por *default*, todos os atributos.
Árvores de Decisão	**DecisionStump**	Habitualmente utilizado em conjunto com algoritmos que recorrem à técnica de *boosting*. Aplica as tarefas de regressão (baseado na métrica *mean-squared error*) ou classificação (baseado na entropia).
	J48	Implementação em Java do algoritmo C4.5. Gera uma árvore de decisão.
	NBTree	Algoritmo que gera uma árvore de decisão com classificadores *naive Bayes* nas folhas. Constrói uma rede bayesiana para cada folha. Algoritmo que apresenta bons resultados para conjuntos de dados bastante grandes.
	RandomForest	Algoritmo responsável pela construção de uma floresta de árvores aleatórias.
	SimpleCart	Algoritmo responsável pela construção de árvores de decisão binárias.
	DTNB	Algoritmo responsável pela construção e utilização de uma árvore de decisão e de um classificador híbrido *naive Bayes*. Algoritmo que apresenta bons resultados para conjuntos de dados pequenos.
Regras de Classificação	**OneR**	Cria uma regra para cada atributo dos dados de treino e selecciona a regra com menor percentagem de erro como regra única.
	PART	Algoritmo que gera uma lista de decisão. Usa a estratégia de divisão e conquista. Constrói de forma parcial uma árvore de decisão C4.5 em cada iteração, transformando a "melhor folha" numa regra.
	ZeroR	Algoritmo extremamente simples que classifica todas as instâncias de acordo com a classe dominante. Utilizado como classificador de referência.

Tabela 1 - Síntese dos doze algoritmos aplicados ao universo de dados alvo de estudo

De entre estes doze algoritmos destacamos três deles, uma vez que serão alvo de análise no capítulo 5 – *Análise de Resultados*.

Sendo assim, iremos descrever em detalhe os algoritmos *naive Bayes*, SMO e J48, relativos respectivamente a redes bayesianas, *support vector machines* e árvores de decisão.

Abordando desde já o algoritmo **J48** [SB05], este permite a criação de modelos de decisão em árvore. Faz uso de uma estratégia *greedy* para induzir árvores de decisão para posterior classificação. O modelo de árvore de decisão é construído pela análise dos dados de treino, sendo posteriormente utilizado para classificar dados ainda não classificados.

Este algoritmo gera árvores de decisão, em que cada nó da árvore avalia a existência ou significância de cada atributo individual. As árvores de decisão caracterizam-se por serem construídas desde o topo até à base, através da escolha do atributo mais apropriado para cada situação. Uma vez escolhido o atributo, os dados de treino são divididos em subgrupos, sendo que o processo é repetido para cada subgrupo até que uma grande parte dos atributos em cada um desses pequenos grupos pertença a uma única classe.

A indução por árvore de decisão é um algoritmo que habitualmente "aprende" um conjunto de regras com elevada acuidade.

Este algoritmo J48 foi essencialmente utilizado para que a sua taxa de precisão fosse alvo de comparação com outros algoritmos.

Relativamente ao algoritmo ***naive Bayes*** [SB05], trata-se de um dos classificadores probabilísticos mais simples.

O modelo construído por este algoritmo é um conjunto de probabilidades. Essas probabilidades são estimadas pela contagem da frequência dos valores de cada característica para as instâncias dos dados de treino. Dada uma nova instância, o classificador estima a probabilidade dessa mesma instância pertencer a uma classe

específica, baseada no produto das probabilidades condicionais individuais para os valores característicos da instância.

O cálculo exacto utiliza o teorema de *Bayes*, sendo por essa mesma razão que o algoritmo é denominado um classificador de *Bayes*. Este algoritmo é igualmente designado *naive*, uma vez que todos os atributos são independentes, dado o valor da variável de classe. Apesar deste pressuposto, o algoritmo apresenta um bom desempenho em muitos dos cenários de previsão de classes. Estudos experimentais revelam a apetência deste algoritmo para "aprender" mais rapidamente que a maioria dos algoritmos de indução, daí a sua utilização no decorrer das experiências.

Por último, o algoritmo **SMO** (***Sequential Minimal Optimization***) é reconhecido como sendo um dos algoritmos mais rápidos e o de mais fácil implementação em software. O algoritmo é iterativo e adopta uma solução analítica para a optimização de um par de Multiplicadores de *Lagrange* em cada iteração, evitando o armazenamento de matrizes de grandes dimensões em memória. O SMO executa três tarefas por iteração, nomeadamente:

- selecção de um par de coeficientes (t_1);
- optimização do par de coeficientes seleccionado (t_2);
- actualização de dados globais (t_3).

O algoritmo executa um total de I_t iterações até que todos os coeficientes satisfaçam um conjunto de condições denominadas de *Karush-Kuhn-Tucker* (KKT) [Pla99]. O tempo total de execução (T_{SMO}) pode ser calculado segundo a igualdade:

$$T_{SMO} = I_t (t_1 + t_2 + t_3),$$

sendo $t_1 + t_2 + t_3$ o tempo médio de execução de cada iteração.

Em modo de conclusão, torna-se essencial referir que devido à sua importância, designadamente pelo facto de providenciar resultados robustos e generalizados, muitos autores têm efectuado optimizações ao algoritmo SMO de forma a reduzirem o seu tempo de execução, uma das principais limitações deste classificador [Her09].

Capítulo 3

Estado da Arte

Este capítulo faz um levantamento do estado da arte relacionada com o conceito de cancro de mama e respectivos estudos ao longo dos últimos anos. Como tal, são inicialmente apresentados alguns dados estatísticos relativos à incidência deste flagelo na sociedade actual. São ainda discutidos os benefícios e respectivos números inerentes aos programas de rastreio introduzidos em meados dos anos 90, com especial destaque para um dos exames mais utilizados neste tipo de prevenção – a mamografia. Por último, é efectuada uma apresentação dos vários trabalhos que têm sido desenvolvidos com vista à resolução de problemas relacionados com o cancro de mama, nomeadamente são descritos alguns estudos em que são aplicados métodos de aprendizagem automática aos dados.

3.1 Cancro de Mama

O organismo humano é constituído por triliões de células que se reproduzem pelo processo de divisão celular. Em condições normais, este é um processo ordenado e controlado, responsável pela formação, crescimento e regeneração de tecidos saudáveis do corpo. Algumas vezes, no entanto, as células perdem a capacidade de limitar e

comandar o seu próprio crescimento, passando então a dividir-se e multiplicar-se muito rapidamente e de maneira aleatória. Como consequência dessa disfunção celular, isto é, desse processo de multiplicação e crescimento desordenado das células, ocorre um desequilíbrio na formação dos tecidos do corpo, levando ao desenvolvimento de nódulos, mais conhecidos como tumores.

O cancro de mama é a nível mundial o tumor maligno mais comum entre as mulheres (excluindo o cancro de pele), correspondendo à segunda causa de morte por cancro no sexo feminino. Nos Estados Unidos da América, uma em cada oito mulheres desenvolve cancro de mama no decorrer da sua vida. Em 2006, data dos últimos dados disponíveis, 191410 mulheres foram diagnosticadas com cancro de mama, sendo que 40820 (≈ 21%) acabaram por não resistir à doença [BSC10].

Em Portugal, por ano são detectados aproximadamente 4500 novos casos e cerca de 1500 óbitos (≈ 33%) [Lig10], sendo a principal causa de morte por neoplasia[14] no sexo feminino. A partir de 1995 começou a verificar-se uma ligeira tendência para a diminuição da mortalidade devido à introdução de **programas de rastreio**[15]. Dos rastreios efectuados até aos dias de hoje[16] concluiu-se que houve uma redução da taxa de mortalidade por cancro de mama na ordem dos 30%, quando comparados com o grupo de controlo após cinco anos. Nenhuma outra medida diagnóstica ou terapêutica permitiu uma redução tão acentuada da taxa de mortalidade.

Nos rastreios que obtiveram os melhores resultados existiu uma influência inequívoca da qualidade da **mamografia** e da experiência dos radiologistas.

A *American Cancer Society* recomenda a realização de um exame clínico e de uma mamografia de dois em dois anos em pessoas que se encontrem na faixa etária dos 40 aos 49. A partir dos 50 anos, segundo a mesma instituição, estes exames deverão ser

[14] Termo que designa alterações celulares que acarretam um crescimento exagerado das células. Proliferação celular anormal e autónoma.

[15] Realização periódica de exames num grupo populacional assintomático com o principal objectivo de detectar o cancro de mama num estado precoce. A mamografia é o exame imagiológico mais utilizado neste tipo de prevenção.

[16] O estudo mais antigo data de 1963 a 1970. Foi realizado nos EUA e rastreou 31000 mulheres entre os 40 e os 64 anos.

anuais. A mamografia anual e o exame clínico podem detectar simultaneamente cerca de 80% de cancros de mama [Orv08].

A mamografia, devido ao seu nível de precisão, permite a um determinado médico conhecer o tamanho, localização e características de nódulos com apenas alguns milímetros, nomeadamente nos casos em que estes ainda não podem ser sentidos por palpação. O sistema **BI-RADS®** (*Breast Imaging Reporting and Data System*) publicado pelo *American College of Radiology* (ACR) em 1993[17], visa padronizar os relatórios médicos relativos a este tipo de exame, reduzindo desta forma as discordâncias existentes na interpretação de mamografias. Este sistema disponibiliza uma série de descritores para os "achados" observados, assim como define categorias que resumem as principais conclusões da parte do radiologista que avalia o exame médico.

O léxico BI-RADS® é composto por 43 descritores organizados numa hierarquia. A Figura 16 ilustra parte desses descritores. De notar que o universo de dados alvo de estudo nesta dissertação poderá não conter todos os descritores, além de que a designação dos mesmos poderá ser diferente.

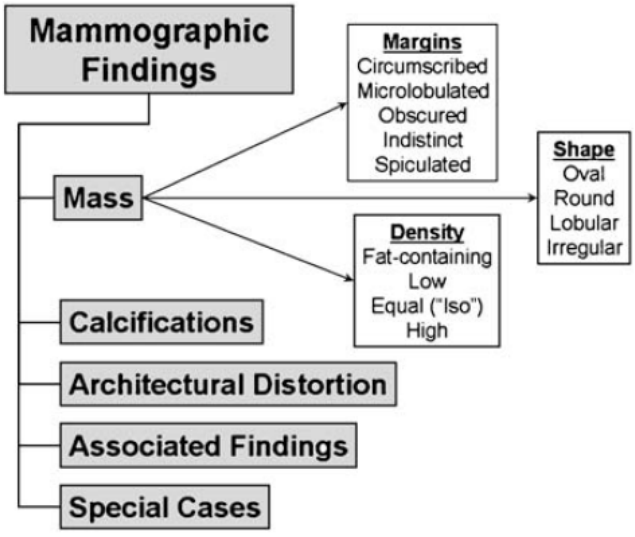


Figura 16 - Descritores BI-RADS® (obtido de [WOS+09])

[17] Novas edições do sistema BI-RADS® foram publicadas em 1995, 1998 e 2003.

Quanto às categorias (Tabela 2), dividem-se em seis tipos que sintetizam a opinião do radiologista para o estudo em questão. Os exames são classificados com base no grau de suspeita das lesões:

Categoria	Interpretação
BI-RADS 0	Lesões que necessitam de informação adicional.
BI-RADS 1	"Achados" negativos: mamografia normal.
BI-RADS 2	"Achados" benignos.
BI-RADS 3	"Achados" provavelmente benignos.
BI-RADS 4	"Achados" suspeitos de malignidade.
BI-RADS 5	"Achados" altamente suspeitos de malignidade.

Tabela 2 - Categorias BI-RADS®

Nos casos em que um determinado médico desconfie que um nódulo seja de origem maligna, poderá sempre efectuar uma **biópsia**. A biópsia trata-se de um procedimento (que poderá ser cirúrgico ou não) em que é recolhida uma amostra do nódulo "suspeito". O tecido retirado é posteriormente analisado por um patologista com o objectivo de confirmar se a origem do nódulo em causa é de natureza maligna. A Figura 17 ilustra precisamente uma situação em que será necessária a realização de uma biópsia a fim de se determinar qual a origem do nódulo presente na mamografia da direita.

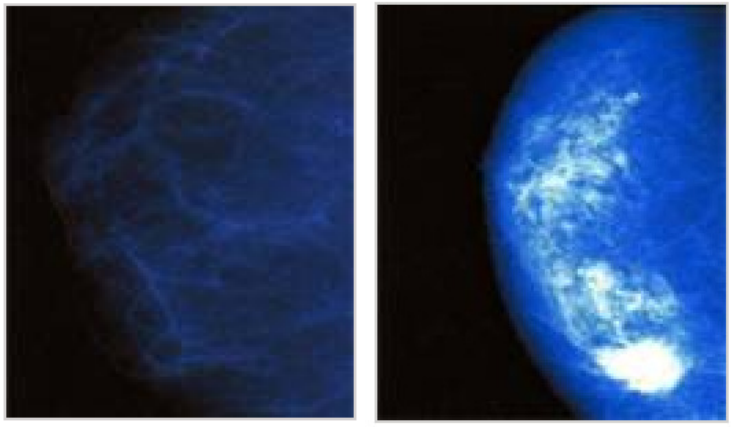


Figura 17 - Imagens referentes a duas mamografias distintas. A mamografia da esquerda apresenta uma mama normal, em que as áreas mais densas (brancas) são os canais galactóforos. A mamografia da direita, por sua vez, apresenta uma área branca densa (canto inferior direito da imagem) que indica a presença de um tumor

É relevante mencionar que desde o início dos rastreios até aos dias de hoje, a percentagem de carcinoma ductal *in situ* (a fase mais precoce do cancro de mama) aumentou de 5% para valores que se situam entre os 20 e 30% de todos os cancros detectados [Orv08].

Torna-se, portanto, fundamental conhecer alguns dos **termos** utilizados **para descrever os tumores mamários**, uma vez que o tratamento e prognóstico variam de doente para doente e em função do tipo de cancro.

Sendo assim, quase todos os tumores malignos da mama têm origem em dois tecidos glandulares: nos ductos ou nos lóbulos da mama, representados na Figura 18.

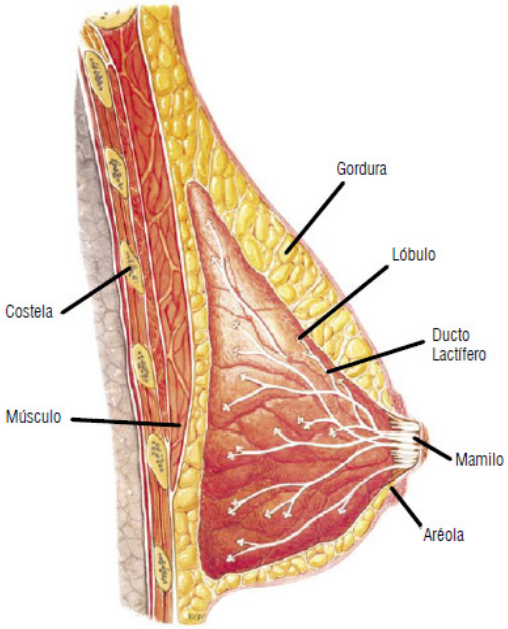


Figura 18 - Anatomia de uma mama saudável

Os tipos mais frequentes de cancro de mama são o carcinoma ductal e o carcinoma lobular. Em seguida passamos a definir os diferentes tipos de tumores mamários e respectivos termos científicos [Por05]:

- ***In situ*:** Este termo define o cancro de mama precoce que se encontra limitado aos ductos (carcinoma ductal *in situ*) ou lóbulos (carcinoma lobular *in situ*), sem invasão dos tecidos mamários vizinhos e de outros órgãos.
- **Carcinoma ductal *in situ* (CDIS):** Trata-se do cancro de mama não invasivo mais frequente. Praticamente todas as mulheres com CDIS têm hipóteses de cura. A mamografia é o melhor método para diagnosticar o cancro de mama nesta fase precoce.
- **Carcinoma lobular *in situ* (CLIS):** Embora não seja verdadeiramente cancro, o CLIS é habitualmente classificado como um cancro de mama não invasivo. Diversos especialistas defendem que o CLIS não se transforma num carcinoma invasor. No entanto, as mulheres com esta neoplasia apresentam um risco maior de desenvolver cancro de mama invasor.

- **Carcinoma ductal invasor (CDI):** Trata-se do cancro de mama invasor mais frequente. Tem origem nos ductos e invade os tecidos vizinhos. Nesta fase pode disseminar-se através dos vasos linfáticos ou do sangue, acabando por atingir outros órgãos. Cerca de 80% dos cancros de mama invasores são carcinomas ductais.

- **Carcinoma lobular invasor (CLI):** Tem origem nas unidades produtoras de leite, ou seja, nos lóbulos. À semelhança do CDI pode disseminar-se (metastizar) para outras partes do corpo. Cerca de 10% dos cancros de mama invasores são carcinomas lobulares.

- **Carcinoma inflamatório da mama:** Trata-se de um cancro extremamente agressivo, mas pouco frequente. Corresponde a cerca de 1 a 3% de todos os cancros de mama.

Existem ainda outros tipos de cancros de mama mais raros, tal como o Carcinoma Medular, o Carcinoma Mucinoso, o Carcinoma Tubular, o Tumor Filóide Maligno, entre outros.

Como nota de conclusão, e de certo modo fazendo uso desta dissertação como forma de consciencializar as pessoas, aproveitamos para sublinhar que o diagnóstico precoce do cancro de mama é absolutamente fundamental para um aumento das hipóteses de cura, sendo a mamografia, o método mais económico e eficiente na detecção prematura deste tipo de tumor.

3.2 Aprendizagem Automática para detecção de Cancro de Mama

Vários trabalhos têm sido desenvolvidos na aplicação de métodos de aprendizagem automática para o estudo do cancro de mama. Na Universidade da Califórnia, em Irvine (UCI), existe um repositório[18] para aprendizagem automática que alberga quatro conjuntos de dados cujo objectivo principal de estudo é o cancro de mama.

Um dos primeiros trabalhos na aplicação de técnicas de aprendizagem automática a dados relativos a cancros de mama, data do início da década de 90. Por esta altura, o primeiro conjunto de dados doado ao repositório da UCI foi criado por Wolberg e Mangasarian após o desenvolvimento de um método multi-superfície de separação de padrões para diagnósticos médicos aplicados à citologia[19] da mama [WM90].

A maioria dos trabalhos presentes na literatura aplica redes neuronais artificiais a dados de mamografias como forma de diagnosticar o cancro de mama [WGD+93, Abb02]. Outros trabalhos focam-se, por sua vez, no prognóstico da doença, recorrendo a métodos de aprendizagem indutiva [SMW95]. Mais recentemente, Ayer *et al*. [AAC+10] avaliaram o modo como uma rede neuronal artificial, treinada num extenso conjunto de dados provenientes de mamografias e recolhidos prospectivamente, poderia diferenciar os dados entre benignos e malignos e, além do mais, conseguir prever com precisão a probabilidade de ocorrência de cancro de mama para casos particulares (pacientes individuais). Outras pesquisas, também relativamente recentes, recaem na extracção de informação a partir dos próprios textos de relatórios médicos de mamografias [NWB+09] e até na influência da idade no cancro de mama não invasivo mais frequente – o carcinoma ductal *in situ* [NPA+10].

O nosso estudo incide, essencialmente, na influência da densidade de massa dos nódulos na previsão de malignidade, no entanto, também abordamos outras questões potencialmente interessantes.

---

[18] Acessível online em: http://archive.ics.uci.edu/ml/datasets.html.

[19] Estudo científico de células.

Trabalhos anteriores, nomeadamente de Jackson *et al*. [JDB+91] e de Cory e Linden [CL93] concluíram que, apesar da maior parte dos nódulos que apresentam densidade elevada serem malignos, a presença de tumores com densidades de massa baixas, assim como uma série de outros indicadores importantes (como a forma das margens do nódulo, por exemplo) fazem da densidade de massa um indicador pouco fiável de malignidade. Em 1991, Sickles [Sic91] publicou um estudo em que refere o mesmo. No entanto, uma investigação levada a cabo por Davis *et al*. [DBD+05] em 2005 revela que a densidade de massa dos nódulos tem efectivamente uma maior importância do que alguns estudos anteriores sugeriram. Num outro trabalho, já em 2009, Woods *et al*. [WOS+09] aplicaram programação lógica indutiva (PLI) a um conjunto de dados referentes a cancro de mama, chegando às mesmas conclusões. Woods e Burnside [WB10], por sua vez, aplicaram regressão logística e estatística *Kappa* a um outro conjunto de dados, concluindo que quer a densidade de massa dos nódulos como a malignidade estão de certa forma relacionados.

Nesta dissertação, fazemos uso do mesmo universo de dados utilizado por Woods e Burnside [WB10], no entanto aplicamos métodos de aprendizagem automática aos dados. Mesmo utilizando uma metodologia diferente, confirmamos que densidade de massa e malignidade estão de facto relacionados. Além do mais, demonstramos que os classificadores gerados neste trabalho são capazes de prever densidade de massa e malignidade com um nível qualitativo semelhante à previsão efectuada por um especialista, assumindo-se como óptimas plataformas de apoio a médicos e radiologistas.

Capítulo 4

Experiências

Este capítulo introduz, inicialmente, os dados fornecidos para a execução das experiências em que são aplicados métodos de aprendizagem automática em tarefas de classificação. Em seguida, é descrita a forma como os atributos relativos a esses mesmos dados foram seleccionados. Por último, é efectuada uma explicação do modo como foi aplicada a aprendizagem *10-fold cross-validation* ao longo das diferentes experiências, assim como a forma de aplicação dos modelos gerados a conjuntos de dados desconhecidos.

4.1 Dados

Os dados utilizados nesta dissertação foram facultados pela Dra. Elizabeth Burnside[20] e pelo Dr. Ryan Woods[21], na altura membros do Departamento de Radiologia da Universidade de Wisconsin nos EUA. Os dados são compostos por 348 casos relativos

[20] *Department of Radiology, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA. EBurnside@uwhealth.org.*

[21] *Department of Radiology, Johns Hopkins Hospital, Baltimore, MD, USA. ryan_woods@alumni.bowdoin.edu.*

a pacientes que foram sujeitos ao rastreio de cancro de mama através de exames imagiológicos, nomeadamente mamografias [WB10]. Tratam-se de dados recolhidos entre Outubro de 2005 e Dezembro de 2007 em 328 mulheres. Cada caso corresponde a um nódulo da mama e foi classificado **retrospectivamente** por um grupo de radiologistas de acordo com o sistema BI-RADS®. Dos 348 nódulos, 180 (≈ 52%) foram classificados **prospectivamente** em termos de densidade por um único radiologista, fazendo uso do mesmo sistema BI-RADS®. Os restantes nódulos, ou seja, 168 casos (≈ 48%) do universo de 348, não foram alvo de classificação prospectiva.

Relativamente aos termos retrospectivo e prospectivo, é extremamente importante clarificar o conceito por detrás desta terminologia, uma vez que será utilizada ao longo deste capítulo e do próximo. Estes termos foram introduzidos pelos próprios médicos norte-americanos e designam o seguinte:

- **Classificação Prospectiva (*Prospective Classification*)**

  A classificação do atributo densidade de massa (*mass density*) relativa aos 180 casos é proveniente de uma espécie de relatório médico breve e superficial elaborado por apenas um radiologista sem qualquer informação relativa a biópsias. Trata-se de uma **classificação** efectuada **sob stress** no momento da mamografia. Uma vez que algumas imagens radiológicas revelavam mais do que um "achado", foi fornecida ao radiologista, nesta sua avaliação prospectiva, a localização dos nódulos, nomeadamente o lado onde se encontravam (*breast laterality*), a posição (*clock face position*) e profundidade (*depth*). O mesmo radiologista avaliou a densidade destes nódulos tendo como base de comparação a densidade de um pedaço semelhante de tecido fibroglandular, associando a cada nódulo um dos seguintes descritores BI-RADS®: "densidade baixa" (*low density*), "densidade média" (*iso-dense*) e "densidade alta" (*high density*).

- **Classificação Retrospectiva (*Retrospective Classification*)**

A classificação retrospectiva é obtida numa espécie de reunião periódica entre radiologistas e médicos experientes em que estes reavaliam uma série de exames, sendo também revista a classificação de densidade de massa (*mass density*) efectuada pelo radiologista aquando da classificação prospectiva. A classificação retrospectiva de densidade pode ser diferente da classificação prospectiva. Trata-se de uma **classificação** obtida **sem stress** e por um grupo de médicos e radiologistas experientes, sendo por isso usada nesta dissertação como valores de referência para o atributo densidade de massa (*mass density*).

Dos 348 nódulos, 118 são malignos (≈ 34%) (Figura 19), além de que 84 destes 348 casos apresentam densidade elevada (≈ 24%), retrospectivamente anotados.

Benign Diagnosis	N
Fibroadenoma	93 (40.4)
Fibrocystic change	46 (20)
Other benign mass	91 (39.6)
Total	230

Malignant Diagnosis	N
Invasive ductal carcinoma	90 (76.3)
Invasive lobular carcinoma	14 (11.9)
Ductal carcinoma in situ	11 (9.3)
Metastatic carcinoma	2 (1.7)
Lymphoma	1 (0.8)
Total	118

Figura 19 - Distribuição dos 348 nódulos em termos de malignidade: 230 benignos (≈ 66%) e 118 malignos (≈ 34%). Na figura, os números entre parêntesis representam percentagens referentes aos diferentes tipos de malignidade (obtido de [WB10])

Torna-se importante referir as conclusões médicas a que os investigadores, que nos forneceram estes dados, chegaram.

Os objectivos principais da equipa norte-americana consistiam em determinar se a densidade de massas sólidas não calcificadas da mama seria um indicador de

malignidade, assim como medir a concordância entre observadores (estatística *Kappa*) [WB10].

Para tal, cingiram-se a associar a densidade dos nódulos ao número de tumores malignos encontrados (Figura 20). Concluíram que no estudo retrospectivo, 70.2% dos nódulos que apresentavam densidade elevada (*high density*) eram malignos, e que 22.3% com densidade média/baixa (*iso-dense*) também eram de natureza cancerígena [WB10].

	Predictor	Total	Benign	Malignant
Mass Density	High	84	25 (29.8)	59 (70.2)
	Low/Iso	264	205 (77.7)	59 (22.3)

Figura 20 - Conclusões obtidas pelos investigadores norte-americanos no que respeita à relação entre densidade e malignidade no estudo retrospectivo. Na figura, os números entre parêntesis representam percentagens (obtido de [WB10])

No modelo prospectivo, por sua vez, a densidade elevada (*high density*) dos nódulos, a forma irregular (*irregular shape*), a margem espiculada (*spiculated margin*) e a idade (*age*) previram significativamente a probabilidade de malignidade. O valor da concordância entre observadores para a densidade de massa foi de k = 0.53 [WB10].

Sendo assim, estes investigadores defendem que a densidade elevada dos nódulos (*high density*) é um indicador significativo de malignidade, quer no estudo retrospectivo como no estudo prospectivo. Além do mais, chamam a atenção para o valor moderado de estatística *Kappa*. Ressalvam portanto que os radiologistas deverão considerar a densidade de massa dos nódulos como um descritor fundamental que poderá ajudar a quantificar o risco de malignidade [WB10].

Nesta dissertação, o principal objectivo é "aprender" modelos que auxiliem os médicos na análise de mamografias. Para tal, recorremos a técnicas de aprendizagem automática que permitirão corroborar (ou não) as conclusões dos médicos norte-americanos.

De seguida, passamos a descrever o modo como os atributos relativos aos dados fornecidos foram seleccionados.

4.1.1 Atributos

O conjunto de dados original referente ao estudo retrospectivo é composto por 35 atributos, enquanto o estudo prospectivo é composto por 33.

A Tabela 3 apresenta todos os atributos e respectivas descrições.

Atributos	Descrição
MRN_scrubbed	*Medical Record Number*. Identificador de registo.
PATIENT_SEX	Sexo dos pacientes.
rnd_num	Número aleatório que identifica uma mamografia.
reread_group	Radiologista que classificou retrospectivamente a densidade de massa dos nódulos.
biopsy_date	Data em que uma determinada biópsia foi efectuada.
ID_MATCH_NMD	Número que identifica um determinado exame na *National Mammography Database* (NMD).
ASSESSMENT	Define a categoria BI-RADS® em que um determinado nódulo se insere.
PENRAD_MAMMO_ID	Número que identifica um determinado exame no sistema PenRad®[22].
MAMMO_STUDY_DATE	Data em que uma determinada mamografia foi efectuada.
age_at_mammo	Idade dos pacientes aquando da realização da mamografia.
PENRAD_ABNORMALITY_ID	Número identificador relativo ao sistema PenRad®.
CLOCKFACE_LOCATION_OR_REGION	Localização dos nódulos.
MASS_SHAPE	Forma dos nódulos.
MASS_MARGINS	Classificação relativa às margens dos nódulos.

[22] Sistema automático de análise de mamografias habitualmente utilizado por médicos e radiologistas.

SIDE	Mama onde os nódulos foram encontrados.
DEPTH	Profundidade dos nódulos (em mm), medida desde a superfície da pele até ao centro da lesão.
MASS_SHAPE_def	Forma dos nódulos.
MASS_MARGINS_def	Classificação relativa às margens dos nódulos.
MASS_MARGINS_worst	Nos casos em que o atributo *MASS_MARGINS* apresenta duas características em simultâneo, este atributo *MASS_MARGINS_worst* identifica a mais preocupante dessas duas características.
ARCHITECTURAL_DISTORTION_def	Define se existe distorção de um determinado nódulo.
CLOCKFACE_def	Localização dos nódulos.
QUADRANT_LOCATION_def	Quadrante onde se localizam os nódulos.
SIDE_def	Mama onde os nódulos foram encontrados.
DEPTH_def	Profundidade dos nódulos (em mm), medida desde a superfície da pele até ao centro da lesão.
SIZE	Largura máxima transversal dos nódulos (em mm).
OVERALL_BREAST_COMPOSITION	Tipo de densidade dos nódulos.
Density_num	Densidade de massa dos nódulos prospectivamente anotada.
retro_density	Densidade de massa dos nódulos retrospectivamente anotada.
outcome_num	Classificação dos nódulos em termos de malignidade baseada em resultados de biópsias.
lb_finding	Diagnóstico patológico de um determinado nódulo.
digital_sub	Sem descrição.
digital	Tipo de estudo mamográfico.
lb_technique	Tipo de biópsia efectuada a um determinado nódulo.
REASON_FOR_THIS_MAMMOGRAM	Sem descrição.
FUmonths	*Follow-Up months.* Número de meses em que um determinado paciente foi alvo de acompanhamento médico.

Tabela 3 - Conjunto de atributos relativos aos dados originais com respectiva descrição

Do conjunto de atributos presentes na Tabela 3, seleccionamos todos aqueles que consideramos relevantes para o nosso estudo (Tabela 4).

Atributos Utilizados
reread_group
age_at_mammo
CLOCKFACE_LOCATION_OR_REGION
MASS_SHAPE
MASS_MARGINS
SIDE
DEPTH
MASS_MARGINS_worst
QUADRANT_LOCATION_def
SIZE
OVERALL_BREAST_COMPOSITION
Density_num
retro_density
outcome_num

Tabela 4 - Conjunto de atributos utilizados para o estudo em questão

Por outro lado, certos atributos como identificadores, atributos redundantes ou mesmo atributos que apresentavam o mesmo valor para todas as instâncias, foram removidos[23].

Torna-se importante referir que o atributo *mass_margins*, visto tratar-se de um atributo que para alguns casos apresentava duas características, foi desdobrado em dois sub-atributos (*mass_margins_1*, *mass_margins_2*) para que não se perdesse informação. A Figura 21 representa precisamente essa manipulação:

[23] Ver *Apêndice B* – Tabela 15 com atributos descartados e com respectivo motivo pelo qual não foram utilizados.

M	N	O	P
MASS_SHAPE	MASS_MARGINS	SIDE	DEPTH
X	S	R	M
O	D	L	P
R		R	A
R	U,S	R	M
L	S	R	
R	D	R	M
L	D,I	L	M
O	D	R	

E	F	G	H	I
MASS_SHAPE	MASS_MARGINS_1	MASS_MARGINS_2	SIDE	DEPTH
X	S	S	R	M
O	D	D	L	P
R			R	A
R	U	S	R	M
L	S	S	R	
R	D	D	R	M
L	D	I	L	M
O	D	D	R	

Figura 21 - Atributo *MASS_MARGINS* desdobrado em dois sub-atributos.

Nota: U,S – *Obscured & Spiculated*

De acordo com um dos objectivos principais desta dissertação – previsão de malignidade – o nosso *atributo classe* designa-se por *outcome_num* e assume os valores "maligno" (*malignant*) e "benigno" (*benign*), tendo sido determinado após análise dos resultados de biópsias. Inicialmente, aquando da facultação dos dados, este atributo apresentava três classes distintas, nomeadamente: "maligno" (*malignant*), "benigno" (*benign*) e "benigno, porém com elevado risco de se tornar maligno" (*high risk benign*). No entanto, devido a um escasso número de instâncias do tipo *high risk benign*, estas acabaram por ser incluídas na classe *benign*.

Tal como ilustrado na Tabela 4, os restantes atributos distribuem-se entre formas do nódulo (*mass shape*), margens do nódulo (*mass margins*), profundidade (*depth*), tamanho (*size*), entre outros.

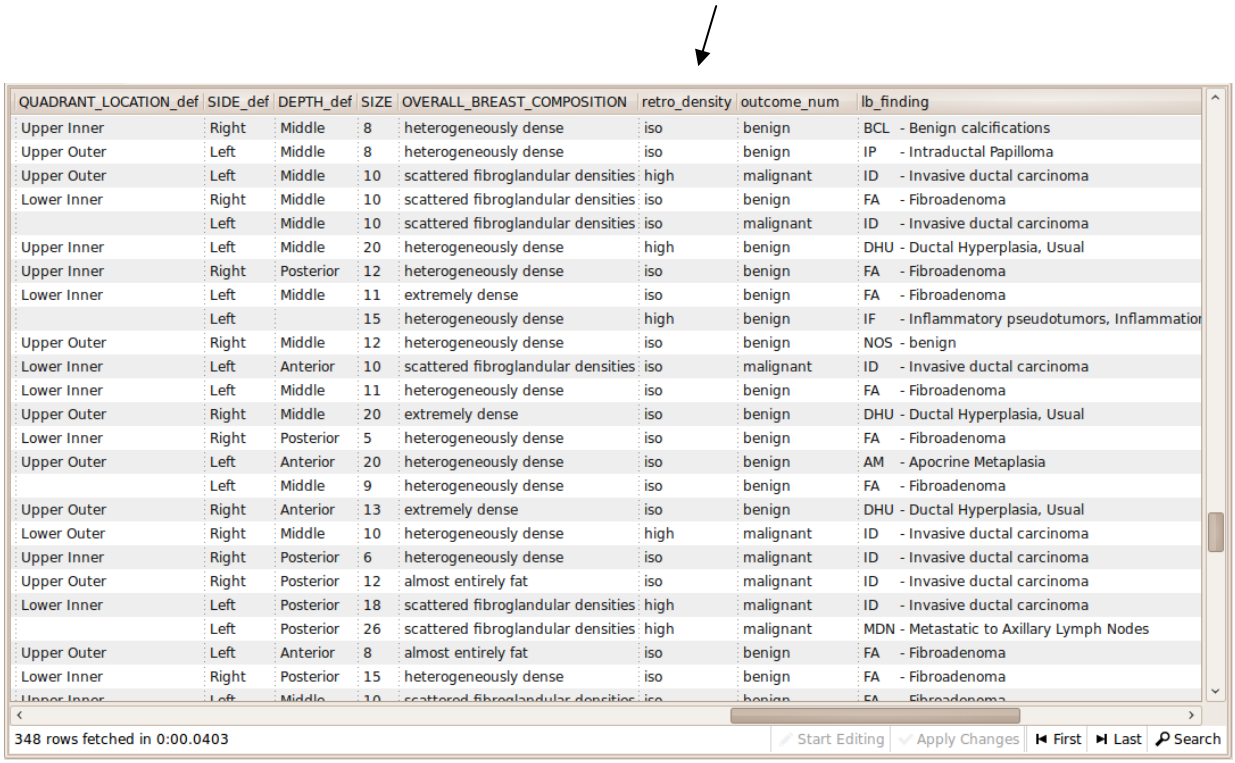
No nosso estudo temos dois atributos que representam as mesmas características para um mesmo "achado", mas com diferentes interpretações. Referimo-nos aos termos *retro_density* e *density_num*. O atributo ***retro_density*** foi **retrospectivamente anotado**,

enquanto o termo ***density_num*** foi **prospectivamente classificado**. Ambos representam densidades de massa que podem assumir os valores "alto" (*high*) ou "médio/baixo" (*iso/low*). Quer nos dados retrospectivos como nos dados prospectivos (Figura 22), a quantidade de instâncias do tipo *low* é bastante baixa para justificar serem colocadas numa classe separada, portanto a classe *low* foi agregada à classe *iso*, por recomendação dos próprios médicos/especialistas.

Predictor		Total
Mass Density		
	Low	7 (2)
	Iso	92 (26.4)
	High	81 (23.3)
	Not Reported	168 (48.3)

Figura 22 - Distribuição original dos dados em termos de densidade de massa no estudo prospectivo. De notar o número bastante baixo de instâncias do tipo *low*, sendo posteriormente associadas à classe *iso*. Na figura, os números entre parêntesis representam percentagens sobre o número total de casos (348) (obtido de [WB10])

A Figura 23 ilustra parte dos dados do modelo retrospectivo. Através de uma base de dados MySQL, criada na altura da recepção das instâncias, foi possível efectuar uma série de consultas que nos auxiliaram na compreensão do universo de dados. Nesta figura é possível visualizar o atributo *retro_density* cujos valores foram atribuídos por um grupo de radiologistas e médicos experientes.

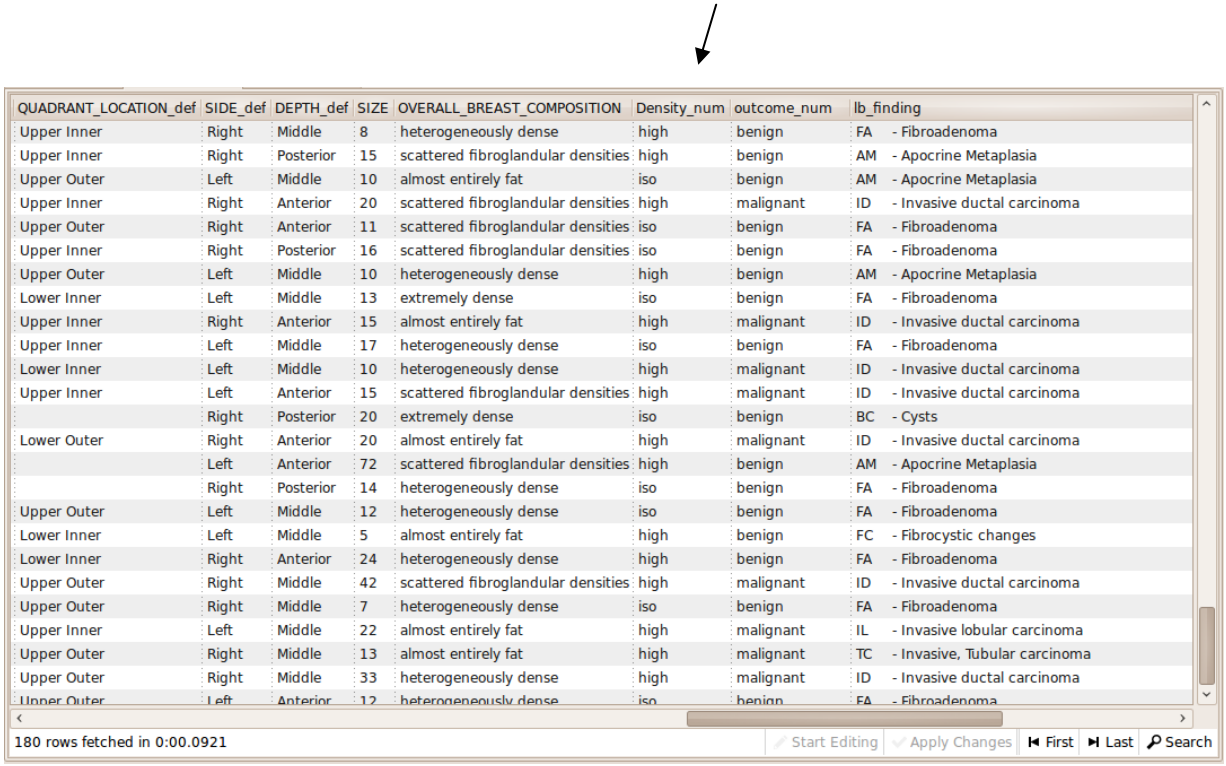


QUADRANT_LOCATION_def	SIDE_def	DEPTH_def	SIZE	OVERALL_BREAST_COMPOSITION	retro_density	outcome_num	lb_finding
Upper Inner	Right	Middle	8	heterogeneously dense	iso	benign	BCL - Benign calcifications
Upper Outer	Left	Middle	8	heterogeneously dense	iso	benign	IP - Intraductal Papilloma
Upper Outer	Left	Middle	10	scattered fibroglandular densities	high	malignant	ID - Invasive ductal carcinoma
Lower Inner	Right	Middle	10	scattered fibroglandular densities	iso	benign	FA - Fibroadenoma
	Left	Middle	10	scattered fibroglandular densities	iso	malignant	ID - Invasive ductal carcinoma
Upper Inner	Left	Middle	20	heterogeneously dense	high	benign	DHU - Ductal Hyperplasia, Usual
Upper Inner	Right	Posterior	12	heterogeneously dense	iso	benign	FA - Fibroadenoma
Lower Inner	Left	Middle	11	extremely dense	iso	benign	FA - Fibroadenoma
	Left		15	heterogeneously dense	high	benign	IF - Inflammatory pseudotumors, Inflammatior
Upper Outer	Right	Middle	12	heterogeneously dense	iso	benign	NOS - benign
Lower Inner	Left	Anterior	10	scattered fibroglandular densities	iso	malignant	ID - Invasive ductal carcinoma
Lower Inner	Left	Middle	11	heterogeneously dense	iso	benign	FA - Fibroadenoma
Upper Outer	Right	Middle	20	extremely dense	iso	benign	DHU - Ductal Hyperplasia, Usual
Lower Inner	Right	Posterior	5	heterogeneously dense	iso	benign	FA - Fibroadenoma
Upper Outer	Left	Anterior	20	heterogeneously dense	iso	benign	AM - Apocrine Metaplasia
	Left	Middle	9	heterogeneously dense	iso	benign	FA - Fibroadenoma
Upper Outer	Right	Anterior	13	extremely dense	iso	benign	DHU - Ductal Hyperplasia, Usual
Lower Outer	Right	Middle	10	heterogeneously dense	high	malignant	ID - Invasive ductal carcinoma
Upper Inner	Right	Posterior	6	heterogeneously dense	iso	malignant	ID - Invasive ductal carcinoma
Upper Outer	Right	Posterior	12	almost entirely fat	iso	malignant	ID - Invasive ductal carcinoma
Lower Inner	Left	Posterior	18	scattered fibroglandular densities	high	malignant	ID - Invasive ductal carcinoma
	Left	Posterior	26	scattered fibroglandular densities	high	malignant	MDN - Metastatic to Axillary Lymph Nodes
Upper Outer	Left	Anterior	8	almost entirely fat	iso	benign	FA - Fibroadenoma
Lower Inner	Right	Posterior	15	heterogeneously dense	iso	benign	FA - Fibroadenoma

348 rows fetched in 0:00.0403 Start Editing Apply Changes First Last Search

Figura 23 - Base de Dados MySQL. Representação de parte dos dados do modelo retrospectivo (destaque para o atributo *retro_density*)

A Figura 24, por sua vez, representa parte do conjunto prospectivo, com especial destaque para o atributo *density_num*, o qual foi preenchido por apenas um radiologista sem qualquer informação relativa a biópsias.



QUADRANT_LOCATION_def	SIDE_def	DEPTH_def	SIZE	OVERALL_BREAST_COMPOSITION	Density_num	outcome_num	lb_finding
Upper Inner	Right	Middle	8	heterogeneously dense	high	benign	FA - Fibroadenoma
Upper Inner	Right	Posterior	15	scattered fibroglandular densities	high	benign	AM - Apocrine Metaplasia
Upper Outer	Left	Middle	10	almost entirely fat	iso	benign	AM - Apocrine Metaplasia
Upper Inner	Right	Anterior	20	scattered fibroglandular densities	high	malignant	ID - Invasive ductal carcinoma
Upper Outer	Right	Anterior	11	scattered fibroglandular densities	iso	benign	FA - Fibroadenoma
Upper Inner	Right	Posterior	16	scattered fibroglandular densities	iso	benign	FA - Fibroadenoma
Upper Outer	Left	Middle	10	heterogeneously dense	high	benign	AM - Apocrine Metaplasia
Lower Inner	Left	Middle	13	extremely dense	iso	benign	FA - Fibroadenoma
Upper Inner	Right	Anterior	15	almost entirely fat	high	malignant	ID - Invasive ductal carcinoma
Upper Inner	Left	Middle	17	heterogeneously dense	iso	benign	FA - Fibroadenoma
Lower Inner	Left	Middle	10	heterogeneously dense	high	malignant	ID - Invasive ductal carcinoma
Upper Inner	Left	Anterior	15	scattered fibroglandular densities	high	malignant	ID - Invasive ductal carcinoma
	Right	Posterior	20	extremely dense	iso	benign	BC - Cysts
Lower Outer	Right	Anterior	20	almost entirely fat	high	malignant	ID - Invasive ductal carcinoma
	Left	Anterior	72	scattered fibroglandular densities	high	benign	AM - Apocrine Metaplasia
	Right	Posterior	14	heterogeneously dense	iso	benign	FA - Fibroadenoma
Upper Outer	Left	Middle	12	heterogeneously dense	iso	benign	FA - Fibroadenoma
Lower Inner	Left	Middle	5	almost entirely fat	high	benign	FC - Fibrocystic changes
Lower Inner	Right	Anterior	24	heterogeneously dense	high	benign	FA - Fibroadenoma
Upper Outer	Right	Middle	42	scattered fibroglandular densities	high	malignant	ID - Invasive ductal carcinoma
Upper Outer	Right	Middle	7	heterogeneously dense	iso	benign	FA - Fibroadenoma
Upper Inner	Left	Middle	22	almost entirely fat	high	malignant	IL - Invasive lobular carcinoma
Upper Outer	Right	Middle	13	almost entirely fat	high	malignant	TC - Invasive, Tubular carcinoma
Upper Outer	Right	Middle	33	heterogeneously dense	high	malignant	ID - Invasive ductal carcinoma
Upper Outer	Left	Anterior	12	heterogeneously dense	iso	benign	FA - Fibroadenoma

180 rows fetched in 0:00.0921 Start Editing Apply Changes First Last Search

Figura 24 - Base de Dados MySQL. Representação de parte dos dados do modelo prospectivo (destaque para o atributo *Density_num*)

Por último, nas Tabelas 5, 6, 7 e 8 apresentamos o modo como os dados estão distribuídos, de acordo com a malignidade (*outcome_num*) e a densidade de massa (retrospectiva e prospectiva) dos nódulos.

Sendo assim, a Tabela 5 mostra o panorama geral para os 348 casos. As Tabelas 6 e 7, por sua vez, exibem a distribuição para os 180 casos prospectivos. Finalmente, a Tabela 8 apresenta a distribuição para os 168 casos que não foram alvo da classificação do radiologista (estudo prospectivo).

348	retro_density		Total
outcome_num	high	iso	
malignant	59 (70.2%)	59 (22.3%)	**118 (33.9%)**
benign	25 (29.8%)	205 (77.7%)	**230 (66.1%)**
Total	**84 (24.1%)**	**264 (75.9%)**	

Tabela 5 - Distribuição dos 348 casos em termos de densidade retrospectivamente anotada e malignidade

180	retro_density		Total
outcome_num	high	iso	
malignant	42 (75.0%)	29 (23.4%)	**71 (39.4%)**
benign	14 (25.0%)	95 (76.6%)	**109 (60.6%)**
Total	**56 (31.1%)**	**124 (68.9%)**	

Tabela 6 - Distribuição dos 180 casos em termos de densidade retrospectivamente anotada e malignidade

180	density_num		Total
outcome_num	high	iso	
malignant	51 (63.0%)	20 (20.2%)	**71 (39.4%)**
benign	30 (37.0%)	79 (79.8%)	**109 (60.6%)**
Total	**81 (45.0%)**	**99 (55.0%)**	

Tabela 7 - Distribuição dos 180 casos em termos de densidade prospectivamente anotada e malignidade

168	retro_density		Total
outcome_num	high	iso	
malignant	17 (60.7%)	30 (21.4%)	**47 (28.0%)**
benign	11 (39.3%)	110 (78.6%)	**121 (72.0%)**
Total	**28 (16.7%)**	**140 (83.3%)**	

Tabela 8 - Distribuição dos 168 casos em termos de densidade retrospectivamente anotada e malignidade

De seguida, passamos a descrever a metodologia utilizada na execução das experiências.

4.2 Métodos

O nosso estudo preliminar consistia em calcular simples frequências a partir dos dados, assim como determinar se existiria algum tipo de relação entre atributos.

Tal como acima mencionado, dos 348 nódulos, 118 são malignos (≈ 34%), além de que 84 desses 348 casos apresentam densidade elevada (≈ 24%).

Tomemos em consideração a hipótese de densidade de massa e malignidade serem variáveis independentes. Pegando 84 casos de forma aleatória dos 348 nódulos, e assumindo que a distribuição é uniforme, a probabilidade destes nódulos serem malignos deverá continuar a ser aproximadamente 34%. No entanto, caso aconteça que todos os 84 casos seleccionados aleatoriamente apresentem densidade de massa elevada (*high density*), então a percentagem de casos malignos subirá para os 70.2% (percentagem de casos simultaneamente malignos e com densidade de massa elevada – valor retirado dos 348 dados[24]), sendo que a probabilidade deste facto ser uma coincidência é bastante

[24] Ver Figura 20.

baixa, de acordo com a distribuição dos dados. Esta simples suposição é, desde já, indício de que a densidade de massa elevada (*high density*) está directamente relacionada com o conceito de malignidade, tal como uma série de outros atributos, nomeadamente a idade dos pacientes (*age_at_mammo*), as formas e margens dos nódulos (*mass shape, mass margins*), entre outros.

Um dos objectivos do nosso estudo é confirmar se estes atributos têm alguma relação com a variável *outcome_num* (atributo alusivo à malignidade dos nódulos).

Como referido anteriormente, um subconjunto de 180 casos (≈ 52%) do universo de 348 foi classificado em termos de densidade por um especialista, que não teve qualquer informação relativa aos resultados das biópsias efectuadas aos nódulos. Isto significa que os restantes casos, ou seja, 168 (≈ 48%), não foram alvo de classificação por parte deste radiologista. Como tal, utilizamos estes dois subconjuntos para a aplicação do conceito de aprendizagem automática, em que os 180 dados são o nosso conjunto de treino e os restantes 168 casos o nosso conjunto de teste.

Todas as experiências foram executadas fazendo uso da ferramenta de mineração de dados WEKA. Em cada um dos ensaios foram aplicados uma série de algoritmos[25] de aprendizagem automática que o sistema WEKA disponibiliza, sendo que apenas os algoritmos que apresentaram os melhores resultados serão alvo de discussão no próximo capítulo (5) – *Análise de Resultados*. É importante mencionar que aquando da aplicação dos algoritmos, os parâmetros definidos internamente foram os parâmetros *default* do próprio WEKA.

De seguida, apresentamos os passos essenciais referentes à ferramenta WEKA para a aprendizagem nos 180 dados.

[25] Ver Tabela 1 da subsecção 2.4.2.

4.2.1 Aprendizagem

A aplicação *Experimenter* é a interface mais adequada para a execução de experiências, uma vez que permite a escolha simultânea de várias tarefas e técnicas a serem testadas num único ensaio. Além do mais, a experiência é executada sem ser necessária a intervenção do utilizador, tendo este posteriormente acesso aos resultados guardados num determinado ficheiro.

Sendo assim, a interface *Experimenter* proporciona ao utilizador três painéis distintos: *Setup*, *Run* e *Analyse*.

As experiências têm início no painel *Setup*, onde serão configuradas por parte do utilizador, sendo que este poderá optar por um de dois modos: simples (*Simple*) ou avançado (*Advanced*). Para as nossas experiências utilizamos o modo *Simple*.

No modo *Simple* é possível configurar uma nova experiência (*New*) definindo o ficheiro de destino dos resultados para posterior análise em *Results Destination*. Em *Experiment Type*, escolhe-se entre *Cross-validation* ou *Train/Test Percentage Split*. Além do mais, é possível optar entre o método de classificação (*Classification*) ou regressão (*Regression*). Torna-se importante referir que em todas as nossas experiências de classificação escolhemos a técnica de *cross-validation* com o valor por omissão de *10-fold*, ou seja, onde são utilizados 10 desdobramentos.

Em *Datasets* são adicionados os conjuntos de dados que serão alvo de estudo[26].

Em *Iteration Control* define-se o número de vezes que cada técnica será testada, sendo possível alterar a ordem da iteração entre *Data sets first* ou *Algorithms first*. Ao longo das nossas experiências optamos por 10 repetições (*runs*) e por *Data sets first*.

Por último, em *Algorithms* é possível escolher uma série de algoritmos de aprendizagem automática para serem aplicados aos conjuntos de dados que se pretendem

[26] O *dataset* presente na Figura 25 diz respeito à experiência para a previsão de densidade de massa (*mass density*) baseada na densidade anotada pelo radiologista (*density_num*) no modelo prospectivo.

estudar. No nosso caso particular, seleccionamos 12 algoritmos[27] que são baseados em árvores de decisão, regras de classificação, SVM's e redes bayesianas. Tal como acima mencionado, para todos os algoritmos mantiveram-se as definições (parâmetros) inicialmente sugeridas pela própria ferramenta de mineração de dados WEKA.

O aspecto da interface *Experimenter*, configurado tal como foi descrito, pode ser observado na Figura 25.

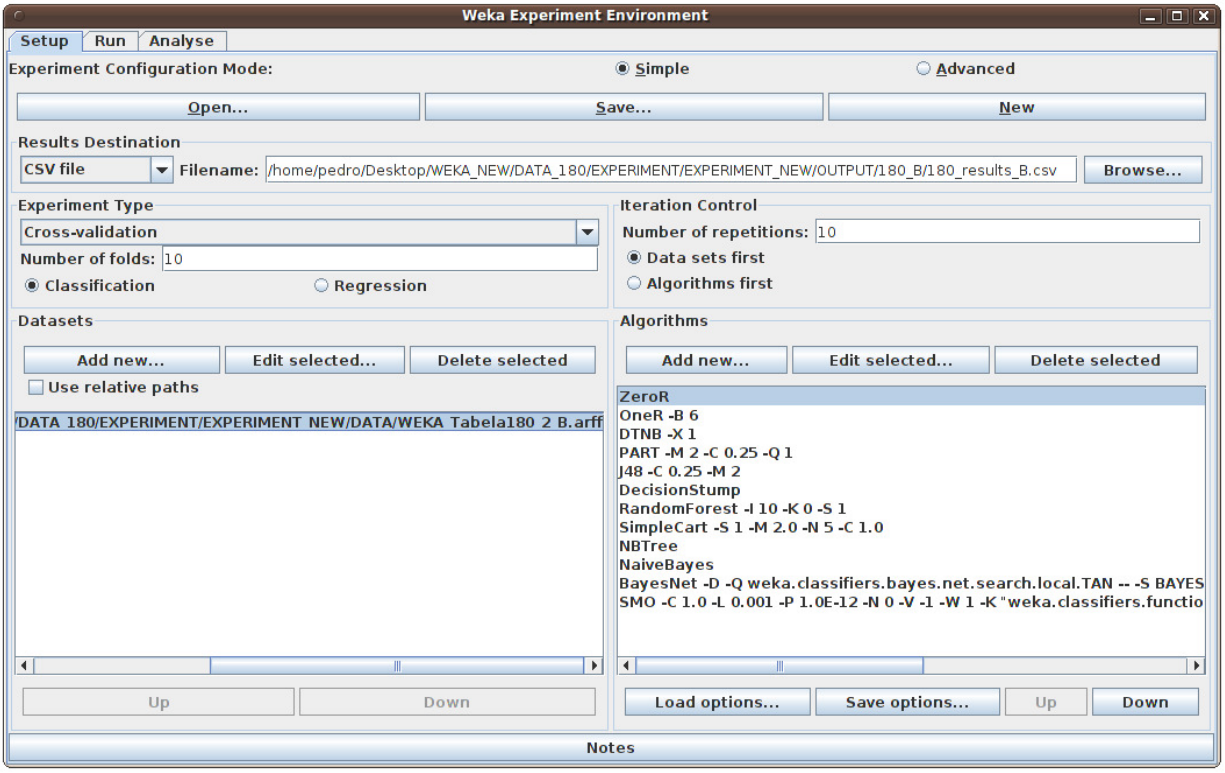


Figura 25 - Experimenter configurado para classificação com *10-fold cross-validation*

Seleccionando o painel *Run* é possível dar início a uma experiência (através do botão *Start*). Enquanto um determinado ensaio decorre é apresentado ao utilizador uma espécie de relatório que o informa sobre o estado da experiência. É igualmente possível interromper um ensaio a qualquer momento (através do botão *Stop*). Quando uma

[27] Além de estarem presentes na Figura 25, é possível consultar informação relativa a estes 12 algoritmos na Tabela 1 da subsecção 2.4.2.

experiência termina, os resultados são gravados num ficheiro previamente escolhido, além de que este painel informa o utilizador que a experiência em questão foi finalizada com sucesso. Após o término de uma experiência, é possível efectuar uma análise dos resultados através do painel *Analyse*. O botão *Experiment* permite analisar os resultados da experiência que acaba de ser executada.

Alternativamente, é possível especificar um ficheiro com os resultados. Configura-se o teste (*Configure test*) começando por se definir o que se pretende nas linhas (*Row*) e nas colunas (*Column*). No caso da Figura 26, surge na linha o conjunto de dados treinado, enquanto as colunas exibem a percentagem de instâncias correctamente classificadas por cada um dos 12 algoritmos aplicados, segundo um nível de significância (*Significance*) de 0.01. Aliás, em todas as nossas experiências, os resultados foram testados de acordo com um dos testes standard de significância do WEKA, nomeadamente o teste *Paired corrected T-Tester* (disponível em *Testing with*). O nível de significância utilizado em todos os ensaios foi, precisamente, de 0.01. Ao utilizarmos este valor significa que a análise estatística ao conjunto de dados tratados nesta dissertação gera um número que é estatisticamente significativo caso seja inferior a 1%, o qual é designado por nível de confiança. Por outras palavras, se a probabilidade de ocorrência de um evento é estatisticamente significativa, poderemos estar 99% seguros de que os resultados não acontecerem por acaso.

Ao lado das percentagens de instâncias correctamente classificadas surgem, entre parêntesis, os desvios-padrão respectivos (*Show std. deviations*). É importante salientar que esta espécie de relatório (*Test output*) com os resultados obtidos é apresentado depois de pressionado o botão *Perform test*.

Em *Test base* é possível definir qual o algoritmo que se pretende comparar em termos de significância com todos os outros. Caso os resultados do algoritmo seleccionado sejam estatisticamente significativos relativamente aos resultados de um ou mais algoritmos, o símbolo "*" surge por baixo dos valores obtidos por esses mesmos algoritmos. Por sua vez, na base da tabela apresentada no *Test output* (a partir da 2ª coluna) surge o número de vezes em que um algoritmo é melhor, igual ou pior (v/ /*) que o algoritmo da 1ª coluna.

Na Figura 26, é apresentado como *Test base* o algoritmo *naive Bayes*, uma vez que apresentou os melhores resultados na previsão de densidade de massa (*mass density*) baseada na densidade anotada pelo radiologista (*density_num*) no modelo prospectivo.

Por último, o campo *Comparison field* permite seleccionar de um conjunto extremamente vasto de métricas, aquela que se pretende comparar quando aplicados diferentes algoritmos (tal como acima referido, na Figura 26 são exibidas as percentagens de instâncias correctamente classificadas (*Percent_correct*)).

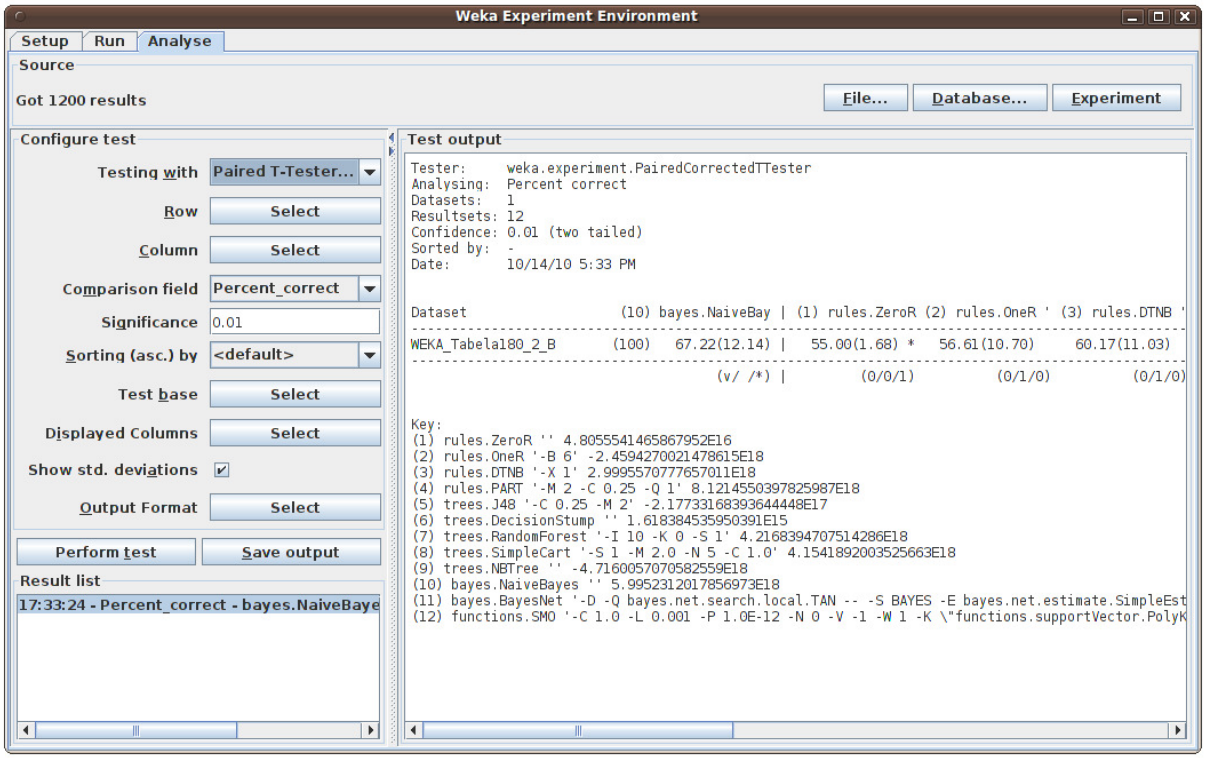


Figura 26 - Resultado de uma experiência de classificação com *10-fold cross-validation*

4.2.2 Teste

Dos 348 casos, o subconjunto de 180 (≈ 52%) foi utilizado como conjunto de treino. Os restantes 168 casos (≈ 48%) foram usados como dados de teste, de forma a avaliar a performance de alguns classificadores. De seguida passamos a descrever o modo como utilizamos os modelos gerados nas experiências de aprendizagem (relativas aos 180 casos) para a classificação de instâncias num conjunto de dados desconhecidos (168).

Apesar da aplicação *Experimenter* ser a interface mais adequada para a execução de experiências, no WEKA *Explorer* é também possível desenvolver tarefas de classificação.

Deste modo, para a utilização de um determinado modelo para classificar instâncias de um conjunto de dados desconhecidos, a aplicação *Explorer* é muito provavelmente a plataforma mais fácil e rápida para a execução de uma tarefa deste tipo.

Inicialmente procede-se ao *upload* do *dataset* de treino, fazendo uso do botão *Open file* no painel *Preprocess*. Na Figura 27 está presente o conjunto de dados treinado com *10-fold cross-validation* relativo à experiência para a previsão de densidade de massa (*mass density*) baseada na densidade anotada pelo radiologista (*density_num*) no modelo prospectivo.

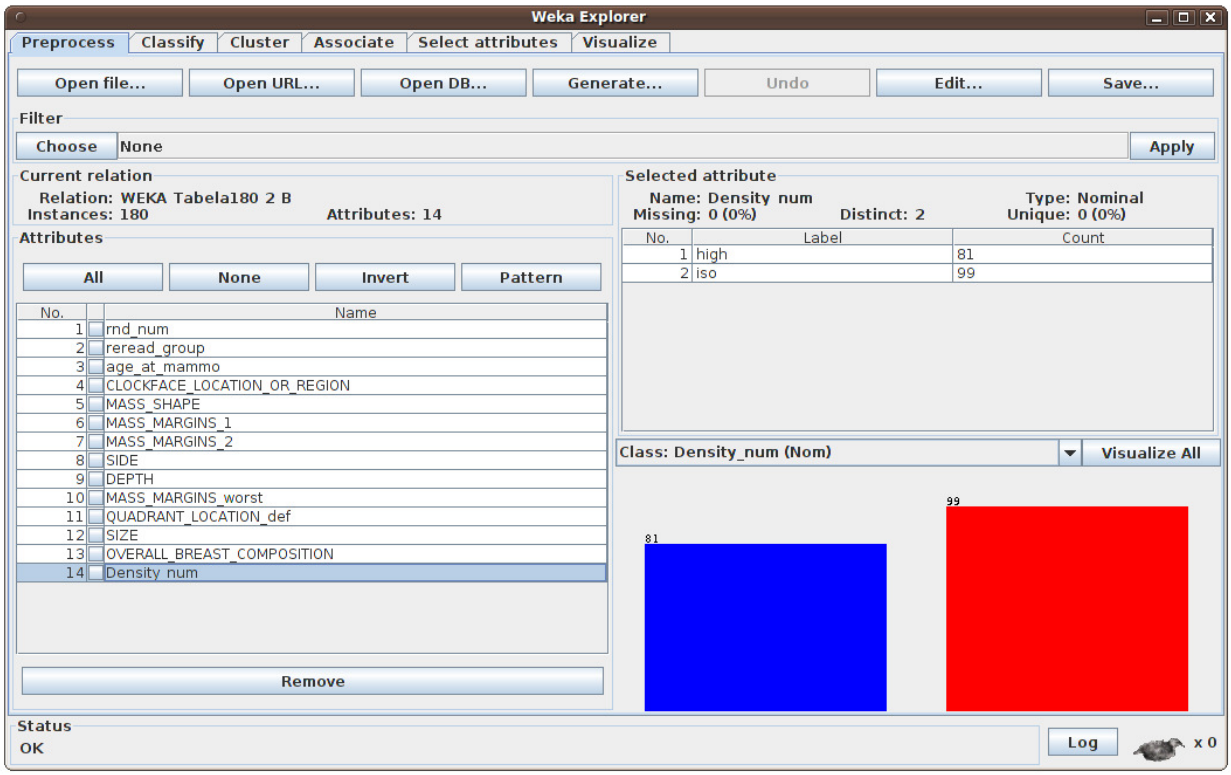


Figura 27 - *Dataset* de treino que servirá como modelo para a classificação de instâncias de um conjunto de dados desconhecidos

Depois de carregado o conjunto de treino, no painel *Classify* é necessário definir o algoritmo que para esse mesmo conjunto (aquando da aprendizagem com *10-fold cross-validation*) apresentou os melhores resultados. Será portanto o modelo desta experiência. De acordo com a Figura 28, em *Classifier* e pressionando o botão *Choose* é possível seleccionar o algoritmo em causa. No caso da experiência retratada nessa mesma figura trata-se do algoritmo *naive Bayes*.

A Figura 28 representa portanto o resultado de uma experiência de classificação em que foi utilizado um modelo *naive Bayes* para prever instâncias da classe *density_num* num novo conjunto de dados.

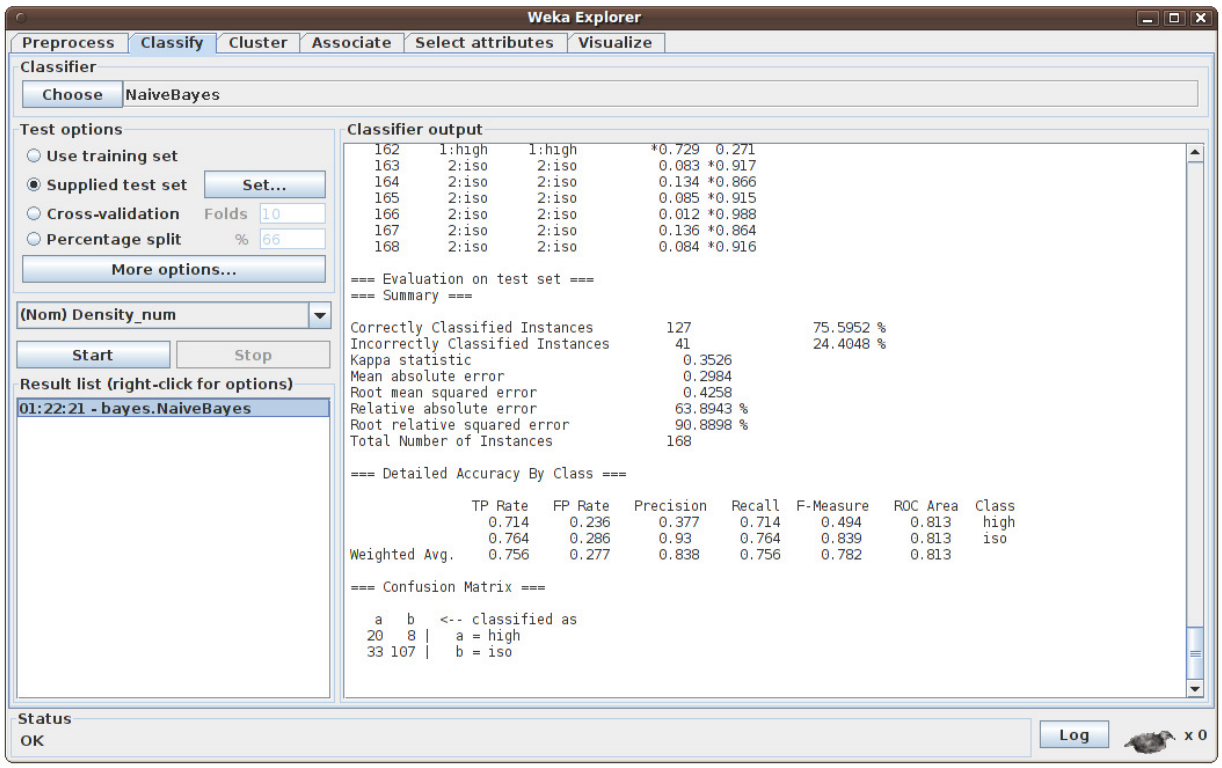


Figura 28 - Resultado de uma experiência de classificação em que foi utilizado um modelo *naive Bayes* para prever instâncias da classe *Density_num* num novo conjunto de dados

Para a validação do modelo gerado é necessário escolher qual a opção de teste em *Test options*. Uma vez que o nosso conjunto de teste se trata de um conjunto de instâncias desconhecidas, a opção a seleccionar será *Supplied test set*. Clicando no botão *Set* é possível carregar o conjunto de teste. Torna-se importante referir que é aconselhável que o próprio ficheiro com os dados de teste apresente como último atributo os dados reais relativos à classe que se pretende prever, uma vez que o próprio WEKA, no momento da classificação das novas instâncias, coloca ao lado dos valores reais, em *Classifier output*, os valores previstos com as respectivas probabilidades de acerto. Trata-se de uma informação bastante útil para o utilizador, visto que assim poderá ter uma ideia da fiabilidade dos resultados.

A presença das listas de instâncias reais e previstas lado a lado, assim como a apresentação das listas de probabilidades, apenas estarão visíveis, caso depois de pressionado o botão *More options*, se seleccione a opção *Output predictions*.

O botão *Start* permite iniciar a experiência de classificação.

A janela *Classifier output*, além das instâncias previstas, dispõe de uma série de métricas de desempenho que atestam a qualidade dos resultados na classificação de instâncias de um conjunto de dados desconhecidos (168) a partir de um modelo de aprendizagem (180).

Os resultados e a análise de todas as experiências de classificação estarão presentes no próximo capítulo (5) – *Análise de Resultados*.

Capítulo 5

Análise de Resultados

Neste capítulo são apresentados os resultados obtidos. A análise a esses mesmos resultados é efectuada através da tentativa de resposta a três questões fundamentais:

1. Será densidade de massa um factor relevante no diagnóstico de cancro de mama?

2. Será possível obter classificadores capazes de preverem densidade de massa com um nível qualitativo semelhante ao de um radiologista?

3. Qual o comportamento dos classificadores gerados num conjunto de dados desconhecidos?

5.1 Será densidade de massa um factor relevante no diagnóstico de cancro de mama?

Tomamos em consideração pelo menos duas formas de investigar se a densidade de massa é efectivamente um indicador de malignidade. A primeira tratar-se-ia de aplicar regras de associação ou regressão logística aos 348 casos, e posteriormente reportar a

relação entre *retro_density* e *outcome_num*. Esta tarefa, no entanto, já foi executada por Woods e Burnside [WB10] num trabalho anterior, fazendo uso de regressão logística e estatística *Kappa*. Os seus resultados revelaram que a densidade de massa elevada (*high density*) é um indicador de malignidade relativamente importante com uma taxa de concordância entre observadores (estatística *Kappa*) de 0.53.

A segunda via para averiguar se densidade de massa é de facto um indicador de malignidade passaria pela utilização de métodos de classificação, de forma a preverem o atributo *outcome_num* quer com informação relativa a densidade como sem qualquer tipo de informação sobre densidade de massa; e em seguida comparar os resultados.

Como nos nossos dados temos dois tipos de densidades de massa – um relativo aos dados retrospectivos (*retro_density*) e um outro de acordo com os dados prospectivos (*density_num*) – utilizamos ambos para construir classificadores.

Antes de mais, torna-se importante referir que para todas as experiências aplicamos o método de ***10-fold cross-validation***, com um valor de ***Paired corrected T-Tester*** de **0.01**.

Sendo assim, a primeira experiência (E_1) consiste em gerar um classificador para prever *outcome_num* com *retro_density* (densidade de massa retrospectiva). A segunda experiência (E_2), por sua vez, consiste em gerar um classificador para prever *outcome_num* fazendo uso de *density_num* (densidade de massa anotada prospectivamente).

De modo a apurar se a densidade de massa de um nódulo é um indicador de malignidade geramos também um classificador (E_3) para prever *outcome_num* sem qualquer espécie de informação sobre densidade de massa.

A Tabela 9 apresenta uma série de métricas consideradas relevantes que permitem resumir os resultados alcançados. Nas três experiências, os melhores classificadores encontrados são baseados em SVM's [Pla98].

180	Previsão de *outcome_num*		
	com densidade de massa		E_3
Métrica	E_1 Retrospectiva (*retro_density*)	E_2 Prospectiva (*density_num*)	sem densidade de massa
Instâncias Correctamente Classificadas	84.78% (7.96)	82.72% (8.32)	81.39% (8.81)
Estatística *Kappa*	0.68 (0.17)	0.63 (0.17)	0.60 (0.18)
Precisão	0.84 (0.12)	0.82 (0.13)	0.81 (0.14)
Recall	0.78 (0.15)	0.75 (0.15)	0.72 (0.15)
F-Measure	0.80 (0.11)	0.77 (0.11)	0.75 (0.12)

Tabela 9 - Previsão de *outcome_num* em 180 casos. Os valores entre parêntesis representam desvios-padrão

Os resultados obtidos revelam que densidade de massa tem alguma influência sobre o atributo *outcome_num*, acima de tudo, quando a densidade é a observada nos dados retrospectivos (E_1).

O classificador (E_3) treinado sem informação relativa a densidade de massa apresenta uma performance global de 81.39% (+/- 8.81) enquanto o classificador (E_1) treinado com a densidade retrospectiva (*retro_density*) revela uma performance global de 84.78% (+/- 7.96). Estes resultados são estatisticamente diferentes (p=0.01). Além do mais, se observarmos os valores de estatística *Kappa*, podemos confirmar que a relação entre densidade de massa e malignidade não é por acaso, tendo em conta o nível de concordância relativamente alto observado entre os dados reais e os valores previstos pelos classificadores.

Quanto à precisão, os resultados também são positivos, com apenas 16% de casos a serem incorrectamente classificados como malignos, aquando da utilização de um classificador (E_1) treinado com *retro_density* (densidade de massa retrospectiva).

A métrica *recall*, por sua vez, apresenta uma taxa relativamente razoável de casos malignos correctamente classificados, no entanto ainda com margem de progressão para aperfeiçoamentos.

Resumindo, estes resultados revelam que se adicionarmos informação alusiva à densidade de massa dos nódulos a outros atributos já de si importantes, a performance de um classificador aumenta.

Outro indício bastante forte da importância de densidade de massa na previsão de malignidade são as árvores de decisão (Figuras 29 e 30) geradas pelo algoritmo J48, em que colocam *retro_density* e *density_num* nas suas raízes.

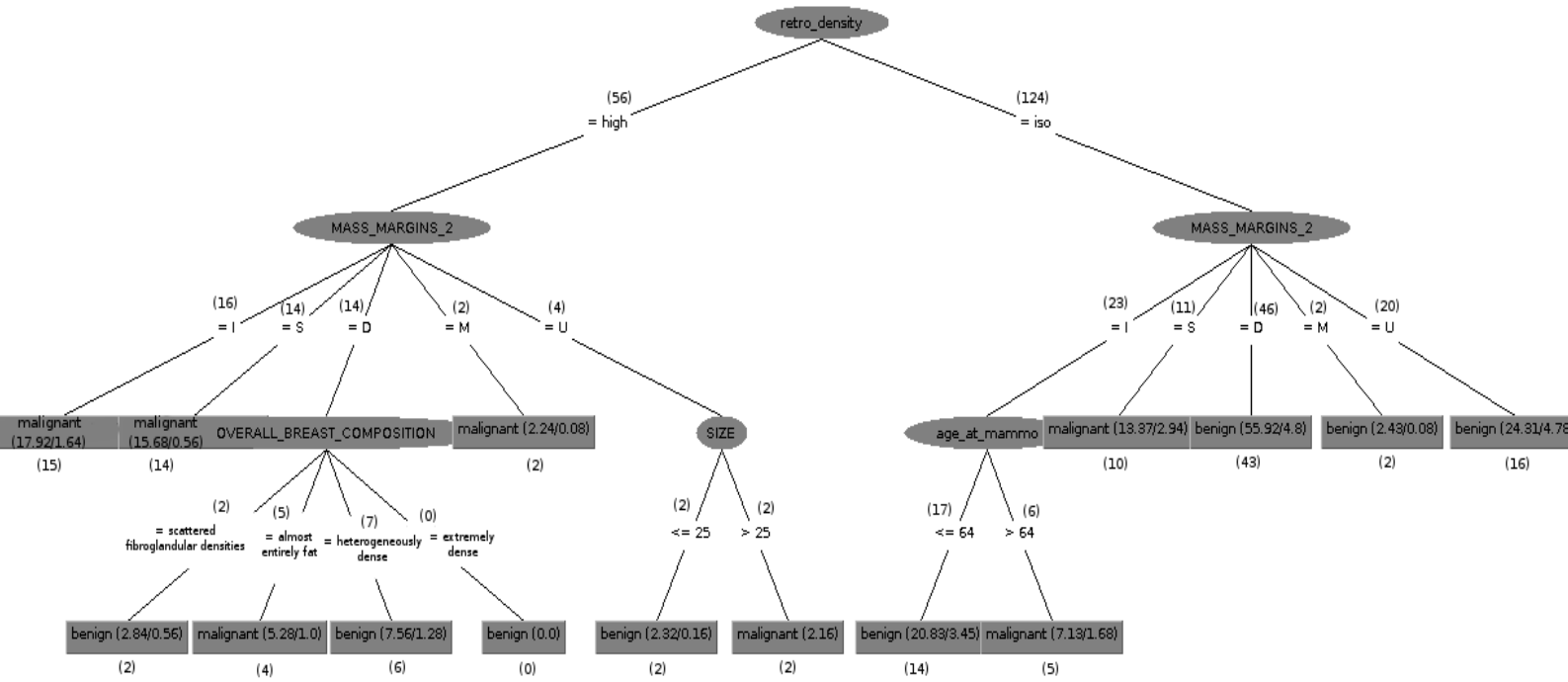


Figura 29 - Árvore de decisão gerada pelo algoritmo J48 relativa à experiência E_1: previsão de *outcome_num* com *retro_density*. Os números entre parêntesis representam o número de instâncias na realidade naqueles pontos da árvore

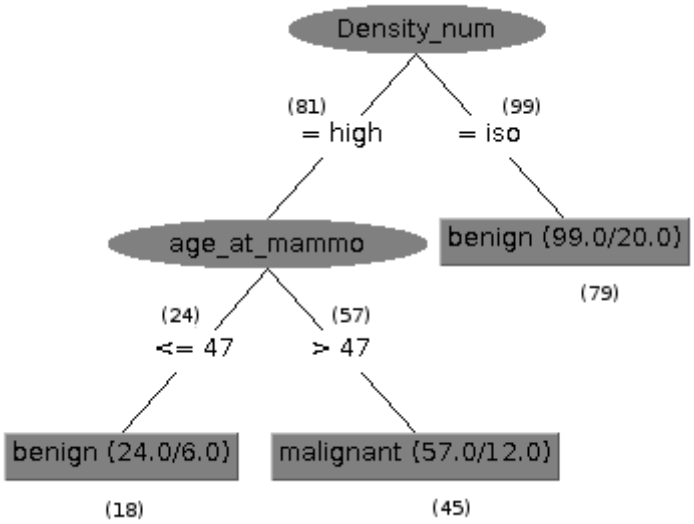


Figura 30 - Árvore de decisão gerada pelo algoritmo J48 relativa à experiência E_2: previsão de *outcome_num* com *Density_num*. Os números entre parêntesis representam o número de instâncias na realidade naqueles pontos da árvore

Tal como mencionamos em situações anteriores, o nosso estudo incidiu sobre 12 algoritmos, sendo que apenas apresentamos os resultados relativos aos algoritmos com melhor performance[28]. No entanto, o algoritmo J48, apesar de não apresentar os índices mais elevados em termos de instâncias correctamente classificadas, estatística *Kappa* ou *F-Measure*, gerou árvores de decisão extremamente interessantes que reforçam a relevância de densidade de massa como factor preponderante no diagnóstico de cancro de mama.

Estes resultados confirmam os indícios presentes na literatura relativos à importância da densidade de massa dos nódulos, e mostram também que é possível obter bons classificadores para a previsão de *outcome_num* (quer com uma percentagem elevada de instâncias correctamente classificadas como com valores de estatística *Kappa*, precisão e *recall* bastante satisfatórios).

[28] Os resultados de todos os algoritmos para as diversas experiências encontram-se para consulta em *Apêndice C*.

5.2 Será possível obter classificadores capazes de preverem densidade de massa com um nível qualitativo semelhante ao de um radiologista?

A nossa segunda questão está relacionada com a qualidade de previsão de um classificador relativamente à avaliação de um especialista.

Visto que temos dois tipos de densidades de massa – um para o estudo retrospectivo (*retro_density*) e um outro para o estudo prospectivo (*density_num*) – geramos dois classificadores: um (E_4) é treinado utilizando os valores retrospectivos de densidade de massa (*retro_density*), enquanto o outro (E_5) é treinado sobre os valores prospectivos de densidade (*density_num*). Uma vez mais, utilizamos os 180 casos como conjunto de treino e aplicamos o método de ***10-fold cross-validation***.

O melhor classificador obtido pelo WEKA *Experimenter* para estas duas tarefas baseia-se no algoritmo *naive Bayes* [JL95]. A Tabela 10 ilustra os resultados para estas experiências como uma média das métricas para os 10 *folds*.

180	Previsão de densidade de massa	
Métrica	E_4 ***retro_density***	E_5 ***density_num***
Instâncias Correctamente Classificadas	72.83% (9.89)	67.22% (12.14)
Estatística *Kappa*	0.37 (0.23)	0.33 (0.25)
Precisão	0.58 (0.20)	0.66 (0.16)
Recall	0.58 (0.22)	0.60 (0.17)
F-Measure	0.56 (0.18)	0.62 (0.15)

Tabela 10 - Previsão de densidade de massa em 180 casos. Os valores entre parêntesis representam desvios-padrão

Ao longo do estudo prospectivo (para os 180 casos) o radiologista classificou de acordo com a classificação retrospectiva (padrão de referência) exactamente 70% das instâncias, isto é, classificou de forma correcta, em termos de densidade, 126 dos 180 nódulos.

Id_Exam	reread_group	age_at_mammo	MASS_MARGINS_worst	SIZE	OVERALL_BREAST_COMPOSITION	retro_density	Density_num
94	salkowski	41	Circumscribed	13	scattered fibroglandular densities	iso	iso
95	salkowski	69		12	scattered fibroglandular densities	iso	iso
96	burnside	40		16	extremely dense	iso	iso
97	burnside	66	Indistinct	13	almost entirely fat	high	high
98	salkowski	55	Obscured	12	scattered fibroglandular densities	iso	iso
100	sisney	48	Indistinct	12	heterogeneously dense	iso	iso
101	burnside	40	Circumscribed	17	heterogeneously dense	high	high
103	sisney	42	Circumscribed	8	heterogeneously dense	iso	iso
104	burnside	37	Spiculated	14	extremely dense	iso	iso
107	burnside	64	Circumscribed	7	almost entirely fat	iso	iso
108	sisney	41	Indistinct	20	heterogeneously dense	iso	iso
109	sisney	75	Circumscribed	17	almost entirely fat	high	high
110	sisney	67	Spiculated	10	almost entirely fat	high	high
113	salkowski	46	Circumscribed	8	heterogeneously dense	iso	iso
116	burnside	61		34	heterogeneously dense	iso	iso
119	salkowski	49	Obscured	19	heterogeneously dense	iso	iso
120	sisney	39	Circumscribed	20	scattered fibroglandular densities	iso	iso
124	sisney	50		12	heterogeneously dense	iso	iso
127	burnside	78	Indistinct	5	almost entirely fat	high	high
134	salkowski	49	Indistinct	13	heterogeneously dense	iso	iso
137	sisney	46		20	heterogeneously dense	high	high
138	salkowski	66	Circumscribed	20	almost entirely fat	high	high
145	salkowski	49	Circumscribed	13	scattered fibroglandular densities	iso	iso
146	burnside	35	Indistinct	28	heterogeneously dense	iso	iso
148	sisney	56	Indistinct	6	heterogeneously dense	iso	iso

126 rows fetched in 0:00.1131 Start Editing Apply Changes

Figura 31 - Excerto da Base de Dados MySQL. Representação de parte das instâncias correctamente classificadas pelo radiologista no modelo prospectivo (*Density_num*). O nosso padrão de referência é o modelo retrospectivo, nomeadamente o atributo *retro_density*. A informação relativa ao total de instâncias correctamente classificadas (126) no modelo prospectivo surge no canto inferior esquerdo da imagem

O classificador *naive Bayes* previu aproximadamente 73% (+/- 9.89) de instâncias correctas quando treinado sobre os nódulos retrospectivamente anotados (E_4) e cerca de 67% (+/- 12.14) quando treinado sobre os casos prospectivamente classificados por um radiologista (E_5).

Estes resultados são consideravelmente bons e indicam que o classificador bayesiano gerado neste estudo poderá ser aplicado a novos exames como ferramenta de auxílio médico na previsão de densidade de massa dos nódulos.

No entanto, os valores de estatística *Kappa*, precisão, *recall* e *F-Measure* para estas duas experiências não são tão elevados como os resultados obtidos aquando da previsão de malignidade (*outcome_num*) apresentados na Tabela 9. Mesmo assim, os valores de estatística *Kappa* presentes na Tabela 10 revelam que o classificador *naive Bayes* apresenta um certo nível de concordância com os dados actuais.

Um facto interessante também a observar é que, apesar do classificador (E_4) treinado com os valores retrospectivos de densidade de massa (*retro_density*) apresentar uma percentagem superior de instâncias correctamente classificadas, exibe valores inferiores de precisão, *recall* e *F-Measure* relativamente ao classificador (E_5) treinado sobre os valores prospectivos de densidade (*density_num*). Este pormenor parece indiciar que o classificador da experiência E_5 poderá apresentar melhor performance a classificar dados que contenham erros prévios de classificação.

- **Curvas *Precision-Recall*: Comportamento do classificador *naive Bayes* durante a aprendizagem (180 casos)**

Ao treinar uma rede bayesiana, o algoritmo normalmente atribui probabilidades aos exemplos classificados. Com estas probabilidades é possível construir uma curva (ROC ou PR) cujo objectivo é analisar como é que o classificador se comporta com a variação destas probabilidades utilizadas como *threshold*[29]. Os resultados mostrados nas Tabelas 9 e 10 são obtidos usando o valor de *threshold* de omissão do WEKA (0.5).

Na Figura 32, podemos observar um espaço PR que apresenta um panorama mais alargado do comportamento dos classificadores bayesianos encontrados pelo WEKA para a previsão de densidade de massa (retrospectiva (E_4) e prospectiva (E_5)) em relação à

[29] Limite; limiar; valor mínimo relativo a uma determinada quantidade.

classe *high density*, quando variamos os *thresholds*. Nesta figura, também apresentamos a performance do radiologista (ponto azul).

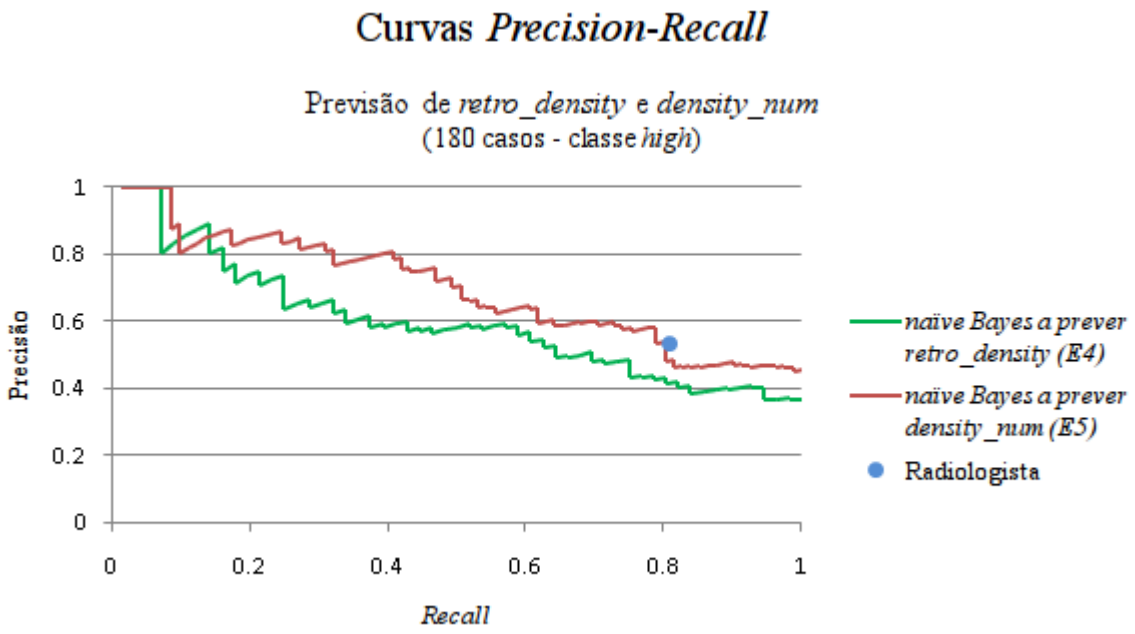


Figura 32 - Espaço PR: Previsão de densidade de massa (retrospectiva e prospectiva) em relação à classe *high density* por classificadores bayesianos em 180 casos[30]

Na curva relativa à previsão de *retro_density* (E_4), para um valor de *recall* igual a 0.8, o radiologista alcança melhor valor de precisão que o classificador obtido. Porém, variando o *threshold* do classificador, podemos alcançar valores melhores de *recall* com um custo pela perda de precisão. No contexto clínico, perder precisão (na prática, classificar incorrectamente instâncias negativas) pode ser tolerado desde que não implique um custo elevado.

[30] Espaço ROC equivalente em *Apêndice D* – ver Figura 38.

O classificador tem um desempenho visivelmente superior quando é treinado com os dados fornecidos pelo próprio radiologista, ou seja, quando treinado com *density_num* (E_5). Trata-se de um desempenho muito semelhante ao do especialista, podendo atingir valores melhores de precisão (aumento de 0.53 para 0.58) com uma redução muito pequena de *recall* (0.81 para 0.79). Por outro lado, se o radiologista julgar que um aumento em *recall* é mais importante, com uma perda de aproximadamente 15% de precisão (redução de 0.53 para 0.45) em relação ao valor do especialista, poderíamos ter *recall* perfeito (melhoria de 23% em relação ao radiologista) e desta forma classificar todos os nódulos com densidade alta (*high density*) de forma correcta.

Se analisarmos o problema inverso e construirmos a curva PR em relação à classificação dos nódulos de densidade média/baixa (*iso-dense*) (Figura 33), obtemos um perfil melhor do classificador quando "aprende" com os dados de *retro_density* (E_4), estando este classificador muito próximo do desempenho do radiologista.

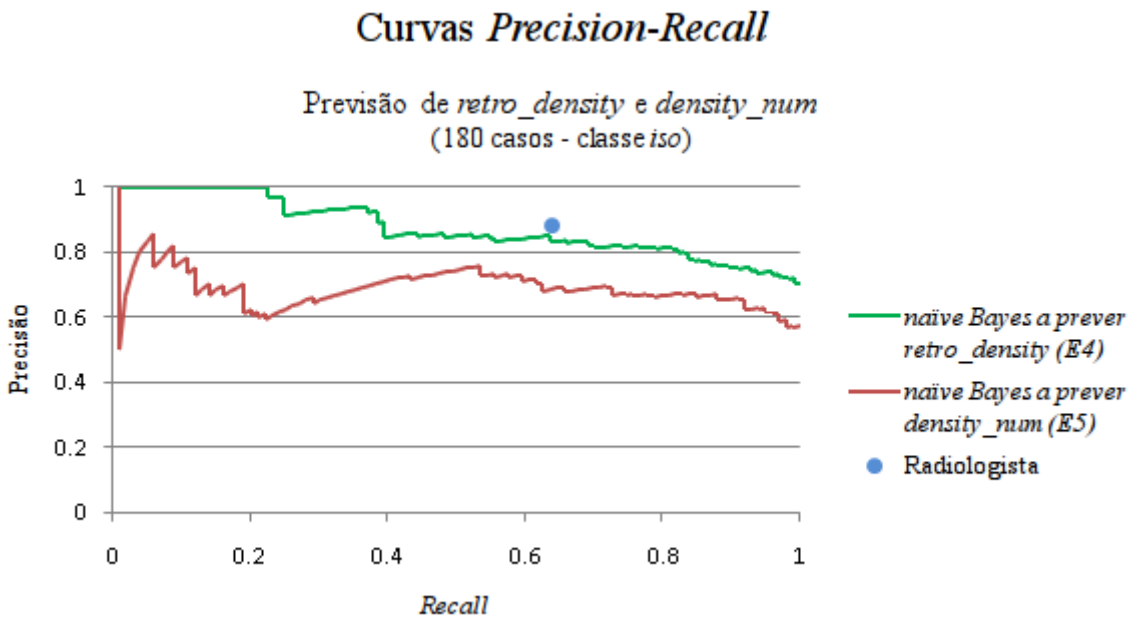


Figura 33 - Espaço PR: Previsão de densidade de massa (retrospectiva e prospectiva) em relação à classe *iso-dense* por classificadores bayesianos em 180 casos[31]

[31] Espaço ROC equivalente em *Apêndice D* – ver Figura 39.

Quando o classificador é obtido através da aprendizagem com *density_num* (E_5), apresenta um desempenho inferior, porém não muito longe da performance do especialista. Este facto parece indicar que para a obtenção de um classificador que faça a correcta classificação de nódulos nas classes média/baixa (*iso-dense*) é importante ter resultados de um estudo retrospectivo dos exames dos pacientes.

5.3 Qual o comportamento dos classificadores gerados num conjunto de dados desconhecidos?

A nossa última questão está relacionada com o modo como classificadores que foram sujeitos a aprendizagem poderão prever malignidade e densidade de massa num conjunto de dados totalmente novo.

De modo a responder a esta questão necessitamos novamente de considerar quer os classificadores gerados que fazem uso da densidade de massa retrospectivamente anotada, como os classificadores que utilizam a densidade de massa prospectiva.

O primeiro classificador (E_1), baseado nos valores retrospectivos de densidade de massa, foi construído após a aprendizagem para os 180 casos, de modo a responder à nossa primeira questão (5.1):

- "Será densidade de massa um factor relevante no diagnóstico de cancro de mama?"

Trata-se de um classificador baseado em SVM's. No entanto, podemos utilizar ainda um outro classificador, baseado nos valores de densidade de massa prospectivamente anotados, para prever os 168 casos que sobraram do universo de 348 instâncias. Como os 168 novos casos não possuem qualquer tipo de densidade de massa prospectivamente anotada, preenchemos estes valores em falta recorrendo aos classificadores gerados aquando da resposta à questão (5.2):

- "Será possível obter classificadores capazes de preverem densidade de massa com um nível qualitativo semelhante ao de um radiologista?"

Nestas experiências foram gerados dois classificadores para prever densidade de massa: um treinado sobre *retro_density* (***E*$_6$**) e um outro treinado sobre *density_num* (***E*$_7$**). Ambos são classificadores bayesianos.

Uma vez preenchidos estes valores, é possível aplicar um classificador "aprendido" para prever *outcome_num* para este conjunto de 168 novos casos.

Os resultados da previsão de densidade de massa no novo conjunto de dados estão representados na Tabela 11. Estes resultados foram obtidos pelo melhor classificador que, em ambos os casos, tratou-se do algoritmo *naive Bayes*.

168	Previsão de densidade de massa	
Métrica	E_6 ***retro_density***	E_7 ***density_num***
Instâncias Correctamente Classificadas	82.14%	75.60%
Estatística *Kappa*	0.45	0.35
Precisão	0.48	0.38
Recall	0.68	0.71
F-Measure	0.56	0.49

Tabela 11 - Previsão de densidade de massa num conjunto de 168 novos casos

Estes resultados são bastante bons, tendo em conta que ambos os classificadores apresentam para um conjunto de dados desconhecidos (168 casos) uma performance em termos de instâncias correctamente classificadas muito acima da observada para os dados de treino (180 casos) (Tabela 10). Os valores de estatística *Kappa* e *recall* também são superiores no conjunto de novos casos.

Observamos, no entanto, a partir dos valores de precisão e *F-Measure* da Tabela 11, uma ligeira quebra na performance aquando da previsão de casos de densidade média/baixa (*iso-dense*). A taxa de falsos positivos aumenta no conjunto de dados desconhecidos. Por outro lado, o algoritmo apresenta melhor desempenho na classificação de instâncias de densidade alta (*high density*).

- **Curvas *Precision-Recall*: Comportamento do classificador *naive Bayes* no conjunto de teste (168 casos)**

As Figuras 34 e 35 mostram as curvas PR para a classificação das 168 instâncias não classificadas pelo radiologista. O classificador de *retro_density*, para a classe *high* (Figura 34), tem um desempenho melhor na classificação das 168 instâncias do que obteve durante a aprendizagem com os 180 casos (na Figura 32, a curva do classificador de *density_num* domina a do classificador de *retro_density*). Apenas para valores de *recall* mais altos (a partir de 0.8) o classificador de *retro_density* mantém-se abaixo do desempenho do classificador de *density_num*.

Para alcançar a performance do radiologista no que diz respeito ao valor de *recall*, os dois classificadores sofrem uma queda de desempenho na precisão, que desce de 0.53 (radiologista) para 0.45 (classificador de *retro_density*). Se, por outro lado, quisermos manter o mesmo nível de desempenho do especialista em relação à precisão, comprometemos o valor de *recall* que desce de 0.81 para aproximadamente 0.6. Os classificadores conseguem alcançar o mesmo nível de *recall* do radiologista (com uma perda em precisão), tratando-se de um resultado muito bom, considerando que são classificadores aprendidos automaticamente.

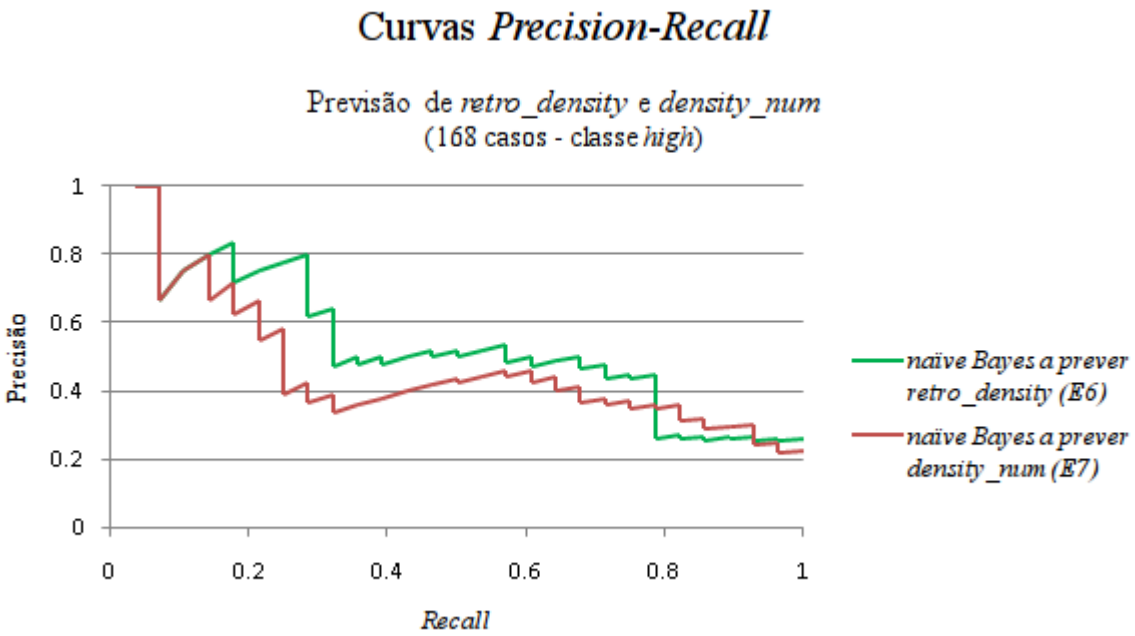


Figura 34 - Espaço PR: Previsão de densidade de massa (retrospectiva e prospectiva) em relação à classe *high density* por classificadores bayesianos em 168 novos casos[32]

[32] Espaço ROC equivalente em *Apêndice D* – ver Figura 40.

Ambos os classificadores têm desempenho quase perfeito nas 168 instâncias no que diz respeito à classe *iso* (Figura 35), apresentando resultados muito superiores em novas instâncias do que nos dados de treino.

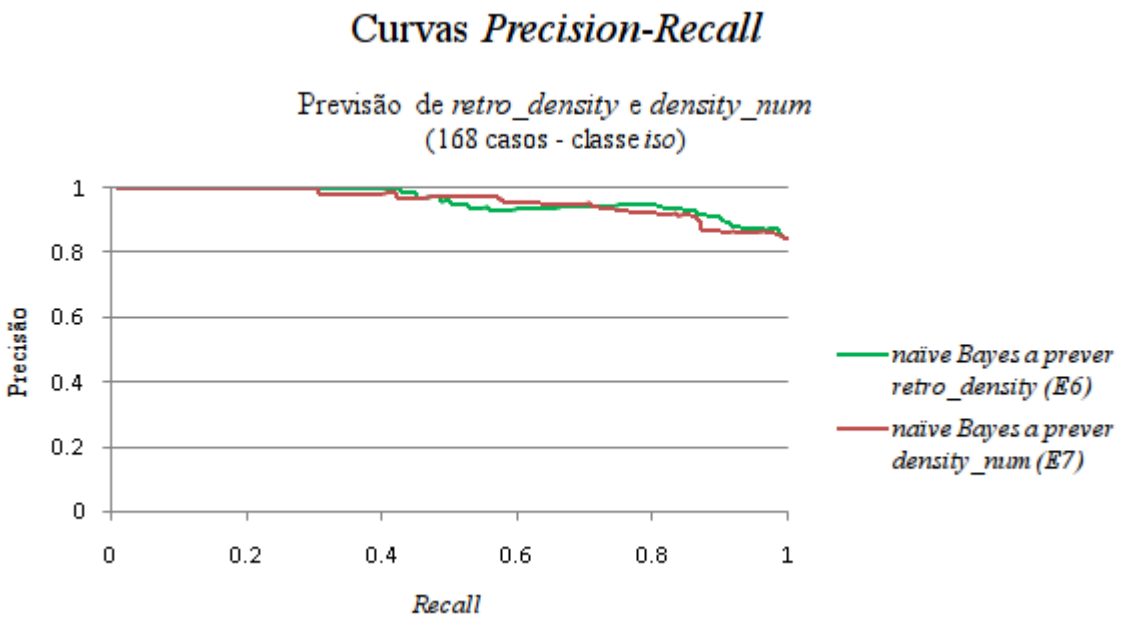


Figura 35 - Espaço PR: Previsão de densidade de massa (retrospectiva e prospectiva) em relação à classe *iso-dense* por classificadores bayesianos em 168 novos casos[33]

[33] Espaço ROC equivalente em *Apêndice D* – ver Figura 41.

Como considerações finais, relativamente a estes gráficos, podemos concluir o seguinte:

a) Para a previsão de instâncias do tipo *high*, o facto de se utilizar informação relativa ao estudo retrospectivo (*retro_density*) não auxilia o desempenho do classificador ao longo da aprendizagem, no entanto poderá ajudar na fase de classificação de novas instâncias (fase de teste);

b) Para a previsão de instâncias do tipo *high*, ambos os classificadores estão muito próximos do desempenho do radiologista, podendo alcançar valores de *recall* mais elevados, caso seja possível comprometer algumas das instâncias negativas (alguns pacientes que, à partida, não se teriam que submeter a exames adicionais, teriam que o fazer);

c) Para a previsão de instâncias do tipo *iso*, ambos os classificadores, independentemente de utilizarem ou não informação relativa ao estudo retrospectivo, apresentam uma performance quase perfeita na presença de novos dados, e além do mais superior relativamente ao desempenho do radiologista.

A partir do momento em que são preenchidos os valores de densidade de massa previstos para os 168 pacientes, passamos para o próximo passo, que consiste na previsão de *outcome_num* para este mesmo conjunto de novos dados. Os resultados dessas experiências poderão ser consultados na Tabela 12.

168	**Previsão de *outcome_num***			
	com densidade de massa			E_{11} **sem densidade de massa**
Métrica	E_8 **Retrospectiva (*retro_density*)** (actual)	E_9 **Retrospectiva (*retro_density*)** (preenchida pelo classificador *naive Bayes*)	E_{10} **Prospectiva (*density_num*)** (preenchida pelo classificador *naive Bayes*)	
Instâncias Correctamente Classificadas	81.55%	79.76%	79.17%	77.38%
Estatística *Kappa*	0.52	0.48	0.46	0.42
Precisão	0.70	0.65	0.65	0.61
Recall	0.60	0.60	0.55	0.53
F-Measure	0.64	0.62	0.60	0.57

Tabela 12 - Previsão de *outcome_num* num conjunto de 168 novos casos

Nesta tabela são apresentadas três previsões diferentes para *outcome_num*, fazendo uso de três tipos de densidades de massa distintos. A segunda coluna (***E*$_8$**) da tabela exibe os resultados da previsão de *outcome_num* com densidade de massa do estudo retrospectivo (atributo *retro_density*). A terceira (***E*$_9$**) e quarta (***E*$_{10}$**) colunas mostram as previsões aquando da utilização de densidade de massa preenchida pelos dois classificadores bayesianos (*naive Bayes*) – um treinado sobre o atributo *retro_density* (***E*$_6$**) e um outro treinado sobre *density_num* (***E*$_7$**).

Comparando as três previsões de malignidade (*outcome_num*), é possível constatar que os três classificadores comportam-se relativamente bem no conjunto de dados desconhecidos, classificando correctamente a maioria dos casos malignos e benignos. O valor de estatística *Kappa*, uma vez mais, indica que estes resultados não aconteceram fruto de um simples acaso. Por outras palavras, os classificadores estão de facto a auxiliar na distinção entre casos malignos e benignos. Tal como observado anteriormente, o classificador treinado sobre os dados retrospectivos produz melhores resultados, no entanto, os outros classificadores também revelam bons níveis de performance. Este facto permite afirmar que a ausência de informação relativa ao estudo retrospectivo não prejudica a tarefa de classificação.

Aliás, uma segunda constatação que tiramos destes resultados é que, apesar de fazermos uso de valores previstos de densidade de massa (com erros de previsão

inerentes), os classificadores para *outcome_num* das colunas três (***E₉***) e quatro (***E₁₀***) da Tabela 12, mantêm um desempenho bastante razoável.

Uma última conclusão que retiramos também é o facto de densidade de massa estar de certa forma relacionada com malignidade (*outcome_num*), revelando-se um atributo extremanente importante que contribui para um aumento da performance dos classificadores. Uma simples comparação entre os dados da última coluna (***E₁₁***) da Tabela 12 (previsão de *outcome_num* sem informação relativa a densidade de massa) com os dados das restantes colunas, confirma precisamente esse facto.

Resumindo, e de certo modo recordando a questão 5.2, a Tabela 13 ilustra a performance de todos os classificadores utilizados quer nas experiências de treino como nas experiências de teste, para a previsão de densidade de massa.

180/168	**Previsão de densidade de massa**				
Métrica	**Radiologista (180)**	***E₅*** ***density_num*** **(180)**	***E₇*** ***density_num*** **(168)**	***E₄*** ***retro_density*** **(180)**	***E₆*** ***retro_density*** **(168)**
Instâncias Correctamente Classificadas	70.00%	67.22% (12.14)	75.60%	72.83% (9.89)	82.14%
Estatística *Kappa*	0.52	0.33 (0.25)	0.35	0.37 (0.23)	0.45
Precisão	0.53	0.66 (0.16)	0.38	0.58 (0.20)	0.48
Recall	0.81	0.60 (0.17)	0.71	0.58 (0.22)	0.68
F-Measure	0.64	0.62 (0.15)	0.49	0.56 (0.18)	0.56

Tabela 13 - Previsão de densidade de massa

Esta tabela permite-nos concluir que os classificadores gerados apresentam boas performances, sendo que em alguns casos são mesmo superiores às obtidas pelo próprio radiologista.

A performance nos 168 novos casos é também relativamente satisfatória, tendo em conta os valores de precisão e *recall*.

As Figuras 36 e 37 representam uma síntese do comportamento dos classificadores bayesianos na previsão de densidade de massa para os 180 e 168 casos.

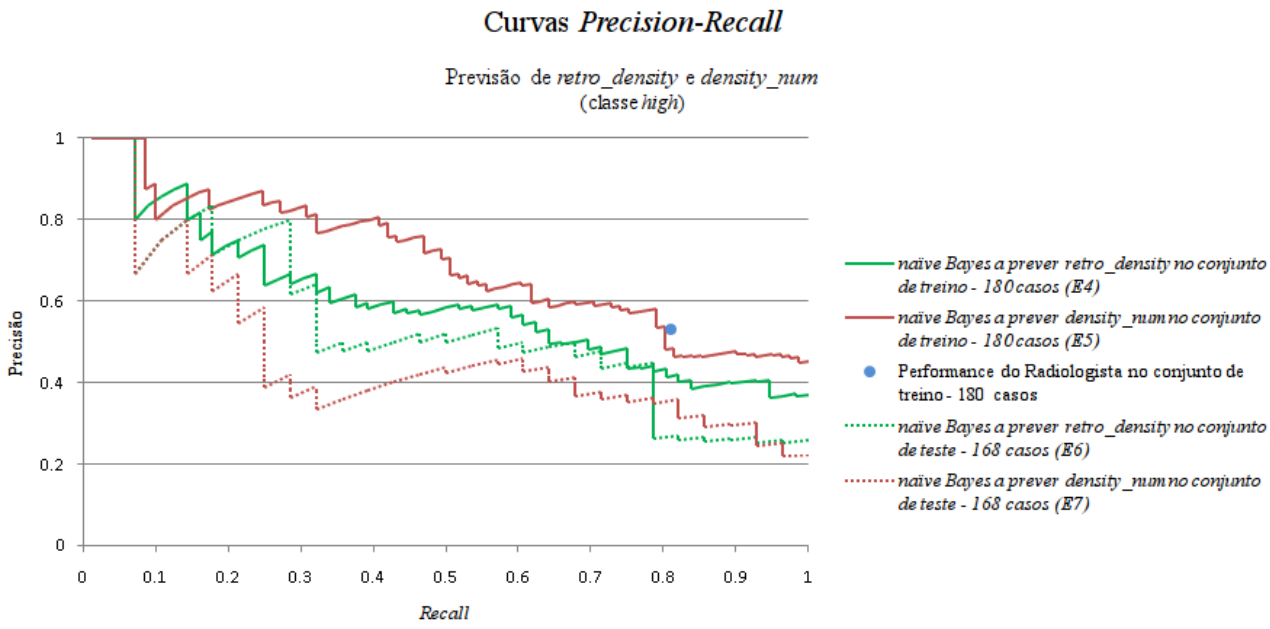


Figura 36 - Espaço PR: Previsão de densidade de massa (retrospectiva e prospectiva) em relação à classe *high density* por classificadores bayesianos em 180 e 168 casos[34]

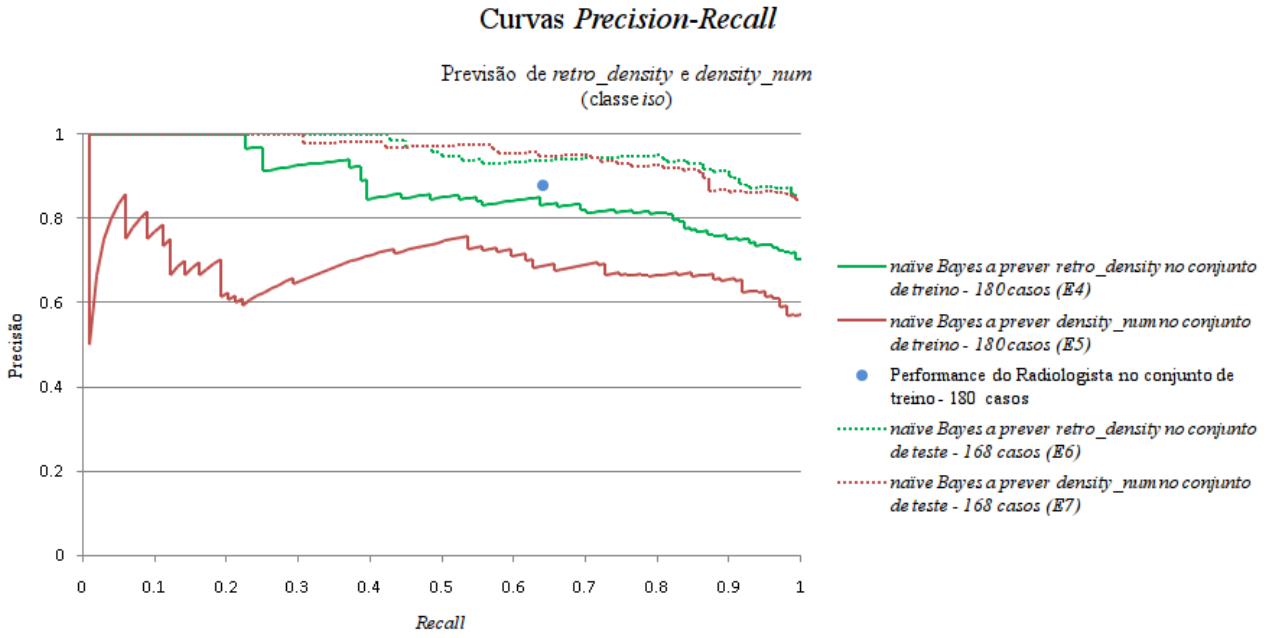


Figura 37 - Espaço PR: Previsão de densidade de massa (retrospectiva e prospectiva) em relação à classe *iso-dense* por classificadores bayesianos em 180 e 168 casos[35]

[34] Espaço ROC equivalente em *Apêndice D* – ver Figura 42.

[35] Espaço ROC equivalente em *Apêndice D* – ver Figura 43.

Por fim, a Tabela 14 apresenta o desempenho de todos os classificadores utilizados quer nas experiências de aprendizagem como nas experiências de teste, para a previsão de malignidade (*outcome_num*).

180/168	**Previsão de *outcome_num***						
	com densidade de massa					**E_3 (180)**	**E_{11} (168)**
Métrica	**E_1 (180)** **Retrospectiva (*retro_density*)**	**E_8 (168)** **Retrospectiva (*retro_density*)** (actual)	**E_9 (168)** **Retrospectiva (*retro_density*)** (preenchida pelo classificador *naive Bayes*)	**E_2 (180)** **Prospectiva (*density_num*)**	**E_{10} (168)** **Prospectiva (*density_num*)** (preenchida pelo classificador *naive Bayes*)	**sem densidade de massa**	**sem densidade de massa**
Instâncias Correctamente Classificadas	84.78% (7.96)	81.55%	79.76%	82.72% (8.32)	79.17%	81.39% (8.81)	77.38%
Estatística *Kappa*	0.68 (0.17)	0.52	0.48	0.63 (0.17)	0.46	0.60 (0.18)	0.42
Precisão	0.84 (0.12)	0.70	0.65	0.82 (0.13)	0.65	0.81 (0.14)	0.61
Recall	0.78 (0.15)	0.60	0.60	0.75 (0.15)	0.55	0.72 (0.15)	0.53
F-Measure	0.80 (0.11)	0.64	0.62	0.77 (0.11)	0.60	0.75 (0.12)	0.57

Tabela 14 - Previsão de *outcome_num*

Estes resultados reforçam a importância da densidade de massa na previsão de *outcome_num*, acima de tudo pela comparação entre os valores relativos às experiências que fazem uso deste atributo e entre os resultados de prever malignidade sem qualquer tipo de informação sobre densidade de massa.

Capítulo 6

Conclusões e Trabalho Futuro

Nesta dissertação foram-nos disponibilizados 348 casos relativos a pacientes que foram sujeitos a exames de rastreio de cancro de mama, nomeadamente mamografias.

Os objectivos deste trabalho eram:

i. Encontrar relações entre os atributos através da aplicação de técnicas de aprendizagem automática aos dados;

ii. "Aprender" modelos capazes de auxiliarem os médicos na avaliação imediata de mamografias.

Para tal, utilizamos a ferramenta de aprendizagem automática WEKA e sempre que possível efectuamos testes estatísticos de significância aos resultados obtidos.

São três as conclusões a que chegamos:

a) A classificação automática de uma mamografia poderá alcançar resultados semelhantes ou mesmo superiores aos obtidos pelos próprios especialistas;

b) A densidade de massa parece ser, efectivamente, um bom indicador de malignidade, tal como estudos anteriores sugeriam;

c) Os classificadores de aprendizagem automática são capazes de prever densidade de massa com um nível qualitativo tão bom como o de um especialista sem qualquer tipo de informação relativa a biópsias.

Como trabalho futuro, planeamos estender este estudo a universos de dados maiores e geograficamente distintos, assim como aplicar outras técnicas de aprendizagem automática baseadas em aprendizagem estatística relacional.

Bibliografia

[AAC+10] T. Ayer, O. Alagoz, J. Chhatwal, J. W. Shavlik, C. E. J. Kahn, and E. S. Burnside. *Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration*. Vol. Cancer. 2010.

[Abb02] H. A. Abbass. *An evolutionary artificial neural networks approach for breast cancer diagnosis*. Artificial Intelligence in Medicine, 2002.

[AS94] R. Agrawal and R. Srikant. *Fast Algorithms for Mining Association Rules*. Proceedings of the 20th Int'l Conference on Very Large Databases. Santiago, Chile, Set. 1994.

[BA96] Ronald J. Brachman and Tej Anand. *The process of knowledge discovery in databases*. Advances in Knowledge Discovery and Data Mining. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.

[BCS10] *Breast Cancer Statistics*. [Online]. Disponível em: http://www.cdc.gov/cancer/breast/statistics/. CDC – Centers for Disease Control and Prevention, Out. 2010.

[BKML+05] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and David L. Wheeler. *Genbank*. Nucleic Acids Research, 2005.

[BWF+00] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. *The protein data bank*. Nucleic Acids Research, 2000.

[CL93] R. C. Cory and S. S. Linden. *The mammographic density of breast cancer*. AJR Am J Roentgenol, 1993.

[Cru07] A. J. R. Cruz. *Data Mining via Redes Neuronais Artificiais e Máquinas de Vectores de Suporte*. Universidade do Minho, Escola de Engenharia, Departamento de Sistemas de Informação, 2007.

[CS04] Adriana C. G. Corrêa and Homero Schiabel. *Descoberta de Conhecimento em Base de Imagens Mamográficas*. Departamento de Engenharia Elétrica da EESC/USP, Universidade de São Paulo, Brasil, 2004.

[DBD+05] J. Davis, E. S. Burnside, I. C. Dutra, D. Page, and V. S. Costa. *Knowledge discovery from structured mammography reports using inductive logic programming*. American Medical Informatics Association 2005 Annual Symposium, 2005.

[DCO+04] J. Davis, V. S. Costa, I. M. Ong, D. Page, and I. Dutra. *Using Bayesian Classifiers to Combine Rules*. Department of Biostatistics and Medical Informatics, University of Madison-Wisconsin, 2004.

[DG06] Jesse Davis and Mark Goadrich. *The Relationship Between Precision-Recall and ROC Curves*. Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA. Department of Computer Sciences and Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, 2006.

[DKG00] N. A. Diamantidis, D. Karlis, and E. A. Giakoumakis. *Unsupervised stratification of cross-validation for accuracy estimation*. Vol. 116. 2000.

[FDF+11] P. Ferreira, I. Dutra, N. A. Fonseca, R. Woods, and E. Burnside. *Studying the relevance of Breast Imaging Features*, in *Proceedings of the international Conference on Health Informatics (HealthInf)*, Jan. 2011.

[Fon06] Nuno A. Fonseca. *Parallelism in Inductive Logic Programming Systems*, PhD thesis, Faculdade de Ciências da Universidade do Porto, 2006.

[FPSS96] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. *From data mining to knowledge discovery: An overview*. Advances in Knowledge Discovery and Data Mining, 1996.

[Gam99] J. M. P. Gama. *Combining Classification Algorithms*. Porto, 1999.

[Her09] R. A. Hernandez. *MP-SMO: um algoritmo para a implementação VSLI do treinamento de máquinas de vetores de suporte*. Dissertação (Mestrado) – Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Sistemas Eletrônicos. São Paulo, 2009.

[HK06] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. 2nd Edition, ELSEVIER, 2006.

[Hos05] Véronique Hoste. *Optimization Issues in Machine Learning of Coreference Resolution*, PhD thesis, University of Antwerpen – Belgium, 2005.

[JD88] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data.* Englewood Cliffs: N.J.: Prentice-Hall, 1988.

[JDB+91] V. P. Jackson, K. A. Dines, L.W. Bassett, R. H. Gold, and H. E. Reynolds. *Diagnostic importance of the radiographic density of noncalcified breast masses: analysis of 91 lesions*. AJR Am J Roentgenol, 1991.

[JL95] G. H. John and P. Langley. *Estimating continuous distributions in bayesian classifiers*. San Mateo: Morgan Kaufmann, 1995.

[KP98] R. Kohavi and F. Provost. *Machine Learning*. Vol. Glossary of Terms. 1998.

[Lee05] Huei Diana Lee. *Seleção de atributos importantes para a extração de conhecimento de bases de dados*, PhD thesis, Instituto de Ciências Matemáticas e de Computação – ICMC-USP, 2005.

[Lig10] *Liga Portuguesa Contra o Cancro – Cancro da Mama*. [Online]. Disponível em: http://www.ligacontracancro.pt/gca/index.php?id=14. Out. 2010.

[LR06] D. S. Leite and L. H. M. Rino. *A migração do SuPor para o WEKA: potencial e abordagens*. Universidade de São Paulo – USP, Universidade Federal de São Carlos – UFSCar, Universidade Estadual Paulista – UNESP. São Paulo, Brasil, 2006.

[MBK98] R. S. Michalski, I. Bratko, and M. Kubat. *Machine Learning and Data Mining: Methods and Applications*. West Sussex, England: John Wiley and Sons, 1998.

[Mit99] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1999.

[MMC09] A. C. Martins, J. M. Marques, and P. D. Costa. *Estudo Comparativo de Três Algoritmos de Machine Learning na Classificação de Dados Electrocardiográficos*. Faculdade de Medicina da Universidade do Porto, Mestrado em Informática Médica. Porto, 2009.

[MST94] D. Michie, D. J. Spiegelhalter, and C. C. Taylor. *Machine learning, neural and statistical classification (edited collection)*. New York: Ellis Horwood, 1994.

[NPA+10] H. Nassif, D. Page, M. Ayvaci, J. Shavlik, and E. S. Burnside. *Uncovering age-specific invasive and dcis breast cancer rules using inductive logic programming*. Proceedings of 2010 ACM International Health Informatics Symposium (IHI 2010), ACM Digital Library, 2010.

[NWB+09] H. Nassif, R. Woods, E. Burnside, M. Ayvaci, J. Shavlik, and D. Page. *Information extraction for clinical data mining: A mammography case study*. ICDMW'09: Proceedings of the 2009 IEEE International Conference on Data Mining Workshops, Washington, DC, USA: IEEE Computer Society, 2009.

[Orv08] Lurdes Orvalho. *Cancro da mama: detectar é fundamental*. Centro de Imagiologia do Hospital da Luz. 2008.

[PK95] U. Pompe and I. Kononenko. *Naive Bayesian classifier within ILP-R*. Proceedings of the 5th International Workshop on Inductive Logic Programming. Department of Computer Science, Katholieke Universiteit Leuven: L. De Raedt. 1995.

[Pla98] J. C. Platt. *Sequential minimal optimization: A fast algorithm for training support vector machines*. Microsoft Research. Technical Report MSR-TR-98-14. 1998.

[Pla99] J. C. Platt. *Fast training of support vector machines using sequencial minimal optimization*. [book auth.] B. Schölkopf, C. J. C. Burges and A. J. Smola. *Advances in kernel methods: support vector learning*. 1st Edition. Cambridge: MIT Press, 1999.

[Por05] *Portal da Saúde – Cancro da mama*. [Online]. Disponível em: http://www.portaldasaude.pt/portal/conteudos/enciclopedia+da+saude/doencas/cancro/cancro+mama.htm. Ministério da Saúde, Out. 2005.

[Pyl99] D. Pyle. *Data Preparation for Data Mining*. California: Morgan Kaufmann Publishers, 1999.

[Rae08] Troy Raeder. *Model Monitor User's Guide version 1.0*. Department of Computer Science and Engineering, University of Notre Dame, 2008.

[RN03] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. 2nd Edition. Upper Saddle River, New Jersey: Prentice-Hall, 2003.

[RPMP03] S. O. Rezende, J. B. Pugliesi, E. A. Melanda, and M. F. Paula. *Mineração de Dados*. 1ª Edição. pp. 307-336. Vol. I. 2003.

[SB05] T. Soman and P. O. Bobbie. *Classification of Arrhythmia Using Machine Learning Techniques*. Proceedings of the 4th International Conference on System Science and Engineering. Rio de Janeiro, Brasil, 2005.

[Sic91] E. A. Sickles. *Periodic mammographic follow-up of probably benign lesions: results in 3,184 consecutive cases*. Vol. Radiology. 1991.

[Sil04] M. P. S. Silva. *Mineração de Dados – Conceitos, Aplicações e Experimentos com Weka*. Mossoró, RN, Brasil: Universidade do Estado do Rio Grande do Norte, 2004.

[SMW95] W. N. Street, O. L. Mangasarian, and W. H. Wolberg. *An inductive learning approach to prognostic prediction*. ICML, 1995.

[TSM85] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite-Mixture Distributions*. Chichester: U.K.: Wiley, 1985.

[VG05] Anthony J. Viera, MD, and Joanne M. Garrett, PhD. *Understanding Interobserver Agreement: The Kappa Statistic*. University of North Carolina, USA, 2005.

[WB10] Ryan Woods and Elizabeth Burnside. *The mammographic density of a mass is a significant predictor of breast cancer*. Radiology, USA, 2010.

[WF00] I. H. Witten and E. Frank. *Data Mining: Pratical Machine Learning Tools and Techniques with Java Implementations*. California: Morgan Kaufmann Publishers, 2000.

[WF05] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd Edition. San Francisco: Elsevier, 2005.

[WGD+93] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz. *Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer*. Vol. Radiology. 1993.

[WKQ+07] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. *Top 10 algorithms in data mining*. London: Springer-Verlag, 2007.

[WM90] W. H. Wolberg and O. L. Mangasarian. *Multisurface method of pattern separation for medical diagnosis applied to breast cytology*. Proceedings of the National Academy of Sciences, 1990.

[WOS+09] R. Woods, L. Oliphant, K. Shinki, D. Page, J. Shavlik, and E. Burnside. *Validation of results from knowledge discovery: Mass density as a predictor of breast cancer*. J Digit Imaging, 2009.

Apêndice A

Artigo *Studying the relevance of Breast Imaging Features*

STUDYING THE RELEVANCE OF BREAST IMAGING FEATURES

Pedro Ferreira, Inês Dutra
Department of Computer Science & CRACS-INESC Porto LA, University of Porto, Porto, Portugal
pedroferreira@dcc.fc.up.pt, ines@dcc.fc.up.pt
Nuno A. Fonseca
CRACS-INESC Porto LA, Porto, Portugal
nunofonseca@acm.org
Ryan Woods
Department of Radiology, Johns Hopkins Hospital, Baltimore, MD, USA
ryan_woods@alumni.bowdoin.edu
Elizabeth Burnside
Department of Radiology, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA
EBurnside@uwhealth.org

Keywords:
Mass Density, Breast Cancer, Mammograms, Classification Methods, Data Mining, Machine Learning

Abstract:

Breast screening is the regular examination of a woman's breasts to find breast cancer in an initial stage. The sole exam approved for this purpose is mammography that, despite the existence of more advanced technologies, is considered the cheapest and most efficient method to detect cancer in a preclinical stage.
We investigate, using machine learning techniques, how attributes obtained from mammographies can relate to malignancy. In particular, this study focus is on how mass density can influence malignancy from a data set of 348 patients containing, among other information, results of biopsies. To this end, we applied different learning algorithms on the data set using the WEKA tools, and performed significance tests on the results. The conclusions are threefold: (1) automatic classification of a mammography can reach equal or better results than the ones annotated by specialists, which can help doctors to quickly concentrate on some specific mammogram for a more thorough study; (2) mass density seems to be a good indicator of malignancy, as previous studies suggested; (3) we can obtain classifiers that can predict mass density with a quality as good as the specialist blind to biopsy.

1 INTRODUCTION

Breast screening is the regular examination of a woman's breasts to find breast cancer earlier. The sole exam approved for this purpose is mammography. Usually, findings are annotated through the Breast Imaging Reporting and Data System (BIRADS) created by the American College of Radiology. The BI-RADS system determines a standard lexicon to be used by radiologists when studying each finding. Despite the existence of more advanced technologies, mammography is considered the cheapest and most efficient method to detect cancer in a preclinical stage.

In this work, we were provided with 348 cases of patients that went through mammography screening. Our main objective is to apply machine learning techniques to these data in order to find non trivial relations among attributes, and learn models that can help medical doctors to quickly assess mammograms.

Much work has been done on applying machine learning techniques to the study of breast cancer, which is one of the most common kinds of cancer in the world. In the UCI (University of California, Irvine) machine learning repository (`http://archive.ics.uci.edu/ml/datasets.html`) there are four data sets whose main target of study is breast cancer. One of the first works on applying machine learning techniques to breast cancer data dates from 1990. The data set used in this study, donated to the UCI repository, was created by Wolberg and Mangasarian after their work on a multisurface method of pattern separation for medical diagnosis applied to breast cytology (Wolberg and Mangasarian, 1990). Most works in the literature applies artificial neural networks to the problem of diagnosing breast cancer (e.g., (Wu et al., 1993) and (Abbass, 2002)). Others focus on prognostic of the disease using inductive learning methods (e.g., (Street et al., 1995)). More recently, Ayer *et al.* (Ayer et al., 2010) have evaluated whether an artificial neural network, trained on

a large prospectively collected data set of consecutive mammography findings, could discriminate between benign and malignant disease, and accurately predict the probability of breast cancer for individual patients. Other recent studies focus on extracting information from free text that appears in medical records of mammography screenings (Nassif et al., 2009), and on the influence of age in ductal carcinoma in situ (DCIS) findings (Nassif et al., 2010).

Our study is focused on the influence of mass density on predicting malignancy, but we also uncover other interesting complementary findings. Previous works by Jackson *et al.* (Jackson et al., 1991) and Cory and Linden (Cory and Linden, 1993) have argued that, although the majority of high density masses are malignant, the presence of low density cancers and more important indicators (like margins, shape, and associated findings) make mass density a less reliable indicator or predictor of malignancy. Sickles (Sickles, 1991) has the same opinion. A study carried out by Davis *et al.* (Davis et al., 2005) indicated that mass density could have more importance and relevance than previous works had reported. In another work, Woods *et al.* (Woods et al., 2009) applied inductive logic programming to a set of breast cancer data and concluded the same thing. Woods and Burnside (Woods and Burnside, 2010) also applied logistic regression and kappa statistics to another set of breast cancer data and concluded that mass density and malignancy are somewhat related.

In this work, we use the same data set used by Woods and Burnside (Woods and Burnside, 2010), but we apply machine learning methods and confirm the findings of Woods and Burnside. In addition, we show that the learned classifiers generated in this work can predict mass density and outcome (classification of a mammography) with a quality as good as a specialist, proving to be good helpers to medical doctors when evaluating mammograms.

2 BREAST CANCER DATA

Our study analyzes 348 consecutive breast masses that underwent image guided or surgical biopsy performed between October 2005 and December 2007 on 328 female subjects. All 348 biopsy masses were randomized and assigned to a radiologist blinded to biopsy results for retrospective assessment using the Breast Imaging Reporting and Data System (retrospectively-assessed data set). Clinical radiologists prospectively assessed the density of 180 of these masses (prospectively-assessed data set). Pathology result at biopsy was the study endpoint.

The atributes included in our study are very much the ones collected by the radiologists from the mammograms, and are based on the BIRADS lexicon. We selected from the original database all the attributes considered relevant by the specialists and removed some attributes such as identifiers, redundant attributes and attributes that had the same value for all instances. For our main task, to predict malignancy, our class attribute was the outcome binary variable assuming values benign or malignant.

From the 348 cases, 118 are malignant ($\approx$ 34%), and 84 cases have high mass density ($\approx$ 24%) retrospectively assessed. Other attributes are mass shape, mass margins, depth, size, among others. For the purpose of our study, we have two attributes that represent the same characteristics of the finding, but with different interpretations. These are retro_density and density_num. Both represent mass densities that can assume values *high* or *iso/low*. Retro_density was retrospectively assessed while density_num was prospectively (at the time of imaging) assessed.

3 EXPERIMENTS AND RESULTS

Our first preliminary study was to calculate simple frequencies from the data and to determine if there was some evidence of relationship between attributes, specially, the main focus of our study:

Is mass density related to malignancy?

As mentioned above, from the 348 breast masses, 118 are malignant ($\approx$ 34%), and 84 have high mass density ($\approx$ 24%). If we consider that mass density and malignancy are independent, and take 84 cases from the 348 at random, the probability of these being malignant should still be $\approx$ 34%. However, if it happens that all 84 cases selected at random have high density, then the percentage of malignant cases raises to 70.2% (this is the percentage of cases that are both malignant and have high mass density). The probability of this being coincidence is very low, given the data distribution. This simple calculation may already imply that high density has some relation to malignancy. So may imply that other attributes such as age, mass shape and mass margins can have some relation to malignancy. One of the objectives of our study is then to confirm if these attributes have some relation to the outcome variable.

3.1 Methods

As mentioned before, the data set used in the experiments contains 348 findings that include data related

to biopsies. A subset of 180 was annotated by a specialist blind to the biopsies results. The task of this specialist was to annotate the mass density. The remaining findings, 168 cases, were not annotated by this specialist.

All experiments were performed using the WEKA tool, developed at Waikato University, New Zealand (Hall et al., 2009). We experimented with several classification algorithms, but report only for the algorithms that produced the best results. The experiments were performed in WEKA using the Experimenter module, where we set several parameters, including the statistical significance test and confidence interval, and the algorithms we wanted to use (we used OneR as reference, ZeroR, PART, J48, SimpleCart, DecisionStump, Random Forests, SMO, Naive Bayes, Bayes with TAN, NBTree and DTNB). The WEKA experimenter produces a table with the performance metrics of all algorithms with an indication of statistical differences, using one of the algorithms as a reference. The significance tests were performed using standard corrected t-test with a significance level of 0.01. The parameters used for the learning algorithms are the WEKA defaults. In the tables, the numbers between parentheses represent standard deviations. From the 348 cases, we trained on the 180 annotated cases. We used the remaining 168 as unseen/test data to evaluate the performance of the classifiers. During the training, we used 10-fold stratified cross validation and reported the results for the average metrics obtained among all folds.

3.2 Is mass density predictive of malignancy?

We considered at least two ways of investigating if mass density is predictive of malignancy. The first one is to apply association rules or logistic regression to the 348 findings, and report the relation between retro_density and outcome. This was already done by Woods and Burnside (Woods and Burnside, 2010), in a previous work, using logistic regression and kappa statistics. Their results showed that high mass density is a relatively important indicator of malignancy with an inter-observer agreement of 0.53.

The second way is to use a classification method and predict outcome using mass density and without using mass density and compare results. As we have two kinds of mass density: one for the retrospective data and another one for the prospective data, we used both to build classifiers. Our first experiment was then to generate a classifier to predict outcome with retro_density using 10-fold cross-validation on the 180 findings. Our second experiment was to generate a classifier to predict outcome with density_num (prospectively assessed), also using 10-fold cross-validation on the 180 findings.

In order to investigate if mass density is predictive of malignancy, we also generated a classifier to predict outcome without any information about density using 10-fold cross-validation on the 180 findings.

In the three experiments, the best classifiers found were based on Support Vector Machines (Platt, 1998). Table 1 summarizes the results obtained using the metrics we found more relevant to the task. CCI is the percentage of Correctly Classified Instances. K is the k-value of kappa statistics. Prec is the Precision, and F is the F-measure. These results show that mass density has some influence on the outcome, specially when mass density is the one observed on the retrospective data. The classifier trained without mass density has an overall performance of 81.39% while the classifier trained with the retrospective assessed mass has an overall performance of 84.78%. These results are statistically different ($p=0.01$). If we look at the K value, we can confirm that the relation between mass density and outcome is not by chance, given the relatively high observed agreement between the real data and the classifier's predicted values. With respect to Precision, the results also seem to be quite good with only 16% of cases being incorrectly classified as malignant when using the retrospective data. The Recall also gives a reasonable rate of correctly classified cases of malignancy, although there is still scope for improvement. The f-measure balances the values of Precision and Recall and also indicates that the classifiers are behaving reasonably well.

Summarising, these results show that attributes other than mass density are also important, but if we add mass density, the classifier's performance improves.

These results also confirm findings in the literature regarding the relevance of mass density, and show that good classifiers can be obtained to predict outcome (with a high percentage of correctly classified instances and good values of K, precision and recall).

3.3 Can we obtain a classifier that predicts mass density as well as the radiologist?

Our second question is related to the quality of the classifier related to a specialist. As we have two annotated mass densities, one for the prospective study and another one for the retrospective, we generated 2 classifiers: one is trained on the prospective values of mass density (density_num), and another one is trained on the retrospective (retro_density) values

Table 1: Prediction of outcome using 180 findings. Standard deviation values are between parentheses.

Metric	with mass density		without mass density
	retro_density	density_num	
CCI	84.78% (7.96)	82.72% (8.32)	81.39% (8.81)
K	0.68 (0.17)	0.63 (0.17)	0.60 (0.18)
Prec	0.84 (0.12)	0.82 (0.13)	0.81 (0.14)
Recall	0.78 (0.15)	0.75 (0.15)	0.72 (0.15)
F	0.80 (0.11)	0.77 (0.11)	0.75 (0.12)

of mass density. Once more, we used the 180 cases as training set and 10-fold stratified cross-validation. The best classifier obtained by the WEKA Experimenter for these two tasks was based on Naive-Bayes (John and Langley, 1995). Table 2 shows the results of these experiments as an average of the metrics for the 10 folds.

Table 2: Prediction of mass density using 180 findings. Standard deviation values are between parentheses.

Metric	retro_density	density_num
CCI	72.83% (9.89)	67.22% (12.14)
K	0.37 (0.23)	0.33 (0.25)
Precision	0.58 (0.20)	0.66 (0.16)
Recall	0.58 (0.22)	0.60 (0.17)
F-Measure	0.56 (0.18)	0.62 (0.15)

70% of masses annotated by the specialist on the 180 findings agreed to the annotated masses of the retrospective study. The Naive Bayes classifier predicted $\approx$ 73% of correct instances when training on the retrospective annotated mass (retro_density) and $\approx$ 67% when training on prospective masses annotated by a radiologist. These results are quite good and indicate that the Bayesian classifier generated in this study can be well applied as a support tool to help doctors predicting mass density for unseen mammograms. The values of K, Precision, Recall and f-measure for this experiment are not so good as the ones obtained when trying to learn outcome. However, the K value indicates that the Naive Bayes classifier has some level of agreement with the actual data, which is not by chance. One interesting thing to observe is that, although the classifier trained on the retrospective data has a higher rate of correctly classified instances, it has lower values for Precision, Recall and f-measure than the classifier trained on the prospective data. This may indicate that this could be a better classifier to be used when one does not have information about the biopsy data.

Our last question is related to how well a learned classifier can predict the outcome (malignant or benign) on unseen data blind to the result of the biopsy.

3.4 Can the generated classifiers behave well on unseen data?

In order to answer this question we need again to consider classifiers generated using the retrospective mass density attribute and the prospective mass density attribute. The first classifier, based on the retrospective values of mass density was generated when training on the 180 findings to answer our first question: "is mass density related to malignancy?" This is a classifier based on Support Vector Machines. However, we can use yet another classifier, based on the prospective values of mass density to predict the 168 unseen cases. As the 168 unseen cases do not have any prospective annotated mass density, we will fill up these missing values using the classifiers generated when answering our question 2 (Subsection 3.3). In those experiments, we generated two classifiers to predict mass density: one that was trained on retro_density and another one that was trained on density_num. Both are Bayesian classifiers. Once we fill up these values, we can apply a classifier learned to predict outcome to this unseen data set.

Results of the prediction of mass density on the unseen data are shown in Table 3. These results were produced by the best classifier that was, in both cases, a naive Bayes network.

Table 3: Prediction of mass_density on unseen data.

Metric	retro_density	density_num
CCI	82.14%	75.60%
K	0.45	0.35
Prec	0.48	0.38
Recall	0.68	0.71
F	0.56	0.49

These results are very good, given that both classifiers have a prediction performance on the unseen data well above the one obtained on the training set (180 cases) with respect to CCI. The K-statistics and the Recall also improved on the unseen data. We see a slightly fall in performance when predicting benign

cases, and this is observed by the precision and f-measure values in the unseen data. The rate of false positives increases on the unseen data. On the other hand, the algorithm performs better on classifying the malignant cases.

Once the predicted values of mass densities of the 168 findings are filled, we move to the next step, which is to predict outcome for the unseen data. Results of this experiment can be found in Table 4.

In Table 4 we show three different predictions for outcome, using three different sources for the mass density. The second column in Table 4 shows the results of predicting outcome using the attribute for mass density available on the retrospective data (retro_density attribute). The third and fourth columns show the predictions when using the mass density filled up by the two Naive Bayes classifiers (one that was trained on the retro_density attribute and another that was trained on the prospective density_num attribute).

Regarding the comparison among these three predictions we can observe that the three classifiers behaved relatively well on the unseen data, capturing most of the malignant and benign cases. The K value, once more, indicates that those results are not by chance. In other words, the classifiers are actually helping to distinguish between malignant and benign cases. As observed before, the classifier trained on the actual retrospective data yields better performance, but the other classifiers are not performing that far, which indicates that the lack of biopsy data is not harming the classification task.

A second observation we take from these results is that, even using predicted values for mass density (with prediction errors), the classifiers for outcome in columns three and four, can maintain a reasonable performance.

The last conclusion we take from these results is that mass density is somehow related to outcome, and is an important attribute that contributes to improve the performance of the classifiers. A comparison between the figures on the last column of Table 4 (prediction without mass density) with the figures on the other columns confirms that fact.

Summarizing, and getting back to our third question "3. Can we obtain a classifier that predicts mass_density as well as the radiologist?", Table 5 shows the performance of all classifiers used for this task on the training data and on unseen data.

Table 5 summarizes our results for predicting mass density and shows that the classifiers generated have a good performance that in some cases is better than the one given by the radiologist. The performance on unseen cases is also quite reasonable regarding the precision and recall values.

4 CONCLUSIONS AND FUTURE WORK

In this work, we were provided with 348 cases of patients that went through mammography screening. The objective of this work was twofold: i) find non trivial relations among attributes by applying machine learning techniques to these data, and; ii) learn models that could help medical doctors to quickly assess mammograms. We used the WEKA machine learning tool and whenever applicable performed statistical tests of significance on the results.

The conclusions are threefold: (1) automatic classification of a mammography can reach equal or better results than the ones annotated by specialists; (2) mass density seems to be a good indicator of malignancy, as previous studies suggested; (3) machine learning classifiers can predict mass density with a quality as good as the specialist blind to biopsy.

As future work, we plan to extend this work to larger data sets, and apply other machine learning techniques based on statistical relational learning, since classifiers that fall in this category provide a good explanation of the predicted outcomes as well as can consider the relationship among mammograms of the same patient. We would also like to investigate how other attributes can affect malignancy or are related to the other attributes.

ACKNOWLEDGMENTS

This work has been partially supported by the projects HORUS (PTDC/EIA-EIA/100897/2008) and Digiscope (PTDC/EIA-CCO/100844/2008) and by the Fundação para a Ciência e Tecnologia (FCT/Portugal). Pedro Ferreira has been supported by an FCT BIC scholarship.

REFERENCES

Abbass, H. A. (2002). An evolutionary artificial neural networks approach for breast cancer diagnosis. *Artificial Intelligence in Medicine*, 25:265.

Ayer, T., Alagoz, O., Chhatwal, J., Shavlik, J. W., Kahn, C. E. J., and Burnside, E. S. (2010). Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration. *Cancer*, 116(14):3310–3321.

Table 4: Prediction of outcome on unseen data.

Metric	with mass density			w/o mass density
	retro_density (actual)	retro_density (fill up by NB)	density_num (fill up by NB)	
CCI	81.55%	79.76%	79.17%	77.38%
K	0.52	0.48	0.46	0.42
Prec	0.70	0.65	0.65	0.61
Recall	0.60	0.60	0.55	0.53
F	0.64	0.62	0.60	0.57

Table 5: Prediction of mass density.

Metric	Mass Density				
	Radiologist (180)	density_num (180)	density_num (168)	retro_density (180)	retro_density (168)
CCI	70.00%	67.22% (12.14)	75.60%	72.83% (9.89)	82.14%
K	–	0.33 (0.25)	0.35	0.37 (0.23)	0.45
Prec	–	0.66 (0.16)	0.38	0.58 (0.20)	0.48
Recall	–	0.60 (0.17)	0.71	0.58 (0.22)	0.68
F	–	0.62 (0.15)	0.49	0.56 (0.18)	0.56

Cory, R. C. and Linden, S. S. (1993). The mammographic density of breast cancer. *AJR Am J Roentgenol*, 160:418–419.

Davis, J., Burnside, E. S., Dutra, I. C., Page, D., and Costa, V. S. (2005). Knowledge discovery from structured mammography reports using inductive logic programming. In *American Medical Informatics Association 2005 Annual Symposium*, pages 86–100.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explorations*, 11:263–286.

Jackson, V. P., Dines, K. A., Bassett, L. W., Gold, R. H., and Reynolds, H. E. (1991). Diagnostic importance of the radiographic density of noncalcified breast masses: analysis of 91 lesions. *AJR Am J Roentgenol*, 157:25–28.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufmann, San Mateo.

Nassif, H., Page, D., Ayvaci, M., Shavlik, J., and Burnside, E. S. (2010). Uncovering age-specific invasive and dcis breast cancer rules using inductive logic programming. In *Proceedings of 2010 ACM International Health Informatics Symposium (IHI 2010)*. ACM Digital Library.

Nassif, H., Woods, R., Burnside, E., Ayvaci, M., Shavlik, J., and Page, D. (2009). Information extraction for clinical data mining: A mammography case study. In *ICDMW '09: Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, pages 37–42, Washington, DC, USA. IEEE Computer Society.

Platt, J. C. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report MSR-TR-98-14, Microsoft Research.

Sickles, E. A. (1991). Periodic mammographic follow-up of probably benign lesions: results in 3,184 consecutive cases. *Radiology*, 179:463–468.

Street, W. N., Mangasarian, O. L., and Wolberg, W. H. (1995). An inductive learning approach to prognostic prediction. In *ICML*, page 522.

Wolberg, W. H. and Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In *Proceedings of the National Academy of Sciences, 87*, pages 9193–9196.

Woods, R. and Burnside, E. (2010). The mammographic density of a mass is a significant predictor of breast cancer. *Radiology*. to appear.

Woods, R., Oliphant, L., Shinki, K., Page, D., Shavlik, J., and Burnside, E. (2009). Validation of results from knowledge discovery: Mass density as a predictor of breast cancer. *J Digit Imaging*, pages 418–419.

Wu, Y., Giger, M. L., Doi, K., Vyborny, C. J., Schmidt, R. A., and Metz, C. E. (1993). Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. *Radiology*, 187:81–87.

Apêndice B

Atributos descartados

Atributos Descartados	Motivos da Não Utilização
MRN_scrubbed	Atributo descartado por se tratar de um identificador de registo.
PATIENT_SEX	Atributo descartado por incidir apenas sobre uma mesma classe (Feminina).
rnd_num	Atributo descartado por se tratar de um identificador de registo.
biopsy_date	Atributo descartado, uma vez que as datas apenas situam um determinado acontecimento no tempo, não acrescentando qualquer tipo de informação importante aos dados.
ID_MATCH_NMD	Atributo descartado por se tratar de um identificador de registo.
ASSESSMENT	Atributo descartado, uma vez que poderia influenciar em demasia os resultados obtidos.
PENRAD_MAMMO_ID	Atributo descartado por se tratar de um identificador de registo.
MAMMO_STUDY_DATE	Atributo descartado, uma vez que as datas apenas situam um determinado acontecimento no tempo, não acrescentando qualquer tipo de informação importante aos dados.
PENRAD_ABNORMALITY_ID	Atributo descartado por se tratar de um identificador de registo.
MASS_SHAPE_def	Atributo descartado por se tratar de um atributo duplicado com *MASS_SHAPE* (atributo utilizado).
MASS_MARGINS_def	Atributo descartado por se tratar de um atributo duplicado com *MASS_MARGINS* (atributo utilizado).
ARCHITECTURAL_DISTORTION_def	Atributo descartado por possuir apenas quatro instâncias com valor definido (*Yes*) e todas iguais.
CLOCKFACE_def	Atributo descartado por se tratar de um atributo duplicado com *CLOCKFACE_LOCATION_OR_REGION* (atributo utilizado).
SIDE_def	Atributo descartado por se tratar de um atributo duplicado com *SIDE* (atributo utilizado).
DEPTH_def	Atributo descartado por se tratar de um atributo duplicado com *DEPTH* (atributo utilizado).
lb_finding	Atributo descartado, uma vez que está directamente relacionado com malignidade e como tal poderia influenciar em demasia os resultados obtidos.
digital_sub	Atributo descartado pelo facto de todas as instâncias não apresentarem qualquer tipo de valor definido.
digital	Atributo descartado por considerarmos que a informação relativa à técnica utilizada não é importante para este estudo.

lb_technique	Atributo descartado por considerarmos que a informação relativa ao tipo de biópsia aplicada não é importante para este estudo.
REASON_FOR_THIS_MAMMOGRAM	Atributo descartado pelo facto de todas as instâncias incidirem sobre um mesmo valor (V).
FUmonths	Atributo descartado por considerarmos que o número de meses em que um determinado paciente foi alvo de acompanhamento médico não é importante para o estudo em questão.

Tabela 15 - Conjunto de atributos descartados com respectivo motivo pelo qual não foram utilizados

Apêndice C

Experiências

Experiências

Aprendizagem - 180 casos

(E_1) Previsão de malignidade (*outcome_num*) com densidade de massa retrospectiva (*retro_density*);

(E_2) Previsão de malignidade (*outcome_num*) com densidade de massa prospectiva (*density_num*);

(E_3) Previsão de malignidade (*outcome_num*) sem densidade de massa;

(E_4) Previsão de densidade de massa retrospectiva (*retro_density*);

(E_5) Previsão de densidade de massa prospectiva (*density_num*);

Teste - 168 casos

(E_6) Previsão de densidade de massa retrospectiva (*retro_density*);

(E_7) Previsão de densidade de massa prospectiva (*density_num*);

(E_8) Previsão de malignidade (*outcome_num*) com densidade de massa retrospectiva (*retro_density*);

(E_9) Previsão de malignidade (*outcome_num*) com densidade de massa retrospectiva (*retro_density*) prevista em E_6;

(E_{10}) Previsão de malignidade (*outcome_num*) com densidade de massa prospectiva (*density_num*) prevista em E_7;

(E_{11}) Previsão de malignidade (*outcome_num*) sem densidade de massa;

180 casos

(E_1) PREVISÃO DE *outcome_num* COM *retro_density*

	SMO	Naive Bayes	DTNB	Bayes Net (TAN)	PART (1)	NB Tree (1)	Simple Cart (1)	Random Forest	J48 (1)	OneR	Decision Stump	ZeroR
Correctly Classified Instances	84.78% (7.96)	81.33% (9.46)	81.06% (9.08)	80.11% (8.91)	79.50% (8.85)	78.33% (9.33)	77.67% (9.77)	76.39% (9.52)	76.39% (9.68)*	71.61% (10.08)*	67.78% (8.09)*	60.56% (1.68)*
Kappa Statistic	0.68 (0.17)	0.61 (0.20)	0.59 (0.20)	0.58 (0.19)	0.57 (0.19)	0.54 (0.19)	0.52 (0.21)	0.49 (0.21)*	0.49 (0.22)*	0.39 (0.22)*	0.27 (0.17)*	0.00 (0.00)*
Precision (2)	0.84 (0.12)	0.79 (0.14)	0.81 (0.14)	0.79 (0.14)	0.78 (0.14)	0.75 (0.14)	0.76 (0.15)	0.75 (0.15)	0.75 (0.18)	0.67 (0.17)*	0.71 (0.26)	0.00 (0.00)*
F-Measure (2)	0.80 (0.11)	0.76 (0.12)	0.74 (0.13)	0.73 (0.13)	0.73 (0.13)	0.72 (0.12)	0.69 (0.16)	0.66 (0.15)*	0.66 (0.16)	0.61 (0.16)*	0.47 (0.17)*	0.00 (0.00)*
TP Rate (2)	0.78 (0.15)	0.75 (0.16)	0.70 (0.18)	0.70 (0.16)	0.71 (0.17)	0.71 (0.16)	0.66 (0.21)	0.61 (0.18)*	0.62 (0.20)	0.58 (0.20)*	0.39 (0.18)*	0.00 (0.00)*
** Confusion Matrix	a b 55 16 | a = mal. 9 100 | b = ben. CCI = 155	a b 53 18 | a = mal. 17 92 | b = ben. CCI = 145	a b 47 24 | a = mal. 12 97 | b = ben. CCI = 144	a b 47 24 | a = mal. 12 97 | b = ben. CCI = 144	a b 49 22 | a = mal. 14 95 | b = ben. CCI = 144	a b 46 25 | a = mal. 19 90 | b = ben. CCI = 136	a b 47 24 | a = mal. 13 96 | b = ben. CCI = 143	a b 44 27 | a = mal. 14 95 | b = ben. CCI = 139	a b 41 30 | a = mal. 14 95 | b = ben. CCI = 136	a b 42 29 | a = mal. 22 87 | b = ben. CCI = 129	a b 28 43 | a = mal. 17 92 | b = ben. CCI = 120	a b 0 71 | a = mal. 0 109 | b = ben. CCI = 109

↓
TP FN
FP TN

	SMO	Bayes Net (K2)	PART (10) (1)	Simple Cart (10) (1)	J48 (10) (1)							
Correctly Classified Instances	84.78% (7.96)	78.39% (9.17)	79.50% (8.85)	77.67% (9.61)	76.39% (9.68)*							
Kappa Statistic	0.68 (0.17)	0.54 (0.20)	0.57 (0.19)	0.52 (0.20)	0.49 (0.22)*							
Precision (2)	0.84 (0.12)	0.76 (0.14)	0.78 (0.14)	0.75 (0.15)	0.75 (0.18)							
F-Measure (2)	0.80 (0.11)	0.71 (0.13)	0.73 (0.13)	0.70 (0.14)	0.66 (0.16)							
TP Rate (2)	0.78 (0.15)	0.68 (0.16)	0.71 (0.17)	0.68 (0.18)	0.62 (0.20)							
** Confusion Matrix	a b 55 16 | a = mal. 9 100 | b = ben. CCI = 155	a b 48 23 | a = mal. 17 92 | b = ben. CCI = 140	a b 49 22 | a = mal. 14 95 | b = ben. CCI = 144	a b 49 22 | a = mal. 16 93 | b = ben. CCI = 142	a b 41 30 | a = mal. 14 95 | b = ben. CCI = 136							

↓
TP FN
FP TN

NOTA:

Todos os algoritmos encontram-se ordenados da esquerda para a direita de forma decrescente de resultados, ou seja, a coluna mais à esquerda relativa ao classificador resultante recorrendo ao algoritmo **SMO** apresenta os melhores resultados comparativamente com os restantes classificadores (tendo em conta, acima de tudo, os valores de **"Correctly Classified Instances"**, **"Kappa Statistic"** e **"F-Measure"**);

Todos os valores que se encontram entre parêntesis representam desvios-padrão;

* Valor de **"Paired Corrected T-Tester"** significativo para **0.01**;

(2) Os valores relativos às métricas: **"Precision"**, **"F-Measure"** e **"TP Rate"** dizem respeito à classe **"malignant"**;

(1) Os classificadores resultantes recorrendo aos algoritmos **PART**, **NBTree**, **SimpleCart** e **J48** apesar de não apresentarem os índices mais elevados em termos de **"Correctly Classified Instances"**, **"Kappa Statistic"** e **"F-Measure"**, geraram regras/árvores interessantes (ver págs. anexas). Conclusões tiradas depois da análise dos respectivos **"Classifiers outputs"** no **WEKA Explorer**;

** **"Confusion Matrix"** - Dados obtidos depois de gerados os **"Classifiers outputs"** para cada um dos classificadores no **WEKA Explorer**.
CCI – Número de **"Correctly Classified Instances"**;

Os **valores relativos à tabela da pág. 2 são valores de teste**, uma vez que foi estudado o comportamento de certos classificadores que recorrem a algoritmos com **parâmetros diferentes** dos presentes na tabela da pág. 1. Sendo assim, temos:

Tabela pág. 1:

BayesNet – "searchAlgorithm" (TAN)
PART – "numFolds" (3 – "default value")
SimpleCart – "numFoldsPruning" (5 – "default value")
J48 – "numFolds" (3 – "default value")

Tabela pág. 2:

BayesNet – "searchAlgorithm" (K2)
PART – "numFolds" (10)
SimpleCart – "numFoldsPruning" (10)
J48 – "numFolds" (10)

PART:

```
=== Classifier model (full training set) ===

PART decision list
------------------

retro_density = iso AND
MASS_MARGINS_2 = D AND
age_at_mammo <= 59: benign (43.31/0.45)

retro_density = high AND
MASS_MARGINS_2 = I: malignant (17.92/1.64)

MASS_MARGINS_2 = S: malignant (29.99/4.41)

age_at_mammo <= 62 AND
reread_group = burnside: benign (16.39)

retro_density = high AND
CLOCKFACE_LOCATION_OR_REGION = C: malignant (3.51/1.0)

reread_group = sisney AND
OVERALL_BREAST_COMPOSITION = heterogeneously dense: benign (13.33/2.51)

reread_group = sisney AND
OVERALL_BREAST_COMPOSITION = almost entirely fat: benign (3.75)

age_at_mammo <= 62 AND
CLOCKFACE_LOCATION_OR_REGION = 12.0: benign (5.67)

retro_density = iso AND
SIZE <= 13 AND
DEPTH = A: benign (10.16/3.0)

reread_group = burnside: malignant (10.75)

OVERALL_BREAST_COMPOSITION = heterogeneously dense AND
MASS_SHAPE = X: benign (4.23/1.41)

MASS_SHAPE = R: benign (6.35/0.3)

: malignant (14.65/6.44)

Number of Rules  :      13
```

NBTree:

```
=== Classifier model (full training set) ===

NBTree
------------------

SIDE = R
|   MASS_SHAPE = X
|   |   MASS_MARGINS_2 = I: NB 3
|   |   MASS_MARGINS_2 = S: NB 4
|   |   MASS_MARGINS_2 = D: NB 5
|   |   MASS_MARGINS_2 = M: NB 6
|   |   MASS_MARGINS_2 = U: NB 7
|   MASS_SHAPE = R: NB 8
|   MASS_SHAPE = O
|   |   age_at_mammo <= 62.5: NB 10
|   |   age_at_mammo > 62.5: NB 11
|   MASS_SHAPE = L: NB 12
SIDE = L
|   QUADRANT_LOCATION_def = Upper Outer: NB 14
|   QUADRANT_LOCATION_def = Upper Inner: NB 15
|   QUADRANT_LOCATION_def = Lower Inner: NB 16
|   QUADRANT_LOCATION_def = Lower Outer: NB 17
```

SimpleCart:

```
=== Classifier model (full training set) ===

CART Decision Tree

retro_density=(iso)
|  MASS_SHAPE=(R)|(L)|(O)
|  |  age_at_mammo < 73.5: benign(79.76/5.75)
|  |  age_at_mammo >= 73.5: malignant(5.75/2.0)
|  MASS_SHAPE!=(R)|(L)|(O)
|  |  CLOCKFACE_LOCATION_OR_REGION=(7.0)|(6.0)|(5.0)|(8.0): benign(5.49/0.0)
|  |  CLOCKFACE_LOCATION_OR_REGION!=(7.0)|(6.0)|(5.0)|(8.0)
|  |  |  SIZE < 14.5: malignant(15.49/2.48)
|  |  |  SIZE >= 14.5
|  |  |  |  SIZE < 19.5: benign(5.0/0.0)
|  |  |  |  SIZE >= 19.5: malignant(2.0/0.24)
retro_density!=(iso)
|  MASS_MARGINS_2=(D)|(U)
|  |  age_at_mammo < 65.0: benign(11.36/3.71)
|  |  age_at_mammo >= 65.0: malignant(4.72/0.35)
|  MASS_MARGINS_2!=(D)|(U): malignant(33.56/2.28)

Number of Leaf Nodes: 9

Size of the Tree: 17
```

J48:

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

retro_density = high
|   MASS_MARGINS_2 = I: malignant (17.92/1.64)
|   MASS_MARGINS_2 = S: malignant (15.68/0.56)
|   MASS_MARGINS_2 = D
|   |   OVERALL_BREAST_COMPOSITION = scattered fibroglandular densities: benign (2.84/0.56)
|   |   OVERALL_BREAST_COMPOSITION = almost entirely fat: malignant (5.28/1.0)
|   |   OVERALL_BREAST_COMPOSITION = heterogeneously dense: benign (7.56/1.28)
|   |   OVERALL_BREAST_COMPOSITION = extremely dense: benign (0.0)
|   MASS_MARGINS_2 = M: malignant (2.24/0.08)
|   MASS_MARGINS_2 = U
|   |   SIZE <= 25: benign (2.32/0.16)
|   |   SIZE > 25: malignant (2.16)
retro_density = iso
|   MASS_MARGINS_2 = I
|   |   age_at_mammo <= 64: benign (20.83/3.45)
|   |   age_at_mammo > 64: malignant (7.13/1.68)
|   MASS_MARGINS_2 = S: malignant (13.37/2.94)
|   MASS_MARGINS_2 = D: benign (55.92/4.8)
|   MASS_MARGINS_2 = M: benign (2.43/0.08)
|   MASS_MARGINS_2 = U: benign (24.31/4.78)

Number of Leaves  :     15

Size of the tree :      21
```

SimpleCart(10):

```
=== Classifier model (full training set) ===

CART Decision Tree

retro_density=(iso)
|  MASS_SHAPE=(R)|(L)|(O)
|  |  age_at_mammo < 73.5: benign(79.76/5.75)
|  |  age_at_mammo >= 73.5: malignant(5.75/2.0)
|  MASS_SHAPE!=(R)|(L)|(O)
|  |  CLOCKFACE_LOCATION_OR_REGION=(7.0)|(6.0)|(5.0)|(8.0): benign(5.49/0.0)
|  |  CLOCKFACE_LOCATION_OR_REGION!=(7.0)|(6.0)|(5.0)|(8.0)
|  |  |  SIZE < 14.5: malignant(15.49/2.48)
|  |  |  SIZE >= 14.5: benign(5.24/2.0)
retro_density!=(iso)
|  MASS_MARGINS_2=(D)|(U)
|  |  age_at_mammo < 65.0: benign(11.36/3.71)
|  |  age_at_mammo >= 65.0: malignant(4.72/0.35)
|  MASS_MARGINS_2!=(D)|(U): malignant(33.56/2.28)

Number of Leaf Nodes: 8

Size of the Tree: 15
```

180 casos

(E_2) PREVISÃO DE *outcome_num* COM *density_num*

	SMO	Naive Bayes	Bayes Net (TAN)	NB Tree (1)	DTNB	PART (1)	J48 (1)	OneR	Simple Cart (1)	Random Forest	Decision Stump	ZeroR
Correctly Classified Instances	82.72% (8.32)	80.33% (9.26)	78.78% (8.66)	77.22% (10.43)	76.94% (10.03)	76.22% (9.21)	75.00% (10.37)	74.28% (10.35)	73.39% (8.91)*	73.33% (9.24)*	67.61% (8.76)*	60.56% (1.68)*
Kappa Statistic	0.63 (0.17)	0.59 (0.19)	0.54 (0.19)	0.52 (0.22)	0.50 (0.22)	0.49 (0.20)	0.46 (0.23)	0.45 (0.22)	0.43 (0.19)*	0.42 (0.21)*	0.27 (0.18)*	0.00 (0.00)*
Precision (2)	0.82 (0.13)	0.77 (0.14)	0.78 (0.14)	0.73 (0.16)	0.74 (0.15)	0.73 (0.17)	0.71 (0.16)	0.70 (0.18)	0.69 (0.15)	0.71 (0.16)	0.76 (0.29)	0.00 (0.00)*
F-Measure (2)	0.77 (0.11)	0.75 (0.12)	0.70 (0.13)	0.70 (0.15)	0.68 (0.15)	0.67 (0.15)	0.65 (0.16)	0.65 (0.16)	0.64 (0.13)*	0.62 (0.15)*	0.46 (0.17)*	0.00 (0.00)*
TP Rate (2)	0.75 (0.15)	0.75 (0.15)	0.66 (0.17)	0.69 (0.18)	0.65 (0.19)	0.65 (0.18)	0.63 (0.20)	0.63 (0.19)	0.62 (0.18)	0.58 (0.19)*	0.38 (0.21)*	0.00 (0.00)*
** **Confusion Matrix**	a b 51 20 | a = mal. 12 97 | b = ben. CCI = 148	a b 54 17 | a = mal. 18 91 | b = ben. CCI = 145	a b 44 27 | a = mal. 14 95 | b = ben. CCI = 139	a b 49 22 | a = mal. 20 89 | b = ben. CCI = 138	a b 48 23 | a = mal. 22 87 | b = ben. CCI = 135	a b 47 24 | a = mal. 19 90 | b = ben. CCI = 137	a b 44 27 | a = mal. 21 88 | b = ben. CCI = 132	a b 47 24 | a = mal. 19 90 | b = ben. CCI = 137	a b 49 22 | a = mal. 24 85 | b = ben. CCI = 134	a b 41 30 | a = mal. 17 92 | b = ben. CCI = 133	a b 29 42 | a = mal. 16 93 | b = ben. CCI = 122	a b 0 71 | a = mal. 0 109 | b = ben. CCI = 109

↓
TP FN
FP TN

	SMO	Bayes Net (K2)	PART (10) (1)	J48 (10) (1)	Simple Cart (10) (1)							
Correctly Classified Instances	82.72% (8.32)	79.67% (9.11)	76.22% (9.21)	75.00% (10.37)	73.72% (9.14)*							
Kappa Statistic	0.63 (0.17)	0.56 (0.20)	0.49 (0.20)	0.46 (0.23)	0.44 (0.19)*							
Precision (2)	0.82 (0.13)	0.78 (0.14)	0.73 (0.17)	0.71 (0.16)	0.69 (0.14)							
F-Measure (2)	0.77 (0.11)	0.72 (0.14)	0.67 (0.15)	0.65 (0.16)	0.65 (0.14)							
TP Rate (2)	0.75 (0.15)	0.68 (0.17)	0.65 (0.18)	0.63 (0.20)	0.63 (0.17)							
** Confusion Matrix	a b 51 20 | a = mal. 12 97 | b = ben. CCI = 148	a b 49 22 | a = mal. 16 93 | b = ben. CCI = 142	a b 47 24 | a = mal. 19 90 | b = ben. CCI = 137	a b 44 27 | a = mal. 21 88 | b = ben. CCI = 132	a b 46 25 | a = mal. 25 84 | b = ben. CCI = 130							

↓

TP FN

FP TN

NOTA:

Todos os algoritmos encontram-se ordenados da esquerda para a direita de forma decrescente de resultados, ou seja, a coluna mais à esquerda relativa ao classificador resultante recorrendo ao algoritmo **SMO** apresenta os melhores resultados comparativamente com os restantes classificadores (tendo em conta, acima de tudo, os valores de **"Correctly Classified Instances"**, **"Kappa Statistic"** e **"F-Measure"**);

Todos os valores que se encontram entre parêntesis representam desvios-padrão;

* Valor de **"Paired Corrected T-Tester"** significativo para **0.01**;

(2) Os valores relativos às métricas: **"Precision"**, **"F-Measure"** e **"TP Rate"** dizem respeito à classe **"malignant"**;

(1) Os classificadores resultantes recorrendo aos algoritmos **NBTree**, **PART**, **J48** e **SimpleCart** apesar de não apresentarem os índices mais elevados em termos de **"Correctly Classified Instances"**, **"Kappa Statistic"** e **"F-Measure"**, geraram regras/árvores interessantes (ver págs. anexas). Conclusões tiradas depois da análise dos respectivos **"Classifiers outputs"** no **WEKA Explorer**;

** **"Confusion Matrix"** - Dados obtidos depois de gerados os **"Classifiers outputs"** para cada um dos classificadores no **WEKA Explorer**.
CCI – Número de **"Correctly Classified Instances"**;

Os **valores relativos à tabela da pág. 2 são valores de teste**, uma vez que foi estudado o comportamento de certos classificadores que recorrem a algoritmos com **parâmetros diferentes** dos presentes na tabela da pág. 1. Sendo assim, temos:

Tabela pág. 1:

BayesNet – "searchAlgorithm" (TAN)
PART – "numFolds" (3 – "default value")
J48 – "numFolds" (3 – "default value")
SimpleCart – "numFoldsPruning" (5 – "default value")

Tabela pág. 2:

BayesNet – "searchAlgorithm" (K2)
PART – "numFolds" (10)
J48 – "numFolds" (10)
SimpleCart – "numFoldsPruning" (10)

NBTree:

```
=== Classifier model (full training set) ===

NBTree
------------------

SIDE = R
|   DEPTH = P: NB 2
|   DEPTH = M: NB 3
|   DEPTH = A
|   |   SIZE <= 6.5: NB 5
|   |   SIZE > 6.5: NB 6
SIDE = L: NB 7
```

PART:

```
=== Classifier model (full training set) ===

PART decision list
------------------

Density_num = iso: benign (99.0/20.0)

age_at_mammo > 47 AND
MASS_SHAPE = X: malignant (25.33/1.44)

age_at_mammo > 62 AND
reread_group = burnside: malignant (7.0/1.0)

MASS_MARGINS_2 = D AND
age_at_mammo <= 62: benign (17.46/1.98)

SIZE <= 16: benign (17.64/6.56)

: malignant (13.57/1.0)

Number of Rules  :      6
```

J48:

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

Density_num = high
|   age_at_mammo <= 47: benign (24.0/6.0)
|   age_at_mammo > 47: malignant (57.0/12.0)
Density_num = iso: benign (99.0/20.0)

Number of Leaves  :     3

Size of the tree :      5
```

SimpleCart:

```
=== Classifier model (full training set) ===

CART Decision Tree

MASS_MARGINS_2=(D)|(U)
|  age_at_mammo < 62.5: benign(71.84/6.21)
|  age_at_mammo >= 62.5
|  |  CLOCKFACE_LOCATION_OR_REGION=(1.0)|(10.0)|(2.0)|(4.0): benign(7.55/2.55)
|  |  CLOCKFACE_LOCATION_OR_REGION!=(1.0)|(10.0)|(2.0)|(4.0): malignant(9.65/1.65)
MASS_MARGINS_2!=(D)|(U)
|  Density_num=(iso)
|  |  SIDE=(L): benign(14.92/3.44)
|  |  SIDE!=(L): malignant(9.89/4.78)
|  Density_num!=(iso): malignant(39.23/8.23)

Number of Leaf Nodes: 6

Size of the Tree: 11
```

SimpleCart (10):

```
=== Classifier model (full training set) ===

CART Decision Tree

MASS_MARGINS_2=(D)|(U)
|  age_at_mammo < 62.5: benign(71.84/6.21)
|  age_at_mammo >= 62.5
|  |  CLOCKFACE_LOCATION_OR_REGION=(1.0)|(10.0)|(2.0)|(4.0): benign(7.55/2.55)
|  |  CLOCKFACE_LOCATION_OR_REGION!=(1.0)|(10.0)|(2.0)|(4.0): malignant(9.65/1.65)
MASS_MARGINS_2!=(D)|(U)
|  Density_num=(iso)
|  |  SIDE=(L): benign(14.92/3.44)
|  |  SIDE!=(L)
|  |  |  MASS_SHAPE=(R)|(L): benign(2.03/0.0)
|  |  |  MASS_SHAPE!=(R)|(L): malignant(9.89/2.75)
|  Density_num!=(iso): malignant(39.23/8.23)

Number of Leaf Nodes: 7

Size of the Tree: 13
```

180 casos

(E_3) PREVISÃO DE *outcome_num* SEM DENSIDADE DE MASSA

	SMO	Naive Bayes	Simple Cart (1)	Bayes Net (TAN)	DTNB	PART (1)	OneR	NB Tree (1)	Random Forest	J48 (1)	Decision Stump	ZeroR
Correctly Classified Instances	81.39% (8.81)	76.22% (9.90)	75.67% (9.50)	75.67% (10.17)	74.78% (9.78)	74.56% (9.32)	74.28% (10.35)	73.44% (9.69)	72.22% (10.08)*	71.94% (9.72)*	68.56% (8.98)*	60.56% (1.68)*
Kappa Statistic	0.60 (0.18)	0.51 (0.20)	0.48 (0.20)	0.48 (0.22)	0.46 (0.21)	0.45 (0.20)	0.45 (0.22)	0.45 (0.20)	0.40 (0.22)*	0.40 (0.21)*	0.29 (0.19)*	0.00 (0.00)*
Precision (2)	0.81 (0.14)	0.70 (0.14)	0.73 (0.15)	0.73 (0.16)	0.72 (0.16)	0.74 (0.17)	0.70 (0.18)	0.67 (0.15)	0.69 (0.16)	0.69 (0.17)	0.78 (0.29)	0.00 (0.00)*
F-Measure (2)	0.75 (0.12)	0.71 (0.13)	0.68 (0.13)	0.67 (0.15)	0.65 (0.15)	0.64 (0.15)	0.65 (0.16)	0.66 (0.13)	0.61 (0.15)*	0.61 (0.15)*	0.47 (0.18)*	0.00 (0.00)*
TP Rate (2)	0.72 (0.15)	0.74 (0.17)	0.66 (0.17)	0.64 (0.18)	0.61 (0.19)	0.60 (0.18)	0.63 (0.19)	0.67 (0.17)	0.56 (0.18)	0.58 (0.19)	0.39 (0.22)*	0.00 (0.00)*
** Confusion Matrix	a b 49 22 | a = mal. 13 96 | b = ben. CCI = 145	a b 53 18 | a = mal. 24 85 | b = ben. CCI = 138	a b 45 26 | a = mal. 18 91 | b = ben. CCI = 136	a b 44 27 | a = mal. 19 90 | b = ben. CCI = 134	a b 47 24 | a = mal. 24 85 | b = ben. CCI = 132	a b 41 30 | a = mal. 16 93 | b = ben. CCI = 134	a b 47 24 | a = mal. 19 90 | b = ben. CCI = 137	a b 51 20 | a = mal. 23 86 | b = ben. CCI = 137	a b 39 32 | a = mal. 20 89 | b = ben. CCI = 128	a b 38 33 | a = mal. 18 91 | b = ben. CCI = 129	a b 30 41 | a = mal. 15 94 | b = ben. CCI = 124	a b 0 71 | a = mal. 0 109 | b = ben. CCI = 109

↓
TP FN
FP TN

	SMO	Bayes Net (K2)	Simple Cart (10) (1)	PART (10) (1)	J48 (10) (1)							
Correctly Classified Instances	81.39% (8.81)	77.61% (9.59)	75.33% (9.58)	74.56% (9.32)	71.94% (9.72)*							
Kappa Statistic	0.60 (0.18)	0.52 (0.21)	0.48 (0.20)	0.45 (0.20)	0.40 (0.21)*							
Precision (2)	0.81 (0.14)	0.75 (0.15)	0.72 (0.15)	0.74 (0.17)	0.69 (0.17)							
F-Measure (2)	0.75 (0.12)	0.70 (0.14)	0.67 (0.13)	0.64 (0.15)	0.61 (0.15)*							
TP Rate (2)	0.72 (0.15)	0.67 (0.18)	0.65 (0.16)	0.60 (0.18)	0.58 (0.19)							
** Confusion Matrix	a b 49 22 | a = mal. 13 96 | b = ben. CCI = 145	a b 48 23 | a = mal. 19 90 | b = ben. CCI = 138	a b 43 28 | a = mal. 18 91 | b = ben. CCI = 134	a b 41 30 | a = mal. 16 93 | b = ben. CCI = 134	a b 38 33 | a = mal. 18 91 | b = ben. CCI = 129							

↓
TP FN
FP TN

NOTA:

Todos os algoritmos encontram-se ordenados da esquerda para a direita de forma decrescente de resultados, ou seja, a coluna mais à esquerda relativa ao classificador resultante recorrendo ao algoritmo **SMO** apresenta os melhores resultados comparativamente com os restantes classificadores (tendo em conta, acima de tudo, os valores de **"Correctly Classified Instances"**, **"Kappa statistic"** e **"F-Measure"**);

Todos os valores que se encontram entre parêntesis representam desvios-padrão;

* Valor de **"Paired Corrected T-Tester"** significativo para **0.01**;

(2) Os valores relativos às métricas: **"Precision"**, **"F-Measure"** e **"TP Rate"** dizem respeito à classe **"malignant"**;

(1) Os classificadores resultantes recorrendo aos algoritmos **SimpleCart**, **PART**, **NBTree** e **J48** apesar de não apresentarem os índices mais elevados em termos de **"Correctly Classified Instances"**, **"Kappa statistic"** e **"F-Measure"**, geraram regras/árvores interessantes (ver págs. anexas). Conclusões tiradas depois da análise dos respectivos **"Classifiers outputs"** no **WEKA Explorer**;

** **"Confusion Matrix"** - Dados obtidos depois de gerados os **"Classifiers outputs"** para cada um dos classificadores no **WEKA Explorer**.
CCI – Número de **"Correctly Classified Instances"**;

Os **valores relativos à tabela da pág. 2 são valores de teste**, uma vez que foi estudado o comportamento de certos classificadores que recorrem a algoritmos com **parâmetros diferentes** dos presentes na tabela da pág. 1. Sendo assim, temos:

Tabela pág. 1:

BayesNet – "searchAlgorithm" (TAN)
SimpleCart – "numFoldsPruning" (5 – "default value")
PART – "numFolds" (3 – "default value")
J48 – "numFolds" (3 – "default value")

Tabela pág. 2:

BayesNet – "searchAlgorithm" (K2)
SimpleCart – "numFoldsPruning" (10)
PART – "numFolds" (10)
J48 – "numFolds" (10)

SimpleCart:

```
=== Classifier model (full training set) ===

CART Decision Tree

MASS_MARGINS_2=(D)|(U)
|  age_at_mammo < 62.5: benign(71.84/6.21)
|  age_at_mammo >= 62.5
|  |  CLOCKFACE_LOCATION_OR_REGION=(1.0)|(10.0)|(2.0)|(4.0): benign(7.55/2.55)
|  |  CLOCKFACE_LOCATION_OR_REGION!=(1.0)|(10.0)|(2.0)|(4.0): malignant(9.65/1.65)
MASS_MARGINS_2!=(D)|(U)
|  age_at_mammo < 48.5
|  |  MASS_SHAPE=(L)|(O)|(R): benign(11.15/1.44)
|  |  MASS_SHAPE!=(L)|(O)|(R): malignant(5.89/2.42)
|  age_at_mammo >= 48.5
|  |  CLOCKFACE_LOCATION_OR_REGION=(2.0)|(5.0)|(7.0)|(6.0)|(4.0)|(8.0)
|  |  |  age_at_mammo < 68.0: benign(8.23/1.0)
|  |  |  age_at_mammo >= 68.0: malignant(2.89/0.0)
|  |  CLOCKFACE_LOCATION_OR_REGION!=(2.0)|(5.0)|(7.0)|(6.0)|(4.0)|(8.0): malignant(41.34/6.13)

Number of Leaf Nodes: 8

Size of the Tree: 15
```

PART:

```
=== Classifier model (full training set) ===

PART decision list
------------------

MASS_MARGINS_2 = D AND
age_at_mammo <= 62: benign (54.89/2.58)

MASS_MARGINS_2 = S: malignant (29.79/4.26)

age_at_mammo > 71: malignant (16.81/2.52)

OVERALL_BREAST_COMPOSITION = heterogeneously dense AND
MASS_MARGINS_1 = D: benign (6.06/0.14)

OVERALL_BREAST_COMPOSITION = extremely dense: benign (4.92)

MASS_MARGINS_2 = U: benign (22.12/5.81)

QUADRANT_LOCATION_def = Upper Outer AND
MASS_SHAPE = X: malignant (13.29/2.54)

reread_group = burnside: benign (12.89/5.44)

reread_group = sisney: benign (8.22/0.44)

OVERALL_BREAST_COMPOSITION = almost entirely fat: malignant (3.31/0.31)

: benign (7.7/3.02)

Number of Rules  :      11
```

NBTree:

```
=== Classifier model (full training set) ===

NBTree
------------------

DEPTH = P
|   age_at_mammo <= 62.5
|   |   rnd_num <= 0.255: NB 3
|   |   rnd_num > 0.255: NB 4
|   age_at_mammo > 62.5: NB 5
DEPTH = M
|   SIZE <= 24
|   |   OVERALL_BREAST_COMPOSITION = scattered fibroglandular densities: NB 8
|   |   OVERALL_BREAST_COMPOSITION = almost entirely fat: NB 9
|   |   OVERALL_BREAST_COMPOSITION = heterogeneously dense: NB 10
|   |   OVERALL_BREAST_COMPOSITION = extremely dense: NB 11
|   SIZE > 24: NB 12
DEPTH = A: NB 13
```

J48:

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

MASS_MARGINS_2 = I
|   OVERALL_BREAST_COMPOSITION = scattered fibroglandular densities
|   |   QUADRANT_LOCATION_def = Upper Outer: malignant (10.37/2.12)
|   |   QUADRANT_LOCATION_def = Upper Inner: benign (1.22/0.09)
|   |   QUADRANT_LOCATION_def = Lower Inner: benign (2.47/0.69)
|   |   QUADRANT_LOCATION_def = Lower Outer: malignant (0.0)
|   OVERALL_BREAST_COMPOSITION = almost entirely fat: malignant (10.03/2.51)
|   OVERALL_BREAST_COMPOSITION = heterogeneously dense: benign (20.59/8.51)
|   OVERALL_BREAST_COMPOSITION = extremely dense: benign (1.51)
MASS_MARGINS_2 = S: malignant (29.61/4.29)
MASS_MARGINS_2 = D
|   age_at_mammo <= 62: benign (54.89/2.58)
|   age_at_mammo > 62
|   |   OVERALL_BREAST_COMPOSITION = scattered fibroglandular densities
|   |   |   SIDE = R: benign (3.39/1.0)
|   |   |   SIDE = L: malignant (2.18/0.39)
|   |   OVERALL_BREAST_COMPOSITION = almost entirely fat: malignant (7.18/1.39)
|   |   OVERALL_BREAST_COMPOSITION = heterogeneously dense: benign (3.39)
|   |   OVERALL_BREAST_COMPOSITION = extremely dense: malignant (0.0)
MASS_MARGINS_2 = M
|   age_at_mammo <= 76: benign (2.66/0.16)
|   age_at_mammo > 76: malignant (2.08/0.03)
MASS_MARGINS_2 = U
|   SIZE <= 25: benign (25.95/4.95)
|   SIZE > 25: malignant (2.47/0.16)

Number of Leaves  :     18

Size of the tree :      26
```

SimpleCart(10):

```
=== Classifier model (full training set) ===

CART Decision Tree

MASS_MARGINS_2=(D)|(U)
|  age_at_mammo < 62.5: benign(71.84/6.21)
|  age_at_mammo >= 62.5
|  |  CLOCKFACE_LOCATION_OR_REGION=(1.0)|(10.0)|(2.0)|(4.0)
|  |  |  DEPTH=(A)|(M): benign(7.55/1.0)
|  |  |  DEPTH!=(A)|(M): malignant(1.55/0.0)
|  |  CLOCKFACE_LOCATION_OR_REGION!=(1.0)|(10.0)|(2.0)|(4.0): malignant(9.65/1.65)
MASS_MARGINS_2!=(D)|(U)
|  age_at_mammo < 48.5
|  |  MASS_SHAPE=(L)|(O)|(R)
|  |  |  CLOCKFACE_LOCATION_OR_REGION=(11.0)|(12.0)|(C)|(7.0)|(1.0)|(10.0)|(3.0)|(2.0)|(4.0)|(5.0)|(8.0)|(9.0): benign(11.15/0.0)
|  |  |  CLOCKFACE_LOCATION_OR_REGION!=(11.0)|(12.0)|(C)|(7.0)|(1.0)|(10.0)|(3.0)|(2.0)|(4.0)|(5.0)|(8.0)|(9.0): malignant(1.44/0.0)
|  |  MASS_SHAPE!=(L)|(O)|(R): malignant(5.89/2.42)
|  age_at_mammo >= 48.5
|  |  CLOCKFACE_LOCATION_OR_REGION=(2.0)|(5.0)|(7.0)|(6.0)|(4.0)|(8.0)
|  |  |  age_at_mammo < 68.0: benign(8.23/1.0)
|  |  |  age_at_mammo >= 68.0: malignant(2.89/0.0)
|  |  CLOCKFACE_LOCATION_OR_REGION!=(2.0)|(5.0)|(7.0)|(6.0)|(4.0)|(8.0)
|  |  |  CLOCKFACE_LOCATION_OR_REGION=(C)|(12.0)
|  |  |  |  SIZE < 8.5: benign(2.89/0.0)
|  |  |  |  SIZE >= 8.5: malignant(10.44/1.78)
|  |  |  CLOCKFACE_LOCATION_OR_REGION!=(C)|(12.0): malignant(30.89/1.44)

Number of Leaf Nodes: 12

Size of the Tree: 23
```

180 casos

(E_4) PREVISÃO DE *retro_density*

	Naive Bayes	OneR	J48 (1)	Simple Cart (1)	DTNB	Decision Stump	Random Forest	SMO	PART (1)	Bayes Net (TAN)	NB Tree (1)
Correctly Classified Instances	72.83% (9.89)	77.78% (7.28)	74.44% (8.79)	73.11% (9.69)	72.06% (8.95)	71.78% (9.13)	71.50% (8.46)	71.28% (8.79)	70.78% (9.20)	70.61% (9.29)	68.33% (10.19)
Kappa Statistic	0.37 (0.23)	0.41 (0.20)	0.32 (0.24)	0.33 (0.23)	0.32 (0.22)	0.36 (0.20)	0.26 (0.22)	0.25 (0.24)	0.30 (0.20)	0.26 (0.24)	0.24 (0.23)
Precision (2)	0.58 (0.20)	0.75 (0.25)	0.67 (0.29)	0.61 (0.26)	0.58 (0.22)	0.57 (0.18)	0.58 (0.28)	0.58 (0.29)	0.59 (0.21)	0.54 (0.25)	0.52 (0.22)
F-Measure (2)	0.56 (0.18)	0.54 (0.19)	0.47 (0.21)	0.49 (0.20)	0.50 (0.18)	0.56 (0.15)	0.42 (0.20)	0.41 (0.21)	0.49 (0.15)	0.44 (0.20)	0.46 (0.17)
TP Rate (2)	0.58 (0.22)	0.45 (0.20)	0.39 (0.21)	0.45 (0.21)	0.48 (0.20)	0.60 (0.21)	0.36 (0.20)	0.35 (0.21)*	0.47 (0.19)	0.41 (0.22)	0.45 (0.20)
** Confusion Matrix	a b 33 23 | a = high 24 100 | b = iso CCI = 133	a b 26 30 | a = high 8 116 | b = iso CCI = 142	a b 24 32 | a = high 13 111 | b = iso CCI = 135	a b 23 33 | a = high 20 104 | b = iso CCI = 127	a b 29 27 | a = high 24 100 | b = iso CCI = 129	a b 35 21 | a = high 30 94 | b = iso CCI = 129	a b 18 38 | a = high 14 110 | b = iso CCI = 128	a b 19 37 | a = high 15 109 | b = iso CCI = 128	a b 27 29 | a = high 22 102 | b = iso CCI = 129	a b 21 35 | a = high 23 101 | b = iso CCI = 122	a b 30 26 | a = high 28 96 | b = iso CCI = 126

↓
TP FN
FP TN

	Naive Bayes	J48 (10) (1)	Simple Cart (10) (1)	PART (10) (1)	Bayes Net (K2)							ZeroR
Correctly Classified Instances	72.83% (9.89)	74.44% (8.79)	73.44% (9.85)	70.78% (9.20)	69.39% (10.43)							68.89% (2.74)
Kappa Statistic	0.37 (0.23)	0.32 (0.24)	0.33 (0.25)	0.30 (0.20)	0.28 (0.24)							0.00 (0.00)*
Precision (2)	0.58 (0.20)	0.67 (0.29)	0.62 (0.28)	0.59 (0.21)	0.52 (0.20)							0.00 (0.00)*
F-Measure (2)	0.56 (0.18)	0.47 (0.21)	0.49 (0.22)	0.49 (0.15)	0.50 (0.18)							0.00 (0.00)*
TP Rate (2)	0.58 (0.22)	0.39 (0.21)	0.45 (0.23)	0.47 (0.19)	0.51 (0.22)							0.00 (0.00)*
** Confusion Matrix	a b 33 23 | a = high 24 100 | b = iso CCI = 133	a b 24 32 | a = high 13 111 | b = iso CCI = 135	a b 23 33 | a = high 16 108 | b = iso CCI = 131	a b 27 29 | a = high 22 102 | b = iso CCI = 129	a b 30 26 | a = high 29 95 | b = iso CCI = 125							a b 0 56 | a = high 0 124 | b = iso CCI = 124

↓

TP FN

FP TN

NOTA:

Todos os algoritmos encontram-se ordenados da esquerda para a direita de forma decrescente de resultados, ou seja, a coluna mais à esquerda relativa ao classificador resultante recorrendo ao algoritmo ***naive Bayes*** apresenta os melhores resultados comparativamente com os restantes classificadores (tendo em conta, acima de tudo, os valores de **"Correctly Classified Instances"** e **"Kappa statistic"**);

Todos os valores que se encontram entre parêntesis representam desvios-padrão;

* Valor de **"Paired Corrected T-Tester"** significativo para **0.01**;

(2) Os valores relativos às métricas: **"Precision"**, **"F-Measure"** e **"TP Rate"** dizem respeito à classe **"high"**, relativo a **"high density"**;

(1) Os classificadores resultantes recorrendo aos algoritmos **J48**, **SimpleCart**, **PART** e **NBTree** apesar de não apresentarem os índices mais elevados em termos de **"Correctly Classified Instances"**, **"Precision"** e **"F-Measure"**, geraram regras/árvores interessantes (ver págs. anexas). Conclusões tiradas depois da análise dos respectivos **"Classifiers outputs"** no **WEKA Explorer**;

** **"Confusion Matrix"** - Dados obtidos depois de gerados os **"Classifiers outputs"** para cada um dos classificadores no **WEKA Explorer**.
CCI – Número de **"Correctly Classified Instances"**;

Os **valores relativos à tabela da pág. 2 são valores de teste**, uma vez que foi estudado o comportamento de certos classificadores que recorrem a algoritmos com **parâmetros diferentes** dos presentes na tabela da pág. 1. Sendo assim, temos:

Tabela pág. 1:

BayesNet – "searchAlgorithm" (TAN)
J48 – "numFolds" (3 – "default value")
SimpleCart – "numFoldsPruning" (5 – "default value")
PART – "numFolds" (3 – "default value")

Tabela pág. 2:

BayesNet – "searchAlgorithm" (K2)
J48 – "numFolds" (10)
SimpleCart – "numFoldsPruning" (10)
PART – "numFolds" (10)

J48:

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

SIZE <= 16
|   age_at_mammo <= 61: iso (86.0/7.0)
|   age_at_mammo > 61
|   |   SIZE <= 12: iso (33.0/8.0)
|   |   SIZE > 12: high (7.0)
SIZE > 16
|   OVERALL_BREAST_COMPOSITION = scattered fibroglandular densities: high (14.0/3.0)
|   OVERALL_BREAST_COMPOSITION = almost entirely fat: high (7.0)
|   OVERALL_BREAST_COMPOSITION = heterogeneously dense
|   |   SIDE = R: high (11.0/2.0)
|   |   SIDE = L: iso (17.0/6.0)
|   OVERALL_BREAST_COMPOSITION = extremely dense: iso (5.0/1.0)

Number of Leaves  :     8

Size of the tree :      13
```

SimpleCart:

```
=== Classifier model (full training set) ===

CART Decision Tree

SIZE < 16.5: iso(104.0/22.0)
SIZE >= 16.5: high(34.0/20.0)

Number of Leaf Nodes: 2

Size of the Tree: 3
```

PART:

```
=== Classifier model (full training set) ===

PART decision list
------------------

SIZE <= 16 AND
age_at_mammo <= 61: iso (86.0/7.0)

SIZE <= 9 AND
OVERALL_BREAST_COMPOSITION = scattered fibroglandular densities: iso (7.0)

SIZE <= 9 AND
OVERALL_BREAST_COMPOSITION = heterogeneously dense: iso (5.0)

CLOCKFACE_LOCATION_OR_REGION = C AND
rnd_num > 0.25: high (7.0)

CLOCKFACE_LOCATION_OR_REGION = 11.0 AND
age_at_mammo > 46 AND
SIZE > 12: high (11.0)

CLOCKFACE_LOCATION_OR_REGION = 11.0: iso (9.0/1.0)

CLOCKFACE_LOCATION_OR_REGION = 6.0: high (6.0/2.0)

CLOCKFACE_LOCATION_OR_REGION = 12.0 AND
MASS_SHAPE = X: high (5.0/1.0)

CLOCKFACE_LOCATION_OR_REGION = 10.0: iso (7.0/3.0)

SIDE = R AND
OVERALL_BREAST_COMPOSITION = heterogeneously dense: high (7.0)

DEPTH = A: iso (11.38/4.38)

DEPTH = M AND
CLOCKFACE_LOCATION_OR_REGION = 1.0: high (4.48/1.0)

CLOCKFACE_LOCATION_OR_REGION = 12.0 AND
MASS_SHAPE = O AND
SIDE = L: iso (3.0)

CLOCKFACE_LOCATION_OR_REGION = 12.0 AND
reread_group = burnside: high (2.0)

: iso (9.14/2.14)

Number of Rules  :      15
```

NBTree:

```
=== Classifier model (full training set) ===

NBTree
------------------

SIDE = R
|   SIZE <= 16.5
|   |   QUADRANT_LOCATION_def = Upper Outer: NB 3
|   |   QUADRANT_LOCATION_def = Upper Inner: NB 4
|   |   QUADRANT_LOCATION_def = Lower Inner: NB 5
|   |   QUADRANT_LOCATION_def = Lower Outer: NB 6
|   SIZE > 16.5: NB 7
SIDE = L
|   SIZE <= 14.5
|   |   rnd_num <= 0.525: NB 10
|   |   rnd_num > 0.525: NB 11
|   SIZE > 14.5: NB 12
```

180 casos

(E_5) PREVISÃO DE *density_num*

	Naive Bayes	NB Tree (1)	SMO	Simple Cart (1)	Bayes Net (TAN)	PART (1)	DTNB	J48 (1)	Random Forest	OneR	Decision Stump	ZeroR
Correctly Classified Instances	67.22% (12.14)	66.17% (11.17)	64.50% (11.95)	63.61% (10.98)	60.89% (12.68)	60.39% (11.46)	60.17% (11.03)	59.11% (10.28)	59.11% (10.81)	56.61% (10.70)	56.28% (7.14)*	55.00% (1.68)*
Kappa Statistic	0.33 (0.25)	0.31 (0.23)	0.27 (0.24)	0.25 (0.23)	0.20 (0.26)	0.20 (0.23)	0.18 (0.23)	0.16 (0.21)	0.17 (0.22)	0.10 (0.22)	0.07 (0.16)*	0.00 (0.00)*
Precision (2)	0.66 (0.16)	0.65 (0.17)	0.64 (0.18)	0.61 (0.20)	0.57 (0.17)	0.58 (0.17)	0.57 (0.17)	0.57 (0.16)	0.55 (0.15)	0.55 (0.21)	0.51 (0.27)	0.00 (0.00)*
F-Measure (2)	0.62 (0.15)	0.60 (0.14)	0.57 (0.15)	0.54 (0.18)	0.54 (0.17)	0.54 (0.15)	0.51 (0.16)	0.49 (0.15)	0.51 (0.15)	0.41 (0.17)*	0.30 (0.17)*	0.00 (0.00)*
TP Rate (2)	0.60 (0.17)	0.57 (0.17)	0.54 (0.18)	0.51 (0.20)	0.53 (0.20)	0.53 (0.19)	0.48 (0.19)	0.46 (0.17)	0.49 (0.18)	0.36 (0.18)*	0.24 (0.15)*	0.00 (0.00)*
** Confusion Matrix	a b 49 32 | a = high 28 71 | b = iso CCI = 120	a b 42 39 | a = high 24 75 | b = iso CCI = 117	a b 38 43 | a = high 29 70 | b = iso CCI = 108	a b 39 42 | a = high 28 71 | b = iso CCI = 110	a b 40 41 | a = high 35 64 | b = iso CCI = 104	a b 37 44 | a = high 34 65 | b = iso CCI = 102	a b 40 41 | a = high 30 69 | b = iso CCI = 109	a b 36 45 | a = high 24 75 | b = iso CCI = 111	a b 42 39 | a = high 36 63 | b = iso CCI = 105	a b 26 55 | a = high 24 75 | b = iso CCI = 101	a b 13 68 | a = high 12 87 | b = iso CCI = 100	a b 0 81 | a = high 0 99 | b = iso CCI = 99

↓
TP FN
FP TN

	Naive Bayes	Bayes Net (K2)	Simple Cart (10) (1)	PART (10) (1)	J48 (10) (1)							
Correctly Classified Instances	67.22% (12.14)	64.56% (11.60)	64.17% (11.43)	60.39% (11.46)	59.11% (10.28)							
Kappa Statistic	0.33 (0.25)	0.27 (0.24)	0.27 (0.23)	0.20 (0.23)	0.16 (0.21)							
Precision (2)	0.66 (0.16)	0.63 (0.17)	0.64 (0.19)	0.58 (0.17)	0.57 (0.16)							
F-Measure (2)	0.62 (0.15)	0.57 (0.16)	0.56 (0.16)	0.54 (0.15)	0.49 (0.15)							
TP Rate (2)	0.60 (0.17)	0.54 (0.19)	0.53 (0.19)	0.53 (0.19)	0.46 (0.17)							
** Confusion Matrix	a b 49 32 | a = high 28 71 | b = iso CCI = 120	a b 43 38 | a = high 28 71 | b = iso CCI = 114	a b 40 41 | a = high 26 73 | b = iso CCI = 113	a b 37 44 | a = high 34 65 | b = iso CCI = 102	a b 36 45 | a = high 24 75 | b = iso CCI = 111							

↓

TP FN

FP TN

NOTA:

Todos os algoritmos encontram-se ordenados da esquerda para a direita de forma decrescente de resultados, ou seja, a coluna mais à esquerda relativa ao classificador resultante recorrendo ao algoritmo ***naive Bayes*** apresenta os melhores resultados comparativamente com os restantes classificadores (tendo em conta, acima de tudo, os valores de **"Correctly Classified Instances"**, **"Kappa Statistic"** e **"F-Measure"**);

Todos os valores que se encontram entre parêntesis representam desvios-padrão;

* Valor de **"Paired Corrected T-Tester"** significativo para **0.01**;

(2) Os valores relativos às métricas: **"Precision"**, **"F-Measure"** e **"TP Rate"** dizem respeito à classe **"high"**, relativo a **"high density"**;

(1) Os classificadores resultantes recorrendo aos algoritmos **NBTree**, **SimpleCart**, **PART** e **J48** apesar de não apresentarem os índices mais elevados em termos de **"Correctly Classified Instances"**, **"Kappa Statistic"** e **"F-Measure"**, geraram regras/árvores interessantes (ver págs. anexas). Conclusões tiradas depois da análise dos respectivos **"Classifiers outputs"** no **WEKA Explorer**;

** **"Confusion Matrix"** - Dados obtidos depois de gerados os **"Classifiers outputs"** para cada um dos classificadores no **WEKA Explorer**.
CCI – Número de **"Correctly Classified Instances"**;

Os **valores relativos à tabela da pág. 2 são valores de teste**, uma vez que foi estudado o comportamento de certos classificadores que recorrem a algoritmos com **parâmetros diferentes** dos presentes na tabela da pág. 1. Sendo assim, temos:

Tabela pág. 1:

BayesNet – "searchAlgorithm" (TAN)
SimpleCart – "numFoldsPruning" (5 – "default value")
PART – "numFolds" (3 – "default value")
J48 – "numFolds" (3 – "default value")

Tabela pág. 2:

BayesNet – "searchAlgorithm" (K2)
SimpleCart – "numFoldsPruning" (10)
PART – "numFolds" (10)
J48 – "numFolds" (10)

NBTree:

```
=== Classifier model (full training set) ===

NBTree
------------------

SIZE <= 16.5: NB 1
SIZE > 16.5
|   MASS_SHAPE = X: NB 3
|   MASS_SHAPE = R: NB 4
|   MASS_SHAPE = O: NB 5
|   MASS_SHAPE = L: NB 6
```

SImpleCart:

```
=== Classifier model (full training set) ===

CART Decision Tree

SIZE < 16.5
|  age_at_mammo < 61.5: iso(66.0/20.0)
|  age_at_mammo >= 61.5
|  |  CLOCKFACE_LOCATION_OR_REGION=(10.0)|(2.0)|(1.0)|(4.0): iso(11.0/5.0)
|  |  CLOCKFACE_LOCATION_OR_REGION!=(10.0)|(2.0)|(1.0)|(4.0): high(20.0/4.0)
SIZE >= 16.5
|  OVERALL_BREAST_COMPOSITION=(extremely dense): iso(5.0/0.0)
|  OVERALL_BREAST_COMPOSITION!=(extremely dense): high(36.0/13.0)

Number of Leaf Nodes: 5

Size of the Tree: 9
```

PART:

```
=== Classifier model (full training set) ===

PART decision list
------------------

OVERALL_BREAST_COMPOSITION = extremely dense: iso (13.0/1.0)

SIZE > 16 AND
OVERALL_BREAST_COMPOSITION = scattered fibroglandular densities: high (14.0/2.0)

OVERALL_BREAST_COMPOSITION = almost entirely fat AND
MASS_MARGINS_1 = S: high (8.0/0.25)

age_at_mammo <= 69 AND
SIZE <= 16: iso (94.5/25.75)

CLOCKFACE_LOCATION_OR_REGION = C: high (5.75)

CLOCKFACE_LOCATION_OR_REGION = 10.0: iso (5.0/1.0)

SIDE = R: high (18.75/3.0)

OVERALL_BREAST_COMPOSITION = heterogeneously dense: iso (15.0/6.0)

: high (6.0)

Number of Rules  :      9
```

J48:

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

OVERALL_BREAST_COMPOSITION = scattered fibroglandular densities
|   SIZE <= 17: iso (48.0/15.0)
|   SIZE > 17: high (14.0/2.0)
OVERALL_BREAST_COMPOSITION = almost entirely fat: high (32.0/9.0)
OVERALL_BREAST_COMPOSITION = heterogeneously dense
|   MASS_MARGINS_2 = I: iso (21.03/7.15)
|   MASS_MARGINS_2 = S: high (8.66/2.19)
|   MASS_MARGINS_2 = D
|   |   age_at_mammo <= 40: high (4.36)
|   |   age_at_mammo > 40: iso (21.63/6.07)
|   MASS_MARGINS_2 = M: high (3.71/1.51)
|   MASS_MARGINS_2 = U: iso (13.61/3.75)
OVERALL_BREAST_COMPOSITION = extremely dense: iso (13.0/1.0)

Number of Leaves  :     10

Size of the tree :      14
```

SimpleCart (10):

```
=== Classifier model (full training set) ===

CART Decision Tree

SIZE < 16.5
|  age_at_mammo < 61.5: iso(66.0/20.0)
|  age_at_mammo >= 61.5: high(25.0/15.0)
SIZE >= 16.5: high(36.0/18.0)

Number of Leaf Nodes: 3

Size of the Tree: 5
```

168 casos

(E_6) PREVISÃO DE *retro_density*

	Naive Bayes	OneR	J48	Simple Cart	DTNB	Decision Stump	Random Forest	SMO	PART	Bayes Net (TAN)	NB Tree	ZeroR
Correctly Classified Instances (Accuracy)	82.14% (138)	81.55% (137)	79.17% (133)	78.57% (132)	79.17% (133)	78.57% (132)	79.76% (134)	86.31% (145)	78.57% (132)	83.33% (140)	75.00% (126)	83.33% (140)
Incorrectly Classified Instances	17.86% (30)	18.45% (31)	20.83% (35)	21.43% (36)	20.83% (35)	21.43% (36)	20.24% (34)	13.69% (23)	21.43% (36)	16.67% (28)	25.00% (42)	16.67% (28)
Kappa Statistic	0.45	0.31	0.32	0.31	0.32	0.31	0.31	0.45	0.23	0.38	0.25	0.00
Mean absolute error	0.23	0.18	0.28	0.30	0.30	0.30	0.30	0.14	0.26	0.23	0.27	0.38
Precision (1)	0.48	0.44	0.40	0.39	0.40	0.39	0.41	0.62	0.36	0.50	0.33	0.00
F-Measure (1)	0.56	0.42	0.44	0.44	0.44	0.44	0.43	0.53	0.36	0.48	0.40	0.00
TP Rate (1) (Recall)	0.68	0.39	0.50	0.50	0.50	0.50	0.46	0.46	0.36	0.46	0.50	0.00
FP Rate (1)	0.15	0.10	0.15	0.16	0.15	0.16	0.14	0.06	0.13	0.09	0.20	0.00
** **Confusion Matrix**	a b 19 9 | a=high 21 119 | b=iso CCI = 138	a b 11 17 | a=high 14 126 | b=iso CCI = 137	a b 14 14 | a=high 21 119 | b=iso CCI = 133	a b 14 14 | a=high 22 118 | b=iso CCI = 132	a b 14 14 | a=high 21 119 | b=iso CCI = 133	a b 14 14 | a=high 22 118 | b=iso CCI = 132	a b 13 15 | a=high 19 121 | b=iso CCI = 134	a b 13 15 | a=high 8 132 | b=iso CCI = 145	a b 10 18 | a=high 18 122 | b=iso CCI = 132	a b 13 15 | a=high 13 127 | b=iso CCI = 140	a b 14 14 | a=high 28 112 | b=iso CCI = 126	a b 0 28 | a=high 0 140 | b=iso CCI = 140

↓
TP FN
FP TN

	Naive Bayes	J48 (10)	Simple Cart (10)	PART (10)	Bayes Net (K2)							
Correctly Classified Instances (Accuracy)	82.14% (138)	79.17% (133)	78.57% (132)	78.57% (132)	82.74% (139)							
Incorrectly Classified Instances	17.86% (30)	20.83% (35)	21.43% (36)	21.43% (36)	17.26% (29)							
Kappa Statistic	0.45	0.32	0.31	0.23	0.42							
Mean absolute error	0.23	0.28	0.30	0.26	0.23							
Precision (1)	0.48	0.40	0.39	0.36	0.49							
F-Measure (1)	0.56	0.44	0.44	0.36	0.53							
TP Rate (1) (Recall)	0.68	0.50	0.50	0.36	0.57							
FP Rate (1)	0.15	0.15	0.16	0.13	0.12							
** Confusion Matrix	a b 19 9 | a=high 21 119 | b=iso CCI = 138	a b 14 14 | a=high 21 119 | b=iso CCI = 133	a b 14 14 | a=high 22 118 | b=iso CCI = 132	a b 10 18 | a=high 18 122 | b=iso CCI = 132	a b 16 12 | a=high 17 123 | b=iso CCI = 139							

↓
TP FN
FP TN

NOTA:

Todos os algoritmos encontram-se ordenados de acordo com a experiência E_4 para os **180 casos**, ou seja, de acordo com a experiência que serviu de **modelo** a este estudo (E_4);

Os **valores** a **encarnado** apresentam **melhores índices** de diferentes classificadores (para as métricas mais relevantes) comparativamente com os valores do classifiicador (***naive Bayes***) que evidenciou **resultados mais altos** a todos os níveis aquando da **previsão de *retro_density*** para os **180 casos**;

Os **valores** a **bege** apresentam **índices iguais** de diferentes classificadores (para as métricas mais relevantes) comparativamente com os valores do classifiicador (***naive Bayes***) que evidenciou **resultados mais altos** a todos os níveis aquando da **previsão de *retro_density*** para os **180 casos**;

Todas as **métricas** encontram-se **arredondadas** às **duas casas decimais**;

(1) Os valores relativos às métricas: **"Precision"**, **"F-Measure"**, **"TP Rate"** e **"FP Rate"** dizem respeito à classe **"high"**, relativo a **"high density"**;

**** CCI** – Número de **"Correctly Classified Instances"**;

Os **valores relativos à tabela da pág. 2 são valores de teste**, uma vez que foi estudado o comportamento de certos classificadores que recorrem a algoritmos com **parâmetros diferentes** dos presentes na tabela da pág. 1. Sendo assim, temos:

Tabela pág. 1:

BayesNet – "searchAlgorithm" (TAN)
J48 – "numFolds" (3 – "default value")
SimpleCart – "numFoldsPruning" (5 – "default value")
PART – "numFolds" (3 – "default value")

Tabela pág. 2:

BayesNet – "searchAlgorithm" (K2)
J48 – "numFolds" (10)
SimpleCart – "numFoldsPruning" (10)
PART – "numFolds" (10)

168 casos

(E_7) PREVISÃO DE *density_num*

	Naive Bayes	NB Tree	SMO	Simple Cart	Bayes Net (TAN)	PART	DTNB	J48	Random Forest	OneR	Decision Stump	ZeroR
Correctly Classified Instances (Accuracy)	75.60% (127)	72.62% (122)	75.60% (127)	69.05% (116)	64.88% (109)	75.60% (127)	75.00% (126)	76.79% (129)	75.60% (127)	77.38% (130)	81.55% (137)	83.33% (140)
Incorrectly Classified Instances	24.40% (41)	27.38% (46)	24.40% (41)	30.95% (52)	35.12% (59)	24.40% (41)	25.00% (42)	23.21% (39)	24.40% (41)	22.62% (38)	18.45% (31)	16.67% (28)
Kappa Statistic	0.35	0.36	0.22	0.25	0.11	0.29	0.30	0.28	0.29	0.27	0.09	0.00
Mean absolute error	0.30	0.31	0.24	0.40	0.39	0.38	0.41	0.38	0.38	0.23	0.44	0.47
Precision (1)	0.38	0.36	0.32	0.31	0.23	0.36	0.35	0.36	0.36	0.36	0.33	0.00
F-Measure (1)	0.49	0.51	0.37	0.42	0.31	0.44	0.45	0.42	0.44	0.41	0.16	0.00
TP Rate (1) (Recall)	0.71	0.86	0.43	0.68	0.46	0.57	0.61	0.50	0.57	0.46	0.11	0.00
FP Rate (1)	0.24	0.30	0.18	0.31	0.31	0.21	0.22	0.18	0.21	0.16	0.04	0.00
** Confusion Matrix	a b 20 8 | a=high 33 107 | b=iso CCI = 127	a b 24 4 | a=high 42 98 | b=iso CCI = 122	a b 12 16 | a=high 25 115 | b=iso CCI = 127	a b 19 9 | a=high 43 97 | b=iso CCI = 116	a b 13 15 | a=high 44 96 | b=iso CCI = 109	a b 16 12 | a=high 29 111 | b=iso CCI = 127	a b 17 11 | a=high 31 109 | b=iso CCI = 126	a b 14 14 | a=high 25 115 | b=iso CCI = 129	a b 16 12 | a=high 29 111 | b=iso CCI = 127	a b 13 15 | a=high 23 117 | b=iso CCI = 130	a b 3 25 | a=high 6 134 | b=iso CCI = 137	a b 0 28 | a=high 0 140 | b=iso CCI = 140

↓
TP FN
FP TN

	Naïve Bayes	Bayes Net (K2)	Simple Cart (10)	PART (10)	J48 (10)							
Correctly Classified Instances (Accuracy)	75.60% (127)	73.81% (124)	65.48% (110)	75.60% (127)	76.79% (129)							
Incorrectly Classified Instances	24.40% (41)	26.19% (44)	34.52% (58)	24.40% (41)	23.21% (39)							
Kappa Statistic	0.35	0.28	0.26	0.29	0.28							
Mean absolute error	0.30	0.33	0.39	0.38	0.38							
Precision (1)	0.38	0.34	0.30	0.36	0.36							
F-Measure (1)	0.49	0.44	0.44	0.44	0.42							
TP Rate (1) (Recall)	0.71	0.61	0.82	0.57	0.50							
FP Rate (1)	0.24	0.24	0.38	0.21	0.18							
** Confusion Matrix	a b 20 8 | a=high 33 107 | b=iso CCI = 127	a b 17 11 | a=high 33 107 | b=iso CCI = 124	a b 23 5 | a=high 53 87 | b=iso CCI = 110	a b 16 12 | a=high 29 111 | b=iso CCI = 127	a b 14 14 | a=high 25 115 | b=iso CCI = 129							

↓
TP FN
FP TN

NOTA:

Todos os algoritmos encontram-se ordenados de acordo com a experiência E_5 para os **180 casos**, ou seja, de acordo com a experiência que serviu de **modelo** a este estudo (E_7);

Os **valores** a **encarnado** apresentam **melhores índices** de diferentes classificadores (para as métricas mais relevantes) comparativamente com os valores do classifiicador (***naive Bayes***) que evidenciou **resultados mais altos** a todos os níveis aquando da **previsão de *density_num*** para os **180 casos**;

Os **valores** a **bege** apresentam **índices iguais** de diferentes classificadores (para as métricas mais relevantes) comparativamente com os valores do classifiicador (***naive Bayes***) que evidenciou **resultados mais altos** a todos os níveis aquando da **previsão de *density_num*** para os **180 casos**;

Todas as **métricas** encontram-se **arredondadas** às **duas casas decimais**;

(1) Os valores relativos às métricas: **"Precision"**, **"F-Measure"**, **"TP Rate"** e **"FP Rate"** dizem respeito à classe **"high"**, relativo a **"high density"**;

**** CCI** – Número de **"Correctly Classified Instances"**;

Os **valores relativos à tabela da pág. 2 são valores de teste**, uma vez que foi estudado o comportamento de certos classificadores que recorrem a algoritmos com **parâmetros diferentes** dos presentes na tabela da pág. 1. Sendo assim, temos:

Tabela pág. 1:

BayesNet – "searchAlgorithm" (TAN)
SimpleCart – "numFoldsPruning" (5 – "default value")
PART – "numFolds" (3 – "default value")
J48 – "numFolds" (3 – "default value")

Tabela pág. 2:

BayesNet – "searchAlgorithm" (K2)
SimpleCart – "numFoldsPruning" (10)
PART – "numFolds" (10)
J48 – "numFolds" (10)

168 casos

(E8) PREVISÃO DE *outcome_num* COM *retro_density*

	SMO	Naive Bayes	DTNB	Bayes Net (TAN)	PART	NB Tree	Simple Cart	Random Forest	J48	OneR	Decision Stump	ZeroR
Correctly Classified Instances (Accuracy)	81.55% (137)	78.57% (132)	79.76% (134)	80.95% (136)	64.88% (109)	73.81% (124)	79.76% (134)	79.76% (134)	81.55% (137)	72.02% (121)	75.60% (127)	72.02% (121)
Incorrectly Classified Instances	18.45% (31)	21.43% (36)	20.24% (34)	19.05% (32)	35.12% (59)	26.19% (44)	20.24% (34)	20.24% (34)	18.45% (31)	27.98% (47)	24.40% (41)	27.98% (47)
Kappa Statistic	0.52	0.49	0.46	0.50	0.22	0.36	0.50	0.47	0.51	0.34	0.31	0.00
Mean absolute error	0.18	0.24	0.29	0.23	0.34	0.30	0.26	0.32	0.28	0.28	0.36	0.45
Precision (1)	0.70	0.60	0.69	0.69	0.41	0.53	0.64	0.67	0.72	0.50	0.61	0.00
F-Measure (1)	0.64	0.64	0.59	0.63	0.47	0.54	0.64	0.61	0.63	0.54	0.45	0.00
TP Rate (1) (Recall)	0.60	0.68	0.51	0.57	0.55	0.55	0.64	0.55	0.55	0.57	0.36	0.00
FP Rate (1)	0.10	0.17	0.09	0.10	0.31	0.19	0.14	0.11	0.08	0.22	0.09	0.00
** Confusion Matrix	a b 28 19 | a=mal. 12 109 | b=ben. CCI = 137	a b 32 15 | a=mal. 21 100 | b=ben. CCI = 132	a b 24 23 | a=mal. 11 110 | b=ben. CCI = 134	a b 27 20 | a=mal. 12 109 | b=ben. CCI = 136	a b 26 21 | a=mal. 38 83 | b=ben. CCI = 109	a b 26 21 | a=mal. 23 98 | b=ben. CCI = 124	a b 30 17 | a=mal. 17 104 | b=ben. CCI = 134	a b 26 21 | a=mal. 13 108 | b=ben. CCI = 134	a b 26 21 | a=mal. 10 111 | b=ben. CCI = 137	a b 27 20 | a=mal. 27 94 | b=ben. CCI = 121	a b 17 30 | a=mal. 11 110 | b=ben. CCI = 127	a b 0 47 | a=mal. 0 121 | b=ben. CCI = 121

↓
TP FN
FP TN

	SMO	Bayes Net (K2)	PART (10)	Simple Cart (10)	J48 (10)							
Correctly Classified Instances (Accuracy)	81.55% (137)	76.19% (128)	64.88% (109)	79.76% (134)	81.55% (137)							
Incorrectly Classified Instances	18.45% (31)	23.81% (40)	35.12% (59)	20.24% (34)	18.45% (31)							
Kappa Statistic	0.52	0.40	0.22	0.49	0.51							
Mean absolute error	0.18	0.24	0.34	0.26	0.28							
Precision (1)	0.70	0.58	0.41	0.64	0.72							
F-Measure (1)	0.64	0.57	0.47	0.63	0.63							
TP Rate (1) (Recall)	0.60	0.55	0.55	0.62	0.55							
FP Rate (1)	0.10	0.16	0.31	0.13	0.08							
** **Confusion Matrix**	a b 28 19 | a=mal. 12 109 | b=ben. CCI = 137	a b 26 21 | a=mal. 19 102 | b=ben. CCI = 128	a b 26 21 | a=mal. 38 83 | b=ben. CCI = 109	a b 29 18 | a=mal. 16 105 | b=ben. CCI = 134	a b 26 21 | a=mal. 10 111 | b=ben. CCI = 137							

↓
TP FN
FP TN

NOTA:

Todos os algoritmos encontram-se ordenados de acordo com a experiência E_1 para os **180 casos**, ou seja, de acordo com a experiência que serviu de **modelo** a este estudo (E_s);

Os **valores** a **encarnado** apresentam **melhores índices** de diferentes classificadores (para as métricas mais relevantes) comparativamente com os valores do classifiicador **(SMO)** que evidenciou **resultados mais altos** a todos os níveis aquando da **previsão de *outcome_num* com *retro_density*** para os **180 casos**;

Os **valores** a **bege** apresentam **índices iguais** de diferentes classificadores (para as métricas mais relevantes) comparativamente com os valores do classifiicador **(SMO)** que evidenciou **resultados mais altos** a todos os níveis aquando da **previsão de *outcome_num* com *retro_density*** para os **180 casos**;

Todas as **métricas** encontram-se **arredondadas** às **duas casas decimais**;

(1) Os valores relativos às métricas: **"Precision"**, **"F-Measure"**, **"TP Rate"** e **"FP Rate"** dizem respeito à classe **"malignant"**;

**** CCI** – Número de **"Correctly Classified Instances"**;

Os **valores relativos à tabela da pág. 2 são valores de teste**, uma vez que foi estudado o comportamento de certos classificadores que recorrem a algoritmos com **parâmetros diferentes** dos presentes na tabela da pág. 1. Sendo assim, temos:

Tabela pág. 1:

BayesNet – "searchAlgorithm" (TAN)
PART – "numFolds" (3 – "default value")
SimpleCart – "numFoldsPruning" (5 – "default value")
J48 – "numFolds" (3 – "default value")

Tabela pág. 2:

BayesNet – "searchAlgorithm" (K2)
PART – "numFolds" (10)
SimpleCart – "numFoldsPruning" (10)
J48 – "numFolds" (10)

168 casos

(E_9) PREVISÃO DE *outcome_num* COM *retro_density* (PREVISTA EM E_6)

	SMO	Naive Bayes	DTNB	Bayes Net (TAN)	PART	NB Tree	Simple Cart	Random Forest	J48	OneR	Decision Stump	ZeroR
Correctly Classified Instances (Accuracy)	79.76% (134)	76.79% (129)	80.36% (135)	80.36% (135)	64.29% (108)	73.81% (124)	80.95% (136)	78.57% (132)	79.17% (133)	72.02% (121)	77.97% (131)	72.02% (121)
Incorrectly Classified Instances	20.24% (34)	23.21% (39)	19.64% (33)	19.64% (36)	35.71% (60)	26.19% (44)	19.05% (32)	21.43% (36)	20.83% (35)	27.98% (47)	22.02% (37)	27.98% (47)
Kappa Statistic	0.48	0.45	0.49	0.49	0.22	0.37	0.53	0.44	0.46	0.34	0.43	0.00
Mean absolute error	0.20	0.25	0.29	0.24	0.35	0.31	0.25	0.32	0.29	0.28	0.35	0.45
Precision (1)	0.65	0.57	0.68	0.68	0.40	0.53	0.65	0.64	0.65	0.50	0.63	0.00
F-Measure (1)	0.62	0.61	0.62	0.62	0.47	0.56	0.67	0.58	0.60	0.54	0.58	0.00
TP Rate (1) (Recall)	0.60	0.66	0.57	0.57	0.57	0.60	0.68	0.53	0.55	0.57	0.53	0.00
FP Rate (1)	0.12	0.19	0.11	0.11	0.33	0.21	0.14	0.12	0.12	0.22	0.12	0.00
** Confusion Matrix	a b 28 19 | a=mal. 15 106 | b=ben. CCI = 134	a b 31 16 | a = mal. 23 98 | b = ben. CCI = 129	a b 27 20 | a=mal. 13 108 | b=ben. CCI = 135	a b 27 20 | a=mal. 13 108 | b=ben. CCI = 135	a b 27 20 | a=mal. 40 81 | b=ben. CCI = 108	a b 28 19 | a = mal. 25 96 | b = ben. CCI = 124	a b 32 15 | a=mal. 17 104 | b=ben. CCI = 136	a b 25 22 | a=mal. 14 107 | b=ben. CCI = 132	a b 26 21 | a=mal. 14 107 | b=ben. CCI = 133	a b 27 20 | a = mal. 27 94 | b = ben. CCI = 121	a b 25 22 | a=mal. 15 106 | b=ben. CCI = 131	a b 0 47 | a=mal. 0 121 | b=ben. CCI = 121

↓
TP FN
FP TN

	SMO	Bayes Net (K2)	PART (10)	Simple Cart (10)	J48 (10)							
Correctly Classified Instances (Accuracy)	79.76% (134)	76.79% (129)	64.29% (108)	80.36% (135)	79.17% (133)							
Incorrectly Classified Instances	20.24% (34)	23.21% (39)	35.71% (60)	19.64% (33)	20.83% (35)							
Kappa Statistic	0.48	0.41	0.22	0.52	0.46							
Mean absolute error	0.20	0.24	0.35	0.25	0.29							
Precision (1)	0.65	0.59	0.40	0.65	0.65							
F-Measure (1)	0.62	0.57	0.47	0.65	0.60							
TP Rate (1) (Recall)	0.60	0.55	0.57	0.66	0.55							
FP Rate (1)	0.12	0.15	0.33	0.14	0.12							
** Confusion Matrix	a b 28 19 | a=mal. 15 106 | b=ben. CCI = 134	a b 26 21 | a=mal. 18 103 | b=ben. CCI = 129	a b 27 20 | a=mal. 40 81 | b=ben. CCI = 108	a b 31 16 | a=mal. 17 104 | b=ben. CCI = 135	a b 26 21 | a=mal. 14 107 | b=ben. CCI = 133							

↓
TP FN
FP TN

NOTA:

Todos os algoritmos encontram-se ordenados de acordo com a experiência E_1 para os **180 casos**, ou seja, de acordo com a experiência que serviu de **modelo** a este estudo (E_2);

Os **valores** a **encarnado** apresentam **melhores índices** de diferentes classificadores (para as métricas mais relevantes) comparativamente com os valores do classifiicador **(SMO)** que evidenciou **resultados mais altos** a todos os níveis aquando da **previsão de *outcome_num* com *retro_density*** para os **180 casos**;

Os **valores** a **bege** apresentam **índices iguais** de diferentes classificadores (para as métricas mais relevantes) comparativamente com os valores do classifiicador **(SMO)** que evidenciou **resultados mais altos** a todos os níveis aquando da **previsão de *outcome_num* com *retro_density*** para os **180 casos**;

Todas as **métricas** encontram-se **arredondadas** às **duas casas decimais**;

(1) Os valores relativos às métricas: **"Precision"**, **"F-Measure"**, **"TP Rate"** e **"FP Rate"** dizem respeito à classe **"malignant"**;

**** CCI** – Número de **"Correctly Classified Instances"**;

Os **valores relativos à tabela da pág. 2 são valores de teste**, uma vez que foi estudado o comportamento de certos classificadores que recorrem a algoritmos com **parâmetros diferentes** dos presentes na tabela da pág. 1. Sendo assim, temos:

Tabela pág. 1:

BayesNet – "searchAlgorithm" (TAN)
PART – "numFolds" (3 – "default value")
SimpleCart – "numFoldsPruning" (5 – "default value")
J48 – "numFolds" (3 – "default value")

Tabela pág. 2:

BayesNet – "searchAlgorithm" (K2)
PART – "numFolds" (10)
SimpleCart – "numFoldsPruning" (10)
J48 – "numFolds" (10)

168 casos

(E_{10}) PREVISÃO DE *outcome_num* COM *density_num* (PREVISTA EM E_7)

	SMO	Naive Bayes	Bayes Net (TAN)	NB Tree	DTNB	PART	J48	OneR	Simple Cart	Random Forest	Decision Stump	ZeroR
Correctly Classified Instances (Accuracy)	79.17% (133)	76.19% (128)	80.36% (135)	73.81% (124)	73.81% (124)	77.38% (130)	77.38% (130)	72.02% (121)	74.40% (125)	77.38% (130)	76.19% (128)	72.02% (121)
Incorrectly Classified Instances	20.83% (35)	23.81% (40)	19.64% (33)	26.19% (44)	26.19% (44)	22.62% (38)	22.62% (38)	27.98% (47)	25.60% (43)	22.62% (38)	23.81% (40)	27.98% (47)
Kappa Statistic	0.46	0.43	0.48	0.40	0.37	0.42	0.44	0.34	0.43	0.40	0.22	0.00
Mean absolute error	0.21	0.25	0.24	0.27	0.33	0.33	0.34	0.28	0.32	0.31	0.39	0.45
Precision (1)	0.65	0.57	0.68	0.53	0.53	0.61	0.60	0.50	0.53	0.62	0.89	0.00
F-Measure (1)	0.60	0.60	0.61	0.59	0.56	0.58	0.60	0.54	0.61	0.55	0.29	0.00
TP Rate (1) (Recall)	0.55	0.64	0.55	0.66	0.60	0.55	0.60	0.57	0.72	0.49	0.17	0.00
FP Rate (1)	0.12	0.19	0.10	0.23	0.21	0.14	0.16	0.22	0.25	0.12	0.01	0.00
** Confusion Matrix	a b 26 21 | a=mal. 14 107 | b=ben. CCI = 133	a b 30 17 | a=mal. 23 98 | b=ben. CCI = 128	a b 26 21 | a=mal. 12 109 | b=ben. CCI = 135	a b 31 16 | a=mal. 28 93 | b=ben. CCI = 124	a b 28 19 | a=mal. 25 96 | b=ben. CCI = 124	a b 26 21 | a=mal. 17 104 | b=ben. CCI = 130	a b 28 19 | a=mal. 19 102 | b=ben. CCI = 130	a b 27 20 | a=mal. 27 94 | b=ben. CCI = 121	a b 34 13 | a=mal. 30 91 | b=ben. CCI = 125	a b 23 24 | a=mal. 14 107 | b=ben. CCI = 130	a b 8 39 | a=mal. 1 120 | b=ben. CCI = 128	a b 0 47 | a=mal. 0 121 | b=ben. CCI = 121

↓
TP FN
FP TN

	SMO	Bayes Net (K2)	PART (10)	J48 (10)	Simple Cart (10)							
Correctly Classified Instances (Accuracy)	79.17% (133)	76.79% (129)	77.38% (130)	77.38% (130)	75.00% (126)							
Incorrectly Classified Instances	20.83% (35)	23.21% (39)	22.62% (38)	22.62% (38)	25.00% (42)							
Kappa Statistic	0.46	0.41	0.42	0.44	0.44							
Mean absolute error	0.21	0.24	0.33	0.34	0.32							
Precision (1)	0.65	0.59	0.61	0.60	0.54							
F-Measure (1)	0.60	0.57	0.58	0.60	0.62							
TP Rate (1) (Recall)	0.55	0.55	0.55	0.60	0.72							
FP Rate (1)	0.12	0.15	0.14	0.16	0.24							
** Confusion Matrix	a b 26 21 | a=mal. 14 107 | b=ben. CCI = 133	a b 26 21 | a=mal. 18 103 | b=ben. CCI = 129	a b 26 21 | a=mal. 17 104 | b=ben. CCI = 130	a b 28 19 | a=mal. 19 102 | b=ben. CCI = 130	a b 34 13 | a=mal. 29 92 | b=ben. CCI = 126							

↓

TP FN

FP TN

NOTA:

Todos os algoritmos encontram-se ordenados de acordo com a experiência E_2 para os **180 casos**, ou seja, de acordo com a experiência que serviu de **modelo** a este estudo (E_{180});

Os **valores** a **encarnado** apresentam **melhores índices** de diferentes classificadores (para as métricas mais relevantes) comparativamente com os valores do classifiicador **(SMO)** que evidenciou **resultados mais altos** a todos os níveis aquando da **previsão de *outcome_num* com *density_num*** para os **180 casos**;

Os **valores** a **bege** apresentam **índices iguais** de diferentes classificadores (para as métricas mais relevantes) comparativamente com os valores do classifiicador **(SMO)** que evidenciou **resultados mais altos** a todos os níveis aquando da **previsão de *outcome_num* com *density_num*** para os **180 casos**;

Todas as **métricas** encontram-se **arredondadas** às **duas casas decimais**;

(1) Os valores relativos às métricas: **"Precision"**, **"F-Measure"**, **"TP Rate"** e **"FP Rate"** dizem respeito à classe **"malignant"**;

**** CCI** – Número de **"Correctly Classified Instances"**;

Os **valores relativos à tabela da pág. 2 são valores de teste**, uma vez que foi estudado o comportamento de certos classificadores que recorrem a algoritmos com **parâmetros diferentes** dos presentes na tabela da pág. 1. Sendo assim, temos:

Tabela pág. 1:

BayesNet – "searchAlgorithm" (TAN)
PART – "numFolds" (3 – "default value")
J48 – "numFolds" (3 – "default value")
SimpleCart – "numFoldsPruning" (5 – "default value")

Tabela pág. 2:

BayesNet – "searchAlgorithm" (K2)
PART – "numFolds" (10)
J48 – "numFolds" (10)
SimpleCart – "numFoldsPruning" (10)

168 casos

(E_{11}) PREVISÃO DE *outcome_num* SEM DENSIDADE DE MASSA

	SMO	Naive Bayes	Simple Cart	Bayes Net (TAN)	DTNB	PART	OneR	NB Tree	Random Forest	J48	Decision Stump	ZeroR
Correctly Classified Instances (Accuracy)	77.38% (130)	73.81% (124)	73.81% (124)	78.57% (132)	73.81% (124)	80.36% (135)	72.02% (121)	75.00% (126)	72.62% (122)	73.21% (123)	76.19% (128)	72.02% (121)
Incorrectly Classified Instances	22.62% (38)	26.19% (44)	26.19% (44)	21.43% (36)	26.19% (44)	19.64% (33)	27.98% (47)	25.00% (42)	27.38% (46)	26.79% (45)	23.81% (40)	27.98% (47)
Kappa Statistic	0.42	0.39	0.40	0.46	0.37	0.49	0.34	0.40	0.25	0.30	0.22	0.00
Mean absolute error	0.23	0.27	0.32	0.26	0.33	0.31	0.28	0.30	0.35	0.35	0.39	0.45
Precision (1)	0.61	0.53	0.53	0.62	0.53	0.68	0.50	0.55	0.52	0.53	0.89	0.00
F-Measure (1)	0.57	0.58	0.59	0.61	0.56	0.62	0.54	0.57	0.43	0.48	0.29	0.00
TP Rate (1) (Recall)	0.53	0.64	0.66	0.60	0.60	0.57	0.57	0.60	0.36	0.45	0.17	0.00
FP Rate (1)	0.13	0.22	0.23	0.14	0.21	0.11	0.22	0.19	0.13	0.16	0.01	0.00
** Confusion Matrix	a b 25 22 | a=mal. 16 105 | b=ben. CCI = 130	a b 30 17 | a = mal. 27 94 | b = ben. CCI = 124	a b 31 16 | a = mal. 28 93 | b = ben. CCI = 124	a b 28 19 | a=mal. 17 104 | b=ben. CCI = 132	a b 28 19 | a = mal. 25 96 | b = ben. CCI = 124	a b 27 20 | a=mal. 13 108 | b=ben. CCI = 135	a b 27 20 | a = mal. 27 94 | b = ben. CCI = 121	a b 28 19 | a = mal. 23 98 | b = ben. CCI = 126	a b 17 30 | a=mal. 16 105 | b=ben. CCI = 122	a b 21 26 | a=mal. 19 102 | b=ben. CCI = 123	a b 8 39 | a=mal. 1 120 | b=ben. CCI = 128	a b 0 47 | a=mal. 0 121 | b=ben. CCI = 121

↓
TP FN
FP TN

	SMO	Bayes Net (K2)	Simple Cart (10)	PART (10)	J48 (10)							
Correctly Classified Instances (Accuracy)	77.38% (130)	75.60% (127)	73.81% (124)	80.36% (135)	73.21% (123)							
Incorrectly Classified Instances	22.62% (38)	24.40% (41)	26.19% (44)	19.64% (33)	26.79% (45)							
Kappa Statistic	0.42	0.40	0.38	0.49	0.30							
Mean absolute error	0.23	0.25	0.31	0.31	0.35							
Precision (1)	0.61	0.56	0.53	0.68	0.53							
F-Measure (1)	0.57	0.57	0.57	0.62	0.48							
TP Rate (1) (Recall)	0.53	0.57	0.62	0.57	0.45							
FP Rate (1)	0.13	0.17	0.22	0.11	0.16							
** Confusion Matrix	a b 25 22 | a=mal. 16 105 | b=ben. CCI = 130	a b 27 20 | a=mal. 21 100 | b=ben. CCI = 127	a b 29 18 | a = mal. 26 95 | b = ben. CCI = 124	a b 27 20 | a=mal. 13 108 | b=ben. CCI = 135	a b 21 26 | a=mal. 19 102 | b=ben. CCI = 123							

↓

TP FN
FP TN

NOTA:

Todos os algoritmos encontram-se ordenados de acordo com a experiência E_3 **para os 180 casos**, ou seja, de acordo com a experiência que serviu de **modelo** a este estudo (E_{11});

Os **valores** a **encarnado** apresentam **melhores índices** de diferentes classificadores (para as métricas mais relevantes) comparativamente com os valores do classifiicador **(SMO)** que evidenciou **resultados mais altos** a todos os níveis aquando da **previsão de *outcome_num* sem densidade de massa** para os **180 casos**;

Os **valores** a **bege** apresentam **índices iguais** de diferentes classificadores (para as métricas mais relevantes) comparativamente com os valores do classifiicador **(SMO)** que evidenciou **resultados mais altos** a todos os níveis aquando da **previsão de *outcome_num* sem densidade de massa** para os **180 casos**;

Todas as **métricas** encontram-se **arredondadas** às **duas casas decimais**;

(1) Os valores relativos às métricas: **"Precision"**, **"F-Measure"**, **"TP Rate"** e **"FP Rate"** dizem respeito à classe **"malignant"**;

**** CCI** – Número de **"Correctly Classified Instances"**;

Os **valores relativos à tabela da pág. 2 são valores de teste**, uma vez que foi estudado o comportamento de certos classificadores que recorrem a algoritmos com **parâmetros diferentes** dos presentes na tabela da pág. 1. Sendo assim, temos:

Tabela pág. 1:

BayesNet – "searchAlgorithm" (TAN)
SimpleCart – "numFoldsPruning" (5 – "default value")
PART – "numFolds" (3 – "default value")
J48 – "numFolds" (3 – "default value")

Tabela pág. 2:

BayesNet – "searchAlgorithm" (K2)
SimpleCart – "numFoldsPruning" (10)
PART – "numFolds" (10)
J48 – "numFolds" (10)

Apêndice D

Gráficos Área ROC

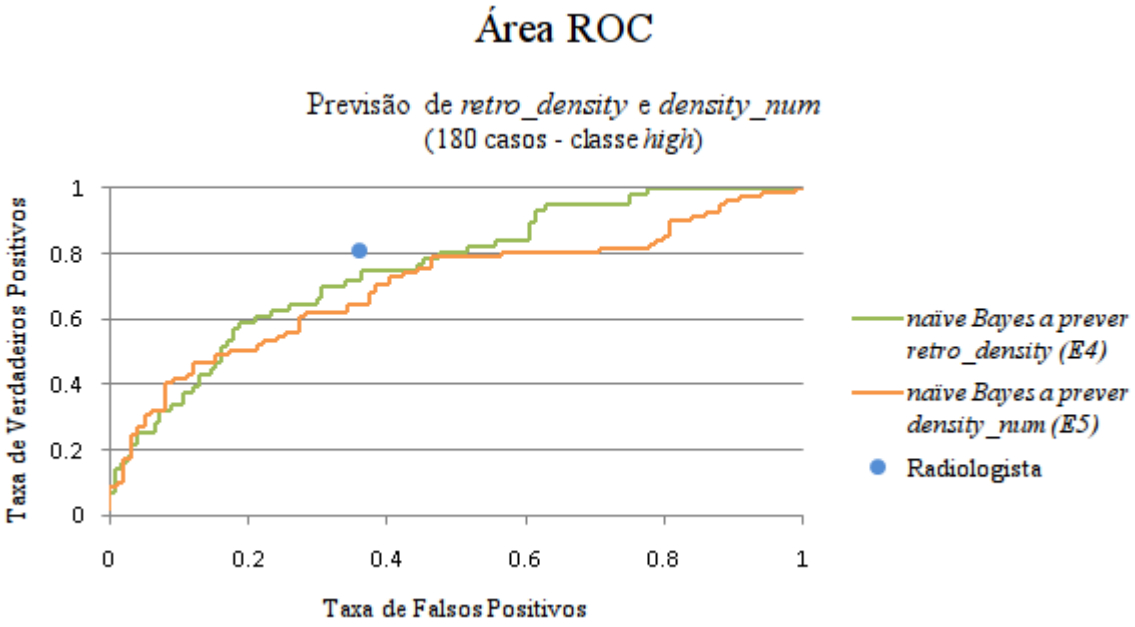


Figura 38 - Espaço ROC: Previsão de densidade de massa (retrospectiva e prospectiva) em relação à classe *high density* por classificadores bayesianos em 180 casos

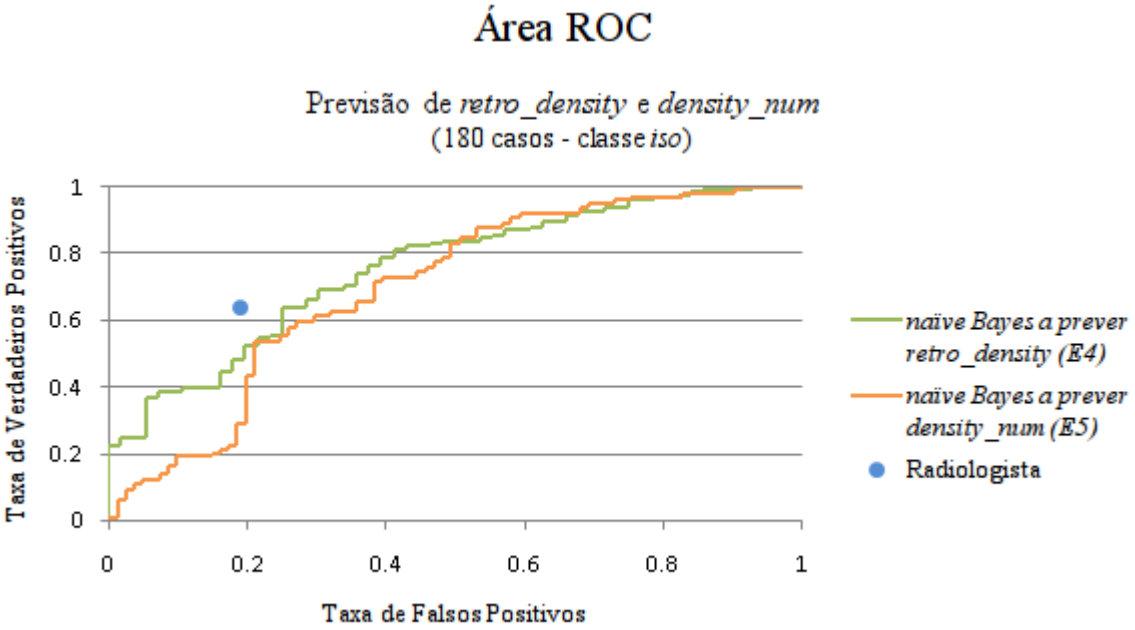


Figura 39 - Espaço ROC: Previsão de densidade de massa (retrospectiva e prospectiva) em relação à classe *iso-dense* por classificadores bayesianos em 180 casos

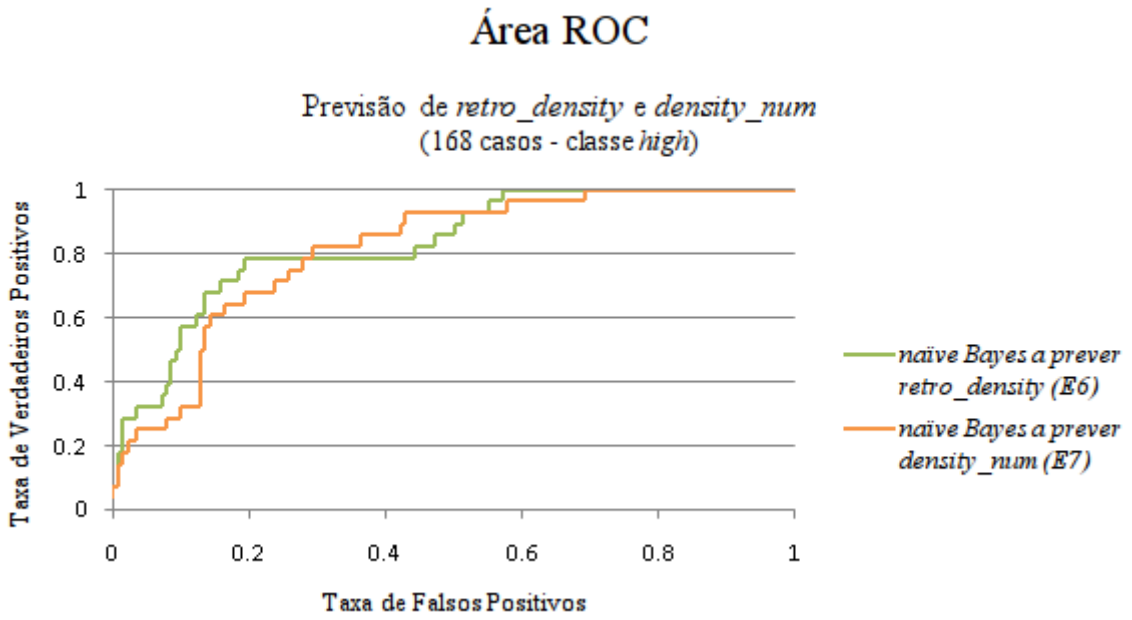


Figura 40 - Espaço ROC: Previsão de densidade de massa (retrospectiva e prospectiva) em relação à classe *high density* por classificadores bayesianos em 168 novos casos

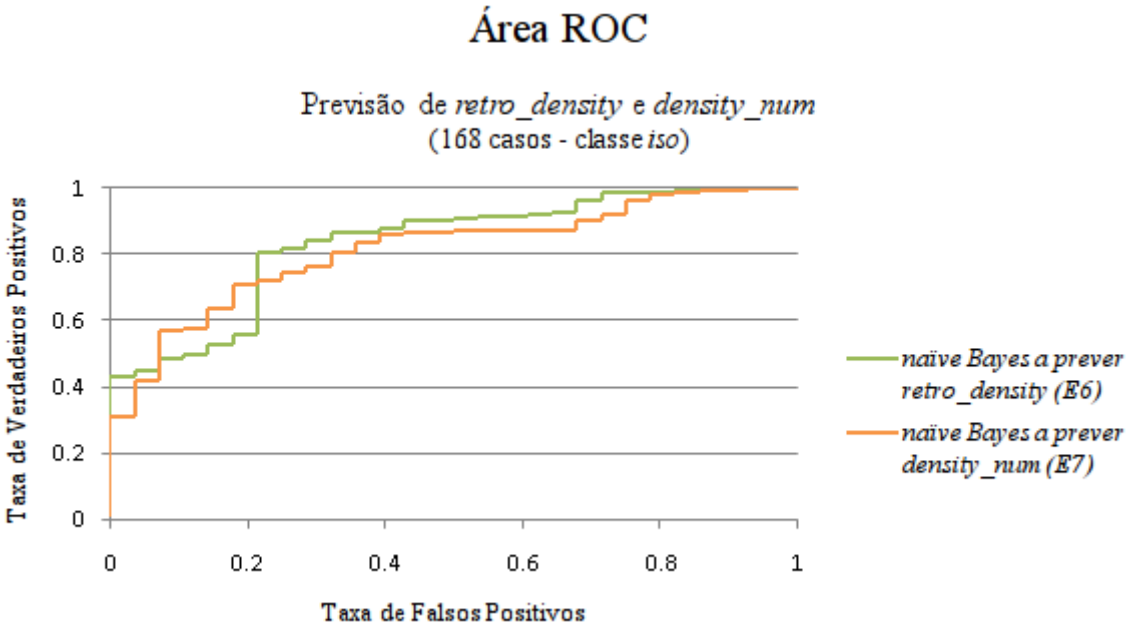


Figura 41 - Espaço ROC: Previsão de densidade de massa (retrospectiva e prospectiva) em relação à classe *iso-dense* por classificadores bayesianos em 168 novos casos

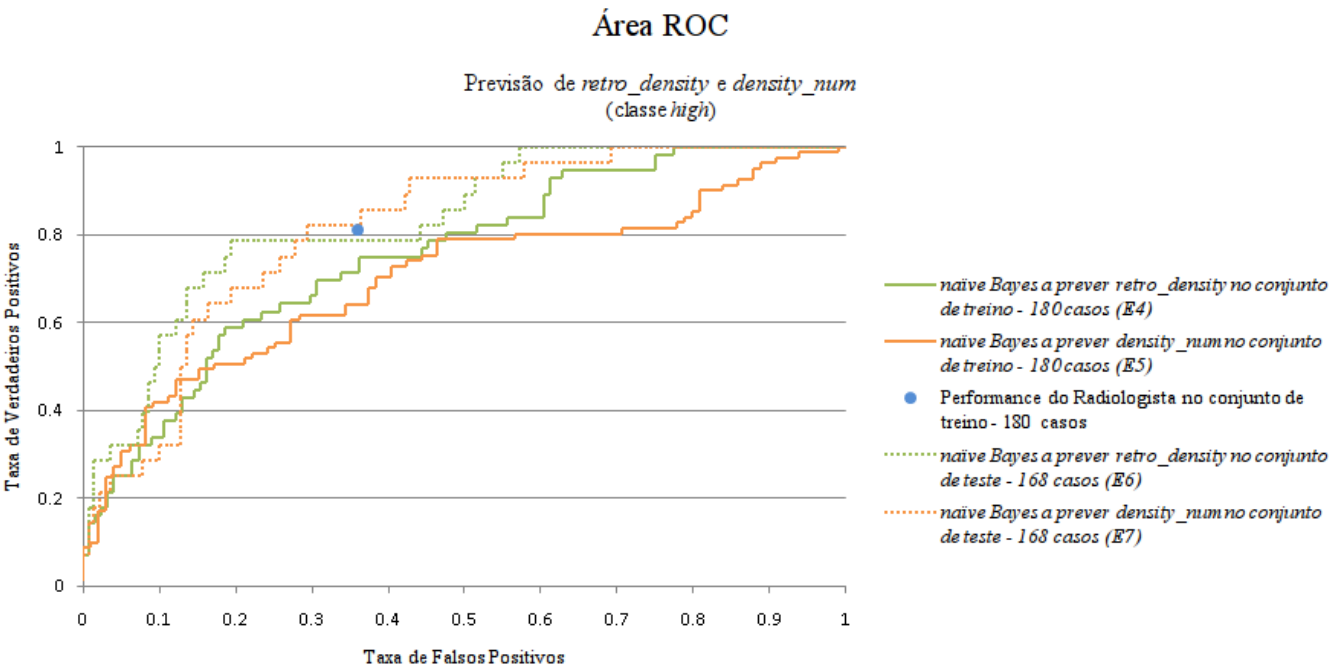


Figura 42 - Espaço ROC: Previsão de densidade de massa (retrospectiva e prospectiva) em relação à classe *high density* por classificadores bayesianos em 180 e 168 casos

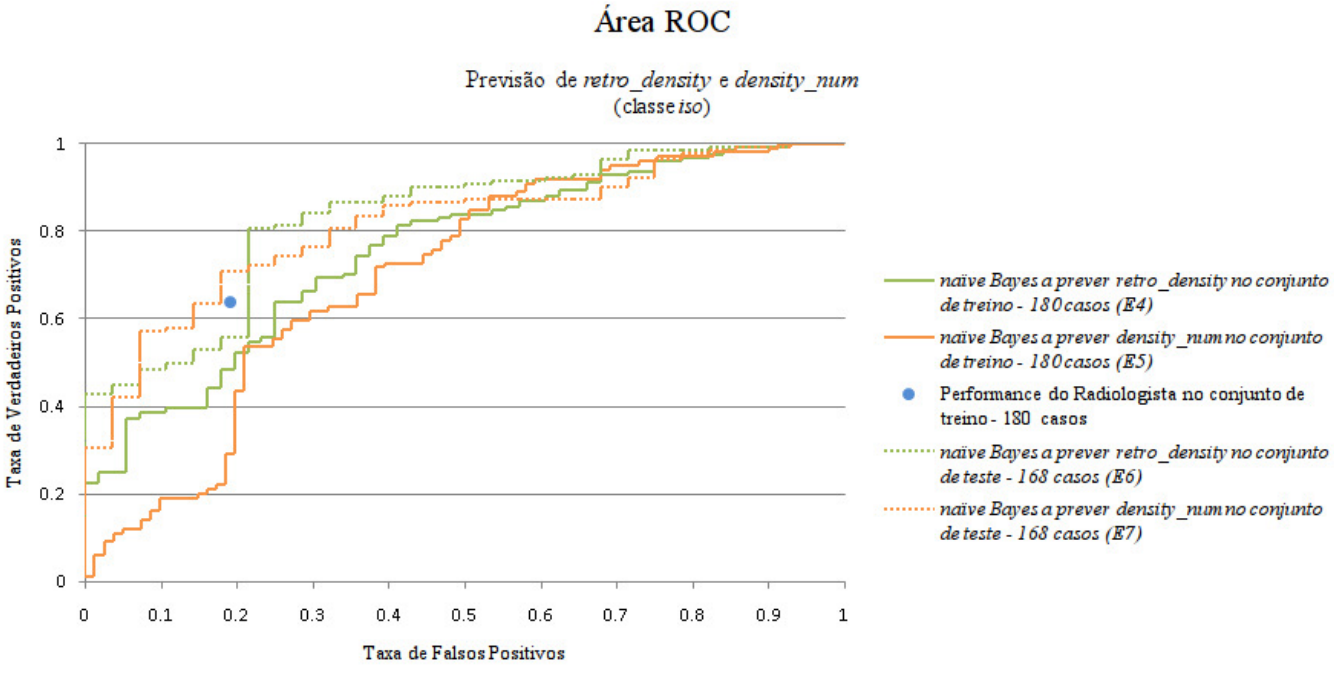


Figura 43 - Espaço ROC: Previsão de densidade de massa (retrospectiva e prospectiva) em relação à classe *iso-dense* por classificadores bayesianos em 180 e 168 casos