• U  C •

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Diogo Fabião Passadouro

Métodos de Inteligência Computacional para Estimar Dados em Falta e Classificação em Dados de Cancro da Mama

Dissertação apresentada à Universidade de Coimbra
para cumprimento dos requisitos necessários à obtenção
do grau de Mestre em Engenharia Biomédica

Orientadores:
Professor Doutor **Pedro H. Abreu**
Professor Doutor **Pedro J. García-Laencina**

Coimbra, 2015

Este trabalho foi desenvolvido em colaboração com:

Centro de Informática e Sistemas da Universidade de Coimbra

Instituto Português de Oncologia do Porto

Abstract

Breast cancer is the most common type of cancer worldwide for women. Even with the recent breakthroughs in treatment methods, it is still one of the major causes of the highest female cancer mortality rates. With the goal of reducing the impact of this disease, computational intelligence methods based on clinical information have been developed over the years. These methods are capable of predicting patient survival and assessing the most effective treatment for the patient.

It is common to observe missing data in datasets. This can lead to the decrease of the number of observations for analysis or to a distortion of models intended to be used. The issue can be solved by ignoring the cases with missing data or predicting the required values based on the remaining data – imputation.

The algorithm *Optimally Pruned Extreme Learning Machine* (OP-ELM) has been used as a classifier in this context due to its good results and good performance and generaliation capabilities. This work proposes its implementation as an imputation method and its adaptation to deal with categorical variables, comparing its performance with other methods used for the same purpose.

Until the development of this work, the OP-ELM had never been used as an imputation method for missing data, therefore the approach presented in this project is completely new. The results that were obtained show that the OP-ELM imputation method leads to similar performances compared to the most widely used imputation algorithms. Moreover, OP-ELM needs smaller computation times, even for datasets with a high percentage of missing data.

Resumo

O cancro da mama é o mais comum no sexo feminino a nível mundial. Apesar dos avanços ao nível das técnicas de tratamento, esta doença é uma das principais causas para o elevado número de mortes por cancro em mulheres. Com o objetivo de reduzir o impacto desta doença, têm-se desenvolvido métodos de inteligência computacional baseados em informação clínica, capazes de prever a sobrevivência dos pacientes e adaptar-lhes o tratamento mais eficaz.

É frequente existirem dados em falta nos *datasets*, o que poderá levar à redução do número de observações para análise ou a uma distorção dos modelos que se pretendem criar. Para resolver este problema existem várias abordagens que podem passar por ignorar os casos com valores em falta ou prever esses valores de acordo com os restantes dados – imputação.

O algoritmo *Optimally Pruned Extreme Learning Machine* (OP-ELM) tem vindo a ser usado como classificador e tem permitido bons resultados nestes contextos, devido aos seus bons resultados e devido ao seu bom desempenho quanto ao tempo de computação e capacidade de generalização. Neste trabalho propomos a implementação deste algoritmo como método de imputação e a sua adaptação para lidar com variáveis categóricas, comparando a sua performance com outros métodos usados para este efeito.

Até ao momento do desenvolvimento deste trabalho não existia qualquer abordagem com o algoritmo OP-ELM para imputação de dados em falta, pelo que a implementação deste algoritmo neste trabalho é uma abordagem completamente nova. Os resultados obtidos mostram que a imputação com OP-ELM conduz a performances semelhantes às obtidas com os algoritmos de imputação mais usados até ao momento, com tempos computacionais inferiores, mesmo em *datasets* com elevadas percentagens de dados em falta.

Agradecimentos

Quero agradecer aos meus orientadores, o Prof. Pedro Abreu e o Prof. Pedro Laencina, por todo o apoio e disponibilidade que demonstraram durante a minha tese. O Professor Pedro Abreu, pela confiança depositada, pela coragem que constantemente transmitia e pela honestidade que sempre assumiu. O Professor Pedro Laencina, pelo enorme profissionalismo e o empenho que demonstrou ao longo deste projeto, a quem dedico esta tese com um pesado sentimento de vermos os bons partirem cedo.

Durante o meu percurso académico tive a sorte de conviver com pessoas que jamais farão parte do meu passado. Ao Diogo Martins, à Mariana Nogueira, à Inês Barroso, à Carolina Silveira, à Heloísa Sobral, à Bruna Nogueira e ao Daniel Osório agradeço os momentos que passámos, que por terem sido tão diferentes e tão genuínos levarei comigo para a vida. À Miriam Santos e ao Bruno Andrade pela ajuda durante este trabalho e pela sua humildade. Ao Ricardo Mendes, ao Pedro Duarte, ao Ricardo Simões e ao Luís Henriques que sempre ajudaram a animar os tempos de convívio.

Quero agradecer à Teresa Ourives pelo seu carinho. A ela devo a grande ajuda para superar algumas etapas ao longo dos últimos dois anos.

Finalmente, estou verdadeiramente grato à minha família. Aos meus pais agradeço terem sempre conseguido ajudar muito com o pouco que foram tendo, e permitirem que este sonho se tornasse realidade. Aos meus irmãos com quem nunca consegui falar a sério, mas com quem eu conto incondicionalmente.

All models are wrong, but some are useful.

George E. P. Box

Conteúdo

Lista de Abreviaturas

ACC Análise de Casos Completos.

ANN *Artificial Neural Network*.

AUC *Area Under the Curve*.

CD *Critical Difference*.

CV *Cross Validation*.

DT *Decision Trees*.

ELM *Extreme Learning Machine*.

EM *Expectation-Maximization*.

FFNN *FeedForward Neural Network*.

FN Falso Negativo.

FP Falso Positivo.

HEOM *Heterogeneous Euclidean-Overlap Metric*.

IG *Information Gain*.

IM Informação Mútua.

IMC Índice de Massa Corporal.

IPO Instituto Português de Oncologia.

KNN K-*Nearest Neighbors*.

LDA *Linear Discriminant Analysis*.

LOO *Leave-One-Out*.

MAR *Missing at Random*.

MCAR *Missing Completely at Random*.

MD *Missing Data*.

MDI *Median Imputation*.

MI *Mean Imputation*.

MImp Imputação Múltipla.

ML *Machine Learning*.

MLP *Multi-Layer Perceptron*.

MNAR *Missing Completely Not at Random*.

MRSR *Multiresponse Sparse Regression*.

MSE *Mean Square Error*.

OP-ELM *Optimally Pruned Extreme Learning Machine*.

P-ELM *Pruned Extreme Learning Machine*.

PRESS *PREdiction Sum of Squares*.

RBF *Radial Basis Function*.

RF *Random Forest*.

SLFN *Single-Layer Feedforward Neural Network*.

SOM *Self-Organizing Map*.

SVD *Singular Value Decomposition*.

SVM *Support Vector Machine*.

TVN Taxa de Verdadeiros Negativos.

TVP Taxa de Verdadeiros Positivos.

UCI *University of California, Irvine*.

VN Verdadeiro Negativo.

VP Verdadeiro Positivo.

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Este projeto foi desenvolvido no Departamento de Engenharia Informática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra, integrado no programa do Mestrado em Engenharia Biomédica. O principal objetivo deste capítulo é fornecer uma visão global deste projeto: as duas primeiras secções baseiam-se no enquadramento e motivação deste projeto e na terceira e quarta secções são definidos os objetivos e o planeamento do projeto. Finalmente é apresentada a estrutura seguida em todo o documento.

1.1 Enquadramento

O cancro é segunda causa de morte em todo o mundo, logo a seguir às doenças cardíacas [1]. Segundo Siegel et al. [1], o cancro da mama ocupa o primeiro lugar nas previsões de novos cancros e o segundo lugar nas estatísticas do número de mortes por cancro em mulheres, prevendo-se que a esta doença se associem 15% das mortes e 29% dos novos casos de cancro em mulheres no ano de 2015, mesmo com os recentes avanços dos tratamentos.

Prover os clínicos de modelos de prognóstico de sobrevivência ao cancro da mama é um dos grandes objetivos da aplicação de técnicas de inteligência computacional nesta matéria [2]. Estas técnicas podem ser úteis, também, para identificação de

subgrupos de pacientes com cancro da mama, com o objetivo de poder estabelecer um tratamento mais personalizado e por isso mais eficaz e menos invasivo [3, 4]. Estes modelos computacionais são construídos com base na informação clínica e histológica do paciente. No entanto, estes conjuntos de dados apresentam, muitas vezes, lacunas, os dados em falta (*Missing Data* (MD)).

1.2 Motivação

Lidar com MD não é uma tarefa simples. Primeiro, é importante existir um bom conhecimento dos dados em estudo e do MD que apresentam. Tratar MD de forma correta, é ainda mais crucial quando estamos a lidar com dados do foro médico, dos quais se pretende fazer inferências clínicas, já que um tratamento errado dos dados em falta pode ter consequências graves na decisão clínica. A abordagem mais comum para lidar com MD é a Análise de Casos Completos (ACC) onde são ignorados quaisquer casos onde ocorra MD. No entanto, esta forma de lidar com o problema dos MD leva a elevadas perdas de informação [5]. Surgiu assim uma nova abordagem: imputação de MD, onde as lacunas deixadas pelo MD são preenchidas com valores estimados de acordo com os dados completos.

Vários são os trabalhos que têm aparecido na literatura para endereçar este problema. A abordagem mais comum para MD é a eliminação de casos [5]. No entanto esta abordagem, por levar à eliminação de muita informação relevante tem vindo a dar lugar às abordagens de imputação com base em aprendizagem automática (*Machine Learning* (ML)). Algoritmos comuns nas áreas de classificação e regressão, foram implementados com o intuito da imputação de MD. Os algoritmos *Multi-Layer Perceptron* (MLP), *Support Vector Machine* (SVM), K-*Nearest Neighbors* (KNN) e *Self-Organizing Map* (SOM) têm vindo a ser implementados para imputação, com um grande contributo no aumento nas performances obtidas comparativamente aos métodos de eliminação de casos ou de imputação pela média ou por regressão [6–9]. Também existem trabalhos na literatura onde os métodos computacionais são aplicados em dados de cancro da mama, com resultados significativos no aumento da performance no prognóstico desta doença [4, 7]. No entanto não existe, até ao que foi possível apurar, nenhum trabalho onde se recorra ao OP-ELM para imputação de MD o que prova o contributo deste trabalho nesta área.

1.3 Objetivos

O principal objetivo deste projeto foi a adaptação do algoritmo OP-ELM a variáveis categóricas e a sua implementação como método de imputação, já que até ao momento apenas foi usado para tarefas de regressão e classificação. A performance deste novo método foi testada e comparada com o desempenho dos métodos de imputação MLP, SVM, KNN e SOM (já utilizados para este efeito), em dois contextos distintos:

- Recorrendo a *datasets* completos avaliar o desempenho do algoritmo de forma individualizada mas também em comparação com os outro métodos de imputação. Nesta abordagem existiu a inserção de MD de forma artificial;
- Na segunda abordagem foram usados *datasets* que originalmente contêm MD. Com estes dados pretende-se avaliar a resposta dos métodos de imputação e classificação que foram testados no cenário anterior, mas num contexto onde não se dispõe dos valores reais para o MD.

1.4 Planeamento

Nesta secção apresenta-se a comparação entre o planeamento inicialmente definido e a calendarização conseguida durante o desenvolvimento deste projeto. Na Figura 1.1 estão esquematizadas a previsão temporal de cada atividade e aquela que foi a duração real de cada uma. O projeto iniciou-se com a definição do trabalho a ser desenvolvido. Uma vez definidos os objetivos, foi necessário fazer um estudo das técnicas que já existiam, suas vantagens de desvantagens, de modo a orientar o trabalho no sentido de acrescentar novidade à arte. Na terceira etapa, o algoritmo OP-ELM foi implementado de modo a poder ser usado como método de imputação. Nas quartas e quintas etapas implementaram-se os algoritmos de imputação e classificação, incluindo o algoritmo OP-ELM proposto, em duas abordagens distintas: uma com *datasets* completos e uma segunda abordagem com *datasets* que originalmente contêm MD. Por último regista-se a escrita deste relatório final.

Olhando para a Figura 1.1 é possível verificar que a etapa de simulação com *datasets* completos adiou a data prevista de conclusão deste trabalho. Durante estas simulações existiram redefinições da abordagem que estava a ser seguida o que levou

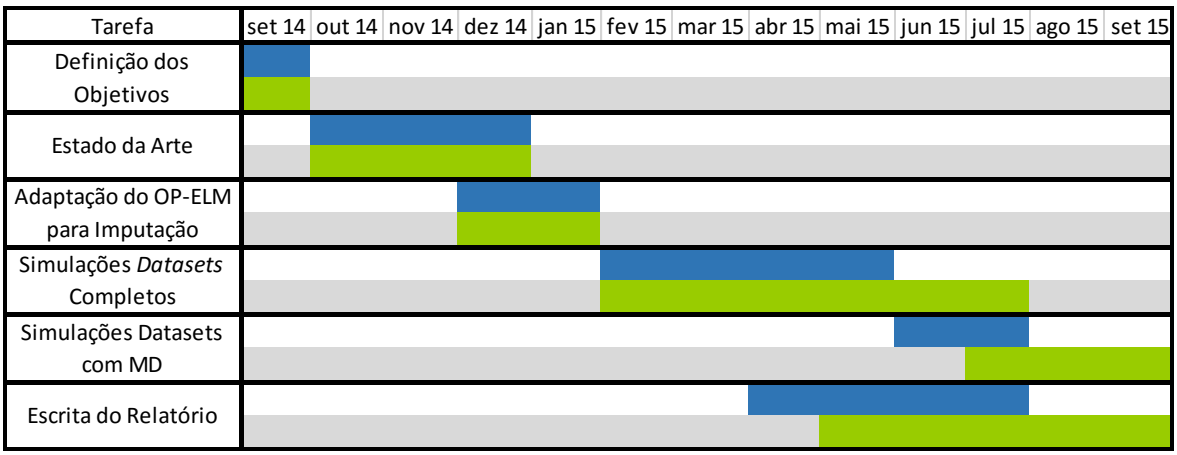

Figura 1.1: Planeamento definido (a azul) VS conseguido (a verde) neste trabalho

a um acréscimo do tempo previsto para esta etapa e a um atraso das simulações com os *datasets* incompletos, que não seria possível iniciar sem os resultados da primeira abordagem. Contudo, é de salientar que os objetivos propostos foram conseguidos.

1.5 Estrutura do Documento

O documento está organizado como se segue: no Capítulo 2 serão descritos alguns fundamentos teóricos essenciais para a compreensão do restante documento; no Capítulo 3 será apresentada uma revisão de alguns estudos que existem relacionados com os métodos para lidar com valores em falta num conjunto de dados; no Capítulo 4 serão dados a conhecer os *datasets*, métodos de imputação e de classificação usados neste projeto apresentando, de forma detalhada, a estrutura seguida durante todas as simulações; no Capítulo 5 serão abordados e discutidos os resultados alcançados; por último, no Capítulo 6 serão apresentadas as principais conclusões e trabalho futuro a ser desenvolvido.

Capítulo 2

Fundamentos Teóricos

Com o desenvolvimento da tecnologia houve um aumento incontornável da quantidade de informação a ser gerada, transmitida e armazenada a todo o momento. No entanto surge frequentemente um problema: a falta de valores nos *datasets* recolhidos. Estas lacunas nos conjuntos de dados, os MD, podem dever-se a múltiplos fatores tais como o responsável pela aquisição ou a problemas com sensores a que se recorre para a obtenção dos dados. Antes, muita da aquisição de dados era feita por alguém e de uma forma manual, o que levava, por vezes, à existência de MD nos dados adquiridos. Exemplo disso são os questionários, que por serem escritos não permitiam o controlo instantâneo das respostas. Estes dados em falta são um problema para muitos investigadores, quando pretendem extrair informação deles, já que os métodos estatísticos convencionais aplicam-se a bases de dados completas (com valores em todas as variáveis e observações) [10]. Para ultrapassar este problema existem duas abordagens distintas que podem ser seguidas. A primeira é a estratégia de análise com os casos completos, onde são descartadas todas as observações que contenham dados em falta. Esta abordagem é usada frequentemente dada a facilidade que apresenta. No entanto, dado que são descartados dados, é descartada informação que pode ser importante. No sentido de preservar essa informação, desenvolveram-se métodos que permitem prever os melhores valores para os MD, de acordo com os dados completos no *dataset*. Estes processos de imputação podem ser divididos em dois tipos: modelos que são baseados em análise estatística e os que têm por base as técnicas de aprendizagem automática (ML).

Durante este capítulo serão dados a conhecer os tipos de MD quanto à sua origem. Num segundo ponto serão apresentadas as principais abordagens para lidar com MD. Depois apresentam-se os classificadores a que se recorreu durante a componente experimental. Por último apresentam-se as técnicas de amostragem e medidas de performance que foram usadas para ilustrar os resultados alcançados neste trabalho.

2.1 Classificação dos Dados Incompletos

O correto conhecimento da origem das lacunas nos *datasets* é essencial para uma escolha acertada da estratégia para lidar com os MD, de forma a evitar distorção dos dados. Assim, antes de se decidir a que estratégia se irá recorrer, deve-se identificar a situação que levou ao aparecimento de MD.

Ao longo do tempo vários foram os autores que propuseram uma classificação para a origem do MD [5, 10, 11]. Rubin et al. [11] classificaram os MD em três categorias: *Missing Completely at Random* (MCAR), *Missing at Random* (MAR) ou *Missing Completely Not at Random* (MNAR).

São MCAR os MD onde a falta de dados numa determinada variável não está relacionada nem com os valores observados, nem com os valores em falta nessa variável, ou com qualquer valor das restantes variáveis. Não existe nenhuma relação entre o acontecimento que leva à falta de dados e os dados em si. Exemplo deste mecanismo é qualquer resposta num questionário que não seja respondida, apenas, por esquecimento de quem o está a preencher.

A origem do MD é classificada como MAR se a existência de dados em falta numa variável não tem relação com os valores que a variável assume nas outras observações, podendo ou não ter relação com as restantes variáveis. Ou seja, se um grupo de observações onde uma variável com MD tiver, em média, maiores ou menores valores nas outras variáveis do que no grupo de observações completas, sem que haja qualquer correlação com os valores da variável em si, então estaremos perante dados MAR.

Por último, se a falta de valores depender da variável onde existe MD, então o mecanismo diz-se MNAR. Ao contrário do que acontece no mecanismo MAR, onde se podem estimar os dados em falta a partir das variáveis completas, no caso de MD do tipo MNAR, tal estimativa não pode ser feita, já que existe uma relação entre a falta de dados e os valores que existiriam nessas lacunas. Exemplo destes dados em falta é

um *dataset* de doentes de cancro da mama, onde ficam por preencher valores referentes a um tratamento que o paciente não realizou. Portanto, os dados em falta não poderão ser estimados corretamente, já que há uma relação entre a falta dos valores e os valores que a variável com MD assumiria.

2.2 Estratégias de Imputação

2.2.1 Métodos por Eliminação de Casos

Uma abordagem simples e que é frequentemente usada para solucionar o problema do MD é a Análise de Casos Completos (ACC), onde é eliminada qualquer observação onde pelo menos o valor de uma das variáveis está em falta [5,10]. A principal vantagem desta abordagem é a possibilidade de recorrer a métodos estatísticos convencionais para extrair informação do conjunto de dados que já não apresenta lacunas, pelo que não requer nenhum método computacional específico. No entanto, dado que existem dados que são descartados, há uma perda de informação associada. Esta perda de informação não é tão prejudicial quando o mecanismo que originou os MD é MCAR, ao contrário do que acontece com mecanismos MAR e MNAR. Se o mecanismo que leva à ocorrência de MD não for MCAR, quando se ignoram as observações que contenham MD, ignora-se também informação relevante, dado que pode existir uma correlação entre a falta de dados e as variáveis do conjunto de dados. Desta forma, quando se extrai alguma informação a partir do conjunto de dados completo o resultado será uma inferência que apresentará um viés ao que seria obtido com os dados originalmente completos. Pela mesma razão, esta abordagem também não é a mais correta quando os dados completos não são uma amostra representativa de todos os dados. Este será o caso quando há MD num grande número de observações e variáveis, o que resultará numa reduzida percentagem de observações completas. Neste caso, ainda que os dados perdidos sejam MCAR, existe uma grande perda de informação, com consequências na qualidade das análises estatísticas aos dados completos.

Incluída ainda nos métodos por eliminação de casos está a abordagem com os casos disponíveis, em que são considerados os casos completos das variáveis necessárias para a estatística que se pretende. Caso um método estatístico não recorra a todas as variáveis, então o método usa, não os casos completos de todo o conjunto de dados, mas

sim os casos completos para as variáveis a que recorre. Uma determinada observação com o valor de uma variável específica em falta, pode ser usada caso não se recorra a essa variável num determinado método estatístico de análise dos dados. Uma vantagem desta abordagem com os casos disponíveis é a utilização de mais dados do que na ACC, já que pode conter observações com MD para variáveis que não são usadas. Ainda assim, apresenta a desvantagem de recorrer, para diferentes estatísticas, a conjuntos de observações diferentes.

Ao contrário do que acontece nos métodos por eliminação de casos, nos métodos de imputação não se eliminam casos, antes pelo contrário, imputam-se os valores em falta. Os métodos de imputação podem dividir-se essencialmente em dois grupos: os métodos baseados em análise estatística e os baseados em ML. As estratégias de imputação apresentadas a seguir, pelo facto de preencherem as lacunas existentes, têm a vantagem de acrescentar informação ao conjunto de dados, ao invés do que acontecia ao remover as observações com casos incompletos. No entanto, esta informação que é acrescentada com a imputação do MD pode contribuir para o aumento do ruído, levando à distorção dos dados, efeito contrário ao que se deseja.

2.2.2 Imputação Baseada em Análise Estatística

Nos métodos de imputação baseados em análise estatística recorre-se a uma medida ou método estatístico para realizar a imputação. Pode recorrer-se apenas à variável cujos valores se pretendem imputar ou recorrer a mais do que uma variável para estimar esses valores. A imputação pela média é um exemplo de uma situação em que se recorre apenas à variável que contém o MD que se pretende imputar. Outro caso de imputação com base na análise estatística é a imputação por regressão, onde se recorre às variáveis que não contêm MD para estimar os valores em falta.

Imputação Baseada na Média/Moda

A imputação pela média (*Mean Imputation* (MI)) consiste no preenchimento das lacunas de uma variável com o valor da média dos casos observados nessa variável. Existe uma variante deste método que consiste na imputação pela média, mas condicionada à classe a que uma observação pertence, ou seja, o valor que será atribuído à lacuna será a média das observações com valor para a variável em questão

e que pertencem à mesma classe. Para evitar a influência que os *outliers* poderão ter na MI, é normal recorrer-se à imputação pela mediana (*Median Imputation* (MDI)), seguindo a mesma lógica que na MI. No caso das variáveis categóricas os dados em falta serão substituídos pela moda dos valores observados para essa variável, em vez da média. Estes métodos de imputação acarretam consigo alguma redução da variabilidade natural dos dados, já que aumentam o número de observações completas disponíveis sem aumentar a informação estatisticamente diferente, uma vez que a média de valores na variáveis onde existe MD permanecerá inalterada. Este método não é muito recomendado quando existem muitos casos a serem imputados na mesma variável ou quando os coeficientes de correlação já estão muito próximos de zero, sendo um método que pode diminuir ainda mais esta correlação entre as variáveis [5, 12].

Imputação Baseada em Modelos de Regressão

A imputação por regressão é um modelo que recorre à modelação das variáveis completas para estimar os valores que irão substituir os dados em falta. No caso dos dados incompletos se concentrarem apenas numa variável, será necessário apenas um modelo de imputação como função de aproximação [5]. Já para os casos em que existe mais do que uma variável com dados incompletos é necessário criar um modelo de regressão multi-variável por cada combinação possível de variáveis com MD. A escolha do método de regressão depende do tipo de dados a que se pretende imputar o modelo. Em dados que não respeitem uma tendência linear não é correto recorrer a um método de regressão linear, mas sim a um método de regressão que ajuste uma curva aos dados. Este tipo de imputação apresenta bons resultados quando as variáveis com MD estão correlacionadas com as variáveis completas. Como vantagem, a imputação por regressão preserva a variância e covariância das variáveis incompletas. No entanto, se as variáveis com MD forem originalmente independentes, a aplicação dos modelos de regressão para a imputação levam à multicolinearidade, dado que os dados imputados irão estar correlacionados com os restantes dados. Assim como na imputação pela média, a imputação por regressão não consegue recuperar a variabilidade original dos dados, já que os dados imputados seguirão uma única curva de regressão.

2.2.3 Imputação Baseada em Técnicas de Aprendizagem Automática

Outra abordagem para imputação são as técnicas de *Machine Learning* (ML). Genericamente, estas técnicas recorrem aos valores disponíveis e aprendem esses valores, permitindo depois dar como entrada os dados para as variáveis completas de uma determinada observação e estimar os valores para as variáveis onde existem MD. A seguir serão dadas a conhecer as seguintes técnicas de ML para imputação: K-vizinhos próximos (KNN), perceptrão multi-camada (MLP), máquina de vetores de suporte (SVM), mapas auto-organizáveis (SOM) e máquina de aprendizagem extrema com base na poda/corte (OP-ELM).

Imputação Baseada em KNN

O método KNN é um método que tem vindo a ser muito usado como método de imputação. A imputação de MD é feita de acordo com o valor tomado pelos K vizinhos mais próximos, vizinhos estes que são encontrados de acordo com a métrica previamente definida. Após encontrado o conjunto dos vizinhos é necessário estimar o valor a colocar nas lacunas. Dependendo do tipo de dados esta estimativa pode ser feita pela moda (no caso de variáveis discretas) ou pela média dos valores que a variável toma nos K vizinhos. Pode ainda ter-se em consideração a distância de cada vizinho e garantir maior contribuição aos vizinhos que estão mais próximos do caso que se pretende imputar, sendo que neste caso se considera uma média pesada, onde o peso associado ao valor é o inverso da distância a esse vizinho:

$$\tilde{x}_j = \frac{1}{K} \sum_{k=1}^{K} w_k \upsilon_{kj} \tag{2.1}$$

Na expressão 2.1, $\tilde{x}_j$ representa o valor que irá substituir o MD. Este valor é calculado como uma média pesada dos K vizinhos. w_k pesará a contribuição que o valor do vizinho k (υ_{kj}) terá no valor com que o MD será imputado. O peso pode ser definido como sendo igual para todos os vizinhos, mas é mais comum que seja o inverso da distância entre a observação a imputar e o vizinho k. Neste caso, quanto mais próximo o vizinho estiver, maior será a sua contribuição para o valor que será imputado.

Ao invés de se considerarem apenas os casos completos no conjunto das variáveis, pode recorrer-se a todos os casos que não estejam incompletos na variável que se pretende imputar. Para que esta situação seja possível é necessário recorrer a uma métrica para a distância que consiga lidar tanto com variáveis discretas ou contínuas como com dados em falta. A métrica *Heterogeneous Euclidean-Overlap Metric* (HEOM) atribui à distância o valor um (distância máxima) entre dois casos de uma variável quando um dos valores está em falta. No caso da variável ser discreta, o valor atribuído para a distância é um ou zero, caso os valores para os dois casos nesta variável sejam iguais ou diferentes, respetivamente. Já no caso de variáveis contínuas, a distância entre os dois casos para esta variável é dada pela razão entre o módulo da diferença dos valores de cada caso e a diferença entre o máximo e o mínimo da variável em causa, isto é, normalizada no intervalo de valores entre zero e um [13]. A vantagem deste método relativamente ao método de imputação utilizando a média (MI) é o facto de ter em consideração apenas os casos mais similares e não todos os casos presentes [6]. No entanto, o método KNN para a imputação, pela necessidade de percorrer todo o conjunto de dados para encontrar os K vizinhos mais próximos, pode levar a um elevado custo computacional, principalmente para grandes conjuntos de dados.

Imputação Baseada em MLP

O MLP é a rede neuronal artificial (*Artificial Neural Network* (ANN)) mais usada em tarefas de classificação ou imputação, pelo facto de ser um aproximador universal [14]. No algoritmo de imputação com base num MLP, recorre-se aos casos completos para treinar o modelo que será utilizado para estimar os valores em falta. Na fase de treino as variáveis completas servem de entrada e as incompletas de saída, com o intuito de estimar os melhores parâmetros da rede, desde o número de neurónios na camada escondida aos pesos que são atribuídos às entradas. Podem distinguir-se duas abordagens distintas: pode ser criado um MLP para cada combinação de variáveis incompletas ou pode ser criado um único modelo para todas as variáveis incompletas. Há ainda quem aplique o MLP a todo o conjunto de dados, com todas as variáveis como entrada e ao mesmo tempo saída do neurónio [15]. Neste caso, o *dataset* é dividido num conjunto de treino e num conjunto de teste aleatoriamente quanto à quantidade de

dados incompletos, com o intuito de lidar com os dados em falta no conjunto de treino e imputar os que faltam no conjunto de teste. A maior desvantagem da utilização do MLP na sua abordagem de treinar uma rede por cada combinação de variáveis incompletas é a quantidade de redes que é necessário gerar e guardar, que será tanto maior quanto maior for o número de variáveis incompletas e quanto maior for o número de variáveis do conjunto de dados.

Imputação Baseada em SVM

O algoritmo SVM é muito usado em tarefas de classificação e regressão, dado o desempenho que apresenta com diversos tipos de dados e com conjuntos de dados de elevadas dimensões [14]. O algoritmo SVM ajusta, durante o treino, a complexidade do modelo e qualidade da tarefa para que foi desenhado, permitindo um balanço que é responsável pela generalização destes modelos a novos dados. Em 2005, Honghai et al. [8] propuseram a aplicação do algoritmo SVM para imputação de MD. Na imputação com SVM os modelos são treinados com os casos completos e tendo como entrada as variáveis completas e como saída as variáveis onde há MD. Desta maneira, semelhante ao que acontece na imputação por regressão ou com o método KNN, os casos com dados em falta serão imputados com os valores da saída do modelo, após a fase de treino, e dando como entrada os valores nas variáveis completas.

Imputação Baseada em SOM

Os mapas auto-organizáveis (*Self-Organizing Map* (SOM)) são um tipo de ANN propostas por Teuvo Kohonen em 1982, que tiveram por base os neurónios do córtex cerebral [16]. Segundo o trabalho deste autor, nos SOMs existe uma interface – mapa – constituída por nodos que são treinados para fazer a transformação dos dados de entrada para uma dimensão menor. Os mapas mais comuns são bidimensionais e, aquando do treino destas redes, mantêm a ordem que existia no espaço original associado aos dados, sendo esta ordem refletida numa relação local entre os nodos próximos. Este aspeto é o que mais diferencia as redes SOM das restantes ANNs [16]. O treino do SOM divide-se em duas fases principais: encontrar o nodo que mais se relaciona com os dados de entrada e a atualização dos pesos associado a este nodo e aos vizinhos mais próximos.

O algoritmo SOM foi aplicado para imputação de MD. Para imputar os dados a rede é treinada com os dados completos, obtendo a estrutura que os mapeia. Uma vez obtido o mapa treinado, dão-se as observações incompletas como entrada para encontrar o nodo que melhor caracteriza os valores em falta, atribuído-lhe assim um valor de acordo com os pesos associado a esse nodo [17, 18].

Imputação Baseada em OP-ELM

Recentemente foi introduzido na área de ML um novo algoritmo: máquina de aprendizagem extrema (*Extreme Learning Machine* (ELM)). Este algoritmo tem vindo a ser estudado como substituto aos métodos convencionais de treino das redes neuronais *feed-forward* (*FeedForward Neural Network*s (FFNNs)). O treino destas redes consiste no ajuste dos parâmetros de entrada dos diversos neurónios da rede de acordo com os dados de entrada e de saída que lhes são fornecidos [19]. As FFNNs, representadas esquematicamente na Figura 2.1, apresentam a grande vantagem de permitirem mapeamentos não lineares entre os dados de entrada e os dados de saída. Apresentam no entanto desvantagens aquando do treino: os algoritmos de treino convencionais levam a um elevado custo computacional e podem atingir mínimos locais distantes do mínimo global da função de avaliação, colocando em causa a sua capacidade de generalização.

Ao contrário do que acontece com os algoritmos de treino convencionais, onde os pesos e os desvios de entrada têm de ser atualizados até encontrar os melhores valores, com o algoritmo ELM é feita uma inicialização aleatória destes parâmetros de entrada das redes neuronais *feed-forward* uni-camada (*Single-Layer Feedforward Neural Network*s (SLFNs)). Assim, o algoritmo ELM permite obter um treino muito mais rápido e uma boa capacidade de generalização, mesmo com funções de ativação não diferenciáveis, superando ainda o problema dos mínimos locais [20].

No que diz respeito ao funcionamento do algoritmo ELM, o passo inicial é a definição do número de neurónios da camada escondida ($\boldsymbol{M}$). De seguida são gerados aleatoriamente os pesos ($\boldsymbol{w}$) e os desvios ($\boldsymbol{b}$) que ligam a camada escondida e a camada de entrada, e é definida a função de ativação ($\boldsymbol{f}$). Para $\mathbf{N}$ casos distintos de

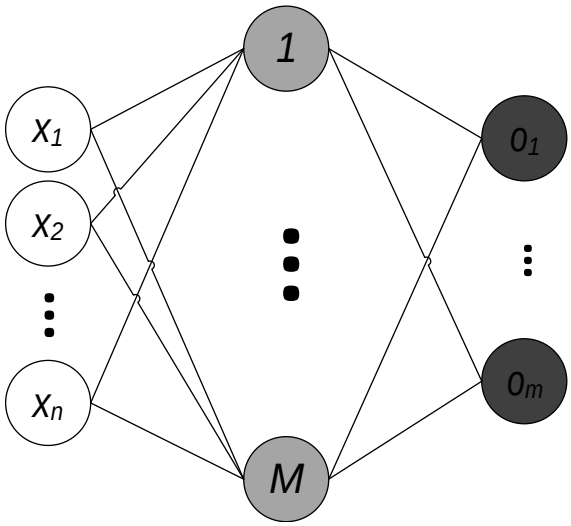

Figura 2.1: Estrutura das FFNN, com apenas uma camada com M neurónios, n valores de entrada e m valores de saída.

treino, a matriz de saída da camada escondida ($\boldsymbol{H}$) é assim representada por:

$$H(\mathbf{w}_1,\ldots,\mathbf{w}_M,b_1,\ldots,b_M,\mathbf{x}_1,\ldots,\mathbf{x}_N) = \begin{bmatrix} f(\mathbf{w}_1\cdot\mathbf{x}_1+b_1) & \ldots & f(\mathbf{w}_M\cdot\mathbf{x}_1+b_M \\ \vdots & \ldots & \vdots \\ f(\mathbf{w}_1\cdot\mathbf{x}_N+b_1) & \ldots & f(\mathbf{w}_M\cdot\mathbf{x}_N+b_M \end{bmatrix}_{N*M} \tag{2.2}$$

Por último falta calcular a matriz dos pesos da camada de saída ($\boldsymbol{B}$). Dado que $\mathbf{HB}=\mathbf{T}$, a matriz dos pesos da camada de saída pode ser obtida por: $B = H^{\dagger}T$, onde $H^{\dagger}$ é a matriz inversa generalizada de Moore-Penrose.

O número de neurónios que constitui a camada escondida da FFNN continua a ser uma variável que necessita ser previamente definida. O valor escolhido para este parâmetro nem sempre é o mais correto, ou porque está subestimado e não é suficiente para mapear totalmente o conjunto de dados, ou porque está acima do necessário e teremos uma rede mais extensa que o que seria desejável, podendo comprometer a capacidade de generalização deste modelo. Surgem assim duas abordagens para solucionar este problema e, iterativamente, encontrar o melhor valor do número de neurónios da camada escondida: (1) os algoritmos construtivos e (2) os algoritmos baseados no corte (*Pruned Extreme Learning Machine* (P-ELM)). Os primeiros são caracterizados por, de forma iterativa, se ir aumentando o número de

neurónios e voltar a treinar a rede até que seja atingindo o critério de paragem previamente definido – seja atingindo o número máximo de neurónios aceite ou o erro fique abaixo do limiar pré-estabelecido. Dado o seu carácter incremental, este tipo de estratégia tende a atingir redes mais pequenas que nos algoritmos por poda, revelando um menor custo computacional no treino das redes neuronais. Este custo computacional também é menor nestes algoritmos dado que se começa com redes pequenas e assim existem poucos parâmetros das redes a definir. A desvantagem mais evidente deste método é a sensibilidade á inicialização aleatória, que pode levar a soluções em mínimos locais [21–23]. Já a segunda abordagem consiste no contrário dos algoritmos por incremento: inicia-se com um número de neurónios maior do que o necessário e iterativamente vão-se eliminando os neurónios menos contributivos de acordo com uma métrica previamente definida. Ainda que o custo computacional do treino destas redes neuronais seja avultado, a vantagem é o facto da convergência destas redes ser mais rápida e menos sensível à inicialização aleatória das condições iniciais [24].

A abordagem mais simples dos algoritmos P-ELM consiste na definição do número inicial de neurónios e na eliminação dos nodos menos relevantes de acordo com a sua contribuição para a classificação final. O limiar de relevância de cada neurónio é definido pelo utilizador e tem influência no tamanho da rede bem como na qualidade da classificação. O teste do χ^2 (qui-quadrado) é normalmente usado para quantificar a relação entre o vetor das classes (o vetor de saída da rede) e os valores nos nodos da rede neuronal, ao mesmo tempo que o ganho de informação (*Information Gain* (IG)) permite quantificar a informação que cada nodo acrescenta à saída da rede neuronal. O algoritmo P-ELM está apenas preparado para tarefas de classificação [22, 25]. Assim, surgiu o algoritmo OP-ELM que é capaz de resolver também tarefas de regressão e onde a eliminação dos neurónios associados às variáveis irrelevantes não exige a definição prévia de um limiar de erro [26].

Na Figura 2.2 estão representados os três passos principais do algoritmo OP-ELM: (1) construir a SLFN, (2) ordenar os neurónios de acordo com a sua utilidade e (3) eliminar os neurónios de acordo com o critério deixar-um-fora (*Leave-One-Out* (LOO))

- A construção da SLFN é feita com a geração aleatória dos pesos e com um número de neurónios acima do que é necessário para o problema, como é feito nos

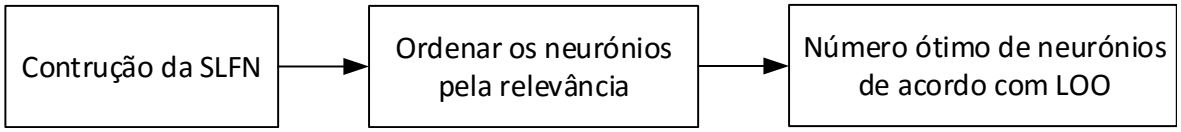

Figura 2.2: Três principais passos do algoritmo OP-ELM

algoritmos P-ELM. A diferença desta SLFN com as que são usadas no algoritmo base ELM é a escolha das funções de ativação. Enquanto que a metodologia ELM recorria apenas a um tipo de funções de ativação (sigmoide, gaussiana ou linear), no caso do OP-ELM podem combinar-se estas funções para obter um modelo mais robusto [26];

- Na segunda fase deste algoritmo o objetivo é ordenar os neurónios de acordo com a sua utilidade. Neste ponto recorre-se à técnica de Regressão Multi-Resposta Esparsa (*Multiresponse Sparse Regression* (MRSR)) definida por Simila *et al.* em [27]. No algoritmo MRSR as colunas da matriz de saída da camada oculta são adicionadas iterativamente a uma nova matriz H^N com os neurónios ordenados. No início do processo, a matriz H^N é uma matriz de zeros com tantas colunas como o número de neurónios da camada escondida. Durante o processo, a coluna de H que representa o neurónio a que corresponde a maior correlação com os resíduos do modelo é adicionada à matriz H^N. Assim, no final teremos uma matriz de saída da camada oculta organizada por ordem de relevância dos neurónios. Dado que nos algoritmos ELM há uma relação linear entre a saída da rede e os nodos iniciados aleatoriamente, a ordem obtida pelo MRSR é exata [20];

- A última fase do algoritmo OP-ELM consiste em definir o melhor número de neurónios da camada escondida. Para tal é comum recorrer ao critério LOO, com tantas iterações quantas o número de variáveis, onde em cada iteração uma das amostras é deixada de fora do treino para ser usada para validação. O maior problema deste critério é o custo computacional em conjuntos de dados com um elevado número de amostras [14]. Assim, no caso do OP-ELM recorre-se a um método exato para calcular o erro do LOO usando a estatística de predição da soma dos quadrados (*PREdiction Sum of Squares* (PRESS)). O número ótimo de neurónios para a rede SLFN é o que permitir o valor mais baixo para o erro

estimado. Comparativamente ao algoritmo P-ELM, este modelo é preferível dado que não há necessidade de definir nenhum limiar do erro para encontrar o melhor número de neurónios. O facto de se recorrer à ordenação por relevância com o algoritmo MRSR permite obter uma convergência mais rápida do algoritmo LOO e leva a redes mais pequenas com a mesma performance [26].

O algoritmo OP-ELM pode assim resumir-se nos seguintes passos:

Algoritmo *Optimally Pruned Extreme Learning Machine* (OP-ELM)

Dado: Conjunto de treino, funções de ativação e o número de neurónios M elevado:

1: Geração aleatória dos pesos e desvios de entrada $\{\mathbf{w}_i, b_i\}_{i=1}^{M}$;

2: Cálculo da matriz de saída da camada escondida $\mathbf{H}$, de acordo com a expressão 2.2;

3: *Ranking* de $\mathbf{H}$ de acordo com o algoritmo MRSR;

4: Para $k = 1$ a M:

– Adicionar o neurónio k ao modelo $H^k = [H^{k-1}, h_k]$, onde h_k é a coluna k de $\mathbf{H}$ após *ranking*;

– Cálculo do erro LOO com H^k;

5: Escolha do número de neurónios $\mathbf{M^*}$ que conduziu ao menor erro LOO;

6: Cálculo da matriz dos pesos da camada de saída: $B = H^{\dagger}T$;

O modelo seguido neste projeto para imputação com recurso ao algoritmo OP-ELM está representado na Figura 2.3. A imputação com OP-ELM segue os seguintes passos:

1. Divisão do *dataset* em observações completas e observações com MD;

2. Treinar, para cada combinação de variáveis com MD, uma SLFN com o algoritmo OP-ELM. Como entrada são dadas as variáveis completas e como saída a variável que contém MD.

3. Dar como entrada às redes treinadas no ponto anterior os valores completos das observações que contêm MD e preencher as lacunas com os valores obtidos como saída das redes.

Dado que se pretende a aplicação do algoritmo OP-ELM para imputação de variáveis categóricas, são necessárias algumas adaptações. O algoritmo OP-ELM está preparado para lidar com variáveis categóricas como entrada, devolvendo, no entanto, valores contínuos para o MD dessas variáveis. Portanto, o OP-ELM necessita ser adaptado no sentido de devolver valores categóricos para o MD dessa variáveis. Para

tal preparou-se o algoritmo OP-ELM para lidar com estas variáveis recorrendo à codificação 1-de-C. Desta forma, cada variável categórica dá origem a C variáveis binárias, onde a variável C^j tomará o valor 1 se a observação pertencer à categoria j e o valor 0 caso não pertença a esta categoria. Depois de ser realizada a previsão dos melhores valores para as observações com MD, é feita a codificação no sentido contrário e o MD será substituído com a categoria que lhe é atribuída.

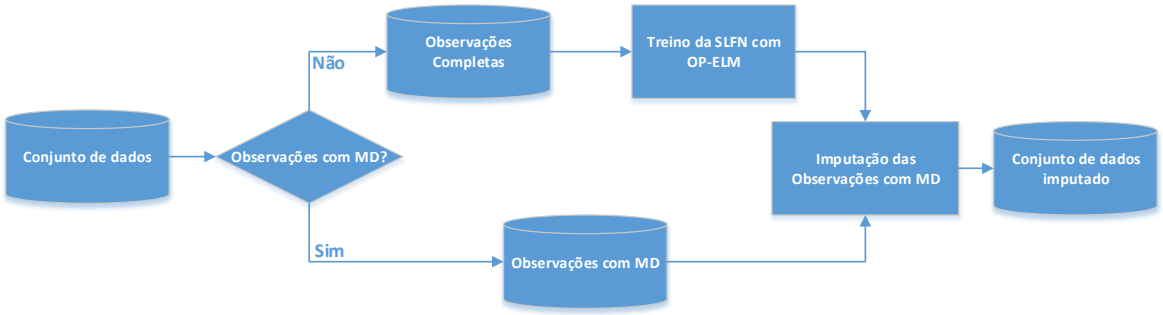

Figura 2.3: Representação do modelo de imputação com redes neuronais treinadas com OP-ELM

2.3 Métodos de Classificação

Ao longo deste projeto, quer seja na revisão da literatura no Capítulo 3, quer na componente experimental deste trabalho, será necessário recorrer a classificadores. A classificação consiste na atribuição de uma classe a um caso, ou conjunto de casos, que são fornecidos como entrada do modelo. Numa primeira fase é necessário treinar o classificador, com o conjunto de dados de treino e o vetor que contém as classes que lhes são previamente atribuídas. Normalmente durante o processo de treino são estabelecidos os parâmetros do classificador que levam aos melhores resultados, pelo que há uma amostragem do conjunto de treino no sentido de ter dados que validem cada parâmetro a testar. Numa segunda fase, fornecem-se dados de teste a este classificador, que devolverá as classes a que cada caso ou observação pertencem. Estes dados de teste são dados que nunca foram 'apresentados' ao classificador, no sentido de entender qual o comportamento do modelo a dados novos. Esta classificação pode ser comparada com o vetor das classes de teste, se disponível, para perceber a qualidade da classificação.

Neste projeto escolheu-se recorrer a quatro classificadores: KNN, SVM, *Random Forest* (RF) e MLP, apresentados a seguir.

2.3.1 K-Vizinhos Próximos

A classificação recorrendo ao KNN tem por base a mesma ideia que foi apresentada na Subsecção 2.2.3:

1. O primeiro passo consiste no cálculo da distância de cada caso que se pretende classificar aos restantes casos;

2. Recorrendo às distância calculadas, encontram-se os K vizinhos mais próximos, com a menor distância à observação a ser classificada;

3. Atribuí-se a classe que estiver em maioria nesses K vizinhos, podendo ser considerada uma abordagem que atribua mais peso às classes dos casos mais próximos, mediando esse peso pela distância, como foi também apresentado no algoritmo KNN para imputação.

Com este classificador há a necessidade de definir, por validação cruzada (*Cross Validation* (CV)), o valor de K que conduz aos melhores resultados.

2.3.2 Máquina de Vetores de Suporte

O algoritmo SVM consiste num modelo supervisionado de aprendizagem muito usado para classificação. Este modelo 'aprende' os dados de treino, criando uma superfície que ao mesmo tempo maximiza a divisão das classes minimizando o erro de classificação desses dados. Este método recorre a *kernels* para mapear os dados num espaço de maior dimensão, dado que muitas vezes não são separáveis linearmente. Portanto, é necessário definir o tipo de *kernel* que otimiza a classificação, ao mesmo tempo que se procuram os melhores parâmetros para essa função de mapeamento. Dando como entrada o conjunto de dados e saída o vetor de classes, a superfície de separação das classes é ajustada com o objetivo de permitir a classificação com melhor resultado.

2.3.3 Árvores de Decisão

Na área de ML existem estruturas que são frequentemente usadas tanto para tarefas de classificação como as de regressão: árvores de decisão (*Decision Trees* (DT)). Estas árvores caracterizam-se por ligar as folhas da árvore (que representam as classes no caso de uma classificação) às variáveis de entrada, por intermédio de ramos que são obtidos por divisões sucessivas dessas variáveis ou de combinações delas. Estas divisões que originam os vários ramos definem uma regra que divide os dados de acordo com o seu valor. Entretanto têm surgido muitas abordagens paras as DT. Neste trabalho recorre-se ao classificador 'floresta aleatória' (*Random Forest* (RF)). Este algoritmo foi proposto por Breiman em 2001 [28]. A lógica de base do RF é a criação de conjunto de DTs e dar-lhe como entrada um subconjunto aleatório das variáveis dos conjunto de dados original. No fim o resultado da classificação é dado pela moda das classificações de cada árvore individualmente [28]. O parâmetro que tem de ser definido neste algoritmo é o número de árvores usado.

2.3.4 Perceptrão Multi-Camada

O percetrão é uma ANN mais básica, sendo constituído apenas por uma camada de neurónios. O maior inconveniente desta estrutura é que apenas está preparada para resolver problemas que sejam linearmente separáveis. Para problemas não lineares é necessário recorrer a mais que uma camada de neurónios, os percetrões multi-camada (MLPs). Dado que possuem mais que uma camada de neurónios conseguem criar uma superfície não linear capaz de separar os dados. A estrutura mais comum para o MLP é constituída por três camadas de neurónios: a camada de entrada, com tantos neurónios quantas as variáveis de saída; a camada escondida, que terá um número de neurónios consoante as características dos dados; e a camada de saída, que terá um número de neurónios consoante o problema para que é definido – para efeitos de classificação terá um neurónio que devolve a classe. O treino destas redes consiste na escolha do número de neurónios da camada escondida e no ajuste dos pesos de cada um dos neurónios. Com um ajuste correto deste parâmetros consegue-se definir um aproximador universal de funções contínuas, a base do classificador com este algoritmo [29]. Assim, para o treino destas redes fornecem-se o *dataset* de treino como entrada e o vetor de classes como saída e com o algoritmo de treino adequado (retro-propagação é o mais comum)

ajustam-se os pesos da rede. O número ótimo de neurónios da camada escondida tem de ser definido por CV.

2.4 Estratégias de Amostragem

Para avaliar um classificador é necessário encontrar a taxa de erro verdadeira, taxa esta que corresponde a toda a população em estudo. Esta situação é ideal, uma vez que na realidade não se consegue apurar este erro para toda a população, já que os dados que existem são apenas uma amostra de toda a população. Uma abordagem errada consiste na utilização de todos os dados como treino e teste, uma vez que o erro que se obtém com o conjunto de teste será sobre-estimado, já que o classificador foi treinado com os mesmo dados. Assim, surgem estratégias de amostragem, em que o objetivo é ter dois conjuntos de dados: os dados de treino e os dados de teste.

Hold-out Este método consiste na divisão do conjunto de dados original em dois conjuntos de dados disjuntos de acordo com uma percentagem definida. É muito comum a divisão do conjunto de dados em dois conjuntos, um com 70% e outro com 30% dos dados, para treino e teste, respetivamente. Ainda que este método seja o mais simples, acarta alguma limitações, principalmente quando estamos perante *datasets* pequenos, já que os dados que servirão para treino podem não ser representativos de toda a população, informação que ficará retida nos dados de teste.

Validação cruzada com K-*folds* (CV) Esta técnica de amostragem consiste, inicialmente, na divisão do *dataset* em K subconjuntos. O conjunto de treino será constituído por K-1 subconjuntos enquanto que o conjunto sobrante será utilizado para teste. Serão assim realizadas K simulações de modo a que cada um dos subconjuntos possa servir uma vez de conjunto de teste. O erro final será uma média dos erros obtidos em cada uma das K simulações. Neste tipo de divisão do conjunto de dados o valor a atribuir ao K é o maior entrave. Um valor alto para este parâmetro permite obter um erro médio baixo mas com grande variância, ao contrário do que acontece para um valor de K pequeno [19]. Existe uma estratégia comum ao nível da CV, que consiste na divisão do *dataset* original em partições de uma forma equilibrada quanto ao número de classes que cada partição tem. Esta

estratégia de CV estratificada mantém em todas as partições a mesma razão de classes que no conjunto de dados inicial, evitando qualquer contribuição de um não balanceamento de classes.

Leave-One-Out Esta técnica é um caso especial do método anterior, quando o número de partições K é igual ao número de observações N do conjunto de dados original. Dado as consequências da escolha do valor de K descrita no método anterior, é evidente que neste caso teremos um erro médio baixo, mas uma elevada variância, o que leva a aceitar este método principalmente quando se está perante um conjunto de dados com poucos casos, por isso baixo valor de N.

Normalização Ainda que não sejam estratégias de amostragem, as técnicas de normalização/estandardização enquadram-se no pré-processamento dos dados. Esta etapa do processamento é crucial para que variáveis que inicialmente sejam representadas em escalas e intervalos de valores diferente sejam niveladas na contribuição que irão ter em tarefas como a de classificação [30]. Uma das estratégias mais usadas no pré-processamento de dados é a normalização ou transformação mínimo-máximo, que consiste em redimensionar os valores da cada variável para que integrem o intervalo [0 1]:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2.3}$$

Uma abordagem não menos comum é a estandardização, em que os dados são transformados de modo a que sigam uma distribuição normal, ou seja, com média zero e desvio padrão um:

$$X' = \frac{X - \mu}{\sigma} \tag{2.4}$$

Informação Mútua Outro ponto que se enquadra, no pré-processamento, é o cálculo das variáveis mais importantes para a classificação. Neste projeto recorreremos à Informação Mútua (IM), que mostra ser uma boa medida para a medida da relevância das variáveis para a classificação [31, 32]. Nesta métrica, é devolvido um valor de IM para cada variável de acordo com a contribuição que têm para a classificação, e por isso, a contribuição que têm para o vetor das classes. Assim, quanto maior o valor da IM mais relevante é variável associada a esse valor.

2.5 Medidas de Performance

A avaliação da performance de um classificador tem normalmente por base a matriz de confusão, representada na Tabela 2.1. Nesta tabela são cruzados os vetores das classes que são originalmente conhecidas com as que são obtidas depois da classificação, associando a cada observação um dos atributos da tabela.

Tabela 2.1: Matriz de Confusão

		Classe Real	
		Positivo	Negativo
Classe Prevista	Positivo	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	Negativo	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Existem algumas métricas que têm por base os valores da matriz de confusão. A precisão corresponde à razão entre os casos corretamente preditos positivos e todos os casos classificados como positivos:

$$Precisão = \frac{VP}{VP + FP} \tag{2.5}$$

A Taxa de Verdadeiros Positivos (TVP), ou sensibilidade, devolve a capacidade que o classificador tem em detetar os elementos que são realmente positivos:

$$TVP = \frac{VP}{VP + FN} \tag{2.6}$$

A mesma ideia é seguida para a especificidade ou Taxa de Verdadeiros Negativos (TVN), mas neste caso pretende-se medir a habilidade para detetar corretamente os casos negativos:

$$TVN = \frac{VN}{VN + FP} \tag{2.7}$$

Como medida que engloba as duas anteriores surge a *accuracy*, que representa a taxa de predições que foram classificadas corretamente:

$$Accuracy = \frac{VN + VN}{VP + VN + FP + FN} \tag{2.8}$$

Outra medida a que se pode recorrer para comparar dois vetores diferentes, quer seja em tarefas de classificação ou regressão, é o erro médio quadrático (*Mean Square*

Error (MSE)) que quantifica a diferença entre o vetor obtido ($\hat{Y}$) e o vetor real (Y), para todas os N casos em análise:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i - Y_i)^2 \tag{2.9}$$

2.6 Comparação de Algoritmos

Neste projeto, a comparação dos diferentes algoritmos será realizada com base no teste não paramétrico de Friedman, seguido do teste de Nemenyi [33]. O teste de Friedman é utilizado para comparar a performance de k algoritmos em N *datasets* distintos, com base em duas etapas:

1. Num primeiro passo, os k algoritmos são ordenados (*ranked*) para cada um dos N *datasets*, separadamente, de acordo com a sua performance. Em caso de empate, á atribuído a média dos *ranks* envolvidos.

2. A estatística de Friedman (F_f) é calculada de acordo com os *ranks* atribuídos:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right], \quad F_f = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2} \tag{2.10}$$

   onde R_j corresponde ao valor médio dos *ranks* para os N datasets em causa.

O valor desta estatística é comparado com o valor crítico para a distribuição de F com $k-1$ e $(k-1)(N-1)$ graus de liberdade. A hipótese nula que estabelece que todos os algoritmos são estatisticamente equivalentes. Neste teste, a hipótese nula pode ser rejeitada caso o valor de F_f seja superior ao valor crítico $F(k-1,(k-1)(N-1))$.

Uma vez rejeitada a hipótese nula, pode realizar-se o teste de Nemenyi para comparar todos os classificadores entre si [33]. Para este teste é necessário calcular o valor da diferença crítica (*Critical Difference* (CD)), onde o valor crítico q_α é baseado na estatística *studentized range* dividida por $\sqrt{2}$:

$$CD = q_\alpha\sqrt{\frac{k(k+1)}{6N}} \tag{2.11}$$

Este teste estabelece que dois algoritmos apresentam uma performance estatisticamente diferente se os *ranks* médios para os algoritmos considerados apresentarem uma diferença de, pelo menos, o valor da CD.

Capítulo 3

Revisão da Literatura

Para o sucesso deste projeto é importante que se conheça o estado da arte. Por um lado, é crucial perceber que abordagens são seguidas no estudo de métodos de imputação e classificação e, por outro lado, perceber que resultados foram conseguidos. Ao longo deste capítulo serão dados a conhecer alguns trabalhos experimentais onde são testados algoritmos de imputação de MD e de classificação. Haverá o cuidado de evidenciar as técnicas, os *datasets* e os resultados a que os autores chegaram.

3.1 Trabalhos Relacionados

Em 2001, Troyanskaya et al. [6] testaram três técnicas de imputação em dados de *microarrays* de ADN. O objetivo destes investigadores era entender qual o melhor método para preencher as lacunas dos dados de que dispunham. Para tal recorreram às observações completas e eliminaram dados numa percentagem de 1 a 20%, de modo a poderem verificar a influência da quantidade de MD na qualidade da imputação. Dado que utilizaram conjuntos de dados que estão, inicialmente, completos, puderam comparar o conjunto de dados original com os dados imputados. Como métodos de imputação usaram os métodos KNN, MI e *Singular Value Decomposition* (SVD), para os quais testaram vários valores para os parâmetros de cada algoritmo. O algoritmo SVD devolve uma transformação dos dados que permite obter os valores em falta por regressão dos valores próprios desta transformação. Neste método, o número ou percentagem de valores a considerar para a regressão é o parâmetro a definir.

Neste trabalho, os autores verificaram que o método KNN apresenta bons resultados mesmo para MD de observações que se encontram expressas em *clusters* pequenos, algo que não acontece para os outros dois métodos testados, onde uma minoria perde a expressão como na imputação pela média. É notável, para a imputação com KNN, que houve apenas um pequeno aumento do erro da imputação mesmo para a percentagem máxima de MD testada – 20%. O método KNN apresentou resultados similares para valores de K no intervalo 10–20, apresentando maus resultados para valores abaixo ou acima deste intervalo. O algoritmo SVD apresenta imputações com menor erro, mas esta técnica é mais sensível ao tipo de dados usados. Dados com maior ruído corrompem mais a qualidade da imputação com SVD em comparação com KNN. A imputação pela média apresenta os piores resultados deste estudo ainda que seja uma boa alternativa, segundo os autores, à simples substituição dos MD por zeros.

Em 2002, Batista et al. [13] compararam a técnica de imputação com base no KNN com outras duas técnicas de ML: CN2 e C4.5. Os autores recorrem a três *datasets* do repositório da *University of California, Irvine* (UCI) [34]: *Bupa*, *Cmc* e *Pima*, conjuntos de dados que não apresentam, originalmente, MD. Na análise experimental deste trabalho, os autores dividem os *datasets* em 10% pares de dados de treino e teste pela aplicação de validação cruzada com 10 partições. São criadas lacunas nos dados de treino em percentagens que vão dos 10 aos 60%. Na execução experimental são feitas duas abordagens: (1) uma em que o algoritmo KNN é usado para imputar as lacunas dos dados de treino, dados que depois de completos darão entrada nos algoritmos CN2 e C4.5; e (2) uma abordagem em que os dados não são corrigidos quanto aos MD, dados com que os algoritmos de classificação já referidos lidam internamente. Em ambas as abordagens os algoritmos CN2 e C4.5 são os classificadores usados para fazer as comparações pretendidas.

Os resultados mostram que na generalidade das experiências realizadas o algoritmo KNN para a imputação supera os mecanismos internos dos algoritmos CN2 e C4.5 para lidar com MD, já que permite obter um menor erro na classificação com o conjunto de dados de teste. Há uma evidência importante neste trabalho: o algoritmo KNN apresenta boa capacidade quando lida com altas percentagens de MD, que chegam, neste trabalho, a 60% do número total de observações do conjunto de dados.

Também no ano de 2002, Fessant et al. [18] apresentaram um estudo em que aplicavam o método SOM, por um lado, para imputação e, por outro lado, como

algoritmo para detetar dados errados. Como método de imputação, o algoritmo SOM é aplicado como se encontra descrito no Capítulo 2. Para a deteção de dados errados o esquema é simples: primeiro a observação é 'apresentada' ao mapa e é escolhido, por minimização da distância, o nodo que mais se relaciona com essa observação. É depois definido um *coeficiente de representação* que relaciona a distância da observação ao nodo que melhor a define com a distância média das observações associadas ao mesmo nodo. Dependendo do *threshold* que o utilizador define para o coeficiente de representação, a observação pode ser considerada correta ou errada. Durante a execução experimental, recorreram aos métodos MI, *hot-deck*, MLP, com a criação de tantos modelos quantas as variáveis com MD, ao contrário da imputação com SOM, onde apenas criaram um modelo para todo o conjunto de dados. Decidiram usar um mapa bidimensional e quadrado $(7 * 7)$, para evitar o efeito borda. As simulações foram realizadas com um conjunto de dados referente a um inquérito sobre os transportes pessoais em França.

Os resultados para os métodos de imputação testados mostram que a imputação com SOM atinge, com apenas um modelo para todo o conjunto de dados, resultados muito semelhantes ao restantes métodos. A imputação com MLP conduziu, no caso das variáveis numéricas, aos melhores resultados, o que os autores associam à capacidade do MLP como função de aproximação. No caso da deteção de erros, o algoritmo SOM apresenta bons resultados, mas necessita de haver bom senso na definição do *threshold*. Neste trabalho ficam assentes dois pontos importantes a favor do SOM: a sua capacidade de lidar com vários tipos de variáveis (numéricas, categóricas) e também o facto de ser um modelo que permite a visualização para interpretação dos resultados, ao contrário, por exemplo, da imputação com MLP.

Já em 2004, Acuña et al. [9] testaram quatro métodos de imputação em doze *datasets* do repositório UCI [34]. Foram testados os algoritmos de imputação ACC, MI, MDI e KNN. Para classificação os autores recorreram aos métodos *Linear Discriminant Analysis* (LDA) e KNN. Para comparar os resultados dos diferentes algoritmos de imputação, os autores basearam-se na taxa de erro de classificação obtida com os dois classificadores usados. Na implementação experimental para quatro *datasets* com MD apenas foram utilizadas as variáveis mais relevantes, selecionadas de acordo com um método de filtragem dos mesmos autores. Foram ainda eliminadas as variáveis cuja percentagem de MD fosse maior que 30% e todas

as observações com mais de 50% de dados em falta. Nos restantes *datasets* foram criadas lacunas, de modo aleatório e proporcional com o número de observações por classe, em percentagens entre 1 e 20%. Para as duas situações, a estimação do erro de classificação realizou-se por validação cruzada com 10 partições.

Os resultados mostram que para *datasets* com menor percentagem de observações com MD não há uma diferença muito evidente entre o algoritmo ACC e os métodos de imputação, diferença esta que se torna evidente em conjuntos com maior número de observações. Não se verificou uma grande diferença entre MI e MDI, o que pode ser explicado pela existência de *outliers* nas duas direções da distribuição dos dados, levando a um efeito anulador do que se esperaria na presença desses valores extremos. A ACC apresenta os piores resultados para os *datasets* *Sonar*, *Breast* e *German* [34] comparativamente ao outros dois métodos de imputação, algo que pode ser explicado pela distribuição dos MD numa maior percentagem de observações. Por fim, os autores concluem que a imputação com KNN é a que permite, no geral, os melhores resultados e explicam que se relaciona com uma maior robustez ao viés que se verifica com o aumento da percentagem de MD.

Honghai et al. [8] apresentaram, em 2005, uma abordagem para imputação de MD com base no algoritmo SVM para regressão. O algoritmo SVM para a imputação, que está apresentado no Capítulo 2, consiste na criação de um modelo por cada variável que contenha MD e, treinado com os dados completos, preveja os melhores dados para os valores em falta nesse variável. No treino do modelo, a variável com MD será tomada como a saída do modelo e as restantes como entrada. Os autores deste artigo testaram o algoritmo que propuseram no *dataset* SARS, onde trocavam a variável com MD pelo vetor das classes. A imputação com SVM aqui proposta foi comparada com MI, MDI e dois KNN, um com $K = 2$ e outro com $K = 1$. No final, a imputação que obteve maior precisão foi a executada com SVM. Os autores defendem, no entanto, que deve haver dados de treino com observações completas suficientes para o treino do modelo, já que se houver poucos dados é evidente que haverá influência na qualidade da imputação com SVM.

Richman et al. [35] publicaram em 2007 um trabalho em que compararam o método de imputação múltipla, com recurso a um algoritmo SVM e a ANNs para regressão, com os métodos convencionais de regressão e MI. Realizaram as simulações em dois *datasets* de climatologia: um com dados de precipitação e outro com valores de temperatura.

Dado que pretendiam auferir a qualidade da imputação, recorreram aos dois conjuntos de dados já citados mas com inclusão de 5, 10 e 20% de MD e realizaram dez simulações para cada conjunto de dados gerados. Neste estudo foram ainda avaliadas as diferenças na variância das variáveis antes e após a imputação, e ainda realizada a imputação pela Análise de Casos Completos.

Para uma percentagem de 5% de MD os métodos de ML revelaram-se os melhores dado que permitiram obter os valores mais baixos de MSE, sendo que o método da imputação pela média apresentou os piores valores para este parâmetro. Para os outros dois valores de percentagem de MD testados, a tendência foi de um aumento do valor do MSE, com um notório pior resultado para a imputação com ANN. Na imputação pela ACC verifica-se a maior diferença de variância das variáveis entre o *dataset* imputado e os dados originais, sendo que a MI ocupa o segundo pior algoritmo na manutenção da variância original dos dados. Estes são dois algoritmos que pelas suas formas de atuação levam naturalmente à perda de variância: no caso da eliminação de casos isto deve-se à eliminação de dados e por isso de informação. Na *Mean Imputation* (MI), não há informação estatisticamente diferente acrescentada, por isso não se recupera a informação perdida com os MD. A imputação múltipla com SVM apresentou os melhores resultados na recuperação dos dados originais.

Jerez et al. [7] apresentaram em 2010 um trabalho em que testam 6 métodos de imputação distintos para preencher as lacunas num conjunto com dados referentes a pacientes de cancro da mama. Como métodos de imputação, recorreram à MI, *hot-deck*, Imputação Múltipla (MImp), MLP, SOM e KNN. Já como classificador, neste trabalho foi usado um modelo ANN. O objetivo dos autores foi verificar a influência da imputação de MD na qualidade de classificação de cada caso em recaída ao cancro ou não.

Para poderem ter uma base de comparação, os autores deste trabalho decidiram obter um valor de classificação pela ACC. Todos os métodos de imputação superaram a área debaixo da curva (*Area Under the Curve* (AUC)) da Análise de Casos Completos, com a maior diferença a ser obtida com KNN. Um teste ANOVA permitiu concluir que os métodos de imputação baseados em ML foram os que apresentaram significativamente maiores aumentos na qualidade da classificação, sendo que os restantes não tiveram efeito significativo. Para perceber a influência do tamanho do conjunto de treino na qualidade da imputação e classificação, neste trabalho fizeram

uma simulação com eliminação sucessiva de 10% dos dados de treino. Os resultados mostram um decréscimo na qualidade da classificação com a diminuição dos dados de treino, observação que se torna mais evidente a partir dos 60% de dados de treino eliminados. Ainda que a qualidade da classificação diminua, é de notar que a diferença entre a predição com os dados imputados com a ACC aumenta com o aumento dos dados de treino descartados, isto é, torna-se mais evidente a influência da imputação no aumento da qualidade da classificação, sendo uma observação importante na aplicação de imputação a conjuntos de dados pequenos.

Em 2011 Silva-Ramírez et al. [15] propuseram uma abordagem para imputação com MLP diferente da que foi apresentada no Capítulo 2. Neste trabalho os autores propõem uma abordagem com MLP que recorre a todas as variáveis como entrada e, ao mesmo tempo, como saída da ANN. Dado que o número de neurónios de entrada e saída da rede é igual, o processo de treino necessita ser realizado apenas uma vez, em lugar de correr uma vez por cada variável que contenha MD. As experiências são realizadas recorrendo a *datasets* completos e introduzindo-lhes lacunas de forma aleatória com uma probabilidade de 5% em cada variável. Os autores definiram também um limite de metade das variáveis com MD para cada observação. A amostragem foi feita segundo o método *hold-out*, com uma razão de 2/3 dos dados para treino e o restante para teste. Simultaneamente é realizada validação cruzada com 10 partições para encontrar a melhor combinação de parâmetros para o MLP. O treino do MLP para imputação é feito com o *dataset* com MD como entrada e o mesmo *dataset* completo como saída. Com o MLP treinado recorre-se aos dados de teste para calcular o erro entre os dados imputados e o conjunto de dados de teste original.

Dada a estrutura do algoritmo, os dois objetivos principais deste trabalho são perceber, por um lado, a influência dos diversos parâmetros do MLP na sua performance e, por outro lado, perceber a robustez deste algoritmo ao lidar com MD. Para tal recorreram aos métodos MI, regressão e *hot-deck* para comparar com os resultados do método de imputação proposto. O algoritmo Levenberg-Marquardt revelou-se o melhor para o treino dos MLPs, método de imputação que apresentou os melhores resultados em oito dos quinze conjuntos de dados testados, com maior ênfase em *datasets* com variáveis categóricas. O maior contributo deste trabalho é a possibilidade de realizar imputação com dados que contêm MD e não apenas com as observações que estão completas, revelando-se mais útil em situações reais. Foi feita

ainda uma experiência com um classificador, uma variante do KNN, de onde se concluiu que as reduções da capacidade de classificação atingiam um máximo de 5%.

Já em 2015, García-Laencina et al. [4] desenvolveram um trabalho com o objetivo de avaliar o efeito de diversas abordagens aos MD na predição da sobrevivência de pacientes de cancro da mama. Neste trabalho os autores recorreram a um conjunto de dados de paciente de cancro da mama do Instituto Português de Oncologia (IPO) do Porto. Este *dataset* caracteriza-se por ter elevadas percentagens de MD: apenas 3% dos casos não apresentam MD em nenhuma das variáveis e há variáveis que apresentam 80% dos casos sem valor para a mesma, havendo apenas 3 variáveis completas no conjunto das 16 que constituem o *dataset*. Neste estudo, os autores recorrem a 3 métodos de imputação e a três métodos de classificação: imputação pela moda, *Expectation-Maximization* (EM) ou KNN e classificação com KNN, DT, regressão logística ou SVM. Numa primeira abordagem os autores recorrem a métodos de classificação capazes de lidar com MD: KNN e DT. Numa segunda abordagem foram usados os classificadores supra-referidos, mas com o conjunto de dados imputado.

Para a primeira abordagem, onde era realizada a classificação com o *dataset* original (com MD), os resultados não foram bons. Devido à alta percentagem de MD associada a algumas variáveis era atribuída a mesma classe a todos os casos. Para colmatar este problema, os autores decidiram ignorar as variáveis com maiores percentagens de MD, excluindo aquelas onde mais de 40% dos casos estivesse sem valor. No entanto o problema na classificação repetiu-se, levando à conclusão de que era indispensável a imputação de MD para obter resultados melhores. Na abordagem com imputação antes da classificação fica claro que a imputação permite melhorar os resultados da classificação, já que, logo com o método mais simples de imputação, a imputação pela moda, se obtém uma classificação mais válida que na abordagem anterior. O método de imputação KNN permitiu obter os melhores resultados para a classificação. Quanto aos métodos de classificação, a decisão com DT levou aos piores resultados, muito influenciados pelo viés que a classe maioritária, sendo que o método de classificação com KNN permitiu as classificações de maior qualidade.

3.2 Conclusão

Na Tabela 3.1 encontram-se, de uma forma resumida, as principais técnicas de imputação e classificação usadas na literatura revista. Apresentam-se os valores de percentagens de MD e os conjuntos de dados com que foram testados os métodos e ainda as técnicas utilizadas para a amostragem dos dados para treino e teste das diferentes abordagens.

Da literatura revista as principais conclusões que se podem retirar são a influência dos métodos de imputação na obtenção de conjuntos de dados completos que levam, por um lado, a melhores classificações, mas também a uma manutenção da variabilidade dos dados ou mesmo à contribuição para a restituição da variabilidade inicial. Há ainda a observar que os métodos de imputação com base em modelos de ML conduzem, geralmente, às melhores imputações, isto quando comparado com os métodos estatísticos convencionais.

Tabela 3.1: *Datasets*, algoritmos de imputação, e classificação, percentagens de MD e estratégias de amostragem usadas nos estudos analisados

Publicações	***Datasets***	**Algoritmos**	**% de MD**	**Classificação**	**Amostragem**
Troyanskaya et al., 2001 [6]	3 *datasets* de *microarrays* de ADN	KNN, SVD, MI	1-20%	–	–
Batista et al., 2002 [13]	Bupa, Cmc, Pima (UCI)	KNN	10-60%	C4.5, CN2	CV 10-*fold*
Fessant et al., 2002 [18]	Inquérito transportes pessoais em França	MI, *hot-deck*, MLP	–	–	–
Acuña et al., 2004 [9]	12 *datasets* (UCI)	ACC, MI, MDI, KNN	1-20%	LDA, KNN	CV 10-*fold*
Honghai et al., 2005 [8]	SARS	MI, MDI, KNN, SVM	–	–	–
Richman et al., 2007 [35]	2 *datasets*: precipitação e temperetura	MImp, ACC, MI	5, 10 e 20%	–	–
Jerez et al., 2010 [7]	*Dataset* Cancro da Mama	ACC, MI, *hot-deck*, MImp, MLP, SOM e KNN	Originalmente com MD	ANN	CV 10-*fold*
Silva-Ramírez et al., 2011 [15]	14 *datasets* (UCI), 1 Artificial	MI, regression, *hot-deck*, MLP	5%	KNN	CV 10-*fold* *Hold-out* 70/30%
García-Laencina et al., 2015 [4]	*Dataset* Cancro da Mama do IPO Porto	Moda, EM, KNN	97%	KNN, DT, regressão logística, SVM	CV

Capítulo 4

Componente Experimental

O objetivo primordial deste trabalho é, como foi referido no Capítulo 1, implementar o algoritmo OP-ELM como método de imputação e comparar a sua performance com outros métodos de imputação já utilizados para este efeito. Nas duas secções que se seguem serão apresentados os conjuntos de dados e descritas, detalhadamente, as simulações para as duas abordagens definidas. Num primeiro contexto recorre-se a *datasets* completos, aos quais vai ser, artificialmente, inserido MD, para poder utilizar os diversos métodos de imputação. Por último é feita a classificação com os *datasets* imputados. Na segunda abordagem pretende-se comparar a performance dos diversos métodos de imputação em *datasets* que originalmente contêm MD.

4.1 Conjuntos de Dados Completos

4.1.1 *Datasets*

A estrutura de um conjunto de dados seguida ao longo da execução experimental deste trabalho é apresentada na Figura 4.1. Um conjunto de dados é assim caracterizado por uma matriz X que contém N linhas, correspondendo aos diversos casos/observações, com 'd' variáveis. Existe ainda um vetor T que contém a classe a que cada uma das observações pertence. Um valor em falta no conjunto de dados é representado por '?'. Durante as simulações recorre-se a uma matriz binária (M) que identifica os valores em falta. Essa matriz, que é do mesmo tamanho da matriz que contém os dados, é uma matriz cujo valor é igual a 1 para os valores em falta e 0 nos

restantes casos.

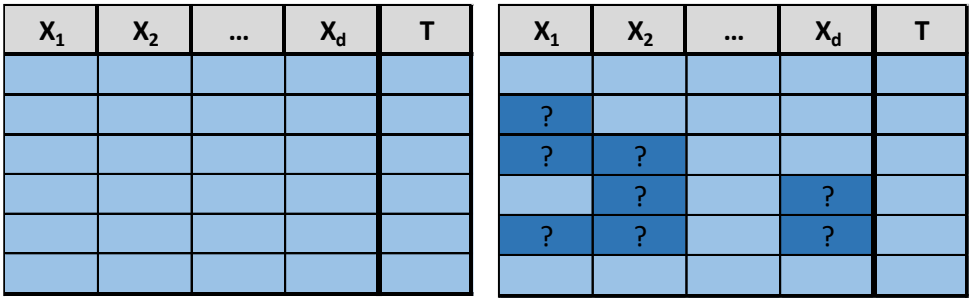

Figura 4.1: Representação esquemática de um *dataset* completo (à esquerda) e de um *dataset* com MD (à direita). Neste esquema '?' representa uma lacuna nos dados. X representa a matriz dos dados, com 'd' variáveis. O vetor T representa o vetor das classes.

Nesta etapa do projeto utilizámos os conjuntos de dados completos apresentados na Tabela 4.1. A escolha dos conjuntos de dados pretende estudar diversos cenários possíveis: *datasets* com um elevado número de observações e outros conjuntos de dado mais pequenos; *datasets* com diferentes dimensionalidades, i.e., com um maior ou menor número de variáveis; situações de classificação binária e também situações multi-classe e os diversos tipos de variáveis que é possível encontrar nos *datasets*, desde variáveis contínuas, inteiras/discretas ou categóricas. Os conjuntos de dados presentes neste estudo foram obtidos no repositório de dados UCI e num repositório de dados da Universidade de Bangor.

4.1.2 Simulações com *datasets* completos

Para que sejam válidas as comparações entre os vários métodos de imputação e classificação há o cuidado, durante a componente experimental deste projeto, de manter uma estrutura comum para os diferentes cenários que se pretendem abordar e para os diversos conjuntos de dados em causa.

Na Figura 4.2 encontra-se esquematizada a estrutura seguida na utilização dos *datasets* completos para validar os vários métodos de imputação e classificação:

1) O primeiro passo que é feito nas simulações é a divisão dos conjuntos de dados originalmente completos em dois conjuntos disjuntos: um para treino e outro para teste. A divisão foi feita recorrendo ao método *hold-out*, reservando para 70%

Tabela 4.1: Conjuntos de dados completos utilizados na avaliação de métodos de imputação e classificação

Dataset	**# Casos**	**# Atributos (I,C,R)**	**# Classes**
Pima	768	8 (1/0/7)	2
Indian Liver	583	10 (1/1/8)	2
Iris	150	4 (0/0/3)	3
Banknote	1372	4 (0/0/4)	2
Seeds	210	7 (0/0/7)	3
Laryngeal	353	16 (1/0/15)	3
Voice	238	10 (0/1/9)	3
Transfusion	748	4 (0/0/4)	2
Telugu	871	3 (3/0/0)	6
Red Wine	1599	11 (0/0/11)	6
Bupa	345	6 (0/0/6)	2

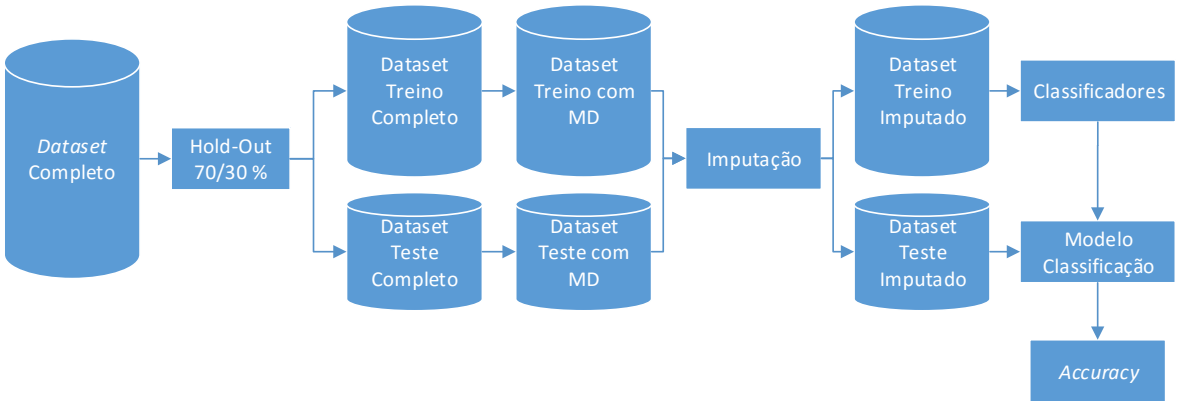

Figura 4.2: Esquema representativo do estudo feito em *datasets* completos.

dos dados para treino e os restantes 30% para teste. É de realçar que a proporção de classes verificada no *dataset* original foi preservada nos dois conjuntos obtidos após a divisão, com o objetivo de não enviesar os resultados. Por sua vez, o *dataset* de treino é dividido em 10 partições com o intuito de realizar validação cruzada estratificada, quer seja na obtenção dos melhores parâmetros/modelos de imputação, quer no processo de classificação. O facto desta validação ser

estratificada implica que a divisão preserve em cada uma das partições uma proporção dos casos de cada classe igual à do conjunto inicial;

2) O segundo passo consiste na normalização dos dados. Nas simulações realizadas recorreu-se à normalização utilizando média zero e desvio padrão um. Optou-se por normalizar os dados de treino e guardar a transformação, como o intuito de normalizar os dados de teste com a mesma transformação dos dados de treino;

3) Neste ponto procede-se à inserção de MD do tipo MCAR nos dados de treino e teste completos de modo a obter os conjuntos de dados que deverão ser imputados. Dado que se pretende verificar a influência dos métodos de imputação nos resultados das classificações para diferentes percentagens de MD, existe a necessidade de abrir lacunas nos *datasets* em percentagens que vão dos 5 aos 70% –5, 10, 20, 30, 50 e 70%– de casos. Assim como na divisão dos dados inicias em dados de treino e teste, a inserção de MD é feita de modo a obter a mesma proporção de observações com MD em cada classe. As variáveis em que os MD são inseridos são as que são mais contributivas para o processo de classificação. A razão pela qual se escolheram as variáveis mais relevantes para a classificação é poder ter um maior efeito nos valores obtidos para *accuracy* aquando do teste de diferentes percentagens de MD. Para escolher as variáveis ditas relevantes optou-se pelo cálculo da IM entre cada uma das variáveis e o vetor das classes (Tabela 4.2). Deste modo são escolhidas as variáveis que apresentam os maiores valores para esta métrica. Para cada conjunto de dados completo, e para poder verificar a influência de MD em mais do que uma variável, inseriram-se, de forma incremental, MD na primeira, segunda e terceira variáveis mais importantes (se a dimensão do *dataset* permitir). No caso do esquema da Figura 4.1, as variáveis que contêm MD são as variáveis 1, 2 e 'd';

4) No processo de imputação foram usados, nas simulações, 5 métodos diferentes: MLP, SVM, KNN, SOM e OP-ELM (proposto neste projeto). Na Figura 4.3 encontra-se esquematizada a lógica de imputação seguida neste projeto.

Os processos de imputação iniciam-se com a divisão do conjunto de treino em dois conjuntos: um com as observações completas e outro com todos os casos em que pelo menos uma variável esteja em falta. No protótipo da Figura 4.1 a divisão

Tabela 4.2: Informação Mútua calculada entre cada variável e o vetor de classes. Os valores para esta métrica encontram-se normalizados à unidade e foram calculados com os *datasets* inicialmente completos

Dataset	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}
Pima	5,77	**28,41**	5,54	7,63	**25,88**	11,74	1,86	**13,16**	–	–	–	–	–	–	–	–
Indian Liver	13,97	0,59	6,80	5,82	**31,49**	**18,16**	**20,74**	0,43	1,51	0,48	–	–	–	–	–	–
Iris	19,53	7,27	**39,00**	**34,20**	–	–	–	–	–	–	–	–	–	–	–	–
Banknote	**52,56**	**32,73**	**13,36**	1,36	–	–	–	–	–	–	–	–	–	–	–	–
Seeds	**26,92**	**24,56**	0,00	13,52	11,08	6,72	**17,20**	–	–	–	–	–	–	–	–	–
Lanryngeal	0,13	3,46	**15,99**	3,86	4,09	7,43	**10,86**	**10,44**	9,77	6,34	7,85	1,21	9,66	2,12	2,14	4,67
Voice	0,64	**57,52**	0,00	0,25	0,00	0,00	**8,21**	**25,35**	0,35	7,69	–	–	–	–	–	–
Transfusion	**39,66**	26,43	**33,91**	–	–	–	–	–	–	–	–	–	–	–	–	–
Telugu	**32,07**	**42,43**	25,50	–	–	–	–	–	–	–	–	–	–	–	–	–
Red Wine	5,85	8,43	1,61	3,86	0,29	**14,91**	**35,87**	0,00	0,65	3,90	**24,64**	–	–	–	–	–
Bupa	5,69	**26,73**	**19,80**	10,78	**28,54**	8,47	–	–	–	–	–	–	–	–	–	–

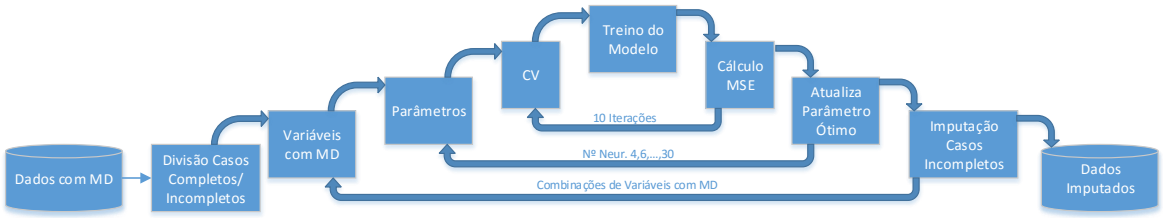

Figura 4.3: Representação esquemática do processo de imputação seguido.

referida incluiria o primeiro e último caso nas observações completas e as restantes seriam consideradas incompletas. Uma vez obtidas as observações completas, iniciar-se-ia a procura dos melhores parâmetros/modelos para a imputação do conjunto de dados específico. Dado que podem existir várias variáveis com MD, é necessário criar um modelo para cada combinação dessas variáveis com MD. No caso do esquema da Figura 4.1, temos 3 variáveis com MD, por isso podem existir até 6 combinações de variáveis com dados em falta. No caso da observação da segunda linha, apenas a primeira variável está em falta, por isso as restantes variáveis serviriam como entrada ao modelo de imputação e a variável X_1 seria incluída como saída do modelo, durante o treino com as observações completas. No caso em que mais de uma variável na mesma observação contenha MD, é necessário criar um modelo para cada uma das variáveis com MD como saída desse modelo.

Para cada combinação de variáveis incompletas, e para cada parâmetro que se

pretende avaliar, é criado um modelo de imputação com nove das partições obtidas na divisão para CV. A última servirá para testar o modelo e obter o valor do MSE entre os dados imputados e os dados originais, que, no caso dos *datasets* completos, estão disponíveis. Este processo é repetido 10 vezes, de forma a cada partição servir uma vez como partição de validação. Uma vez testados todos os parâmetros, escolhe-se o que conduziu ao menor valor médio do MSE para as 10 iterações. Com o melhor parâmetro escolhido, cria-se um modelo de imputação e treina-se com todas as observações completas para imputar os MD, processo que será repetido para todas as combinações de variáveis com MD.

Para a imputação com MLP, o parâmetro que é decidido por CV é o número de neurónios da camada escondida. Para este parâmetro são testados valores ente 4 e 30. Na imputação com SVM, dependendo do *kernel* a que se recorre podem existir parâmetros diferentes a definir. O valor de C tem de ser definido para qualquer *kernel*, mas no caso de se usar um *kernel Radial Basis Function* (RBF) há ainda a definir o valor do parâmetro γ (gama), que define a largura da RBF. Neste projeto são testados valores de C entre 10^{-5} e 100 e valores de γ que variam entre 2^{-10} e 2^{10}. Na imputação com KNN, o parâmetro a decidir é o valor de K vizinhos que são considerados para fazer a imputação. Os valores que são testados para este parâmetro vão desde K=1 até ao valor da raiz quadrada do número de observações do conjunto de treino. Para a imputação com SOM, é necessário definir o melhor tamanho para o mapa bidimensional. Neste projeto assumimos que o mapa é quadrado e são testados valores de 3 até 12 para o tamanho de cada lado do mapa, ou seja, são testados mapas de dimensão $3 * 3$ até $12 * 12$.

Para o conjunto de dados de teste, uma vez que já se dispõe dos modelos que levaram aos melhores resultados de validação, é necessário apenas dar como entrada as observações completas para obter os valores a preencher nas lacunas dos MD, havendo na mesma a necessidade de percorrer cada uma das combinações de variáveis em falta.

5) Para o efeito de classificação recorreu-se, neste projeto, aos métodos KNN, SVM, RF e MLP. A estrutura seguida para a classificação encontra-se representada

na Figura 4.4. Para os diversos classificadores há parâmetros que devem ser ajustados de acordo com o *dataset* que se pretende classificar. No caso do KNN é necessário definir o valor de K. Já para o algoritmo SVM é necessário encontrar o melhor valor de C e gama. Para o classificador RF o parâmetro a definir é o número de árvores de decisão. Por fim, na classificação com MLP é necessário procurar o número de neurónios da camada oculta. A procura dos valores, para estes parâmetros, que contribuem para a melhor classificação é feita por CV com 10 partições. Para cada partição é treinado o classificador que é usado para classificar a partição de validação. É escolhido o parâmetro que conduzir ao maior valor médio de *accuracy* das 10 iterações. Depois de encontrado o melhor parâmetro para o classificador, procede-se ao treino do modelo, mas agora com todo o *dataset* de treino imputado, e classifica-se o conjunto de teste, classificação que é comparada com a atribuição de classes originais.

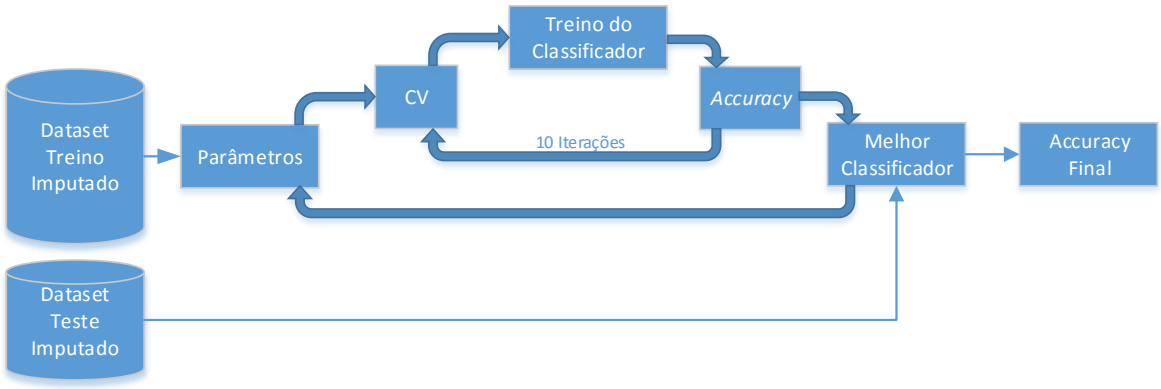

Figura 4.4: Representação esquemática do processo de classificação seguido com os diversos métodos de classificação.

4.2 Conjuntos de Dados com Valores em Falta

Um dos objetivos deste trabalhos é testar os métodos de imputação que apresentam melhores resultados nas simulações com conjuntos de dados completos em *datsets* incompletos. A seguir apresentam-se as características dos *datasets* a que se recorreu para esta etapa do projeto, e fazer-se-á a descrição da metodologia implementada nas simulações.

4.2.1 *Datasets*

Dataset de Cancro da Mama do IPO

Este *dataset* contém dados de 399 pacientes de cancro da mama do IPO do Porto. Cada paciente corresponde a um caso do conjunto de dados e é caracterizado por 16 variáveis, apresentadas na Tabela 4.3: a idade; a localização do tumor; a topografia; envolvimento contralateral da mama; tipo histológico; o grau de diferenciação; tamanho, número de nodos e de metástases; estado do tumor; expressão de recetores hormonais; expressão de HER2; e o tipo de tratamento, desde o tipo de cirurgia, regime de quimioterapia e o tipo de hormonoterapia.

Tabela 4.3: Variáveis que constituem o conjunto de dados de pacientes de cancro da mama do IPO do Porto. A variável idade é a única variável contínua do conjunto de dados, sendo que o número de categorias de cada uma das restantes variáveis se encontra indicado na segunda coluna.

Variável	Categorias	% MD
Idade	–	0
Localização do Tumor	3	15,53
Topografia	9	0
Contralateral	2	0
Tipo histológico	4	9,77
Grau de diferenciação	3	16,79
Tamanho	4	1,5
Nodos	4	7,77
Metástases	2	14,04
Estado	7	22,06
Recetores hormonais	2	9,77
HER2	2	80,45
Cirurgia	5	13,78
Tipo de tratamento	8	8,02
Quimioterapia	6	43,36
Hormonoterapia	3	47,12

No trabalho de García-Laencina [4], onde o mesmo conjunto de dados é usado, os autores deixam explícito que as 16 variáveis que constituem este *dataset foram selecionadas pelo staff médico do IPO com base em linhas orientadores internacionais, experiência profissional, conhecimento, decisões prévias e resultados observados*. O objetivo deste conjunto de dados é potenciar uma base para modelos de previsão da sobrevivência a 5 anos, baseados nas variáveis apresentadas. Dado isto, os casos são divididos nas classes 0 e 1, conforme se preveja que o paciente não sobreviva ou sobreviva, respetivamente, no espaço de 5 anos. Neste conjunto de dados existem 117 casos da classe 0 e 282 casos da classe 1.

O maior problema que este conjunto de dados apresenta é a elevada percentagem de MD. A variável que codifica a expressão de HER2 apresenta valor apenas em 19,55% dos casos deste conjunto de dados. Apenas 3 das 16 variáveis que constituem este *dataset* têm valor para todos os pacientes. Dos 399 casos que constituem este *dataset* apenas 3% apresenta valor para todas as variáveis, um valor que é crítico quando se pretende fazer imputação com base nos casos completos.

PIMA

Este conjunto de dados é constituído por 300 casos de mulheres com antecedentes no povo Pima. Este povo apresenta uma alta taxa de doentes de diabetes, o que levou investigadores a estudar as suas características. Neste conjunto de dados são recolhidos alguns dados, desde o número de gravidezes, a concentração oral de glucose, a pressão diastólica, tamanho das dobra de pele nos trícipes, Índice de Massa Corporal (IMC), valor da função de ascendência para a Diabetes e a idade, todas estas variáveis que estão correlacionadas com o aparecimento da diabetes. Os indivíduos são depois classificados como sujeitos à doença ou não [36].

Como é possível observar na tabela 4.4, neste *dataset* em 3 das 7 variáveis que o constituem há MD. A percentagem máxima de MD observada é de 32,67%. Dos 300 casos que constituem este conjunto de dados 33% apresenta algum valor em falta.

Os dois conjuntos de dados apresentados acima apresentam originalmente MD. No entanto as suas características são distintas: no caso do conjunto de dados de cancro da mama, apenas 3 das variáveis não possuem MD, o que corresponde a 19% das variáveis. Quanto ao tipo de variáveis, o primeiro *dataset* apresentado contém

Tabela 4.4: Variáveis que constituem o conjunto de dados Pima e respetivas percentagens de MD.

Variável	% MD
Nº gravidezes	0
Glucose	0
Pressão diastólica	4,33
Pele trícipes	32,67
IMC	1
Ascendência Diabetes	0
Idade	0

essencialmente variáveis categóricas, enquanto que o conjunto de dados Pima apresenta apenas variáveis contínuas. Já no *dataset* Pima, existem 57% das variáveis sem qualquer valor em falta. Nos dados de pacientes de cancro da mama, a variável com maior percentagem de MD apresenta, aproximadamente, 80% de lacunas, enquanto que o valor máximo para o *dataset* Pima é de apenas 33%. Assim, é possível avaliar a respostas de diferentes métodos de imputação e de classificação em abordagens distintas.

4.2.2 Simulações com *datasets* com MD

Ao contrário do que acontece com os conjuntos de dados completos, nos *datasets* que originalmente contêm MD não existem os valores originais, para que possamos comparar com os obtidos com a imputação. Por esta razão, e devido à reduzida percentagem de casos completos de que se dispõe nestes conjuntos de dados, não há casos suficientes para encontrar, por CV, os parâmetros dos métodos de imputação que levam aos melhores resultados nas classificações. Portanto, nas simulações desta abordagem é necessário imputar os conjuntos de dados com cada um dos parâmetros possíveis para o método de imputação. Nesta segunda abordagem, dado os resultados que serão apresentados no Capítulo 5, foram utilizados os algoritmos de imputação SVM, KNN, SOM e OP-ELM. Os classificadores a que se recorreu foram os mesmos que foram utilizados com os conjuntos de dados completos: KNN, SVM, RF e MLP.

Para cada método de imputação, em lugar do que acontecia na imputação de

datasets completos, não há um processo iterativo de procura dos melhores parâmetros dos métodos de imputação. Nesta abordagem o conjunto de dados é todo imputado para cada um dos parâmetros do método de imputação, sendo que se mantém a divisão em combinações de variáveis incompletas. Assim, na imputação com KNN os parâmetros que são testados são os valores de K vizinhos, valor que será testado de 1 até à raiz quadrada do número de casos de treino. Para cada valor de K testado é treinado um modelo para cada combinação de variáveis com MD. São imputados os dados de teino em falta e depois, com o mesmo modelo, imputam-se os dados de teste. Na imputação com SOM serão testados mapas quadrados de lado 3 até 12. Na imputação com SVM surge um problema nesta abordagem: existem muitos valores de C e γ para testar. O valor de C varia entre 10^{-5} e 100 e os valores de γ variam entre 2^{-10} e 2^{10}, o que levaria a testar 861 combinações destes dois parâmetros. Dado o tempo que tal simulação exigiria, decidiu-se testar valores de C e γ que fossem representativos da gama de valores testados na abordagem com casos completos. Portanto, foram testados os valores 1e-3, 1, 10 e 100 para o parâmetro C e 2^{-10}, 1 e 2^{10} para o parâmetro γ. No caso da imputação com o algoritmo OP-ELM não é necessário definir nenhum parâmetro, já que a procura do número ótimo de neurónios é definida internamente.

Capítulo 5

Resultados

Neste capítulo são apresentados os resultados das simulações realizadas. Depois serão dados a conhecer os resultados da imputação e classificação nos conjuntos de dados completos nas diferentes abordagens seguidas. Por último apresentam-se os resultados obtidos com os *datasets* que originalmente contêm MD.

Na Tabela 5.1 são apresentados os valores da *accuracy* obtidos com os quatro métodos de classificação usados neste projeto para os diversos conjuntos de dados completos. Como é possível verificar, a necessidade de recorrer a diferentes classificadores deve-se ao facto de se poderem obter os melhores resultados de classificação nos diferentes *datasets* para diferentes classificadores. Estes valores servem também de base à comparação com os resultados das classificações com os conjuntos de dados imputados.

5.1 Variáveis Relevantes para a Classificação

No Capítulo 4 ficou explícito que, na abordagem com os conjuntos de dados completos, a inserção de MD foi feita nas variáveis 'mais relevantes' para a classificação. Na Tabela 5.2 são apresentados os valores de *accuracy* obtidos com o conjunto de dados Iris, valores médios das 10 divisões *hold-out* nos conjuntos de treino e teste. No caso do *dataset* Iris, é possível observar na Tabela 4.2 que as variáveis mais relevantes para a classificação são a 3 e a 4, sendo as variáveis 1 e 2 as menos relevantes, de acordo com o seu valor para a IM. As variáveis 3 e 4 do *dataset*

Tabela 5.1: Valores de *accuracy* obtidos com os conjuntos de dados completos. A negrito encontra-se realçado o maior valor de *accuracy* para cada conjunto de dados.

Dataset	**KNN**	**SVM**	**RF**	**MLP**
Pima	74,00	**77,30**	75,65	76,48
Indian Liver	69,77	70,00	69,60	**70,80**
Iris	94,89	96,00	95,56	**96,67**
Banknote	99,85	**100**	99,20	99,27
Seeds	92,06	90,63	90,95	**92,54**
Laryngeal	72,45	**72,83**	70,94	70,75
Voice	**79,72**	79,15	78,31	76,90
Transfusion	79,29	79,20	73,66	**77,95**
Telugu	87,25	**87,90**	87,52	82,14
Red Wine	61,50	63,53	**67,81**	58,06
Bupa	62,52	70,87	**71,36**	67,57

Iris acumulam 73,20% da informação relevante para a classificação deste conjunto de dados, sendo que se espera que a inserção de MD e consequente imputação nestas variáveis tenha um efeito significativo nos resultados da classificação.

Tabela 5.2: Valores de *accuracy* obtidos com o classificador KNN, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Iris.

		Variáveis 1 e 2					**Variáveis 3 e 4**				
Métodos de Imputação		**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**	**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**
% MD	**5**	94,44	93,56	94,00	94,00	**95,11**	93,78	94,22	93,11	**94,44**	93,56
	10	**95,11**	94,00	94,44	94,44	94,67	92,67	91,56	**92,89**	92,44	92,67
	20	95,56	94,44	94,22	95,56	**96,00**	91,56	90,89	91,33	91,11	**92,67**
	30	**94,89**	94,44	94,00	93,11	92,89	**90,44**	87,11	89,78	88,22	89,78
	50	94,00	94,89	**96,44**	94,89	94,00	84,44	81,56	84,00	81,78	**85,56**
	70	94,44	**96,00**	84,89	95,11	92,44	77,33	78,67	67,78	77,56	**79,11**

Pelos resultados apresentados na Tabela 5.2 e na Figura 5.1 é possível estabelecer que a inserção de MD nas variáveis menos relevantes para a classificação no conjunto de dados Iris seguida da sua imputação, não têm um efeito tão destrutivo nos resultado da

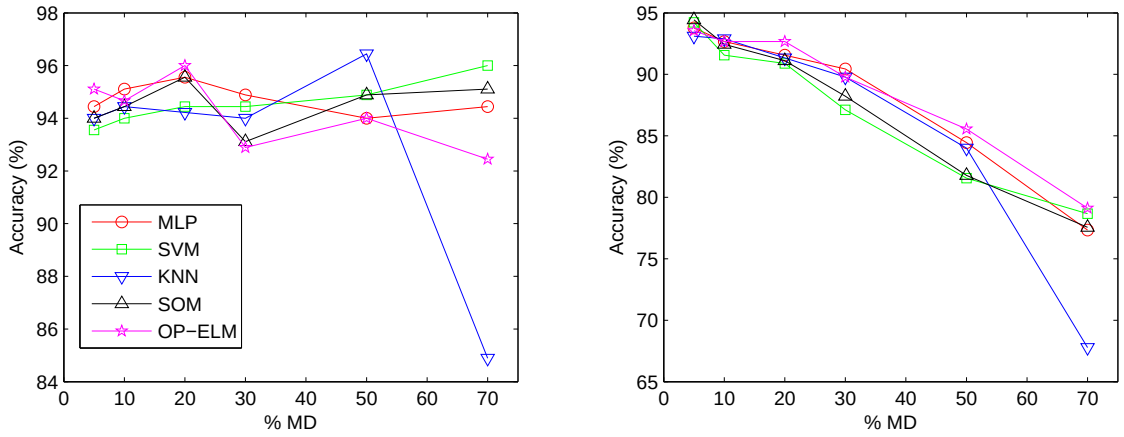

Figura 5.1: Valores de *accuracy* para a classificação com KNN, para o *dataset* Iris, em função da percentagem de MD: na subfigura **(a)**, a inserção de MD foi feita nas variáveis 1 e 2 enquanto que na subfigura **(b)** foram inseridos MD nas variáveis 3 e 4, as mais relevantes para a classificação.

classificação como o que acontece nas variáveis 3 e 4. A inserção de MD nas variáveis 3 e 4 (as variáveis mais relevantes) leva a uma diminuição da qualidade da classificação com o aumento da percentagem de MD. No caso da imputação das variáveis menos relevantes, a inserção de cada vez maiores percentagens de MD não tem um efeito significativo na diminuição dos valores de *accuracy* e não leva a uma ordem evidente na diminuição destes valores da classificação. No caso das variáveis imputadas 3 e 4, há um decréscimo da *accuracy* com o aumento da percentagem de MD. No caso em que a percentagem de MD é 70%, é notável que se atingem valores de 67,78% de *accuracy* com a imputação com KNN dos MD nas variáveis relevantes, sendo que para as variáveis 1 e 2 o valor mais baixo atingido é 84,89%, para o mesmo método de imputação. Olhando para os valores de *accuracy* para os dois conjuntos de variáveis imputadas e para a percentagem de MD de 5%, verifica-se que, devido à reduzida quantidade de dados que foram imputados, não há uma grande diferença entre os valores de *accuracy* para as duas situações e nota-se, ainda, que os valores estão próximos dos obtidos com o conjunto de dados completo (Tabela 5.1). Nas Tabelas A.11, A.12 e A.13 estão presentes os valores de *accuracy* obtidos pelos classificadores SVM, RF e MLP para o *dataset* Iris, onde também são válidas as conclusões apresentadas acima.

5.2 Divisão Treino/Teste

Nas primeiras simulações a imputação era realizada apenas para uma divisão *hold-out* do conjunto de dados em conjuntos de treino e teste, como se encontra representado na Figura 4.2. No entanto percebeu-se que os resultados não estavam de acordo com os obtidos por diversos autores para os mesmos conjuntos de dados que se usaram nestas simulações. Era esperado que a *accuracy* para os classificadores testados diminuísse com o aumento da percentagem de MD. Na Tabela 5.3 apresentam-se os resultados obtidos com apenas uma divisão treino/teste. Ao contrário do que acontece na Tabela 5.2, onde os resultados são uma média dos valores de *accuracy* para dez divisões treino/teste diferentes, com apenas uma simulação observam-se, por vezes, os mesmos valores de *accuracy* para percentagens diferentes de MD. Esta situação não é totalmente desprezável caso se verificasse pontualmente, devido a um conjunto de dados específico. No entanto observa-se que, por exemplo para a imputação com SVM nas variáveis 1 e 2, para os 6 valores de percentagens testados se obtém o mesmo valor de *accuracy*, ainda que estejamos a falar de MD nas variáveis menos relevantes. De acordo com a literatura revista, onde em alguns trabalhos se realizavam 10 simulações com divisões treino/teste diferentes, neste projeto seguiu-se o mesmo raciocínio, permitindo obter resultados mais concordantes com o esperado. Tal pode dever-se à eliminação do efeito que apenas uma divisão pode ter nos resultados.

Tabela 5.3: Valores de *accuracy* obtidos com o classificador SVM, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Iris, resultados de apenas uma simulação.

		Variáveis 1 e 2					**Variáveis 3 e 4**				
Métodos de Imputação		**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**	**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**
% MD	**5**	95,56	95,56	95,56	95,56	95,56	95,56	93,33	95,56	95,56	93,33
	10	95,56	95,56	97,78	95,56	97,78	93,33	95,56	93,33	93,33	93,33
	20	95,56	95,56	97,78	97,78	95,56	93,33	95,56	93,33	93,33	88,89
	30	95,56	95,56	97,78	93,33	95,56	95,56	91,11	95,56	95,56	88,89
	50	93,33	95,56	97,78	95,56	97,78	91,11	82,22	84,44	88,89	88,89
	70	95,56	95,56	97,78	97,78	95,56	86,67	86,67	75,56	77,78	84,44

A escolha dos melhores parâmetros dos classificadores é feita por validação cruzada

(CV). Desta forma, com os dados de treino encontram-se os parâmetros que permitem os melhores resultados de validação e depois classificam-se os dados de teste com o melhor modelo obtido. Para colmatar o efeito que se verifica na Tabela 5.3, onde para diferentes percentagens de MD se observam os mesmos valores de *accuracy*, pensou-se definir uma estrutura fixa para a classificação. Esta estrutura fixa consistia na utilização de classificadores com os parâmetros obtidos para os *datasets* completos. Para as diferentes percentagens de MD e para os diversos métodos de imputação, não era necessária validação cruzada para obter os melhores parâmetros. Seria apenas necessário treinar o classificador com os dados de treino e classificar os dados de teste. Esta abordagem mostrou-se infrutífera, e a situação patente na Tabela 5.3 acontecia da mesma forma. Assim, a abordagem seguida nas simulações apresentadas a seguir foi de dez simulações com dez divisões diferentes em dados de treino e teste, numa proporção 70/30%, e com os parâmetros do classificador a serem definidos por CV para cada situação simulada.

5.3 Dados Completos

A imputação com o algoritmo MLP apresentou, em todos os conjuntos de dados estudados, os maiores tempos de computação, para todas as percentagens de MD independentemente do número de variáveis onde eram eliminados valores. Este resultado não é surpreendente, já que em todos os trabalhos que foram revisto ao longo deste projeto, este algoritmo foi sempre apresentando como sua maior desvantagem o seu tempo de computação. Ainda assim, os resultados das classificações com os dados imputados com MLP não são os que apresentam pior *accuracy*. Nas tabelas que são apresentadas com os resultados das simulações, estão realçados, a negrito, os maiores valores de *accuracy* de entre os vários métodos de imputação para cada valor de percentagem de MD. Na Tabela A.1, a classificação com os dados cujas variáveis imputadas são a 2 e a 5, apresenta os melhores valores com a imputação com MLP para todas as percentagens de MD, exceto para 5%, o que demonstra que, apresar do tempo de imputação, se obtêm bons resultados de classificação comparativamente aos outros métodos.

A imputação com o algoritmo KNN é descrita na literatura como uma abordagem simples e rápida, exceto para conjuntos de dados de maiores dimensões. Nos

resultados obtidos com os diversos conjuntos de dados verifica-se que ao algoritmo KNN se associam, muitas vezes, os menores tempos de imputação. Como é possível observar na Tabela A.34, com o conjunto de dados Voice, a imputação com KNN é a mais rápida para todas as percentagens de MD. No algoritmo KNN é necessário calcular as distâncias entre cada um dos casos, o que implica que para conjuntos de dados maiores este algoritmo apresente maiores tempos de computação. Como é possível verificar nos resultados, aumentado a percentagem de MD obtém-se menores tempos de imputação, uma vez que há sucessivamente menos casos entre os quais se necessita calcular a distância.

O algoritmo OP-ELM apresenta uma grande vantagem comparativamente aos restantes métodos de imputação: não há a necessidade de encontrar um parâmetro por CV. A seleção do valor ótimo para a camada de neurónios escondida é feita pela avaliação do erro da imputação, sem a necessidade da definição de um limiar do erro, já que antes da escolha do número ótimo de neurónios, estes foram ordenados pela sua relevância, para que no final os neurónios que levam aos melhores sejam mantidos e os restantes eliminados. Assim, não há a necessidade de percorrer um conjunto de parâmetros possíveis para encontrar o que conduz aos melhores resultados, como acontece nos restantes métodos de imputação testados. O algoritmo de imputação baseado apresenta as vantagens de ter por base ANN, e de ser, ao contrário do que acontece com o MLP, um algoritmo bastante rápido para as mesmas imputações. Nas simulações que realizámos, não é possível estabelecer que este algoritmo seja o mais rápido, ja que é muitas vezes ultrapassado pelo SVM e KNN em relação aos tempos de computação. Ainda assim, é possível afirmar que o algoritmo OP-ELM é robusto quanto à dimensão dos conjuntos de dados, uma vez que mesmo para conjuntos de dados grandes apresenta baixos tempos de computação, ao contrário do que acontece com o SVM e KNN, apresentado nesta medida uma enorme vantagem.

Na Figura 5.2 estão representados os valores médios do MSE para os 5 métodos de imputação testados, nas diferentes percentagens de MD inseridas e para as 10 simulações. Este erro é calculado entre os valores originais e os valores que são estimados na imputação para preencher as lacunas dos MD. Nas Figuras 5.2a e 5.2b estão representados os resultados do MSE para o *dataset* Transfusion, para as variáveis 1 e 1-3, respetivamente, enquanto que nas Figuras 5.2c e 5.2d estão representados os resultados do MSE para o *dataset* Telugu, para as variáveis 2 e 2-1.

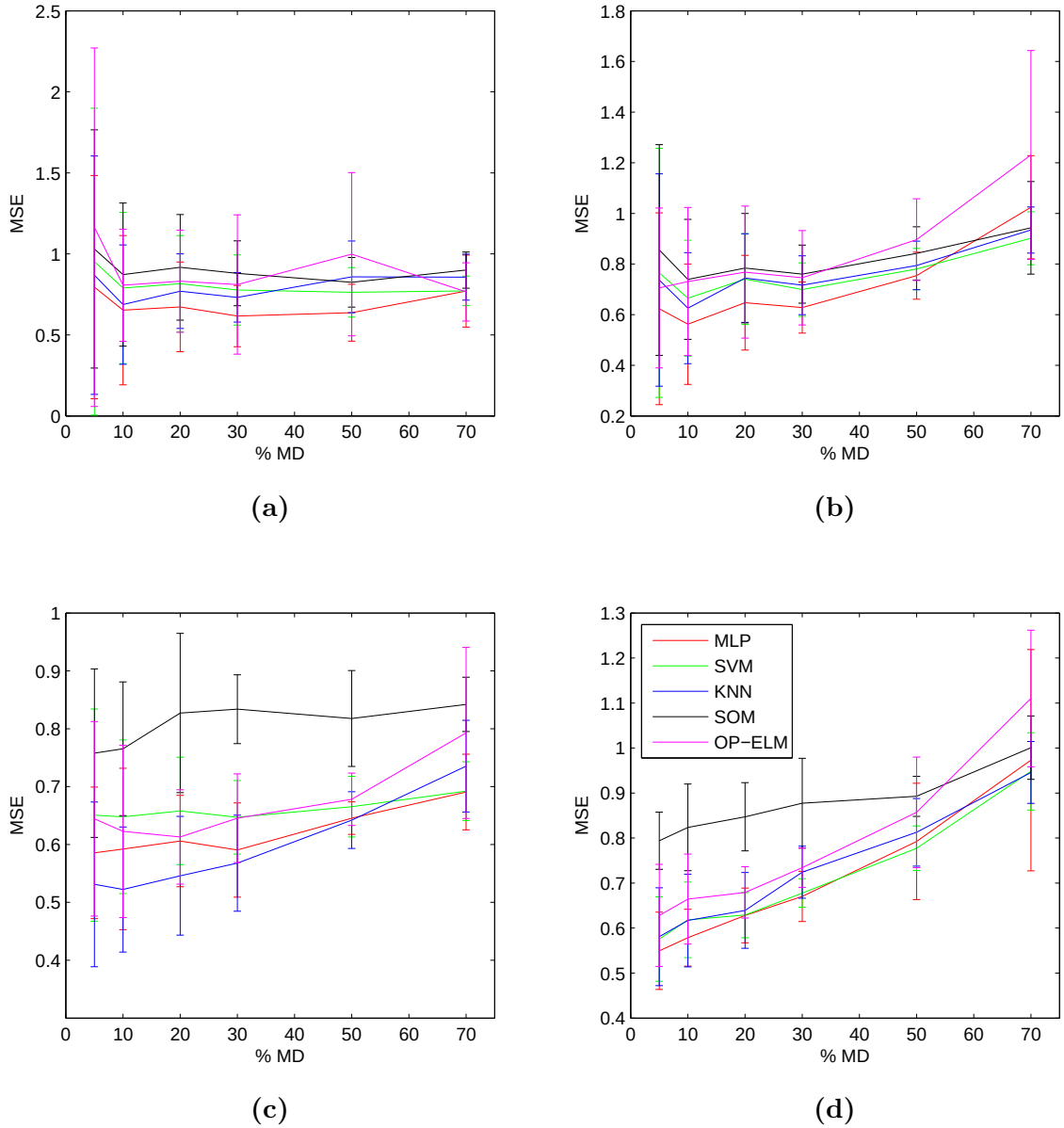

Figura 5.2: Valores médios e desvio padrão do MSE, calculados entre os valores imputados e os valores originais. Nas subfiguras **(a)** e **(b)** encontram-se os resultados com o conjunto de dados Transfusion para as duas combinações de variáveis com MD 1 e 1–3, respetivamente. Já nas subfiguras **(c)** e **(d)** encontram-se os resultados para o MSE com o conjunto de dados Telugu para as duas combinações de variáveis com MD 2 e 2–1, respetivamente.

O aumento do número de variáveis onde é inserido MD leva a um aumento mais rápido do valor do MSE, para os dois *datasets* em causa. Este efeito é expectável, dado que a presença de MD em mais variáveis leva a menos casos completos disponíveis para a imputação. Nas Figuras 5.2a e 5.2b verifica-se que o algoritmo MLP é o que leva a um menor erro de imputação.

Para comparar estatisticamente os valores de *accuracy* obtidos com os vários métodos de imputação e classificação testados, recorreu-se à estatística de Friedman (F_f), apresentada na Secção 2.6. Foram considerados $k = 20$ algoritmos (4 métodos de classificação $*$ 5 métodos de imputação) e $N = 11$ *datasets*, os usados nesta abordagem. Para cada *dataset* atribuí-se um *rank* a cada algoritmo de acordo com a *accuracy* obtida. Este processo foi realizado para cada percentagem de MD testada. Na Tabela 5.4 encontram-se os valores de *accuracy* e respetivo *rank* entre parênteses.

Consultando tabelas da distribuição de F, obtém-se, para 5% de significância, que F(19,190)=1,65. Para as percentagens de 5%, 10%, 20%, 30%, 50% e 70%, e de acordo com os valores apresentados nas Tabelas 5.4, B.1, B.2, B.3, B.4 e B.5, respetivamente, os resultados para F_f são 3,517; 3,474; 3,020; 3,039; 2,607 e 3,852. Assim, dado que todos os valores obtidos para F_f são superiores ao valor tabelado da distribuição de F para esta situação, rejeita-se a hipótese nula, o que implica que os resultados para os diferentes algoritmos são estatisticamente diferentes. Assim sendo, prossegue-se com o teste de Nemenyi, para comparar a performance dos diversos algoritmos. O valor crítico (q_α) para $k = 20$ e 5% de significância é 3,544, levando ao valor de $CD = 8,940$. Dados isto, consideram-se estatisticamente diferentes os algoritmos cuja diferença de *ranks* for superior a 8,940. Dos *ranks* apresentados na última coluna da Tabela 5.4 é possível retirar as seguintes conclusões:

1. A combinação do método de imputação MLP com o classificador SVM foi a que permitiu obter o melhor *rank*;

2. Os melhores *ranks* são obtidos com o classificador SVM, para qualquer um métodos de imputação testados;

3. A imputação com OP-ELM permite obter, para os diversos classificadores, *ranks* muito próximos dos obtidos, para os mesmos classificadores mas com métodos de imputação diferentes, não havendo uma diferença estatisticamente relevante,

Tabela 5.4: Comparação das diversas combinações de métodos de imputação e classificação. Entre parênteses apresentam-se os valores que irão ser usados no teste de Friedman. Os valores de *accuracy* são os obtidos quando foram inseridos 5% de MD

Imp.	**Class.**	***Datasets***											**Rank**
		Pima	**Indian**	**Iris**	**Banknote**	**Seeds**	**Laryngeal**	**Voice**	**Transfusion**	**Telugu**	**Red**	**Bupa**	**Médio**
MLP	KNN	73,87 (20)	70,40 (11)	93,78 (18)	99,13 (7,5)	92,06 (3)	73,58 (2)	79,01 (6,5)	77,90 (6,5)	85,23 (10)	61,52 (12)	62,50 (17)	10,32
	SVM	77,00 (4)	72,27 (1)	94,89 (11)	99,37 (1)	91,16 (11)	73,90 (1)	80,00 (1)	78,17 (2)	86,22 (1)	64,72 (6)	72,09 (1)	3,64
	RF	75,30 (12)	68,25 (20)	96,00 (1,5)	98,83 (9)	91,84 (6)	71,70 (9,5)	78,59 (10,5)	73,88 (17)	85,34 (8)	67,01 (2)	69,66 (8)	9,41
SVM	MLP	76,87 (5,5)	70,83 (9)	95,11 (8)	98,79 (10)	90,93 (14)	70,13 (16)	75,92 (19)	77,41 (14)	81,03 (17)	59,92 (16)	64,56 (15)	13,05
	KNN	74,17 (17)	69,25 (18,5)	94,22 (16)	99,30 (3,5)	90,95 (13)	72,17 (7)	77,89 (15)	78,04 (3)	86,15 (2)	60,56 (15)	61,36 (20)	11,82
	SVM	77,39 (1,5)	71,84 (3)	95,11 (8)	99,32 (2)	91,90 (4,5)	71,70 (9,5)	79,58 (2)	77,90 (6,5)	86,03 (3)	63,70 (7)	71,26 (4)	4,64
	RF	75,48 (10)	70,55 (10)	94,67 (13,5)	98,67 (12)	89,84 (19,5)	70,09 (17)	78,59 (10,5)	73,48 (19,5)	85,61 (5)	67,08 (1)	70,58 (5,5)	11,23
KNN	MLP	74,96 (15)	69,25 (18,5)	94,00 (17)	98,37 (16)	91,90 (4,5)	69,34 (19)	78,59 (10,5)	77,72 (8)	79,66 (20)	58,48 (19)	66,50 (11)	14,41
	KNN	74,04 (18,5)	69,40 (17)	93,11 (20)	99,15 (5,5)	91,11 (12)	71,60 (11)	77,75 (16)	78,53 (1)	85,50 (6)	61,38 (13)	62,33 (18)	12,55
	SVM	77,39 (1,5)	72,13 (2)	96,00 (1,5)	99,30 (3,5)	90,16 (17,5)	72,92 (3)	79,15 (5)	77,68 (9,5)	84,92 (12)	62,59 (10)	71,55 (2,5)	6,18
	RF	76,04 (7)	69,68 (15,5)	94,89 (11)	98,57 (14)	89,84 (19,5)	71,13 (12)	78,31 (13,5)	73,48 (19,5)	85,31 (9)	66,43 (4)	70,00 (7)	12,00
SOM	MLP	75,61 (9)	71,41 (4)	95,56 (5)	98,35 (17)	92,22 (2)	68,87 (20)	76,62 (17,5)	77,99 (4,5)	80,19 (19)	58,43 (20)	66,31 (12)	11,82
	KNN	74,04 (18,5)	69,97 (14)	94,44 (15)	98,59 (13)	91,43 (9,5)	72,36 (6)	78,87 (8)	77,59 (11,5)	84,35 (15)	61,04 (14)	63,79 (16)	12,77
	SVM	77,35 (3)	71,12 (6)	94,89 (11)	98,71 (11)	90,79 (15)	72,55 (4,5)	79,01 (6,5)	77,68 (9,5)	84,77 (13)	63,07 (9)	71,55 (2,5)	8,27
	RF	75,17 (13)	70,26 (12,5)	95,56 (5)	98,16 (18,5)	90,63 (16)	70,75 (13)	78,59 (10,5)	74,29 (16)	84,47 (14)	66,62 (3)	69,22 (9)	11,86
OP-ELM	MLP	75,09 (14)	70,98 (7,5)	95,78 (3)	97,77 (20)	91,75 (7,5)	70,38 (15)	76,62 (17,5)	77,54 (13)	80,34 (18)	58,60 (18)	64,85 (14)	13,41
	KNN	74,26 (16)	69,68 (15,5)	93,56 (19)	99,13 (7,5)	91,43 (9,5)	72,08 (8)	78,31 (13,5)	77,59 (11,5)	85,92 (4)	61,54 (11)	62,04 (19)	12,23
	SVM	76,87 (5,5)	71,26 (5)	94,67 (13,5)	99,15 (5,5)	90,16 (17,5)	72,55 (4,5)	79,30 (4)	77,99 (4,5)	85,38 (7)	63,34 (8)	70,58 (5,5)	7,32
	RF	75,39 (11)	70,26 (12,5)	95,56 (5)	98,45 (15)	91,75 (7,5)	70,57 (14)	79,44 (3)	73,62 (18)	85,19 (11)	66,24 (5)	68,54 (10)	10,18
	MLP	75,96 (8)	70,98 (7,5)	95,11 (8)	98,16 (18,5)	92,38 (1)	69,43 (18)	75,77 (20)	77,32 (15)	81,34 (16)	59,54 (17)	65,15 (13)	12,91

segundo o teste de Nemenyi, já que a diferença de *ranks* é sempre menor que o valor da CD.

Estabelece-se, assim, que o algoritmo OP-ELM, ainda que não conduza aos melhores valores de *accuracy*, não é estatisticamente diferente dos restantes métodos de imputação em análise, apresentando desde logo uma grande vantagem à imputação com MLP, já que apresenta um tempo de computação inferior.

5.4 Dados Incompletos

Dados os resultados obtidos com os conjuntos de dados completos, foi excluída, nesta segunda abordagem, a imputação com o algoritmo MLP. Esta exclusão deveu-se ao tempo de computação deste algoritmo que, atendendo ao tempo deste trabalho, tornou o seu uso impraticável.

A seguir apresentam-se os resultados obtidos para a imputação do *dataset* de pacientes de cancro da mama do IPO do Porto e do *dataset* Pima.

Numa análise prévia do conjunto de dados disponibilizado pelo IPO do Porto, verificou-se que 3 das 16 variáveis tinham o mesmo valor em todos os casos completos. Estas variáveis, que fornecem informação sobre a expressão de recetores hormonais, sobre o número de metástases e sobre o tipo de cirurgia. Dado que assumem o mesmo valor em todos os casos, não irão ser relevantes na classificação, isto porque após a imputação dos valores em falta para estas variáveis, os valores imputados iriam ser iguais ao que a variável tem em todas as observações completas.

Os tempos de imputação para este conjunto de dados corroboram os resultados obtidos com os conjuntos de dados completo: os tempos de imputação com SVM e KNN encontram-se próximos, o tempo de imputação do algoritmo SOM apresenta um valor muito elevado e o algoritmo OP-ELM o valor mais baixo.

Para a imputação com SVM foram testados os 12 pares de valores de C e γ apresentados na Tabela 5.6. Para o conjunto de dados de pacientes de cancro da mama do IPO do Porto, para os diferentes classificadores há diferentes valores de K a conduzirem aos melhores valores de *accuracy*, na imputação com KNN. Neste caso, o algoritmo de classificação SVM é o que permite atingir a maior percentagem de *accuracy* (82,50%), para um valor de K=9. Já para a imputação com SOM, o maior

Tabela 5.5: Valores de *acuracy* obtidos com os diferentes classificadores e métodos de imputação testados com o *dataset* de pacientes de cancro da mama do IPO do Porto. Entre parênteses, ao lado de cada método de imputação, encontram-se os valores do tempo de simulação para os métodos de imputação.

		Imputação			
		SVM (23,00)	**KNN** (34,16)	**SOM** (47435,09)	**OP-ELM (0.17)**
Classificação	**KNN**	77,50	**79,17**	75,83	72,50
	SVM	80,00	**82,50**	79,17	75,83
	RF	**81,67**	80,83	80,83	65,83
	MLP	76,67	**78,33**	**78,33**	75,83

valor de *accuracy* é obtido com o classificador RF, com um mapa de tamanho 3*3 nodos. No caso da imputação com SVM, os melhores resultados para as diferentes classificações são obtidos para os maiores valores de γ, sendo relativamente constantes quanto ao valor de C. Com os dados imputados com SVM o melhor valor de *accuracy* obtido foi 81,67% para C=10 e $\gamma = 2^{10}$.

Tabela 5.6: Parâmetros testados com o algoritmo SVM para imputação.

		C			
		1e-3	**1**	**10**	**100**
γ	2^{-10}	1	4	7	10
	1	2	5	8	11
	2^{10}	3	6	9	12

Este conjunto de dados apresenta apenas uma variável não categórica. Ainda que tenham sido estudados conjuntos de dados com variáveis categóricas, como é exemplo o *dataset* Telugu, que é constituído apenas por este tipo de variáveis, não houve nenhum *dataset* com variáveis categóricas com a dimensão do que é estudado nestas simulações. Nos resultados com o conjunto de dados Telugu a imputação com OP-ELM conduziu a valores de *accuracy* semelhantes aos que se obtiveram com os restantes métodos de imputação, havendo mesmo situações em que a imputação com OP-ELM levou aos melhores valores de *accuracy*.

Na Tabela 5.7 são apresentados os resultados da imputação do conjunto de dados Pima. Ao contrário do que aconteceu com o conjunto de dados de pacientes de cancro

da mama, com este *dataset* o algoritmo OP-ELM não foi o que imputou os dados em menor tempo. O algoritmo SVM foi o que demorou menos tempo na imputação, seguido do algoritmo KNN. Já o algoritmo SOM é, novamente, o que demora mais tempo na imputação. Estas diferenças no tempo demorado, relativamente ao algoritmo anterior, podem ser explicadas pelo tamanho e características dos *datasets*. O *dataset* Pima apresenta apenas 7 variáveis, enquanto que o *dataset* de cancro da mama apresenta 13, durante as simulações feitas. Como aconteceu com os resultados apresentados na Tabela A.45, o algoritmo KNN é sensível tanto ao número de casos como à dimensionalidade do *dataset*, apresentado tempos de imputação maiores para um maior número de variáveis, quando em *datasets* de baixa dimensionalidade leva aos melhores tempos de imputação. Com o conjunto de dados imputado com OP-ELM, não se conseguiu, de novo, que fosse este algoritmo que conduzisse aos melhores valores de *accuracy*, ainda que neste conjunto de dados apresente valores mais próximos dos que se obtêm com os restantes *datasets* imputados. O facto de existir uma maior percentagens de casos completos neste conjunto de dados pode ter influenciado estes valores de *accuracy* a não serem tão diferentes dos obtidos com os restantes conjuntos de dados.

Tabela 5.7: Valores de *acuracy* obtidos com os diferentes classificadores e métodos de imputação testados com o *dataset* Pima. Entre parênteses, ao lado de cada método de imputação, encontram-se os valores do tempo de simulação para os métodos de imputação.

		Imputação			
		SVM (0,20)	**KNN** (0,38)	**SOM** (13,09)	**OP-ELM** (2,26)
Classificação	**KNN**	72,56	73,44	**73,67**	73,33
	SVM	72,67	73,89	**73,67**	71,00
	RF	**74,11**	72,67	73,56	72,67
	MLP	72,89	**73,33**	72,56	73,11

Reunindo os resultados dos dois *datasets*, compararam-se estatisticamente os valores de *accuracy* obtidos com os vários métodos de imputação e classificação testados, recorrendo à estatística de Friedman (F_f). Foram considerados $k = 4$ algoritmos (4 métodos de imputação) e $N = 8$ *datasets* (2 *datasets* * 4 classificadores). Para cada *dataset* atribuí-se um *rank* a cada algoritmo de acordo com a *accuracy* obtida. Na Tabela 5.8 encontram-se os valores de *accuracy* e

respetivo *rank* entre parênteses, com a última coluna a apresentar os valores médios para cada algoritmo de imputação testado.

Tabela 5.8: Comparação dos quatro métodos de imputação testados com os algoritmos incompletos. Entre parênteses apresentam-se os valores que irão ser usados no teste de Friedman.

Imputação	Cancro da Mama				Pima				Rank
	KNN	SVM	RF	MLP	KNN	SVM	RF	MLP	Médio
SVM	77,50 (2)	80,00 (2)	81,67 (1)	76,67 (1,5)	72,56 (4)	72,67 (3)	74,11 (1)	72,89 (3)	2,1875
KNN	79,17 (1)	82,50 (1)	80,83 (2,5)	78,33 (1,5)	73,44 (2)	73,89 (1)	72,67 (3,5)	73,33 (1)	1,6875
SOM	75,83 (3)	79,17 (3)	80,83 (2,5)	78,33 (3)	73,67 (1)	73,67 (2)	73,56 (2)	72,56 (4)	2,5625
OP-ELM	72,50 (4)	75,83 (4)	65,83 (4)	75,83 (4)	73,33 (3)	71,00 (4)	72,67 (3,5)	73,11 (2)	3,5625

Consultando tabelas da distribuição de F, obtém-se, para 5% de significância, que F(3,21)=3,07. Para os resultados da Tabela 5.8, o valor de F_f é 4,26. Dado que o valor obtidos para F_f é superior ao valor tabelado da distribuição de F para esta situação, rejeita-se a hipótese nula, o que implica que os resultados para os diferentes algoritmos são estatisticamente diferentes. Assim sendo, prossegue-se com o teste de Nemenyi, para comparar a performance dos diversos algoritmos. O valor crítico (q_α) para $k = 3$ e 5% de significância é 2,569, levando ao valor de $CD = 1,66$. Dado este valor, verifica-se que, para o teste de Nemenyi, o algoritmo que se destaca dos restantes pela sua performance é o algoritmo KNN, sendo que os algoritmos SVM, SOM e OP-ELM apresentam performances que não são estatisticamente diferentes.

O algoritmo OP-ELM é um método a aplicar quando se pretende imputar conjuntos de dados de grandes dimensões, já que apresenta o valor mais baixo para o tempo computacional, comparativamente aos restantes métodos. Ainda que não apresente valores de *accuracy* superiores aos observados nos restantes métodos de imputação, apresenta valores que não são estatisticamente diferentes.

Capítulo 6

Conclusões e Trabalho Futuro

Os MD são um problema muito comum no tratamento de dados. Lidar com esta problemática requer especial atenção quando se tratam de *datasets* médicos dos quais se pretende fazer inferências dos melhores tratamentos a aplicar. Por isso, o primeiro esforço a lidar com dados em falta deve ser perceber a sua origem e tentar corrigir esse problema.

Neste trabalho foram implementados 5 métodos de imputação: MLP, SVM, KNN, SOM e OP-ELM paralelamente à utilização dos algoritmos KNN, SVM, RF e MLP como métodos de classificação. As simulações foram realizadas em 11 *datasets* completos e 2 *datasets* que, originalmente, apresentam MD. Os algoritmos foram também testados com variáveis discretas.

Até ao momento da realização deste relatório, e segundo a literatura revista, não foi encontrada nenhuma referência onde tivesse sido alguma vez implementado o algoritmo OP-ELM como método de imputação. Neste projeto ficou assente que este algoritmo consegue imputar os conjuntos de dados independentemente do tipo de variáveis e da percentagem de MD que contenham, com bons resultados mesmo para *datasets* com maior número de casos e maior dimensionalidade. Os resultados das classificações com os conjuntos imputados mostram que a imputação com OP-ELM leva a resultados estatisticamente similares aos obtidos com os restantes métodos de imputação. Há um desfasamento entre os valores de *accuracy* obtidos com a imputação com OP-ELM e os que foram obtidos com os restantes métodos de imputação, principalmente com os *datasets* que contêm originalmente MD. Ainda assim, a imputação com OP-ELM

mostrou permitir melhores tempos de imputação do que outros métodos de imputação, com valores de *accuracy* próximos, sem a necessidade de definir qualquer parâmetro. Esta conclusão é mais evidente quando comparamos o tempo de imputação do MLP com OP-ELM: sendo ambos métodos com base em ANN, o segundo apresenta tempos de imputação muito mais baixos, permitindo em certas situações melhores valores de classificação.

Dada a limitação de tempo deste trabalho não foi possível comparar estes algoritmos num maior número de *datasets* incompletos, trabalho que deve ser desenvolvido no futuro. Para além de testar com mais *datasets*, é importante salientar que seria interessante, em termos de imputação, estudar a natureza de MD que estão presentes nos *datasets* e tentar inferir qual é o algoritmo que lida melhor com cada realidade.

Bibliografia

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer Statistics, 2015," *CA Cancer J Clin*, vol. 65, no. 1, pp. 5–29, 2015.

[2] P. H. Abreu, H. Amaro, D. C. Silva, P. Machado, M. H. Abreu, N. Afonso, and A. Dourado, "Overall survival prediction for women breast cancer using ensemble methods and incomplete clinical data," in *XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013*, pp. 1366–1369, Springer, 2014.

[3] P. H. Abreu, H. Amaro, D. C. Silva, P. Machado, and M. H. Abreu, "Personalizing breast cancer patients with heterogeneous data," in *The International Conference on Health Informatics*, pp. 39–42, Springer, 2014.

[4] P. J. García-Laencina, P. H. Abreu, M. H. Abreu, and N. Afonoso, "Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values," *Computers in Biology and Medicine 59*, pp. 125–133, 2015.

[5] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. Wiley Series in Probability and Mathematical Statistics, Wiley, second ed., 1987.

[6] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays.," *Bioinformatics (Oxford, England)*, vol. 17, no. 6, pp. 520–525, 2001.

[7] J. M. Jerez, I. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín, and L. Franco, "Missing data imputation using statistical and machine learning

methods in a real breast cancer problem," *Artificial Intelligence in Medicine*, vol. 50, no. 2, pp. 105–115, 2010.

[8] F. Honghai, C. Guoshun, and Y. Cheng, "A SVM Regression Based Approach to Filling in Missing Values," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (R. Khosla, R. Howlett, and L. Jain, eds.), vol. 3683 of *Lecture Notes in Computer Science*, pp. 581–587, Springer Berlin Heidelberg, 2005.

[9] E. Acuña and C. Rodriguez, "The Treatment of Missing Values and its Effect on Classifier Accuracy," in *Classification, Clustering, and Data Mining Applications* (D. Banks, F. McMorris, P. Arabie, and W. Gaul, eds.), Studies in Classification, Data Analysis, and Knowledge Organisation, pp. 639–647, Springer Berlin Heidelberg, 2004.

[10] P. D. Allison, *Missing Data*. SAGE Publications, Inc, first ed., 2001.

[11] D. B. Rubin, "Inference and Missing Data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[12] A. L. Cool, "A Review of Methods for Dealing with Missing Data," in *Annual meeting of the Southwest Educational Research Association, Dallas*, no. 1, 2000.

[13] G. Batista and M. C. Monard, "A Study of K-Nearest Neighbour as an Imputation Method," *In Second International Conference on Hybrid Intelligent Systems, Soft Computing Systems: Design, Management and Applications*, pp. 251–260, 2002.

[14] M. C. Bishop, *Neural Networks for Pattern Recognition.* Oxford: Clarendon Press, first ed., 1995.

[15] E.-L. Silva-Ramírez, R. Pino-Mejías, M. López-Coello, and M.-D. Cubiles-de-la Vega, "Missing value imputation on missing completely at random data using multilayer perceptrons.," *Neural networks : the official journal of the International Neural Network Society*, vol. 24, no. 1, pp. 121–129, 2011.

[16] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[17] T. Samad and S. Harp, "Self–organization with partial data," *Network: Computation in Neural Systems*, vol. 3, no. 2, pp. 205–212, 1992.

[18] F. Fessant and S. Midenet, "Self-organising map for data imputation and correction in surveys," *Neural Computing and Applications*, vol. 10, no. 4, pp. 300–310, 2002.

[19] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, Inc., second ed., 2000.

[20] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006.

[21] S. K. Sharma and P. Chandra, "Constructive neural networks: a review," *International journal of engineering science and technology*, vol. 2, no. 12, pp. 7847–7855, 2010.

[22] P. J. García-Laencina, A. Bueno-Crespo, and J.-L. Sancho-Gómez, "Design and Training of Neural Architectures Using Extreme Learning Machine," *Neurocomputing: Learning, Architectures and Modeling*, pp. 119–145, 2012.

[23] G. B. Huang, L. Chen, and C. K. Siew, "Universal approximation using incremental constructive feedforward networks with random hidden nodes," *IEEE Transactions on Neural Networks*, vol. 17, no. 4, pp. 879–892, 2006.

[24] R. Reed, "Pruning algorithms-a survey.," *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, vol. 4, no. 5, pp. 740–747, 1993.

[25] H.-J. Rong, Y.-S. Ong, A.-H. Tan, and Z. Zhu, "A fast pruned-extreme learning machine for classification problem," *Neurocomputing*, vol. 72, no. 1-3, pp. 359–366, 2008.

[26] Y. Miche, A. Sorjamaa, P. Bas, O. Simula, C. Jutten, and A. Lendasse, "OP-ELM: Optimally pruned extreme learning machine," *IEEE Transactions on Neural Networks*, vol. 21, no. 1, pp. 158–162, 2010.

[27] T. Similä and J. Tikka, "Multiresponse sparse regression with application to multidimensional scaling," in *Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005*, pp. 97–102, Springer, 2005.

[28] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[29] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[30] L. Al Shalabi and Z. Shaaban, "Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix," *IEEE proceedings of the international conference on dependability of computer systems (Depcos - Relcomex 2006)*, pp. 207–214, 2006.

[31] N. Kwak and C. H. Choi, "Input feature selection by mutual information based on Parzen window," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1667–1671, 2002.

[32] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[33] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[34] M. Lichman, "UCI Machine Learning Repository," [Acessed on: 30 Aug 2015]. [Online]. Available at: `http://archive.ics.uci.edu/ml`.

[35] M. B. Richman, T. B. Trafalis, and I. Adrianto, "Multiple imputation through machine learning algorithms," in *Fifth conference on artificial intelligence applications to environmental science*, 2007.

[36] J. W. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes, "Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus," *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pp. 261–265, 1988.

Apêndice A

Resultados com Conjuntos de Dados Completos

Pima

Tabela A.1: Valores de *accuracy* obtidos com o classificador KNN, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Pima.

		Variável 2					Variáveis 2 e 5					Variáveis 2, 5 e 8				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	73,87	73,57	74,00	73,65	**74,22**	73,35	73,43	74,22	**74,61**	74,04	73,87	74,17	74,04	74,04	**74,26**
	10	**73,83**	73,43	73,65	73,48	73,78	**74,39**	73,91	73,61	74,26	73,83	74,00	73,52	**73,96**	73,43	72,91
	20	**73,48**	72,39	73,09	73,13	73,39	**73,43**	73,04	73,00	72,74	72,09	72,78	**73,04**	72,78	72,65	72,83
	30	72,00	72,30	**72,70**	72,09	71,96	**72,48**	72,00	71,57	71,96	70,96	71,09	72,00	72,43	**72,78**	72,65
	50	70,00	70,17	70,70	69,83	**71,13**	**69,96**	69,87	69,52	69,48	69,52	**69,91**	69,43	69,30	69,48	68,96
	70	**71,00**	69,52	70,39	69,57	68,78	**68,52**	67,57	67,00	68,26	67,22	67,78	66,35	66,48	**68,35**	66,87

Tabela A.2: Valores de *accuracy* obtidos com o classificador SVM, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Pima.

		Variável 2					Variáveis 2 e 5					Variáveis 2, 5 e 8				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	76,91	77,13	77,22	77,35	**77,70**	76,74	**77,48**	77,00	77,26	77,13	77,00	**77,39**	**77,39**	77,35	76,87
	10	76,13	76,70	77,04	**77,39**	76,48	76,74	76,70	76,87	**77,04**	76,78	**76,91**	76,43	76,70	76,87	76,74
	20	**77,09**	76,13	75,61	76,00	76,17	**76,57**	75,74	75,26	75,91	75,57	**76,52**	76,09	75,61	76,04	75,09
	30	**75,83**	75,09	75,00	74,39	75,04	75,04	74,43	74,57	**75,43**	74,91	75,13	74,74	74,74	**75,22**	75,17
	50	72,57	72,83	72,91	72,48	**73,22**	72,43	**73,30**	72,91	73,00	72,43	72,39	72,39	**72,87**	71,70	71,83
	70	71,22	**71,35**	70,61	70,96	70,78	69,48	70,83	69,35	**71,09**	69,83	68,30	**71,04**	69,43	70,39	69,30

Tabela A.3: Valores de *accuracy* obtidos com o classificador RF, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Pima.

		Variável 2					Variáveis 2 e 5					Variáveis 2, 5 e 8				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	**75,91**	74,70	75,57	75,39	75,48	75,39	**76,04**	75,91	75,70	76,00	75,30	75,48	**76,04**	75,17	75,39
	10	75,43	74,26	75,26	75,04	**76,00**	75,04	**75,26**	75,04	74,52	74,78	74,43	74,39	**75,91**	75,48	74,22
	20	**74,48**	73,91	74,17	74,13	73,65	**75,00**	74,39	73,91	73,83	73,83	74,22	74,61	**75,04**	73,83	74,52
	30	**73,22**	73,09	73,04	72,57	72,61	73,70	**74,09**	73,61	72,87	73,83	72,43	72,78	71,96	72,39	**72,91**
	50	72,30	72,52	**72,61**	72,30	72,57	70,96	72,65	72,09	**73,13**	71,70	70,96	70,61	**71,17**	**71,17**	70,70
	70	71,26	70,65	70,39	**72,04**	70,39	69,57	69,83	69,48	**70,09**	69,13	68,04	68,91	67,35	**69,87**	68,35

Tabela A.4: Valores de *accuracy* obtidos com o classificador MLP, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Pima.

		Variável 2					Variáveis 2 e 5					Variáveis 2, 5 e 8				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	**76,65**	75,61	76,26	75,52	76,26	**75,87**	74,35	75,57	75,83	75,65	**76,87**	74,96	75,61	75,09	75,96
	10	74,65	74,70	75,04	75,96	**76,30**	75,13	74,91	**76,57**	74,48	74,35	75,30	75,91	76,57	**76,61**	75,17
	20	**75,57**	73,61	73,48	75,04	75,17	75,17	75,30	74,22	**75,57**	74,43	**76,65**	74,26	75,43	72,04	75,04
	30	73,65	73,48	**75,13**	74,43	73,65	73,87	74,09	73,17	**74,35**	72,83	73,22	**74,70**	73,96	**74,70**	74,13
	50	**73,91**	72,61	72,61	72,35	72,74	72,83	71,09	71,13	72,61	**73,00**	70,78	**71,78**	70,87	71,52	70,22
	70	71,17	70,87	70,26	**71,91**	70,17	69,17	**71,57**	69,26	70,39	69,91	67,96	**70,22**	68,00	**70,22**	69,74

Tabela A.5: Tempo, em segundos, que os diferentes métodos de imputação demoraram, para todas as percentagens de MD testadas com o *dataset* Pima.

		Variável 2					Variáveis 2 e 5					Variáveis 2, 5 e 8				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	3513,294	29,798	**14,414**	17,039	26,748	10085,368	109,496	**49,137**	62,882	94,225	20952,726	285,972	**144,165**	202,995	267,471
	10	3373,571	26,470	**12,703**	15,993	24,209	9197,474	88,054	**40,949**	60,647	84,856	18857,699	204,586	**103,710**	193,013	221,686
	20	3094,839	21,167	**10,497**	15,565	21,357	8018,080	55,394	**20,738**	61,985	63,706	15600,452	102,585	**41,740**	194,580	150,275
	30	2800,297	16,358	**8,615**	15,427	18,399	7404,228	33,447	**13,824**	66,306	49,003	22533,818	51,356	**22,932**	234,098	110,283
	50	2466,713	9,022	**3,896**	16,002	12,773	13807,323	10,843	**5,619**	91,425	32,835	23767,392	11,935	**7,968**	437,296	39,234
	70	6152,349	3,703	**1,859**	19,252	8,878	7814,813	2,737	**2,341**	187,860	8,434	8190,895	**4,661**	4,964	1402,440	9,972

Indian

Tabela A.6: Valores de *accuracy* obtidos com o classificador KNN, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Indian.

		Variável 5					Variáveis 5 e 7					Variável 5, 7 e 6				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	5	69,11	69,25	69,11	69,11	**69,40**	**69,68**	68,68	69,11	68,97	69,54	**70,40**	69,25	69,40	69,97	69,68
	10	69,11	68,97	68,53	69,68	**70,11**	68,97	**69,83**	68,68	68,97	69,40	69,54	69,40	69,54	**70,26**	69,68
	20	**69,54**	69,83	68,82	68,82	68,68	70,40	70,26	69,11	**70,69**	**70,69**	68,68	70,55	**70,69**	69,54	68,97
	30	**69,97**	69,25	69,83	68,97	68,68	67,96	69,25	69,25	69,68	**69,97**	**70,11**	69,97	**70,11**	69,25	68,97
	50	**70,40**	68,53	70,11	69,54	69,25	69,11	69,97	70,11	69,97	**70,40**	69,11	69,25	**70,55**	69,54	70,26
	70	68,97	70,55	**70,83**	70,69	70,69	70,69	69,40	**71,12**	70,40	69,25	68,68	**70,98**	70,26	68,82	70,55

Tabela A.7: Valores de *accuracy* obtidos com o classificador SVM, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Indian.

		Variável 5					Variáveis 5 e 7					Variáveis 5, 7 e 6				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	5	70,83	70,26	70,40	70,55	**71,70**	**71,98**	71,41	69,97	70,55	70,40	**72,27**	71,84	72,13	71,12	71,26
	10	71,12	71,55	**71,98**	70,11	71,12	71,12	71,26	71,26	**71,84**	69,68	70,83	**71,98**	**71,98**	71,26	70,69
	20	70,69	**71,55**	70,26	70,11	71,26	70,11	**71,55**	69,54	70,40	69,97	71,12	70,98	**71,98**	69,97	70,83
	30	70,11	70,83	70,55	70,69	**70,98**	71,12	70,83	70,55	70,26	**71,55**	70,83	71,12	69,83	**71,26**	70,69
	50	68,82	70,26	69,83	**70,83**	69,54	70,55	70,98	69,97	70,26	**71,55**	71,12	70,98	**71,84**	69,83	68,82
	70	**71,70**	71,41	70,40	70,83	70,26	71,41	**71,84**	70,55	70,83	71,12	70,55	70,11	70,69	70,98	**71,41**

Tabela A.8: Valores de *accuracy* obtidos com o classificador RF, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Indian.

		Variável 5					Variáveis 5 e 7					Variáveis 5, 7 e 6				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	5	69,68	**72,13**	70,40	70,98	69,97	68,82	70,26	70,26	70,11	**70,40**	68,25	**70,55**	69,68	70,26	70,26
	10	70,83	69,68	**71,26**	69,40	70,40	68,10	69,11	69,83	**69,97**	69,83	68,10	**72,56**	70,69	68,97	68,82
	20	70,69	69,83	70,40	**72,13**	71,98	**71,55**	70,11	70,11	70,11	67,82	70,11	69,83	**72,27**	69,25	70,69
	30	**71,84**	71,70	69,83	69,68	70,83	**72,99**	70,55	70,98	71,41	70,83	70,11	**70,98**	70,83	69,40	68,68
	50	70,69	70,40	68,39	**71,55**	**71,55**	**71,26**	67,96	69,68	69,83	68,82	68,10	**71,12**	70,40	69,40	68,68
	70	71,12	**72,41**	69,54	71,55	70,40	70,83	69,97	70,40	**70,98**	70,69	68,39	69,11	69,68	**69,97**	69,11

Tabela A.9: Valores de *accuracy* obtidos com o classificador MLP, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Indian.

		Variável 5					Variáveis 5 e 7					Variáveis 5, 7 e 6				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	69,54	70,55	70,11	**71,12**	70,55	68,39	**72,70**	70,26	**72,70**	70,83	70,83	69,25	**71,41**	70,98	70,98
	10	70,11	69,83	69,68	**71,26**	69,97	68,68	**71,12**	**71,12**	70,69	69,25	**70,98**	69,25	70,69	70,69	69,68
	20	**70,98**	69,25	69,97	69,54	**70,98**	70,83	**71,41**	71,12	70,83	69,25	71,12	70,83	**72,13**	69,25	69,54
	30	**72,41**	70,98	70,26	70,11	70,69	68,53	69,97	70,11	**71,98**	68,10	**70,83**	69,83	69,25	69,40	69,11
	50	70,11	69,54	**71,70**	70,40	70,26	70,69	69,11	70,26	70,69	**71,12**	**70,69**	70,55	70,26	69,40	70,26
	70	69,25	69,25	**70,55**	69,97	69,83	70,26	68,68	**70,69**	69,97	69,68	**71,98**	69,54	70,40	68,68	69,54

Tabela A.10: Tempo, em segundos, que os diferentes métodos de imputação demoraram, para todas as percentagens de MD testadas com o *dataset* Indian.

		Variável 5					Variáveis 5 e 7					Variáveis 5, 7 e 6				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	3663,113	17,853	**7,528**	19,136	16,753	13772,483	60,635	**30,686**	71,567	62,327	30781,008	163,160	**64,435**	199,596	169,722
	10	3350,954	15,629	**7,712**	17,108	15,679	13738,905	48,658	**23,315**	68,093	55,373	28999,063	122,334	**41,114**	198,194	146,156
	20	3240,087	12,409	**5,694**	17,235	13,690	12889,275	32,948	**12,292**	67,218	46,704	26614,968	63,352	**24,424**	208,545	118,138
	30	3111,116	10,037	**4,030**	18,751	12,521	12938,787	19,757	**8,769**	72,538	39,010	32341,880	30,919	**13,274**	247,469	99,450
	50	3166,728	5,590	**2,473**	18,856	10,311	18211,058	5,991	**3,716**	107,833	27,184	30447,226	6,465	**5,424**	499,317	24,142
	70	3886,158	2,445	**1,419**	22,874	8,382	9546,856	**1,619**	1,850	238,310	5,980	8015,714	**2,334**	4,238	2099,174	6,586

Iris

Tabela A.11: Valores de *accuracy* obtidos com o classificador SVM, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Iris.

		Variáveis 1 e 2					Variáveis 3 e 4				
Métodos de Imputação		**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**	**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**
% MD	**5**	96,00	96,00	**96,22**	95,78	95,78	94,89	95,11	**96,00**	94,89	94,67
	10	96,22	95,78	96,22	**96,67**	96,22	**94,89**	93,11	94,44	94,00	94,00
	20	94,67	95,11	96,00	**96,67**	95,56	**94,22**	92,00	93,56	93,11	94,00
	30	**96,22**	96,00	**96,22**	96,00	95,11	**92,89**	89,56	90,22	90,89	**92,89**
	50	94,44	**96,00**	95,11	95,78	94,00	86,89	84,67	81,78	84,00	**87,56**
	70	93,33	**96,22**	82,67	95,11	92,22	79,11	78,89	69,33	79,33	**79,56**

Tabela A.12: Valores de *accuracy* obtidos com o classificador RF, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Iris.

		Variáveis 1 e 2					Variáveis 3 e 4				
Métodos de Imputação		**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**	**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**
% MD	**5**	95,56	**95,78**	95,11	**95,78**	95,33	**96,00**	94,67	94,89	95,56	95,56
	10	**96,22**	95,56	95,33	95,56	95,56	**94,89**	93,78	93,78	93,78	93,33
	20	95,78	95,56	**96,00**	95,78	94,89	**93,78**	93,11	92,89	92,22	93,11
	30	94,89	95,56	**96,44**	94,89	95,56	91,56	89,33	90,00	90,22	**91,78**
	50	**96,67**	95,78	96,22	95,78	95,56	**87,78**	86,67	82,22	85,11	87,33
	70	95,33	96,44	**96,67**	95,78	96,00	79,78	79,56	66,22	79,11	**80,22**

Tabela A.13: Valores de *accuracy* obtidos com o classificador MLP, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Iris.

		Variáveis 1 e 2					**Variáveis 3 e 4**				
Métodos de Imputação		**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**	**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**
% MD	**5**	92,22	94,22	94,44	94,67	**95,56**	95,11	94,00	95,56	**95,78**	95,11
	10	95,78	**96,22**	95,78	96,00	**96,22**	94,00	93,11	**94,44**	**94,44**	93,78
	20	94,67	96,22	**96,89**	94,89	94,67	**93,56**	92,22	92,89	91,11	93,11
	30	**96,00**	95,33	95,11	94,22	95,33	90,67	88,00	91,33	**92,22**	91,56
	50	94,22	**95,11**	93,33	93,56	**95,11**	85,56	79,11	83,33	**85,78**	85,11
	70	95,56	95,11	82,22	**95,78**	95,56	75,78	77,56	62,89	79,11	**80,00**

Tabela A.14: Tempo, em segundos, que os diferentes métodos de imputação demoraram, para todas as percentagens de MD testadas com o *dataset* Iris.

		Variáveis 1 e 2					**Variáveis 3 e 4**				
Métodos de Imputação		**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**	**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**
% MD	**5**	129505,01	643,97	**69,59**	4420,51	940,83	146465,15	327,97	**69,09**	4482,90	973,29
	10	114317,40	506,04	**62,59**	4868,79	734,77	135197,42	269,21	**65,01**	4892,20	758,76
	20	107197,86	335,90	**48,21**	5462,14	501,16	118578,32	193,54	**52,19**	5700,08	522,86
	30	102807,79	178,44	**39,74**	7207,46	313,24	122552,28	110,11	**40,30**	7349,46	350,15
	50	194881,27	63,39	**27,96**	12537,88	167,34	178013,65	64,51	**31,02**	13445,06	190,53
	70	121472,42	40,90	**29,17**	34855,37	89,41	92901,14	45,81	**26,54**	41753,42	94,91

Banknote

Tabela A.15: Valores de *accuracy* obtidos com o classificador KNN, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Banknote.

		Variável 1					Variáveis 1 e 2					Variáveis 1, 2 e 3				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	99,42	**99,59**	**99,59**	99,20	**99,59**	99,47	**99,51**	99,44	99,10	99,42	99,13	**99,30**	99,15	98,59	99,13
	10	99,37	99,42	99,32	98,88	**99,44**	98,74	98,69	**98,83**	98,11	98,79	98,35	**98,40**	98,33	97,52	98,23
	20	**98,93**	**98,93**	98,35	97,82	98,62	97,09	**97,55**	96,33	95,75	97,38	**96,21**	96,12	95,36	93,79	95,46
	30	98,35	**98,64**	97,60	97,57	98,01	**94,95**	94,81	93,86	93,33	94,49	91,80	**92,60**	91,97	90,32	92,43
	50	97,43	**97,72**	96,00	96,48	97,11	88,16	**88,47**	86,36	86,94	87,89	82,62	**83,23**	80,92	81,48	82,43
	70	96,65	**97,09**	93,11	96,70	96,46	**81,12**	80,39	74,59	79,64	79,85	68,64	**70,90**	66,43	69,17	69,32

Tabela A.16: Valores de *accuracy* obtidos com o classificador SVM, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Banknote.

		Variável 1					Variáveis 1 e 2					Variáveis 1, 2 e 3				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	99,66	99,66	**99,68**	99,37	99,54	**99,56**	**99,56**	99,49	99,13	99,37	**99,37**	99,32	99,30	98,71	99,15
	10	99,42	99,39	99,44	98,83	**99,47**	**98,98**	98,96	98,88	98,25	98,74	**98,64**	98,52	98,47	97,67	98,45
	20	99,03	**99,05**	98,54	97,96	98,54	97,09	**97,33**	96,67	95,73	97,28	**96,26**	96,19	95,41	94,44	95,85
	30	98,50	**98,59**	97,67	97,38	97,91	94,95	**95,22**	94,22	93,88	94,66	92,48	**92,96**	91,97	90,51	92,55
	50	97,38	**97,43**	95,92	96,36	96,72	89,03	**89,17**	86,82	87,84	88,35	83,69	**83,81**	80,85	82,45	83,06
	70	**96,89**	96,84	93,06	96,04	96,04	80,68	**80,95**	73,57	80,39	79,76	71,46	**72,31**	65,87	70,10	69,76

Tabela A.17: Valores de *accuracy* obtidos com o classificador RF, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Banknote.

		Variável 1					Variáveis 1 e 2					Variáveis 1, 2 e 3				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	**98,93**	98,91	98,86	98,69	98,91	**98,93**	98,91	98,81	98,64	98,96	**98,83**	98,67	98,57	98,16	98,45
	10	98,79	**98,88**	98,69	98,33	98,81	**98,45**	**98,45**	98,11	97,96	98,06	**98,30**	98,08	97,74	97,50	97,91
	20	**98,54**	98,52	97,91	97,48	97,94	96,89	**97,11**	96,17	95,61	96,84	**95,87**	95,66	95,27	94,25	95,56
	30	**98,30**	98,23	97,33	96,92	97,62	94,66	**95,15**	93,57	93,28	94,68	91,46	**92,91**	92,04	90,49	92,72
	50	96,75	**97,31**	95,58	95,87	96,58	88,45	**89,13**	87,72	87,84	88,76	82,86	82,82	80,70	**83,54**	82,84
	70	96,41	**96,55**	93,40	95,56	96,07	**83,25**	81,38	77,06	81,31	81,89	68,83	**71,92**	67,77	**71,92**	70,83

Tabela A.18: Valores de *accuracy* obtidos com o classificador MLP, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Banknote.

		Variável 1					Variáveis 1 e 2					Variáveis 1, 2 e 3				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	98,79	98,86	**98,98**	98,40	98,67	**99,03**	98,83	98,52	98,40	98,81	**98,79**	98,37	98,35	97,77	98,16
	10	**98,69**	98,62	**98,69**	97,96	98,59	**98,40**	97,77	97,77	96,94	97,91	**97,86**	97,43	97,60	96,72	97,40
	20	**98,20**	98,11	97,60	96,41	97,74	96,21	**96,46**	96,09	94,85	96,09	95,39	**95,66**	95,44	92,94	94,76
	30	97,62	**97,67**	97,01	95,44	97,14	93,93	**94,44**	93,83	91,77	93,74	91,41	**92,01**	91,31	89,59	91,97
	50	**96,75**	96,43	95,63	93,30	95,44	86,60	**87,84**	86,38	84,93	87,43	81,21	**83,01**	80,56	80,07	78,88
	70	94,61	**95,17**	92,69	91,36	93,52	77,57	**78,45**	73,67	75,10	76,21	68,45	**68,79**	66,55	67,96	67,14

Tabela A.19: Tempo, em segundos, que os diferentes métodos de imputação demoraram, para todas as percentagens de MD testadas com o *dataset* Banknote.

		Variável 1					Variáveis 1 e 2					Variáveis 1, 2 e 3				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	17594,625	693,550	44,140	**12,717**	35,705	43835,672	1726,956	160,335	**47,924**	123,967	86071,697	3168,108	394,439	**134,080**	281,776
	10	16121,218	595,028	39,786	**11,777**	32,615	38395,191	1328,638	121,917	**43,521**	102,127	70399,885	2236,212	305,259	**134,067**	234,300
	20	14476,512	441,545	31,968	**11,327**	28,246	28452,276	739,786	76,432	**41,855**	79,507	43915,643	1055,035	154,683	**125,238**	165,254
	30	11766,456	314,887	24,201	**12,157**	23,519	20251,288	388,168	47,212	**42,426**	58,916	28171,175	440,611	**58,586**	129,088	117,457
	50	7225,089	120,800	13,043	**10,751**	16,667	8731,807	76,200	**11,574**	47,911	32,788	10679,591	46,771	**13,824**	192,579	66,676
	70	4178,333	30,509	**4,639**	12,397	10,722	3232,668	6,588	**3,690**	85,795	15,686	11654,917	**3,379**	6,204	639,312	9,422

Seeds

Tabela A.20: Valores de *accuracy* obtidos com o classificador KNN, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Seeds.

		Variável 1					Variáveis 1 e 2					Variáveis 1, 2 e 7				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	5	**92,52**	91,75	92,06	91,90	92,06	**92,52**	91,11	92,06	91,27	92,06	**92,06**	90,95	91,11	91,43	91,43
	10	**92,52**	91,11	91,90	91,59	92,06	**92,52**	91,43	91,59	91,43	92,06	**91,61**	91,27	90,63	90,79	91,11
	20	**92,52**	91,43	91,90	91,90	92,06	**92,06**	90,79	91,59	91,59	91,90	90,25	**91,27**	90,95	90,32	90,79
	30	**92,06**	90,95	91,75	**92,06**	91,75	**91,61**	91,59	91,11	91,59	91,59	90,25	89,84	**91,11**	89,37	90,63
	50	**92,06**	91,59	91,59	91,75	91,90	91,38	**91,59**	90,63	91,27	90,95	88,89	89,68	89,68	**90,00**	89,84
	70	**92,06**	91,11	91,43	91,27	**92,06**	**91,84**	91,43	90,79	91,75	89,52	89,12	89,68	73,33	**90,79**	88,89

Tabela A.21: Valores de *accuracy* obtidos com o classificador SVM, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Seeds.

		Variável 1					Variáveis 1 e 2					Variáveis 1, 2 e 7				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	5	**92,06**	**92,06**	90,63	90,95	90,63	91,61	**91,90**	90,79	91,27	90,63	91,16	**91,90**	90,16	90,79	90,16
	10	91,61	91,43	90,79	**91,75**	90,63	**92,06**	91,59	91,43	91,27	90,79	89,57	**91,27**	91,11	90,79	90,00
	20	92,06	**92,54**	90,63	91,11	90,95	92,29	91,43	91,75	**92,54**	90,63	**90,48**	90,00	**90,48**	90,32	90,16
	30	92,06	**92,38**	91,75	91,59	90,79	92,29	92,38	92,54	**92,86**	91,59	90,25	90,63	89,52	90,00	**91,11**
	50	92,06	**92,54**	92,06	92,22	90,79	91,84	**92,70**	90,95	92,22	90,79	87,30	89,84	**90,63**	90,16	88,25
	70	91,61	92,06	**92,38**	91,43	91,11	**93,42**	92,22	89,21	91,75	91,27	**88,66**	88,57	74,44	89,37	88,41

Tabela A.22: Valores de *accuracy* obtidos com o classificador RF, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Seeds.

		Variável 1					Variáveis 1 e 2					Variáveis 1, 2 e 7				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	5	91,84	90,16	90,32	91,11	**91,90**	90,93	**92,54**	90,63	92,22	90,95	**91,84**	89,84	89,84	90,63	91,75
	10	91,38	**91,43**	90,48	91,27	90,95	91,84	**91,90**	90,16	90,32	91,27	**90,93**	89,68	90,79	90,48	90,32
	20	**91,61**	91,43	90,48	91,59	91,11	90,48	90,95	**91,27**	90,79	90,79	89,34	88,57	89,21	**90,00**	89,21
	30	90,93	90,79	90,63	89,84	**91,75**	90,93	90,16	90,95	90,16	**91,59**	89,57	87,62	89,68	**90,16**	88,89
	50	91,38	90,63	91,27	**91,43**	90,95	90,48	**91,27**	89,05	90,79	90,00	87,76	87,78	83,97	88,10	**88,25**
	70	90,70	**91,27**	90,32	90,48	**91,27**	89,80	**90,63**	89,21	90,48	90,32	86,17	87,78	76,51	**88,10**	87,14

Tabela A.23: Valores de *accuracy* obtidos com o classificador MLP, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Seeds.

		Variável 1					Variáveis 1 e 2					Variáveis 1, 2 e 7				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	92,29	**93,17**	92,70	93,02	92,86	90,02	**92,70**	92,38	92,06	92,38	90,93	91,90	92,22	91,75	**92,38**
	10	90,93	92,22	92,70	92,22	**93,02**	89,80	92,06	92,54	92,70	**93,81**	90,25	89,52	**92,06**	91,59	**92,06**
	20	90,48	92,70	91,43	92,06	**93,17**	89,57	91,27	92,38	91,43	**93,33**	89,12	89,21	**91,11**	90,32	90,48
	30	91,16	92,38	92,86	92,06	**93,02**	90,70	**93,17**	92,70	91,59	**93,17**	88,21	88,89	**90,48**	**90,48**	89,52
	50	89,12	91,75	**93,49**	92,38	93,33	90,70	91,11	**92,70**	91,27	92,54	87,76	**90,79**	87,94	89,21	88,73
	70	90,02	**93,49**	92,70	92,38	93,17	90,70	91,59	90,95	91,27	**91,90**	87,30	88,89	72,54	**89,05**	88,41

Tabela A.24: Tempo, em segundos, que os diferentes métodos de imputação demoraram, para todas as percentagens de MD testadas com o *dataset* Seeds.

		Variável 1					Variáveis 1 e 2					Variáveis 1, 2 e 7				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	8802,444	2,925	**0,846**	20,645	9,848	19117,460	9,476	**3,000**	90,076	34,636	25361,723	23,637	**7,712**	257,503	96,624
	10	8746,805	2,394	**0,853**	21,877	9,109	19160,638	7,353	**2,454**	98,645	33,200	25510,445	17,423	**6,005**	290,408	90,455
	20	8537,061	1,896	**0,699**	23,772	8,781	18772,618	5,188	**1,936**	117,146	24,418	25607,876	9,985	**4,353**	378,261	46,152
	30	8500,831	1,642	**0,567**	39,465	8,197	18922,757	3,877	**1,559**	143,119	14,789	23445,574	6,277	**3,167**	523,194	24,824
	50	8264,098	1,188	**0,530**	34,373	4,421	11984,409	1,789	**0,983**	252,200	6,322	15248,463	3,357	**2,198**	1154,108	9,664
	70	6003,784	0,750	**0,385**	49,600	2,614	3141,169	1,271	**0,841**	596,737	2,969	2849,145	3,452	**2,105**	3473,826	5,125

Laryngeal

Tabela A.25: Valores de *accuracy* obtidos com o classificador KNN, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Laryngeal.

		Variável 3					Variáveis 3 e 7					Variáveis 3, 7 e 8				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	5	**73,27**	73,11	72,26	71,98	72,26	**72,96**	72,17	72,08	72,64	71,70	**73,58**	72,17	71,60	72,36	72,08
	10	72,33	72,26	71,70	71,89	**72,64**	**73,27**	72,36	71,60	72,36	72,64	71,38	**72,26**	71,79	**72,26**	71,89
	20	**73,90**	72,17	71,89	73,68	72,26	**74,21**	73,21	72,36	71,98	72,64	71,70	**74,06**	72,55	72,26	71,98
	30	**72,64**	72,08	71,42	72,26	72,45	72,01	**73,21**	72,36	72,92	72,17	**72,64**	**72,64**	71,51	71,70	71,13
	50	70,75	**72,64**	71,60	**72,64**	71,89	**73,27**	72,83	71,98	72,45	73,02	**73,58**	71,60	71,98	72,55	71,13
	70	72,96	**73,30**	71,89	72,64	72,26	**73,27**	72,45	71,98	72,26	73,02	72,33	71,98	58,87	**72,45**	72,17

Tabela A.26: Valores de *accuracy* obtidos com o classificador SVM, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Laryngeal.

		Variável 3					Variáveis 3 e 7					Variáveis 3, 7 e 8				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	5	**73,27**	72,36	72,08	72,55	72,83	**72,64**	71,79	72,36	72,17	72,45	**73,90**	71,70	72,92	72,55	72,55
	10	**73,58**	70,94	72,83	71,23	72,36	**73,27**	72,74	72,26	71,98	72,55	**73,27**	72,64	71,79	71,70	72,26
	20	**74,53**	72,17	71,79	71,89	72,26	72,64	72,36	**73,02**	72,45	**73,02**	72,33	**72,74**	72,36	71,79	72,64
	30	**74,21**	72,36	72,64	73,02	71,04	**73,58**	72,92	72,17	72,74	72,45	**73,58**	72,55	72,74	72,55	72,64
	50	**73,27**	72,83	72,08	71,98	72,45	72,96	72,74	72,08	**73,30**	71,98	**74,53**	71,23	72,26	71,89	71,79
	70	72,33	**72,92**	71,79	72,74	71,42	**74,21**	72,08	71,89	72,08	73,02	71,70	72,17	61,70	**73,49**	71,23

Tabela A.27: Valores de *accuracy* obtidos com o classificador RF, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Laryngeal.

		Variável 3					Variáveis 3 e 7					Variáveis 3, 7 e 8				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	5	**71,70**	71,23	70,28	70,85	69,72	71,07	70,00	70,57	71,13	**71,23**	**71,70**	70,09	71,13	70,75	70,57
	10	**75,16**	71,42	71,13	72,08	70,47	**71,70**	70,47	71,04	70,28	70,19	**71,70**	71,42	70,09	70,38	70,09
	20	**72,96**	69,15	69,34	69,72	70,85	71,38	68,96	70,66	69,91	**71,70**	70,44	70,38	**70,85**	70,09	69,72
	30	**71,70**	71,60	69,53	71,04	69,91	**72,01**	70,47	70,57	70,28	70,94	**72,01**	71,70	70,85	71,32	69,62
	50	**72,33**	70,19	70,75	69,43	72,08	**72,64**	70,38	69,91	71,42	69,62	**73,90**	70,57	71,42	70,94	70,75
	70	**72,33**	70,66	70,28	70,38	70,47	**72,01**	70,94	69,91	70,66	71,60	69,81	**71,32**	69,81	70,94	69,72

Tabela A.28: Valores de *accuracy* obtidos com o classificador MLP, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Laryngeal.

		Variável 3					Variáveis 3 e 7					Variáveis 3, 7 e 8				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	**73,27**	69,72	70,94	69,25	69,62	**73,58**	68,96	69,34	69,91	68,68	70,13	69,34	68,87	**70,38**	69,43
	10	**71,70**	71,51	71,04	70,00	68,87	**73,27**	69,34	71,04	71,32	68,77	**72,64**	69,15	70,66	71,51	70,38
	20	**72,96**	69,81	68,96	71,51	66,79	**71,70**	67,64	69,06	69,43	70,09	**72,33**	68,96	69,62	69,53	71,51
	30	**72,01**	69,15	71,23	69,72	69,43	**72,64**	68,87	69,91	69,43	67,92	**70,13**	68,68	68,87	69,53	68,58
	50	**73,58**	68,40	71,13	68,96	69,53	**72,64**	69,62	67,92	68,49	68,30	**71,70**	69,25	69,53	70,28	68,02
	70	72,33	68,30	70,57	**73,02**	66,32	70,13	**72,26**	70,57	69,91	68,49	**72,64**	69,72	57,64	66,23	69,06

Tabela A.29: Tempo, em segundos, que os diferentes métodos de imputação demoraram, para todas as percentagens de MD testadas com o *dataset* Laryngeal.

		Variável 3					Variáveis 3 e 7					Variáveis 3, 7 e 8				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	5965,650	5,474	**1,792**	15,951	10,947	14581,453	17,354	**6,060**	62,768	40,270	31534,064	44,368	**13,681**	179,873	112,861
	10	6461,352	4,788	**1,567**	15,720	10,850	16791,014	14,698	**4,819**	65,885	35,988	32197,995	36,005	**11,373**	203,434	102,609
	20	8702,139	3,768	**1,356**	16,663	9,769	22617,759	9,790	**3,653**	75,821	32,347	43710,686	18,575	**7,800**	251,229	94,125
	30	9909,826	2,920	**0,940**	18,370	9,098	25089,794	5,619	**2,266**	89,555	28,731	46644,490	10,215	**4,898**	335,019	52,869
	50	10751,642	1,557	**0,695**	20,958	8,354	17313,814	2,045	**1,219**	156,011	8,826	20424,911	2,880	**2,335**	805,943	11,878
	70	5386,622	0,812	**0,489**	30,833	4,139	4711,887	**0,893**	0,907	386,047	3,160	2211,367	**1,795**	2,504	3476,356	4,162

Voice

Tabela A.30: Valores de *accuracy* obtidos com o classificador KNN, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Voice.

		Variável 2					Variáveis 2 e 8					Variáveis 2, 8 e 7				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	**79,01**	78,17	78,59	78,73	78,59	78,73	78,17	78,17	77,61	**79,30**	**79,01**	77,89	77,75	78,87	78,31
	10	77,32	**78,45**	78,31	78,17	78,31	77,89	**78,45**	77,75	78,31	**78,45**	78,03	**78,31**	77,61	**78,31**	78,03
	20	**78,87**	77,61	78,73	78,45	77,75	76,90	78,31	78,31	76,48	**78,87**	78,03	**78,31**	78,03	77,46	77,18
	30	75,77	**77,04**	75,35	75,63	76,90	75,92	75,21	75,77	76,48	**76,76**	76,20	76,34	75,92	76,90	**77,32**
	50	74,37	**75,21**	74,37	74,93	74,79	74,93	**75,07**	74,93	74,08	**75,07**	73,80	74,65	74,65	75,07	**76,34**
	70	73,38	**75,35**	73,94	73,52	74,08	75,07	**75,21**	71,83	73,66	73,52	73,94	**74,37**	73,52	73,66	71,69

Tabela A.31: Valores de *accuracy* obtidos com o classificador SVM, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Voice.

		Variável 2					Variáveis 2 e 8					Variáveis 2, 8 e 7				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	**80,28**	79,44	79,15	79,58	79,30	79,44	79,30	79,30	79,44	**79,72**	**80,00**	79,58	79,15	79,01	79,30
	10	78,45	79,01	**79,72**	79,15	78,45	79,30	79,30	78,03	**79,44**	77,89	78,59	78,03	**79,15**	79,01	78,45
	20	78,17	**79,15**	78,03	78,87	77,04	78,45	**79,44**	78,45	79,01	78,45	78,31	**79,44**	78,31	78,59	78,73
	30	**77,46**	77,04	77,18	76,62	76,90	**78,87**	76,90	77,04	77,18	76,48	77,46	77,75	77,32	**78,31**	77,75
	50	**77,04**	76,34	75,21	76,06	76,62	75,92	76,20	76,34	76,48	**77,18**	75,77	**76,90**	74,93	76,34	75,07
	70	75,63	**76,06**	74,65	74,37	75,35	**76,62**	**76,62**	76,06	76,34	76,48	75,35	76,20	**76,48**	**76,48**	75,49

Tabela A.32: Valores de *accuracy* obtidos com o classificador RF, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Voice.

		Variável 2					Variáveis 2 e 8					Variáveis 2, 8 e 7				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	78,73	77,75	78,31	78,31	**79,58**	77,18	78,45	**79,72**	78,45	78,59	78,59	78,59	78,31	78,59	**79,44**
	10	**79,58**	78,45	78,59	79,01	78,59	78,59	78,59	78,31	78,87	**79,58**	77,18	79,72	**80,56**	78,45	77,75
	20	78,03	78,59	77,04	**78,59**	76,90	**78,87**	78,31	77,32	78,59	**78,87**	78,73	78,45	79,01	79,58	**80,00**
	30	78,17	78,31	78,03	**79,01**	76,34	79,15	78,59	77,75	**80,14**	77,89	78,03	78,87	79,72	**80,14**	78,59
	50	78,87	78,45	78,59	77,32	**79,01**	78,87	78,45	**79,30**	77,75	78,17	77,89	78,17	**79,86**	79,15	78,73
	70	**78,59**	78,03	76,76	78,45	77,75	**79,58**	78,59	79,44	78,73	77,89	**80,14**	79,15	77,89	79,30	78,45

Tabela A.33: Valores de *accuracy* obtidos com o classificador MLP, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Voice.

		Variável 2					Variáveis 2 e 8					Variáveis 2, 8 e 7				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	1908,534	2,597	**1,086**	16,265	8,334	4534,135	9,048	**2,699**	71,418	29,578	14561,784	28,480	**6,912**	212,283	84,827
	10	1802,019	2,291	**0,724**	15,907	7,185	4350,656	8,541	**2,288**	77,674	28,444	14450,833	21,123	**5,678**	244,765	83,593
	20	1776,247	2,004	**0,597**	18,640	7,274	4248,293	5,618	**1,811**	94,352	25,405	13351,930	10,834	**4,014**	302,902	49,447
	30	1778,068	1,620	**0,535**	20,311	7,258	4258,087	3,740	**1,405**	109,166	16,019	13862,536	5,825	**3,418**	442,635	26,328
	50	2446,621	1,013	**0,434**	27,117	4,211	5115,336	1,640	**0,817**	207,125	6,251	8313,994	2,530	**1,834**	1094,053	7,209
	70	1905,029	0,607	**0,284**	40,242	1,985	2440,650	0,895	**0,638**	482,037	2,300	1277,814	1,764	**1,706**	4246,630	3,096

Tabela A.34: Tempo, em segundos, que os diferentes métodos de imputação demoraram, para todas as percentagens de MD testadas com o *dataset* Voice.

		Variável 2					Variáveis 2 e 8					Variáveis 2, 8 e 7				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	1908,534	2,597	**1,086**	16,265	8,334	4534,135	9,048	**2,699**	71,418	29,578	14561,784	28,480	**6,912**	212,283	84,827
	10	1802,019	2,291	**0,724**	15,907	7,185	4350,656	8,541	**2,288**	77,674	28,444	14450,833	21,123	**5,678**	244,765	83,593
	20	1776,247	2,004	**0,597**	18,640	7,274	4248,293	5,618	**1,811**	94,352	25,405	13351,930	10,834	**4,014**	302,902	49,447
	30	1778,068	1,620	**0,535**	20,311	7,258	4258,087	3,740	**1,405**	109,166	16,019	13862,536	5,825	**3,418**	442,635	26,328
	50	2446,621	1,013	**0,434**	27,117	4,211	5115,336	1,640	**0,817**	207,125	6,251	8313,994	2,530	**1,834**	1094,053	7,209
	70	1905,029	0,607	**0,284**	40,242	1,985	2440,650	0,895	**0,638**	482,037	2,300	1277,814	1,764	**1,706**	4246,630	3,096

Transfusion

Tabela A.35: Valores de *accuracy* obtidos com o classificador KNN, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Transfusion.

		Variável 1					Variáveis 1 e 3				
Métodos de Imputação		**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**	**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**
% MD	**5**	78,26	78,53	**78,75**	78,57	78,39	77,90	78,04	**78,53**	77,59	77,59
	10	78,13	78,30	78,21	**78,44**	78,35	**77,46**	77,14	77,01	77,41	77,37
	20	77,54	76,65	77,59	**77,63**	77,14	76,03	75,89	75,89	**76,38**	76,29
	30	**77,63**	77,01	77,19	77,01	77,14	75,89	75,85	76,38	**76,43**	76,07
	50	77,10	77,14	76,70	76,61	**77,77**	**75,76**	75,45	75,45	75,71	75,63
	70	76,21	76,83	**77,41**	76,29	76,25	75,09	75,40	**75,71**	75,27	74,69

Tabela A.36: Valores de *accuracy* obtidos com o classificador SVM, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Transfusion.

		Variável 1					Variáveis 1 e 3				
Métodos de Imputação		**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**	**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**
% MD	**5**	78,48	**78,71**	78,13	78,53	78,57	**78,17**	77,90	77,68	77,68	77,99
	10	78,53	78,30	78,13	**78,57**	77,77	77,46	**78,26**	77,41	77,81	77,77
	20	**78,48**	76,96	77,23	77,90	77,37	76,16	76,21	76,43	76,29	**76,56**
	30	**77,46**	76,43	76,70	76,38	77,01	76,25	76,12	76,25	75,94	**76,38**
	50	**77,68**	77,59	76,52	76,43	76,88	76,25	**76,52**	76,12	76,03	76,47
	70	76,34	76,34	**77,23**	76,16	76,16	75,54	75,98	76,07	**76,25**	75,71

Tabela A.37: Valores de *accuracy* obtidos com o classificador RF, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Transfusion.

		Variável 1					Variáveis 1 e 3				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	73,57	73,88	74,06	**74,42**	74,15	73,88	73,48	73,48	**74,29**	73,62
	10	74,15	73,84	74,02	**74,60**	73,39	73,97	73,93	73,62	**74,51**	73,17
	20	73,71	74,02	**74,29**	73,66	73,35	74,11	**74,82**	73,26	74,42	74,29
	30	73,48	73,39	73,30	**73,97**	73,57	73,44	73,75	**74,06**	73,26	73,13
	50	73,48	73,57	**74,20**	73,84	73,75	73,48	**73,75**	72,68	73,35	73,62
	70	73,97	**74,46**	74,24	73,71	73,88	74,51	73,93	**75,31**	73,35	74,11

Tabela A.38: Valores de *accuracy* obtidos com o classificador MLP, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Transfusion.

		Variável 1					Variáveis 1 e 3				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	77,63	**77,90**	77,77	77,59	77,01	77,41	77,72	**77,99**	77,54	77,32
	10	**77,99**	77,50	77,54	77,46	77,37	**77,68**	**77,68**	76,96	77,23	77,46
	20	77,59	76,88	76,47	**77,63**	77,19	76,61	76,70	76,96	**77,05**	76,96
	30	**77,59**	77,10	76,88	76,88	76,47	**76,56**	76,21	76,38	76,43	76,38
	50	**77,32**	76,96	76,47	76,25	77,23	76,29	**76,52**	75,67	75,94	76,25
	70	76,83	**77,32**	75,80	76,79	77,01	75,94	**76,70**	76,61	75,80	75,89

Tabela A.39: Tempo, em segundos, que os diferentes métodos de imputação demoraram, para todas as percentagens de MD testadas com o *dataset* Transfusion.

		Variável 1					Variáveis 1 e 3				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	5431,122	119,344	15,382	**13,401**	16,444	10096,518	180,017	**47,874**	51,623	46,005
	10	4962,015	106,287	13,235	**13,170**	15,269	9022,601	140,930	**38,684**	52,678	42,222
	20	4170,056	83,195	**10,293**	13,048	13,824	7012,501	93,614	**19,581**	53,920	35,135
	30	3773,693	64,924	**8,678**	13,482	12,037	5390,276	58,620	**13,206**	58,578	31,682
	50	2582,776	34,039	**3,968**	15,121	9,137	3050,786	13,346	**5,083**	82,516	26,429
	70	1587,269	11,607	**1,977**	18,526	7,243	1692,355	2,810	**2,419**	173,791	7,380

Telugu

Tabela A.40: Valores de *accuracy* obtidos com o classificador KNN, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Telugu.

		Variável 2					Variáveis 2 e 1				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	**86,95**	86,26	86,30	86,60	86,45	85,23	85,50	**86,15**	84,35	85,92
	10	84,96	84,96	84,66	84,77	84,50	82,63	83,47	**83,59**	81,87	82,82
	20	**83,40**	82,98	83,13	81,87	81,98	78,40	78,13	**78,85**	76,95	78,70
	30	80,57	81,07	**81,15**	79,54	79,96	73,02	73,02	**73,55**	70,65	72,79
	50	**76,68**	75,31	76,45	75,34	76,15	61,07	60,53	**63,70**	59,62	63,47
	70	71,68	68,51	**73,13**	70,69	72,02	48,47	44,24	49,31	**49,43**	48,59

Tabela A.41: Valores de *accuracy* obtidos com o classificador SVM, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Telugu.

		Variável 2					Variáveis 2 e 1				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	**87,25**	86,87	86,68	86,41	87,21	86,22	86,03	84,92	84,77	**85,38**
	10	84,62	**85,42**	85,15	**85,42**	85,04	**83,17**	83,13	83,09	82,67	82,44
	20	83,55	**83,97**	82,98	82,18	82,33	79,24	79,47	78,02	76,83	**79,77**
	30	80,92	**81,72**	81,49	80,08	80,46	**74,24**	**74,24**	72,60	71,34	74,05
	50	77,25	**77,52**	75,65	76,45	76,37	**65,11**	65,08	62,06	63,09	64,58
	70	73,93	**74,62**	69,96	71,72	74,50	55,15	55,46	48,21	**56,18**	55,08

Tabela A.42: Valores de *accuracy* obtidos com o classificador RF, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Telugu.

		Variável 2					Variáveis 2 e 1				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	5	85,99	85,92	86,11	**86,15**	85,76	85,34	**85,61**	85,31	84,47	85,19
	10	85,27	85,15	85,08	85,31	**85,34**	82,40	**83,24**	82,98	82,33	82,71
	20	**83,51**	82,86	83,24	81,91	82,98	**79,24**	78,78	78,13	77,71	78,21
	30	**81,18**	81,15	80,04	80,88	80,50	72,98	73,17	73,17	72,10	**73,89**
	50	76,91	76,98	72,75	**77,44**	77,33	**65,15**	63,59	61,41	64,01	63,97
	70	**73,74**	73,70	67,86	73,28	73,13	57,02	**57,60**	48,85	56,11	55,11

Tabela A.43: Valores de *accuracy* obtidos com o classificador MLP, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Telugu.

		Variável 2					Variáveis 2 e 1				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	5	82,56	**82,67**	81,91	81,53	81,60	81,03	79,66	80,19	80,34	**81,34**
	10	80,69	80,92	81,18	80,73	**81,83**	79,35	79,24	78,24	78,74	**79,62**
	20	79,12	**79,85**	79,16	79,73	79,27	75,38	**75,92**	74,43	73,66	74,96
	30	77,60	**77,75**	77,02	76,53	76,72	69,58	**70,61**	70,08	68,28	70,23
	50	73,47	**74,12**	71,26	72,63	73,89	58,97	**62,40**	58,70	57,37	60,92
	70	70,88	**71,18**	66,83	67,10	69,01	49,77	**51,15**	45,57	50,00	49,66

Tabela A.44: Tempo, em segundos, que os diferentes métodos de imputação demoraram, para todas as percentagens de MD testadas com o *dataset* Telugu.

		Variável 2					Variáveis 2 e 1				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	5	6205,451	33,296	19,605	**14,086**	19,830	15925,408	153,837	68,307	**50,858**	59,522
	10	5715,340	29,847	17,595	**14,095**	18,819	13836,690	124,413	53,404	**50,713**	51,618
	20	5165,922	24,647	14,334	**13,913**	15,624	10241,672	79,296	**37,180**	51,283	43,714
	30	4219,725	18,480	**11,125**	13,649	13,655	7723,986	48,909	**18,088**	54,165	36,016
	50	3011,391	9,696	**5,199**	13,992	10,294	4235,821	13,036	**6,981**	72,759	28,097
	70	1867,316	3,679	**2,607**	16,756	8,061	2316,084	**2,662**	2,968	146,796	9,247

Red Wine

Tabela A.45: Tempo, em segundos, que os diferentes métodos de imputação demoraram, para todas as percentagens de MD testadas com o *dataset* Red Wine.

		Variável 7					Variáveis 7 e 11					Variáveis 7, 11 e 6				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	11930,646	110,446	96,695	**23,475**	72,451	27027,072	414,974	324,979	**88,107**	250,085	63249,300	1110,519	852,660	**254,858**	676,877
	10	11284,700	100,573	80,967	**22,490**	62,971	25892,568	339,002	265,599	**83,199**	222,341	53391,105	829,277	591,127	**235,568**	564,999
	20	9615,864	82,555	60,837	**20,990**	54,534	20673,945	225,191	164,019	**75,201**	159,819	39256,192	419,462	325,610	**209,441**	366,549
	30	8615,189	65,169	49,350	**19,588**	45,466	16841,018	128,551	93,230	**69,169**	119,240	27975,610	197,154	**150,845**	196,061	232,314
	50	6352,916	32,832	25,808	**17,216**	31,480	13851,437	37,614	**22,652**	66,636	56,919	38566,221	32,177	**23,785**	245,777	113,250
	70	5299,595	13,105	**9,399**	16,808	17,602	19230,772	7,648	**6,903**	106,138	29,109	14786,640	**6,607**	9,221	844,341	15,655

Tabela A.46: Valores de *accuracy* obtidos com o classificador KNN, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Red Wine.

		Variável 7					Variáveis 7 e 11					Variáveis 7, 11 e 6				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	61,52	61,50	**61,71**	60,96	60,94	**62,00**	61,00	61,80	60,77	61,77	61,52	60,56	61,38	61,04	**61,54**
	10	60,89	60,46	**61,21**	60,54	60,96	61,17	60,23	60,75	60,27	**61,48**	**61,17**	60,19	60,69	59,96	60,46
	20	59,85	59,71	60,58	61,11	**61,15**	59,15	59,81	**60,52**	59,69	60,25	59,64	58,60	59,58	58,98	**59,81**
	30	60,47	59,19	**60,77**	59,75	60,19	59,43	59,12	59,77	59,04	**59,85**	**59,43**	58,10	57,75	58,39	58,35
	50	60,13	**60,31**	60,15	59,67	59,96	**60,19**	57,60	57,35	57,24	59,35	**58,94**	56,47	56,58	56,93	57,18
	70	60,40	**60,58**	59,71	59,90	60,19	**60,68**	58,12	56,43	57,31	57,91	54,91	56,60	54,22	**56,70**	55,20

Tabela A.47: Valores de *accuracy* obtidos com o classificador SVM, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Red Wine.

		Variável 7					Variáveis 7 e 11					Variáveis 7, 11 e 6				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	**64,93**	63,42	63,80	63,42	63,11	**65,83**	63,28	63,19	62,80	63,47	**64,72**	63,70	62,59	63,07	63,34
	10	**65,83**	63,42	62,90	63,26	63,13	**65,41**	62,92	62,76	62,86	63,65	**63,67**	63,36	63,15	62,90	63,15
	20	**65,14**	62,88	63,34	63,11	63,40	**64,30**	61,63	62,46	62,13	63,13	**64,30**	62,09	62,11	61,46	63,19
	30	**64,44**	63,15	62,78	63,30	63,07	**64,09**	61,65	61,09	61,94	62,09	**61,52**	60,88	61,09	61,27	60,96
	50	**65,14**	62,84	62,38	62,23	62,36	**64,02**	60,67	59,75	60,77	61,80	**60,68**	60,17	59,44	59,62	59,81
	70	**65,07**	62,21	61,54	62,40	63,90	**62,84**	60,86	59,21	60,27	60,90	57,20	**59,44**	56,41	58,94	56,95

Tabela A.48: Valores de *accuracy* obtidos com o classificador RF, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Red Wine.

		Variável 7					**Variáveis 7 e 11**					**Variáveis 7, 11 e 6**				
Métodos de Imputação		**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**	**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**	**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**
% MD	**5**	**68,89**	66,28	66,85	66,30	66,83	**67,43**	66,41	66,58	67,06	**67,43**	67,01	**67,08**	66,43	66,62	66,24
	10	**67,57**	67,04	67,54	66,66	67,01	**68,20**	65,95	66,74	66,08	66,05	**68,48**	66,47	66,81	66,30	67,35
	20	**67,64**	67,60	66,53	66,91	66,64	**67,15**	65,53	65,49	64,93	66,53	65,83	64,97	65,72	65,66	**66,10**
	30	**67,22**	66,81	66,05	66,60	66,87	**66,81**	64,38	64,70	64,76	65,43	**66,88**	64,53	64,66	64,47	64,78
	50	67,36	**67,72**	66,53	66,14	66,76	**65,76**	64,74	62,94	63,88	64,91	**65,48**	62,90	61,09	62,69	63,36
	70	**68,20**	65,97	66,28	66,64	66,20	**64,93**	63,61	62,15	62,57	64,13	62,84	**63,09**	61,02	62,51	62,38

Tabela A.49: Valores de *accuracy* obtidos com o classificador MLP, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Red Wine.

		Variável 7					**Variáveis 7 e 11**					**Variáveis 7, 11 e 6**				
Métodos de Imputação		**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**	**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**	**MLP**	**SVM**	**KNN**	**SOM**	**OP-ELM**
% MD	**5**	59,57	59,48	**59,73**	58,94	59,06	**59,08**	59,04	58,87	58,91	58,23	**59,92**	58,48	58,43	58,60	59,54
	10	59,08	59,14	59,19	57,60	**59,42**	**59,15**	58,87	58,33	58,79	59,10	58,04	58,58	**59,10**	58,50	59,02
	20	**60,13**	58,89	58,10	59,23	59,10	**58,59**	57,87	58,50	58,33	58,10	58,66	58,02	57,91	57,72	**58,75**
	30	**59,85**	58,35	57,81	59,21	58,71	57,34	57,91	**58,10**	57,77	57,97	55,95	57,37	**58,20**	57,72	57,27
	50	**59,36**	58,35	57,93	59,14	58,66	**59,08**	57,52	57,72	57,33	56,53	**58,59**	56,66	56,64	57,66	57,10
	70	**59,57**	58,81	57,81	58,29	58,98	**57,97**	56,60	57,70	56,95	57,22	54,98	56,45	56,37	55,87	**56,60**

Bupa

Tabela A.50: Valores de *accuracy* obtidos com o classificador KNN, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Bupa.

		Variável 5					Variáveis 5 e 2					Variáveis 5, 2 e 3				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	62,86	61,17	**63,40**	62,72	62,72	62,14	62,72	62,72	62,91	**63,88**	62,50	61,36	62,33	**63,79**	62,04
	10	**63,59**	62,91	61,84	62,62	63,11	60,19	62,33	62,43	**62,62**	61,36	60,07	62,43	**62,72**	61,94	60,87
	20	**62,86**	60,68	60,49	61,84	62,14	62,26	**63,88**	62,82	60,97	60,78	**62,50**	62,43	62,33	61,65	62,14
	30	**62,62**	61,65	60,68	58,54	60,00	60,07	63,20	63,59	**63,69**	61,46	60,44	61,36	**61,46**	59,51	59,51
	50	61,04	61,75	61,84	**62,82**	60,87	60,92	**63,20**	60,10	62,52	60,29	55,58	**60,39**	58,64	59,90	59,03
	70	58,62	**62,04**	60,10	60,39	60,49	59,83	60,58	59,81	**61,55**	59,42	56,55	60,39	58,25	**60,58**	59,90

Tabela A.51: Valores de *accuracy* obtidos com o classificador SVM, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Bupa.

		Variável 5					Variáveis 5 e 2					Variáveis 5, 2 e 3				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	72,33	**73,30**	72,33	72,62	72,43	72,57	**73,30**	**73,30**	72,52	72,23	**72,09**	71,26	71,55	71,55	70,58
	10	71,84	72,04	**72,91**	71,94	72,23	69,17	71,26	71,55	**71,65**	70,29	**71,97**	69,61	71,17	70,00	69,32
	20	71,60	71,55	70,58	**72,52**	71,55	69,17	**72,14**	69,71	69,51	70,19	66,63	67,38	**69,51**	69,13	67,67
	30	**71,60**	70,78	70,39	69,61	69,51	69,30	69,81	69,03	**70,78**	69,42	64,56	64,56	**65,63**	64,37	62,43
	50	68,08	**69,03**	67,48	67,67	67,77	64,81	**66,50**	65,05	65,73	65,24	58,74	60,78	**61,46**	59,03	59,13
	70	64,56	65,73	65,15	**65,83**	65,73	66,75	66,70	65,73	**66,89**	64,66	59,34	**62,14**	58,25	62,04	58,74

Tabela A.52: Valores de *accuracy* obtidos com o classificador RF, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Bupa.

		Variável 5					Variáveis 5 e 2					Variáveis 5, 2 e 3				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	69,30	70,00	70,78	**71,17**	71,07	69,78	**70,29**	**70,29**	69,51	69,90	69,66	**70,58**	70,00	69,22	68,54
	10	67,72	69,42	70,19	**70,29**	69,81	67,72	68,93	**69,32**	69,03	68,16	68,33	67,48	**68,83**	**68,83**	67,86
	20	66,63	68,93	**70,68**	69,90	68,35	67,84	67,09	67,38	**69,03**	67,38	64,93	66,41	**66,89**	66,02	64,76
	30	68,57	66,31	68,06	**68,93**	66,41	66,99	66,31	67,18	66,70	**67,48**	63,35	63,20	63,20	**63,40**	61,84
	50	65,90	65,53	**66,12**	62,72	64,47	61,89	61,36	64,85	63,88	**65,83**	58,62	59,22	**61,26**	60,19	59,32
	70	63,47	63,79	62,23	**65,63**	64,95	64,20	**65,44**	63,40	64,17	60,19	57,77	58,54	**60,10**	60,00	59,81

Tabela A.53: Valores de *accuracy* obtidos com o classificador MLP, para todas as percentagens de MD testadas e para os diferentes métodos de imputação em análise, para o *dataset* Bupa.

		Variável 5					Variáveis 5 e 2					Variáveis 5, 2 e 3				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	63,59	**70,00**	69,71	67,96	66,80	64,44	64,95	67,86	67,28	**68,64**	64,56	**66,50**	66,31	64,85	65,15
	10	64,93	65,73	**66,89**	65,73	65,53	64,56	63,30	65,15	66,89	**67,86**	61,41	66,02	66,41	61,46	**66,60**
	20	63,59	64,66	66,31	**67,28**	66,70	60,80	66,70	**66,89**	65,05	64,37	64,08	**64,27**	61,55	63,40	63,69
	30	66,38	65,53	**66,60**	65,53	65,05	58,98	**65,34**	**65,34**	64,37	64,08	62,38	**63,11**	60,10	59,03	60,39
	50	61,41	64,66	64,47	63,79	**65,05**	62,26	63,50	**65,15**	64,56	63,88	**59,47**	59,42	58,64	58,83	57,57
	70	62,62	62,91	62,14	**65,24**	62,62	62,74	**64,85**	60,49	63,11	62,62	56,67	**58,74**	51,17	56,02	57,86

Tabela A.54: Tempo, em segundos, que os diferentes métodos de imputação demoraram, para todas as percentagens de MD testadas com o *dataset* Bupa.

		Variável 5					Variáveis 5 e 2					Variáveis 5, 2 e 3				
Métodos de Imputação		MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM	MLP	SVM	KNN	SOM	OP-ELM
% MD	**5**	1913,509	3,654	**2,317**	10,368	5,499	5417,542	12,682	**3,487**	35,927	17,418	12276,762	40,941	**9,449**	106,111	51,149
	10	1922,943	3,184	**0,936**	9,014	4,676	5304,213	10,628	**2,887**	37,600	16,818	11994,834	29,561	**7,118**	110,305	47,790
	20	2032,452	2,660	**0,762**	9,459	4,408	5587,189	6,914	**2,041**	40,947	15,335	14882,219	15,079	**4,329**	125,388	43,340
	30	2427,608	2,074	**0,631**	9,843	4,164	7336,365	3,989	**1,443**	46,955	14,503	15517,054	6,537	**2,862**	166,842	27,242
	50	3870,481	1,197	**0,437**	11,403	4,237	7463,413	1,389	**1,083**	80,447	6,082	15351,319	1,724	**1,443**	393,010	7,692
	70	3431,496	0,528	**0,261**	16,497	2,072	5216,231	**0,465**	0,556	190,524	2,086	4341,438	**0,933**	1,416	1624,853	3,148

Apêndice B

Testes Estatísticos

Tabela B.1: Comparação das diversas combinações de métodos de imputação e classificação. Entre parênteses apresentam-se os valores que irão ser usados no teste de Friedman. Os valores de *accuracy* são os obtidos quando foram inseridos 10% de MD

Imp.	Class.	*Datasets*											Rank Médio
		Pima	Indian	Iris	Banknote	Seeds	Laryngeal	Voice	Transfusion	Telugu	Red	Bupa	
MLP	KNN	74,00 (16)	69,54 (14,5)	92,67 (17,5)	98,35 (6)	91,61 (3)	71,38 (14)	78,03 (11)	77,46 (7)	82,63 (11)	61,17 (11)	60,07 (20)	11,91
	SVM	76,91 (1)	70,83 (6)	94,89 (1,5)	98,64 (1)	89,57 (19)	73,27 (1)	78,59 (5)	77,46 (7)	83,17 (4)	63,67 (6)	71,97 (1)	4,77
	RF	74,43 (13)	68,10 (20)	94,89 (1,5)	98,30 (8)	90,93 (9)	71,70 (10,5)	77,18 (17)	73,97 (17)	82,40 (13)	68,48 (1)	68,33 (8)	10,73
	MLP	75,30 (11)	70,98 (5)	94,00 (7)	97,86 (12)	90,25 (16)	72,64 (2,5)	75,92 (19)	77,68 (4,5)	79,35 (17)	58,04 (20)	61,41 (18)	12,00
SVM	KNN	73,52 (18)	69,40 (16)	91,56 (20)	98,40 (5)	91,27 (5,5)	72,26 (5)	78,31 (8,5)	77,14 (13)	83,59 (1)	60,19 (14)	62,43 (15)	11,00
	SVM	76,43 (7)	71,98 (2,5)	93,11 (14,5)	98,52 (2)	91,27 (5,5)	72,64 (2,5)	78,03 (11)	78,26 (1)	83,13 (5)	63,36 (7)	69,61 (4)	5,64
	RF	74,39 (14)	72,56 (1)	93,78 (10,5)	98,08 (10)	89,68 (18)	71,42 (13)	79,72 (2)	73,93 (18)	83,24 (3)	66,47 (4)	67,48 (10)	9,41
	MLP	75,91 (8,5)	69,25 (17)	93,11 (14,5)	97,43 (18)	89,52 (20)	69,15 (20)	75,63 (20)	77,68 (4,5)	79,24 (18)	58,58 (18)	66,02 (13)	15,59
KNN	KNN	73,96 (17)	69,54 (14,5)	92,89 (16)	98,33 (7)	90,63 (13)	71,79 (8,5)	77,61 (15,5)	77,01 (14)	83,47 (2)	60,69 (12)	62,72 (14)	12,14
	SVM	76,70 (4)	71,98 (2,5)	94,44 (4)	98,47 (3)	91,11 (7,5)	71,79 (8,5)	79,15 (3)	77,41 (9,5)	83,09 (6)	63,15 (8,5)	71,17 (2)	5,32
	RF	75,91 (8,5)	70,69 (8,5)	93,78 (10,5)	97,74 (13)	90,79 (11)	70,09 (18,5)	80,56 (1)	73,62 (19)	82,98 (7)	66,81 (3)	68,83 (6,5)	9,68
	MLP	76,57 (6)	70,69 (8,5)	94,44 (4)	97,60 (15)	92,06 (1,5)	70,66 (15)	77,89 (13)	76,96 (15)	78,24 (20)	59,10 (16)	66,41 (12)	11,45
SOM	KNN	73,43 (19)	70,26 (11)	92,44 (19)	97,52 (16)	90,79 (11)	72,26 (5)	78,31 (8,5)	77,41 (9,5)	81,87 (15)	59,96 (15)	61,94 (16)	13,18
	SVM	76,87 (2)	71,26 (4)	94,00 (7)	97,67 (14)	90,79 (11)	71,70 (10,5)	79,01 (4)	77,81 (2)	82,67 (10)	62,90 (10)	70,00 (3)	7,05
	RF	75,48 (10)	68,97 (18)	93,78 (10,5)	97,50 (17)	90,48 (14)	70,38 (16,5)	78,45 (6,5)	74,51 (16)	82,33 (14)	66,30 (5)	68,83 (6,5)	12,18
	MLP	76,61 (5)	70,69 (8,5)	94,44 (4)	96,72 (20)	91,59 (4)	71,51 (12)	77,61 (15,5)	77,23 (12)	78,74 (19)	58,50 (19)	61,46 (17)	12,36
OP-ELM	KNN	72,91 (20)	69,68 (12,5)	92,67 (17,5)	98,23 (9)	91,11 (7,5)	71,89 (7)	78,03 (11)	77,37 (11)	82,82 (8)	60,46 (13)	60,87 (19)	12,32
	SVM	76,74 (3)	70,69 (8,5)	94,00 (7)	98,45 (4)	90,00 (17)	72,26 (5)	78,45 (6,5)	77,77 (3)	82,44 (12)	63,15 (8,5)	69,32 (5)	7,23
	RF	74,22 (15)	68,82 (19)	93,33 (13)	97,91 (11)	90,32 (15)	70,09 (18,5)	77,75 (14)	73,17 (20)	82,71 (9)	67,35 (2)	67,86 (9)	13,23
	MLP	75,17 (12)	69,68 (12,5)	93,78 (10,5)	97,40 (19)	92,06 (1,5)	70,38 (16,5)	76,76 (18)	77,46 (7)	79,62 (16)	59,02 (17)	66,60 (11)	12,82

Tabela B.2: Comparação das diversas combinações de métodos de imputação e classificação. Entre parênteses apresentam-se os valores que irão ser usados no teste de Friedman. Os valores de *accuracy* são os obtidos quando foram inseridos 20% de MD

Imp.	Class.	*Datasets*											Rank Médio
		Pima	Indian	Iris	Banknote	Seeds	Laryngeal	Voice	Transfusion	Telugu	Red	Bupa	
MLP	KNN	72,78 (17,5)	68,68 (20)	91,56 (16)	96,21 (2)	90,25 (11)	71,70 (11)	78,03 (12,5)	76,03 (13)	78,40 (8)	59,64 (12)	62,50 (15)	12,55
	SVM	76,52 (2)	71,12 (4,5)	94,22 (1)	96,26 (1)	90,48 (6)	72,33 (6,5)	78,31 (10)	76,16 (12)	79,24 (3,5)	64,30 (6)	66,63 (6)	5,32
	RF	74,22 (13)	70,11 (12)	93,78 (3)	95,87 (5)	89,34 (15)	70,44 (14)	78,73 (5,5)	74,11 (19)	79,24 (3,5)	65,83 (2)	64,93 (9)	9,18
	MLP	76,65 (1)	71,12 (4,5)	93,56 (4,5)	95,39 (13)	89,12 (19)	72,33 (6,5)	76,62 (17,5)	76,61 (5)	75,38 (17)	58,66 (16)	64,08 (12)	10,55
SVM	KNN	73,04 (15)	70,55 (11)	90,89 (20)	96,12 (4)	91,27 (1)	74,06 (1)	78,31 (10)	75,89 (14,5)	78,85 (5)	58,60 (17)	62,43 (16)	10,41
	SVM	76,09 (3)	70,98 (6)	92,00 (15)	96,19 (3)	90,00 (13,5)	72,74 (2)	79,44 (3)	76,21 (11)	79,47 (2)	62,09 (9)	67,38 (4)	6,50
	RF	74,61 (10)	69,83 (14)	93,11 (7,5)	95,66 (7,5)	88,57 (20)	70,38 (15)	78,45 (8)	74,82 (16)	78,78 (6)	64,97 (5)	66,41 (7)	10,55
	MLP	74,26 (12)	70,83 (7,5)	92,22 (13,5)	95,66 (7,5)	89,21 (17)	68,96 (20)	75,92 (19)	76,70 (4)	75,92 (16)	58,02 (18)	64,27 (11)	13,23
KNN	KNN	72,78 (17,5)	70,69 (9,5)	91,33 (17)	95,36 (14)	90,95 (3)	72,55 (4)	78,03 (12,5)	75,89 (14,5)	78,13 (10,5)	59,58 (13)	62,33 (17)	12,05
	SVM	75,61 (5)	71,98 (3)	93,56 (4,5)	95,41 (12)	90,48 (6)	72,36 (5)	78,31 (10)	76,43 (7)	78,02 (12)	62,11 (8)	69,51 (1)	6,68
	RF	75,04 (8,5)	72,27 (1)	92,89 (10,5)	95,27 (15)	89,21 (17)	70,85 (13)	79,01 (4)	73,26 (20)	78,13 (10,5)	65,72 (3)	66,89 (5)	9,77
	MLP	75,43 (6)	72,13 (2)	92,89 (10,5)	95,44 (11)	91,11 (2)	69,62 (18)	76,62 (17,5)	76,96 (2,5)	74,43 (19)	57,91 (19)	61,55 (20)	11,59
SOM	KNN	72,65 (19)	69,54 (15,5)	91,11 (18,5)	93,79 (19)	90,32 (9)	72,26 (8)	77,46 (14)	76,38 (8)	76,95 (14)	58,98 (14)	61,65 (19)	14,36
	SVM	76,04 (4)	69,97 (13)	93,11 (7,5)	94,44 (17)	90,32 (9)	71,79 (10)	78,59 (7)	76,29 (9,5)	76,83 (15)	61,46 (10)	69,13 (2)	9,45
	RF	73,83 (14)	69,25 (17,5)	92,22 (13,5)	94,25 (18)	90,00 (13,5)	70,09 (16)	79,58 (2)	74,42 (17)	77,71 (13)	65,66 (4)	66,02 (8)	12,41
	MLP	72,04 (20)	69,25 (17,5)	91,11 (18,5)	92,94 (20)	90,32 (9)	69,53 (19)	76,90 (16)	77,05 (1)	73,66 (20)	57,72 (20)	63,40 (14)	15,91
OP-ELM	KNN	72,83 (16)	68,97 (19)	92,67 (12)	95,46 (10)	90,79 (4)	71,98 (9)	77,18 (15)	76,29 (9,5)	78,70 (7)	59,81 (11)	62,14 (18)	11,86
	SVM	75,09 (7)	70,83 (7,5)	94,00 (2)	95,85 (6)	90,16 (12)	72,64 (3)	78,73 (5,5)	76,56 (6)	79,77 (1)	63,19 (7)	67,67 (3)	5,45
	RF	74,52 (11)	70,69 (9,5)	93,11 (7,5)	95,56 (9)	89,21 (17)	69,72 (17)	80,00 (1)	74,29 (18)	78,21 (9)	66,10 (1)	64,76 (10)	10,00
	MLP	75,04 (8,5)	69,54 (15,5)	93,11 (7,5)	94,76 (16)	90,48 (6)	71,51 (12)	74,65 (20)	76,96 (2,5)	74,96 (18)	58,75 (15)	63,69 (13)	12,18

Tabela B.3: Comparação das diversas combinações de métodos de imputação e classificação. Entre parênteses apresentam-se os valores que irão ser usados no teste de Friedman. Os valores de *accuracy* são os obtidos quando foram inseridos 30% de MD

Imp.	Class.	*Datasets*											Rank Médio
		Pima	Indian	Iris	Banknote	Seeds	Laryngeal	Voice	Transfusion	Telugu	Red	Bupa	
MLP	KNN	71,09 (20)	70,11 (9)	90,44 (10)	91,80 (13)	90,25 (7,5)	72,64 (4)	76,20 (15)	75,89 (14)	73,02 (8,5)	59,43 (11)	60,44 (15)	11,55
	SVM	75,13 (3)	70,83 (5)	92,89 (1,5)	92,48 (6)	90,25 (7,5)	73,58 (1)	77,46 (10)	76,25 (8,5)	74,24 (1,5)	61,52 (6)	64,56 (2,5)	4,77
	RF	72,43 (15,5)	70,11 (9)	91,56 (5,5)	91,46 (14)	89,57 (13)	72,01 (8)	78,03 (6)	73,44 (18)	72,98 (10)	66,88 (1)	63,35 (6)	9,64
	MLP	73,22 (10)	70,83 (5)	90,67 (9)	91,41 (15)	88,21 (19)	70,13 (15)	76,06 (16)	76,56 (1)	69,58 (19)	55,95 (20)	62,38 (11)	12,73
SVM	KNN	72,00 (18)	69,97 (11)	87,11 (20)	92,60 (4)	89,84 (11)	72,64 (4)	76,34 (14)	75,85 (15)	73,55 (5)	58,10 (15)	61,36 (14)	11,91
	SVM	74,74 (4,5)	71,12 (2)	89,56 (16)	92,96 (1)	90,63 (3,5)	72,55 (6,5)	77,75 (7,5)	76,12 (11)	74,24 (1,5)	60,88 (10)	64,56 (2,5)	6,00
	RF	72,78 (12,5)	70,98 (3)	89,33 (17)	92,91 (2)	87,62 (20)	71,70 (9,5)	78,87 (3)	73,75 (17)	73,17 (6,5)	64,53 (4)	63,20 (7,5)	9,27
	MLP	74,70 (6,5)	69,83 (12,5)	88,00 (19)	92,01 (9)	88,89 (17,5)	68,68 (19)	73,80 (20)	76,21 (10)	70,61 (16)	57,37 (18)	63,11 (9)	14,23
KNN	KNN	72,43 (15,5)	70,11 (9)	89,78 (14,5)	91,97 (11)	91,11 (1,5)	71,51 (11)	75,92 (17,5)	76,38 (5,5)	73,02 (8,5)	57,75 (16)	61,46 (13)	11,18
	SVM	74,74 (4,5)	69,83 (12,5)	90,22 (11,5)	91,97 (11)	89,52 (14,5)	72,74 (2)	77,32 (11,5)	76,25 (8,5)	72,60 (12)	61,09 (8)	65,63 (1)	8,82
	RF	71,96 (19)	70,83 (5)	90,00 (13)	92,04 (8)	89,68 (12)	70,85 (14)	79,72 (2)	74,06 (16)	73,17 (6,5)	64,66 (3)	63,20 (7,5)	9,64
	MLP	73,96 (9)	69,25 (16,5)	91,33 (7)	91,31 (16)	90,48 (5,5)	68,87 (18)	74,93 (19)	76,38 (5,5)	70,08 (18)	58,20 (14)	60,10 (17)	13,23
SOM	KNN	72,78 (12,5)	69,25 (16,5)	88,22 (18)	90,32 (19)	89,37 (16)	71,70 (9,5)	76,90 (13)	76,43 (2,5)	70,65 (15)	58,39 (12)	59,51 (18,5)	13,86
	SVM	75,22 (1)	71,26 (1)	90,89 (8)	90,51 (17)	90,00 (10)	72,55 (6,5)	78,31 (5)	75,94 (13)	71,34 (14)	61,27 (7)	64,37 (4)	7,86
	RF	72,39 (17)	69,40 (14,5)	90,22 (11,5)	90,49 (18)	90,16 (9)	71,32 (12)	80,14 (1)	73,26 (19)	72,10 (13)	64,47 (5)	63,40 (5)	11,36
	MLP	74,70 (6,5)	69,40 (14,5)	92,22 (3)	89,59 (20)	90,48 (5,5)	69,53 (17)	77,61 (9)	76,43 (2,5)	68,28 (20)	57,72 (17)	59,03 (20)	12,27
OP-ELM	KNN	72,65 (14)	68,97 (19)	89,78 (14,5)	92,43 (7)	90,63 (3,5)	71,13 (13)	77,32 (11,5)	76,07 (12)	72,79 (11)	58,35 (13)	59,51 (18,5)	12,45
	SVM	75,17 (2)	70,69 (7)	92,89 (1,5)	92,55 (5)	91,11 (1,5)	72,64 (4)	77,75 (7,5)	76,38 (5,5)	74,05 (3)	60,96 (9)	62,43 (10)	5,09
	RF	72,91 (11)	68,68 (20)	91,78 (4)	92,72 (3)	88,89 (17,5)	69,62 (16)	78,59 (4)	73,13 (20)	73,89 (4)	64,78 (2)	61,84 (12)	10,32
	MLP	74,13 (8)	69,11 (18)	91,56 (5,5)	91,97 (11)	89,52 (14,5)	68,58 (20)	75,92 (17,5)	76,38 (5,5)	70,23 (17)	57,27 (19)	60,39 (16)	13,82

Tabela B.4: Comparação das diversas combinações de métodos de imputação e classificação. Entre parênteses apresentam-se os valores que irão ser usados no teste de Friedman. Os valores de *accuracy* são os obtidos quando foram inseridos 50% de MD

Imp.	Class.	*Datasets*											Rank
		Pima	Indian	Iris	Banknote	Seeds	Laryngeal	Voice	Transfusion	Telugu	Red	Bupa	Médio
MLP	KNN	69,91 (16)	69,11 (17)	84,44 (12)	82,62 (10)	88,89 (10)	73,58 (3)	73,80 (16,5)	75,76 (10)	61,07 (14)	58,94 (11)	55,58 (20)	12,68
	SVM	72,39 (2,5)	71,12 (2,5)	86,89 (4)	83,69 (2)	87,30 (19)	74,53 (1)	75,77 (9)	76,25 (5,5)	65,11 (2)	60,68 (6)	58,74 (15)	6,23
	RF	70,96 (10)	68,10 (20)	87,78 (1)	82,86 (7)	87,76 (17,5)	73,90 (2)	77,89 (5)	73,48 (18)	65,15 (1)	65,48 (1)	58,62 (18)	9,14
	MLP	70,78 (12)	70,69 (5)	85,56 (7,5)	81,21 (14)	87,76 (17,5)	71,70 (9)	73,80 (16,5)	76,29 (4)	58,97 (18)	58,59 (12)	59,47 (7)	11,14
SVM	KNN	69,43 (18)	69,25 (16)	81,56 (19)	83,23 (4)	89,68 (7,5)	71,60 (10)	74,65 (13,5)	75,45 (14,5)	63,70 (7)	56,47 (20)	60,39 (4)	12,14
	SVM	72,39 (2,5)	70,98 (4)	84,67 (11)	83,81 (1)	89,84 (5,5)	71,23 (12)	76,90 (6)	76,52 (1,5)	65,08 (3)	60,17 (7)	60,78 (3)	5,14
	RF	70,61 (14)	71,12 (2,5)	86,67 (5)	82,82 (9)	87,78 (16)	70,57 (16)	78,17 (4)	73,75 (16)	63,59 (8)	62,90 (3)	59,22 (10)	9,41
	MLP	71,78 (5)	70,55 (6,5)	79,11 (20)	83,01 (6)	90,79 (1)	69,25 (19)	73,52 (18)	76,52 (1,5)	62,40 (11)	56,66 (17)	59,42 (8)	10,27
KNN	KNN	69,30 (19)	70,55 (6,5)	84,00 (13,5)	80,92 (15)	89,68 (7,5)	71,98 (6)	74,65 (13,5)	75,45 (14,5)	60,53 (16)	56,58 (19)	58,64 (16,5)	13,36
	SVM	72,87 (1)	71,84 (1)	81,78 (17,5)	80,85 (16)	90,63 (2)	72,26 (5)	74,93 (12)	76,12 (7)	62,06 (12)	59,44 (10)	61,46 (1)	7,68
	RF	71,17 (8,5)	70,40 (8)	82,22 (16)	80,70 (17)	83,97 (20)	71,42 (11)	79,86 (1)	72,68 (20)	61,41 (13)	61,09 (5)	61,26 (2)	11,05
	MLP	70,87 (11)	70,26 (10)	83,33 (15)	80,56 (18)	87,94 (15)	69,53 (18)	73,38 (19)	75,67 (12)	58,70 (19)	56,64 (18)	58,64 (16,5)	15,59
SOM	KNN	69,48 (17)	69,54 (13)	81,78 (17,5)	81,48 (13)	90,00 (4)	72,55 (4)	75,07 (10,5)	75,71 (11)	59,62 (17)	56,93 (16)	59,90 (6)	11,73
	SVM	71,70 (6)	69,83 (12)	84,00 (13,5)	82,45 (11)	90,16 (3)	71,89 (7)	76,34 (7,5)	76,03 (8)	63,09 (10)	59,62 (9)	59,03 (12,5)	9,05
	RF	71,17 (8,5)	69,40 (14,5)	85,11 (9,5)	83,54 (3)	88,10 (14)	70,94 (14)	79,15 (2)	73,35 (19)	64,01 (5)	62,69 (4)	60,19 (5)	8,95
	MLP	71,52 (7)	69,40 (14,5)	85,78 (6)	80,07 (19)	89,21 (9)	70,28 (17)	74,23 (15)	75,94 (9)	57,37 (20)	57,66 (13)	58,83 (14)	13,05
OP-ELM	KNN	68,96 (20)	70,26 (10)	85,56 (7,5)	82,43 (12)	89,84 (5,5)	71,13 (13)	76,34 (7,5)	75,63 (13)	63,47 (9)	57,18 (14)	59,03 (12,5)	11,27
	SVM	71,83 (4)	68,82 (18)	87,56 (2)	83,06 (5)	88,25 (12,5)	71,79 (8)	75,07 (10,5)	76,47 (3)	64,58 (4)	59,81 (8)	59,13 (11)	7,82
	RF	70,70 (13)	68,68 (19)	87,33 (3)	82,84 (8)	88,25 (12,5)	70,75 (15)	78,73 (3)	73,62 (17)	63,97 (6)	63,36 (2)	59,32 (9)	9,77
	MLP	70,22 (15)	70,26 (10)	85,11 (9,5)	78,88 (20)	88,73 (11)	68,02 (20)	72,96 (20)	76,25 (5,5)	60,92 (15)	57,10 (15)	57,57 (19)	14,55

Tabela B.5: Comparação das diversas combinações de métodos de imputação e classificação. Entre parênteses apresentam-se os valores que irão ser usados no teste de Friedman. Os valores de *accuracy* são os obtidos quando foram inseridos 70% de MD

Imp.	Class.	*Datasets*											Rank
		Pima	Indian	Iris	Banknote	Seeds	Laryngeal	Voice	Transfusion	Telugu	Red	Bupa	Médio
MLP	KNN	67,78 (16)	68,68 (18,5)	77,33 (15)	68,64 (13)	89,12 (4)	72,33 (4)	73,94 (12)	75,09 (15)	48,47 (17)	54,91 (19)	56,55 (18)	13,77
	SVM	68,30 (12)	70,55 (6,5)	79,11 (8,5)	71,46 (4)	88,66 (8)	71,70 (8)	75,35 (10)	75,54 (11)	55,15 (6)	57,20 (8)	59,34 (9)	8,27
	RF	68,04 (13)	68,39 (20)	79,78 (3)	68,83 (11)	86,17 (16)	69,81 (12,5)	80,14 (1)	74,51 (17)	57,02 (2)	62,84 (2)	57,77 (16)	10,32
	MLP	67,96 (15)	71,98 (1)	75,78 (16)	68,45 (14)	87,30 (14)	72,64 (2)	73,24 (15)	75,94 (6)	49,77 (11)	54,98 (18)	56,67 (17)	11,73
SVM	KNN	66,35 (20)	70,98 (3,5)	78,67 (12)	70,90 (5)	89,68 (2)	71,98 (7)	74,37 (11)	75,40 (12)	49,31 (14)	56,60 (11,5)	60,39 (4)	9,27
	SVM	71,04 (1)	70,11 (10)	78,89 (11)	72,31 (1)	88,57 (9)	72,17 (5,5)	76,20 (8)	75,98 (5)	55,46 (5)	59,44 (6)	62,14 (1)	5,68
	RF	68,91 (9)	69,11 (15,5)	79,56 (4,5)	71,92 (2,5)	87,78 (13)	71,32 (9)	79,15 (3)	73,93 (19)	57,60 (1)	63,09 (1)	58,54 (12)	8,14
	MLP	70,22 (3,5)	69,54 (13,5)	77,56 (13,5)	68,79 (12)	88,89 (6,5)	69,72 (14,5)	71,97 (18)	76,70 (1)	51,15 (9)	56,45 (13)	58,74 (10,5)	10,45
KNN	KNN	66,48 (19)	70,26 (9)	67,78 (18)	66,43 (19)	73,33 (19)	58,87 (19)	73,52 (14)	75,71 (9,5)	44,24 (20)	54,22 (20)	58,25 (13,5)	16,36
	SVM	69,43 (7)	70,69 (5)	69,33 (17)	65,87 (20)	74,44 (18)	61,70 (18)	76,48 (6,5)	76,07 (4)	48,21 (18)	56,41 (14)	58,25 (13,5)	12,82
	RF	67,35 (17)	69,68 (12)	66,22 (19)	67,77 (16)	76,51 (17)	69,81 (12,5)	77,89 (5)	75,31 (13)	48,85 (15)	61,02 (5)	60,10 (5)	12,41
	MLP	68,00 (14)	70,40 (8)	62,89 (20)	66,55 (18)	72,54 (20)	57,64 (20)	70,99 (20)	76,61 (2)	45,57 (19)	56,37 (15)	51,17 (20)	16,00
SOM	KNN	68,35 (10,5)	68,82 (17)	77,56 (13,5)	69,17 (10)	90,79 (1)	72,45 (3)	73,66 (13)	75,27 (14)	49,43 (13)	56,70 (10)	60,58 (3)	9,82
	SVM	70,39 (2)	70,98 (3,5)	79,33 (6)	70,10 (7)	89,37 (3)	73,49 (1)	76,48 (6,5)	76,25 (3)	56,18 (3)	58,94 (7)	62,04 (2)	4,00
	RF	69,87 (5)	69,97 (11)	79,11 (8,5)	71,92 (2,5)	88,10 (12)	70,94 (11)	79,30 (2)	73,35 (20)	56,11 (4)	62,51 (3)	60,00 (6)	7,73
	MLP	70,22 (3,5)	68,68 (18,5)	79,11 (8,5)	67,96 (15)	89,05 (5)	66,23 (17)	72,25 (16)	75,80 (8)	50,00 (10)	55,87 (16)	56,02 (19)	12,41
OP-ELM	KNN	66,87 (18)	70,55 (6,5)	79,11 (8,5)	69,32 (9)	88,89 (6,5)	72,17 (5,5)	71,69 (19)	74,69 (16)	48,59 (16)	55,20 (17)	59,90 (7)	11,73
	SVM	69,30 (8)	71,41 (2)	79,56 (4,5)	69,76 (8)	88,41 (10,5)	71,23 (10)	75,49 (9)	75,71 (9,5)	55,08 (8)	56,95 (9)	58,74 (10,5)	8,09
	RF	68,35 (10,5)	69,11 (15,5)	80,22 (1)	70,83 (6)	87,14 (15)	69,72 (14,5)	78,45 (4)	74,11 (18)	55,11 (7)	62,38 (4)	59,81 (8)	9,41
	MLP	69,74 (6)	69,54 (13,5)	80,00 (2)	67,14 (17)	88,41 (10,5)	69,06 (16)	72,11 (17)	75,89 (7)	49,66 (12)	56,60 (11,5)	57,86 (15)	11,59