U.PORTO

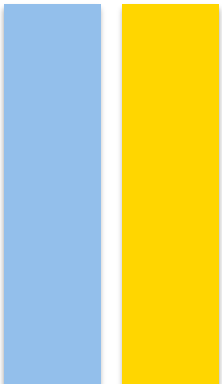
2º CICLO DE ESTUDOS
MESTRADO EM INFORMÁTICA MÉDICA

Aplicação de técnicas de *data mining* para prever a mortalidade e o período de estadia hospitalares no contexto do enfarte agudo do miocárdio

Ana Raquel Santos Oliveira

M

2020



U.PORTO

2º CICLO DE ESTUDOS
MESTRADO EM INFORMÁTICA MÉDICA

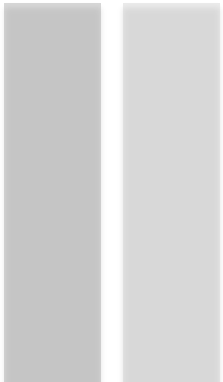
Aplicação de técnicas de *data mining* para prever a mortalidade e o período de estadia hospitalares no contexto do enfarte agudo do miocárdio

Ana Raquel Santos Oliveira

Orientadores:
Professor Doutor Alberto Freitas
Doutor Júlio Souza

M

2020



Agradecimentos

Em primeiro lugar gostaria de agradecer à minha família por todo o apoio prestado ao longo do meu percurso académico. Aos meus pais agradeço todas as palavras de conforto e a incessável dedicação à minha evolução académica.

Ao Professor Doutor Alberto Freitas presto breves e sentidas palavras de agradecimento pelas oportunidades de desenvolvimento académico e pela sua dedicação ao ensino nesta área com grande potencial de expansão científica.

O presente trabalho tornou-se um grande desafio em termos de aprendizagem, tendo sido agilizado pelo incansável Doutor Júlio Souza, a quem agradeço profundamente por todo o conhecimento transmitido, pela paciência em responder a todas as dúvidas, e no fundo por estar sempre disposto a ajudar como se do seu próprio trabalho se tratasse.

Entre tantos estudos de sucesso realizados por investigadores igualmente bem sucedidos e inspiradores da Faculdade de Medicina da Universidade do Porto, tive o prazer de conhecer a Doutora Mariana Lobo e os seus trabalhos. Agradeço-lhe pela sua generosidade e voto de confiança ao ter fornecido os dados que tornaram possível esta dissertação.

A toda a equipa docente que está por detrás da organização do Mestrado em Informática Médica agradeço singelamente pela transmissão de conhecimentos e partilha de experiências que enriqueceram estes dois anos letivos de uma forma ímpar.

Presto o meu humilde agradecimento a toda a equipa técnica, sempre disposta a ajudar e a tornar esta passagem mais confortável com a preparação dos *coffe breaks*.

Ao longo da frequência deste ciclo de estudos foram várias as pessoas que me acompanharam, o trabalho em equipa e interajuda foram fulcrais para cada etapa, não podendo deixar de lhes agradecer.

Findando, todas as palavras serão poucas face ao quão agradecida estou a todos os que se cruzaram comigo de um modo tão enobrecedor.

Abstract

Introduction: Acute myocardial infarction (AMI) is a common, fatal, and onerous disease. Because it is a common cause of both mortality and long lengths of hospital stay, its initial evaluation can be crucial for patient management. Considering that multiple data mining techniques have shown to be efficient in the identification of patterns amongst data and, presumably, in the prediction of events, the association of these two premises resulted in the main objective of the present study: the construction of models to predict mortality and length of hospital stay (LOS) using data from AMI patients.

Methods: The employed data originated in another study regarding AMI hospitalizations were subjected to pre-processing, namely class balancing techniques due to the low number of deaths compared to survival. For the construction of the mortality models, several data mining, logistic regression, random forest, support vector machine, K-nearest neighbors, Naive Bayes, and classification and regression tree techniques were applied. Considering the possibility of selecting the most relevant variables, a feature selection technique was tested. Concerning the construction of LOS prediction models, the Poisson regression and negative binomial model techniques were applied to data previously processed and subjected to different feature selection approaches.

Results: Generally, the logistic regression technique allowed for a better model of mortality prediction, with performance nearing 80% (area under the receiver operating characteristic curve). Concerning the LOS models, the negative binomial was the best at data modeling. However, the prediction capacity was the same in both models.

Discussion: Upon analysis of the obtained data, it can be inferred that the feature selection technique did not significatively modify the mortality prediction results. Class balancing, however, was critical. Results from other studies were good albeit hard to compare due to sample heterogeneity and the considered samples. In regard to the LOS models, the negative binomial model presented the best results but led to a 10-day prediction error which could be related to data quality, namely the absence of variables of critical importance to such prediction. Besides that, the

considered variables only relate to admissions, and others might arise during hospitalization.

Conclusion: The present dissertation presents models that can be improved in performance through the acquisition of more homogeneous data or the test of other automated learning techniques. Given the common interest of this theme, its investigation is progressively suggested.

Keywords: acute myocardial infarction, data mining, prediction, mortality, length of stay.

Resumo

Introdução – O enfarte agudo do miocárdio é uma doença comum, fatal e onerosa. Sendo uma causa comum de mortalidade e longos períodos de estadia hospitalares, a sua avaliação inicial pode ser crucial para a gestão do doente. Uma vez que variadas técnicas de aprendizagem automática se têm demonstrado eficazes na descoberta de padrões entre os dados, e presumivelmente na previsão de eventos, a associação destas duas premissas resultou no objetivo principal deste estudo, a construção de modelos de previsão da mortalidade e período de estadia hospitalares utilizando dados de doentes com enfarte agudo do miocárdio.

Metodologia – Os dados utilizados provieram de outro estudo no contexto das hospitalizações por enfarte agudo do miocárdio, tendo sido sujeitos a um pré-processamento, nomeadamente a técnicas de balanceamento das classes devido ao pequeno súmero de mortes face ao de sobrevivências. Na construção dos modelos de mortalidade utilizaram-se diferentes técnicas de *data mining,* regressão logística, *random forest, support vector machine*, *K-nearest neighbors,* Naive Bayes e classification and regression tree. Tendo em vista a possibilidade de selecionar as variáveis mais relevantes testou-se uma técnica de *feature selection*. Quanto à elaboração dos modelos de previsão do período de estadia utilizaram-se as técnicas de regressão de Poisson e o modelo da distribuição binomial negativa para os dados previamente processados e sujeitos a diferentes técnicas de *feature selection*.

Resultados – No geral, a técnica de regressão logística permitiu obter o melhor modelo de previsão da mortalidade, com uma performance perto de 80 % (área sob a curva característica de operação do recetor). No que diz respeito aos modelos de previsão do período de estadia, o modelo da binomial negativa modelou melhor os dados, contudo, a capacidade de previsão foi igual em ambos os modelos.

Discussão – Ao analisar os resultados obtidos, pode ser inferido que a técnica de *feature selection* no contexto da previsão da mortalidade não modificou significativamente os resultados, contudo o balanceamento das classes foi crítico. Os resultados obtidos por outros estudos foram bons no entanto dificeis de

comparar dada a heterogeneidade da amostra e variáveis consideradas. Quanto ao modelo de previsão do período de estadia o modelo com melhores resultados foi o modelo da binomial negativa, contudo resultou num erro de previsão de 10 dias que pode estar relacionado com a qualidade dos dados, nomeadamente a ausência de variáveis que poderiam ser críticas nesta previsão, além disso as variáveis consideradas são apenas na admissão, podendo aparecer outras ao longo do internamento.

Conclusão – A presente dissertação apresenta modelos que podem ser melhorados a nível de performace através da obtenção de dados mais homogéneos, ou do teste de outras técnicas de aprendizagem automática. Dado o interesse comum deste tema, a sua investigação é progressivamente sugerida.

Palavras-chave: enfarte agudo do miocárdio, *data mining*, previsão, mortalidade e período de estadia.

Preâmbulo

Quando no Mestrado de Informática Médica me foi abordado o tema de extração de conhecimento a partir de bases de dados, tendo a possibilidade de gerar conhecimento clinicamente útil utilizando regras estatísticas, não tive dúvidas no caminho a seguir.

Rapidamente entendi a dificuldade enfrentada pelos clínicos ao serem confrontados com imensa informação, não tendo por vezes a capacidade de gerir os cuidados prestados aos doentes de uma forma personalizada. Esse obstáculo estaria presente de uma forma mais gravosa em situações urgentes, como a condição de enfarte agudo do miocárdio.

O Professor Doutor Alberto Freitas sugeriu um tema muito semelhante ao que pretendia, e agilmente me apresentou a um grupo de trabalho sugerindo a utilização de uma base de dados utilizada por esse grupo num estudo anterior, no âmbito do enfarte.

A partir daí, juntei-me ao grupo de trabalho envolvido nesta temática, com quem aprendi bastante ao longo deste ano letivo, tendo tido oportunidade de cooperar na realização de um estudo, "*Protocol for Analysis of Root Causes of Problems Affecting the Quality of the Diagnosis Related Group-Based Hospital Data: A Rapid Review and Delphi Process*".

Índice

Lista de tabelas

Lista de figuras

Acrónimos

ACSS – Administração Central do Sistema de Saúde (ACSS);

ACTION – Registo Global de Eventos Coronários Agudos;

AUC – Área sob a Curva ROC;

AVC – Acidente Vascular Cerebral;

CART – Classification And Regression Trees;

IA – Inteligência Artificial;

ICD – International Classification of Diseases/Classificação Internacional de Doenças;

DPOC – Doença Pulmonar Obstrutiva Crônica;

EAM – Acute Myocardial Infartaction ou Enfarte Agudo do Miocárdio;

ECD – Extração de Conhecimento de Dados;

ECDS – Extração de Conhecimento de Dados em Saúde;

GRACE - registo global de eventos coronários agudos;

KNN – K Nearest Neighbors;

LR – *Logistic regression* ou Regressão Logística;

NB – Naive Bayes;

NUTSII – Nomenclatura Comum das Unidades Territoriais Estatísticas;

TIMI – Trombólise em Enfarte do Miocárdio;

RF – Random Forest;

RFE – Recursive Feature Elimination;

ROC – Receiver Operating Curve/Curva característica de Operação do recetor;

ROSE – Random Over-Sampling Examples;

SMOTE – Synthetic Minority Over-sampling Technique;

SVM – Support Vector Machine/Máquinas de Vetores de Suporte.

Introdução

1. Introdução

O enfarte agudo do miocárdio (EAM) é definido como diminuição ou interrupção do fluxo sanguíneo para uma parte do coração, o que pode levar à necrose do tecido devido à escassez de oxigénio. Esta doença é uma das causas mais comuns de morte e/ou internamento hospitalar, sendo o número de episódios de EAM 32,4 milhões por ano em todo o mundo. Apesar das recentes reduções das taxas de mortalidade hospitalar, o EAM continua a ser a principal causa de morte cardiovascular na Europa, sendo ainda assim uma prioridade alcançar novas melhorias. A mortalidade hospitalar a 30 dias é considerada um indicador importante de qualidade hospitalar, ao nível da atuação médica quanto aos procedimentos e intervenções. Sublinha-se que a mortalidade por EAM é influenciada não só pela qualidade dos cuidados hospitalares, mas também pelas diferenças relativamente às transferências hospitalares, severidade do EAM e período de estadia médio.

O reconhecimento e entendimento das características do paciente são críticos para prever a mortalidade e outras medidas de resultado nos episódios de EAM, dando ao clínico informação para determinar o prognóstico do doente, guiar o seu tratamento e seguimento. Se o doente tiver um enfarte anterior ou apresentar fatores de risco como hipertensão, diabetes, hiperlipemia ou falha de rim, a probabilidade de morte aumenta. Assim como, as comorbilidades e outras condições de saúde, os dados demográficos também contribuem para a deteção do aumento do risco de morte, nomeadamente idade, género, e fatores socioeconómicos. Por último, a conjuntura da estrutura hospitalar pode conjuntamente contribuir para o prognóstico de AMI.

A inteligência artificial compreende diversos conceitos e subáreas, como *data mining*, aprendizagem automática (*machine learning*), probabilidade e estatística. Estas técnicas envolvem tipicamente a descoberta de padrões entre os dados de forma a descrever ou aplicar funções de aprendizagem para prever resultados de interesse. O processo de *data mining* pode ajudar na previsão de eventos futuros com base em dados clínicos. Como observado na literatura, a maioria dos modelos tem bons resultados, potenciando uma possível associação entre a informática e a medicina.

As técnicas de *data mining* podem ser úteis na construção de modelos preditivos de mortalidade e período de estadia de doentes com EAM. A existência destes modelos preditivos pode ajudar os profissionais de saúde a programar e gerir recursos hospitalares de modo garantir melhor assistência ao doente.

A principal motivação do presente trabalho consiste em prever com exatidão o prognóstico do EAM, o que poderia guiar a triagem dos cuidados a serem prestados ao doente e auxiliar no processo de tomada de decisão informada. A hipótese é que a utilização de técnicas de *data mining* contemporâneas, nomeadamente aprendizagem automática, pode melhorar a estratificação de risco no contexto do EAM e identificar as relações complexas entre os vários potenciais preditores e variáveis de desfecho. O objetivo principal deste estudo é construir modelos de previsão da mortalidade e período de estadia hospitalares de doentes com EAM através da utilização de técnicas de *data mining*. Desta forma, serão exploradas, avalidas e comparadas diferentes técnicas de aprendizagem automática e análises de regressão nos dados clínicos de doentes diagnosticados com EAM, construindo modelos que possam vir a ser relevantes para estabelecer o prognóstico e o consumo de recursos (sendo o período de estadia um *proxy* deste último fator). Adicionalmente, a performance dos modelos poderá ser comparada com os já existentes na literatura no contexto do AMI.

Fundamentação teórica

2. Fundamentação teórica

2.1 Enfarte agudo do miocárdio

O EAM, é definido em termos leigos como ataque cardíaco, mais frequentemente causado pela diminuição ou interrupção do fluxo sanguíneo para uma parte do coração, levando à necrose do músculo cardíaco. Por norma, este acontecimento resulta de um coágulo sanguíneo na artéria epicárdica que fornece a região afetada. Contudo, é agora reconhecido que nem todos os casos se devem necessariamente a um coágulo. Uma vez que o equilíbrio entre a necessidade e o suprimento de oxigénio deve ser mantido, em casos de desequilíbrio como uma frequência cardíaca muito rápida (suprimento excessivo) ou uma queda da pressão arterial (suprimento insuficiente) podem levar a danos no miocárdio sem a presença de um coágulo. [1]

Os pacientes com enfarte prévio são um grupo de risco para novos enfartes, e nesses novos enfartes têm um risco aumentado de mortalidade.[2] Assim como doentes que apresentem fatores de risco como hipertensão, diabetes, hiperlipemia, insuficiência renal, doença pulmonar obstrutiva crónica (DPOC), distúrbios do sono (síndrome de apneia obstrutiva e central), cancro antecedente, pneumonia, ataque vascular-cerebral (AVC) prévio, entre outros.[3,4,5,6] Mas, os dados demográficos, como a idade do paciente, interferem no risco de mortalidade, sendo o risco de mortalidade tanto maior quanto maior for a idade do paciente, assim como o género, os homens são propícios a desencadear enfartes numa idade mais precoce.[7]

O EAM é uma doença comum, fatal e onerosa. Apesar dos declínios significativos na mortalidade durante a última década, devido ao aparecimento de novas tecnologias na área da saúde, este continua a ser um considerável "fardo" social.[8]

Esta patologia está entre as causas comuns de mortalidade e de um longo período de estadia hospitalares.[9,10] Sendo que globalmente, existem cerca de 32.4 milhões de enfartes por ano. Em Portugal, a mortalidade em 30 dias após a admissão por EAM em 2017 foi de 7.3 por 100 pacientes com mais de 45 anos ligeiramente abaixo da média dos países da Organização para a Cooperação e

Desenvolvimento Econômico (OECD) de 7.6 por 100 pacientes com mais de 45 anos, 30 dias após a admissão por EAM.[11]

O reconhecimento e compreensão das características dos doentes para a previsão da mortalidade e período de estadia perante um EAM fornecem aos médicos informações valiosas para determinar o prognóstico, orientar o tratamento e monitorizar o doente.[12]

Existem na literatura inúmeros estudos que investigaram o desenvolvimento de modelos de risco em diferentes áreas clínicas, sendo que a maioria dos modelos de estratificação de risco no contexto do EAM utilizam informações demográficas e clínicas que existem antes da hospitalização, para além das características clínicas apresentadas na admissão, para estimar o risco de mortalidade, e período de estadia do paciente.[13]

2.2 Registo clínico

Entende-se por registo clínico um conjunto de documentos que contém toda a informação clínica de um paciente, ou seja, anotações realizadas por médicos e outros profissionais de saúde.

As principais funções deste registo são o acesso rápido aos dados do doente, a anotação contínua dos problemas de saúde, a frequência de consultas, e o registo de medidas de ações preventivas. Os dados contemplam a história clínica do doente, exame(s) físico(s), diagnósticos, e tratamentos efetuados, sendo, assim, complementados por testes laboratoriais e relatórios de meios complementares de diagnóstico.

Os dados clínicos constantes no prontuário podem ser organizados de várias formas, pela sua ordem cronológica, pela sua origem ou pelo tipo de problema, sendo classificados os registos clínicos como orientados cronologicamente, quanto à fonte, ou quanto ao problema, respetivamente.[14]

Os propósitos dos registos clínicos são variados, podendo: auxiliar na identificação de casos para estudos retrospetivos e ensaios clínicos; melhorar a gestão e manutenção de registos de pacientes; auxiliar na investigação, avaliação e planeamento dos serviços de saúde; documentar a eficácia das terapias em ambientes reais; gerir fatores de risco para resultados adversos; fornecer informações sobre a natureza da doença e os benefícios dos tratamentos em

subgrupos de pacientes; e fornecer relatórios comprovativos de referência às partes interessadas, incluindo hospitais, equipa clínica, pacientes, consumidores, financiadores e seguradoras.[15]

Apesar de serem diversos os propósitos de um registo de saúde, existe um único objetivo comum, promover a aplicação das ciências da saúde de forma a que melhorem o bem-estar dos doentes, incluindo a condução de investigação e atividades de saúde pública que abordem a saúde populacional.[16]

Dada a importância destes registos, a qualidade dos mesmos deve ser assegurada, por isso existem indicadores da qualidade como, a exatidão, completude e adequação dos dados.[15,17]

A qualidade pode prender-se com o tipo de registo, dado que este pode ser feito em formato de papel ou em suporte eletrónico.[18]

2.2.1 Registo clínico em papel

O registo clínico em papel é um arquivo de informações em formato de papel, suporte físico que pode ser acedido apenas num local e possivelmente por não mais que uma pessoa em simultâneo. Este tipo de registo permite que haja flexibilidade no que é registado e não necessita de treino de adaptação.

Contudo, este formato livre acarreta certos prejuízos de qualidade, como a ilegibilidade da caligrafia do profissional de saúde, falta de estruturação, duplicação da informação e podem ser facilmente destruídos em caso de catástrofe natural.[16]

Dadas as desvantagens significativas do registo clínico em papel, surge preferencialmente o registo clínico eletrónico.[16]

2.2.2 Registo clínico eletrónico

Um registo clínico eletrónico corresponde a um repositório de informações sobre o estado e cuidados de saúde de um indivíduo, armazenadas eletronicamente de forma a que possam servir vários usos e utilizadores legítimos. As informações constantes neste repositório eletrónico são contínuas,

isto é, abrangem a história clínica do paciente desde o seu nascimento até à sua morte.

As vantagens dos registos eletrónicos em relação aos registos em papel são várias, nomeadamente:

- Formato interativo de inserção de dados;
- Imposição de captura de dados.
- Possibilidade de aplicação de medidas de validade à medida que os dados são inseridos;
- Apresentação dos dados em vários formatos;
- Possibilidade de redução de erros tipográficos através de menus de entrada restritivos e revisão ortográfica;
- Incorporação de informações de multimédia, como imagens e vídeos;
- Acessibilidade em vários locais e por vários utilizadores;
- Possibilidade da realização de várias cópias de segurança;

Através da enumeração das vantagens, sabe-se que o conteúdo dos registos clínicos eletrónicos é legível e bastante organizado.

Contudo, dos benefícios destes registos podem advir alguns custos, tais como a compressividade da informação, e o elevado grau de estruturação do sistema de registo que pode restringir os dados inseridos.

Embora que poucas, o registo eletrónico tem algumas desvantagens, nomeadamente, o grande investimento inicial comparativamente ao suporte de papel, devido aos custos de *hardware*, *software*, treino e suporte; e a adequação do fluxo de trabalho dos profissionais de saúde ao sistema de registo.

Apesar do elevado custo e do tempo de habituação despendido, os sistemas de registo clínico eletrónico são uma ferramenta importante na assistência clínica, reguladora e de negócio da prática da medicina.[16]

Sendo que as informações contidas nos registos de pacientes são sistematicamente abstraídas, codificadas e agrupadas em grupos relacionados ao diagnostico, gerando bases de dados recetivas a diferentes usos.[19]

2.3 Dados secundários em saúde

Define-se dados secundários em saúde como dados gerados por grandes instituições de saúde, como parte dos registos guardados organizacionalmente. A informação pode ser extraída de variados arquivos de dados, nomeadamente registos clínicos eletrónicos.[20]

2.3.1 Codificação de dados clínicos

Segundo a Administração Central do Sistema de Saúde (ACSS), a codificação clínica traduz-se na codificação, mediante nomenclaturas e sistemas de codificação adequados, dos procedimentos, diagnósticos e atos que caracterizam o contacto do utente com o hospital, seja em internamento, hospital dia, consultas externas ou urgências.[21] Sendo a base dessa codificação os registos clínicos realizados pelos profissionais de saúde.[22]

Os codificadores são responsáveis por traduzir as informações contidas nos registos clínicos em códigos adequados, e para que este processo seja conseguido tem de existir um processo de abstração, isto é, a seleção da informação útil à codificação.[23]

No que diz respeito à codificação de diagnósticos e procedimentos realizados em internamento, ou ambulatório é utilizado um sistema de classificação, a Classificação Internacional de Doenças (ICD).[21]

2.3.2 Qualidade dos dados

Os registos clínicos são a base da codificação clínica, portanto a qualidade das informações registadas afeta a qualidade dos dados codificados. Não sendo os códigos clínicos bem atribuídos caso os registos sejam inadequados. Admite-se que características como a coerência, precisão, integridade e consistência da informação sejam regras a seguir por quem regista auxiliando quem codifica.

Dado que muitas vezes as regras de registo não são cumpridas, a codificação clínica enfrenta alguns obstáculos, nomeadamente, variações na descrição de diagnóstico pelos médicos, falta de clareza nos registos, documentação incompleta, uso de sinónimos e abreviações, ou falta de comunicação entre os profissionais de saúde.[19]

Neste contexto, aquando o processo de codificação torna-se importante a retificação de problemas de integridade, particularmente a normalização de valores, identificação de *missing values* e de informação redundante, entre outros processos de gestão da base de dados.[24]

Não só a gestão de bases de dados é uma importante medida de qualidade, como também a utilidade dos dados. Sendo a qualidade dos dados tanto maior quanto a sua utilidade, o que coloca em evidência o ponto de vista dos utilizadores da informação e a sua perceção da assertividade das informações.[25]

Particularmente, as bases de dados médicas reforçam o facto da qualidade dos dados depender do seu tipo de uso, uma vez que estas podem servir para análise económica, mas serem insuficientes para análise clínica ou epidemiológica.[26]

Apesar do conhecimento das mais variadas medidas de qualidade, estas tornam-se difíceis de tomar e controlar em grandes unidades de saúde como os hospitais.[26]

2.3.3 Finalidades dos dados

> Sendo a codificação clínica em ICD uma linguagem comum da mortalidade e da morbilidade, esta permite ao mundo comparar e compartilhar informações, como tendências e estatísticas de saúde.

Estes dados permitem monitorizar a incidência e a prevalência de doenças, o reembolso hospitalar, a alocação de recursos, e diretrizes de segurança e qualidade.[27]

Deste modo, as doenças, desordens, danos e outras condições relacionadas com a saúde estão listadas numa base de dados compreensiva, permitindo um fácil armazenamento, recuperação e análise, compartilhando e comparando informações entre os hospitais, regiões e países, para além de permitir a comparação de dados na mesma localidade durante certos períodos de tempo. Estas informações tornam-se preciosas na área da investigação científica, pois são retiradas da realidade rotineira.

2.3.3.1 Investigação

O aumento da disponibilidade de dados administrativos potencia o número de estudos a utilizar dados secundários em saúde.[28] As vantagens são evidentes, pois a amostra tem um grande tamanho, heterogeneidade e cobertura populacional, permitindo que as investigações reflitam o "mundo real".[29] Outras vantagens inerentes são a ausência de outros gastos na reunião dos dados, períodos de longa observação, e a possibilidade de ligar várias bases de dados com informações do paciente (tabela 2).[30,31]

Contudo, existem uma série de desvantagens, a duvidosa qualidade dos dados, a ausência de informação de interesse para o fim investigacional, a dificuldade sentida ao retirar conclusões causais devido à presença de vieses e a possível classificação errada do diagnóstico ou exposição (tabela 2).[32]

Tabela 2 *Vantagens e desvantagens do uso de dados secundários em saúde*

Vantagens	
	Cobertura populacional
	Amostra de elevado tamanho
	Heterogeneidade populacional
	Longos períodos de observação
	Informação atualizada
	Sem custos adicionais na recolha de dados
	Possibilidade de ligar várias fontes de informação
Desvantagens	
	Necessidade de experiência específica na base de dados para garantir o uso correto dos dados
	Não são concebidos para investigações em saúde, não tendo informações clínicas específicas
	A qualidade dos dados é afetada pelo uso administrativo, por exemplo para reembolso
	Seleção dos casos com base em códigos de diagnóstico
	São necessários estudos de validação
	Possibilidade de classificação incorreta de resultados ou exposição
	Dificuldades para controlar fatores de exposição e estabelecer relações causais
	O significado estatístico é facilmente alcançado, a relevância clínica deve ser levada em consideração na análise estatística e na discussão dos resultados

2.4 Extração de conhecimento a partir de dados em saúde

No contexto investigacional, surge o conceito de extração de conhecimento de dados (ECD), que de entre várias definições na literatura, foi definido por Mitchell (1997) como:

'A capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência.'

Cada vez mais, a abstração de informação dos dados em saúde oferece a oportunidade de aprender algoritmicamente o conhecimento médico, através do reconhecimento de padrões a partir dos dados observados.[33] Sendo que, através de recursos computacionais, é realizada uma aprendizagem a partir de experiência passada. O princípio aplicado nesta aprendizagem é o da inferência, em que é formulada uma hipótese ou induzido um modelo que seja capaz de tirar conclusões generalizáveis a outros contextos, a partir de um conjunto de exemplos dados, e que possa ser aplicado a outro conjunto dados, sendo obtido um determinado resultado descritivo ou preditivo.

Esta extração de conhecimento pode servir para fins epidemiológicos, de previsão de medidas de resultado, de avaliação da qualidade dos cuidados prestados ou de estudo de eventos adversos devido a fármacos.[31]

No contexto da previsão de medidas de resultado, apesar da facilidade de acesso e recolha de dados, existem aspetos a considerar nesta conjuntura, nomeadamente, o grande tamanho da amostra, uma vez que as associações entre variáveis de resultado e exposições ou características do paciente são facilmente consideradas estatisticamente significativas, mesmo que o significado clínico não o seja.[34]

A ECDS engloba a preparação de dados (descrição, análise estatística e pré-processamento dos dados) e as tarefas de aprendizagem, sendo que esta sequência de etapas foi também cosniderada na realização do presente trabalho.[35]

2.4.1 Preparação dos dados

Uma investigação deve começar pela avaliação da qualidade dos dados por parte do investigador, sendo a sua qualidade tanto maior quanto a sua usabilidade.

No que se refere à descrição e análise estatística, estas permitem a descoberta de padrões e tendências que podem fornecer informações importantes no processo que gera os dados. Nesta etapa de descrição e exploração dos dados, são estimadas estatísticas simples de dispersão e sumário, como a frequência, média, mediana, âmbito interquartil, desvio padrão, entre outras. Se oportuno, também,

poderão ser estudadas relações entre as diferentes variáveis (análises bivariadas).[35]

De forma a tornar o conjunto de dados mais apropriados à construção de algoritmos de ECDS, são frequentemente utilizadas técnicas de pré-processamento tais como:

- Eliminação manual de atributos;
- Integração de dados;
- Amostragem de dados;
- Balanceamento de dados;
- Limpeza de dados ;
- Transformação de dados;
- Redução da dimensionalidade (*feature selection*).

A eliminação manual de atributos serve para eliminar atributos que não são necessários ou irrelevantes para análise, num determinado contexto de estudo, por exemplo, variaveis de identificaçãode um doente (por exemplo o nome e número de identificação). Normalmente esta etapa está dependente na contextualização do problema em questão, tendo em conta a opinião de especialistas, a revisão da literatura existente, hipóteses formuladas e o conhecimento do domínio clínico e dos próprios dados.

A integração de dados refere-se à obtenção de uma estrutura uniforme dos dados a partir de diferentes bases de dados, sendo por vezes necessário identificar os casos (objetos a representar a unidade de análise, como o paciente) presentes nas diferentes bases de dados a serem integradas (utilizando por exemplo a identificação do doente).

A amostragem de dados implica que certos atributos e/ou objetos não sejam utilizados na sua totalidade, pois pode ser mais eficiente utilizar apenas parte do conjunto de dados original. Os algoritmos de extração de conhecimento podem ter dificuldade em lidar com grandes quantidades de dados, contudo a amostra selecionada deve representar o conjunto de dados originais. A amostragem pode ser realizada de diferentes formas: a amostragem aleatória simples (com ou sem reposição), amostragem estratificada (em relação à classe, manter o mesmo número de casos para cada classe, ou manter a mesma proporção de casos do

conjunto original), e a amostragem progressiva (aumentar progressivamente o tamanho da amostra, enquanto a taxa de acerto continua a melhorar).

Quanto ao balanceamento de dados, é oportuno quando o conjunto de dados reais apresenta um número de casos muito discrepante para as diferentes classes (por exemplo, em 10% dos casos a doença está presente, enquanto que em 90% a doença está ausente). Para corrigir esta discrepância existem técnicas de balanceamento de dados, como a redefinição do tamanho do conjunto de dados (acréscimo de casos à classe minoritária ou com a eliminação de casos da classe maioritária; contudo na primeira situação poderá ocorrer *overfitting*, e na segunda *underfitting*), a utilização de diferentes custos de classificação para as diferentes classes (possível dificuldade em definir custos e de os incorporar em alguns algoritmos), e a indução de um modelo para uma classe (aprendizagem em separado para cada classe).

Não menos importante é a limpeza dos dados, este processo prende-se com problemas de qualidade de dados, dados com ruído (valores diferentes do esperado – *outliers*), dados inconsistentes (contradizem valores de outros atributos do mesmo caso), dados redundantes (quando dois ou mais objetos têm os mesmos valores para todos os atributos), e dados incompletos (ausência de valores).

Outra técnica é a transformação dos dados, tendo em conta que algumas técnicas algorítmicas exigem uma consistência de valores simbólicos, apenas, ou valores numéricos, isto é converter variáveis categóricas em numéricas e vice-versa, embora que por vezes tal não seja possível. Para além disto, estudos de associação ou resultados de modelos preditores construidos através de algumas técnicas tendem a ser mais robustos se as variáveis dependentes e independentes estiverem normalmente distribuídas, sendo por vezes necessário recorrer à transformações dos dados para obter a normalização dos mesmos.

Por último, quando as bases de dados apresentam um elevado número de atributos, a capacidade de algumas técnicas algorítmicas fica reduzida, para além de potencialmente aumentar a redundância e multicolinearidade entre as variáveis preditoras. Neste contexto, surgem técnicas de áreas como o reconhecimento de padrões, estatística e teoria da informação que poderão ser utilizadas na redução do número de atributos, sugerindo abordagens de

agregação e seleção dos atributos que explicam a maior parte da variabilidade dos dados, processo de redução da dimensionalidade conhecido do inglês como "*feature selection*".[35]

2.4.2 Inteligência artificial

A inteligência artificial (IA) visa a habilitar os sistemas computacionais a agirem e pensarem de forma inteligente. O objetivo científico desta área é entender quais os princípios que permitem um conhecimento inteligente. O propósito da engenharia da IA é a criação de sistemas computacionais capazes de resolver problemas do mundo real tão ou mais eficazes e eficientes que o ser humano. A metodologia por detrás destas premissas é diversa, como uma caixa de ferramentas.

2.4.2.1 Aprendizagem automática

A aprendizagem automática é um subdomínio da IA que se dedica à resolução de tarefas específicas. O termo aprendizagem automática refere-se à aprendizagem a partir dos dados. A aprendizagem surge de um computador procurar relações entre os dados através de interseções estatísticas, através do uso de diferentes algoritmos. Esta interface entre a matemática e a ciência dos computadores é atualmente impulsionada pelo maior poder computacional adquirido, o que permite construir modelos estatísticos a partir de um conjunto massivo de dados. Este processo pode ser classificado em supervisionado e não supervisionado.

A aprendizagem supervisionada trata-se de uma previsão de um output (classe ou *outcome* de interesse) conhecido. Deste modo, os algoritmos desta categoria constroem um modelo matemático que contém os *inputs* (variáveis preditoras) e *outputs* desejados. O papel do cientista incide no ensinamento ao algoritmo das conclusões que este deve apresentar.

A aprendizagem não supervisionada trata-se de uma previsão de um output desconhecido. Ou seja, não existem outputs a prever, e o algoritmo procura padrões ou agrupamentos que ocorrem naturalmente nos dados. O desempenho

destes modelos é avaliado por tarefas de aprendizagem supervisionadas posteriores, de modo a verificar a utilidade dos novos padrões, se existirem.[36]

Na conjuntura da ECDS, as aprendizagens supervisionada e não supervisionada são classificadas como preditivas e descritivas, respetivamente. Na aprendizagem preditiva, o supervisor externo pode avaliar a capacidade de a hipótese induzida predizer o valor de saída para novos exemplos. Em tarefas de descrição, o objetivo consiste na exploração ou descrição de um conjunto de dados, por exemplo descrever um agrupamento de dados, encontrando casos semelhantes ou encontrar regras de associação que relacionam grupos de atributos.[35]

Estes modelos estatísticos podem ser úteis onde o sucesso humano pode ser limitado. No caso da aprendizagem supervisionada, um exemplo comum é a interpretação automatizada de eletrocardiogramas, já no caso da aprendizagem não supervisionada, há uma inclinação crescente para observar padrões fisiopatológicos que poderão estar por detrás de uma doença.[37,38] São variados os exemplos em que ambas as categorias de aprendizagem computacional podem intervir na medicina, desde o desenvolvimento de modelos de risco à redefinição de classes de pacientes. No entanto, existem limitações para a utilização de aprendizagem automática de larga escala em contextos clínicos. Alguns obstáculos podem estar relacionados com questões paradigmáticas envolvidas neste setor, como a responsabilidade da garantia do bem-estar dos pacientes, para além da baixa interpretabilidade e transparência de alguns algoritmos de aprendizagem supervisionada (os chamados algoritmos *black-box*). Deste modo, são criadas alternativas em que os médicos podem interagir com os sistemas aumentando a precisão e produtividade, reduzindo os custos.[37]

2.4.2.2 Data mining

O *data mining* pode ser definido como um processo que visa gerar conhecimento através dos dados. O conhecimento gerado neste contexto pode ser obtido através da descoberta de padrões, relações e tendências novas e não triviais num conjunto de dados. O processo de *data mining* envolve o recolhimento e seleção de dados, o pré-processamento dos dados, incluindo a visualização e interpretação dos resultados e a aplicação de conhecimento.

De forma a pré-processar e analisar os dados, os métodos de aprendizagem automática e estatísticos são implementados no contexto do *data mining*, fazendo parte dele. Os resultados deste processo podem ser distinguidos em descritivos, o conhecimento é representado na forma de modelos que representam padrões e relações no conjunto de dados, ou em preditivos, sendo o conhecimento apresentado em modelos de previsão de condições futuras, tendências e relações.

Problema clínico

3. Problema clínico

A previsão do prognóstico de um paciente e estratificação de risco são passos essenciais para definir o tratamento e tomar decições apropiradas no contexto do EAM.[39] No entanto, vários sistemas de pontuação de risco convencionais e amplamente aceites, como trombólise em enfarte do miocárdio (TIMI), o registo global de eventos coronários agudos (GRACE), e o tratamento agudo coronário e rede de resultados de intervenção (ACTION), têm limitações importantes na prática, tendo em conta que são modelos que foram desenvolvidos há mais de 2 décadas, além de utilizarem apenas variáveis selecionadas com base em técnicas estatísticas convencionais, resultando possivelmente em perda de informação relevante.[40]

De modo a averiguar que desenvolvimentos existem na literatura acerca desta temática, tornou-se oportuno construir uma pergunta de investigação (*query*). Com a leitura de estudos semelhantes, será possível tomar conhecimento das técnicas mais utilizadas e das suas vantagens e limitações, o que poderá ser uma ferramenta útil na realização deste estudo e/ou na sua análise final.

Query: *((("data mining" OR "machine learning" OR "regression") AND ("predict*")) AND ("mortality" OR "death" OR "length of stay" OR "hospital stay")) AND ("Acute myocardial infarction")*

Esta *query* foi inserida nas bases de dados *scopus* e *pubmed* com a seleção dos estudos de 2010 a 2020. Os critérios de seleção basearam-se na semelhança ao tema da presente dissertação, tendo sido selecionados apenas os que contemplassem a:

- Área do EAM especificamente;
- Construção de modelos de previsão;
- Previsão da mortalidade hospitalar ou previsão do período de estadia hospitalar;

Sendo que na *pubmed* foram encontrados 700 resultados dos quais foram selecionados 8 estudos de interesse, e na *scopus* obtiveram-se 771 resultados dos quais se selecionaram 9 estudos, tendo sido 8 dos quais selecionados na *pubmed.*

3.1 Estado da arte

Vários estudos têm vindo a demonstrar que o *data mining* pode vir a influenciar positivamente a prática clínica. A nível hospitalar, entre as variáveis de resultado mais procuradas estão a mortalidade e o período de estadia, no contexto do EAM estas medidas são consideravelmente valiosas num determinado momento, dada a conjuntura da doença.

Chenxi Song et al. construíram uma ferramenta de previsão da mortalidade hospitalar através dos registos de EAM chineses de Janeiro de 2013 a Setembro de 2014. Com a análise dos dados foram selecionadas 16 variáveis, tendo sido feito um modelo de risco através da técnica de regressão logística (LR), também foi realizado uma coorte de validação para validar o risco atribuído a cada variável. Os investigadores concluíram que a capacidade de discriminação e calibração do modelo são elevados, podendo ser útil aos clínicos para prever a mortalidade com precisão e optimizar a gestão dos cuidados ao doente. Contudo foram apontadas algumas limitações, como a indicação de validação com uma coorte separada numa escala maior, assim como o facto de este modelo poder não ser transversal a outras etnias, dado que os dados são apenas de população chinesa.[41]

Benjamin Goldstein et al. reconheceram a necessidade de testar técnicas de *machine learning* na previsão da mortalidade após o diagnóstico de EAM. Na construção dos modelos de previsão, foram utilizadas técnicas de regressão como a logística, seleção *forward*, LASSO, *ridge*, PCR e modelo aditivo generalizado, além destas também foram testadas técnicas baseadas em árvores de decisão, como Classification And Regression Tree (CART), *random forest* (RF) e *Boosting,* e outras como*, K-Nearest Neighbours* (K-NN) e *Neural Networks.* Nenhum dos modelos obteve a área sobre a curva ROC superior a 80%, tendo sido o melhor a técnica de Boosting a criar o melhor modelo. De um modo conclusivo, os autores assumem que em dados homogéneos, onde é possível construir um modelo linear, as técnicas de *machine learning* não são as eleitas. Estas técnicas mais robustas

também não têm recursos para lidar com dados temporais, ou para reconhecer uma relação de causalidade (um preditor poderá simplesmente ser um marcador útil), tal não é um problema de interpretação de prognostico, mas sim da causa. Contudo, é assumido que quando o objetivo é gerar um modelo que preveja precisamente a medida de resultado, e quando existem vários preditores relacionados ou relações não lineares, devem ser eleitas estas técnicas.[42]

Robert L. McNamara et al. desenvolveram um modelo baseado na técnica de LR para prever a mortalidade hospitalar de doentes com EAM. No entanto, as limitações são várias, como a fonte de dados ser uma base de dados voluntária, mais informação clínica também poderia enriquecer o modelo, assim como uma extensão do período de observação ser alargado para mortalidade a 30 dias e não apenas hospitalar, entre outras.[43]

Yu Qi et al. criaram uma ferramenta de previsão do risco de mortalidade através de dados de mulheres com EAM provenientes de bases de dados da avaliação centrada paciente de eventos cardíacos (China PEACE). A pertinência deste estudo advém das diferenças entre homens e mulheres na condução do seu quadro clínico. A LR *backward stepwise* foi utilizada para examinar a relação entre as variáveis e a mortalidade hospitalar. *Wami score* destaca-se em relação a outros modelos de previsão do risco de mortalidade. Contudo, existem limitações apontadas, referem-se à qualidade dos dados utilizados, viés de seleção das pacientes, análises laboratoriais duvidosas quanto à sua utilização e pacientes com morte muito precoce podem estar sub-representados, entre outras.[44]

Jonn-myoung Kwon et al. desenvolveram um modelo de estratificação do risco de mortalidade de doentes com EAM baseado em *deep learning*. Os dados utilizados provieram dos registos do grupo de trabalho do EAM coreano. Para além de ter sido testado o *deep learning*, também foram testadas a LR e RF como meio de comparação. O modelo de *deep learning* revelou-se excelente para predizer o prognóstico, denotando-se entre as outras técnicas. Contudo estão presentes limitações, como a falta de interpretação dos modelos, assim como a representabilidade dos dados (ao contrário do conhecimento médico, o *deep learning* usa apenas a relação entre variáveis), tendo o modelo de ser validado noutro conjunto de dados.[40]

Salman desenvolveu modelos de previsão da mortalidade utilizando dados de pacientes de dois países diferentes, sendo que cerca de 89% sobreviveram ao EAM (os dados encontram-se tendenciosamente na sobrevivência). O autor utilizou técnicas como LR, árvores de decisão e *Naive Bayes* (NB). O modelo com maior poder de discriminação foi a árvore aumentada de NB. Contudo este tipo de técnica pode não saber lidar com dados incompletos e é reconhecido pelos autores existirem muitos *missings* nos dados.[45]

Chee Tang Chin et al. ao obterem dados de pacientes com EAM, criaram um modelo de LR para prever a mortalidade hospitalar. O poder preditivo obtido foi elevado. Contudo, o estudo apresenta limitações, como o facto de os dados utilizados não contemplarem todos os hospitais, podendo haver deste modo uma padronização; os pacientes transferidos não foram analisados; ao serem consideradas apenas as mortes hospitalares, poderá haver um enviesamento por não haver uma análise mais longa, e algumas variáveis que poderiam ser importantes não foram consideradas.[46]

Zhi Qu et al. começaram por calcular os pesos de cada variável, através da técnica de LR, usando deste modo a função de probabilidade construída para prever a probabilidade de morte. O objetivo do estudo de conseguir um bom modelo de previsão do risco para uma possível aplicação no momento de admissão hospitalar foi atingido, embora que os diagnósticos secundários contemplados na base de dados utilizada para a construção do modelo tenham a possibilidade de ter sido descobertos ao longo da hospitalização e não na admissão.[47]

Em Portugal foi realizado um estudo por Magalhães *et al.* que contempla um modelo de previsão do período de estadia através da técnica de LR. O modelo foi construído baseado nos dados de 755 doentes, de entre as variáveis destaca-se o período de estadia extenso. Os resultados foram positivos, uma boa habilidade discriminativa e calibração, contudo foram detetados alguns problemas na base de dados, sendo a qualidade do registo e codificação colocada em dúvida.[48]

Conclui-se que os estudos que existem até ao momento no que diz respeito a modelos de previsão da mortalidade se baseiam maioritariamente na técnica de LR, contudo, também se verificou a utilização de outras técnicas de *machine learning*. O presente trabalho pretende estudar a construção de modelos de

previsão de uma forma conjunta, podendo comparar várias técnicas aprendizagem supervisionada, criando algo que não foi encontrado na literatura no contexto do enfarte agudo do miocárdio.

Com a aplicação da *query* acima citada, não foram obtidos artigos que contemplassem a previsão do período de estadia na conjuntura específica do EAM, com a exceção de um estudo português que utilizou a técnica de LR. Nesse sentido foi realizada uma pesquisa arbitrária de estudos fora do contexto do enfarte, obtendo assim uma perceção melhorada das técnicas utilizadas para prever o período de estadia no campo da medicina em geral.

3.2 Objetivos

O objetivo major desta dissertação é a construção de modelos de previsão do risco de mortalidade e período de estadia através de técnicas de *data mining.*

À medida que o estudo avança vão sendo concretizados objetivos intermédios inerentes ao objetivo final nomeadamente, a compreensão das técnicas utilizadas, o seu modo de avaliação e reprodutibilidade no contexto dos dados, assim como, de um modo conclusivo, o reconhecimento de limitações do estudo e a perspetivação para trabalhos futuros.

O principal propósito destes modelos de previsão é agilizar a intervenção médica de forma a evitar a morte e ajudar a gerir os processos de cuidados ao nível hospitalar. Frequentemente, os médicos não têm ferramentas suficientes para delinear cuidados personalizados, mas alguns modelos apresentam uma boa *performance* na previsão de medidas de resultado com base em fatores relevantes. Com este estudo, pretende-se resolver esta lacuna através do desenvolvimento de modelos que possam prever a mortalidade hospitalar e tempo de internamento entre os doentes com EAM, explorando métodos de *data mining* para encontrar a relação entre distintos fatores em ambas as medidas de resultado.

Metodologia

4. Metodologia

4.1 Compreensão dos dados

4.1.1 Fonte de dados

Para a elaboração deste estudo, foi utilizada uma base de dados de morbilidade hospitalar. Esta base de dados inclui características demográficas dos pacientes, período de estadia hospitar, tipo de admissão, destino após a alta, e códigos de diagnósticos e procedimentos (codificados em ICD-9-CM). Este conjunto de dados está claramente definido, e cobre a maioria das hospitalizações (Tabela 1).[49]

Tabela 1 *Conteúdo da base de dados de morbilidade hospitalar.*

Dados administrativos		
	Dados de admissão	Dados do paciente
	• Código do hospital • Código de admissão e data • Tipo de admissão • Departamentos e transferências • Tipo de destino após a alta e data	• Data de nascimento • Género • Cidadania • Residência principal • Entidade pagadora
Dados médicos		
	• Diagnóstico principal (ICD-9) • Diagnósticos secundários (ICD-9) • Quaisquer serviços médicos do catálogo de procedimentos	

A base de dados para o presente estudo foi fornecida pela ACSS e compreende todas as hospitalizações públicas ocorridas entre Janeiro de 2012 e 31 de Dezembro de 2015 em Portugal continental.[50]

Os hospitais públicos representam 79% de todas as admissões hospitalares e 52% de todos os hospitais nacionais, podendo os seus dados representar o panorama nacional.[51,52,53]

A proteção de identidade dos pacientes foi garantida por um identificador único e anónimo constante numa variável da base.[50]

Para além dos conteúdos apresentados na tabela 1, também foram obtidas informações acerca dos hospitais responsáveis por cada episódio, como o número de camas, área geográfica, presença ou não de unidade hemodinâmica e cirurgia cardiotorácica, estado de ensino (hospital universitário) e volume de doentes com AMI.[50]

O número de camas foi disponibilizado pela ACSS e/ou nos *websites* dos hospitais, referente a 2010/2011 (ou no ano mais próximo).[54]

A informação relativa à classificação geográfica dos hospitais foi obtida através do *Eurostat*. A localização geográfica foi atribuída com base no segundo nível da Nomenclatura das Unidades Territoriais da Estatística (NUTSII), que divide Portugal continental em 5 regiões (Norte, Centro, Lisboa, Alentejo e Algarve).[55] A região do Algarve foi combinada com a do Alentejo, porque havia apenas um centro hospitalar nessa região e o Alentejo faz fronteira com a região a sul de Portugal.

O Programa nacional das doenças cardiovasculares e cerebrais, assim como a comunicação direta com os hospitais, permitiu a confirmação da presença ou não de unidade hemodinâmica e cirurgia cardiotorácica no período em que os dados em estudo foram construídos.[56]

O estado de ensino foi determinado de acordo com a Federação Nacional Médica.[57]

Quanto às possíveis fusões hospitalares, estas foram identificadas através dos documentos oficiais de leis relacionadas. Durante o período do estudo ocorreram duas fusões hospitalares. Nesses casos, foi considerado o centro hospitalar durante todo o período, portanto as suas características resultam da combinação das características individuais do hospital (por exemplo, o tamanho corresponde à soma dos leitos dos hospitais incluídos no centro hospitalar).[50]

O volume de EAM foi estimado a partir dos dados de pacientes internados em 2011.[50]

Reunindo todas as informações, foi possível construir uma base de dados robusta, que contemplasse informações que pudessem vir a ser úteis não só no estudo de *Lobo et al. (2020)*[50], mas também noutras investigações pertinentes, como a presente dissertação.

Tendo em vista o contexto do EAM, este último conjunto de dados foi sujeito a critérios de seleção de episódios, pacientes e as suas características, assim como características hospitalares e medidas de resultado.[50]

4.1.2 Definição de episódio

A definição de episódio entende-se como hospitalização constante na base de dados, porém estes episódios podem estar relacionados através de transferências que possam ter ocorrido dentro ou fora do mesmo episódio de atendimento, tendo tais sido identificadas. Estas hospitalizações partilhavam o mesmo código de identificação do paciente, mas também foram identificados os seguintes aspetos: 1) as datas de internamento e alta de um episódio foram incluídas nas datas de internamento e alta do outro episódio; 2) a data de alta hospitalar inicial e a data do internamento subsequente diferiram em 1 dia ou menos, existindo identificação do hospital de e para onde o doente foi transferido; podendo neste último, a data de admissão hospitalar subsequente ter ocorrido no mesmo dia que a alta de hospitalização inicial.

No caso 1), o hospital no qual o paciente teve maior tempo de permanência foi considerado responsável pela admissão e alta, consumando-se num único episódio.

No caso 2), foi, também, considerado um único episódio sendo definido o destino do paciente após a alta no último hospital.[50]

4.1.3 Seleção dos pacientes: critérios de inclusão e exclusão

Na seleção dos episódios da base de dados inicial, foram incluídos todos os pacientes hospitalizados (com idade igual ou superior a 20 anos) de todos os hospitais públicos do país do ano de 2012 ao de 2015.

A partir desta base de dados foram selecionados todos os pacientes com diagnóstico principal de IAM.

No caso de episódios com várias hospitalizações, foi definido um episódio de enfarte se o diagnóstico principal das hospitalizações iniciais e finais fosse de EAM, tendo sido para esse efeito selecionados os códigos de ICD9-CM 410.xx,

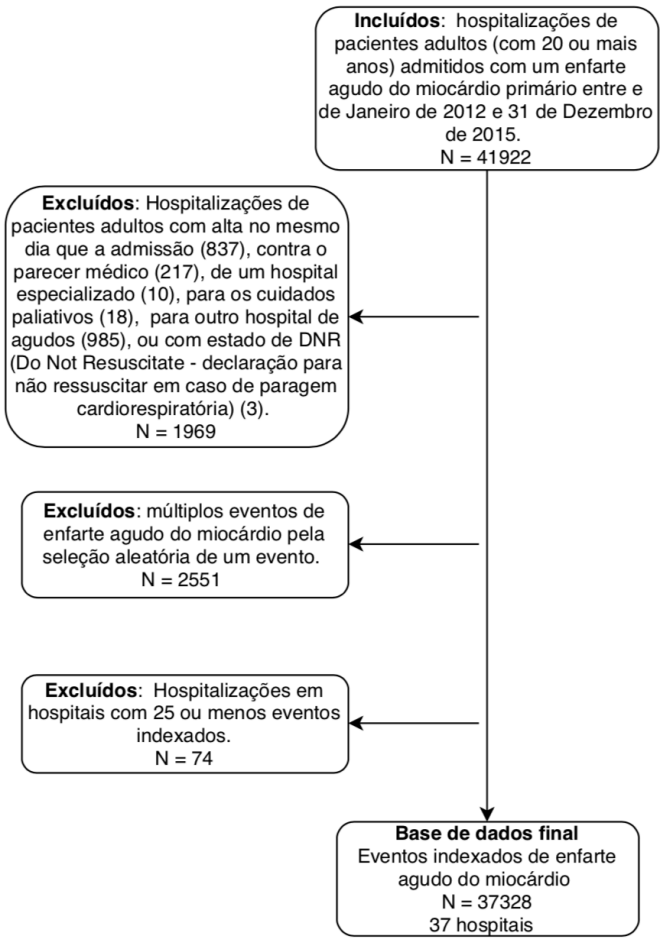
excepto se no quinto digito surgisse um 2. No caso das transferências, o hospital responsável pelo episódio é o que admite o paciente e não o que lhe confere a alta.

Foram excluídas hospitalizações de pacientes que receberam alta no mesmo dia em que foram admitidos, que saíram contra parecer médico ou que receberam alta de um hospital especializado (oncológico, psiquiátrico ou maternidade) por ser pouco provável que representem um EAM. Os episódios de doentes que receberam alta hospitalar para cuidados paliativos também foram excluídos, uma vez que têm maior probabilidade de morrer no tempo imediato após a alta.

Mesmo depois da combinação de todas as hospitalizações num único episódio, poucos internamentos permaneceram na condição de transferência para outros cuidados agudos. Por não ser fácil lidar com problemas de identificação de pacientes, pode acontecer que, por exemplo o hospital de admissão transfira temporariamente o paciente para outro hospital para se sujeitar a um procedimento de revascularização retornando ao hospital inicial, e ambos os hospitais podem ser documentados como estado de transferência, não se sabendo assim o que aconteceu efetivamente ao doente, se este morreu ou não, por exemplo. Sendo assim, para evitar a introdução de potenciais vieses, foram excluídos todos os episódios com destino após a alta de transferência.

Além dos critérios de exclusão anteriores, excluíram-se pacientes admitidos em hospitais com 25 ou menos episódios. Os critérios de inclusão e exclusão encontram-se sumariados no diagrama 1.[50]

Figura 1Esquema de obtenção da base de dados final. Adaptado de Lobo et. al (2020).



Em concordância com o que está descrito acima (figura 1), de modo a assegurar a qualidade dos dados, estes foram sujeitos a uma limpeza, tendo sido realizados testes de integridade (pré-processamento dos dados). Uma vez que os dados deste estudo são provenientes de *Lobo et al* (2020), boa parte da preparação dos dados foi realizada por este grupo, o que inclui a identificação de *outliers,* dados inconsistentes, redundantes e incompletos. Neste contexto, episódios com data de admissão após a data de alta, ou hospitalizações pertencentes ao mesmo paciente com um estado de morte na hospitalização anterior, ou pacientes com género desconhecido foram excluídos, tendo sido contabilizados 37328 pacientes.[50]

4.1.4 Características dos pacientes e do hospital

Após a obtenção da base de dados final fornecida pelo grupo de investigação de *Lobo et al.* (2020) procedeu-se à análise das variáveis, averiguando sua conformidade com o objetivo de produzir modelos de previsão.

A caracterização dos pacientes deve estar em concordância com as medidas de resultado, neste contexto a mortalidade e o período de estadia, sendo que quanto mais severo for o caso, maior probabilidade de morte e elevado período de estadia terá. As variáveis de interesse tornam-se assim informações que contribuem para a severidade de cada caso, desde características demográficas, ou hospitalares (qualidade de atendimento), a dados clínicos. Nesse sentido, foram selecionadas as informações demográficas dos pacientes, fatores de risco como diabetes ou demência, entre outras comorbilidades, informações recolhidas ao longo de 1 ano antes da hospitalização e no próprio episósido (tabela 2). As variáveis incluídas no estudo estão alinhadas com estudos que averiguaram o impacto das variáveis nas medidas de resultado.[58,59] Os diagnósticos secundários foram também convertidos num índice de Charlson em cada episódio, conceito explicado adiante.

Tabela 2 *Variáveis utilizadas na construção dos modelos de previsão de mortalidade e período de estadia.*

Dados demográficos do paciente	• Idade • Género • Concatenação das variáveis distrito e concelho • Identificação do doente • Tipo de admissão (urgente ou não)
Dados hospitalares	• Estado de aprendizagem • Número de camas • Região do hospital de admissão • Volume de doentes admitidos com enfarte agudo do miocárdio em 2011 • Unidade hemodinâmica (presente ou ausente) • Cirurgia cardiotorácica (possível ou não)
Dados médicos	• Diagnóstico principal de enfarte agudo do miocárdio • Fatores de risco - História de angioplastia coronária transluminal percutânea - História de enxerto de *bypass* da artéria coronária

	- Outra localização de enfarte do miocárdio - Enfarte do miocárdio anterior - Hipertensão - Choque e falha cardiorrespiratória - Trauma no último ano - Doença crónica de fígado - Doenças psiquiátricas major - Desnutrição - Demência ou outras desordens cerebrais - Cancro metastizado leucemia aguda ou outros cancros severos - Pneumonia - Doença pulmonar obstrutiva crónica - Falha renal - Diabetes - Desabilidade funcional, paralisia, hemiplegia, paraplegia - Doença e complicações vasculares - Ataque cardíaco - Doença cerebrovascular - Doença reumática cardíaca ou valvular - Formas subagudas de doença isquémica do coração - Angina ou aterosclerose coronária - Falha congestiva do coração • Índice de Charlson • Sobrevivência ou morte • Período de estadia

Os procedimentos médicos não foram considerados por ter sido assumido não contribuírem significativamente para prever as medidas de resultado no momento da admissão, uma vez que são identificados no próprio episódio e não contemplados na história clínica.

Sumariando, o conjunto de variáveis contempla a informação histórica de cada paciente 1 ano antes da hospitalização e a informação adquirida no início do próprio episódio de internamento, podendo assim ser realizada uma previsão no momento da admissão.

4.1.5 Medidas de resultado

As medidas de resultado constantes na base de dados são a mortalidade intra-hospitalar (obtida através da variável de destino após a alta) e o período de estadia. A mortalidade foi fornecida diretamente pela base de dados, o período de estadia (tempo de internamento) foi calculado através da diferença entre as datas de admissão e de alta hospitalar.

4.2 Processamento dos dados

As etapas de pré-processamento que incluem a identificação e tratamento de *outliers,* dados inconsistentes, redundantes e incompletos foram realizadas no âmbito do estudo conduzido por *Lobo et al. (2020).*[50]

O risco de mortalidade e período de estadia estão estreitamente relacionados com diagnósticos de outras doenças pré ou coexistentes, sendo que técnicas que permitam contabilizar a carga de doença, como o Índice de *Charlson,* podem contribuir com uma nova e poderosa variável para a construção dos modelos objetivados.

Dada a grande discrepância entre o número de óbitos e o número de vivos, e sabendo que tal pode enviesar os modelos de previsão da mortalidade, foi testado um balanceamento das classes de modo a prevenir essas possíveis limitações.

Por último, anteriormente à implementação das técnicas de *data mining* procedeu-se a uma seleção das variáveis que mais contribuíam para as medidas de resultado em estudo através de técnicas de *feature selection*.

4.2.1 Acréscimo de variáveis

O índice de comorbilidade de Charlson (ICC) é uma classificação de gravidade que pode ser aplicada em bases de dados administrativos, ou seja, um método que emprega condições clínicas selecionadas, registadas como diagnósticos secundários (comorbilidades) no cálculo do risco de óbito. O índice calcula a carga de morbilidade do paciente, independentemente do diagnóstico principal. Para constituir o ICC, foram definidas 17 condições clínicas. Para cada uma das condições clínicas, é estabelecida uma pontuação com base no risco relativo, com pesos variando de zero a seis. A pontuação de gravidade de cada doente é o resultado da soma dos pesos de todas as comorbilidades selecionadas.[60]

4.3 Modelos de previsão da mortalidade

4.3.1 Balanceamento das classes

No âmbito do pré-processamento dos dados, implementaram-se técnicas de balanceamento das classes para a construção dos modelos de previsão do risco de mortalidade, na tentativa de minimizar as discrepâncias entre o número de mortes e de sobrevivências (3206 mortes e 34122 sobrevidas) de forma a não enviesar os modelos.[61]

As técnicas utilizadas foram, *oversampling*, *SMOTE* e *ROSE*, contudo também foi testado o não balanceamento das classes.

A técnica de *oversampling* produz amostras aleatórias com substituição (ou seja, replicação) da classe minoritária (por exemplo, episódios com óbito). A vantagem desta técnica é que não leva à perda de informações. Esta técnica adiciona casos replicados do conjunto de dados original, sendo a sua desvantagem a redundância do conjunto de dados de treino, que culmina em vários módulos semelhantes, levando possivelmente a um sobre ajuste do modelo aos dados de treino (*overfiting)*.[61]

Entende-se como sobre ajuste quando o modelo é demasiado adequado ao conjunto de dados de treino, não conseguindo fazer previsões confiáveis quando aplicado num novo conjunto de dados, sendo mais tendencioso (a medida de resultado mais frequente é favorecida).

A técnica SMOTE cria dados artificiais com base nas semelhanças dos casos da classe minoritária, tendo em conta todas as variáveis existentes para essa classe. Para cada caso da classe minoritária presente no conjunto de dados de treino, o SMOTE executa as seguintes etapas:

(1) Calcula os k-vizinhos mais próximos.

(2) Seleciona os N casos da classe maioritária com base na menor magnitude das distâncias euclidianas obtidas dos k-vizinhos mais próximos.

Finalmente, o SMOTE combina a super amostragem sintética dos casos da classe minoritária com a subamostra dos casos da classe maioritária.

A técnica ROSE (*Random Over-Sampling Examples*) extrai amostras artificiais da vizinhança espacial em torno da classe minoritária. O ROSE combina

superamostragem e subamostragem, gerando uma amostra aumentada dos dados (especialmente pertencentes à classe rara).[62]

A técnica ROSE consiste em quatro etapas:

(1) Cria uma nova amostra dos dados da classe maioritária usando uma técnica de reamostragem com *bootstrapping* para remover os casos da classe maioritária de modo a se obter uma a proporção de 50% (subamostragem).

(2) Reamostrar os dados da classe minoritária usando uma técnica de reamostragem de auto inicialização para repetir os módulos da classe minoritária de modo a se obter uma proporção de 50% (sobre amostragem).

(3) Combina os dados das classes maioritária e minoritária das etapas (1) e (2) numa nova amostra de treinamento.

(4) Gera um novo conjunto de dados sintéticos para as classes maioritária e minoritária na sua vizinhança com base nos dados combinados na etapa (3).

Estas quatro etapas são repetidas para cada amostra de treino, a fim de reproduzir uma nova amostra de treino sintética de tamanho aproximadamente igual ao do conjunto de dados original, onde o número de módulos de ambas as classes é representado igualmente (isto é, uma proporção de quase 50%).[63]

4.3.2 Feature Selection

Como pré-processamento dos dados, aplicou-se a técnica de *Recursive Feature Elimination* (RFE) previamente à aplicação dos algoritmos de previsão da mortalidade.

A técnica RFE assume inicialmente todas as variáveis como preditores válidos para uma dada classe, e computa uma pontuação de importância para cada uma das variáveis. As variáveis com pontuações mais baixas são eliminadas, o modelo é novamente construído e são computadas novas pontuações de importância, este processo repete-se até que seja encontrado o subconjunto de variáveis de previsão ideal. Ao longo do processo, o analista especifica o número de variáveis de cada subconjunto, sendo esta característica um parâmetro de sintonização deste método (RFE). O tamanho do subconjunto de variáveis que otimizar os critérios de performance é usado para selecionar as variáveis consoante a sua pontuação formando o subconjunto final.

Este método de seleção é frequentemente utilizado com o modelo de RF por dois motivos. Primeiramente, este último tende a não excluir variáveis da equação de previsão, o que se transpõe na possibilidade de usar todas as variáveis para prever a classe. A técnica RF faz uma seleção aleatória de n variáveis e constrói uma árvore de decisão a partir do melhor nó (da variável com melhor pontuação de importância), sendo este processo repetido k vezes, até que sejam obtidas k árvores distintas. Portanto, cada variável é testada no mínimo uma vez. Tendo em conta que são selecionadas aleatoriamente variáveis como um subconjunto antes de construir cada árvore, existe a obtenção de árvores mais diversificadas, evitando deste modo que as variáveis com uma pontuação de importância maior sejam sobrevalorizadas sem que as demais tenham sido testadas.

O segundo motivo pelo qual o algoritmo RF é usado com o RFE, é porque este possui um método interno conhecido por medir a importância das variáveis, podendo ser usado como primeiro ajuste de modelo do RFE, onde todas as variáveis de previsão do conjunto têm uma pontuação atribuída.

O modelo inicial é construído com todas as variáveis, após tal é eliminada a variável com menor significância estatística, e recursivamente é contruído um novo modelo RF com todas as outras variáveis, repetindo-se o processo sucessivamente até restar apenas 1 variável.

Por último, é selecionado o modelo que tiver melhor *accuracy*, e as variáveis contemplantes neste modelo serão as elegidas para ingressar no algoritmo propriamente dito.[64,65]

4.3.3 Modelos de classificação

Cerca de 80% do conjunto de dados disponível, foi usado para construir os modelos. Na construção dos modelos de previsão da mortalidade por EAM, foram utilizados vários algoritmos de classificação. Tendo sido selecionados um conjunto de algoritmos de aprendizagem automática populares em domínios médicos e conjunto de dados. Seguem-se os algoritmos utilizados, todos disponíveis no pacote R "caret":

1) Regressão logística

A LR é um método de classificação supervisionado bem estabelecido que pode ser considerado uma extensão dos modelos de regressão comuns, usados principalmente para classificar variáveis dicotómicas (ocorrência ou não ocorrência de um determinado evento). Os modelos de LR estimam probabilidades de um caso pertencer a uma determinada classe e consideram um limite de probabilidade para diferenciar as classes. Ao contrário da regressão linear ou de outros modelos baseados em mínimos quadrados ordinários (ou seja, relação linear entre as variáveis independentes e dependentes, em termos de erro (resíduos) que são normalmente distribuidos e de homocedasticidade), não existe uma pré-suposição para fazer modelos de LR. No entanto, geralmente os modelos de LR requerem uma grande amostra, e suposições como pouca ou nenhuma multicolineariedade entre as variáveis independentes (as variábeis independentes não devem ser altamente correlacionadas) e linearidade entre as mesmas. Este algoritmo de classificação tem sido amplamente utilizado na literatura para predizer a mortalidade hospitalar, inclusivamente em doentes com EAM. Foi utilizada a função GLM do R com a função " logit ", junto com o pacote "caret", para implementar modelos de LR.[66]

2) Classification And Regression Trees

As árvores de decisão e regressão são modelos de previsão construídos ao particionar de forma recorrente um conjunto de dados e ajustando um modelo simples a cada partição. O particionamento pode ser representado graficamente como uma árvore de decisão. As árvores de classificação são projetadas para variáveis dependentes que recebem um número finito de valores não ordenados, com um erro de previsão medido em termos de custo de classificação incorreta. As árvores de regressão são para variáveis dependentes que assumem variáveis discretas continuas ou ordenadas, com um erro de previsão medido, normalmente, pela diferença entre os valores observados e os valores previstos.[67]

As árvores produzidas são fáceis de interpretar, a preparação dos dados é relativamente simples (dados numéricos, nominais e categóricos). Contudo, este

algoritmo não tem um desempenho tão bom quanto outros, exige uma ordem dos dados e não sabe lidar com *missings*.[66]

3) K-Nearest Neighbors (kNN)

K-NN é um algoritmo simples que armazena todos os casos disponíveis e classifica novos com base em medidas de similaridade, sendo baseado no princípio de que as amostras semelhantes se localizam próximas umas das outras. Dada uma amostra não classificada, o modelo procura no espaço multidimensional dos atributos os casos mais próximos aos novos dados a serem classificados, de modo a atribuir a classe que seja mais frequente entre os k vizinhos mais próximos [68]

Esta técnica é fácil de entender e implementar, o treino é mais rápido em relação aos outros algoritmos e tem uma boa *performance* no caso em que a amostra tem muitas classes. Entretanto, requer um grande custo computacional quando tem que comparar uma grande amostra não classificada, sendo muito lento a classificar: este método atribui o mesmo peso a cada atributo, o que pode ser problemático se existirem muitos atributos irrelevantes o que resulta numa precisão pobre. Além disso, é sensível à estrutura local dos dados, têm elevados requerimentos de armazenamento e são sensíveis à similaridade da função de distância escolhida para comparar as instâncias.[66]

Neste estudo, o pacote "caret" fornece uma função que escolhe o valor óptimo de k com base na *accuracy* e nas métricas de kappa, sendo que este parâmetro é normalmente definido pelo próprio analista e pode ter um impacto considerável nos resultados de classificação.

4) Naive Bayes

O teorema de Bayes pode ser utilizado para prever algo baseado no conhecimento prévio e evidencia corrente. A previsão é baseada em um conhecimento anterior que pode estar relacionado com o evento de interesse. O teorema também reflete mudanças na probabilidade de um desfecho quando novas evidências são consideradas. A evidência corrente é expressa como a

verosimilhança que reflete a probabilidade de um preditor dado um determinado desfecho. O teorema de Bayes é expresso na seguinte equação:

$P(A|B) = P(B|A) \times P(A) / P(B)$

Onde, P(A) e P(B) são a probabilidade dos eventos A e B sem se considerarem um ao outro;

P(A|B) é a probabilidade de A condicionada por B;

P(B|A) é a probabilidade de B condicionada por A;

Na classificação de NB, A são eventos de desfecho e B são uma serie de preditores. O termo Naive refere-se à assumpção de que os preditores são independentes uns dos outros para o mesmo valor de desfecho. Pelo que pode ser escrito: $P(b_1|A) \times P(b_2|A) \times P(b_3|A)$ em vez de $P(b_1, b_2, b_3|A)$.[69]

Estes algoritmos são fáceis de implementar, requerem pouco treino dos dados, no entanto criam facilmente relações entre variáveis e assume uma distribuição normal das variáveis numéricas.[66]

5) Random Forest

O algoritmo RF é um na prática um conjunto de classificadores, que opera pelo treino de várias árvores de decisão, e retorna a classe votada pela maioria das árvores no conjunto, ou faz a previsão média das árvores individuais.

Uma árvore de decisão começa com as amostras de treino e os seus rótulos de classes associados. Um conjunto inicial é repartido recursivamente em subconjuntos mais "puros" que os conjuntos-pai. Cada nó interno numa árvore representa um teste no atributo (recurso), cada ramificação representa o resultado do teste e cada nó terminal representa a classe. O nó da raiz da árvore é o que melhor divide os dados de treino, existindo várias medidas para encontrar o recurso que melhor divide os dados de treino. Deste modo, a rota de decisão é o caminho traçado deste o nó da raiz até ao nó final, com várias ramificações pelo meio, e nenhuma medida é significativamente superior a outras.[70]

Estes algoritmos são rápidos, robustos a ruídos nos dados, fáceis de interpretar e lida bem com grandes bases de dados. No entanto, à medida que o número de árvores aumenta, o algoritmo torna-se mais lento para previsão em

tempo real, podendo ocorrer *overfitting*, além de atribuir pesos de importância muitos diferentes às variáveis.[66]

6) Support Vector Machine (SVM)

Os modelos de classificação construídos a partir do algoritmo *Support Vector Machine* (SVM) são baseados em teoria estatística de aprendizagem e minimização estrutural de risco, tendo como objetivo a determinação da localização das fronteiras de decisão, também conhecidas como hiperplano, que produz uma separação óptima das classes no espaço multidimensional. Estes algoritmos giram em torno da noção de "margem", qualquer lado do hiperplano que separa duas classes de dados. [71,72]

Este tipo de algoritmo tem uma boa precisão, é robusto para dados de alta dimensão, e tem uma boa capacidade de generalização. Consegue lidar com várias variáveis, com baixo risco de *overfiting*, e lida bem com dados não-estruturados. Entretanto, é computacionalmente caro, e é sensível a ruídos nos dados, sendo que o peso e impacto das variáveis são frequentemente difíceis de interpretar.[66]

4.3.4 Avaliação do desempenho dos modelos

Para avaliar e validar a performance dos modelos foi utilizado o método de 10-fold cross-validation, no qual os dados são divididos em 10 amostras (*folds*), de tamanho aproximado. Neste método de avaliação, 9 amostras dos dados são utilizados para treinar e induzir os modelos, enquanto que a sobrante é utilizada para testar o modelo construído. O processo é repetido até que cada amostra tenha sido utilizada como conjunto de treino e teste. Em geral, a principal vantagem deste método consiste na variância em comparação com abordagens que utilizam apenas um conjunto de validação, sendo menos sensível a viés de particionamento no conjunto de treino ou de teste.[73] Este método é também útil para evitar o *overfitting*.[74] Além disso, para cada iteração do processo de avaliação (com ou sem seleção das variáveis e com ou sem balanceamento da classe anterior), foram calculadas algumas métricas de desempenho, nomeadamente sensibilidade, especificidade, precisão e exatidão (*accuracy*). Adicionalmente, a

habilidade dos modelos é usualmente determinada pela matriz de confusão e pela curva característica de operação do recetor (curva ROC).

A matriz de confusão é uma forma estruturada de identificar verdadeiros positivos (TP), em que a previsão de morte estava de acordo com os dados. Similarmente, verdadeiros negativos (TN), em que a previsão de sobrevivência estava de acordo com os dados. Os falsos positivos (FP) são os casos negativos (morte) identificados incorretamente como positivos (sobrevivência) pelo modelo, e os falsos negativos (FN) são os casos positivos identificados incorretamente pelo modelo.[66] Existem várias formas de manipular estes valores, como a precisão, sensibilidade, especificidade e exatidão:

- Precisão = $\frac{TP}{TP+FP}$
- Sensibilidade = $\frac{TP}{TP+FN}$
- Especificidade = $\frac{TN}{TN+FP}$
- Exatidão = $\frac{TP+TN}{TP+TN+FP+FN}$

A curva ROC é criada traçando os verdadeiros positivos contra os falsos positivos em várias configurações de limite. A área sob a curva ROC (AUC) é também comumente utilizada para determinar a capacidade de previsão do modelo. Quanto maior for a área maior será a qualidade do modelo, e vice-versa.[75]

De forma a validar criticamente os resultados obtidos e objetivando a capacidade de generalização, separou-se 20% dos dados para testar a performance dos modelos em novos dados, obtendo os valores de avaliação acima mencionados como precisão, especificidade, etc.

4.4 Modelos de previsão do período de estadia

4.4.1 Feature Selection

Quanto aos modelos de previsão do período de estadia, utilizaram-se as técnicas de feature selection *forward elimination*, *backward elimination* e *stepwise elimination*.

A técnica *forward elimination* inicia-se sem variáveis preditoras no modelo. As variáveis são adicionadas uma a uma, o método de seleção direta calcula o valor de p (isto é, a variância) para cada uma das variáveis. Se o valor de p calculado para a variável for menor que o valor critico (p inferior a 0,05 – intervalo de confiança de 95%), o método selecionará a variável para o modelo, caso contrário será eliminada. Este processo é feito iterativamente até que todas as variáveis do modelo tenham um valor de p menor que 0,1.

Quanto à técnica *backward elimination*, este método inicia com todas as variáveis de previsão no modelo e remove uma variável de cada vez usando o valor p. Na primeira etapa o valor p é calculado para todas as variáveis preditivas, e a variável com um valor p excedente ao valor p critico é excluída. Na segunda etapa, o valor p é calculado para as restantes variáveis e, novamente, a variável que for maior que o valor p critico é eliminada. Este processo iterativo é repetido até que o valor p mais alto seja menor que o valor p critico, indicando que a variável correspondente não é redundante na presença de outras variáveis.

Numa combinação das técnicas *forward* e *backward selection* surge a técnica *stepwise selection*. Este método começa sem nenhuma variável, adicionando uma a uma ao modelo, cumprindo os critérios de valor de p ($p<0,1$). Após uma variável ser adicionada ao modelo são examinadas todas as variáveis no modelo e excluídas quaisquer que mostrem um valor de p maior que o valor p crítico. A próxima variável é adicionada ao modelo somente após verificar o modelo e excluir quaisquer variáveis, se necessário. Este processo continua até que todas as variáveis atendam aos critérios de p e não tenham um valor p maior que o valor p crítico.[76]

4.4.2 Modelos de regressão

1) Regressão de Poisson

A regressão de Poisson é apropriada quando a classe é contínua, neste caso o número de dias que os doentes estiveram hospitalizados.

A regressão de Poisson trata-se de um modelo de regressão clássico com apenas uma exceção. Sendo essa exceção, a classe assumir uma distribuição de *Poisson*. Esta técnica faz a assunção de que a variância é igual à média nos dados.

Assumindo uma distribuição de *Poisson*, é utilizado o método de máxima verosimilhança. Desta forma, é possível obter estimativas de parâmetros de regressão desconhecidos β0, β1, β2, βk.

Assumindo Y_i como uma variável dependente contínua, e X_i como as variáveis preditoras que podem ser contínuas ou dicotómicas. Assume-se que o valor esperado de y_i será:

$$E\{y_i|x_i\} = \exp\{x_i^T \beta\}$$

A variável contínua como referido acima assumirá uma distribuição de Poisson com o risco relativo$_i$ = $\exp\{x_i^T \beta\}$. A função de probabilidade de y_i condicional de xi é dada por:

$$P\{y_i = y|x_i\} = \exp\{-\lambda_i\} \lambda_i^y / y!, y = 0, 1, 2, ...,$$

Onde y! expressa y fatorial. Substituindo a forma funcional apropriada para λ_i são produzidas expressões para as probabilidades que podem ser usadas para construir a função de probabilidade de log para este modelo, conhecido como regressão de *Poisson*.[77]

A função GLM do R foi usada para construir o modelo de regressão de Poisson.

2) Regressão binomial negativa

A regressão binomial negativa é uma generalização da regressão de Poisson, contudo não assume que a média seja igual à variância nos dados.

Esta técnica pode ser usada para descrever dados de contagem univariados que exibem sobredispersão. A sobredispersão em relação ao modelo Poisson ocorre quando a variância da amostra é substancialmente superior à média da amostra. A distribuição binomial negativa descreve as probabilidades de ocorrência de números inteiros miores ou iguais a 0 tal como a distribuição de Poisson.[78]

4.4.3 Avaliação do desempenho dos modelos

De modo a avaliar o desempenho de modelos de regressão utiliza-se a Critério de informação de Akaike (AIC). Esta criteriação faz uma previsão de erro dos modelos fora da amostra, quanto maior o valor menor a performance do modelo. O melhor modelo mostra o menor valor de AIC.[79]

Além da AIC, também se observaram os desvio nulo e residual. Os desvios são uma medida de adequação do modelo, quanto mais pequeno o desvio mais adequado é o modelo. Quanto maior a diferença entre o desvio residual e nulo, melhor o modelo.[80]

Resultados

5. Resultados

5.1 Modelo de previsão da mortalidade

Nas tabelas 1-8 estão descritos os melhores modelos de cada técnica de *machine learning* face à aplicação ou não de técnicas de *feature selection* e de balanceamento das classes elegíveis para o conjunto de dados.

O estudo considerou 32 variáveis, tendo sido construídos modelos com essas mesmas variáveis, tabelas 1-4. Mas, também foi selecionado um subconjunto das mesmas através da técnica de *feature selection, recursive feature elimination*, que foi utilizada para a construção dos modelos descritos nas tabelas 6-8.

Para a construção de cada modelo foram consideradas técnicas de balanceamento das classes de três tipos, Oversampling, SMOTE e ROSE (tabelas 2, 3, 4, 6, 7, 8), tendo sido construídos previamente modelos sem recurso a essas técnicas (tabelas 1 e 4).

Primeiramente, não foi aplicada qualquer técnica de *feature selection* nem de balanceamento das classes, estando as métricas de avaliação de cada modelo apresentadas na tabela 1. Os modelos descritos apresentam valores extremos de sensibilidade e especificidade, sendo muito sensíveis ou pouco específicos e vice-versa. No que diz respeito à precisão, esta é baixa (abaixo dos 0,5 em todos os modelos). A exatidão balanceada é mediana, não sendo os modelos muito ou pouco exatos. A performance dos modelos é perto dos 0.7 exceto no modelo de CART, sendo expressivamente maior nos modelos construídos com técnicas de LR e NB(mais próximo dos 0.8).

Tabela 3 Resultados dos modelos de mortalidade construídos sem balanceamento das classes.

	LR	CART	KNN	NB	RF	SVM
Sensibilidade	0,030	0,000	0,008	0,000	0,000	0,000
Especificidade	0,996	1,000	0,997	1,000	1,000	1,000
Precisão	0,177	NA	0,187	NA	0,417	0,333
Exatidão balanceada	0,513	0,500	0,502	0,500	0,500	0,500
AUC	0,788	0,500	0,700	0,782	0,702	0,700

A análise dos resultados anteriores potenciou o teste de técnicas de balanceamento das classes, tendo sido primeiro testada a técnica *oversampling*, estando as métricas de avaliação dos modelos deste modo construídos expostas na tabela 2. Os modelos apresentam sensibilidade e especificidade relativamente equilibradas (exceto os

modelos construídos com as técnicas de *machine learning* NB e RF). A precisão mantém-se em valores baixos em todos os modelos (abaixo de 0.2). No que toca à exatidão, esta aumentou consideravelmente geralmente, exceto nos modelos construídos a partir das técnicas de machine learning NB e RF, mantendo-se mediana. O poder discriminativo dos modelos é acima dos 0.65, sendo maior nos modelos com as técnicas de LR e NB.

Tabela 4 Resultados dos modelos de mortalidade construídos com balanceamento das classes, oversampling.

	LR	CART	KNN	NB	RF	SVM
Sensibilidade	0,750	0,775	0,546	0,000	0,154	0,767
Especificidade	0,715	0,647	0,728	1,000	0,618	0,938
Precisão	0,198	0,171	0,159	NA	0,190	0,159
Exatidão balanceada	0,733	0,711	0,637	0,500	0,546	0,693
AUC	0,793	0,760	0,648	0,775	0,747	0,648

A técnica de balanceamento SMOTE foi testada de seguida, os resultados de avaliação dos modelos construídos estão a seguir mencionados, na tabela 3. Os valores da sensibilidade são baixos a medianos, sendo a sua especificidade alta (0.8 a 1). A precisão varia entre os 0.2 e os 0.48, não sendo expressivamente alta em nenhum modelo. A exatidão retorna a valores próximos dos 50%, exceto no modelo de LR. A AUC dos modelos varia entre os 0.7 e 0.77.

Tabela 5 Resultados dos modelos de mortalidade construídos com balanceamento das classes, SMOTE.

	LR	CART	KNN	NB	RF	SVM
Sensibilidade	0,509	0,131	0,195	0,000	0,017	0,178
Especificidade	0,829	0,969	0,950	1,000	0,998	0,953
Precisão	0,218	0,282	0,268	NA	0,478	0,261
Exatidão balanceada	0,669	0,550	0,572	0,500	0,508	0,565
AUC	0,774	0,715	0,708	0,697	0,765	0,715

Em última análise, testou-se a técnica de balanceamento das classes, ROSE, os modelos construídos neste sustentáculo encontram-se na tabela 4. Os valores de sensibilidade e especificidade distanciam-se de forma extrema em todos os modelos exceto nos de LR e KNN. Todos os modelos são pouco precisos (valores abaixo de 0.26). A exatidão continua mediana em todos exceto no modelo com a técnica de LR. Os modelos com maior AUC são o de LR, NB e RF.

Tabela 6 Resultados dos modelos de mortalidade construídos com balanceamento das classes, ROSE.

	LR	CART	KNN	NB	RF	SVM
Sensibilidade	0,761	1,000	0,457	0,000	0,980	0,184
Especificidade	0,707	0,000	0,674	1,000	0,080	0,950
Precisão	0,196	0,086	0,116	NA	0,091	0,258
Exatidão balanceada	0,734	0.500	0,565	0.500	0,530	0,567
AUC	0,790	0,672	0,572	0,749	0,760	0,572

De modo a testar um subconjunto de variáveis recorreu-se à técnica de *feature selection* RFE, tendo sido avaliados os modelos sem recurso a qualquer técnica de balanceamento, as métricas de avaliação dos modelos obtidos estão discriminadas na tabela 5. Os valores da sensibilidade e especificidade são extremos, próximos de 0 e 1, respetivamente. Os modelos não são precisos (valores próximos de 0), e a sua exatidão apresenta valores medianos. A performance dos modelos foi maior nas técnicas de LR e NB (AUC de 0.76).

Tabela 7 Resultados dos modelos de mortalidade construídos através dos dados sujeitos a feature selection (Recursive Feature Elimination) sem balanceamento das classes.

	LR	CART	KNN	NB	RF	SVM
Sensibilidade	0,003	0,000	0,005	0,000	0,008	0,002
Especificidade	0,998	1,000	0,998	1,000	0,999	0,999
Precisão	0,167	0,000	0,005	0,000	0,008	0,002
Exatidão balanceada	0,501	0,500	0,501	0,500	0,503	0,501
AUC	0,763	0,500	0,701	0,758	0,620	0,500

Após a análise dos resultados imediatamente acima mencionados recorreu-se a várias técnicas de balanceamento das classes, tendo sido a primeira, a técnica de *oversampling*, com esta obtiveram-se os resultados descritos na tabela 6. Os valores de sensibilidade e especificidade foram relativamente próximos (diferença de 0.1 ou menor), exceto na técnica de NB. Sendo que os valores de sensibilidade foram próximos de 0,7, e os de especificidade 0.65, salvo a exceção referida. Os modelos demonstraram ser pouco precisos (menor que 0.2 – exceto a técnica de NB). Quanto à exatidão, os valores não foram tão mediados, tendo sido de 0.7 na técnica de LR, CART, SVM. A AUC foi próxima dos valores de 0.7, tendo atingido 0.76 nas técnicas de LR e NB.

Tabela 8 Resultados dos modelos de mortalidade construídos através dos dados sujeitos a feature selection (Recursive Feature Elimination) com balanceamento das classes, Oversampling.

	LR	CART	KNN	NB	RF	SVM
Sensibilidade	0,750	0,755	0,674	0,062	0,648	0,771
Especificidade	0,645	0,641	0,651	0,980	0,677	0,642
Precisão	0,165	0,165	0,154	0,231	0,159	0,168
Exatidão balanceada	0,698	0,698	0,663	0,521	0,662	0,706
AUC	0,763	0,728	0,684	0,761	0,722	0,728

A técnica de balanceamento das classes, SMOTE, também foi testada, tendo os resultados de avaliação sido colocados na tabela 7. Os valores da sensibilidade são geralmente elevados (acima dos 0.7). A especificidade apresenta valores acima dos 0.7 sendo mais expressiva na técnica de NB. Os modelos são pouco precisos (abaixo dos 0.2 exceto na técnica de NB). Quanto à exatidão, esta apresenta valores acima de 0.6, sendo próxima de 0.7 nas técnicas de LR, CART e SVM. Os modelos com maior poder discriminatório foram as técnicas de NB e LR.

Tabela 9 Resultados dos modelos de mortalidade construídos através dos dados sujeitos a feature selection (Recursive Feature Elimination) com balanceamento das classes, SMOTE.

	LR	CART	KNN	NB	RF	SVM
Sensibilidade	0,638	0,544	0,427	0,042	0,401	0,674
Especificidade	0,741	0,778	0,820	0,992	0,827	0,716
Precisão	0,188	0,187	0,182	0,329	0,178	0,182
Exatidão balanceada	0,690	0,661	0,624	0,5170	0,614	0,695
AUC	0,763	0,736	0,688	0,770	0,682	0,736

Por último, testou-se a técnica de balanceamento das classes ROSE, estando os resultados reportados na tabela 8. Os modelos demonstraram-se medianamente ou pouco sensíveis com a aplicação das técnicas KNN e NB, RF, SVM, CART, respetivamente. A especificidade foi perto de 1 em todas as técnicas exceto nas de LR e KNN (0.64 e 0.69, respetivamente). Os modelos são pouco precisos (abaixo de 0.5), sendo as técnicas de NB e RF as mais precisas. Os valores da exatidão são medianos exceto nas técnicas KNN e LR. A AUC é maior no modelo de LR, sendo em todos os modelos perto dos 0.7.

Tabela 10 Resultados dos modelos de mortalidade construídos através dos dados sujeitos a feature selection (Recursive Feature Elimination) com balanceamento das classes, ROSE.

	LR	CART	KNN	NB	RF	SVM
Sensibilidade	0,752	0,056	0,529	0,012	0,039	0,058
Especificidade	0,643	0,990	0,689	0,998	0,993	0,986
Precisão	0,165	0,281	0,138	0,400	0,342	0,286
Exatidão balanceada	0,698	0,521	0,609	0,505	0,516	0,522
AUC	0,763	0,740	0,672	0,756	0,712	0,740

Os valores de AUC não melhoram com a técnica de *feature selection*, no contexto do estudo. Em geral, a técnica de LR superou os outros algoritmos de aprendizagem automática supervisionada nas combinações de algoritmos com balanceamento das classes, apresentando uma AUC média de 0,786 quando não houve feature selection, o que é ligeiramente superior à obtida após aplicação da técnica de feature selection (AUC=0,763 – tabela 7). A maior AUC foi obtida com a aplicação de LR combinada com o *oversampling*, utilizando todas as variáveis iniciais (tabela 4).

No que toca à exatidão balanceada, a LR também superou os outros classificadores supervisionados, com a aplicação da LR em combinação com o oversampling ou ROSE existe uma exatidão mais equilibrada (0,73 – tabela 4 e 6). A exatidão balanceada média sem feature selection aumentou de 0,50 para 0,64, 0,56 e 0,57 após o uso do *oversampling*, SMOTE e ROSE, respetivamente (tabelas 4, 5 e 6).

No geral, a especificidade obtida pelos modelos foi em média muito maior que os valores de sensibilidade. A aplicação das técnicas de *oversampling*, SMOTE e ROSE melhorou a sensibilidade média de menos de 0,1 para 0,49, 0,17 e 0,56, respetivamente, quando não se aplica feature selection, enquanto que os valores aumentaram de menos de 0,1 para 0,61, 0,45 e 0,24, respetivamente, ao considerar a técnica de feature selection. Além disso, o *oversampling* superou os outros algoritmos de balanceamento das duas classes para a maioria dos modelos de aprendizagem automática testados.

No entanto, o modelo mais sensível foi obtido por CART e RF em combinação com a técnica ROSE, atingindo uma sensibilidade de 1 e 0,98, respetivamente (tabela 6). A combinação de SVM com *oversampling* forneceu o segundo maior valor de sensibilidade (cerca de 0,77 – tabela 4 e 8), que foi ligeiramente superior à sensibilidade obtida na técnica de LR combinada com o *oversampling* ou ROSE (0,75 – tabela 4, 8 e 10) e CART combinada com *oversampling* (cerca de 0,76 – tabela 8).

5.2 Modelo de previsão do período de estadia

O tempo de internamento hospitalar foi modelado pela técnica de regressão de Poisson, considerando o conjunto de variáveis utilizado para modelar a mortalidade hospitalar (32).

No entanto, foram também aplicadas técnicas de *feature selection* (*Backward*, *Forward* e *Stepwise*) de forma a obter modelos ideais. Os resultados obtidos com as 3

técnicas foi o mesmo (modelos compostos pelas mesmas variáveis à exceção de uma ou duas).

Ao analisar os desvios residuais, foi identificada a presença de assimetria (mediana igual a -0,579 para todos os três modelos) e o teste qui-quadrado de *goodness-of-fit* foi estatisticamente significativo, concluindo que os dados não se ajustavam ao modelo de Poisson adequadamente. Além disso, verificou-se a presença de sobredispersão, visto que a variância excede a média, sugerindo que a distribuição de Poisson não é adequada no presente contexto.

Como alternativa para modelar o período de estadia, foi posteriormente conduzida uma regressão binomial negativa, utilizando as mesmas variáveis, e que inclui um parâmetro extra para modelar a sobredispersão (parâmetro teta). O parâmetro de dispersão teta faz com que a variância convirja para o mesmo valor da média, transformando a distribuição binomial negativa em distribuição de Poisson. As técnicas *backward, forward* e *stepwise* foram novamente executadas, e o conjunto selecionado de variáveis com significância estatística no tempo de internamento foram muito semelhantes entre si e às obtidas com a regressão de Poisson, tendo os resultados sido simplificados numa única tabela (11).

Para interpretar os resultados obtidos com a função log link, os coeficientes foram exponenciados e estão representados na tabela 11 juntamente com os repetivos valores de p e os intervalos de confiança de 95%. De acordo com o modelo, a presença da maioria das comorbilidades aumenta o número de dias de internamento, mas a hemiplegia foi o fator que mais contribuiu para o aumento do tempo de internamento entre os pacientes com EAM, com um tempo de internamento 83% superior em relação a um paciente sem essa comorbilidade. Além disso, os doentes com episódios urgentes passam em média 16% mais tempo no hospital. Em termos de diagnóstico principal, os doentes com causa de internamento relacionada com o código da ICD-9-CM 410.20 (EAM da parede ínfero-lateral, episódio de cuidado não especificado) apresentam um tempo de internamento 0,95 vezes superior ao grupo de referência (pacientes com código de diagnóstico principal 41.000 – EAM da parede anterolateral, episódio de cuidados não especificados), em média.

Tabela 11 Coeficientes exponenciados obtidos com um modelo de regressão binomial negativa para prever o período de estadia.

	Coeficientes	97.5% CI	z	p-value
Intercepto	3.84	[3.13;4.72]	12.85	0.00
AGE	1.01	[1.01;1.01]	23.78	0.00
urg1	1.16	[1.12;1.21]	7.67	0.00
AMIvolume	1.00	[1.00;1.00]	16.18	0.00
beds	1.00	[1.00;1.00]	-14.44	0.00
NUT2	0.99	[0.98;1.00]	-2.28	0.02
DIAG141001	0.85	[0.70;1.04]	-1.55	0.12
DIAG141010	0.75	[0.60;0.95]	-2.35	0.02
DIAG141011	0.80	[0.66;0.98]	-2.19	0.03
DIAG141020	0.95	[0.60;1.50]	-0.24	0.81
DIAG141021	0.76	[0.62;0.94]	-2.58	0.01
DIAG141030	0.43	[0.24;0.76]	-2.86	0.00
DIAG141031	0.75	[0.61;0.92]	-2.73	0.01
DIAG141040	0.70	[0.53;0.93]	-2.49	0.01
DIAG141041	0.73	[0.60;0.89]	-3.09	0.00
DIAG141050	0.75	[0.44;1.28]	-1.05	0.29
DIAG141051	0.71	[0.58;0.87]	-3.23	0.00
DIAG141060	0.38	[0.08;1.80]	-1.22	0.22
DIAG141061	0.71	[0.56;0.91]	-2.73	0.01
DIAG141070	0.70	[0.57;0.86]	-3.36	0.00
DIAG141071	0.75	[0.62;0.92]	-2.79	0.01
DIAG141080	0.85	[0.64;1.13]	-1.12	0.26
DIAG141081	0.76	[0.62;0.93]	-2.65	0.01
DIAG141090	0.80	[0.63;1.01]	-1.86	0.06
DIAG141091	0.73	[0.60;0.90]	-3.00	0.00
ami1	0.90	[0.86;0.94]	-5.17	0.00
chf1	1.18	[1.14;1.23]	8.37	0.00
cevd1	1.20	[1.15;1.26]	7.89	0.00
dementia1	0.91	[0.85;0.97]	-2.70	0.01
copd1	1.04	[0.99;1.09]	1.69	0.09
pud1	1.19	[1.07;1.31]	3.28	0.00
mld1	1.18	[1.10;1.27]	4.79	0.00
diab1	0.96	[0.92;0.99]	-2.39	0.02
diabwc1	1.12	[1.06;1.18]	4.00	0.00
hp1	1.83	[1.63;2.05]	10.36	0.00
rend1	1.09	[1.05;1.14]	4.13	0.00
canc1	1.03	[0.96;1.09]	0.75	0.45
msld1	1.30	[1.07;1.58]	2.66	0.01
metacanc1	1.16	[1.03;1.30]	2.53	0.01
aids1	1.25	[1.05;1.49]	2.56	0.01
score	1.14	[1.10;1.18]	8.04	0.00

Avançando comparações entre os dois modelos, após a aplicação da regressão de Poisson, verificou-se que o valor do chi-quadrado ao nível de significância (valor crítico) para 29,823 graus de liberdade foi de 30,225.9, sendo que o desvio residual foi de 131,052.3. Com o modelo da binomial negativa, verificou-se um desvio residual muito inferior em comparação ao modelo de Poisson, de 29,645.28, estando portanto abaixo do valor crítico de 30,225.9. A tabela 12 sumaria algumas medidas para comparar a qualidade dos modelos de Poisson e binomial negativa.

Tabela 12 Valores de Akaike Indice Criterion, desvio residual e do modelo de Poisson e modeloda regressão binomial negativa.

	AIC	Desvio Residual	REQM
Poisson	238,490	131,052.3	10.42
Binomial Negativa	173,606	29,645.3	10.43

A distribuição dos resíduos também foi analisada para se perceber o quão ajustado o modeloda binomial negativa está aos dados. Assim como o observado para o modelo de Poisson, a distribuição residual também se mostrou irregular ("*skewed*") para o modelo da binomial negativa (mediana = -0.3121), e ao analisar o gráfico de dispersão dos resíduos *versus* valores previstos do período de estadia (Figura 2), observou-se um padrão aleatório para ambos os modelos, dentro do que era esperado, em que não foi possível detetar claramente uma tendência, exceto pela redução dos resíduos negativos quanto maior forem os valores previstos do período de estadia.

Outra métrica importante para a comparação da qualidade dos modelos consiste no *score* AIC (*Akaike Information Criterion*), sendo que quanto menor o *score* AIC melhor será o modelo, o que foi observado para o modelo da binomial negativa (AIC = 173,606).[81]

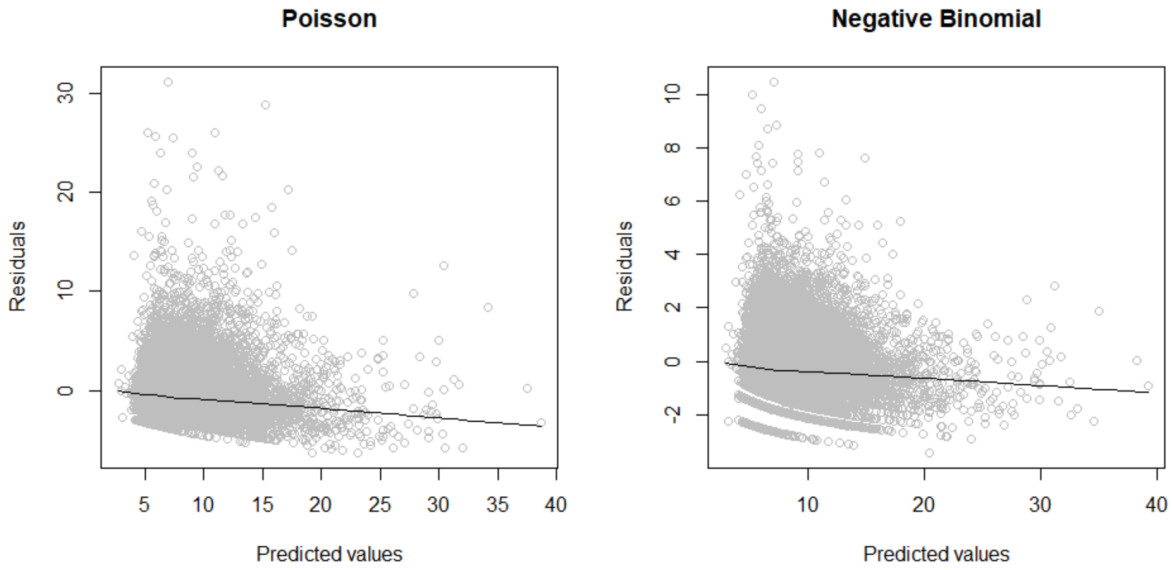


Figura 2 Análise dos resíduos versus valores do período de estadia previstos pelos modelos de Poisson e binomial negativa.

Adicionalmente, para analisar e comparar o quão bem os modelos preveêm o período de estadia em casos novos, foi utilizado um conjunto de teste com 7,464 internamentos e o valor da raíz do erro quadrático médio (REQM) foi estimado para ambos os modelos. Através da figura 3 é possível observar um gráfico de densidade que mostra a distribução dos valores do período de estadia observados e os valores previstos (linha vermelha), tendo sido verificado que a distribuição foi aproximadamente a mesma para ambos os modelos. Adicionalmente, observou-se que o valor da REQM foi praticamente o mesmo para ambos os modelos (erro ao redor de 10 dias).

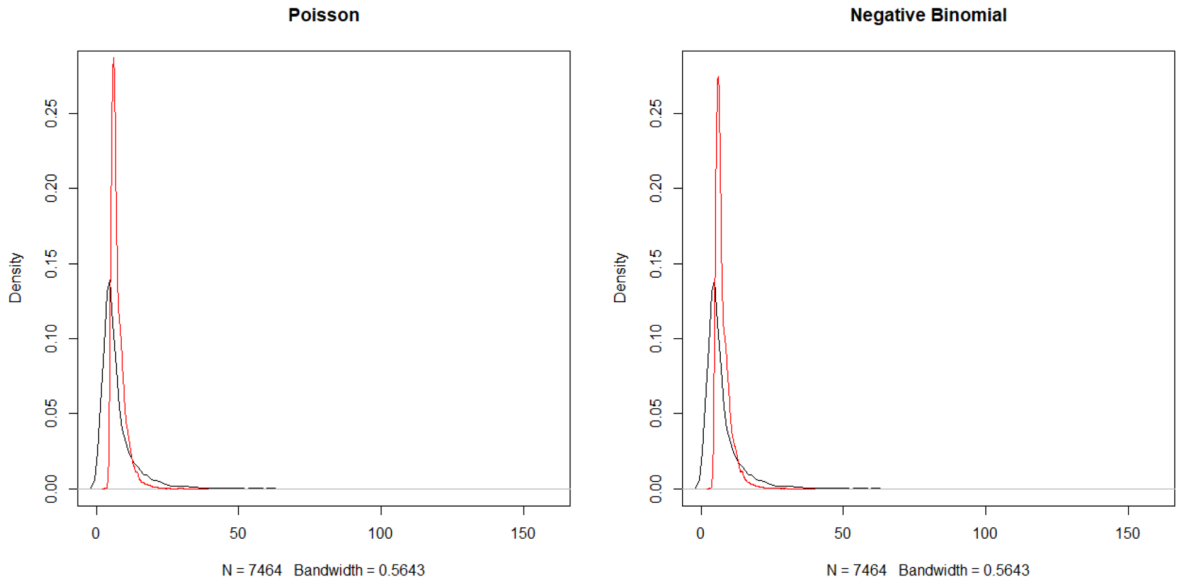


Figura 3 Gráfico de densidade dos valores do período de estadia observados e previstos pelos modelos de Poisson e da binomial negativa.

Discussão

6. Discussão

Recordando o que foi mencionado acima, o desempenho de várias técnicas para construir modelos de previsão da mortalidade e período de estadia hospitalares foi avaliado utilizando uma amostra de 37.328 hospitalizações por EAM extraídas de uma base de dados de pacientes internados em Portugal. Um conjunto de variáveis referentes às características do hospital, além de variáveis demográficas e clínicas habitualmente presentes na admissão, como comorbilidades e outras doenças pré-existentes, foram utilizadas na construção dos modelos.

Este conjunto de variáveis foi selecionado em estudos anteriores usando o mesmo conjunto de dados. Foram testados métodos de *feature selection* de modo a selecionar os melhores preditores de mortalidade e período de estadia hospitalares, o impacto da seleção das variáveis foi avaliado.

A capacidade de previsão da mortalidade hospitalar por EAM dos vários modelos de aprendizagem automática foi avaliada. Os algoritmos utilizados são bem estabelecidos e amplamente aceites, com aplicações anteriores na área médica, tendo elevado desempenho nas mesmas. Vários estudos têm utilizado algoritmos de aprendizagem automática para prever a mortalidade hospitalar por EAM, sendo a regressão Logística a técnica mais recorrente. É importante realçar, no entanto, que o desempenho dos modelos de classificação depende de vários fatores, desde a qualidade, adequação dos dados utilizados para a construção dos modelos, seleção de variáveis relevantes e características da população presente na amostra. Neste estudo, procuramos investigar o efeito da aplicação de métodos como *feature selection* e balanceamento de classes de modo a obter modelos preditivos com utilidade clínica, para a gestão dos cuidados e que possa resultar em benefícios ao próprio paciente.

No presente estudo, o desempenho dos modelos em termos de AUC e exatidão balanceada (taxa de acertos ajustada a ambas as classes, e não apenas globalmente) não melhorou com a aplicação da técnica de *feature selection*. Neste estudo, e de maneira geral, a LR obteve um desempenho superior em comparação aos demais modelos testados em termos de exatidão balanceada e AUC, tendo

fornecido o melhor resultado de classificação quando combinada com o método *oversampling* para tratar o desbalanceamento das classes (AUC igual a 0,786 e exatidão balanceada em torno de 74%). Este resultado vai de encontro ao estudo conduzido por Li et. al (2020), que utilizou e comparou vários algoritmos de classificação para prever a mortalidade por EAM, em que a LR também superou os modelos de aprendizagem supervisionada testados no presente estudo, nomeadamente K-NN, RF e árvores de Decisão. Entretanto, a área sob a curva obtida com os modelos no contexto do estudo de Li et al (2020) variou entre 0,709 a 0,942, que se sobrepõe ao intervalo obtido neste estudo, embora com valores superiores.[82]

No geral, a técnica de *feature selection* não melhorou os resultados de classificação dos algoritmos testados, apesar de produzir modelos menos complexos. Uma possível explicação pode ser na baixa multicolinearidade ou redundância do conjunto inicial das variáveis independentes utilizadas no estudo. Possivelmente estas varáveis já exercem um importante papel preditivo no contexto da mortaliade por EAM, reforçando as evidências encontradas por Lobo et al. (2020) [50]

Visto que a proporção de mortes no conjunto de dados era substancialmente menor que os sobreviventes (aproximadamente 9%), diferentes técnicas de balanceamento das classes foram empregues de modo a ajustar o conjunto de dados, melhorando assim o desempenho dos modelos a preverem mortes. Globalmente, a utilização destas técnicas foi de facto crucial para melhorar o desempenho dos modelos testados. Os valores de precisão foram geralmente muito baixos, nomeadamente em função da pequena proporção de óbitos na base de dados utilizada, o que também foi refletido nos valores de especificidade muito superiores aos valores de sensibilidade. A exatidão média (sem *feature selection*) melhorou de 50.5% para 63.7%, 56.1% e 56.6% ao utilizar os métodos oversampling, SMOTE e ROSE, respetivamente. O mesmo efeito foi observado ao se fazer *feature selection* previamente ao treino dos modelos, em que a exatidão balanceada subiu de 50.1% para 65.8, 63.3 e 56.2% nos métodos oversampling, SMOTE e ROSE, respetivamente.

Com o aumento do número de casos positivos (óbitos) na base de dados, verificou-se uma melhoria substancial na sensibilidade. Sem qualquer aplicação de técnicas de balanceamento, e sem uso de *feature selection*, a sensibilidade média era inferior a 1%, sendo que este valor aumentou para 49%, 17% e 56% com os métodos *oversampling*, SMOTE e ROSE, respetivamente. O mesmo efeito foi observado sem a aplicação prévia de *feature selection*, em que a sensibilidade média dos modelos subiu de menos de 1% para 61%, 45% e 24% ao se utilizar os métodos *oversampling*, SMOTE e ROSE, respetivamente. O método de balanceamento mais simples, *oversampling*, em geral, produziu melhores resultados para a maioria dos modelos testados, embora os modelos mais sensíveis tivessem sido obtidos através da técnica ROSE em combinação com árvore de decisão ou RF, ambos com valores de sensibilidade próximos aos 100%. A utilização do algoritmo SVM em combinação com o método *oversampling* produziu a segunda maior sensibilidade (a volta de 77%), o que foi levemente superior ao terceiro modelo mais sensível, obtido através do algoritmo SVM em combinação com as técnicas de *oversampling* ou ROSE (ambos produziram valores de sensibilidade perto de 75%).

Os resultados obtidos em outros estudos existentes na literatura como o de *Salman et al.,* que obtiveram através do algoritmo de *tree augmented* NB uma AUC de 95% no seu modelo de previsão da mortalidade em doentes com EAM.[45] Com a utilização de variáveis laboratoriais e clínicas *Chee Chin et al.* construíram um modelo com base na técnica de LR multivariada, tendo obtido um bom poder discriminatório (85% - *concordance statistic* ou *c statistic*), este grupo desenvolveu também um modelo de pontuações de risco com uma performance de 84% (*c statistics*).[46]

Yu Qi et al. construíram um modelo de pontuação de risco de mortalidade tendo obtido um poder discriminatório bom (84% *c statistics*), superior ao obtido no presente estudo, e apesar de algumas limitações pode ser replicado e futuramente integrado como ferramenta de previsão de risco.[44] Em alinhamento com os modelos de estratificação de risco, *Joon-Myoungkwon et al.* ao utilizar a técnica de *deep learning* obtiveram um, tendo sido alcançada uma AUC de 90%, apesar das limitações o resultado é valorizável.[40] *Chenxi Song et al.* obtiveram um sistema de pontuação com um poder de discriminação bom (AUC igual a 84%)[41]

utilizando a técnica de LR multivariada. Utilizando a técnica *tree based boosting* *Benjamin Goldstein et al.* obtiveram o melhor modelo, com um desempenho de 76% (*c statistics*).[42] A partir da mesma técnica *Robert McNamara et al.* desenvolveram outro modelo de estratificação, tendo obtido o valor de 88% (*c statistics*) no seu poder discriminatório.[43] Acrescentando por último a obtenção de um modelo por *Zhi Qu et al.* com recurso à técnica de LR, o método mais usual, tendo obtido uma performance acima dos 80% (*c statistics*).[47]

Após a apreciação dos valores apresentados na literatura, e comparando com o presente estudo, percebe-se que os resultados atingidos noutras investigações são geralmente melhores, sendo normal se as populações e as variáveis utilizadas forem muito diferentes. No entanto, a comparação entre alguns estudos torna-se ambígua quando os métodos de avaliação não são os mesmos, como nos estudos que usam o *c statistics* como métrica ao invés da AUC e constroem um modelo de pontuação de risco ao invés de um modelo de previsão da mortalidade.

Para além da mortalidade hospitalar, o período de estadia é um dado crítico na gestão de recursos hospitalares. Considerando o estabelecimento de prioridades, esta abordagem pretende adicionar evidências aos desenvolvimentos existentes no estudo de modelos de previsão do período de estadia.

Através dos resultados acima citados, foi possivel verificar que o modelo da técnica de regressão de Poisson não se adequa tão bem aos dados quanto o modelo da binomial negativa (com uma pontuação de AIC superior), embora que os dois tenham o mesmo desempenho ao prever o período de estadia.

O erro obtido em média em ambos os modelos de previsão do período de estadia foi de 10 dias de internamento (ou seja, existe em média um desvio de 10 dias nas previsões), o que é considerado elevado no contexto do presente estudo. Tal pode estar relacionado com a qualidade dos dados, nomeadamente a ausência de variáveis que não estavam disponíveis na base de dados e teríam um efeito na previsão do período de estadia. Para além disso, o facto de as variáveis serem correspondentes à admissão, eventualmente existem variáveis que surgem ao longo do episódio, como certos procedimentos, que poderiam ter um importante efeito no período de estadia.

No estudo mencionado acima, de Magalhães *et al.*, foi construído um modelo de previsão do período de estadia hospitalar tendo em conta dados laboratoriais

e administrativos. Foram utilizadas variáveis de dados laboratoriais como, a pressão de oxigênio, para além de variáveis de comorbilidades e dados sociodemográficos. A técnica utilizada foi a LR e o resultado foi positivo, o modelo adequa-se aos dados e é um bom preditor (poder discriminatório superior a 80%). A partir da análise dos modelos, pode ser atribuída a respondabilidade à escolha de variáveis, no entanto a população também foi criteriosamente escolhida neste estudo.

Os modelos construídos apresentam algumas limitações, entre elas o facto de o pacote "*caret*" do R *studio* usar o *accuracy* para realizar o ajuste de parâmetro nos algoritmos SVM, CART e RF, bem como para definir o valor de k no modelo de KNN, que por sua vez pode introduzir um viés na escolha do modelo ideal, especialmente num conjunto de dados desbalanceado, como o utilizado neste estudo. A amostra escolhida não é homogénea, pelo que o resultado pode ser inviesado. Outra particularidade, referida ao longo desta discussão, prende-se com as variáveis selecionadas para construir os modelos de previsão da mortalidade e período de estadia estarem presentes na admissão, e poderem não representar bem a população em estudo. Além disso variáveis mais específicas nomeadamente de dados laboratoriais dariam um melhor poder de previsão, algo relatado noutros estudos. E tendo em conta que o tempo é um fator crítico neste contexto, as distâncias entre as residências dos pacientes ao hospital eram de considerar.

Conclusão e perspetivas futuras

7. Conclusão e perspetivas futuras

No que se refere à previsão da mortalidade, no geral, o modelo de LR obteve a melhor performance entre as restantes técnicas, embora a melhor sensibilidade tenha sido alcançada pelo modelo de RF em dados sujeitos a balanceamento das classes (ROSE), mas sem recurso à técnica de *feature selection,* o mesmo aconteceu quanto à precisão, o melhor valor foi obtido pela mesma técnica contudo com a técnica de balanceamento das classes SMOTE.

Uma assumpção realizada pela aplicação da técnica de *feature selection* tratou-se do seu efeito quase nulo no desempenho dos modelos, o que poderá estar associado com a escolha de variáveis selecionadas para este estudo, sendo o conjunto suficiente para explicar a mortalidade, com um baixo nível de redundância entre as variáveis preditoras.

Pelo contrário, o balanceamento das classes foi crítico ou mesmo essencial para melhorar a capacidade de deteção dos casos de morte, destacando-se a técnica de *oversampling* que resultou no maior aumento de performance dos modelos.

A previsão do período de estadia foi melhor modelada pelo modelo da binomial negativa que pelo modelo da regressão de Poisson, conquanto a qualidade para prever o número de dias tenha sido semelhante em ambas as técnicas. Embora os erros de previsão tenham sido consideravalmente elevados neste estudo, algo relevante na sua redução de erros de previsão poderá ter sido o conjunto de variáveis, que poderá ser futuramente ajustado noutro estudo. Assim como nos modelos de previsão da mortalidade, as variáveis selecionadas foram muito parecidas, podendo estar associado, também com o fator acima mencionado.

Sendo que os modelos de previsão do período de estadia e mortalidade permitiriam uma melhoria na gestão hospitalar e do doente, a realização de estudo nesta área é sugestiva de investimento.

Aquando a futura construção de modelos de cariz preditivo da mortalidade e período de estadia de pacientes com EAM, sugere-se o teste de novas variáveis, nomeadamente de índole laboratorial, para que a previsão realizada seja mais

precisa. Sem menosprezar a utilização de novas técnicas de *data mining,* que possam ser mais adequadas ao contexto, ou a seleção de uma população mais homogénea (por exemplo com ajuste de *case-mix*), dada a complexidade da patologia. Tendo em conta os fatores externos, como a área geográfica ou o levar em consideração variações a nível institucional (hospital) e do sistema de saúde como um todo, pelo que a utilização de modelos hierárquicos torna-se pertinente. Apesar deste tipo de modelos ser importante a nível investigacional, a nível hospitalar modelos que prevejam uma pontuação numa determinada escala de risco de mortalidade serão mais apreciados. Quanto mais informativo for o modelo maior será a sua valorização na prática clínica, nesse contexto sugere-se a evolução de modelos binários para modelos de pontuação.

Referências bibliográficas

8. Referências bibliográficas

1. Saleh, M. & Ambrose, J. A. Understanding myocardial infarction [version 1; referees: 2 approved]. *F1000Research* **7**, 1–8 (2018).
2. World Health Organization. Prevention of Recurrences of Myocardial Infarction and Stroke Study. Available at: https://www.who.int/cardiovascular_diseases/priorities/secondary_prevention/country/en/index1.html.
3. Haye H. van der Wal, Vincent M. van Deursen, Peter van der Meer, and A. A. V. Comorbidities in Heart Failure. *Springer* (2017). doi:10.1007/164
4. Rohrmann, S., Witassek, F., Erne, P., Rickli, H. & Radovanovic, D. Treatment of patients with myocardial infarction depends on history of cancer. *Eur. Hear. journal. Acute Cardiovasc. care* **7**, 639–645 (2018).
5. Corrales-Medina, V. F. *et al.* Cardiac complications in patients with community-acquired pneumonia: A systematic review and meta-analysis of observational studies. *PLoS Med.* **8**, (2011).
6. Silva, F. M. orit. F., Pesaro, A. E. duard. P., Franken, M. & Wajngarten, M. Acute management of unstable angina and non-ST segment elevation myocardial infarction. *Einstein (Sao Paulo).* **13**, 454–461 (2015).
7. Rathore, V. Risk Factors of Acute Myocardial Infarction: A Review. *Eurasian J. Med. Investig.* (2018). doi:10.14744/ejmi.2018.76486
8. Nichols, M., Townsend, N., Scarborough, P. & Rayner, M. Cardiovascular disease in Europe 2014: Epidemiological update. *Eur. Heart J.* **35**, 2950–2959 (2014).
9. Li, Q. *et al.* National trends in hospital length of stay for acute myocardial infarction in China. *BMC Cardiovasc. Disord.* **15**, 1–12 (2015).
10. Aso, S. *et al.* Incidence and Mortality of Acute Myocardial Infarction. *Int. Heart J.* **52**, 197–202 (2011).
11. OECD Health Statistics 2019. (2019). Available at: https://www.oecd.org/health/health-data.htm.
12. Castro-dominguez, Y., Dharmarajan, K. & Mcnamara, R. L. Predicting Death after Acute Myocardial Infarction. *Trends Cardiovasc. Med.* (2017).

doi:10.1016/j.tcm.2017.07.011

13. Abdissa Negassa, Ph.D.†, E. Scott Monrad, M.D.‡, Ji Yon Bang, M.Sc†, and V.S. Srinivas, M.B.; B.S., F. A. C. C. . Tree-structured Risk Stratification of In-hospital Mortality Following Percutaneous Coronary Intervention for Acute Myocardial Infarction: A Report From the New York State Percutaneous Coronary Intervention Database. *Am Hear. J.* **32**, 736–740 (2007).
14. Shinn, C. Registo Clínico Integrado.
15. Hoque, D. M. E., Kumari, V., Ruseckaite, R., Romero, L. & Evans, S. M. Impact of clinical registries on quality of patient care and health outcomes: Protocol for a systematic review. *BMJ Open* **6**, 1–7 (2016).
16. Shortliffe, E. H. & Cimino, J. J. *Biomedical informatics: Computer applications in health care and biomedicine: Fourth edition*. *Biomedical Informatics: Computer Applications in Health Care and Biomedicine: Fourth Edition* (2014). doi:10.1007/978-1-4471-4474-8
17. Hoeijmakers, F., Beck, N., Wouters, M. W. J. M., Prins, H. A. & Steup, W. H. National quality registries: how to improve the quality of data? *J. Thorac. Dis.* **10**, S3490–S3499 (2018).
18. Ministério da Saúde (Brasil). Instituto Nacional de Câncer José de Alencar Gomes da Silva (INCA). Câncer de Mama - Tratamento. *J. Am. Med. Informatics Assoc.* **10**, 470–477 (2003).
19. Alonso, V. *et al.* Health records as the basis of clinical coding: Is the quality adequate? A qualitative study of medical coders' perceptions. *Heal. Inf. Manag. J.* (2019). doi:10.1177/1833358319826351
20. Benedictine University. Public Health Research Guide: Primary & Secondary Data Definitions. Available at: https://researchguides.ben.edu/c.php?g=282050&p=4036581. (Accessed: 10th October 2020)
21. Saúde), A. (Administração C. do S. de. Codificação Clínica.
22. Roberts, L., Araromi, S. & Peatman, O. Clinical coding - an insight into healthcare data. *Br. Student Dr.* **2**, 36 (2018).
23. Moghaddasi, H., Rabiei, R. & Sadeghi, N. Improving the quality of clinical coding: a comprehensive audit model. *J. Heal. Manag. Informatics* **1**, 36–40

(2014).

24. Barateiro, José, H. G. A survey of data quality tools. *Datenbank-Spektrum* **5**, 15–21 (2005).
25. Tayi, G. K. & Ballou, D. P. Examining data quality (TayiBallou). *Commun. ACM* **41**, 54–57 (1998).
26. Gaspar, J. Detection of inconsistencies in hospital data. (2012). doi:10.5220/0003757301890194
27. ames H. Stephensa, Gerald R. Ledlowb, and T. V. F. Converting ICD-9 to ICD-10. (2016).
28. Freitas, A., Gaspar, J., Rocha, N., Marreiros, G. & Da Costa-Pereira, A. Quality in hospital administrative databases. *Appl. Math. Inf. Sci.* **8**, 1–6 (2014).
29. Nguyen, L. L. & Barshes, N. R. Analysis of large databases in vascular surgery. *J. Vasc. Surg.* **52**, 768–774 (2010).
30. Lipscombe, L. L. & Hux, J. E. Trends in diabetes prevalence, incidence, and mortality in Ontario, Canada 1995-2005: a population-based study. *Lancet* **369**, 750–756 (2007).
31. Mazzali, C. & Duca, P. Use of administrative data in healthcare research. *Intern. Emerg. Med.* **10**, 517–524 (2015).
32. Ioannidis, J. P. A. Are mortality differences detected by administrative data reliable and actionable? *JAMA - J. Am. Med. Assoc.* **309**, 1410–1411 (2013).
33. Chen, I. Y., Agrawal, M., Horng, S. & Sontag, D. Robustly Extracting Medical Knowledge from EHRs: A Case Study of Learning a Health Knowledge Graph. *Pac. Symp. Biocomput.* **25**, 19–30 (2020).
34. Gavrielov-Yusim, N. & Friger, M. Use of administrative medical databases in population-based research. *J. Epidemiol. Community Health* **68**, 283–287 (2014).
35. Gama, J., Carvalho, A. P. de L., Facelli, K., Márcia, L. & Lorena, A. C. Extração de Conhecimento de Base de dados - Data Mining. *Extração Conhecimento Base dados - Data Min.* 236-242;285-311 (2012).
36. Jordan, M. I. & Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science (80-.).* **349**, 255–260 (2015).
37. Deo, R. C. Machine Learning in Medicine HHS Public Access. *Circulation* **132**, 1920–1930 (2015).

38. Erickson, B. J., Korfiatis, P., Akkus, Z. & Kline, T. L. Machine Learning for Medical Imaging. *Radiographics* 1–11 (2017).
39. Arnold, S. V. *et al.* Risk factors for rehospitalization for acute coronary syndromes and unplanned revascularization following acute myocardial infarction. *J. Am. Heart Assoc.* **4**, 1–8 (2015).
40. Kwon, J. myoung *et al.* Deep-learning-based risk stratification for mortality of patients with acute myocardial infarction. *PLoS One* **14**, 1–15 (2019).
41. Song, C. *et al.* The CAMI-score: A Novel Tool derived from CAMI Registry to Predict In-hospital Death among Acute Myocardial Infarction Patients. *Sci. Rep.* **8**, 1–8 (2018).
42. Goldstein, B. A., Navar, A. M. & Carter, R. E. Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges. *Eur. Heart J.* **38**, 1805–1814 (2017).
43. McNamara, R. L. *et al.* Predicting In-Hospital Mortality in Patients With Acute Myocardial Infarction. *J. Am. Coll. Cardiol.* **68**, 626–635 (2016).
44. Qi, Y. *et al.* Development and validation of Women Acute Myocardial Infarction in-Hospital Mortality Score (WAMI Score). *Int. J. Cardiol.* **259**, 31–39 (2018).
45. Salman, I. Heart attack mortality prediction: An application of machine learning methods. *Turkish J. Electr. Eng. Comput. Sci.* **27**, 4378–4389 (2019).
46. Chin, C. T. *et al.* Risk adjustment for in-hospital mortality of contemporary patients with acute myocardial infarction: The Acute Coronary Treatment and Intervention Outcomes Network (ACTION) Registry®-Get with the Guidelines (GWTG)™ acute myocardial infarction mortality mo. *Am. Heart J.* **161**, 113-122.e2 (2011).
47. Qu, Z., Zhao, L. P., Ma, X. & Zhan, S. Building a patient-specific risk score with a large database of discharge summary reports. *Med. Sci. Monit.* **22**, 2097–2104 (2016).
48. Rocha, Á., Correia, A. M., Tan, F. B. & Stroetmann Editors, K. A. *Advances in Intelligent Systems and Computing 276 New Perspectives in Information Systems and Technologies, Volume 2*. **1**, (2014).
49. Busse, R. *Diagnosis Related Groups in Europe*. *Diagnosis Related Groups in*

Europe (1993). doi:10.1007/978-3-642-78472-9

50. Lobo, M. F. *et al.* Understanding the large heterogeneity in hospital readmissions and mortality for acute myocardial infarction. *Health Policy (New. York).* **124**, 684–694 (2020).
51. INE & Turismo de Portugal, I. Statistics Portugal. *As pessoas - 2018* 35 (2020). Available at: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0007430&contexto=bd&selTab=tab2. (Accessed: 10th October 2020)
52. INE. Statistics Portugal. *Estatísticas da saúde - 2017.* (2012). Available at: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_publicacoes&PUBLICACOESpub_boui=320460040&PUBLICACOESmodo=2. (Accessed: 10th October 2020)
53. Statistics Portugal. *Internments (No.) in hospitals by Geographic localization.* Available at: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0002904&contexto=bd&selTab=tab2. (Accessed: 10th October 2020)
54. Sistema Nacional de Saúde. Direção Geral de Saúde. *Programa Nacional para as Doenças Cérebro-Cardiovasculares* Available at: https://www.dgs.pt/paginas-de-sistema/saude-de-a-a-z/programa-nacional-para-as-doencas-cerebro-cardiovasculares/relatorios-e-publicacoes.aspx. (Accessed: 10th October 2020)
55. Eurostat. *Regions in the European Union: Nomenclature of territorial units for statistics NUTS 2013/EU-28: Manual and guides* Available at: https://ec.europa.eu/eurostat/home? (Accessed: 10th October 2020)
56. Direção Geral da Saúde. Programa nacional para as doenças cérebro-cardiovasculares.
57. Federação Nacional dos Médicos. Available at: http://www.fnam.pt/antigo/ministeriais/ ministeriais files/031128hospuniv.htm. (Accessed: 10th October 2020)
58. QualityNet. *Archived resources - mortality measures*
59. Spencer, F. A., Lessard, D., Gore, J. M., Yarzebski, J. & Goldberg, R. J.

Declining Length of Hospital Stay for Acute Myocardial Infarction and Postdischarge Outcomes: A Community-Wide Perspective. *Arch. Intern. Med.* **164**, 733–740 (2004).

60. Souza, J. *et al.* Importance of coding co-morbidities for APR-DRG assignment: Focus on cardiovascular and respiratory diseases. *Heal. Inf. Manag. J. Heal. Inf. Manag. Assoc. Aust.* **49**, 47–57 (2020).
61. Tantithamthavorn, C., Hassan, A. E. & Matsumoto, K. The Impact of Class Rebalancing Techniques on the Performance and Interpretation of Defect Prediction Models. *IEEE Trans. Softw. Eng.* **PP**, 1 (2018).
62. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
63. Lunardon, N., Menardi, G. & Torelli, N. ROSE: A package for binary imbalanced learning. *R J.* **6**, 79–89 (2014).
64. Chen, Q., Meng, Z., Liu, X., Jin, Q. & Su, R. Decision variants for the automatic determination of optimal feature subset in RF-RFE. *Genes (Basel).* **9**, (2018).
65. Max Kuhn, K. J. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. (Press, Crc, 2019).
66. Uddin, S., Khan, A., Hossain, M. E. & Moni, M. A. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med. Inform. Decis. Mak.* **19**, 1–16 (2019).
67. Soslow, A., Alektiar, K. M., Hensley, M. L. & Jr, M. M. L. Endometrial Carcinoma : Seeing the Forest for the Trees. **130**, 452–456 (2014).
68. Cover, T. M. & Hart, P. E. Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* **13**, 21–27 (1967).
69. Di, P. & Duan, L. New naive Bayes text classification algorithm. *Shuju Caiji Yu Chuli/Journal Data Acquis. Process.* **29**, 71–75 (2014).
70. Lobo, M. F. *et al.* A comparison of in-hospital acute myocardial infarction management between Portugal and the United States: 2000-2010. *Int. J. Qual. Heal. Care* **29**, 669–678 (2017).
71. T., J. Making Large-Scale SVM Learning Practical. *Chimerica* 28 (1998). doi:10.5040/9781784600266.00090006

72. Awad, M. & Khanna, R. Efficient learning machines: Theories, concepts, and applications for engineers and system designers. *Effic. Learn. Mach. Theor. Concepts, Appl. Eng. Syst. Des.* 1–248 (2015). doi:10.1007/978-1-4302-5990-9

73. Sakr, S. *et al.* Comparison of machine learning techniques to predict all-cause mortality using fitness data: The Henry Ford exercIse testing (FIT) project. *BMC Med. Inform. Decis. Mak.* **17**, 1–15 (2017).

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

75. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874 (2006).

76. Heinze, G., Wallisch, C. & Dunkler, D. Variable selection – A review and recommendations for the practicing statistician. *Biometrical J.* **60**, 431–449 (2018).

77. Zamani, H., Faroughi, P. & Ismail, N. Bivariate generalized Poisson regression model: applications on health care data. *Empir. Econ.* **51**, 1607–1621 (2016).

78. Famoye, F. On the bivariate negative binomial regression model. *J. Appl. Stat.* **37**, 969–981 (2010).

79. Obubu, M., A, O. G., N, O. C. & O, O.-I. H. Effect of Akaike Information Criterion on Model Selection in Analyzing Auto-crash Variables. *Int. J. Sci. Basic Appl. Res.* **26**, 98–109 (2016).

80. Sierk, M. L., Smoot, M. E., Bass, E. J. & Pearson, W. R. Improving pairwise sequence alignment accuracy using near-optimal protein sequence alignments. *BMC Bioinformatics* **11**, (2010).

81. Lindsey, J. K. & Jones, B. Choosing among generalized linear models applied to medical data. *Stat. Med.* **17**, 59–68 (1998).

82. Li, Y. M. *et al.* Machine learning to predict the 1-year mortality rate after acute anterior myocardial infarction in Chinese patients. *Ther. Clin. Risk Manag.* **16**, 1–6 (2020).

Anexos

9. Anexos

Anexo I Tabelas dos resultados gerais dos modelos de previsão da mortalidade, precisão, sensibilidade, especificidade, exatidão, valor de kappa e exatidão balanceada.

Precisão							
		Reg Log	CART	K-NN	NB	RF	SVM
Sem *feature selection*	No balancing	0,167	NA	0,188	NA	0,417	0,333
	Oversampling	0,199	0,171	0,159	NA	0,190	0,159
	SMOTE	0,218	0,282	0,268	NA	0,478	0,261
	ROSE	0,196	0,086	0,116	NA	0,091	0,258
Feature selection	No balancing	0,167	0,000	0,005	0,000	0,008	0,002
	Oversampling	0,165	0,165	0,154	0,231	0,158	0,168
	SMOTE	0,188	0,188	0,183	0,329	0,178	0,182
	ROSE	0,165	0,281	0,138	0,400	0,342	0,287

Sensibilidade							
		Log Reg	CART	K-NN	NB	RF	SVM
Sem feature selection	No balancing	0,030	0,000	0,008	0,000	0,000	0,000
	Oversampling	0,750	0,775	0,546	0,000	0,154	0,768
	SMOTE	0,509	0,131	0,195	0,000	0,017	0,178
	ROSE	0,761	1,000	0,457	0,000	0,980	0,184
Feature selection	No balancing	0,003	0,000	0,005	0,000	0,008	0,002
	Oversampling	0,750	0,755	0,674	0,062	0,647	0,771
	SMOTE	0,638	0,544	0,427	0,042	0,401	0,674
	ROSE	0,752	0,056	0,529	0,012	0,039	0,058

Especificidade							
		Log Reg	CART	K-NN	NB	RF	SVM
Sem *feature selection*	No balancing	0,996	1,000	0,997	1,000	1,000	1,000
	Oversampling	0,715	0,647	0,728	1,000	0,618	0,938
	SMOTE	0,829	0,969	0,950	1,000	0,998	0,953
	ROSE	0,708	0,000	0,674	1,000	0,081	0,950
Feature selection	No balancing	0,999	1,000	0,998	1,000	0,999	1,000
	Oversampling	0,644	0,641	0,652	0,981	0,677	0,642
	SMOTE	0,741	0,778	0,820	0,992	0,827	0,716
	ROSE	0,643	0,987	0,689	0,998	0,993	0,987

Exatidão							
		Log Reg	CART	K-NN	NB	RF	SVM
Sem *feature selection*	No balancing	0,914	0,914	0,912	0,914	0,914	0,914
	Oversampling	0,718	0,658	0,713	0,914	0,871	0,631
	SMOTE	0,795	0,893	0,880	0,851	0,913	0,882

	ROSE	0,712	0,086	0,655	0,914	0,158	0,884
Feature selection	No balancing	0,913	0,914	0,913	0,914	0,914	0,914
	Oversampling	0,654	0,651	0,653	0,902	0,674	0,653
	SMOTE	0,732	0,758	0,787	0,910	0,790	0,712
	ROSE	0,653	0,907	0,675	0,914	0,911	0,907

Valor de *Kappa*							
		Log Reg	CART	K-NN	NB	RF	SVM
Com *feature selection*	No balancing	0,045	0,000	0,008	0,000	0,000	0,000
	Oversampling	0,206	0,163	0,130	0,000	0,101	0,141
	SMOTE	0,211	0,132	0,165	0,000	0,027	0,153
	ROSE	0,204	0,000	0,056	0,000	0,011	0,154
Feature selection	No balancing	0,003	0,000	0,005	0,000	0,012	0,002
	Oversampling	0,152	0,151	0,129	0,064	0,135	0,157
	SMOTE	0,182	0,173	0,154	0,056	0,145	0,175
	ROSE	0,152	0,067	0,095	0,019	0,053	0,069

Exatidão balanceada							
		Log Reg	CART	K-NN	NB	RF	SVM
Sem *feature selection*	No balancing	0,513	0,500	0,502	0,500	0,500	0,500
	Oversampling	0,733	0,711	0,637	0,500	0,546	0,693
	SMOTE	0,669	0,550	0,572	0,500	0,508	0,565
	ROSE	0,734	0,500	0,565	0,500	0,530	0,567
Feature selection	No balancing	0,501	0,500	0,501	0,500	0,503	0,501
	Oversampling	0,697	0,698	0,663	0,521	0,662	0,706
	SMOTE	0,689	0,661	0,624	0,517	0,614	0,695
	ROSE	0,698	0,521	0,609	0,505	0,516	0,522

Anexo II Tabelas dos resultados gerais dos modelos de previsão da mortalidade, precisão, sensibilidade, especificidade, exatidão, valor de kappa e exatidão balanceada.

		LG			CART			K-NN		
Sem *feature selection*	Sem balanceamento	Referência			Referência			Referência		
		Previsão	0	1	Previsão	0	1	Previsão	0	1
		0	6798	622	0	6824	641	0	6802	636
		1	26	19	1	0	0	1	22	5

		LG			CART			K-NN		
Sem *feature selection*	*Oversampling*	Referência			Referência			Referência		
		Previsão	0	1	Previsão	0	1	Previsão	0	1
		0	4882	160	0	4416	144	0	4969	291
		1	1942	481	1	2408	497	1	1855	350

		LG			CART			K-NN		
Sem *feature selection*	SMOTE	Referência			Referência			Referência		
		Previsão	0	1	Previsão	0	1	Previsão	0	1
		0	5658	315	0	6610	557	0	6482	516
		1	1166	326	1	214	84	1	342	125

		LG			CART			K-NN		
Sem *feature selection*	ROSE	Referência			Referência			Referência		
		Previsão	0	1	Previsão	0	1	Previsão	0	1
		0	4828	153	0	0	0	0	4597	348
		1	1996	488	1	6824	641	1	2227	293

		LG			CART			K-NN		
Com *feature selection*	Sem balanceamento	Referência			Referência			Referência		
		Previsão	0	1	Previsão	0	1	Previsão	0	1
		0	6814	639	0	6824	641	0	6811	638
		1	10	2	1	0	0	1	13	3

		LG			CART			K-NN		
Com *feature selection*	*Oversampling*	Referência			Referência			Referência		
		Previsão	0	1	Previsão	0	1	Previsão	0	1
		0	4398	160	0	4374	157	0	4446	209
		1	2426	481	1	2450	484	1	2378	432

		LG			CART			K-NN		
Com *feature selection*	SMOTE	Referência			Referência			Referência		
		Previsão	0	1	Previsão	0	1	Previsão	0	1
		0	5054	232	0	5312	292	0	5597	367
		1	1770	409	1	1512	349	1	1227	274

		LG			CART			K-NN		
Com *feature selection*	ROSE	Referência			Referência			Referência		
		Previsão	0	1	Previsão	0	1	Previsão	0	1
		0	4391	159	0	6732	605	0	4703	302
		1	2433	482	1	92	36	1	2121	339

		NB			RF			SVM		
Sem *feature selection*	Sem balanceamento	Referência			Referência			Referência		
		Previsão	0	1	Previsão	0	1	Previsão	0	1
		0	6824	641	0	6824	641	0	6824	641
		1	0	0	1	0	0	1	0	0

		NB			RF			SVM		
Sem *feature selection*	*Oversampling*	Referência			Referência			Referência		
		Previsão	0	1	Previsão	0	1	Previsão	0	1
		0	6824	641	0	4217	149	0	6401	542
		1	0	0	1	2607	492	1	423	99

		NB			RF			SVM		
Sem *feature selection*	SMOTE	Referência			Referência			Referência		
		Previsão	0	1	Previsão	0	1	Previsão	0	1
		0	6824	641	0	6812	630	0	6501	527
		1	0	0	1	12	11	1	323	114

		NB			RF			SVM		
Sem *feature selection*	ROSE	Referência			Referência			Referência		
		Previsão	0	1	Previsão	0	1	Previsão	0	1
		0	6824	641	0	550	13	0	6484	523
		1	0	0	1	6274	628	1	340	118

		NB			RF			SVM		
Com *feature selection*	Sem balanceamento	Referência			Referência			Referência		
		Previsão	0	1	Previsão	0	1	Previsão	0	1
		0	6817	641	0	6817	636	0	6822	640
		1	0	0	1	7	5	1	2	1

		NB			RF			SVM		
Com *feature selection*	*Oversampling*	Referência			Referência			Referência		
		Previsão	0	1	Previsão	0	1	Previsão	0	1
		0	6691	601	0	4619	226	0	4380	147
		1	133	40	1	2205	415	1	2444	494

		NB			RF			SVM		
Com *feature selection*	SMOTE	Referência			Referência			Referência		
		Previsão	0	1	Previsão	0	1	Previsão	0	1
		0	6769	614	0	5641	384	0	4885	209
		1	55	27	1	1183	257	1	1939	432

		NB			RF			SVM		
Com *feature selection*	ROSE	Referência			Referência			Referência		
		Previsão	0	1	Previsão	0	1	Previsão	0	1
		0	6812	633	0	6776	616	0	6732	640
		1	12	8	1	48	25	1	92	37

Anexo III Tabelas com os resultados gerais da técnica de Poisson após aplicação das técnicas de feature selection backward, forward e stepwise

Backward	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	1.459e+00	4.915e-02	29.682	< 2e-16 ***
AGE	7.241e-03	1.878e-04	38.564	< 2e-16 ***
urg1	1.474e-01	1.135e-02	12.987	< 2e-16 ***
AMIvolume	5.020e-04	1.628e-05	30.838	< 2e-16 ***
beds	-1.911e-04	7.259e-06	-26.323	< 2e-16 ***
NUT2	-1.597e-02	2.296e-03	-6.958	3.45e-12 ***
DIAG141001	-2.081e-01	4.758e-02	-4.374	1.22e-05 ***
DIAG141010	-3.995e-01	5.743e-02	-6.956	3.49e-12 ***
DIAG141011	-2.824e-01	4.655e-02	-6.068	1.30e-09 ***
DIAG141020	5.354e-02	9.552e-02	0.561	0.575122
DIAG141021	-3.319e-01	4.849e-02	-6.844	7.68e-12 ***
DIAG141030	-1.167e+00	2.402e-01	-4.859	1.18e-06 ***
DIAG141031	-3.665e-01	4.960e-02	-7.388	1.50e-13 ***
DIAG141040	-4.548e-01	7.770e-02	-5.853	4.83e-09 ***
DIAG141041	-3.337e-01	4.662e-02	-7.157	8.24e-13 ***
DIAG141050	-4.550e-01	1.885e-01	-2.414	0.015761 *
DIAG141051	-4.106e-01	5.025e-02	-8.172	3.04e-16 ***
DIAG141060	-1.006e+00	5.023e-01	-2.003	0.045214 *
DIAG141061	-4.655e-01	6.286e-02	-7.406	1.30e-13 ***
DIAG141070	-4.525e-01	4.939e-02	-9.163	< 2e-16 ***
DIAG141071	-3.320e-01	4.623e-02	-7.182	6.85e-13 ***
DIAG141080	-1.730e-01	7.040e-02	-2.457	0.013993 *
DIAG141081	-3.217e-01	4.758e-02	-6.760	1.38e-11 ***
DIAG141090	-3.735e-01	5.956e-02	-6.270	3.61e-10 ***
DIAG141091	-3.958e-01	4.777e-02	-8.286	< 2e-16 ***
ami1	-1.747e-01	9.389e-03	-18.607	< 2e-16 ***
chf1	1.430e-01	8.958e-03	15.960	< 2e-16 ***
pvd1	-5.507e-02	1.188e-02	-4.635	3.57e-06 ***
cevd1	1.063e-01	1.018e-02	10.445	< 2e-16 ***
dementia1	-2.067e-01	1.687e-02	-12.252	< 2e-16 ***
rheumd1	-1.310e-01	2.453e-02	-5.341	9.22e-08 ***
pud1	2.292e-01	2.410e-02	9.509	< 2e-16 ***
mld1	8.145e-02	1.615e-02	5.044	4.56e-07 ***
diab1	-8.775e-02	7.917e-03	-11.084	< 2e-16 ***
hp1	4.141e-01	2.269e-02	18.249	< 2e-16 ***
rend1	2.714e-02	1.093e-02	2.482	0.013068 *
canc1	-4.933e-02	1.697e-02	-2.907	0.003650 **
msld1	3.463e-01	4.286e-02	8.080	6.49e-16 ***
aids1	-1.452e-01	5.485e-02	-2.648	0.008109 **
score	1.519e-01	1.057e-02	14.372	< 2e-16 ***
wscore	2.060e-02	6.212e-03	3.316	0.000913 ***

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

Forward	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.457e+00	4.914e-02	29.650	< 2e-16 ***
AGE	7.273e-03	1.871e-04	38.871	< 2e-16 ***
urg1	1.473e-01	1.135e-02	12.982	< 2e-16 ***
AMIvolume	5.016e-04	1.628e-05	30.814	< 2e-16 ***
beds	-1.911e-04	7.259e-06	-26.330	< 2e-16 ***
NUT2	-1.604e-02	2.296e-03	-6.985	2.84e-12 ***
DIAG141001	-2.084e-01	4.758e-02	-4.379	1.19e-05 ***
DIAG141010	-3.976e-01	5.743e-02	-6.922	4.44e-12 ***
DIAG141011	-2.826e-01	4.655e-02	-6.072	1.27e-09 ***
DIAG141020	5.154e-02	9.553e-02	0.539	0.58954
DIAG141021	-3.320e-01	4.849e-02	-6.847	7.53e-12 ***
DIAG141030	-1.163e+00	2.402e-01	-4.841	1.29e-06 ***
DIAG141031	-3.665e-01	4.960e-02	-7.389	1.48e-13 ***
DIAG141040	-4.552e-01	7.770e-02	-5.859	4.67e-09 ***
DIAG141041	-3.340e-01	4.662e-02	-7.163	7.91e-13 ***
DIAG141050	-4.546e-01	1.885e-01	-2.412	0.01586 *
DIAG141051	-4.110e-01	5.025e-02	-8.178	2.88e-16 ***
DIAG141060	-1.011e+00	5.023e-01	-2.012	0.04425 *
DIAG141061	-4.655e-01	6.286e-02	-7.405	1.32e-13 ***
DIAG141070	-4.517e-01	4.939e-02	-9.146	< 2e-16 ***
DIAG141071	-3.322e-01	4.623e-02	-7.186	6.66e-13 ***
DIAG141080	-1.735e-01	7.040e-02	-2.464	0.01372 *
DIAG141081	-3.222e-01	4.758e-02	-6.771	1.28e-11 ***
DIAG141090	-3.735e-01	5.956e-02	-6.270	3.61e-10 ***
DIAG141091	-3.962e-01	4.777e-02	-8.293	< 2e-16 ***
ami1	-2.144e-01	8.283e-03	-25.885	< 2e-16 ***
chf1	1.031e-01	7.716e-03	13.368	< 2e-16 ***
pvd1	-9.689e-02	1.110e-02	-8.726	< 2e-16 ***
cevd1	6.583e-02	9.271e-03	7.101	1.24e-12 ***
dementia1	-2.473e-01	1.630e-02	-15.166	< 2e-16 ***
rheumd1	-1.711e-01	2.402e-02	-7.123	1.06e-12 ***
pud1	1.890e-01	2.365e-02	7.992	1.33e-15 ***
mld1	3.768e-01	4.142e-02	9.096	< 2e-16 ***
diab1	-1.258e-01	6.532e-03	-19.261	< 2e-16 ***
hp1	3.953e-01	2.209e-02	17.899	< 2e-16 ***
metacanc1	6.258e-02	3.030e-02	2.065	0.03890 *
canc1	-6.514e-02	1.530e-02	-4.259	2.06e-05 ***
mld1	4.305e-02	1.539e-02	2.798	0.00513 **
aids1	-8.216e-02	4.676e-02	-1.757	0.07891 .
score	2.127e-01	4.702e-03	45.226	< 2e-16 ***
copd1	-4.346e-02	8.950e-03	-4.856	1.20e-06 ***

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

Stepwise	Estimativa	Erro padrão	Valor Z	Pr (>|z|)
(Intercept)	1.459e+00	4.915e-02	29.682	< 2e-16 ***
AGE	7.241e-03	1.878e-04	38.564	< 2e-16 ***
urg1	1.474e-01	1.135e-02	12.987	< 2e-16 ***
AMIvolume	5.020e-04	1.628e-05	30.838	< 2e-16 ***
beds	-1.911e-04	7.259e-06	-26.323	< 2e-16 ***
NUT2	-1.597e-02	2.296e-03	-6.958	3.45e-12***
DIAG141001	-2.081e-01	4.758e-02	-4.374	1.22e-05***
DIAG141010	-3.995e-01	5.743e-02	-6.956	3.49e-12***
DIAG141011	-2.824e-01	4.655e-02	-6.068	1.30e-09***
DIAG141020	5.354e-02	9.552e-02	0.561	0.575122
DIAG141021	-3.319e-01	4.849e-02	-6.844	7.68e-12***
DIAG141030	-1.167e+00	2.402e-01	-4.859	1.18e-06***

DIAG141031	-3.665e-01	4.960e-02	-7.388	1.50e-13***
DIAG141040	-4.548e-01	7.770e-02	-5.853	4.83e-09***
DIAG141041	-3.337e-01	4.662e-02	-7.157	8.24e-13***
DIAG141050	-4.550e-01	1.885e-01	-2.414	0.015761 *
DIAG141051	-4.106e-01	5.025e-02	-8.172	3.04e-16***
DIAG141060	-1.006e+00	5.023e-01	-2.003	0.045214 *
DIAG141061	-4.655e-01	6.286e-02	-7.406	1.30e-13***
DIAG141070	-4.525e-01	4.939e-02	-9.163	< 2e-16 ***
DIAG141071	-3.320e-01	4.623e-02	-7.182	6.85e-13***
DIAG141080	-1.730e-01	7.040e-02	-2.457	0.013993 *
DIAG141081	-3.217e-01	4.758e-02	-6.760	1.38e-11***
DIAG141090	-3.735e-01	5.956e-02	-6.270	3.61e-10***
DIAG141091	-3.958e-01	4.777e-02	-8.286	< 2e-16 ***
ami1	-1.747e-01	9.389e-03	-18.607	< 2e-16 ***
chf1	1.430e-01	8.958e-03	15.960	< 2e-16 ***
pvd1	-5.507e-02	1.188e-02	-4.635	3.57e-06***
cevd1	1.063e-01	1.018e-02	10.445	< 2e-16 ***
dementia1	-2.067e-01	1.687e-02	-12.252	< 2e-16 ***
rheumd1	-1.310e-01	2.453e-02	-5.341	9.22e-08***
pud1	2.292e-01	2.410e-02	9.509	< 2e-16 ***
mld1	8.145e-02	1.615e-02	5.044	4.56e-07***
diab1	-8.775e-02	7.917e-03	-11.084	< 2e-16 ***
hp1	4.141e-01	2.269e-02	18.249	< 2e-16 ***
rend1	2.714e-02	1.093e-02	2.482	0.013068 *
canc1	-4.933e-02	1.697e-02	-2.907	0.003650 **
msld1	3.463e-01	4.286e-02	8.080	6.49e-16***
metacanc1	-1.904e-01	3.718e-02	-5.121	3.04e-07***
aids1	-1.452e-01	5.485e-02	-2.648	0.008109 **
score	1.519e-01	1.057e-02	14.372	< 2e-16 ***
wscore	2.060e-02	6.212e-03	3.316	0.000913***

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

SEDE ADMINISTRATIVA

FACULDADE DE **MEDICINA**

FACULDADE DE **CIÊNCIAS**