ANASTASIYA
DANYLYAK

# Previsão da Admissão de doentes na Unidade de Cuidados Intensivos através de Modelos de *Machine Learning*

# Prediction of Intensive Care Unit Admission using Machine Learning Models

**Universidade de Aveiro**
**2023**

**ANASTASIYA DANYLYAK**

# Previsão da Admissão de doentes na Unidade de Cuidados Intensivos através de Modelos de *Machine Learning*

# Prediction of Intensive Care Unit Admission using Machine Learning Models

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Estatística Médica , realizada sob a orientação científica do Doutor Luís Miguel Almeida da Silva, Professor auxiliar do Departamento de Matemática da Universidade de Aveiro, e do Doutor Bernardo Marques, Data Scientist na Prológica

.

**o júri / the jury**

presidente / president

Prof. Doutora Vera Mónica Almeida Afreixo
professora associada do Departamento de Matemática da Universidade de Aveiro

vogais / examiners committee

Prof. Doutora Ana Helena Marques de Pinho Tavares
professora adjunta convidada da Escola Superior de Tecnologia e Gestão de Águeda

Prof. Doutor Luís Miguel Almeida da Silva
professor auxiliar do Departamento de Matemática da Universidade de Aveiro

**agradecimentos / acknowledgements**

A escrita deste relatório de estágio e a sua conclusão representam momentos importantes no meu percurso académico. Houve altos e baixos ao longo desse percurso, e a superação destes desafios contou não apenas com o meu esforço, mas também com o apoio incansável de várias pessoas que desempenharam papéis essenciais durante estes meses.

Em primeiro lugar, gostaria de expressar um agradecimento especial ao meu orientador, Professor Luís Silva, por todo o apoio, paciência, total disponibilidade e pelo exemplar profissionalismo demonstrado na orientação deste projeto. Também desejo estender o meu agradecimento à Prologica, ao Bernardo e ao Heitor, que estiveram sempre disponíveis durante o período de orientação do estágio. Não posso deixar de expressar um especial obrigada à Professora Vera Afreixo, que desempenhou um papel fundamental não apenas para me ajudar a não desistir nesta fase, mas também por nos acompanhar de maneira incrível ao longo de todo o meu percurso no mestrado.

À minha mãe, ao Daniel, à Rita, aos meus avós e ao senhor Joaquim, expresso a minha gratidão por todo o apoio em todas as situações ao longo da minha jornada académica e não só.

Aos meus amigos, Joana, Diogo, Didi, Bernardo, Beatriz, Lénia, Benedito e Marcos, que estiveram ao meu lado desde o início e nunca me abandonaram, quero agradecer calorosamente por todos os momentos memoráveis que compartilhámos. Estes momentos foram também uma fonte de inspiração e força que me ajudaram a concluir este relatório com sucesso. Um grande obrigado a todos os outros que me acompanharam nesta etapa.

Por fim, mas não menos importante, deixo o meu agradecimento à Universidade de Aveiro e ao Departamento de Matemática por proporcionarem uma instituição fantástica que me permitiu crescer tanto pessoalmente quanto academicamente. Estou grata por todas as oportunidades e aprendizagens que recebi durante esta minha jornada nesta academia.

**Palavras Chave**

Machine Learning, Unidade de Cuidados Intensivos, Classificação, Random Forest, Regressão Logística, Previsão

**Resumo**

Nos últimos anos temos testemunhado avanços significativos no campo da medicina, impulsionados principalmente pelo desenvolvimento e incorporação de novas tecnologias. Estas inovações estão a revolucionar a forma como os profissionais de saúde diagnosticam e tratam as doenças, proporcionando resultados mais precisos e eficientes. No entanto, com o aumento exponencial do volume de dados gerados diariamente, tornou-se fundamental desenvolver novas técnicas e ferramentas que serão capazes de lidar com essa imensa quantidade de informação. Neste contexto, com o desenvolvimento das novas tecnologias surgem novas técnicas que nos podem ajudar a facilitar a automatização do processo de gestão de dados na saúde, e, consequentemente, melhorar os serviços prestados aos doentes e também ajudar os profissionais de saúde nas suas atividades diárias. Uma das áreas que necessita deste auxílio é a Medicina Intensiva, que tem vindo a ser desenvolvida de forma a ajudar a salvar vidas de doentes que se encontram em risco. Assim, torna-se evidente que há necessidade de prever as admissões à UCI, visto que para além de constituírem custos adicionais para as instituições e ocuparem recursos desnecessários, as admissões não planeadas são arriscadas para os doentes que se encontram debilitados.

Neste contexto, após o pré-processamento de dados de uma instituição de saúde portuguesa e uma análise exploratória, foram usados dois modelos de *Machine Learning* de classificação, o Random Forest e a Regressão Logística, com o objetivo de prever a probabilidade de um doente ser admitido na Unidade de Cuidados Intensivos, onde a variável a ser prevista é categórica que indica se o doente foi ou não à UCI.

Como conclusão foi possível analisar que variáveis como a idade, a proveniência dos doentes e as medições dos sinais vitais desempenham um papel significativo e importante na capacidade de prever a admissão dos doentes na Unidade de Cuidados Intensivos. Estas conclusões proporcionam informações que podem ser importantes para os profissionais de saúde, destacando a relevância destas variáveis no processo de triagem e na identificação de casos que requerem encaminhamento para a UCI.

**Keywords** Machine Learning, Intensive Care Unit, Classification, Random Forest, Logistic Regression, Prevision

**Abstract**

In recent years we have witnessed significant advances in the field of medicine, driven mainly by the development and incorporation of new technologies. These innovations are revolutionizing the way healthcare professionals diagnose and treat diseases, providing more accurate and efficient results. However, with the exponential increase in the volume of data generated daily, it has become essential to develop new techniques and tools that will be capable of dealing with this immense amount of information. In this context, with the development of new technologies, new techniques emerge that can help us facilitate the automation of the healthcare data management process, and, consequently, improve the services provided to patients and also help healthcare professionals in their daily activities. One of the areas that needs this help is Intensive Medicine, which has been developed to help save the lives of patients who are at risk. Therefore, it becomes clear that there is a need to predict admissions to the ICU, since in addition to constituting additional costs for institutions and taking up unnecessary resources, unplanned admissions are risky for patients who are weakened.

In this context, after pre-processing data from a Portuguese healthcare institution and an exploratory analysis, two classification *Machine Learning* models were used, Random Forest and Logistic Regression, to predict the probability of a patient being admitted to the Intensive Care Unit, where the variable to be predicted is categorical and indicates whether or not the patient went to the ICU.

In conclusion, it was possible to analyze that variables such as age, patient origin, and vital sign measurements play a significant role in the ability to predict patient admission to the Intensive Care Unit. These conclusions provide information that may be important for health professionals, highlighting the relevance of these variables in the screening process and in identifying cases that require referral to the ICU.

# Conteúdo

# Lista de Figuras

# Lista de Tabelas

CAPÍTULO 1

# Introdução

## 1.1 Contextualização

Nos últimos anos temos testemunhado avanços significativos no campo da medicina, impulsionados principalmente pelo desenvolvimento e incorporação de novas tecnologias. Essas inovações estão a revolucionar a forma como os profissionais de saúde diagnosticam e tratam as doenças, proporcionando resultados mais precisos e eficientes. No entanto, com o aumento exponencial do volume de dados gerados diariamente, tornou-se crucial desenvolver novas técnicas e ferramentas que serão capazes de lidar com essa imensa quantidade de informação. Neste contexto, com o desenvolvimento das novas tecnologias surgem novas técnicas que nos podem ajudar a facilitar a automatização do processo de gestão de dados na saúde, e, consequentemente, melhorar os serviços prestados aos doentes e também auxiliar os profissionais de saúde nas suas atividades diárias. [6]

Uma das áreas que necessita deste auxílio é a Medicina Intensiva, que tem vindo a ser desenvolvida de forma a ajudar a salvar vidas de doentes que se encontram em risco. Aqui, é usado um conjunto de técnicas e procedimentos para a monitorização, diagnóstico e tratamento de doentes, 24 horas por dia. É uma área que engloba as várias áreas da medicina desde a cardiologia à cirurgia, estando focada no tratamento de falência dos órgãos das funções vitais de um doente. Para este auxílio ser prestado surgiram as Unidades de Cuidados Intensivos (UCI), que são locais preparados e equipados com todos os recursos necessários para tratamento dos doentes críticos. [1]

Os gastos em saúde relacionados com doentes internados em Unidades de Cuidados Intensivos (UCI) são uma das principais preocupações das administrações hospitalares. As UCI são serviços que regularmente assumem uma ocupação de 100% das camas disponíveis e, naturalmente, os serviços hospitalares não transferem doentes para as unidades de cuidados intensivos sem que exista uma razão que o justifique. Contudo, a previsão dos doentes com elevado risco de complicações e, portanto, os mesmos serem transferidos para uma UCI poderá permitir uma adaptação na prestação de cuidados que possam evitar a sua transferência. Assim, torna-se evidente que há necessidade de prever as admissões à UCI, visto que para

além de constituírem custos elevados nas instituições e existir a possibilidade de doentes estarem a ocupar recursos desnecessários, as admissões não planeadas são arriscadas para os doentes que se encontram debilitados. As técnicas de *Machine Learning* são uma das opções a ser utilizadas para a resolução deste tipo de problemas. A ideia é antecipar os eventos críticos num doente de forma a prever as admissões à UCI, ou de alguma forma caracterizar os doentes que poderão ser internados nos cuidados intensivos. Existem alguns modelos já criados que permitem agrupar os doentes de acordo com as características que apresentam, para posteriormente serem admitidos na UCI. Desta forma, os médicos poderão verificar se os doentes se inserem em algum desses grupos e terem as noções necessárias sobre o estado do doente para implementar um tratamento. Ter estes modelos é uma vantagem para o planeamento e alocação eficiente de recursos, e ainda, permite que os médicos, doentes e familiares possam estar cientes dos riscos associados às condições de saúde dos doentes. [6]

Neste desafio, pretende-se estudar e desenvolver metodologias de *Machine Learning* para a previsão de risco de transferência para a UCI, garantindo que os profissionais responsáveis são notificados sobre os doentes em situações de risco. Porque outro dos problemas associados ao internamento em UCI é o tempo. Para este desafio serão considerados dados reais de uma instituição hospitalar portuguesa.

Este relatório foi realizado no âmbito do 2º ano curricular do Mestrado de Estatística Médica da Universidade de Aveiro. Este mestrado além da componente teórico-prática, proporciona aos seus estudantes a opção de realizar um estágio curricular que permite aplicar num contexto empresarial os conhecimentos adquiridos ao longo da formação. A escolha de realizar um estágio curricular partiu do interesse de aplicar todos os conhecimentos que foram adquiridos ao longo destes 2 anos em contexto empresarial, e adquirir novos conhecimentos que sirvam como preparação para o mercado de trabalho. O estágio teve início no dia 23 de janeiro de 2023 e terminou no dia 19 de maio de 2023, sob orientação científica do Professor Luís Silva e do Bernardo Marques *Data Scientist* da Prológica.

## 1.2 Objetivos

O principal objetivo deste relatório é desenvolver metodologias de *Machine Learning* para a previsão de risco de transferência para a UCI, de forma a garantir que os profissionais tomem decisões acertadas no decorrer do tratamento dos doentes e que estes mesmos profissionais sejam notificados sempre que um dos seus doentes se encontre em risco.

O objetivo do estágio, para além de desenvolver estas metodologias aplicando todos os conhecimentos adquiridos ao longo da formação é a integração no meio empresarial e aquisição de novas competências de trabalho que em fusão com as competências que já possuímos nos facilitem no mercado de trabalho.

## 1.3 Entidade Acolhedora

A Prologica foi fundada em 1984, e é uma empresa portuguesa que cria e implementa soluções nas áreas da Tecnologias da Informação e Comunicação baseadas em dados da saúde para apoiar nas tomadas de decisão.

O seu propósito é facilitar as organizações de saúde em todo o processo de apoio à gestão e atividade clínica, desenvolvendo novas tecnologias para as mesmas depois puderem implementar no seu dia a dia.

A empresa acredita firmemente que ter uma plataforma de dados bem estruturada para recolher, monitorizar e analisar resultados e custos é fundamental para que os dados se encontrem organizados, não só para futuramente ser mais fácil desenvolver técnicas que facilitem nos processos clínicos mas também para que as organizações de saúde tenham os dados ordenados.

## 1.4 Ferramentas utilizadas

As ferramentas utilizadas foram o *Microsoft Excel* para a organização de uma parte das tabelas e o R para preparação dos dados e para a aplicação das técnicas de ML.

O *Microsoft Excel* é um programa integrado na família de produtos de *software* da *Microsoft 365* e é o principal programa baseado em folhas de cálculo com diversas funcionalidades ao nível da análise e visualização de dados. Neste relatório foi útil para uma análise inicial dos dados, visto que permite uma visualização mais simplificada devido à sua interface intuitiva.

O *software R* é uma ferramenta utilizada para análise de dados, permitindo importar os mesmos e transformá-los, explorá-los, criar gráficos e tabelas e também modelar. Também é uma ferramenta com a qual é possível fazer previsões sobre os dados através de modelos de ML. Neste relatório o R foi utilizado para o pré-processamento dos dados, para análise da base de dados e para a construção dos modelos de *Machine Learning*.

O R é uma linguagem de programação para ciências de dados que possui fortes recursos para visualização e exploração de dados e também para aplicar modelos de ML e avaliar os seus *outputs*.

CAPÍTULO 2

# Revisão de Literatura

## 2.1 Unidade de Cuidados Intensivos

A Unidade de Cuidados Intensivos (UCI) é uma componente essencial e insubstituível do sistema de saúde, que desempenha um papel importante no tratamento de doentes que se encontram em estado grave e necessitam de cuidados médicos intensivos e vigilância constante. Além disso, as UCI's desempenham um papel fundamental no suporte dos doentes que tenham passado por procedimentos cirúrgicos complexos, onde os cuidados intensivos são essenciais para otimizar a sua recuperação. Ao longo do tempo a evolução das Unidades de Cuidados Intensivos é uma história notável de avanços médicos, inovação tecnológica e dedicação incansável da equipa médica e de enfermagem. Estas unidades têm evoluído constantemente para atender às crescentes necessidades da medicina moderna. Com o aparecimento das novas tecnologias existe a possibilidade da monitorização constante dos sinais vitais e o acesso a equipamentos médicos avançados que são capazes de oferecer um tratamento personalizado e intensivo aos doentes em estado crítico. Além disso, a história das UCI's também reflete o compromisso inabalável dos profissionais de saúde, que trabalham incansavelmente para proporcionar o mais alto nível de cuidados e suporte aos doentes e às suas famílias. A equipa das UCI é composta por médicos especializados, enfermeiros, técnicos e outros profissionais de saúde, cuja dedicação desempenha um papel vital na recuperação e no bem-estar dos doentes. [1]

As Unidades de Cuidados Intensivos desempenham um papel insubstituível na medicina moderna, garantindo que os doentes em estado grave recebam o tratamento intensivo necessário para a sua recuperação. A evolução contínua das UCI's, juntamente com a combinação da inovação tecnológica e com os cuidados humanos, demonstra um compromisso da comunidade médica em proporcionar os mais altos padrões de cuidados de saúde. [2]

A história da Unidade de Cuidados Intensivos remonta para meados do século XX, quando a medicina estava a começar a perceber melhor quais as necessidades dos pacientes que se encontravam em estado grave. Antes desse período, os pacientes com doenças graves tinham opções limitadas de tratamento e muitas vezes enfrentavam resultados fatais. Os médicos

perceberam que, para aumentar as probabilidades de sobrevivência, era necessário fornecer cuidados mais intensivos e especializados. [1]

Foi então, que durante a epidemia da poliomielite na década de 1950 que começaram a surgir doentes com insuficiência respiratória aguda e tornou-se evidente que seria necessário existir um espaço onde os mesmos poderiam ter os cuidados que necessitavam. [2]Nesse momento crítico, os ventiladores mecânicos e outras tecnologias de suporte à vida começaram a ser desenvolvidos e aprimorados para fazer os tratamentos que os doentes necessitavam. Esta fase foi marcada como um início de uma abordagem moderna da UCI. À medida que a compreensão médica e a tecnologia avançaram, as Unidades de Cuidados Intensivos foram-se desenvolvendo rapidamente. Atualmente, essas unidades são encontradas em hospitais por todo o mundo e contam com profissionais de saúde especializados. As UCI's contêm equipamentos sofisticados que garantem uma correta vigilância e tratamento, incluindo ventiladores avançados, bombas de infusão e sistemas de suporte à vida que permitem fazer um controlo preciso das funções vitais de cada paciente. Além disso, estas unidades têm protocolos rigorosos para garantir que os pacientes recebem a atenção mais adequada e personalizada possível. [1]

As técnicas de *Machine Learning* (ML) nas UCI's são cada vez mais implementadas tanto no diagnóstico como na previsão de resultados. [3]

## 2.2 BioSign

O BioSign é um sistema automatizado que monitoriza em tempo real o estado de saúde dos pacientes, combinando dados vitais obtidos através de monitores instalados na enfermaria. Este sistema recolhe informações cruciais, como a frequência cardíaca, a frequência respiratória, a saturação arterial de oxigénio (SAO2), a temperatura da pele e a pressão arterial. A partir destes cinco sinais vitais, é gerada uma representação numérica conhecida como Índice do Estado do Paciente (PSI). Este índice atua como um alerta, sendo capaz de acionar a intervenção de uma equipa médica especializada se for detetada a necessidade de cuidados adicionais por parte do paciente. Este sistema representa um modelo probabilístico de normalidade com cinco dimensões, anteriormente aprendido a partir dos sinais vitais recolhidos de uma amostra de pacientes, que designamos como conjunto de treino. O seu propósito é treinar os dados, conforme sugere o próprio nome, a fim de desenvolver um modelo. Este modelo de normalidade é armazenado no sistema e utilizado para avaliar os sinais vitais do paciente em análise, determinando se estes se enquadram nos padrões considerados normais em relação ao conjunto de treino. Sempre que esses sinais vitais se desviam do que é considerado normal, o valor do Índice do Estado do Paciente (PSI) ultrapassa o limite predefinido, desencadeando a emissão de um alerta. [6].

**BioSign – Construção do Modelo**

Foram recolhidos dados dos sinais vitais de 150 pacientes da enfermaria geral do Hospital *John Radcliffe* em Oxford, entre 2001 e 2003. Esses pacientes foram ligados a um monitor multiparamétrico, em média, 24 horas por paciente [6].

Os pacientes foram selecionados dos seguintes grupos de pacientes de "alto risco":

- pacientes que foram analisados por, pelo menos, 24 horas após um enfarte do miocárdio e novamente algumas horas horas 5 dias depois;
- pacientes com insuficiência cardíaca grave;
- pacientes com problemas respiratórios agudos (por exemplo, asma aguda ou pneumonia);
- pacientes idosos com fratura do quadril, foram analisados antes e depois da cirurgia.

O modelo de normalidade *BioSign* é a função de densidade de probabilidade incondicional (fdp), $\hat{p}(x)$, dos dados do conjunto de treino, onde $\mathbf{x} = (x_1, x_2, \ldots, x_5)$ é o vetor de parâmetros dos sinais vitais, com $x_1$= frequência cardíaca, $x_2$=frequência respiratória, etc. Como os cinco parâmetros têm escalas diferentes (um aumento de 0,5ºC na temperatura é muito mais significativo do que um aumento de 0,5 mm Hg na pressão arterial), os dados precisam de ser normalizados antes que possam formar o vetor $\mathbf{x}$. Para essa normalização foi usada uma transformação padrão de média zero e variância 1 [6]. Em primeiro lugar, o algoritmo de *clustering k-means* foi usado para selecionar os centros dos *clusters* dos dados normalizados do conjunto de treino. Cada um dos centros $x_j$ será então um *kernel* no método *Parzen Win-*

*dows*,que permite obter a estimativa contínua da função densidade, dado pela seguinte equação:

$$\hat{p}(x) = \frac{1}{N(2\pi)^{\frac{d}{2}}\sigma^d} \sum_{j=1}^{N} exp\left(\frac{-\|x - x_j\|^2}{2\sigma^2}\right)$$

Foi definido um "PSI"para que os alertas possam ser gerados quando esse índice aumentar acima de um valor limite. O PSI foi calculado através de uma fórmula específica, e sempre que esse índice no paciente ultrapassa o valor estabelecido como limite, um alerta é acionado, resultando na chamada da equipa médica [6].

$$\text{Patient Status Index} = log_e\left[\frac{1}{\hat{p}(x)}\right]$$

Recentemente, o sistema BioSign foi aplicado em diversos estudos clínicos na Europa e nos Estados Unidos, com o objetivo de analisar os dados dos sinais vitais de pacientes submetidos a cirurgias de alto risco ou após internamento de emergência devido a condições agudas não cirúrgicas [6]. De certa forma, o *BioSign* foi usado para criar alertas para que equipas médicas especializadas fossem chamadas quando os sinais vitais do paciente estivessem acima de uma média definida e o mesmo necessitasse de apoio médico [6].

**Sistemas circEWS e circEWS-lite**

Foi criada uma nova estratégia para identificar a insuficiência circulatória nos pacientes das Unidades de Cuidados Intensivos (UCI), utilizando conhecimento médico, análise extensiva de dados e técnicas de Machine Learning. Desenvolveram-se dois sistemas de alerta precoce, denominados circEWS e circEWS-lite, ambos com diferentes níveis de complexidade, destinados a notificar os médicos sobre pacientes em risco de insuficiência circulatória nas próximas 8 horas. Implementou-se uma estrutura abrangente de análise, que engloba desde o pré-processamento e limpeza de dados até a extração e interpretação de características, assim como a escolha de técnicas supervisionadas de Machine Learning em grande escala para construir os sistemas de alerta precoce. Para avaliar a eficácia desses sistemas, estabeleceu-se uma métrica de avaliação baseada em eventos de alarme, que avalia a fração de eventos de insuficiência circulatória corretamente previstos (ou seja, quando um alarme foi ativado para tal evento) e a taxa de alarmes falsos (ou seja, quando um alarme foi ativado, mas nenhum evento ocorreu). [17]

Como linha de base, foi criada uma árvore de decisão utilizando variáveis relacionadas com a definição de insuficiência circulatória. Esta abordagem resultou na construção de um sistema de regras com base em limiares. A análise das áreas sob as curvas dos modelos foi realizada para avaliar o desempenho. Nos modelos criados são geradas pontuações de previsão contínua a cada 5 minutos em relação ao risco de insuficiência circulatória nas próximas 8 horas. Um sistema de alerta baseado em limiares derivados dessa pontuação pode levar a alarmes a cada 5 minutos, causando fadiga do alarme. Portanto desenvolveu-se um sistema de alarme que implementou uma política de silenciamento: uma vez que o alarme é acionado os

alarmes subsequentes serão suprimidos por 30 minutos. O sistema é redefinido se o paciente apresentar falha circulatória e se recuperar [18].

**Sistema EWS**

O sistema EWS original foi criado como uma versão simplificada do sistema de Avaliação do Estado Fisiológico e de Saúde Crónica, utilizado para avaliar a gravidade da doença nas Unidades de Cuidados Intensivos (UCI). O sistema de pontuação EWS normalmente baseia-se em variáveis fisiológicas, como frequência cardíaca, frequência respiratória e pressão arterial, além de uma medida do nível de alerta do paciente. À medida que o paciente se afasta da normalidade, as pontuações aumentam, e valores acima de um limite predefinido indicam a necessidade de chamar uma equipa médica especializada. [6] Foi realizado um estudo prospetivo para validar o tal sistema EWS, calculou os *scores* de 709 admissões de emergência médica até 5 dias e analisou a relação com o resultado. Uma pontuação máxima de 5 ou mais foi associada a maior risco de morte, internamento em UCI e internamento em unidade de alta dependência. [6] Um estudo posterior da mesma equipa acompanhou 1.695 admissões médicas agudas usando o mesmo sistema EWS, com protocolos de suporte para acionar a revisão médica e de cuidados intensivos. Não mostrou alteração no desfecho dos internamentos médicos agudos, nem no desfecho dos pacientes destacados como de risco, embora tenha sido observada uma tendência de encaminhamento e internamento mais precoce à UCI. Resultados mais promissores foram obtidos quando outra versão do EWS foi combinada com critérios de convocação para um serviço de cuidados intensivos. Isso levou a uma redução na taxa de internamentos de emergência na UCI, com menor tempo de permanência e menor mortalidade para os pacientes. Apesar dos benefícios geralmente observados nos estudos observacionais com sistemas EWS, até agora não se tentou efetivamente automatizar o processo de cálculo dos *scores*. Uma abordagem mais próxima talvez tenha sido a utilização de um modelo para analisar dados em tempo real provenientes de diversos dispositivos na Unidade de Cuidados Intensivos (UCI). A ideia subjacente foi introduzir lógica na análise, procurando distinguir leituras clinicamente insignificantes daquelas que são associadas a tendências fisiológicas prejudiciais ao paciente. [6]

## 2.3 Conceitos importantes

### 2.3.1 Pré-Processamento

O pré-processamento de dados é uma etapa essencial e inicial em qualquer processo de análise de dados. Consiste num conjunto de técnicas e procedimentos que têm como objetivo transformar os dados num formato que seja adequado, útil e eficiente para a análise subsequente. A eficácia do pré-processamento de dados é crucial para garantir que os resultados finais da análise sejam precisos e significativos. O processo de pré-processamento de dados envolve uma variedade de tarefas, como limpeza dos dados para lidar com valores ausentes (NA), duplicados ou inconsistentes, normalização para padronizar a escala das diferentes variáveis,

codificação de variáveis categóricas em formatos numéricos compreensíveis para algoritmos de *Machine Learning*, e seleção de características para identificar quais as variáveis que são mais relevantes para os objetivos da análise. [4]

É comum em grandes quantidades de dados existirem *missing values*, por várias razões, desde falhas na aquisição dos mesmos como a erros na importação dos ficheiros e é importante que que se atue sobre os valores, visto que os mesmos podem afetar as análises/modelações posteriores. Algumas formas de atuar são: eliminar as observações onde eles ocorrem, eliminar as variáveis onde ocorrem ou fazer a imputação através de técnicas existentes. [5]

Além disso, o pré-processamento de dados também pode incluir a deteção e tratamento de *outliers*, que são valores atípicos que podem distorcer os resultados, bem como a redução de dimensionalidade para lidar com conjuntos de dados de alta dimensionalidade, e a criação de conjuntos de treino e teste para avaliar a eficácia dos modelos. [4]

A escolha das técnicas de pré-processamento deve ser adaptada aos objetivos específicos da análise final. Por exemplo, num problema de classificação, é importante garantir que os dados estejam equilibrados entre as classes, classes balanceadas, para evitar viés nos modelos. [10] Em problemas de regressão, é comum observar variáveis com escalas diferentes umas das outras. Neste caso, seria difícil comparar coeficientes entre as variáveis, pelo que, é necessário proceder com a normalização dos dados. A normalização consiste em transformar as variáveis todas para a mesma escala. Existem várias formas de normalização. [11]

De forma geral, o pré-processamento de dados desempenha um papel crítico na preparação dos dados para análise, garantindo que os dados estejam limpos, formatados corretamente e prontos para serem usados nos algoritmos de *Machine Learning*. A qualidade do pré-processamento de dados desempenha um papel fundamental na qualidade dos resultados finais da análise e modelagem de dados. [4]

### 2.3.2 *Machine Learning*

ML é um subcampo da Inteligência Artificial que se concentra em algoritmos que permitem os computadores definirem um modelo para encontrar relações ou padrões complexos, a partir de dados empíricos, sem serem explicitamente programados. Os algoritmos de ML podem ser supervisionados ou não supervisionados. A aprendizagem supervisionada é uma abordagem da ciência de dados e do *Machine Learning*, que se concentra na previsão de resultados com base nos dados de um conjunto de treino previamente determinado. [12] Existem dois tipos principais de problemas supervisionados, dependendo da natureza da variável supervisora:

- **Problemas de Classificação**
  Quando estamos perante uma variável resposta do tipo categórico, ou seja, consiste em categorias ou classes discretas, neste caso estamos a lidar com problemas de classificação. O objetivo aqui é atribuir cada observação do conjunto de dados a uma classe específica, isto é, frequentemente utilizado em casos como classificação do diagnóstico médico (por exemplo, identificar se um paciente tem uma doença específica ou não), entre outros. A

avaliação do desempenho do modelo de classificação é geralmente baseada na precisão das previsões. [19]

- **Problemas de Regressão**
  Quando estamos perante uma variável resposta do tipo quantitativo, envolvendo valores numéricos contínuos, neste caso estamos a lidar com problemas de regressão. Aqui, o objetivo é prever um valor numérico que seja uma estimativa precisa da variável supervisora. Isto é frequentemente utilizado em previsões financeiras, estimativas de preços de imóveis, entre outros. A avaliação do desempenho do modelo de regressão é geralmente feita através de medidas de performance como o *Mean Squared Error* (MSE) ou $R^2$. [20]

Para avaliar a capacidade de um modelo de ML supervisionado em fazer previsões precisas sobre novas observações, é fundamental ter um conjunto de dados de teste independente do conjunto de dados de treino. O conjunto de treino ($X_{tr}$) é usado para ajustar os parâmetros do modelo, enquanto o conjunto de teste ($X_{te}$) é reservado para avaliar o desempenho do modelo. Isto ajuda a evitar o sobreajustamento, onde o modelo se adapta demais aos dados de treino e não generaliza bem para novos dados. Deve-se ter cuidado com modelos de alta complexidade visto quando o modelo se ajusta excessivamente aos dados do conjunto de treino perde a capacidade de generalização podendo ocorrer o *overfitting*. [19]

Existem diversos métodos e algoritmos disponíveis para resolver problemas de aprendizagem supervisionada. [21] Alguns dos mais conhecidos incluem:

- **Regressão Linear e Não Linear:** A regressão linear é um método simples e eficaz para problemas de regressão, enquanto as versões não lineares permitem modelar relações mais complexas.
- **Análise Discriminante:** É usada em problemas de classificação para encontrar uma fronteira de decisão que melhor separa as classes.
- **Redes Neurais:** Modelos que são altamente flexíveis e podem lidar com uma ampla gama de problemas, tanto de classificação quanto de regressão.
- **Máquinas de Suporte Vetorial (SVM):** São úteis em problemas de classificação, encontrando o hiperplano que melhor separa as classes, maximizando a margem entre elas.
- **k-Vizinhos Mais Próximos (k-NN):** Um método simples que atribui uma classe com base na maioria das classes dos $k$ vizinhos mais próximos no espaço de recursos.
- **Árvores de Decisão:** São usadas em problemas de classificação e regressão, dividindo o espaço em segmentos para fazer previsões.

É importante observar que não existe um método "melhor"em todos os casos. A escolha do método depende da natureza do problema, dos dados disponíveis e dos requisitos específicos do projeto. A escolha do método é uma parte fundamental do processo de modelagem e deve ser baseada na compreensão do problema e na experiência prática.

Na aprendizagem não supervisionada para cada observação não existe uma resposta $t$, ou seja, estão apenas disponíveis as variáveis independentes. Neste tipo de aprendizagem não

faz sentido falar em conjunto de teste ou de treino, pretende-se encontrar uma estrutura ou organização dos dados usando critérios de semelhança (ou dissemelhança). Alguns métodos utilizados são: Técnicas de *Clustering* (k-médias,k-*medoids*), Regras de associação ou SOM (*self organizing maps*).

### 2.3.3 Random Forest

O *Random Forest* é um algoritmo de *Machine Learning* que pertence à família dos métodos mais conhecidos como Métodos Ensemble. Foi criado para melhorar a precisão e a robustez dos modelos de classificação e regressão. O objetivo principal do *Random Forest* é a combinação de várias árvores de decisão para tomar decisões mais precisas e evitar o *overfitting* do conjunto de treino. [9]

Os métodos de ensemble são uma abordagem no ML onde a ideia é melhorar o desempenho preditivo e a robustez dos modelos combinando várias previsões de diferentes algoritmos num único modelo. Esta abordagem baseia-se no princípio de que a combinação de várias previsões pode produzir resultados mais precisos e menos suscetíveis a *overfitting*. Existem diversos tipos de métodos de ensemble, e eles podem ser divididos em duas categorias principais: ***bagging*** e ***boosting***. [13]

A escolha entre estas técnicas depende do problema em questão e do conjunto de dados que temos disponível. Cada método tem as suas próprias características e é mais apropriado em diferentes situações. [13]

O Random Forest é uma abordagem que envolve a utilização de múltiplas classificações geradas por árvores de decisão aleatórias para determinar uma classificação geral para o conjunto de dados em questão. Em vez de depender de uma única árvore de decisão, o Random Forest aproveita a diversidade e o poder das várias árvores para fornecer previsões mais robustas e precisas. Cada árvore contribui com seu voto e o resultado que o modelo fornece é a classe que ocorre com mais frequência, ou seja, a moda das classes. Em cada nó apenas um subconjunto de $m$ variáveis originais é utilizado para criar a partição. Esta seleção é feita para que não existam árvores correlacionadas, ou seja, permite eliminar o efeito das variáveis preditivas mais fortes que são constantemente escolhidas para uma divisão de nó em quase todas as árvores. Se existirem "árvores" na "floresta" que estão correlacionadas, as mesmas não estarão a ajudar o modelo a prever e a classificar o problema. O Random Forest é um algoritmo muito utilizado devido à sua capacidade de recorrer a várias amostras do conjunto de dados original, resultando na redução da variância do modelo e, consequentemente, na melhoria do seu desempenho. Esta estratégia é especialmente eficaz na prevenção do *overfitting*, que é quando o modelo se adapta excessivamente aos dados do conjunto de treino, tornando-se menos capaz de generalizar para novos dados. Ao criar diversas árvores de decisão independentes, cada uma treinada com uma amostra aleatória dos dados, o Random Forest introduz diversidade no modelo, tornando-o mais robusto e preciso. Ao combinar as previsões de todas as árvores por meio de majority voting (classificação) ou média (regressão), o Random Forest fornece uma previsão final que é menos suscetível a

alterações no conjunto de dados de treino, resultando em modelos mais confiáveis e eficazes. Isto torna o Random Forest uma escolha comum em diversas aplicações do *Machine Learning*. Como já foi referido anteriormente o Random Forest é um modelo baseado em árvores de decisão e não é caracterizado através de fórmulas muito complexas, teoremas ou expressões matemáticas detalhadas no seu funcionamento. No entanto, vamos mostrar uma visão geral do funcionamento de uma árvore de decisão e da forma como as várias árvores de decisão são combinadas num Random Forest. Uma árvore de decisão é uma estrutura hierárquica que divide os dados em subgrupos com base nas regras de decisão. Estas regras são obtidas através de divisões do espaço pelas características do conjunto de dados, ou seja, em cada nó é importante determinar qual será a regra que vai definir a partição dos dados para a separação das classes. O objetivo é criar uma árvore que minimize a impureza ou maximize o ganho de informação. [9]

A medida de impureza, $i(t) \geq 0$, é a medida que permite cada nó $t$ escolher a variável e o valor mais vantajoso através do qual vai fazer a partição. [8]Deverá satisfazer as seguintes propriedades:

- $i(t) = 0$ se algum $p_j(t) = 1$ e os restantes são nulos
- $i(t)$ é máxima se $p_j(t) = \frac{1}{2}, \forall j$

onde $p_j(t)$ é a proporção de observações da classe $C_j$ presentes no nó $t$ e $c$ é o número de classes. [8]

A impureza é mínima quando num determinado nó só existem observações de uma única classe e é máxima quando todas as classes estão representadas em igual proporção num único nó. A medida de impureza pode ser calculada através das seguintes fórmulas:

- **Entropia de Shannon:** $i(t) = -\sum_{j=1}^{c} p_j(t) \log p_j(t)$
- **Índice de Gini:** $i(t) = 1 - \sum_{j=1}^{c} p_j^2(t)$

O ganho de informação é a redução esperada na impureza causada pelo particionamento das observações usando um determinado atributo. É calculado da seguinte forma:

$$Ganho(S, A) = i(S) - \sum_{a \in valores(A)} \frac{|S_a|}{|S|} i(S_a)$$

onde $S$ é o conjunto de observações, $A$ é a variável a testar, $valor(A)$ é o conjunto de todos os valores de $A$, $S_a$ é o subconjunto de $S$ cujo valor de $A$ é $a$, $|.|$ é o cardinal e $i(S)$ é a impureza do conjunto $S$ presente no nó $t$.

Em situações em que as variáveis são contínuas ou têm muitos valores distintos, a árvore de decisão pode tornar-se excessivamente complexa e potencialmente sobreajustada aos dados do conjunto de treino (*overfitting*). Para lidar com esse problema, uma abordagem alternativa é a utilização o "RacioGanho(S,A)", que é um método de seleção dos atributos que considera

não apenas o ganho de informação, mas também a complexidade da partição resultante. O Rácio Ganho permite penalizar partições de alta entropia, o que significa que ele desencoraja a seleção de atributos com muitos valores diferentes. Isso ajuda a obter árvores de decisão menos complexas e, consequentemente, evitar o *overfitting*, tornando o modelo mais equilibrado e mais adequado para a generalização.

A fórmula do Rácio Ganho é geralmente definida como a razão entre o ganho de informação e a entropia da partição resultante. A entropia da partição é uma medida de impureza dos grupos formados após a divisão, e o ganho de informação é a diferença entre a entropia antes da divisão e a entropia após a divisão. Portanto, o Rácio Ganho considera o ganho de informação, mas também o grau de impureza resultante da divisão. A fórmula é a seguinte:

$$RacioGanho(S, A) = \frac{Ganho(S, A)}{InfoPart(S, A)}$$

onde $InfoPart(S, A) = -\sum_{a \in valores(A)} \frac{|S_a|}{|S|} log \frac{|S_a|}{|S|}$.

Embora a fórmula exata do Rácio de Ganho possa variar dependendo da implementação e das especificidades do problema, a ideia principal é incluir uma componente de penalização da complexidade da partição na seleção dos atributos. Esta abordagem pode ser útil em cenários onde a complexidade do modelo deve ser controlada, e a simplicidade da árvore de decisão é valorizada. O processo de construção de uma árvore de decisão acontece nó após nó, onde cada divisão é baseada na escolha do atributo que melhor separa as observações. Este processo continua até que todas as observações pertençam à mesma classe (criando uma folha) ou até que todas as observações tenham os mesmos valores para os atributos, o que é menos comum. [9]

Quando uma árvore chega a um estado em que todas as observações das folhas pertencem à mesma classe, essa árvore é representada como $T_{max}$. Embora $T_{max}$ seja capaz de classificar perfeitamente o conjunto de treino, geralmente não apresenta o desempenho ideal nos dados de teste (validação). Isto ocorre devido a dois problemas principais:

- **Profundidade Excessiva:** $T_{max}$ pode ser excessivamente profunda, o que significa que tem muitos níveis e regras de divisão. Isto torna a árvore muito específica para o conjunto de treino e com possibilidade de gerar mais erros nos dados não vistos.
- 
- **Sobreajustamento (*Overfitting*):** $T_{max}$ ajusta-se excessivamente aos dados do conjunto de treino (*overfitting*). Deteta o ruído nos dados de treino, em vez de aprender os padrões gerais que podem ser aplicados aos dados novos.

Para obter uma árvore com melhor desempenho do que $T_{max}$, é importante interromper o crescimento da árvore no momento em que começa a ocorrer *overfitting*. No entanto, determinar o ponto ideal para interromper o crescimento da árvore pode ser difícil. Uma abordagem comum é construir uma árvore completa até $T_{max}$ e, de seguida, usar um processo de poda para remover ramos da árvore com base num conjunto de validação ou validação

cruzada. Isso ajuda a controlar a complexidade da árvore, evitando o *overfitting*. O uso de validação ou validação cruzada permite ajustar o tamanho da árvore para obter um equilíbrio entre a capacidade de modelagem e a generalização. Desta forma, é possível obter um modelo da árvore de decisão que seja mais robusto e eficaz na classificação de novos dados, evitando problemas associados a árvores muito profundas e sobreajustadas.

A combinação das previsões das várias árvores ajuda a reduzir o *overfitting* e aumentar a precisão. No entanto, não há uma fórmula específica ou teorema matemático que descreva todo o processo, porque o algoritmo depende da construção das várias árvores e da combinação das suas previsões. Em vez de equações complexas, a força do Random Forest reside na implementação prática. O algoritmo é muito utilizado devido à sua eficácia e versatilidade na construção de modelos de ML. Embora a sua base teórica envolva conceitos matemáticos, a implementação não requer fórmulas matemáticas complicadas. O Random Forest é uma ferramenta prática que se concentra na combinação dos resultados das árvores de decisão para obter previsões mais robustas.

### 2.3.4 Regressão Logística

A regressão logística é uma técnica de modelagem estatística que se estende da tradicional regressão linear, permitindo a análise e previsão de variáveis de resposta categóricas. Embora a regressão linear seja amplamente utilizada para prever variáveis numéricas, a regressão logística assume o controlo quando a variável de resposta é binária, ou seja, possui apenas dois valores possíveis, como "sim"ou "não", por exemplo. [14]

A importância da regressão logística está na sua adaptabilidade para problemas onde a variável de resposta é dicotómica, como por exemplo na área da saúde para prever a presença ou ausência de uma doença. No entanto, a aplicabilidade da regressão logística não se limita a problemas binários. Ela pode ser facilmente estendida para lidar com problemas de resposta categórica com mais de duas categorias, tornando-a uma ferramenta valiosa numa variedade de domínios, como classificação de texto, previsão de categorias de produtos e muito mais. [15]

A regressão logística utiliza uma função logística (também conhecida como função sigmoide) para modelar a probabilidade de que uma observação pertença a uma categoria específica. Isto torna a regressão logística uma técnica particularmente útil para compreender as relações entre variáveis independentes e a probabilidade de um evento ocorrer. Com suas aplicações versáteis e interpretação intuitiva, a regressão logística desempenha um papel fundamental na análise de dados e na tomada de decisões informadas numa ampla gama de campos. [14]

O modelo de regressão logística baseia-se na função logística (ou função sigmoide), que é uma curva em forma de "S"que varia de 0 a 1. Esta curva é fundamental na regressão logística, pois permite que as probabilidades sejam modeladas de acordo com a relação entre as variáveis independentes e a variável de resposta. [16]

A equação do modelo logístico é dada por:

$$P(Y = 1) \approx \frac{1}{1 + e^{-}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}, \tag{2.1}$$

,onde $P(Y = 1)$ é a probabilidade de que a variável de resposta $Y$ pertença à categoria "1", $\beta_0$ é a interceção e $\beta_1, \beta_2, ..., \beta_p$ são os coeficientes associados às variáveis independentes $X_1, X_2, ..., X_p$.

A principal interpretação do modelo logístico é que ele fornece a probabilidade de sucesso (ou pertencer à categoria "1") em relação às variáveis independentes. Para problemas de classificação binária, como prever se um e-mail é spam ou não spam, a regressão logística é usada para modelar a probabilidade de ser "spam"com base nas várias características do e-mail, como palavras-chave, remetente, etc. [15]

Os coeficientes $\beta$ na equação indicam o impacto das variáveis independentes nas probabilidades. Um coeficiente positivo aumenta as probabilidades de pertencer à categoria "1", enquanto um coeficiente negativo diminui essas probabilidades. A interpretação dos coeficientes é fundamental para entender a relação entre as variáveis independentes e a variável de resposta. [14]

O método de máxima verosimilhança é utilizado para estimar os parâmetros em modelos estatísticos, incluindo o modelo logístico. Este método é particularmente adequado para a regressão logística, uma vez que tem como objetivo encontrar os valores dos coeficientes que maximizam a verosimilhança dos dados observados sob o modelo. O método consiste em encontrar os valores dos parâmetros do modelo que tornam os dados observados mais prováveis de ocorrer. No caso da regressão logística, isto significa encontrar os coeficientes que tornam mais provável a observação das respostas categóricas dadas as variáveis independentes. O método procura ajustar o modelo para que ele se encaixe melhor nos dados observados. Em termos matemáticos, o método de máxima verosimilhança envolve a maximização da função de verosimilhança, que é uma medida da probabilidade de observar os dados sob o modelo. A função de verosimilhança leva em consideração as probabilidades previstas pelo modelo logístico e compara essas probabilidades com os valores reais das respostas categóricas nos dados do conjunto de treino. [16]

A maximização da função de verosimilhança geralmente é realizada com a ajuda de técnicas computacionais, como o algoritmo de *Newton-Raphson* ou o gradiente descendente, que iterativamente ajustam os valores dos coeficientes até que a verosimilhança seja maximizada.

$$L(\beta) = \prod_{i=1}^{n} P(Y_i|X_i; \beta)$$

, onde $L(\beta)$ é a função verosimilhança, $\beta$ representa o vetor de coeficientes a serem estimados, $n$ é o número de observações e $P(Y_i|X_i; \beta)$ é a probabilidade da observação $i$ pertencer à categoria "1"(ou à categoria que está a ser modelada) com base nas variáveis independentes $X_i$ e nos coeficientes $\beta$.

A fórmula completa de $P(Y_i|X_i; \beta)$ é a função sigmoide já referida anteriormente em 2.1. [16]

De forma geral, o modelo logístico é uma ferramenta valiosa para análise de dados e previsão nas situações em que a variável de resposta é categórica. Ele é amplamente utilizado em áreas como medicina, marketing, ciências sociais e muito mais, onde a compreensão das probabilidades é essencial para a tomada de decisões informadas. [16]

### 2.3.5 Medidas de Performance

As medidas de performance são indicadores utilizados para avaliar o desempenho de um modelo. Elas fornecem uma visão objetiva da eficácia, precisão e confiabilidade da execução, sendo essenciais para a análise crítica dos algoritmos ou modelos estatísticos. Estas medidas podem abranger uma variedade de métricas, como a *accuracy*, sensibilidade, especificidade e *F1-Score*, dependendo do contexto específico da aplicação. O objetivo principal das medidas de performance é oferecer uma avaliação quantitativa e comparativa, permitindo a otimização e aperfeiçoamento contínuo do desempenho do modelo em questão.

A matriz de confusão é uma tabela que é usada para avaliar o desempenho de um modelo de classificação. Ela compara as previsões de um modelo em relação aos valores reais dos dados e categoriza as previsões em quatro quadrantes: verdadeiros positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN). Esta matriz fornece uma visão detalhada de como o modelo classifica as observações em diferentes categorias, o que é essencial para a avaliação de sua eficácia.

A Precisão (*accuracy*) é utilizada para avaliar o desempenho de um modelo de classificação. Mede a proporção de previsões corretas feitas pelo modelo em relação ao número total de observações. De forma mais simples, é a "taxa de acerto"do modelo. A fórmula para calcular a precisão é a seguinte:

$$Precisão = \frac{Número de previsões corretas}{Número total de observações}$$

A precisão é expressa através de um valor entre 0 e 1, ou através de uma percentagem. Quanto mais próxima de 1 (ou 100%), maior é a precisão do modelo, indicando que ele está a fazer um bom trabalho na classificação das observações. No entanto, a precisão pode não ser a mais apropriada em todos os cenários, especialmente quando as classes de dados estão desequilibradas. Em problemas com classes desequilibradas, um modelo pode alcançar uma alta precisão simplesmente a prever a classe que tem mais observações na maioria dos casos, o que não reflete necessariamente um bom desempenho na identificação da classe com menos observações. Portanto, nestes casos, é importante considerar outras medidas, como sensibilidade, especificidade e os valores preditivos, para obter uma avaliação mais completa do desempenho do modelo.

A *Sensitivity* (Sensibilidade), também conhecida como Taxa de Verdadeiros Positivos (*True Positive Rate - TPR*), é uma medida de desempenho que indica a capacidade do modelo de identificar corretamente os casos positivos em relação ao número total de casos positivos reais. Para calcular a sensibilidade usa-se a seguinte fórmula:

$$Sensibilidade = \frac{TP}{TP + FN}$$

, onde $TP$ representa o número de verdadeiros positivos e $FN$ representa o número de falsos negativos.

A Especificidade (*Specificity*) mede a capacidade do modelo em identificar corretamente os casos negativos em relação ao número total de casos negativos reais. Para calcular a especificidade usa-se a seguinte fórmula:

$$Especificidade = \frac{TN}{TN + FP}$$

, onde $TN$ representa o número de verdadeiros negativos e $FP$ representa o número de falsos positivos.

No contexto de saúde uma alta especificidade é relevante, porque indica que o modelo é capaz de minimizar os falsos alarmes, ou seja, reduzir o número de casos negativos classificados de forma errada como positivos. No entanto, é importante lembrar que a interpretação das medidas de desempenho deve ter em consideração o contexto clínico específico e o impacto prático das previsões do modelo. A combinação de alta sensibilidade e alta especificidade é geralmente o objetivo ideal em muitos cenários, equilibrando a capacidade de detetar positivos verdadeiros com a capacidade de evitar falsos positivos.

O *F1-Score* é uma medida de performance utilizada na avaliação de modelos de classificação, especialmente em situações onde o equilíbrio entre a precisão e a sensibilidade é fundamental. Ele combina a precisão e a sensibilidade num único valor para proporcionar uma medida compreensiva do desempenho do modelo. O *F1-Score* é particularmente valioso quando as classes de dados estão desequilibradas (este caso), o que é comum em muitos cenários do mundo real. É calculado pela seguinte fórmula:

$$\textit{F1-Score} = 2 \times \frac{Precisão \times Sensibilidade}{Precisão + Sensibilidade}$$

CAPÍTULO 3

# Recolha e análise inicial dos dados

Este capítulo, vai se centrar na apresentação do processo de recolha e preparação dos dados para este estudo. Vai ser mostrada cada etapa, desde a obtenção até à organização dos dados. Teremos também uma análise descritiva do conjunto de dados final onde serão explicadas todas as variáveis recolhidas. Esta análise permitirá compreender a natureza e a distribuição dos dados em estudo. Por último, será feita uma descrição da análise exploratória realizada, destacando padrões, tendências e *insights* que resultaram desta fase do processo.

## 3.1 Recolha dos dados

Após uma revisão de literatura e uma análise profunda do problema em questão, elaborou-se um pequeno conjunto de questões que desempenharam um papel crucial na escolha das variáveis necessárias para prosseguir com o estudo.

As questões elaboradas foram as seguintes:

- Qual é a proporção de utentes internados na UCI 2022?
- Quais os doentes que mais foram internados na UCI 2022?
- De que especialidade vieram os doentes internados na UCI 2022?
- Os doentes internados na UCI 2022 vieram das urgências ou já estavam internados noutra especialidade do hospital?
- Qual foi o tipo de doentes que foram para a UCI (que tipo de patologias os fez ir lá parar: respiratória, cardiovascular, renal...)?
- Qual foi o diagnóstico de admissão para a UCI?
- Com que sintomas entraram na UCI?
- Há quanto tempo tinham começado a ter os sintomas?
- Quais as medidas de frequência respiratória antes de dar entrada na UCI?
- Quais as medidas de frequência cardíaca antes de dar entrada na UCI?
- Quais as medidas da pressão arterial sistólica e diastólica antes da entrada na UCI?

- Quais as medidas de temperatura antes da entrada na UCI e onde a temperatura do doente foi medida?
- Se o doente se encontrava consciente antes de dar entrada na UCI?
- Se o paciente tinha dores antes de dar entrada na UCI?
- Qual o estado mental do doente antes de dar entrada na UCI?
- Quais os problemas de saúde que o doente tinha antes de dar entrada na UCI?
- Qual a faixa etária do doente?
- O doente tomava alguma medicação devido a algum problema de saúde? Se sim, qual?
- O doente já esteve alguma vez na UCI devido a algum problema? Se sim, qual?

Após a apresentação das questões mencionadas anteriormente, avançou-se para a fase de análise dos dados disponíveis através das diversas instituições de saúde em Portugal que a empresa tem acesso, para posteriormente se escolher uma delas para recolha dos dados. Através desta análise não se obtiveram as respostas a todas as questões propostas, no entanto, fomos capazes de identificar e obter as seguintes variáveis, que representam informações valiosas e desempenharão um papel central nesta pesquisa.

| Variável | Descrição |
|---|---|
| **UtenteKey** | Código associado a cada doente |
| **DataNascimento** | Data de nascimento do doente |
| **FaixaEtaria** | Faixa etária a que pertence o doente |
| **IdadeUtente** | Idade do doente |
| **Sexo** | Género do doente |
| **DataObito** | Data de óbito do doente, caso tenha falecido |
| **ObitoFetalMaiorIgual 28Semanas** | Variável que indica se o feto faleceu com 28 ou mais semanas |
| **ProvenienciaId** | Código da proveniência do doente |
| **ProvenienciaDescricao** | Descrição da proveniência do doente |
| **IsencaoId** | Código da razão de isenção de pagamento de taxas |
| **Isencao** | Descrição da razão de isenção de pagamento de taxas |
| **DataAdmissaoUCI** | Data em que o utente foi admitido à UCI |
| **DataSaidaUCI** | Data em que o utente teve alta da UCI |
| **EspecialidadeAdmissaoKey** | Código da especialidade onde o doente foi admitido |
| **EspecialidadeAdmissaoDescricao** | Descrição da especialidade onde o doente foi admitido |

| Variável | Descrição |
|---|---|
| **EspecialidadeAltaKey** | Código da especialidade que o doente teve alta |
| **EspecialidadeAltaDescricao** | Descrição da especialidade onde o doente teve alta |
| **EspecialidadeAdmissaoPrevistaKey** | Código da especialidade onde o doente deveria ter sido admitido, podem existir casos em que não havia espaço para o doente na especialidade prevista então o mesmo foi admitido noutra, mas tem os tratamentos da especialidade prevista |
| **EspecialidadeAdmissaoPrevistaDesc** | Descrição da especialidade prevista de admissão |
| **EspecialidadeAltaPrevistaKey** | Código da especialidade de alta prevista |
| **EspecialidadeAltaPrevistaDesc** | Descrição da especialidade de alta prevista |
| **Histerectomia** | Indica se o doente teve uma histerectomia ou não |
| **Complicacoes** | Indica se o doente teve complicações durante o internamento |
| **RecemNascido** | Indica se o doente é um recém-nascido |
| **CirurgicoProgramado** | Indica se o doente está internado por motivo de uma cirurgia programada |
| **NadoVivo** | Indica se o doente é um nado-vivo |
| **DataAdmissao** | Data de admissão ao internamento |
| **HoraAdmissao** | Hora de admissão ao internamento |
| **DataAltaClinica** | Data de alta clínica, que é a data em o doente recebe a alta do médico e já pode sair do hospital |
| **DataAltaHospitalar** | Data de alta hospitalar é a data em que o doente saiu do hospital. Existem doente que já possuem alta clínica, no entanto ainda precisam de cuidados que não conseguem ter fora do hospital pelo que ainda ficam a receber cuidados médicos e só depois recebem a alta hospitalar. |
| **HoraAltaHospitalar** | Hora da alta hospitalar |
| **DataCirurgia** | Data da cirurgia, caso o doente tenha passado por alguma |
| **DataInscricaoLIC** | Data que foi referenciado para a Lista de inscritos para cirurgia |

| Variável | Descrição |
|---|---|
| **DataAdmissaoUrgencia** | Data de admissão na Urgência, casos onde o doente deu entrada no internamento pela urgência |
| **HoraAdmissaoUrgencia** | Hora de admissão na urgência |
| **ProvenienciaEpisodio** | Indica de onde o doente foi admitido no internamento |
| **ProvenienciaEspecialidadeDescricao** | Descrição da proveniência do doente |
| **InternamentoTipoAdmissao** | Tipo de internamento do doente |
| **InternamentoResultado** | Resultado do internamento do doente |
| **UltimaAdmissao** | Data da última admissão, caso exista |
| **Saturação de oxigénio** | Medidas da saturação de oxigénio que foram feitas aos doentes durante o internamento |
| **Frequência Respiratória** | Medidas da frequência respiratória que foram feitas aos doentes durante o internamento |
| **Frequência Cardíaca** | Medidas da frequência cardíaca que foram feitas aos doentes durante o internamento |
| **Temperatura** | Medidas da temperatura que foram feitas aos doentes durante o internamento |
| **Pressão diastólica** | Medidas da pressão diastólica que foram feitas aos doentes durante o internamento |
| **Pressão sistólica** | Medidas da pressão sistólica que foram feitas aos doentes durante o internamento |
| **Diagnósticos** | Diagnósticos dos doentes no internamento |

**Tabela 3.1:** Descrição das Variáveis

## 3.2 Pré-Processamento

Inicialmente, após a recolha de dados, fez-se uma análise criteriosa das variáveis obtidas. Esta análise é importante, porque permite identificar as variáveis todas, e a forma como as mesmas nos são apresentadas. Com o intuito de otimizar a análise do estudo, deu-se início à organização da base de dados, uma tarefa que visa assegurar que todas as variáveis recolhidas desempenhem um papel relevante na investigação em curso.

Neste contexto, implementou-se um cuidadoso processo de seleção e categorização, garantindo que cada variável contribuísse de maneira significativa para os objetivos de estudo. Algumas variáveis, devido à necessidade de simplificar a análise, foram criadas ou redefinidas. A introdução de variáveis adicionais à base de dados tem como objetivo facilitar e enriquecer a análise subsequente. A organização cuidadosa da base de dados contribui não só para a eficiência do estudo, mas também é um passo crucial para conseguirmos obter informação e conclusões para o nosso estudo.

A primeira variável analisada foi a "Diagnóstico". Esta variável, caracterizada por diversos códigos ("DiagnosticoID"), revelou-se inicialmente extensa, com cada código vinculado a um diagnóstico específico, resultando numa ampla gama de diagnósticos distintos. Diante desta complexidade, foi tomada a decisão estratégica de categorizar os diagnósticos em capítulos, conforme descrito na Tabela 3.2. Esta abordagem tem como objetivo simplificar a interpretação e análise, agrupando os diagnósticos relacionados em categorias mais amplas e representativas.

O ICD-10, ou Classificação Internacional de Doenças, Décima Revisão, é um sistema de classificação padronizado e internacionalmente reconhecido para categorizar diversas condições de saúde e doenças. Desenvolvido pela Organização Mundial da Saúde (OMS), o ICD-10 é amplamente utilizado em todo o mundo para fins epidemiológicos, estatísticos e administrativos no campo da saúde. Esta classificação é composta por códigos alfanuméricos que representam diferentes diagnósticos, sinais, sintomas e causas externas de lesões ou doenças. Cada código no ICD-10 fornece informações detalhadas sobre uma condição de saúde específica. A estrutura do ICD-10 permite uma organização sistemática das doenças em categorias e subcategorias. [23] Neste contexto, os "DiagnósticosID"estavam representados através desses códigos do ICD-10. A empresa forneceu-me o catálogo na versão de 2017 através do qual foi feita essa divisão dos diagnósticos por capítulos.

Ao optar por esta categorização, não apenas simplificamos a complexidade da variável "Diagnóstico", mas também criamos uma base para análises subsequentes. Esta estratégia permitiu uma compreensão mais clara e eficaz dos diferentes diagnósticos dos doentes, ao mesmo tempo preservou a integridade das informações clínicas. Ao agrupar os diagnósticos em capítulos, não só otimizamos a eficiência da análise, mas também proporcionamos uma visão mais estruturada dos dados, contribuindo assim para uma interpretação mais simples dos resultados obtidos.

Obtivemos assim “Capitulo”, que é uma variável que se encontra dividida em 21 categorias. Os capítulos encontram-se apresentados na Tabela 3.2.

| Capítulo | Código | Descrição |
|---|---|---|
| **Capítulo 1** | A00- B99 | Algumas doenças infeciosas e parasitárias |
| **Capítulo 2** | C00- D49 | Neoplasias |
| **Capítulo 3** | D50- D89 | Doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários |
| **Capítulo 4** | E00- E89 | Doenças endócrinas, nutricionais e metabólicas |

| Capítulo | Código | Descrição |
|---|---|---|
| **Capítulo 5** | F01- F99 | Transtornos mentais, comportamentais e de neuro desenvolvimento |
| **Capítulo 6** | G01- G99 | Doenças do sistema nervoso |
| **Capítulo 7** | H00- H59 | Doenças do olho e anexos |
| **Capítulo 8** | H60- H95 | Doenças do ouvido e da apófise mastóide |
| **Capítulo 9** | I00- I99 | Doenças do aparelho circulatório |
| **Capítulo 10** | J00- J99 | Doenças do aparelho respiratório |
| **Capítulo 11** | K00- K95 | Doenças do aparelho digestivo |
| **Capítulo 12** | L00- L99 | Doenças da pele e do tecido subcutâneo |
| **Capítulo 13** | M00- M99 | Doenças do aparelho osteomuscular e do tecido conjuntivo |
| **Capítulo 14** | N00- N99 | Doenças do aparelho geniturinário |
| **Capítulo 15** | O00- 09A | Gravidez, parto e puerpério |
| **Capítulo 16** | P00- P96 | Algumas condições originadas no período perinatal |
| **Capítulo 17** | Q00- Q99 | Malformações congénitas, deformações e anomalias cromossómicas |
| **Capítulo 18** | R00- R99 | Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados |
| **Capítulo 19** | S00- T88 | Lesões, envenenamento e algumas outras consequências de causas externas |
| **Capítulo 20** | V00- Y99 | Causas externas de morbilidade |
| **Capítulo 21** | Z00- Z99 | Fatores que influenciam o estado de saúde e o contacto com os serviços de saúde |

**Tabela 3.2:** Descrição da variável "Diagnostico"

Neste estudo mantiveram-se apenas os capítulos que tinham mais observações todos os restantes foram agrupados num só capítulo chamado “outros”. Nos outros encontram-se os diagnósticos dos capítulos 1, 3, 4, 5, 7, 8, 12, 15, 16, 20 e 21.

Através da variável "DataObito"obteve-se a variável "Obito", classificada como 1(faleceu) e 0(não faleceu), na qual podemos recolher a informação dos doentes que faleceram durante o internamento. Com as variáveis "DataAdmissaoUCI"e "DataAdmissaoUrgencia"obtiveram-se as variáveis "UCI"e "Urgencia", também classificadas com 1 e 0, através das quais podemos analisar os doentes que foram admitidos na UCI e aqueles que foram admitidos no internamento através das urgências, respetivamente. A variável "TempoInternamentoUCI"também se obteve através de datas, ou seja, da diferença entre a "DataSaidaUCI"e a "DataAdmissaoUCI", esta variável obteve-se em dias.

As variáveis da saturação de oxigénio, frequência respiratória, frequência cardíaca, temperatura, pressão diastólica e pressão sistólica foram fornecidas pela empresa numa folha *excel* com todas as medidas que foram feitas aos doentes desde a sua entrada no hospital até à

saída. Analisando as variáveis da pressão sistólica e pressão diastólica calculou-se a variável da pressão arterial média através da seguinte fórmula [22]:

$$PressãoArterialMédia = PressãoDiastólica + \frac{1}{3} \times (PressãoSistólica - PressãoDiastólica)$$

Para simplificar a análise e como o objetivo do estudo é fazer uma previsão dos doentes que são internados na UCI foram calculadas as médias das medidas 24 horas antes do internamento, 48 horas antes do internamento e 72 horas antes do internamento. Isto foi calculado através da data da medida e da data de internamento. Esta abordagem temporal permitirá uma avaliação mais precisa das condições de saúde dos doentes em períodos críticos que antecedem a decisão de internamento na UCI. As médias obtidas nesses intervalos podem fornecer uma visão consolidada e representativa do estado de saúde dos doentes nesses momentos específicos. A análise das médias pretende identificar padrões e tendências que possam ser indicativos de uma possível admissão na UCI. Além disso, ao comparar as médias das últimas 24 horas antes do internamento, 48 horas antes do internamento e 72 horas antes do internamento, será possível avaliar se existem alterações significativas nesses indicadores ao longo do tempo. Esta abordagem temporal permitirá uma compreensão mais profunda das condições clínicas dos doentes e contribuirá para a identificação de fatores de risco ou sinais precoces que possam influenciar a decisão de internamento na UCI.

Com isto foram criadas as seguintes variáveis "*mediaso24h*", "*mediaso48h*", "*mediaso72h*", "*mediafr24h*", "*mediafr48h*", "*mediafr72h*", "*mediafc24h*", "*mediafc48h*", "*mediafc72h*", "*mediat24h*", "*mediat48h*", "*mediat72h*", "*media_sis24h*", "*media_sis48h*", "*media_sis72h*", "*media_dias24h*", "*media_dias48h*", "*media_dias72h*", "*media_press24h*", "*media_press48h*"e "*media_press72*".

As variáveis de temperatura, pressão diastólica, pressão sistólica, pressão arterial, frequência cardíaca, frequência respiratória e saturação de oxigénio desempenham um papel crucial na monitorização da saúde de um indivíduo. No entanto, é importante reconhecer que estas variáveis frequentemente contêm uma quantidade significativa de valores ausentes, ou "NA"(Not Available), em bases de dados médicas. Isso ocorre devido a um desafio intrínseco na recolha de dados em ambientes de saúde.

Uma das razões para a presença frequente de NA's, que me foi indicada num estágio anterior, é o facto destes valores serem por norma registados manualmente por diferentes médicos ou profissionais de saúde em diferentes momentos. Tradicionalmente, estas informações são anotadas nos registos em papel, antes de serem potencialmente transferidas para sistemas informáticos. Durante esse processo, ocorrem erros de transcrição, perdas de dados e inconsistências, além disso, a natureza dinâmica dos ambientes de saúde, com informações a serem recolhidas numa variedade de situações clínicas e locais, pode aumentar a probabilidade de valores ausentes. Por exemplo, em emergências ou quando a atenção médica é prestada em condições adversas, o registo rigoroso dos dados pode ser desafiador. Portanto, ao lidar com essas variáveis de saúde em análises ou estudos, é essencial considerar cuidadosamente a presença de NA's e implementar abordagens adequadas para lidar com eles, como a imputação de dados ausentes ou a avaliação do seu impacto no estudo. Reconhecer a origem desses

valores ausentes é um passo fundamental para garantir a precisão e a validade das análises realizadas no campo da medicina e da saúde.

A percentagem de NA's nestas variáveis é a representada na Tabela 3.3.

| Variável | Percentagem de NA's |
|---|---|
| mediat24h | 61.7 |
| mediat48h | 59.9 |
| mediat72h | 59.6 |
| mediaso24h | 75.7 |
| mediaso48h | 73.6 |
| mediaso72h | 73.5 |
| mediafr24h | 91.9 |
| mediafr48h | 91.0 |
| mediafr72h | 90.9 |
| mediafc24h | 62.3 |
| mediafc48h | 60.2 |
| mediafc72h | 60 |
| media_press24h | 27.3 |
| media_press48h | 48.9 |
| media_press72h | 60.4 |
| media_sis24h | 27.3 |
| media_sis48h | 48.9 |
| media_sis72h | 60.4 |
| media_dias24h | 27.3 |
| media_dias48h | 48.9 |
| media_dias72h | 60.4 |

**Tabela 3.3:** Percentagem de NA's

A variável da frequência respiratória não foi incluída na análise devido a uma presença elevada de valores em falta (NA's). Dado que a qualidade e integridade dos dados são cruciais para análises estatísticas significativas, optou-se por não incluir esta variável no estudo.

A imputação de valores ausentes é um passo importante, caso seja o escolhido para tratar dos valores em falta, no pré-processamento de dados, uma vez que os valores em falta podem afetar a qualidade e a validade das análises e dos modelos. Existem várias técnicas de imputação disponíveis, como a média, a mediana, a moda, regressão, entre outras. No entanto, neste caso, optou-se por utilizar o algoritmo do KNN (*K-Nearest Neighbors*) para esta finalidade. Além da imputação dos NA's, é importante destacar que havia outra alternativa viável, que consistia na remoção dos dados em falta. No entanto, essa opção foi descartada devido às altas percentagens de valores ausentes no conjunto de dados.

O KNN é uma técnica de imputação que tem em consideração a semelhança entre os casos, em vez de atribuir um único valor constante, como a média, a todos os NA's. Usar a média,

por exemplo, poderia levar a uma simplificação excessiva dos dados, criando repetições de valores, o que seria inadequado, especialmente em situações em que os dados são complexos e heterogéneos. O princípio do KNN é encontrar os "vizinhos"mais próximos do caso com valor NA com base noutras variáveis disponíveis. Ele calcula a distância entre os casos e atribui um valor imputado com base nos valores dos vizinhos mais próximos. Isso ajuda a manter a diversidade nos dados, tornando a imputação mais precisa e adaptada ao conjunto de dados.

A variável 'Proveniencia' descreve a origem ou local de onde os doentes foram admitidos para tratamento médico. Esta variável é essencial para compreender o fluxo de doentes numa instituição de saúde e identificar possíveis padrões de admissão. Ela é categorizada em várias categorias distintas, refletindo diferentes fontes de admissão. A proveniência está descrita da seguinte forma:

- **1-** Urgência
- **2-** Consulta Externa
- **3-** Serviço de Internamento
- **4-** Hospital Dia
- **5-** Exterior
- **6-** Recém-Nascido
- **7-** ARS/ Centro Saúde
- **9-** Outro Hospital
- **10-** Clínicas Privadas
- **12-** CODU INEM
- **13-** Cirurgia Ambulatório Bloco
- **18-** Hospitalização Domiciliária
- **20-** Encaminhada pela Saúde24
- **22-** Interno MFR
- **23-** Delegado Saúde
- **26-** Cama Contratualizada
- **99-** Outras

No entanto observou-se que algumas das categorias tinham valores muito baixos na "Proveniencia"e dicidiu-se agrupar algumas delas numa só categoria chamada "Outras". No final ficaram as seguintes categorias:

- **Consulta Externa:** Doentes que chegam ao hospital através de consultas agendadas ou consultas de rotina.
- **Exterior:** Doentes admitidos de fora do hospital, mas não especificamente relacionados com outras categorias desta lista.
- **Hospital de Dia:** Doentes que recebem tratamento durante o dia, sem internamento noturno.
- **Outro Hospital:** Doentes transferidos de outras instituições de saúde ou hospitais.
- **Recém-Nascido:** Bebés recém-nascidos admitidos para monitorização.
- **Urgência:** Doentes que chegam ao hospital em situações de emergência.
- **Outras:** Categorias adicionais que podem incluir casos especiais ou fontes de admissão não especificadas anteriormente.

## 3.3 Análise Exploratória dos Dados

Vamos iniciar o processo com uma análise dos dados antes de aplicar qualquer técnica estatística ou de ML para que se possa conhecer os dados e as relações entre as várias variáveis existentes na base dados.

Nas idades podemos observar que a média de idades dos doentes desta base de dados é de 52 anos, onde o doente mais velho tem 108 anos e o mais novo 0, ou seja, é um possível recém-nascido que tem meses e ainda não tem um ano de idade. Também podemos observar a maioria dos doentes desta amostra têm mais de 65 anos. A faixa etária que tem menos doentes é entre os 6 e os 10 anos. Analisando apenas os doentes que foram internados na UCI podemos observar que os doentes que mais são internados são os doentes com mais de 65 anos.

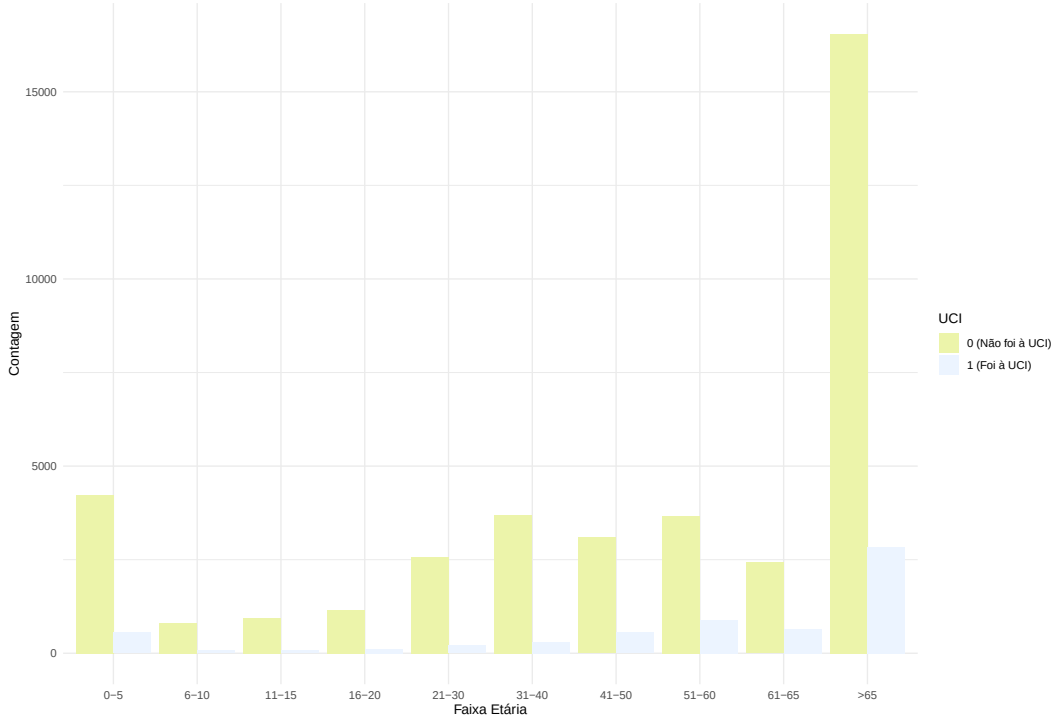


**Figura 3.1:** Comparação das faixas etárias entre os doentes que foram à UCI e os que não foram

O género dos doentes está distribuído de forma semelhante, 51% são mulheres e 49% são homens. Analisando apenas os doentes que foram internados na UCI podemos observar que a percentagem de homens aumentou para 61%.

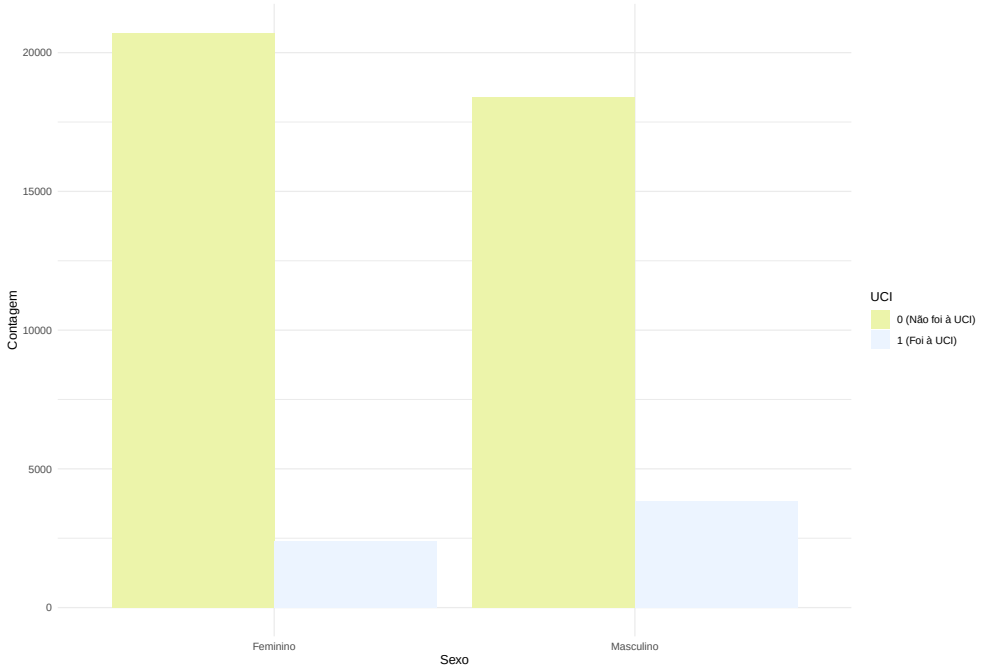


**Figura 3.2:** Comparação do sexo entre os doentes que foram à UCI e os que não foram

Em 45366 episódios podemos observar que ocorreram 2136 mortes das quais 791 foram de doentes que foram à UCI.

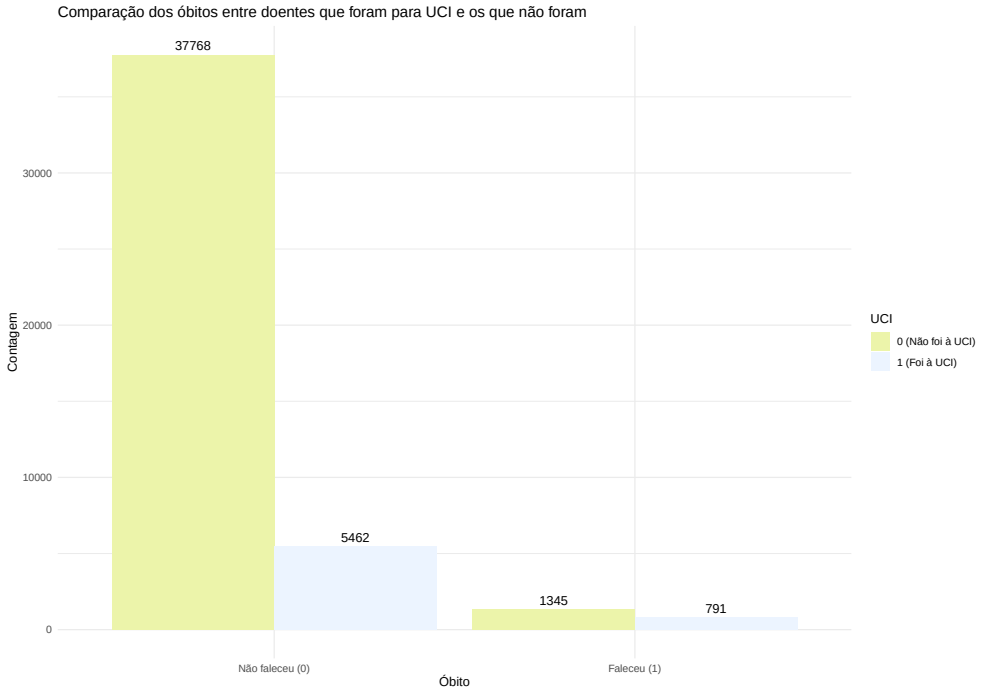


**Figura 3.3:** Comparação dos óbitos entre os doentes que foram à UCI e os que não foram

O capítulo 19 é um dos mais predominantes contém lesões, envenenamento e algumas outras consequências de causas externas. Dos doentes que foram internados na UCI o capítulo

19 (Lesões, envenenamento e algumas outras consequências de causas externas) continua a ser um dos diagnósticos mais predominantes.

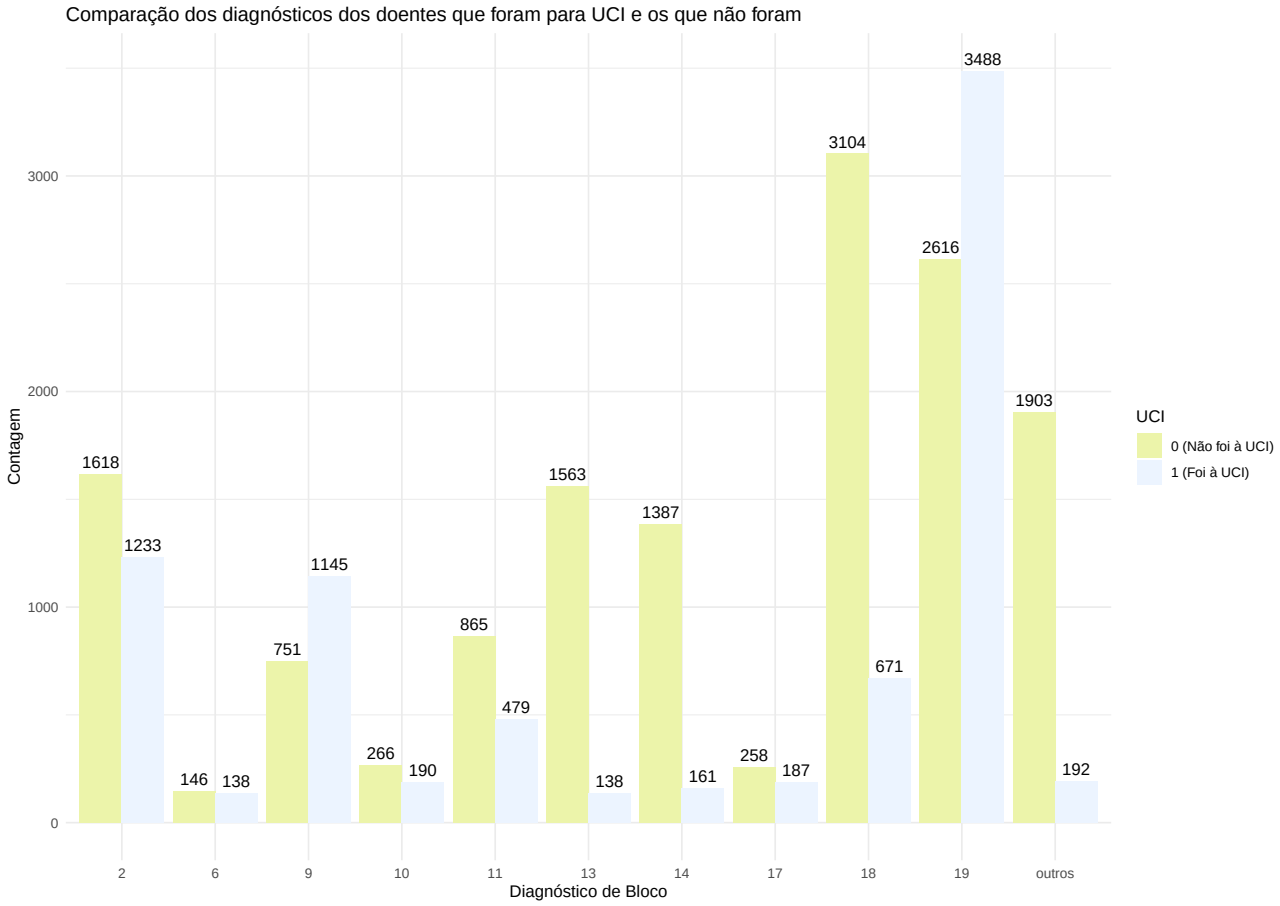


**Figura 3.4:** Comparação dos diagnósticos entre os doentes que foram à UCI e os que não foram

Existiram 25 óbitos fetais iguais ou maiores a 28 semanas, nenhum dos casos foi à UCI.

Um óbito fetal maior ou igual a 28 semanas, também conhecido como morte fetal tardia, ocorre quando um feto, que já atingiu ou ultrapassou a marca das 28 semanas de gestação, não sobrevive até ao parto. Esta variável foi considerada importante colocar no estudo para compreender as causas e os fatores de risco por trás destas situações. Poderiam existir casos destes que foram à UCI e assim posteriormente poderia se arranjar forma de prevenir estas situações para no futuro se melhorar a qualidade da assistência médica e pré-natal. A identificação de fatores de risco e causas subjacentes ajuda os profissionais de saúde a implementar intervenções preventivas e práticas de gestão mais eficazes. Isso pode incluir um acompanhamento mais rigoroso das gestações consideradas de alto risco. No entanto, existiram 25 casos de óbitos fetais iguais ou maiores a 28 semanas nestes dados, mas nenhum deles foi um caso que esteve internado na UCI, pelo que removemos esta variável do nosso estudo.

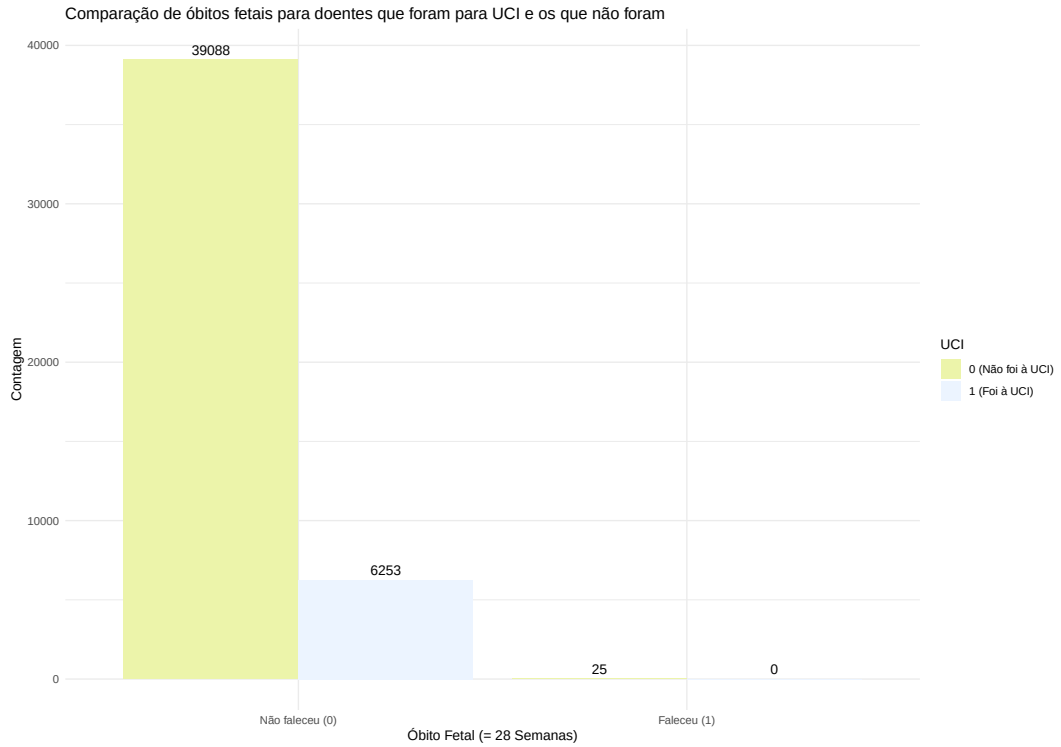


**Figura 3.5:** Comparação dos óbitos fetais maior ou igual a 28 semanas entre as doentes que foram à UCI e as que não foram

Os doentes que foram admitidos com mais frequência no internamento foram aqueles que chegaram pelo serviço de Urgência. Este grupo representa uma parte significativa das admissões, refletindo a imprevisibilidade das necessidades médicas. De seguida, observou-se que os doentes provenientes da Consulta Externa também constituíram um grupo considerável de admissões. Estes doentes tinham consultas agendadas previamente e, posteriormente, foram encaminhados para o internamento conforme necessário. Fazendo a análise da proveniência nos utentes que apenas foram internados na UCI, podemos observar que aqueles que vieram da Urgência e da Consulta Externa continuam a ser a maior parte.

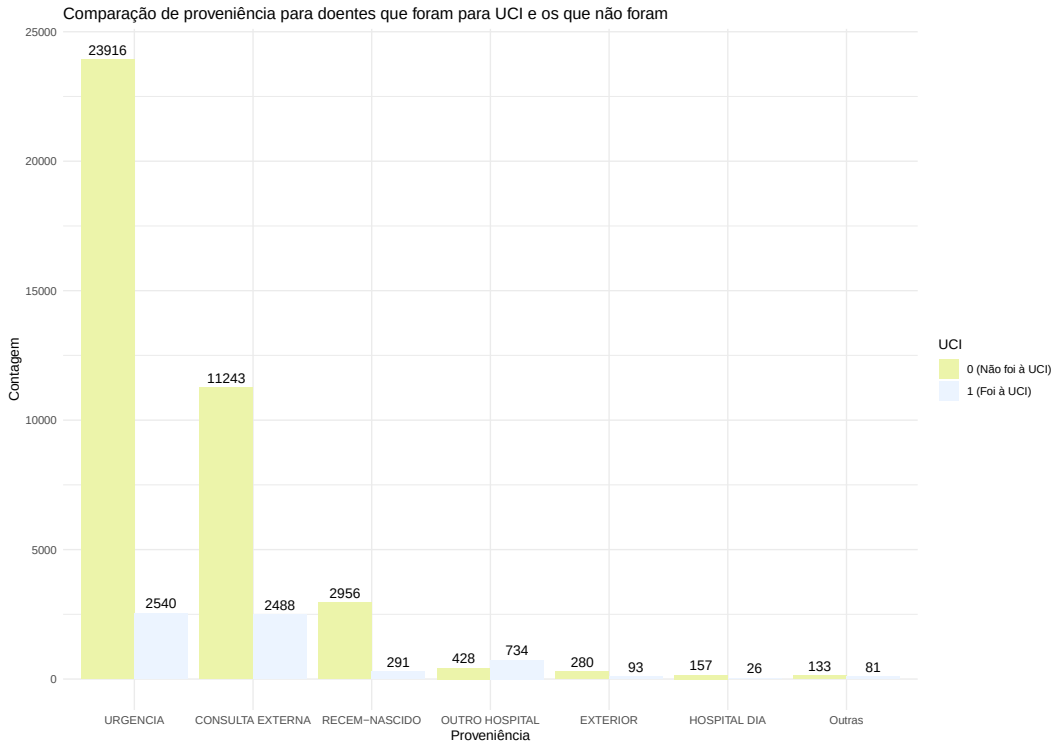


**Figura 3.6:** Comparação da Proveniência entre os doentes que foram à UCI e os que não foram

A variável "Isenção"desempenha consiste na compreensão das razões pelas quais os doentes estão isentos de pagar taxas relacionadas com os seus cuidados de saúde. Esta variável abrange várias categorias que refletem as diversas circunstâncias que levam à isenção. Dentro dessas categorias, destaca-se o grupo de pacientes que se encontra em situação de insuficiência económica. É notável que essa categoria apresenta a taxa mais significativa de isenção. Isso indica que um número considerável de doentes enfrenta desafios económicos que tornam difícil ou impossível assumir com nos custos dos cuidados de saúde.

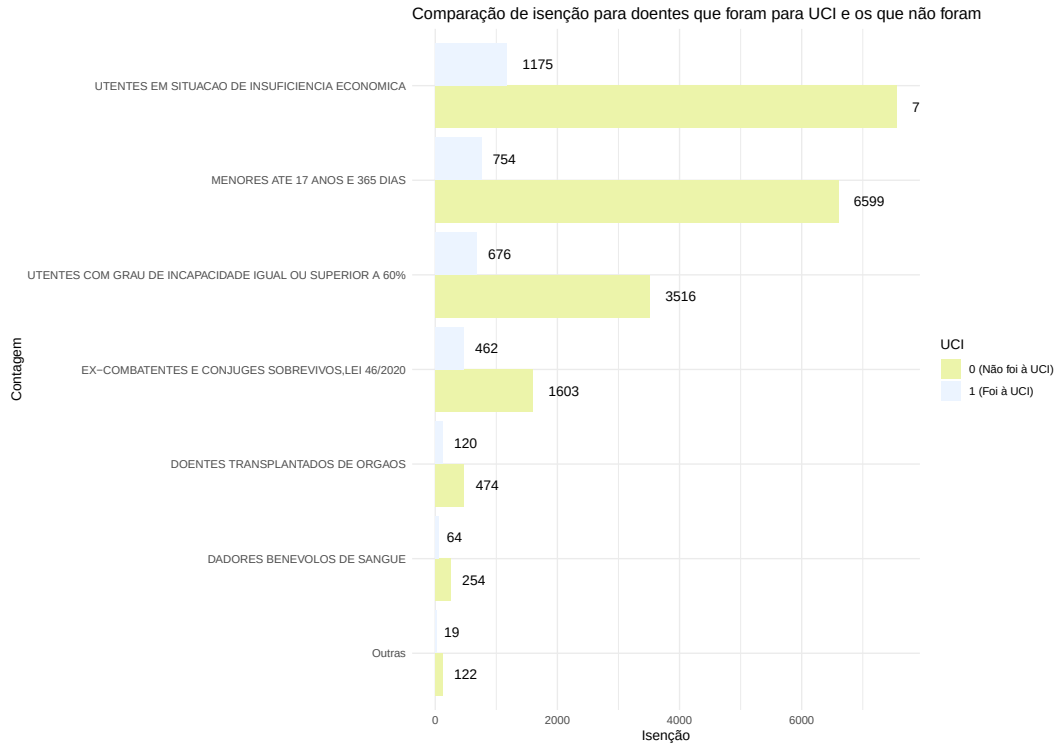


**Figura 3.7:** Comparação da Isenção entre os doentes que foram à UCI e os que não foram

No que diz respeito ao internamento na Unidade de Cuidados Intensivos (UCI), este conjunto de dados mostra que existiu um total de 6.253 casos registados. É evidente que a UCI desempenha um papel fundamental no tratamento de doentes com condições de saúde graves, onde o acompanhamento e cuidados intensivos são essenciais. O tempo médio de internamento na UCI foi de, aproximadamente 4,23 dias, demonstrando a necessidade de um cuidado contínuo e intensivo para estabilizar a condição dos doentes e proporcionar-lhes a assistência médica necessária. No entanto, é de notar que, apesar do tempo médio relativamente curto, alguns pacientes enfrentaram internamentos mais prolongados. O máximo registado foi de 182 dias, um período excecionalmente longo que reflete casos de extrema complexidade e gravidade. Por outro lado, houve casos de internamento de duração mínima, de 0 dias, possivelmente indicando que alguns doentes foram transferidos para a UCI apenas temporariamente ou para avaliação, sem requerer uma permanência prolongada. Estes dados mostram a importância da UCI no tratamento de doentes em estado crítico e a necessidade de uma gestão eficaz do tempo de internamento, adaptada às necessidades clínicas individuais. Além disso, a diversidade de durações no internamento realça a complexidade das condições médicas atendidas na UCI e a necessidade de recursos e técnicos especializados para garantir o melhor cuidado possível para os doentes.

Neste conjunto de dados as especialidades estão divididas em dois grupos distintos: as de admissão e as de alta. Dentro de cada um destes grupos, existe uma subdivisão em especialidade de admissão, que indica o serviço onde o doente foi admitido, e especialidade de admissão prevista, que apontam para o serviço no qual o doente deveria ter sido inicialmente admitido. O mesmo princípio aplica-se às especialidades de alta, onde temos a especialidade

de alta, que regista o serviço de onde o doente teve alta, e a especialidade de alta prevista, que indica o serviço de onde o doente originalmente deveria ter tido alta. A razão para esta distinção reside na realidade hospitalar, onde nem sempre os doentes conseguem ser acomodados no serviço de admissão previsto devido a limitações de capacidade. Como resultado, podem ser admitidos noutras áreas do hospital, embora ainda recebam cuidados e tratamentos da especialidade correspondente. Decidiu-se incluir estas variáveis uma vez que os médicos e enfermeiros priorizam, frequentemente, o atendimento aos doentes dentro dos seus serviços de especialidade designados. Isto pode levar a situações em que doentes admitidos em áreas alternativas não recebem a atenção e cuidados ideais que seriam fornecidos no seu serviço de admissão prevista. Esta distinção permite identificar e monitorizar estes casos, a fim de garantir que todos os doentes recebam a atenção e tratamento adequados para promover o seu bem-estar e recuperação da melhor maneira possível.

| | **Admissão** | | **Alta** | |
|---|---|---|---|---|
| | **Não foi à UCI** | **Foi à UCI** | **Não foi à UCI** | **Foi à UCI** |
| Angiologia e cirurgia vascular | 927 | 187 | 936 | 151 |
| Berçário | 31 | 0 | 2689 | 11 |
| Cardiologia | 1183 | 511 | 1197 | 898 |
| Cardiologia Pediátrica | 203 | 86 | 209 | 92 |
| Cirurgia Cardio-Toracica | 185 | 896 | 183 | 1047 |
| Cirurgia Geral | 3293 | 652 | 3320 | 693 |
| Cirurgia Maxilo-Facial | 608 | 11 | 608 | 35 |
| Cirurgia Pediátrica | 1536 | 133 | 1470 | 144 |
| Cirurgia Plas. e Reconstrutiva | 1078 | 50 | 1031 | 124 |
| Cri. Trat. Cir. Obesidade | 283 | 7 | 304 | 4 |
| Dermatologia | 79 | 3 | 89 | 2 |
| Desconhecido | 842 | 41 | 793 | 54 |
| Doenças Infeciosas | 557 | 34 | 593 | 52 |
| Endocrinologia | 252 | 6 | 242 | 2 |
| Gastrenterologia | 379 | 9 | 387 | 14 |
| Ginecologia | 466 | 1 | 632 | 6 |
| Hematologia Clinica | 245 | 40 | 238 | 13 |
| Med. Física Reabilitação | 16 | 0 | 57 | 30 |
| Medicina Interna | 7138 | 252 | 7250 | 411 |
| Nefrologia | 369 | 14 | 402 | 22 |
| Neonatologia | 10 | 22 | 176 | 251 |
| Neurocirurgia | 661 | 301 | 637 | 425 |
| Neurologia | 861 | 41 | 719 | 37 |
| Obstetricia | 6650 | 301 | 3523 | 3 |
| Oftalmologia | 513 | 4 | 502 | 1 |
| Ortopedia | 2540 | 69 | 2528 | 75 |
| Otorrinolaringologia | 698 | 22 | 705 | 34 |
| Pediatria Médica | 1764 | 64 | 1801 | 157 |

| | Admissão | | Alta | |
|---|---|---|---|---|
| | **Não foi à UCI** | **Foi à UCI** | **Não foi à UCI** | **Foi à UCI** |
| Pedopsiquiatria | 122 | 0 | 156 | 0 |
| Pneumologia | 190 | 12 | 260 | 46 |
| Transplante Hepático | 279 | 106 | 313 | 112 |
| U. Médico Cirúrgica | 3715 | 0 | 3716 | 0 |
| UCI Cardiologia | 0 | 792 | 0 | 281 |
| UCI Cardiotorácica | 0 | 38 | 0 | 44 |
| UCI Neonatologia | 0 | 59 | 0 | 78 |
| UCI Pediatria | 0 | 94 | 0 | 28 |
| UCI Polivalente | 12 | 1275 | 2 | 770 |
| UCI Transplantes | 0 | 1 | 0 | 0 |
| Unidade Queimados | 0 | 72 | 0 | 19 |
| Urologia | 1428 | 47 | 1145 | 87 |

**Tabela 3.4:** Comparação entre os doentes que foram à UCI e os que não foram considerando a sua admissão e alta

A histerectomia é um procedimento cirúrgico que envolve a remoção do útero, é uma intervenção médica realizada com mais frequência em mulheres que se encontram na faixa etária entre os 40 e os 50 anos. Esta cirurgia pode ser indicada para tratar de diversas condições médicas, como miomas uterinos, endometriose, sangramento uterino anormal e cancro do útero. A escolha de realizar uma histerectomia geralmente é baseada nas necessidades médicas específicas da doente. Neste conjunto de dados existem 76 casos onde foi realizada esta cirurgia e só um deles foi à UCI. Considerando assim, esta variável com pouca importância para o objetivo de estudo.

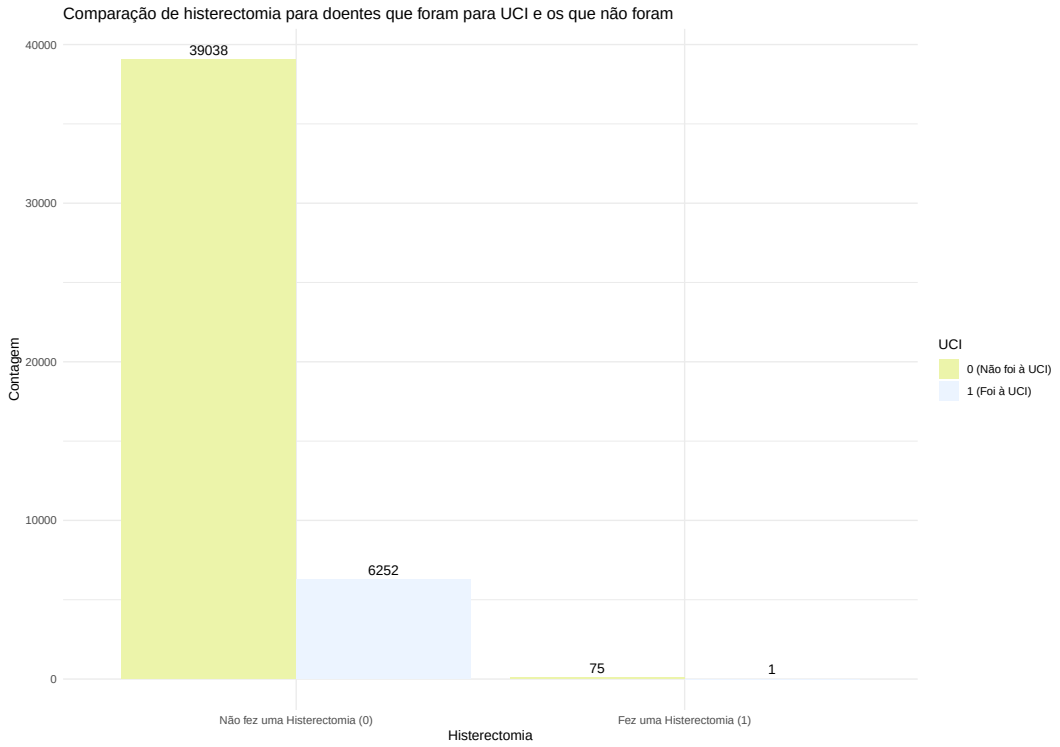


**Figura 3.8:** Comparação entre as doentes que fizeram uma histerectomia e as que não fizeram e entre os casos que foram à UCI e os que não foram

As complicações durante o internamento hospitalar são eventos adversos que podem ocorrer após a admissão do doente e podem não estar relacionados diretamente à condição de saúde que motivou à hospitalização. Estas complicações podem variar na gravidade e na natureza, podendo ser infeções adquiridas no hospital ou até problemas de medicação e reações adversas a tratamentos. Neste conjunto de dados, observou-se que ocorreram complicações num total de 17191 casos durante o período de internamento. É importante realçar que estas complicações podem ter impacto significativo na saúde e no processo de recuperação do doente, exigindo atenção médica adicional e, em alguns casos, prolongando a estadia. Dentro dos casos de complicações, 3.168 deles ocorreram em doentes que estiveram internados na UCI. Com isto podemos notar que é importante a complexidade dos cuidados médicos necessários nos doentes que têm internamentos na UCI, uma vez que estes doentes muitas vezes enfrentam condições de saúde mais críticas e estão sujeitos a riscos mais elevados de complicações durante o seu período de internamento. Portanto, a deteção precoce e a gestão eficaz destas complicações são elementos fundamentais no tratamento destes doentes, ajudando na melhoria da sua condição e do processo de recuperação.

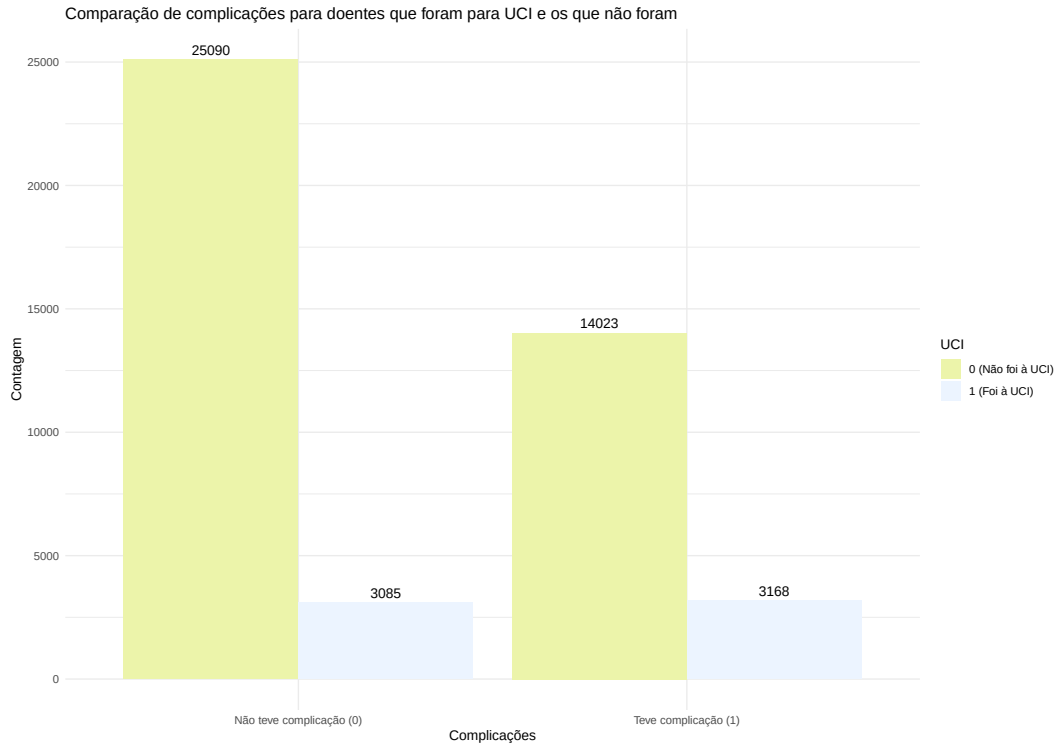


**Figura 3.9:** Comparação entre os doentes que tiveram complicações e os que não tiveram e entre os casos que foram à UCI e os que não foram

Neste conjunto de dados, identificou-se um total de 3.248 recém-nascidos dos quais 291 necessitaram de internamento na UCI durante os primeiros dias ou semanas de vida. Estes números mostram a importância da qualidade da assistência neonatal e dos cuidados médicos de alto nível nas unidades de saúde, garantindo que os recém-nascidos que enfrentam desafios médicos iniciais recebam o apoio e tratamento necessários para uma recuperação bem-sucedida. O acompanhamento cuidadoso e a intervenção médica adequada podem fazer uma diferença significativa na vida desses bebés, assegurando que eles tenham o melhor tratamento para crescerem saudáveis e fortes.

Observou-se um total de 4359 casos correspondentes a cirurgias programadas. Estas cirurgias programadas geralmente envolvem procedimentos que são agendados com antecedência para tratar condições médicas específicas, permitindo aos doentes e profissionais de saúde de se prepararem adequadamente para a intervenção. No entanto, notou-se que 1201 desses casos de cirurgias programadas tenham resultado em internamentos na Unidade de Cuidados Intensivos. Esta informação mostra a complexidade de algumas cirurgias programadas, onde os pacientes podem necessitar de cuidados intensivos no período pós-operatório devido à natureza delicada do procedimento ou às condições médicas às quais foram submetidos.

Já foi referido que um grande número de casos são de doentes que foram internados no hospital através do serviço de urgência, observamos neste conjunto de dados 26332 casos. Esta situação indica a importância das unidades de urgência no sistema de saúde, pois são a porta de entrada para doentes que necessitam de atendimento médico imediato, muitas vezes devido a condições médicas urgentes e graves. No entanto, dos casos que vieram da urgência, apenas

2521 foram internados na UCI. Esta discrepância mostra que nem todos os doentes que chegam à urgência necessitam de cuidados intensivos. Muitos podem ser adequadamente tratados e acompanhados noutros serviços hospitalares, dependendo da natureza da sua condição de saúde. Esta distribuição de casos realça a necessidade da triagem e avaliação adequadas na urgência, para garantir que cada doente seja encaminhado para o local de tratamento mais apropriado. Nem todos os doentes precisam de internamento na UCI, e a capacidade de os direcionar para os serviços corretos é fundamental para otimizar a utilização dos recursos médicos e proporcionar o atendimento adequado a cada indivíduo.

Assim, dentro deste contexto hospitalar, identificamos três tipos distintos de admissão de doentes: Urgente, Programada e Normal. Cada um desses tipos de admissão desempenha um papel fundamental na gestão da assistência médica e na resposta às necessidades dos doentes.

- **Admissão Urgente:** A admissão urgente ocorre quando os pacientes requerem atendimento médico imediato devido à gravidade da sua condição de saúde. Esta categoria é reservada para casos de emergência, onde o tempo é essencial para diagnosticar, estabilizar e tratar do doente.
- **Admissão Programada:** A admissão programada envolve a marcação prévia de um procedimento cirúrgico ou internamento, permitindo que doentes e profissionais de saúde se preparem adequadamente.
- **Admissão Normal:** A admissão normal refere-se à entrada de doentes no hospital devido a condições médicas que não requerem atendimento imediato nem foram previamente programadas. Esses casos podem envolver condições de saúde menos graves que não se qualificam como urgências nem se encaixam em procedimentos programados.

Em cada tipo de admissão é importante atender às diversas necessidades dos doentes e garantir que recebam o atendimento adequado e oportuno. A distinção entre esse tipo de admissão auxilia na organização e alocação dos recursos hospitalares, garantindo que cada um receba a atenção e os cuidados necessários, independentemente da natureza da sua condição médica.

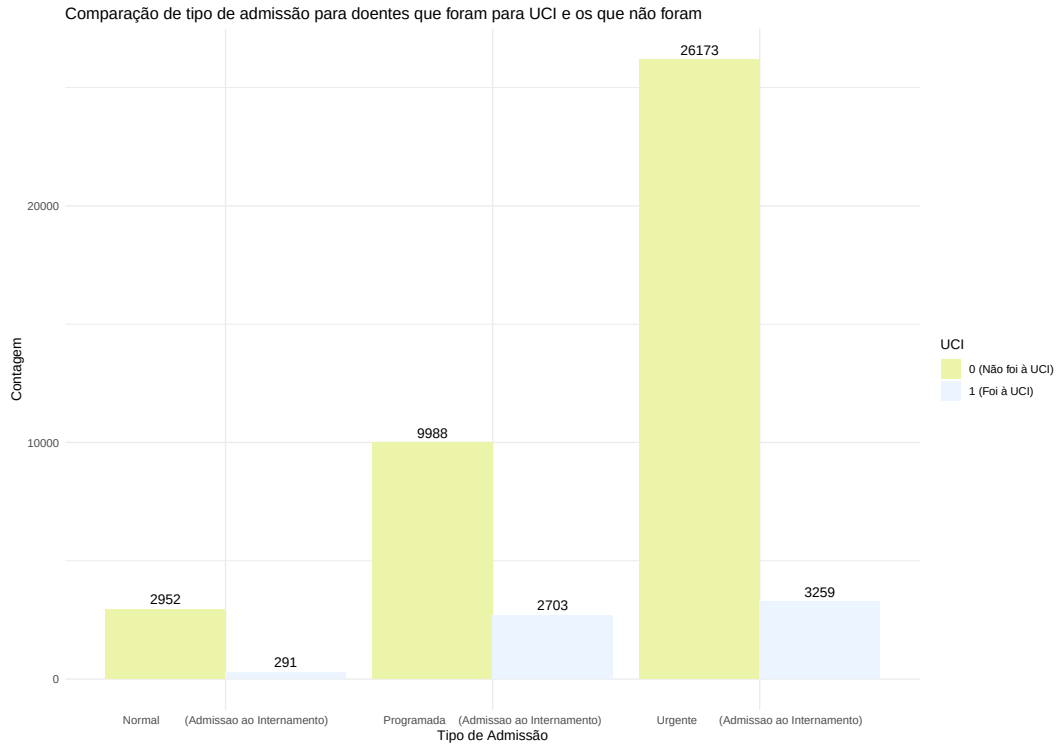


**Figura 3.10:** Comparação entre o tipo de admissão e os casos que foram à UCI e os que não foram

## 3.4 Inferência Estatística

Para analisar a existência de diferenças significativas das temperaturas entre os pacientes que foram admitidos na UCI e aqueles que não foram, avaliou-se as medições de temperatura nas últimas 24 horas, 48 horas e 72 horas antes do internamento. Inicialmente, ia ser realizada uma Análise de Variância (ANOVA) de dois fatores para esse propósito. No entanto, os pressupostos fundamentais de homogeneidade e normalidade não foram cumpridos pelos dados. Devido à violação desses pressupostos, optou-se por uma abordagem não paramétrica, escolhendo a Análise de Variância de Kruskal-Wallis como alternativa. Este teste não paramétrico é adequado quando os pressupostos da ANOVA paramétrica não são cumpridos, e permite a avaliação de diferenças entre grupos sem assumir uma distribuição normal ou homogeneidade de variâncias. Após a realização do teste de Kruskal-Wallis, obtiveram-se resultados que indicaram diferenças estatisticamente significativas nas temperaturas entre os doentes da UCI e aqueles que não foram admitidos. Para identificar quais eram os grupos específicos que apresentavam essas diferenças entre si, aplicou-se o teste de Dunn para comparações múltiplas. Utilizou-se a correção de Bonferroni para controlar o erro tipo I, garantindo assim que as conclusões fossem robustas. Em todas as variáveis de temperatura analisadas, rejeitou-se a hipótese nula. Isso indica que existem diferenças estatisticamente significativas nas medições de temperatura entre os doentes que foram encaminhados para a UCI e aqueles que não foram.

Para as variáveis da saturação de oxigénio, frequência cardíaca, pressão sistólica, pressão diastólica e pressão arterial, realizou-se um procedimento semelhante para avaliar as possíveis diferenças entre os doentes que foram admitidos na UCI e aqueles que não foram. No entanto,

mais uma vez constatou-se que os pressupostos necessários para a aplicação de uma ANOVA de dois fatores não foram cumpridos em nenhuma das variáveis. Diante dessa falha no cumprimento dos pressupostos, optou-se novamente por utilizar a Análise de Variância não paramétrica de Kruskal-Wallis em todas as variáveis. Os resultados revelaram diferenças estatisticamente significativas em quase todas, indicando que os grupos de doentes que foram para a UCI diferem dos que não foram em relação a essas medidas. É importante destacar que apenas nas variáveis 'mediafc24h', 'mediafc48h', 'mediafc72h', 'media_sis24h', 'media_sis72h', 'media_dias48h', 'media_dias72h' e 'media_press72h' não foram encontradas diferenças significativas entre os doentes que foram para a UCI e aqueles que não foram.

CAPÍTULO 4

# Modelos de previsão para o internamento na UCI

Neste capítulo, iremos explorar os modelos de previsão aplicados ao internamento de doentes na Unidade de Cuidados Intensivos. São explorados os modelos do Random Forest e de Regressão Logística em duas vertentes: incluindo todas as variáveis disponíveis (Modelo 1) e com apenas as variáveis consideradas significativas no modelo anterior (Modelo 2).

## 4.1 Random Forest

No contexto deste estudo, e devido à natureza do problema em questão, a escolha inicial do modelo recai sobre o Random Forest, uma abordagem de aprendizagem supervisionada. A nossa variável de interesse é categórica, distinguindo entre os doentes que foram internados na UCI e aqueles que não foram. Com acesso aos dados dos doentes, que incluem diversas variáveis independentes, o nosso objetivo é desenvolver um modelo capaz de prever quais dos doentes podem ser encaminhados para a UCI. O Random Forest destaca-se como uma opção promissora, pois é eficaz na classificação de dados categóricos, considerando múltiplas variáveis independentes quantitativas e qualitativas. Esta abordagem irá ajudar a analisar quais as relações complexas e interações entre as variáveis para tomar decisões informadas e ajudar na gestão de recursos na área da saúde.

Neste problema de classificação, um modelo Random Forest foi construído para prever a variável 'UCI' com base num conjunto de variáveis independentes, consistindo num *ensemble* de 500 árvores de decisão construídas, em cada nó, sobre um conjunto de 5 variáveis independentes escolhidas aleatoriamente do conjunto inicial.

### 4.1.1 Modelo 1

O Modelo 1 corresponde ao modelo onde são consideradas todas as variáveis independentes adequadas ao estudo.

O OOB (*estimate of error rate*) é uma estimativa do erro do modelo que se baseia em observações que não foram usadas durante o treino (*out-of-bag*). Neste caso, a estimativa da taxa de erro OOB é de 12.17%. Isto indica que, em média, o modelo classifica incorretamente cerca de 12.17% das observações.

Aplicando o modelo construído ao conjunto de teste obtém-se a matriz de confusão da Tabela 4.1.

| | | **Referência** | |
|---|---|---|---|
| | | Não admitido na UCI | Admitido na UCI |
| **Previsão** | Não admitido na UCI | 11639 | 1602 |
| | Admitido na UCI | 94 | 273 |

**Tabela 4.1:** Matriz de confusão no conjunto de teste do Modelo 1 do Random Forest

Em face ao não balanceamento das classes do problema e tal como discutido no Capítulo 2 analisamos diferentes medidas de performance do modelo. A Tabela 4.2 apresenta as medidas de performance analisadas.

| **Medidas de Performance** | **Valor** |
|---|---|
| *Accuracy* | 0.875 |
| Sensibilidade | 0.744 |
| Especificidade | 0.879 |
| *F1-Score* | 0.243 |

**Tabela 4.2:** Medidas de Performance para o Modelo 1

A precisão é uma medida através da qual não podemos tirar conclusões, uma vez que estamos perante um problema de classes não balanceadas. Quanto à sensibilidade de 0.854 isto significa que o modelo tem uma sensibilidade muito alta, aproximadamente 85%. Ou seja, o modelo é capaz de identificar corretamente cerca de 99% dos casos que realmente são positivos. Uma sensibilidade de 0.854 indica que o modelo é muito bom em identificar corretamente a maioria dos casos positivos. Neste caso, a especificidade é de aproximadamente 99.2% o que significa que o modelo é eficaz na identificação dos casos negativos, com uma taxa alta de verdadeiros negativos (TN), ou seja, ele classifica a maioria dos casos reais da classe negativa corretamente. Neste modelo o *F1-Score* obtido foi de 0.932, este valor é próximo de 1, o que diz que o modelo tem um bom equilíbrio entre a capacidade de fazer previsões positivas corretas (alta precisão) e identificar corretamente casos positivos (alta sensibilidade). Por outras palavras, este modelo mostra uma capacidade sólida em acertar as previsões positivas (verdadeiros positivos) enquanto mantém os falsos positivos num nível razoavelmente baixo. Um *F1-Score* elevado, como este, é um indicativo positivo de que o modelo está realizar bem a função de classificação e é particularmente relevante em cenários onde é importante evitar tanto falsos positivos como falsos negativos.

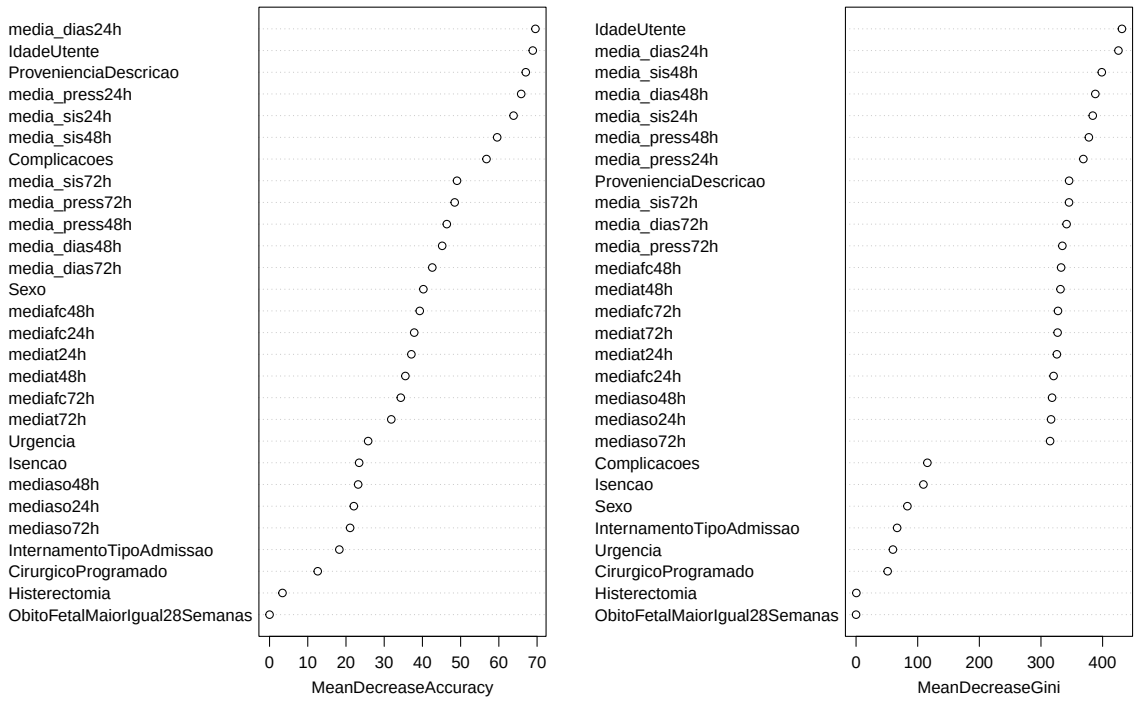


**Figura 4.1:** Importância das variáveis no modelo de Random Forest

O "*MeanDecreaseGini*"é uma medida que indica o quanto cada variável contribui para a redução da impureza nos nós das árvores de decisão do modelo Random Forest. Quanto maior o valor, mais importante a variável é na tomada de decisões do modelo.

A *MeanDecreaseAccuracy* representa a média da diminuição na precisão do modelo ao longo de várias iterações, onde, em cada iteração, uma característica é removida e a precisão do modelo é calculada novamente. A ideia por trás desta medida é que as características mais importantes terão um impacto maior na precisão do modelo quando removidas, enquanto as características menos importantes terão um impacto menor.

A análise da importância das variáveis indica quais características ou variáveis que têm maior impacto nas previsões do modelo. A Figura 4.1 apresenta a importância das variáveis de acordo com dois índices. Para identificar quais as variáveis com maior importância foram observadas as duas medidas e aquelas que estariam em comum como as menos importantes foram removidas no modelo seguinte. Ao analisar os resultados, é evidente que a "Idade do Utente"é uma das variáveis mais influentes, com uma contribuição significativa para a capacidade do modelo tomar decisões precisas. Além disso, variáveis relacionadas a medições de sinais vitais, como pressão arterial e frequência cardíaca, também desempenham um papel importante. Por outro lado, variáveis como "Obito Fetal Maior ou Igual a 28 Semanas", "Histerectomia", "CirurgicoProgramado", "InternamentoTipoAdimssao", "Isencao"e "Urgencia"têm uma importância muito reduzida no modelo.

Esta interpretação sugere que a idade do paciente e as medições dos sinais vitais são os principais fatores que influenciam as previsões do modelo em relação à necessidade de cuidados

intensivos. Portanto, ao aprimorar o modelo ou tomar decisões com base nas suas previsões, é importante considerar o ênfase dado a estas variáveis mais importantes e possivelmente avaliar a inclusão de variáveis com menor impacto.

### 4.1.2 Modelo 2

O modelo aqui apresentado consiste na reaplicação do método Random Forest mas desconsiderando as variáveis independentes identificadas como menos importantes na análise acima. As variáveis que foram removidas são as seguinte: "Obito Fetal Maior ou Igual a 28 Semanas", "Histerectomia", "CirurgicoProgramado", "InternamentoTipoAdimssao", "Isencao"e "Urgencia"considerando a pouca importância das mesmas no modelo anterior.

Neste caso, a *OOB estimate of error rate* é de 12.59%, o que significa que, em média, o modelo classifica incorretamente cerca de 12.59% das observações que não foram usadas durante o treino. Isto diz que que o modelo tem um desempenho geral razoavelmente bom, embora sempre haja espaço para melhoria. O modelo anterior tem esta taxa superior.

A matriz de confusão está representada na Tabela 4.3.

| | | **Referência** | |
|---|---|---|---|
| | | Não admitido na UCI | Admitido na UCI |
| **Previsão** | Não admitido na UCI | 11646 | 1559 |
| | Admitido na UCI | 87 | 316 |

**Tabela 4.3:** Matriz de confusão no conjunto de teste do Modelo 2 do Random Forest

A Tabela 4.4 representa as medidas de performance analisadas para este modelo.

| **Medidas de Performance** | **Valor** |
|---|---|
| *Accuracy* | 0.879 |
| Sensibilidade | 0.784 |
| Especificidade | 0.882 |
| *F1-Score* | 0.277 |

**Tabela 4.4:** Medidas de Performance para o Modelo 1

O modelo de classificação acertou aproximadamente em 87.5% das previsões em relação ao total de observações, ou seja, ele classificou corretamente cerca de 87.5% das observações, independentemente de serem da classe 0 ou da classe 1. Vemos que em comparação com o primeiro modelo o valor aumentou. O valor da sensibilidade medido foi de 0.831 o que significa que o modelo tem uma capacidade alta de identificar corretamente os casos da classe 1 em relação ao número total de casos positivos reais. A especificidade deu um valor de 0.992, o que quer dizer que o modelo tem uma boa capacidade de identificar de forma correta os casos da classe 0 em relação ao número total de casos da classe 0 totais. Da combinação da sensibilidade e da especificidade surge o *F1-Score*, que neste caso tem o valor de 0.934.

Isto significa que o modelo é eficaz na minimização tanto dos falsos positivos como dos falsos negativos, resultando num desempenho geral sólido.

## 4.2 Regressão Logística

A escolha da Regressão Logística como método para prever a probabilidade de um doente ser admitido na Unidade de Cuidados Intensivos (UCI) foi porque, a Regressão Logística é uma técnica muito utilizada em problemas de classificação, especialmente em contexto médico, devido à sua capacidade de lidar com previsões binárias, como "ir para a UCI"ou "não ir para a UCI."Esta abordagem estatística é eficaz em modelar relações entre variáveis independentes, como idade, histórico médico, sinais vitais e outras características clínicas, e a variável dependente que representa a admissão na UCI.

Os resultados da regressão logística fornecem uma análise das variáveis que influenciam a probabilidade de um paciente ser admitido na UCI. Variáveis com coeficientes significativos e *p-values* baixos são consideradas mais influentes nesta previsão.

A Regressão Logística possui uma capacidade de fazer previsões através de probabilidades, o que é fundamental em situações médicas onde é necessário avaliar riscos e tomar decisões informadas. Além disso, a interpretabilidade da Regressão Logística permite compreender como cada variável contribui para as previsões, o que é importante no contexto médico para identificar os fatores de risco e tomar medidas preventivas.

### 4.2.1 Modelo 1

O Modelo 1 corresponde ao modelo onde estão todas as variáveis independentes que foram consideradas no Modelo 1 do Random Forest.

Na Tabela 4.5 podemos observar as estimativas dos coeficientes, os respetivos desvios-padrão e do valores de *p-value*.

| **Coeficientes** | **Estimativa** | **Desvio-Padrão** | ***P-value*** |
|---|---|---|---|
| *Intercept* | 22.637 | 502.143 | 0.964 |
| mediaso72h | 0.015 | 0.013 | 0.249 |
| mediaso48h | 0.010 | 0.013 | 0.459 |
| mediaso24h | 0.006 | 0.013 | 0.607 |
| mediafc72h | 0.008 | 0.002 | 0.001 |
| mediafc48h | 0.002 | 0.002 | 0.341 |
| mediafc24h | 0.004 | 0.002 | 0.061 |
| mediat72h | 0.048 | 0.086 | 0.576 |
| mediat48h | -0.224 | 0.093 | 0.016 |
| mediat24h | -0.209 | 0.089 | 0.019 |
| media_press72h | -0.073 | 0.068 | 0.284 |
| media_press48h | -0.191 | 0.085 | 0.025 |
| media_press24h | -0.023 | 0.072 | 0.745 |
| media_dias72h | 0.054 | 0.045 | 0.233 |
| media_dias48h | 0.137 | 0.057 | 0.016 |

| Coeficientes | Estimativa | Desvio-Padrão | *P-value* |
|---|---|---|---|
| media_dias24h | -0.014 | 0.048 | 0.765 |
| media_sis72h | 0.022 | 0.023 | 0.326 |
| media_sis48h | 0.052 | 0.028 | 0.068 |
| media_sis24h | 0.017 | 0.024 | 0.476 |
| IdadeUtente | 0.004 | 0.001 | 0.0003 |
| Sexo | -0.580 | 0.036 | <2e-16 |
| ObitoFetalMaiorIgual28Semanas1 | -12.529 | 210.665 | 0.953 |
| ProvenienciaDescricaoEXTERIOR | 0.530 | 0.150 | 0.0004 |
| ProvenienciaDescricaoHOSPITAL DIA | -0.316 | 0.262 | 0.228 |
| ProvenienciaDescricaoOutras | 1.204 | 0.185 | 6.97e-11 |
| ProvenienciaDescricaoOUTRO HOSPITAL | 2.190 | 0.087 | <2e-16 |
| ProvenienciaDescricaoRECEM-NASCIDO | -12.605 | 502.136 | 0.980 |
| ProvenienciaDescricaoURGENCIA | 0.061 | 0.298 | 0.837 |
| IsencaoDADORES BENEVOLOS DE SANGUE | 0.149 | 0.180 | 0.406 |
| IsencaoDOENTES TRANSPLANTADOS DE ORGAOS | 0.220 | 0.132 | 0.095 |
| IsencaoEX-COMBATENTES E CONJUGUES SOBREVIVOS | 0.194 | 0.076 | 0.010 |
| IsencaoMENORES ATE 17 ANOS E 365 DIAS | -0.610 | 0.089 | 7.48e-12 |
| IsencaoOutras | -0.629 | 0.358 | 0.079 |
| IsencaoUTENTES COM GRAU DE INCAPACIDADE IGUAL OU SUPERIOR A 60 % | 0.069 | 0.060 | 0.248 |
| IsencaoUTENTES EM SITUACAO DE INSUFICIENCIA ECONOMICA | -0.039 | 0.047 | 0.414 |
| Histerectomia1 | -13.087 | 117.237 | 0.911 |
| Complicacoes1 | 0.554 | 0.041 | <2e-16 |
| CirurgicoProgramado1 | 0.272 | 0.06 | 1.97e-05 |
| Urgencia | -0.569 | 0.294 | 0.053 |
| InternamentoTipoAdmissaoProgramada | -12.505 | 502.136 | 0.980 |
| InternamentoTipoAdmissaoUrgente | -12.627 | 502.136 | 0.980 |

**Tabela 4.5:** Coeficientes estimados e respetivos *p-values* do Modelo 1 da Regressão Logística

Os coeficientes das variáveis independentes representam o efeito que cada variável tem sobre a *log-odds* da admissão na UCI, mantendo todas as outras variáveis constantes. Valores positivos indicam que há um aumento na probabilidade de admissão na UCI com o aumento da variável, enquanto valores negativos indicam que há uma diminuição dessa probabilidade. Os *p-values* estão associados a cada coeficiente e indicam se a variável é estatisticamente significativa na previsão da admissão na UCI. *p-values* menores que 0,05 são geralmente considerados significativos. Variáveis com *p-values* significativamente baixos são consideradas influentes na previsão.

Nas medições de oxigénio dos diferentes intervalos de tempo podemos observar através dos valores dos *p-values* que estas variáveis não são estatisticamente significativas para o nosso modelo. O mesmo podemos observar para as medidas de frequência cardíaca nas últimas

48 e 24 horas antes do internamento, para a medida de temperatura nas 72 horas antes do internamento, para as medidas de pressão arterial sistólica, diastólica e média nas últimas 72 horas e 24 horas antes do internamento, para os casos de óbito fetal com 28 ou mais semanas, para os casos de histerectomia e aqueles que provieram das urgências e para os tipos de admissão no internamento.

A variáveis "mediaso72h", "mediaso48h"e "mediaso24h"representam as medições da saturação de oxigénio nos diferentes intervalos de tempo (72 horas, 48 horas e 24 horas antes do internamento). Os coeficientes estimados são positivos (0.015, 0.010 e 0.006, respetivamente), indicando que um aumento nas medições da saturação de oxigénio está associado a um aumento na probabilidade da admissão na UCI. Os *p-values* são 0.249, 0.459 e 0.607, respetivamente, o que indica que estas variáveis não são estatisticamente significativas na previsão da admissão na UCI.

A variáveis "mediafc72h", "mediafc48h"e "mediafc24h"representam as medições da frequência cardíaca nos diferentes intervalos de tempo (72 horas, 48 horas e 24 horas antes do internamento). Os coeficientes estimados são todos positivos indicando que um aumento nas medições da frequência cardíaca está associado a um aumento na probabilidade de admissão na UCI. Os *p-values* são 0.001, 0.341 e 0.061, respetivamente, o que indica que nestas três variáveis apenas as medições da frequência cardíaca nas últimas 72 horas é uma variável significativa.

Podemos analisar que as variáveis "mediat48h"e "mediat24h"têm o valor do coeficiente negativo (-0.224 e -0.209, respetivamente) indicando que um doente com temperatura corporal mais baixa tem uma probabilidade maior de ser admitido na UCI.

Podemos analisar pelos valores do *p-value* que nenhuma das medições da pressão arterial (média, sistólica e diastólica) é estatisticamente significativa na previsão da admissão de doentes na UCI, com *p-values* acima de 0,05 como representado na Tabela 4.5.

Os resultados da regressão logística indicam que a proveniência dos pacientes pode afetar a probabilidade de admissão na UCI, com algumas proveniências, como "EXTERIOR", "Outras"e "OUTRO HOSPITAL", tendo um impacto significativamente maior, enquanto proveniências, como "HOSPITAL DIA", "RECEM-NASCIDOURGENCIA", não apresentam evidências estatísticas significativas de influência na admissão na UCI. As proveniências "HOSPITAL DIA"e "RECEM-NASCIDO"têm os coeficientes negativos indicando que estas proveniências não tem um impacto significativo na probabilidade de admissão na UCI. No entanto, o alto desvio padrão na proveniência dos recém-nascidos indica uma grande variabilidade nos dados.

A idade do doente é estatisticamente significativa na previsão da admissão na UCI, com um *p-value* de 0.0003.

O coeficiente para a variável "Sexo"na regressão logística é altamente significativo, com um p-*value* muito baixo ($< 2e\text{-}16$), indicando a sua forte influência na previsão da admissão na UCI. O coeficiente negativo (-0,580) indica que ser do sexo feminino está associado a uma probabilidade menor de admissão na UCI em comparação com o sexo masculino. Ou seja, o sexo do paciente é um fator importante na determinação da probabilidade de admissão na UCI, com uma influência negativa, ou seja, os doentes do sexo feminino têm menor probabilidade

de serem admitidos na UCI comparado com os doentes do sexo masculino.

A análise destes resultados indica que algumas categorias de isenção de pagamento, como "EX-COMBATENTES E CONJUGUES SOBREVIVOS"pode estar associada a um aumento estatisticamente significativo na probabilidade de admissão na UCI, enquanto a categoria "MENORES ATE 17 ANOS E 365 DIAS"está associada a uma redução significativa desta probabilidade. Outras categorias de isenção não apresentam impacto estatisticamente significativo na probabilidade de admissão na UCI.

As variáveis "Histerectomia"e "InternamentoTipoAdmissao"não são estatisticamente significativas para este problema, como podemos analisar na Tabela 4.5 através dos valores do *p-value*.

Na variável "Complicações1"existe um coeficiente positivo altamente significativo de 0,554, indicando que a presença de complicações aumenta a probabilidade de admissão na UCI.

Quanto à variável "Cirúrgico Programado1"que tem um coeficiente positivo significativo de 0,272, indicando que os doentes que têm cirurgia programada têm uma probabilidade significativamente maior de serem admitidos na UCI.

Na variável "Urgência"temos um coeficiente negativo de -0,569, no entanto podemos analisar o seu *p-value* que é de 0.053, indicando que os doente que vieram da urgência não têm um impacto estatisticamente significativo na probabilidade de admissão na UCI.

Os coeficientes associados às variáveis "InternamentoTipoAdmissaoProgramada"e "InternamentoTipoAdmissaoUrgente"estão relacionados com os tipos de admissão para o internamento. No entanto, estes coeficientes apresentam valores extremamente altos e negativos, juntamente com *p-values* muito elevados (0.980, em ambos). Isto indica que estas variáveis não têm um impacto estatisticamente significativo na probabilidade de admissão na UCI.

A matriz de confusão para este Modelo 1 está representada na Tabela 4.6.

| | | **Referência** | |
|---|---|---|---|
| | | Não admitido na UCI | Admitido na UCI |
| **Previsão** | Não admitido na UCI | 11613 | 1636 |
| | Admitido na UCI | 120 | 239 |

**Tabela 4.6:** Matriz de confusão no conjunto de teste do Modelo 1 da Regressão Logística

A Tabela 4.10 representa as medidas de performance analisadas para este modelo.

| **Medidas de Performance** | **Valor** |
|---|---|
| *Accuracy* | 0.871 |
| AUC | 0.730 |
| Sensibilidade | 0.666 |
| Especificidade | 0.876 |
| *F1-Score* | 0.214 |

| Medidas de Performance | Valor |
|---|---|

**Tabela 4.7:** Medidas de Performance para o Modelo 1

Como podemos analisar na Tabela 4.10 o valor da *accuracy* é de 0.871, o que significa que de todas as previsões feitas pelo modelo, cerca de 87.10% estão corretas. Isso é geralmente considerado um desempenho razoavelmente bom, especialmente em problemas de classificação. O valor da AUC (Area Under the Curve) é de 0.73 o que indica que o modelo de regressão logística demonstra um desempenho bom na previsão de quais os doentes que podem ser admitidos na Unidade de Cuidados Intensivos com base nas variáveis disponíveis. Quanto mais próximo o valor do AUC estiver de 1, melhor é a capacidade do modelo de discriminar entre doentes que vão para a UCI e aqueles que não vão. Tendo em conta a matriz de confusão da Tabela 4.6 podemos obter os seguintes valores, de precisão aproximadamente 0.127, de sensibilidade aproximadamente 0.666 e portanto de F1 de 0.217. Isto significa que o modelo de regressão logística não tem um desempenho muito bom na classificação. É um valor baixo, o que indica que o modelo pode estar a ter dificuldade em equilibrar a precisão e a sensibilidade, isto pode indicar que o modelo está com tendência de criar mais falsos positivos ou falsos negativos. Idealmente, queremos que o valor de F1 seja o mais próximo de 1 possível, o que indicaria um modelo com alta capacidade de fazer previsões corretas positivas e identificar todos os casos positivos. Portanto, um valor de F1 de aproximadamente 0.217 indica que há espaço para melhorias no desempenho do modelo. Talvez seja necessário ajustar os parâmetros do modelo, usar *features* diferentes ou adotar outras estratégias para melhorar a sua capacidade de classificação.

### 4.2.2 Modelo 2

O modelo que será aqui apresentado consiste na reaplicação da Regressão Logística mas desconsiderando as variáveis independentes identificadas como não significativas na análise anterior. As variáveis que foram removidas são as seguintes: "mediaso72h", "mediaso48h", "mediaso24h", "mediafc24h", "mediafc48h", "mediat72h", "media_press72h", "media_press24h", "media_dias72h", "media_dias24h", "media_sis72h", "media_sis48h", "media_sis24h", "Obito-FetalMaiorIgual28Semanas", "Histerectomia", "Urgencia"e "InternamentoTipoAdmissao".

Na Tabela 4.8 podemos observar as estimativas dos coeficientes, os respetivos desvios-padrão e os valores dos *p-values*.

| **Coeficientes** | **Estimativa** | **Desvio-Padrão** | ***P-value*** |
|---|---|---|---|
| *Intercept* | 12.122 | 2.083 | 5.87e-09 |
| mediafc72h | 0.0.12 | 0.001 | 1.83e-14 |
| mediat24h | -0.264 | 0.085 | 0.002 |
| media_press48h | -0.115 | 0.081 | 0.155 |
| media_dias48h | 0.068 | 0.089 | 0.232 |
| media_sis48h | 0.027 | 0.060 | 0.251 |
| IdadeUtente | 0.005 | 0.036 | 1.44e-07 |
| Sexo | -0.558 | 0.030 | <2e-16 |

| Coeficientes | Estimativa | Desvio-Padrão | *P-value* |
|---|---|---|---|
| ProvenienciaDescricaoEXTERIOR | 0.407 | 0.001 | 0.006 |
| ProvenienciaDescricaoHOSPITAL DIA | -0.325 | 0.147 | 0.192 |
| ProvenienciaDescricaoOutras | 1.218 | 0.249 | 1.99e-11 |
| ProvenienciaDescricaoOUTRO HOSPITAL | 2.047 | 0.081 | <2e-16 |
| ProvenienciaDescricaoRECEM-NASCIDO | -0.078 | 0.105 | 0.454 |
| ProvenienciaDescricaoURGENCIA | -0.558 | 0.043 | <2e-16 |
| IsencaoDADORES BENEVOLOS DE SANGUE | -0.017 | 0.193 | 0.929 |
| IsencaoDOENTES TRANSPLANTADOS DE ORGAOS | 0.194 | 0.135 | 0.151 |
| IsencaoEX-COMBATENTES E CONJUGUES SOBREVIVOS | 0.250 | 0.075 | 0.001 |
| IsencaoMENORES ATE 17 ANOS E 365 DIAS | -0.472 | 0.088 | 8.41e-08 |
| IsencaoOutras | -0.320 | 0.305 | 0.294 |
| IsencaoUTENTES COM GRAU DE INCAPACIDADE IGUAL OU SUPERIOR A 60 % | 0.102 | 0.059 | 0.081 |
| IsencaoUTENTES EM SITUACAO DE INSUFICIENCIA ECONOMICA | -0.005 | 0.047 | 0.917 |
| Complicacoes1 | 0.597 | 0.041 | <2e-16 |
| CirurgicoProgramado1 | 0.291 | 0.06 | 1.30e-06 |

**Tabela 4.8:** Coeficientes estimados e respetivos p-values do Modelo 2 da Regressão Logística

A matriz de confusão para este modelo está representada na Tabela 4.9.

| | | **Referência** | |
|---|---|---|---|
| | | Não admitido na UCI | Admitido na UCI |
| **Previsão** | Não admitido na UCI | 11639 | 1668 |
| | Admitido na UCI | 94 | 207 |

**Tabela 4.9:** Matriz de confusão no conjunto de teste do Modelo 2 da Regressão Logística

Observamos pela Tabela 4.9 11639 casos corretamente classificados como verdadeiros negativos e 207 casos corretamente classificados como verdadeiros positivos. Temos 1668 casos mal classificados chamados de falsos positivos e 94 casos mal classificados chamados de falsos negativos. Esta matriz é útil para avaliar o desempenho do modelo, calculando medidas de performance como *accuracy*, sensibilidade e especificidade.

A Tabela 4.10 representa as medidas de performance analisadas para este modelo.

| **Medidas de Performance** | **Valor** |
|---|---|
| *Accuracy* | 0.871 |
| AUC | 0.701 |
| Sensibilidade | 0.740 |
| Especificidade | 0.875 |
| *F1-Score* | 0.232 |

**Tabela 4.10:** Medidas de Performance para o Modelo 2

A *accuracy* é aproximadamente 0,871, o que indica que o modelo tem uma taxa de previsões corretas de cerca de 87,1%. A sensibilidade é aproximadamente 0,110, o que indica que o modelo consegue identificar corretamente cerca de 11,0% dos casos positivos reais. A especificidade é aproximadamente 0,992, o que indica que o modelo consegue identificar corretamente cerca de 99,2% dos casos negativos reais. O F1-Score é aproximadamente 0,109, o que indica que o modelo não tem um bom equilíbrio entre a precisão e a sensibilidade. Isto significa que o modelo está com tendência a criar mais falsos positivos ou falsos negativos. Pela análise das medidas de performance este modelo não apresenta um bom desempenho para ser considerado.

## 4.3 Comparação entre os modelos da Regressão Logística e do Random Forest

Nesta secção vamos fazer uma análise comparativa entre os modelos obtidos no Random Forest e os modelos obtidos na Regressão Logística. Vamos comparar as medidas de performance obtidas entre todos os modelos e as variáveis que cada um deles considerou importantes.

Na Tabela 4.11 encontram-se todas as medidas de performance dos modelos.

| | **Random Forest** | | **Regressão Logística** | |
|---|---|---|---|---|
| **Medidas de Performance** | **Modelo 1** | **Modelo 2** | **Modelo 1** | **Modelo 2** |
| ***Accuracy*** | 0.875 | 0.879 | 0.871 | 0.871 |
| **Sensibilidade** | 0.744 | 0.784 | 0.666 | 0.740 |
| **Especificidade** | 0.879 | 0.882 | 0.876 | 0.875 |
| ***F1-Score*** | 0.243 | 0.277 | 0.214 | 0.232 |

**Tabela 4.11:** Medidas de Performance dos modelos obtidos

Para o Modelo 1, ambos os algoritmos, Random Forest e Regressão Logística, demonstraram boas performances em termos de *accuracy*, alcançando valores de 0.875 e 0.871, respetivamente. No entanto, quando se observa a Sensibilidade, o Modelo 1 baseado no Random Forest tens melhores resultados que o modelo de Regressão Logística, indicando uma melhor capacidade de identificar corretamente casos positivos. Por outro lado, a Especificidade do Modelo 1 é significativamente superior no caso de Regressão Logística, destacando a capacidade desse modelo em evitar falsos positivos. O *F1-Score*, que considera a precisão e a sensibilidade, também favorece o Modelo 1 para ambas as abordagens. De forma geral, podemos considerar o modelo do Random Forest melhor para retirar conclusões para este problema, apesar da *accuracy* ser idêntica em ambos os modelos, estamos perante um caso de classes não balanceadas, assim a accuracy não é a medida ideal para este tipo de problemas.

Relativamente às variáveis que foram consideradas mais importantes, no modelo do Random Forest destacaram-se as medições dos sinais vitais, a idade do doente e a proveniência

do mesmo, enquanto na Regressão Logística nem todos os sinais vitais foram destacados, foram consideradas como as variáveis mais importantes a médias das medições de frequência cardíaca nas últimas 72 horas, as médias das medidas de temperatura nas últimas 24 e 48 horas, a média das medidas da pressão sistólica e diastólica nas últimas 48 horas, a idade, o sexo do doente, a proveniência, a isenção e as complicações. Podemos observar que em ambos os casos a proveniência do doente e a idade são variáveis que temos de ter em conta neste estudo visto que têm grande importância para a previsão dos casos que precisam de ir à UCI.

Comparando ambos os modelos e as suas medidas de performance podemos considerar o Modelo do Random Forest como uma melhor opção para este estudo.

CAPÍTULO 5

# Conclusões

Este estudo demonstrou o potencial significativo dos modelos de *Machine Learning* na previsão das admissões dos doentes na Unidade de Cuidados Intensivos. Ao longo deste trabalho foram analisados dados de um hospital português e aplicadas duas técnicas de *Machine Learning* para desenvolver modelos de previsão robustos. Os resultados obtidos revelam *insights* valiosos e mostram a eficácia destas abordagens na identificação dos doentes em risco de admissão para a UCI. À medida que surgem avanços na medicina estes modelos têm o poder de melhorar a triagem, o planeamento de recursos e, o mais importante, a qualidade do atendimento aos doentes.

Neste estudo, foi tomada a decisão de usar duas técnicas de *Machine Learning* para a classificação, de forma acumprir os objetivos do desafio da previsão das admissões na Unidade de Cuidados Intensivos. As duas técnicas escolhidas para impulsionar esta análise foram o Random Forest e a regressão logística. A utilização destas duas abordagens distintas permitiu uma investigação mais robusta e abrangente, aproveitando as vantagens únicas de cada uma. O Random Forest, com sua capacidade de lidar com conjuntos de dados complexos e variáveis importantes, ofereceu resultados valiosos sobre as relações entre os diferentes fatores. A regressão logística, com sua interpretabilidade e capacidade de modelar relações lineares, complementou o quadro, fornecendo uma visão mais clara das relações entre as variáveis. Esta abordagem de duas técnicas diferentes aprimorou a robustez dos resultados, reforçando a importância das previsões e da utilidade das mesmas.

Foram desenvolvidos dois modelos de Random Forest como parte do processo de análise. O primeiro modelo, denominado Modelo 1, incluiu todas as variáveis em estudo. Este modelo foi projetado como uma base inicial para a previsão das admissões dos doentes na Unidade de Cuidados Intensivos. De seguida, para uma análise mais aperfeiçoada, foi criado o Modelo 2. Neste segundo modelo, foram removidas as variáveis que foram identificadas como não significativas no Modelo 1. Isto teve como objetivo verificar se a exclusão destas variáveis resultaria em diferenças significativas e se haveria alguma melhoria no desempenho do modelo. A análise comparativa entre os dois modelos revelou resultados valiosos sobre quais as variáveis

que desempenham um papel significativo na previsão das admissões na UCI e quais podem ser consideradas menos influentes. Esta abordagem contribuiu para a otimização do modelo e ajudou a refinar a precisão das previsões, ao mesmo tempo em que economizou recursos computacionais ao eliminar variáveis que não contribuíram significativamente para a melhoria do desempenho do modelo.

No entanto, após a análise, constatou-se que, surpreendentemente, ao remover estas variáveis no Modelo 2, o desempenho do modelo piorou, mas muito pouco, nas medidas de performance analisadas. Isto indica que, embora essas variáveis possam ter sido inicialmente classificadas como não significativas no Modelo 1, elas desempenham, na verdade, um papel importante na precisão das previsões. Estas variáveis aparentemente não significativas, no Modelo 1 contribuíram para a estabilidade e robustez do modelo, destacando a complexidade das relações entre as variáveis em estudo.

Essa descoberta destaca a importância da análise criteriosa e da interpretação dos resultados em projetos de *Machine Learning*. Mostra que, mesmo as variáveis que parecem menos influentes num contexto inicial podem desempenhar um papel importante na precisão das previsões. Portanto, a abordagem de inclusão de todas as variáveis pode ser mais vantajosa do que a exclusão precipitada de variáveis inicialmente consideradas não significativas.

Com base na análise da importância das variáveis neste estudo, é evidente que a 'Idade do Utente' é destacada como a variável mais significativa, tendo assim uma grande importância na capacidade do modelo realizar previsões precisas em relação à necessidade de cuidados intensivos. A idade do paciente é um fator crítico, porque reflete a vulnerabilidade e os riscos associados a diferentes grupos etários, desempenhando um papel de destaque na tomada de decisões clínicas.

Além disso, as variáveis relacionadas com as medições dos sinais vitais, como pressão arterial e frequência cardíaca, também se destacaram como variáveis importantes na previsão da admissão na Unidade de Cuidados Intensivos. Estas medições são indicativos diretos do estado de saúde do doente e desempenham um papel fundamental na identificação de situações que requerem cuidados intensivos.

Por outro lado, variáveis como "Obito Fetal Maior ou Igual a 28 Semanas"e "Histerectomia"apresentam uma importância consideravelmente reduzida no modelo. Isto indica que, no contexto deste estudo, estas variáveis têm um impacto mínimo nas previsões relacionadas à admissão na UCI. Uma das razões pode ter sido a existência de poucas observações onde estes casos ocorreram.

O ênfase adequado às variáveis que tiveram maior importância, como a idade e as medições dos sinais vitais, pode melhorar a precisão das previsões e a capacidade de intervenção precoce em pacientes de risco. No entanto, não se deve descartar completamente as variáveis menos influentes, porque elas podem conter informações valiosas em cenários específicos ou quando combinadas com outras variáveis.

Em resumo, esta análise da importância das variáveis não nos indica apenas a otimização do modelo, mas também realça a complexidade das decisões clínicas e a importância de considerar uma ampla gama de variáveis ao determinar a necessidade de admissão nos

cuidados intensivos. Ela proporciona um resultado valioso para os profissionais de saúde e os investigadores, contribuindo para um atendimento mais informado e personalizado a cada doente.

Com base nas informações obtidas pelos dois modelos de regressão logística e as medidas de performance associadas, podemos concluir que no primeiro modelo a variável "Proveniencia-Descricao"tem um efeito significativo no modelo. As categorias "Exterior", "Outras"e "OUTRO HOSPITAL"têm coeficientes positivos significativos, indicando que estas proveniências estão associadas a um maior risco em relação à categoria de referência (CONSULTA EXTERNA). A *accuracy* do modelo é razoável, mas a baixa precisão e sensibilidade indicam que há espaço para melhorias no equilíbrio entre falsos positivos e falsos negativos. Concluímos assim que a proveniência pode ser um fator importante na admissão na Unidade de Cuidados Intensivos.

No segundo modelo uma das variáveis que tem um efeito significativo no modelo é o "Sexo". O coeficiente negativo indica que o sexo feminino tem uma menor probabilidade de ser admitido no UCI comparado com o sexo masculino.

Ambos os modelos apresentam algumas limitações na identificação correta dos casos positivos, o que é importante em problemas de classes não balanceadas. A melhoria na sensibilidade é fundamental para estes modelos. É importante lembrar que a interpretação dos resultados deve ter em consideração o contexto específico do problema e as necessidades práticas. A escolha do modelo e as estratégias de melhoria devem ser orientadas para a aplicação real e para os objetivos do estudo.

A escolha entre o Random Forest e a Regressão Logística deve ser baseada numa avaliação cuidadosa do desempenho, das medidas de performance, do contexto do problema e dos recursos disponíveis.

Ao concluir que o Random Forest demonstrou resultados melhores em comparação com a Regressão Logística, estamos a tomar esta decisão com base nas inofrmações do desempenho dos modelos. No entanto, isso não significa necessariamente que devemos descartar a Regressão Logística. Cada modelo pode fornecer informações úteis e resultados únicos, e é importante considerar as conclusões de ambos ao tomar decisões para projetos futuros.

A abordagem mais correta será considerar as forças e as fraquezas de cada modelo. O Random Forest pode ser uma escolha sólida quando se procura alta precisão de predição, é menos sensível a *overfitting* e lida bem com dados não lineares e relações complexas entre variáveis. No entanto, pode ser menos interpretável e pode exigir mais recursos computacionais para treino e implementação. A Regressão Logística é um modelo linear simples e altamente interpretável. Pode ser valioso quando se tem como objetivo entender a relação direta entre variáveis e a probabilidade de um evento. É mais fácil de implementar e explicar, mas pode ser menos preciso em casos complexos.

Pode ser feita uma fusão de modelos que combina as previsões de ambos os modelos. Por exemplo, podemos usar uma média ponderada das previsões ou usar métodos *ensemble* mais avançados, como o *Gradient Boosting*.

As conclusões de ambos os modelos ajudam a identificar variáveis mais importantes. Podemos usar essas variáveis em futuros modelos ou criar novas características com base nas

informações extraídas dos modelos.

Temos também de lembrar que o desempenho do modelo pode variar com o tempo e com a mudança de dados. Portanto, é importante manter uma avaliação contínua e adaptar as estratégias à medida que novos dados se tornam disponíveis.

Em suma, podemos concluir que variáveis como a idade, a proveniência dos doentes e as medições dos sinais vitais desempenham um papel significativo e importante na capacidade de prever a admissão dos doentes na Unidade de Cuidados Intensivos. Estas conclusões proporcionam informações que podem ser importantes para os profissionais de saúde, destacando a relevância destas variáveis no processo de triagem e na identificação de casos que requerem encaminhamento para a UCI.

## 5.1 Considerações Finais

No início deste estágio, mergulhei numa pesquisa aprofundada sobre o tema que se tornaria o centro da minha investigação. Esta pesquisa envolveu a análise criteriosa de uma variedade de artigos relacionados, proporcionando-me uma base sólida de conhecimento essencial para orientar o desenvolvimento do meu estudo. A partir desta pesquisa surgiram questões iniciais que se tornariam as questões principais para a procura das variáveis pertinentes disponíveis na empresa, as quais desempenhariam um papel fundamental no estudo.

Contudo, o desafio seguinte revelou-se a etapa mais demorada e muito importante para qualquer estudo que se pretenda realizar, uma vez que implicou a seleção de uma base de dados ideal para atender aos objetivos do estudo. Este processo exigiu muito esforço e paciência, à medida que examinei diversas bases de dados provenientes de diferentes hospitais, a fim de abordar o maior número possível das questões inicialmente formuladas. Foram estudadas quatro bases de dados distintas até finalmente identificar a definitiva, na qual foram então aplicados os modelos estatísticos selecionados.

Nesse trajeto, também explorei como a empresa armazena e processa os seus dados o que contribuiu para eu conhecer como uma empresa trabalha com os dados. Com a aquisição da base de dados final, conduzi uma análise exploratória, seguida da aplicação dos modelos estatísticos escolhidos. Foi um percurso que envolveu pesquisa, análise de dados e conhecimento prático, contribuindo para uma valiosa experiência no campo da pesquisa e análise de dados.

Este estágio proporcionou-me uma oportunidade de aplicar, num ambiente prático e real, os conhecimentos que acumulei ao longo do mestrado. Foi uma lição valiosa, porque aprendi que, num processo de análise de dados no mundo real, nos deparamos com desafios e obstáculos que não nos são apresentados nos exercícios teóricos dados nas salas de aula. Esta experiência mostrou-me que a aplicação do conhecimento teórico em situações reais envolve pormenores, complexidades e decisões que só podem ser verdadeiramente compreendidas e dominadas com a prática e a vivência direta no campo. Esta oportunidade proporcionou-me uma visão mais completa e prática do mundo da análise de dados e do seu papel fundamental nas empresas.

Durante o meu estágio, uma questão que se destacou e que merece atenção é a escassez significativa de dados nas bases de dados hospitalares. Fui confrontada com um grande número

de valores em falta (NA's), o que teve um impacto na minha análise, uma vez que tive que desenvolver estratégias para lidar com essas valores em falta nos dados que podem ter enviesado os resultados finais. Na minha perspetiva, é importante considerar o desenvolvimento de novos métodos para o registo de informações clínicas pelos médicos nas bases de dados.

Por exemplo, no âmbito do meu estudo, as variáveis relacionadas com as medições dos sinais vitais destacaram-se como importantes para a análise dos doentes que têm mais probabilidade de serem admitidos na Unidade de Cuidados Intensivos. Entretanto, foi precisamente nessas variáveis que encontrei o maior número de dados em falta. Este desafio reflete a complexidade e a sobrecarga de trabalho que os médicos enfrentam no seu quotidiano. Por vezes, o registo detalhado dessas informações pode ser negligenciado devido ao facto de darem mais importância à prática do seu trabalho.

Para enfrentar esta questão, é muito importante que os hospitais investiguem as causas deste problema e considerem a implementação de métodos mais eficientes para recolher e registar os dados dos doentes. Isto pode envolver o desenvolvimento de técnicas simplificadas, porém eficazes, que permitam um registo mais ágil e preciso. Garantir a integridade dos dados clínicos é fundamental não apenas para pesquisas, como a que conduzi durante o estágio, mas também para a qualidade global do atendimento aos doentes. Portanto, a busca por soluções inovadoras neste âmbito é uma etapa importante na melhoria dos sistemas de informação na saúde.

Por fim, é claro que ainda há muito a ser explorado e desenvolvido nesta área. Este estudo representa apenas um passo inicial em direção a um objetivo maior: a capacidade de prever com precisão quais os doentes que necessitam de cuidados intensivos. Esta previsão não vai apenas garantir que os doentes recebam o tratamento adequado no momento certo, mas também ajudará a evitar a alocação desnecessária de recursos numa área da saúde que envolve custos muito grandes.

No futuro, a pesquisa e o desenvolvimento nesta área são fundamentais para melhorar a eficiência dos sistemas de saúde e, ao mesmo tempo, melhorar o atendimento aos doentes. Esta visão prospetiva é um lembrete constante de que a procura pelo avanço na medicina e no atendimento médico é um esforço contínuo. À medida que continuamos a explorar novas maneiras de melhorar as nossas práticas e processos, estamos a contribuir para um sistema de saúde mais eficaz, capaz de atender às necessidades de todos de forma mais precisa e eficiente.

# Referências


[1] Fiona E Kelly and Kevin Fong and Nicholas Hirsch and Jerry P Nolan, Clinical Medicine, 376-385, The history of intensive care, 14, 2014

[2] Faisal Masud, M. D., Tina Yaqing Cai Lam, Sahar Fatima, M. B. B. S. (2018). Is 24/7 In-House Intensivist Staffing Necessary in the Intensive Care Unit? Methodist Debakey Cardiovasc J., 134–140. https://doi.org/10.14797/mdcj-14-2-134

[3] Mahanazuddin Syed and Shorabuddin Syed and Kevin Sexton and Hafsa Bareen Syeda and Maryam Garza and Meredith Zozus and Farhanuddin Syed and Salma Begum and Abdullah Usama Syed and Joseph Sanford and Fred Prior, (2021) 'Application of machine learning in intensive care unit (ICU) settings using MIMIC dataset: Systematic review'

[4] Hassler AP, Menasalvas E, García-García FJ, Rodríguez-Mañas L, Holzinger A. Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome. BMC Med Inform Decis Mak. 2019 Feb 18;19(1):33. doi: 10.1186/s12911-019-0747-6. PMID: 30777059; PMCID: PMC6483150

[5] Pugh SL, Brown PD, Enserro D. Missing repeated measures data in clinical trials. Neurooncol Pract. 2021 Jul 16;9(1):35-42. doi: 10.1093/nop/npab043. PMID: 35096402; PMCID: PMC8789297.

[6] Lionel Tarassenko and A. Hann and D. Young,(2006) 'Integrated monitoring and analysis for early warning of patient deterioration'

[7] https://www.nature.com/articles/s41591-020-0789-4

[8] Dell NA, Vaughn MG, Prasad Srivastava S, Alsolami A, Salas-Wright CP. Correlates of cannabis use disorder in the United States: A comparison of logistic regression, classification trees, and random forests. J Psychiatr Res. 2022;151:590-597. doi:10.1016/j.jpsychires.2022.05.021

[9] Rigatti SJ. Random Forest. J Insur Med. 2017;47(1):31-39. doi:10.17849/insm-47-01-31-39.1

[10] Jan Z, Verma B. Multicluster Class-Balanced Ensemble. IEEE Trans Neural Netw Learn Syst. 2021;32(3):1014-1025. doi:10.1109/TNNLS.2020.2979839

[11] Liu H, Zhou G, Zhou Y, Huang H, Wei X. An RBF neural network based on improved black widow optimization algorithm for classification and regression problems. Front Neuroinform. 2023 Jan 10;16:1103295. doi: 10.3389/fninf.2022.1103295. PMID: 36703878; PMCID: PMC9871759.

[12] Deo RC. Machine Learning in Medicine. Circulation. 2015 Nov 17;132(20):1920-30. doi: 10.1161/CIRCULATIONAHA.115.001593. PMID: 26572668; PMCID: PMC5831252.

[13] Lihu A, Holban Ş. A review of ensemble methods for de novo motif discovery in ChIP-Seq data [published correction appears in Brief Bioinform. 2016 Jul;17(4):731]. Brief Bioinform. 2015;16(6):964-973. doi:10.1093/bib/bbv022

[14] Nick TG, Campbell KM. Logistic regression. Methods Mol Biol. 2007;404:273-301. doi:10.1007/978-1-59745-530-5_14

[15] Abreu MN, Siqueira AL, Cardoso CS, Caiaffa WT. Ordinal logistic regression models: application in quality of life studies. Cad Saude Publica. 2008;24 Suppl 4:s581-s591. doi:10.1590/s0102-311x2008001600010

[16] https://smolski.github.io/livroavancado/reglog.html

[17] Fei Tony Liu and Kai Ming Ting and Zhi-Hua Zhou, 'Isolation Forest'

[18] https://www.nature.com/articles/s41591-020-0789-4

[19] Maglogiannis, I. G., Al, E. (2007). Emerging artificial intelligence applications in computer engineering : real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies. Ios Press. https://dl.acm.org/citation.cfm?id=1566770.1566773

[20] Maulud, D., Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. Journal of Applied Science and Technology Trends, 1(4), 140–147. https://doi.org/10.38094/jastt1457

[21] Mahesh, B. (2018). Machine Learning Algorithms-A Review. International Journal of Science and Research

[22] DeMers D, Wachs D. Physiology, Mean Arterial Pressure. 2023 Apr 10. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan–. PMID: 30855814.

[23] ICD-10: History and Context J.A. Hirsch, G. Nicola, G. McGinty, R.W. Liu, R.M. Barr, M.D. Chittle, L. Manchikanti American Journal of Neuroradiology Apr 2016, 37 (4) 596-599; DOI: 10.3174/ajnr.A4696