UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA

PROGRAMA DE PÓS-GRADUAÇÃO

EM COMPUTAÇÃO APLICADA

PREDIÇÃO DE MORTALIDADE EM UTI: APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS

JORGE SOUZA AZEVEDO MONIZ BARRETO

FEIRA DE SANTANA

Julho de 2019

UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA

PROGRAMA DE PÓS-GRADUAÇÃO
EM COMPUTAÇÃO APLICADA

Jorge Souza Azevedo Moniz Barreto

Predição de Mortalidade em UTI: Aplicação de Técnicas de Mineração de Dados

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para obtenção do título de Mestre em Computação Aplicada.

Orientador: Prof. Dr. Angelo C. Loula

Coorientador: Prof. Dr. Hudson F. Golino

Feira de Santana, Julho de 2019

Ficha Catalográfica – Biblioteca Central Julieta Carteado

Barreto, Jorge Souza Azevedo Moniz

B263p Predição de mortalidade em UTI: aplicação de técnicas de mineração de dados./ Jorge Souza Azevedo Moniz Barreto. – 2019.
75f.: il.
Orientador: Angelo C. Loula
Coorientador: Hudson F. Golino
Dissertação (mestrado) – Universidade Estadual de Feira de Santana. Programa de Pós-Graduação em Computação Aplicada, 2019.

1.APACHE-III. 2.Redes neurais. 3.Aprendizado de máquina. 4.UTI. 5.Mineração de dados. I.Loula, Angelo C., orient. II.Golino, Hudson F., coorient. II.Universidade Estadual de Feira de Santana. III.Título.

CDU: 004.89

Maria de Fátima de Jesus Moreira – Bibliotecária – CRB5/1120

Jorge Souza Azevedo Moniz Barreto

Predição de Mortalidade em UTI: Aplicação de técnicas de Mineração de Dados

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Computação Aplicada.

Feira de Santana, 31 de agosto de 2018

BANCA EXAMINADORA

Dr. Angelo Conrado Loula (Orientador)
Universidade Estadual de Feira de Santana

Dr. Ricardo Araújo Rios
Universidade Federal da Bahia

Dr. Rodrigo Tripodi Calumby
Universidade Estadual de Feira de Santana

Este trabalho é dedicado primeiramente a Deus, que direciona e nos permite seguir em frente (mesmo que errando durante o caminho) e a minha Esposa Fábia.

RESUMO

A utilização de escores padronizados para identificar a severidade de estado de pacientes internados em Unidade de Terapia Intensiva - UTI, tais como *Acute Physiology, Age, Chronic Health Evaluation*-APACHE III e *Simplified Acute Physiology Score*-SAPS provém informações utilizadas pela equipe médica para tomada de decisões. Estes escores de severidade passam por constantes revisões que buscam aprimorar sua capacidade de predição. Devido à utilizar metodologias lineares para predição, e os dados utilizados para obtenção dos escores possuem características não lineares, entendemos que possam ser utilizadas outras técnicas e metodologias para melhorar a predição desses escores. Este estudo busca propor a aplicação de métodos de mineração de dados, no pré-processamento da base de dados e na identificação da severidade do estado dos pacientes, utilizando Redes Neurais Artificiais - RNA, *Random Forest* - RF e Regressão Logística, tendo como atributos para análise os registros das variáveis fisiológicas já registradas pela equipe médica para cálculo dos escores mencionados. Os dados utilizados para esse fim, foram obtidos do *Medical Information Mart for Intensive Care*-MIMIC-III , um grande repositório disponível *on-line* para pesquisas e que contém registro de 56.530 pacientes. Além disso, foram analisadas técnicas de imputação de valores ausentes e balanceamento de classe, na busca por uma maior qualidade nos dados. Após aplicação da metodologia descrita no estudo, a Random Forest obteve desempenho melhor que os demais, com a AUC média de 0,780 ($\pm$ 0,005), sensibilidade de 0,712 ($\pm$ 0,012) e especificidade de 0,701 ($\pm$ 0,005) em conjunto com a técnica de imputação de valores padrões em substituição de valores ausentes, e com o balanceamento de classe usando *under sampling*. Mediante seleção de atributos, foi construído modelo com redução de atributos com resultados próximos da classificação com todos atributos, o que pode simplificar a coleta de dados pela equipe médica para gerar um escore de severidade.

Palavras-chaves:APACHE-III, Aprendizado de máquina, Random Forest, Mineração de Dados, UTI, Redes Neurais, Desbalanceamento de classe, Substituição de dados ausentes

ABSTRACT

The use of standardized scores to identify the severity of the condition of patients admitted to the Intensive Care Unit - ICU, such as Acute Physiology, Age, Chronic Health Evaluation - APACHE III and Simplified Acute Physiology Score - provides information used by the medical team to make decisions, these severity scores go through constant revisions that seek to improve their predictive capacity, due to using linear methodologies for prediction and the data used to obtain the scores have non-linear characteristics, we understand that it can be used other techniques and methodologies to improve the prediction of these scores. This study aims to propose the application of data mining methods, in the preprocessing of the entire database and in the identification of the severity of the patient's condition using Neural Networks RNA, Random Forest - RF and Logistic Regression, having as attributes to analyze the records of the physiological variables already registered by the medical team to calculate the mentioned scores. The data used for this purpose were obtained from the Medical Information Mart for Intensive Care (MIMIC-III), a large available on-line repository for searches containing a record of 56,530 patients. In addition, we analyzed techniques of imputation of missing values and class balancing, in the search for a higher quality in the data. After application of the methodology described in the study Random Forest obtained better performance than the others, with mean AUC of 0.780 ($\pm$ 0.005), sensitivity of 0.712 ($\pm$ 0.012) and specificity of 0.701 ($\pm$ 0.005) in conjunction with the technique of imputation of standard values in substitution of missing values, and with class balancing using under sampling. By means of attribute selection, a model with attributes reduction was created with results close to the classification with all attributes, which can simplify the data collection by the medical team to generate a severity score.

Key-words: APACHE-III. Machine Learning. Random Forest. Data mining. ICU. Neural Networks. Imbalanced data. Missing values.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE EQUAÇÕES

SUMÁRIO

1 INTRODUÇÃO

A Unidade de Terapia Intensiva é um ambiente de alta complexidade, onde pacientes severamente debilitados, no limiar entre a vida e a morte, são monitorados constantemente. Decisões rápidas no tratamento e ações de cuidados podem significar salvar uma vida, ou redirecionar recursos para onde possa ser salva (KLUNDERT et al., 2015).

Para avaliar a severidade do estado do paciente e direcionar os recursos, foram criados escores que, através da análise dos valores de atributos fisiológicos, demográficos e exames laboratoriais, pudessem auxiliar a equipe médica nas decisões referentes aos tratamentos dispensados aos pacientes da UTI (GALL; LEMESHOW; SAULNIER, 1993; KNAUS et al., 1981). Paralelo a isso, esses escores também auxiliavam a análise de indicadores das UTI's, fornecendo um parâmetro global utilizado para comparações e melhorias (KUZNIEWICZ et al., 2008).

A análise dos atributos fisiológicos considera cada atributo de forma isolada, pela sua correlação com a não sobrevivência dos pacientes e as correlações entre si, selecionando assim, quais irão formar os escores (GALL; LEMESHOW; SAULNIER, 1993; KNAUS et al., 1981; KNAUS et al., 1985), atribuindo pesos e discretizando em faixa de valores. Estes valores somados formam um índice que indica a gravidade do estado, e a probabilidade de não sobrevivência do paciente (portanto, um modelo linear e sem consideração de interação entre atributos). Os pesos e discretização se basearam na análise do corpo clínico que realizava o estudo de Knaus et al. (1981), e a predição de mortalidade utilizou a regressão logística somente considerando o escore final.

Atualmente, os modelos de predição mais utilizados para prever as condições de pacientes em UTI são o SAPS - Simplified Acute Physiology Score e o APACHE - Acute Physiology and Chronic Health disease Classification System. Esses modelos passam por constantes revisões buscando aprimorar a predição de pacientes individualmente (GALL; LEMESHOW; SAULNIER, 1993; KNAUS et al., 1981; KNAUS et al., 1985), e são utilizados também para avaliar as condições da unidade de terapia intensiva. (KIM; KIM; PARK, 2011; KLUNDERT et al., 2015).

Os modelos de predição de mortalidade do APACHE e do SAPS, utilizam modelos lineares para predição de mortalidade em UTI. Devido à complexidade dos dados dos registros de pacientes (RAGHUPATHI; RAGHUPATHI, 2014), existe espaço para o estudo de aplicação de modelos não lineares e dessa forma, comparar os resultados de predição de mortalidade, identificando possíveis melhorias na acurácia da predição. Assim, serão fornecidas informações mais precisas sobre a gravidade dos pacientes, que auxiliarão a decisão da equipe médica quanto ao tratamento necessário.

Além disso, após levantamento de estudos anteriores, destacou-se a necessidade de aplicar diferentes técnicas para análise do comportamento dos classificadores com bases de

dados com atributos ausentes e desbalanceamento da classe alvo. Segundo Batista, Prati e Monard (2004), o desbalanceamento de classes influencia negativamente a performance do treinamento de classificadores, e o cenário em que a base de dados de estudo possui classes balanceadas é raro quando utilizados repositórios de dados do mundo real. Dentro da pesquisa em medicina, normalmente a classe de interesse vai representar os casos raros ou de menor ocorrência, se fazendo necessário, quando possível, a utilização de técnicas de balanceamento de classe.

De modo similar, registros com ausência de atributos podem invalidar o resultado de pesquisas na área de medicina (STERNE et al., 2009). Dessa forma, avaliar e caracterizar a ausência de dados de modo a reduzir o viés e/ou melhorar a precisão dos classificadores, se tornou um passo importante na metodologia.

Os dados utilizados para cálculo dos escores, gerados pelos sistemas de monitoramento de pacientes das UTI's, registros de exames de laboratórios, avaliação de sinais vitais são, em muitos casos, registrados eletronicamente nos hospitais. Esses registros com o passar dos anos e a maior informatização do setor de saúde e hospitais, se tornaram grandes bases de dados, apresentando um cenário promissor para geração de conhecimento e informação, além de fonte de auxílio para tomadas de decisões e diagnósticos de pacientes (RAGHUPATHI; RAGHUPATHI, 2014).

Por todo o mundo, grandes repositórios de informações em saúde, tais como NICE - National Intensive Care Evaluation na Holanda (ARTS et al., 2002), JIPAD - Japanese Intensive Care Patients Database no Japão (FUJII et al., 2018), MIMIC-III Medical Information Mart for Intensive Care (JOHNSON, 2014), têm sido alvo de pesquisadores para desenvolvimento de várias ferramentas de diagnóstico, pesquisa em saúde e criação de modelos de predição (KIM; KIM; PARK, 2011; HAN; PEI; KAMBER, 2011; WONG; YOUNG, 1999; ALVES et al., 2003; NAVAZ et al., 2016).

O contínuo esforço na análise dos dados dos grandes repositórios fomenta, todos os anos, diversas pesquisas que procuram atualizar modelos já existentes e buscar novas formas de descobrir padrões, correlações e usar o conhecimento adquirido para novas ferramentas mais precisas. O crescimento dessas bases de dados provê cada vez mais registros para análise, o que pode, a cada pesquisa realizada, melhorar a precisão e acurácia das predições de mortalidade. A base do MIMIC em sua versão II possuía dados de 25.328 pacientes, enquanto a versão III possui 53.423. Mais dados disponíveis para a análise pode evidenciar padrões que não puderam ser encontrados em estudos anteriores.

Assim sendo, este estudo se propõe à aplicar técnicas de mineração de dados para predição de mortalidade em UTI, buscando aprimorar as metodologias já existentes através da avaliação de técnicas para tratamento de dados desbalanceados e tratamento para dados ausentes, utilizando os atributos usados para cálculo do APACHE III.

1.1 Objetivos

O objetivo geral deste trabalho é o estudo e a aplicação de métodos de mineração de dados, no pré-processamento e na predição de mortalidade, comparando as técnicas de Redes Neurais Artificiais - RNA, *Random Forest* - RF e Regressão Logística a partir dos atributos definidos pelo APACHE III. Através do estudo dos resultados obtidos na aplicação dos métodos e técnicas acima descritas, busca-se estabelecer uma metodologia na utilização de técnicas de desbalanceamento e tratamento de dados ausentes que possa auxiliar na melhora da predição de mortalidade em UTI, e na análise de grandes bases de dados que apresentem características de desbalanceamento e dados ausentes similares, para desenvolvimento de novas técnicas computacionais.

Serão comparados os resultados dos classificadores em conjunto com técnicas de balanceamento de classe e tratamento de dados ausentes de modo a identificar a influência dessas técnicas nos resultados.

Para alcançar este objetivo geral da pesquisa, serão necessários:

- Identificar e obter base de dados de admissões em UTI;
- Analisar e caracterizar os dados e os atributos dos pacientes;
- Aplicar e avaliar métodos de pré-processamento, para tratamento de dados ausentes e desbalanceamento de classes;
- Ajustar os parâmetros dos modelos de classificação (RNA, RF) na base pré-processada;
- Selecionar medidas de desempenho do processo de mineração de dados;
- Comparar o modelo de predição do APACHE III com técnicas de mineração de dados.

2 FUNDAMENTAÇÃO

Neste capítulo será relacionado o conhecimento que fundamenta o trabalho de modo a embasar a metodologia e os resultados apresentados nos capítulos 3 e 4, respectivamente. A seguir veremos escores usados em UTI's, o SAPS e mais profundamente o APACHE, continuando com o estudo da Mineração de Dados, principais desafios no tratamento de dados, modelos de predição e sua parametrização. De modo a avaliar a eficácia dos classificadores serão mostrados métodos de validação e testes estatísticos que poderão ser usados para identificar o desempenho dos classificadores. Ao final do capítulo 2, serão apresentados alguns trabalhos relacionados a linha de pesquisa.

2.1 Escores em Unidades de Terapia Intensiva

A utilização de escores de severidade de pacientes se tornou padrão em UTI's, seja para identificar a gravidade do estado de saúde de pacientes ou para servir de índice para melhoria da unidade hospitalar (KNAUS et al., 1981). Dentre os escores existentes, os mais utilizados e pesquisados foram o APACHE - Acute Physiologic and Chronic Health Evaluation, atualmente em sua quarta versão, e o SAPS - Simplified Physiologic Score, atualmente em sua terceira versão.

O SAPS foi desenvolvido na França por Le Gall e colaboradores, no Hospital Henri Mondor do Credeil, em 1983 (GALL; LEMESHOW; SAULNIER, 1993) e utiliza atribuição de pontos a 13 variáveis fisiológicas e à idade. Foi modificado para SAPS II em 1993, sendo validado em 13.152 pacientes de 137 UTI's Européias e Americanas, totalizando cerca de 12 países, sendo composto por 12 variáveis fisiológicas mais a idade, tipo de admissão e presença de doença crônica. O escore SAPS II é formado pela soma dos pontos obtidos pelo pior valor de todas as variáveis coletadas durante as primeiras 24 horas, após a admissão na UTI. A escala de ponto de cada variável foi desenvolvida pelos pesquisadores participantes do estudo de Gall, Lemeshow e Saulnier (1993).

2.1.1 APACHE - Acute Physiologic and Chronic Health Evaluation

O APACHE é um sistema de escore que foi desenvolvido, em sua primeira versão APACHE I, em 1981 por Knaus et al. (1981), com o objetivo de fornecer amplas informações sobre as admissões em uma UTI. Era preciso padronizar as informações das UTI's e como até então não estava sendo feito isso, existia uma limitação de informações que auxiliassem no acesso à novas terapias e procedimentos. Neste primeiro momento, ele não foi utilizado nem imaginado para ajudar a tomar decisões individuais e sim classificar grupos de pacientes com base na gravidade de suas doenças.

O sistema de escore do APACHE I foi concebido com 34 possíveis atributos, sendo eles variáveis fisiológicas colhidas nas primeiras 32 horas de admissão na UTI. De forma a padronizar os valores dos atributos, foi criada uma escala de 0 a 4 pontos, onde 0 seria o valor normal e 4 representaria um quadro severo para aquele atributo. Por exemplo, a pressão arterial aferida com um valor normal entre 70 a 100 mm Hg, recebe o valor 0, enquanto uma pressão arterial menor que 50 mm Hg ou maior que 160 mm Hg receberá o valor 4. O escore APACHE é obtido da soma desses valores, sendo que quanto maior o valor, mais grave o estado do paciente considerado.

A escolha dos atributos utilizados foi realizada segundo Knaus et al. (1981), após pesquisa na literatura, procurando identificar no escopo dos 7 maiores sistemas fisiológicos (neurológico, cardiovascular, respiratório, gastrointestinal, renal, metabólico e hematológico), os atributos que melhor definissem a gravidade das doenças que levam os pacientes à UTI. Após essa escolha de atributos, foi formado um grupo com 5 médicos, mais os dois autores do artigo de Knaus et al. (1981), representando as 3 maiores especialidades envolvidas nos cuidados críticos a pacientes em UTI: anestesia, medicina e cirurgia. O grupo era livre para adicionar ou retirar atributos e orientado a aplicar os pesos de 0 a 4 para todos os valores possíveis, de modo que o escore APACHE I fosse um valor cardinal, que classificasse os indivíduos através de uma pontuação final.

Para avaliação do escore APACHE I foram utilizadas 583 admissões de UTI do *George Washington University Medical Center - GWUMC*, em um período de 8 meses. Esse número representou 95% das admissões no período e foram excluídas do estudo as admissões que tiveram menos que 16 horas de permanência na UTI. Foram realizadas análises estatísticas entre o escore fisiológico, terapia e saída da UTI, sendo testadas com equações de regressão simples e múltipla.

A partir do estudo em Knaus et al. (1985) passou-se a analisar a taxa de mortalidade da UTI utilizando o escore APACHE. Como a variável alvo só possuía dois valores, sobrevivente ou não sobrevivente, a matriz de confusão foi muito útil para validação do modelo de predição usando regressão logística. A matriz de confusão obtida no estudo é mostrada na Tabela 1.

Knaus et al. (1981), concluiu que, o APACHE I fornecia um meio objetivo para descrever as características do paciente e estimar a gravidade da doença em uma ampla gama de pacientes de UTI, sendo possível comparar os resultados, avaliar novas tecnologias e planejar tratamentos.

Tabela 1 – Matriz de confusão APACHE I (KNAUS et al., 1981)

		Real	
Preditos		S	NS
	S	467	52
	NS	13	50
S - Sobreviventes			
NS - Não Sobreviventes			

Em 1985, houve necessidade de melhorias no APACHE I e surgiu o APACHE II, que utilizou 12 variáveis fisiológicas das 34 da versão anterior, e incluiu a idade do paciente como

décima terceira variável (KNAUS et al., 1985). Além disso, passou a utilizar os dados das primeiras 24 horas ao invés das 32 horas do estudo anterior em Knaus et al. (1981). Na tabela 2 são mostrados, de forma geral, os atributos considerados no escore.

Tabela 2 – Atributos gerais do APACHE II

Categoria	Atributos
Dados demográficos	Idade e sexo
Informações administrativas	Tipo de admissão (eletiva, urgente, emergência), serviço de UTI[a] (MICU[b], SICU[c], CCU[d], CSRU[e])
Sinais vitais (Medidos os mínimos e máximos a cada 6 horas, nas primeiras 24 horas de admissão na UTI)	Batimento cardíaco, pressão arterial, pressão sistólica, SpO_2, taxa de respiração, temperatura
Resultados laboratoriais (Mínimos e máximos nas 24 horas de internação na UTI)	Hematrócitos, Contagem de células brancas, Glucose, HCO3, Potássio, Sódio, Nitrogênio no Sangue, Creatinina
Intervenções nas primeiras 24 horas	Terapia de vasopressão, ventilação mecânica, pressão positiva de ar contínua
Outros valores	Urina expelida a cada 6 horas, Escala Glasgow

[a]UTI: Unidade de terapia intensiva. [b]MICU: Unidade de cuidados médicos intensivos.
[c]SICU: Unidade de cuidados cirúrgicos. [d]CCU: Unidade de cuidados coronariana.
[e]CSRU: Unidade de recuperação cirúrgica.

A escolha das variáveis que foram utilizadas no estudo de Knaus et al. (1985) foi realizada com base em um processo clínico de avaliação. Os critérios utilizados na seleção foram: redundância das informações, substituição por variáveis mais específicas e comparação multivariada das variáveis. O menor número de variáveis que refletiam na fisiologia do paciente manteve a precisão estatística do escore e facilitou registro pela equipe médica.

Para validação do escore APACHE II foram utilizados os dados de 5.815 admissões em UTI's, obtidas de 13 hospitais nos Estados Unidos. Os pesquisadores visitaram os hospitais e treinaram a equipe para coleta de dados utilizados no estudo. A matriz de confusão é mostrada na Tabela 3.

Tabela 3 – Matriz de confusão APACHE II (KNAUS et al., 1985)

Preditos	Real		
		S	NS
	S	4030	920
	NS	7	73
S - Sobreviventes			
NS - Não Sobreviventes			

Seguindo o processo de evolução do modelo, o APACHE III foi criado em 1991, no estudo de Knaus et al. (1991) e utilizou um banco de dados de 17.000 pacientes de 40 UTI's nos

Estados Unidos. Para estimar os pesos dos atributos foi usada a análise da regressão logística multivariável, para determinar a relação entre a taxa de mortalidade e cada um dos 20 atributos candidatos. Foi também explorada a interação entre os atributos fisiológicos por avaliação combinada e individuais de pesos dos atributos (KNAUS et al., 1991).

O APACHE III possui uma equação para predição de mortalidade que combina a doença e os coeficientes locais do paciente com os pesos das variáveis obtidas. Assim, obteve-se um escore considerado por Knaus et al. (1991), robusto e com maior precisão do que o obtido com o APACHE II, conforme demonstrado na matriz de confusão, presente na tabela 4.

Tabela 4 – Matriz de confusão APACHE III (KNAUS et al., 1991)

Preditos	Real		
		S	NS
	S	14384	2632
	NS	32	392
S - Sobreviventes			
NS - Não Sobreviventes			

Em 2006 foi desenvolvido o APACHE IV, baseado em um estudo de coorte em uma base de 110.558 admissões em 104 UTI's dos Estados Unidos, entre os anos de 2002 e 2003. Foi usada uma equação de Regressão Logística Multivariada para predição de mortalidade, obtendo melhores valores de acurácia quando comparados com o APACHE III. (ZIMMERMAN et al., 2006). O modelo usado de predição usado no APACHE IV, embora tenha mostrado bons resultados, ainda não foi utilizado em uma larga escala como o APACHE III e segundo Keegan, Gajic e Afessa (2012) ambos possuem habilidades similares de predição.

2.2 Mineração de dados

A quantidade de dados gerados por sistemas informatizados em diversas áreas, tais como negócios e transações financeiras, medicina e saúde, vendas de produtos diversos, experimentos científicos, sensoriamento remoto e vigilância eletrônica, criou grandes bancos de dados que devido a sua complexidade não podem mais ser analisados por seres humanos. O uso de ferramentas que possam, através de técnicas e modelos computacionais, analisar esses imensos bancos de dados e obter informações que possam ser utilizadas de alguma forma, se tornou uma necessidade que deu início a mineração de dados (HAN; PEI; KAMBER, 2011), o processo de descobrir padrões e conhecimentos de interesse em grandes quantidades de dados.

Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996) este processo, também conhecido por *Knowledge Discovery in Databases* - KDD, envolve a aplicação de técnicas e conceitos para transformar dados massivos em informação e conhecimento seguindo uma sequência de passos:

1. Conhecer o domínio do problema - para determinar as metas do processo é necessário obter conhecimento relevante sobre a área a ser explorada.

2. Criar uma base de dados: a seleção dos dados a serem analisados fornece o material básico para iniciar o processo de descoberta de conhecimento.

3. Pré-processamento: normalmente os dados no mundo real possuem características que podem dificultar o processo de KDD, uma das principais situações que podem ocorrer é a ausência de valores de atributos. Nesses casos, deve-se analisar os dados para decidir estratégias que irão minimizar os problemas decorrentes dessa situação.

4. Escolha da técnica de Mineração de dados a ser usada: a mineração de dados possui diversos modelos que poderão ser usados para descoberta de padrões, que são úteis em determinados cenários, a saber:

    - Associação: determina se um evento ocorre em conjunto com outro;
    - Classificação: separa os dados em conjuntos, identificados por características de interesse;
    - Agrupamento: agrupa registros que tenham determinadas características, mas ainda não definidas no problema inicial.

5. Interpretação: após descoberta do padrão será realizado uma interpretação do mesmo, traduzindo em informação útil.

6. Utilização da descoberta: Utiliza o conhecimento adquirido, sendo possível tomar decisões ou documentar para reportar a grupos interessados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Com a capacidade da Classificação em distinguir entre classes e conceitos, esse modelo de mineração de dados tem sido utilizado para realizar também predição, ou seja, baseado no padrão obtido após o treinamento no modelo, sua aplicação em novos dados pode classificá-los.

Seguindo a sequência de passos do KDD, após o estudo do problema e criação ou seleção da base de dados a ser analisada, é verificado a qualidade dos dados, à procura de situações que poderão criar dificuldades para os classificadores tais como dados ausentes e desbalanceamento de classes.

2.2.1 Dados ausentes

A análise dos dados ausentes faz parte do KDD, que orienta nos passos necessários para o pré-processamento da base de dados. Identificar as razões e os tipos de dados ausentes é determinante para identificar os impactos ao classificador e aplicar abordagens que melhorem a performance do mesmo. Segundo Schafer e Graham (2002) dados ausentes podem ser classificadas em três grupos:

- Dados ausentes completamente aleatórios - MCAR: Os dados ausentes não têm relação com os dados existentes. A ausência de dados é completamente aleatória, dessa forma, os dados ausentes são independentes dos outros dados ausentes e dos dados existentes do indivíduo;

- Dados ausentes aleatórios - MAR: Os dados que estão ausentes tem alguma relação com algum dado observado do individuo. Consideremos, por exemplo que determinado paciente não teve sua taxa respiratória aferida. Nesse caso, a equipe médica aferiu a pressão e a temperatura e não viu anormalidade nenhuma, decidindo não prolongar o exame;

- Dados ausentes não aleatórios - MNAR: Os dados ausentes têm relação direta com o valor que seria observado para ele, por exemplo, uma família não informa a faixa de renda pois considera muito baixa.

Determinada a razão principal para ausência de dados deve-se escolher qual estratégia aplicar. Algumas técnicas que podem ser utilizadas são:

- Remoção dos exemplos que possuem dados ausentes: Pode diminuir significativamente o número de exemplos disponíveis a depender da quantidade de dados ausentes, mas dessa forma a base não teria valores de atributos ausentes.

- Aplicação de moda ou média do atributo para substituir o dado ausente: Corre-se o risco de uma superadaptação dos dados a depender da quantidade dos dados ausentes e da distribuição dos dados do atributo.

- Substituição dos dados ausentes por valores padrões: Substitui os valores ausentes por um valor que seja padrão para determinado atributo. A depender das características do dados e quantidade de dados ausentes, a base poderá ficar com uma distribuição muito linear.

- Aplicação de modelos preditivos para substituir os valores ausentes: Através da utilização de modelos preditivos pode-se utilizar os dados não ausentes como preditores para os dados ausentes. Dessa forma para cada dado ausente, será feito um modelo de predição utilizando os outros atributos como preditores.

Uma estratégia que surge como método para tratamento de dados ausentes é o MICE - *Multivariate imputation by chained equations* (AZUR et al., 2011). No MICE, uma série de modelos de regressão é executada, sendo cada atributo com dados ausentes modelado de acordo com os outros atributos nos dados. Isso possibilita que os atributos sejam modelados de acordo com sua distribuição, como por exemplo, atributos binários modelados com regressão logística e atributos contínuos modelados usando regressão linear (AZUR et al., 2011). O MICE segue alguns passos para a imputação de valores aos atributos ausentes:

1. É atribuída uma imputação simples aos dados ausentes, tal como a média;
2. Um dos atributos tem seu valor imputado no passo anterior removido;
3. Os valores ausentes são substituídos por predições do modelo de regressão, utilizando os outros atributos;
4. São repetidos os passos 1, 2 e 3 para todos os atributos e esses passos formam um ciclo de imputação;
5. O ciclo de imputação é repetido, sendo o número de repetições definido pelo pesquisador. Os valores dos atributos do último ciclo serão considerados os valores finais dos atributos.

Em todos os casos é necessário analisar os dados ausentes e sua importância para a característica dos exemplos, de modo que aplicando qualquer técnica, não seja prejudicada a eficácia do classificador.

2.2.2 Desbalanceamento de dados

Uma base desbalanceada é identificada desta forma quando há uma classe com muito mais exemplos do que outras na base de dados. Por exemplo, em uma base de dados há 100 exemplos da classe 1 e 20 exemplos da classe 2. Essa situação compromete significativamente a performance da maioria dos modelos de predição (HE; GARCIA, 2009).

Dados desbalanceados são comuns e seu estudo tem sido alvo de inúmeras publicações ao decorrer dos anos, no *Institute of Electrical and Electronics Enginners - IEEE* e na *Association for Computing Machinery - ACM* (HE; GARCIA, 2009). Dessa forma, foram desenvolvidos métodos para tratar dados desbalanceados para uso na mineração de dados:

- *Under sampling:* Consiste na subamostragem da classe dominante de modo que seu número de exemplos se iguale a classe minoritária. A escolha dos exemplos da classe dominante que serão retirados da amostra pode ser aleatória ou informada.

  O *Under sampling* informado pode utilizar algoritmos de *EasyEnsemble*, no qual é desenvolvido um conjunto de classificadores que irão dividir a classe majoritária em subgrupos de tamanhos iguais ou próximos à classe minoritária, e dessa forma montar conjuntos de dados que irão ser aplicados ao classificador; ou *BalancedCascade*, que de forma supervisionada seleciona dados da classe majoritária, (LIU; WU; ZHOU, 2009) verificando os resultados corretos da classificação, e assim selecionando os exemplos que melhor representem as classes.

  Dentre as técnicas de *Under sampling*, temos também o Tomek Link. Essa técnica considera que registros próximos (estabelecido pela análise da distância dos atributos), mas de classe oposta, são ruído ou registros de borda. Dessa forma, a retirada de um ou ambos

pode "limpar os dados", tornando mais fácil a separação do espaço das classes pelo classificador. O Tomek link funciona da seguinte forma: dados os registros x e y pertencentes a classes distintas, e seja a distância d(x, y), x e y são chamados de Tomek link se não houver um registro q, onde a distância (x, q) é menor que a distância (x, y), ou a distância (y, q) é menor que a distância (y, x). Nesses casos é considerado que x ou y ou ambos são ruído, ou ambos são registros de borda. O Tomek link, sendo utilizado como *under sampling* irá remover um desses registros ou ambos, a depender do modo de operação que o algoritmo foi definido. (BATISTA; BAZZAN; MONARD, 2003)

Outra técnica de *under sampling* é o *Neighborhood Cleaning Rule* - NCR. Essa técnica foi trazida por Laurikkala (2001) para o balanceamento de classes através da redução de dados. Uma das vantagens do uso da NCL é o foco maior na qualidade dos dados removidos em vez da redução de dados. A NCL é baseada no conceito de seleção unilateral - SU, que é uma técnica de redução de dados utilizada para classes desbalanceadas (KUBAT; MATWIN et al., 1997). O SU reduz a classe majoritária, deixando a classe minoritária sem alteração. Em um caso que se tenha um conjunto de dados em que I é a classe de interesse e O é a classe majoritária, a NCL usa a regra de vizinhos mais próximos ENN, editada por Wilson e Martinez (2000) para reduzir os dados de O. O ENN compara os três vizinhos mais próximos, e se dois forem de classes diferentes o exemplo é removido. Além disso, a NCL remove os três vizinhos mais próximos que classificam erroneamente os exemplos de I que pertencem à O, aumentando a limpeza. Esta ideia foi proposta para evitar redução de classes muito pequenas.

- *Over sampling:* Consiste na superamostragem da classe minoritária, podendo ser utilizados algoritmos que criem exemplos sintéticos baseados em similaridade, como o SMOTE - *Synthetic Minority Oversampling Technique*.

  O SMOTE utiliza um determinado número de vizinhos de cada exemplo da classe minoritária, e gera novos indivíduos, os quais a variação dos atributos é suficiente para diferenciá-lo, mas não afastá-lo da vizinhança onde foi obtido os atributos. Dessa forma os classificadores têm regiões mais gerais com dados da classe alvo em vez de regiões específicas, como no caso de poucos exemplos da classe. (CHAWLA et al., 2002)

  Outra técnica de *over sampling* é o Adaptive Synthetic - ADASYN. A construção do ADASYN se baseia na ideia de que dados da classe minoritária gerados sinteticamente são mais difíceis de ser absorvidos pelos classificadores do que os dados originais. Dessa forma é usada ponderação para os exemplos criados sinteticamente.

  Segundo He et al. (2008) o aprendizado dos classificadores é melhorado de duas formas: reduzindo o viés criado pelo desbalanceamento de classe, e deslocando o limite de decisão de classificação para os dados gerados sinteticamente, que são mais difíceis de serem aprendidos. De forma similar ao SMOTE serão criados os dados sintéticos da classe minoritária, através da análise dos vizinhos próximos. O ADASYN irá criar junto com os

dados sintéticos, pesos que irá definir regiões de densidade diferenciada de dados. Assim, forçará o classificador a buscar essas regiões onde os dados estão mais concentrados, melhorando o aprendizado dessas classes.

Identificada as soluções que possam reduzir o impacto de dados ausentes e classe desbalanceada, segue-se a aplicação das mesmas na base de dados. A partir desse ponto do KDD, serão testados modelos de classificação em parte da base de dados, sendo avaliados seus resultados através de métricas a serem definidas.

Dentre os modelos de classificação para predição, serão escolhidos para esse estudo a Rede Neural Artificial - RNA, *Random Forest - RF* e Regressão Logística, esses modelos foram escolhidos por representarem um segmento de classificação (conexionista, assembleia e linear).

2.3 Classificadores

2.3.1 Redes Neurais Artificiais

Redes neurais artificias - RNA é uma técnica de mineração de dados, inspirada no funcionamento do cérebro, de modo que consiga alcançar algumas de suas características: velocidade de processamento, capacidade de adaptação, processamento paralelo e aprendizado (HAYKIN et al., 2009). A RNA é composta de nós de rede, assim como o cérebro é composto de neurônios. Essas estruturas se comunicam entre si através de conexões. Cada conexão entre os nós da RNA possui um peso que determina seu grau de aproximação do próximo neurônio a que está conectado. Em seu modo mais simples, temos a representação de um nó da rede (neurônio artificial) na figura 1.

A partir de um valor de entrada, tipicamente o neurônio multiplica esse valor por um peso associado à conexão de entrada. O somatório das entradas ponderadas pelos pesos é aplicado na função de ativação do neurônio, produzindo assim uma saída.

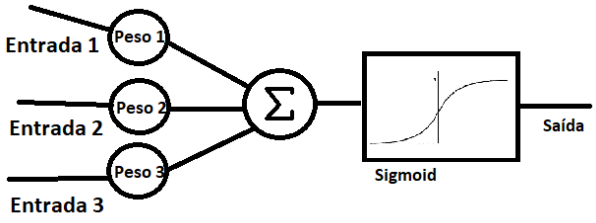

Figura 1 – Perceptron (neurônio artificial).

A capacidade de adaptação da RNA provém da utilização de algoritmos de aprendizado que alteram os pesos da conexão de forma a alterar a saída da rede e seu comportamento. Esses algoritmos analisam a saída da rede para determinar o quanto os pesos serão alterados de modo que a saída se aproxime do valor esperado (HAYKIN et al., 2009).

Nos modelos de aprendizagem supervisionada, é comparado esse valor de saída da rede com o valor esperado, obtendo-se o erro na saída. Este erro é informado à rede, que realizará uma série de operações de ajuste de peso até aproximar o valor de saída ao valor esperado, ajustando então os pesos da rede a partir da saída em direção a entrada (treinamento por *backpropagation*).

Dentre os tipos de RNA destacamos:

- Redes Multilayer Perceptron MLP: Redes criadas para resolução de problemas binários, composta pelo conjunto de neurônios Perceptrons (figura 1), formada normalmente de camada de entrada, camada de saída, e entre elas um número indefinido de camadas intermediárias. A modelagem da rede com a definição de camadas de entrada e camadas intermediárias e número de neurônios em cada camada irá depender do tipo de problema e realização de testes. Bem como, a definição dos parâmetros de configuração da rede, tais como taxa de aprendizado, número de iterações máximas da rede, tipo de otimizador, função de ativação dos neurônios. O treinamento das MLP é feito por *Backpropagation* e sua operação, após treinada, é sempre em uma só direção: da camada de entrada para camada de saída, como mostrado na figura 2.

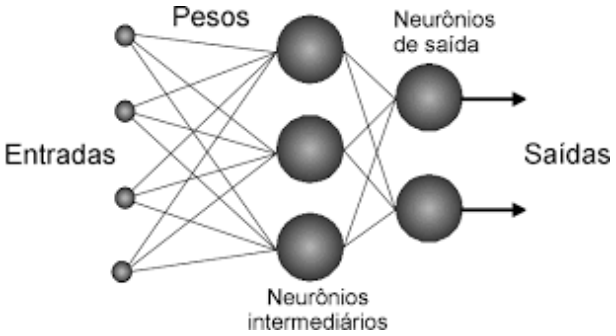

Figura 2 – *Multi-layer Perceptron*.

- Redes Neurais Convolucionais CNR: Essas redes neurais são consideradas redes profundas, por possuírem inúmeras camadas intermediárias. Mais utilizadas para reconhecimento de imagem e dados visuais, são amplamente utilizadas no reconhecimento ótico de caracteres (OCR). A imagem é passada para a CNR como uma caixa retangular, com largura e altura medidas pela quantidade de *pixels*, e a profundidade com três camadas correspondentes a codificação de cores RGB. Essas camadas também são conhecidas como canais e o funcionamento da CNR é mostrado na figura 3.

- *Generative Adversarial Network* GAN: São redes neurais profundas, sendo compostas por duas redes, uma com função gerador e outra discriminador, A rede com função discriminador irá determinar se uma imagem parece natural, isto é, uma imagem do conjunto de dados e o gerador irá criar imagens de aparência tão natural que enganem o discriminador (origem do termo adversárias, funcionamento mostrado na figura 4).

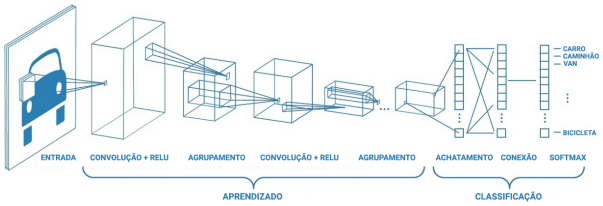

Figura 3 – Redes Neurais Convolucionais.

As redes GAN foram introduzidas no estudo de Goodfellow et al. (2014) e possuem potencial enorme, pois aprendem a imitar qualquer distribuição de dados. Tendo em conta, por exemplo, as palavras em um email, a rede pode ser treinada para prever se a mensagem é spam ou não. O spam é a classe de saída, e as palavras coletadas do email são os recursos que compõem os dados de entrada. Expresso de forma matemática, a classe é chamada y e os recursos são chamados de x.

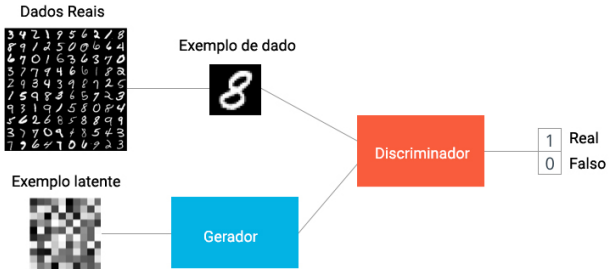

Figura 4 – *Generative Adversarial Network.*

Uma RNA é capaz de trabalhar com métodos lineares e não lineares de regressão, não requer conhecimentos estatísticos para implementação, adapta-se a diferentes conjuntos de dados através de novos treinos e possui tolerância a falhas (HAYKIN et al., 2009)

2.3.2 *Random Forest*

Random Forest - RF, é um método de aprendizado em assembleia para classificação, regressão e predição, que opera construindo N árvores de decisão, distribuindo os atributos de forma aleatória entre elas, sendo que cada árvore terá um voto na assembleia. A saída mais votada será o valor de saída da *Random Forest* (BREIMAN, 2001). Em uma forma simplificada, Breiman (2004) descreve o funcionamento de uma RF seguindo os seguintes passos:

1. O conjunto de treinamento é uma amostra da base de dados.
2. É definido por meio de um parâmetro, e é o número inteiro que representa a quantidade total de árvores da *Random Forest*
3. Outro parâmetro define o número N de atributos que será passado para cada árvore, sendo N é menor do que o número total de atributos.

4. Para cada árvore, um subconjunto aleatório dos atributos é definido por amostragem aleatória, independente e uniforme. Além disso, os dados distribuídos para cada árvore é definido pelo método de *bootstrap*. Conjuntos de treinamento são amostrados aleatoriamente e distribuídos entre as árvores, e dessa forma torna a classificação mais generalizada (BREIMAN, 2001).

5. Após a seleção de atributo a *Random forest* define a importância dos atributos de cada árvore através do índice GINI (BREIMAN, 2001). Esse índice é utilizado para avaliar a distribuição das classes do atributo em cada nó. A divisão de cada nó é feita de maneira a resultar em nós filhos mais "puros"do que o nó original, ou seja, com maiores concentrações de exemplos de certas classes.

   Cada árvore irá crescer até atingir o valor de profundidade definido anteriormente. Na regressão, como um vetor de teste x é colocado em cada árvore, é atribuído o valor médio dos valores de y no nó final. O valor previsto para a classificação da RF é a classe que obtém o maior número de votos (figura 5).

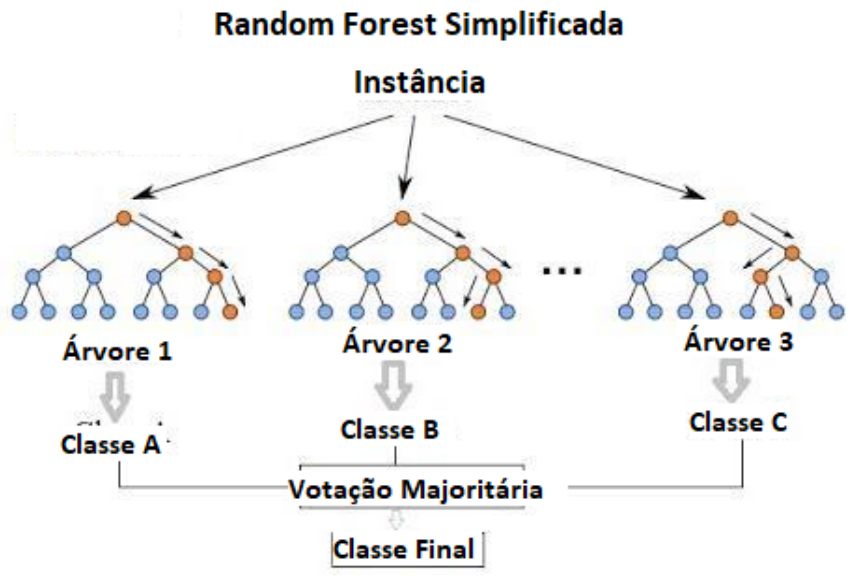

Figura 5 – *Random Forest*. Fonte (JAGANNATH, 2017).

Random Forest é uma ferramenta eficaz na previsão, com menos tendência a *overfit* (BREIMAN, 2001), devido a aleatoriedade na seleção de atributos entre as árvores. Além disso, a estrutura dos preditores individuais e suas correlações fornecem informações sobre a capacidade de predição da *Random Forest*, quanto maior a independência entre as árvores, melhor a generalização possível pelo modelo.

2.3.3 Regressão Logística

A Regressão Logística - RL foi desenvolvida pelo estatístico David Cox em 1958 e é um modelo binário usado para estimar a probabilidade de uma resposta binária com base em uma ou mais variáveis preditoras. Isso permite dizer que a presença de um fator de risco aumenta a probabilidade de um dado resultado por uma porcentagem específica (WALKER; DUNCAN, 1967).

A RL tem sido considerada um dos principais métodos de modelagem estatística de dados. Inclusive, pesquisadores tem discretizado a resposta de seus problemas para transformá-los em sistemas binários, de modo que possam utilizar a RL (PAULA, 2004).

Dentre os modelos de RL podemos nos referir a Regressão Logística Simples e Múltipla. Um modelo de regressão logística simples pode ser usado em regressão com uma variável categórica, enquanto a Regressão Logística Múltipla considera mais de uma variável categórica para realizar a regressão.

Dentre as vantagens no uso da RL, podemos citar:

- Facilidade para lidar com variáveis independentes categóricas:
- Fornece resultados em termos de probabilidade:
- Facilidade de classificação de indivíduos em categorias:
- Requer pequeno número de suposições.

Dessa forma, a RL é um modelo que permite a análise da variável dependente, sendo que a variável pode ser qualitativa e expressa por duas ou mais categorias. Uma boa aproximação é obtida pela RL, o que permite o uso de um modelo de regressão para se calcular ou prever a probabilidade de um evento específico (FIGUEIRA, 2006). Na equação 1 é mostrada uma equação geral de RL.

$$y = \frac{1}{1 + e^{(-f(x))}} \tag{1}$$

2.3.4 Otimização de parâmetros dos classificadores

Cada classificador possui um conjunto de parâmetros que irá guiá-lo no processo de classificação. De modo a melhorar o desempenho, se faz necessário a execução de inúmeros testes com diferentes configurações de parâmetros para encontrar valores ótimos para cada um. Este processo pode ser exaustivo em muitos casos, pela quantidade de parâmetros a serem alterados e as possibilidades de configuração de cada classificador.

Uma das abordagens que pode ser utilizada seria um método guloso de testar as possibilidades de cada parâmetro um de cada vez, mantendo o valor que tenha sido melhor avaliado e seguindo para o próximo parâmetro. Esse método não leva em conta que a correlação entre os parâmetros possa interferir na performance do classificador, por exemplo, a profundidade que uma *Random Forest* foi escolhida, posteriormente foi selecionada a quantidade de árvores da assembleia. Essa combinação escolhida pode não ser a ótima, pois não serão testados os valores anteriores da profundidade com o valor selecionado de árvores.

Outra abordagem é o teste exaustivo das combinações dos parâmetros, avaliando o resultado de cada combinação encontrada. Embora possamos obter um valor ótimo para os parâmetros, levando em conta também a relação entre eles, esse método tem alto custo computacional: dependendo da quantidade de parâmetros e valores possíveis de cada, será executado o treino e teste do classificador o número de vezes descrito pela fórmula de análise combinatória (QNT), mostrada na equação 2, onde x é o número de testes executados, n é a quantidade de parâmetros e m a quantidade de valores possíveis para o parâmetro (BERGSTRA; BENGIO, 2012).

$$QNT = \frac{(n+m-1)!}{(n-1)! * m!} \tag{2}$$

2.3.5 Treinamento e teste

Para construção e avaliação dos classificadores, os mesmos são aplicados a dois conjuntos de dados, que são recortes da base de dados original e: um conjunto para treinamento e outro conjunto para teste. O conjunto de teste só será utilizado ao final do processo, após a otimização de parâmetros e validação dos modelos.

No treinamento, o modelo de predição usará os atributos para determinar a classe alvo e irá comparar o seu resultado com a classe alvo já conhecida, para realizar o ajuste do modelo do classificador. Esse tipo de metodologia de treinamento é chamada de treinamento supervisionado, pois utiliza a informação explícita da classe alvo para acertar a classificação.

Após o treinamento, o modelo de classificação é aplicado ao conjunto de testes com exemplos que ainda não foram vistos pelo classificador, de modo a estimar a capacidade do modelo em dados não vistos. Então é analisado os resultados, permitindo a avaliação da capacidade do modelo na predição em dados gerais. (ZAKI; MEIRA, 2014).

2.3.6 *PCA-Principal Component Analysis*

Conjuntos de dados multidimensionais são difíceis de serem visualizados, assim, frequentemente buscam-se modos de reduzir seu conjunto de variáveis, mas sem perder sua capacidade de informação. Uma das técnicas mais utilizadas é o *PCA-Principal Component Analysis* (ABDI;

WILLIAMS, 2010).Em um caso que se tenha n medições em um vetor x de p variáveis aleatórias, e deseja-se reduzir a dimensão de p para q, sendo q é normalmente muito menor que p. O PCA faz isso encontrando combinações lineares, a1 x, a2 x,. . . , aq x, chamadas de "componentes principais", que têm sucessivamente variância máxima para os dados, sujeitas a não estar correlacionadas com a k x s anteriores. Resolvendo este problema de maximização, descobrimos que os vetores a1, a2, ..., aq são os autovetores da matriz de covariância, S, dos dados, correspondendo aos q maiores autovalores.

2.3.7 Métodos de avaliação

Os modelos de predição utilizam os padrões encontrados a partir da base de treinamento, na qual a classe alvo da predição é conhecida. De modo a validar os modelos de predição, é preciso utilizar metodologias que possam identificar se o modelo consegue predizer corretamente a variável alvo em dados que não foram usados no treinamento. Ao se ajustar o modelo de classificadores buscando aumentar sua performance na predição, pode-se encontrar uma superadaptação do modelo, de modo que se obtenha um classificador extremamente eficiente durante o treinamento, e com resultados ruins quando apresentado a um novo conjunto de dados (WITTEN; FRANK; HALL, 2011).

A fim de avaliar o desempenho de predição, utilizando para testes, uma parte dos dados separados dos dados usados no treinamento, são utilizados técnicas como *Hold-out* e *Cross-Validation* (KOHAVI et al., 1995).

2.3.7.1 Hold-out

O método *Hold-out* consiste em dividir o conjunto de dados em 2 partes exclusivas (os exemplos não se repetem entre as partes), uma parte para treinamento dos modelos de predição e outra parte para validação do modelo (KOHAVI et al., 1995). É realizada essa divisão para que possa ser avaliado o desempenho da predição em dados diferentes dos utilizados no treino. É selecionado um percentual da base original, sendo que, mantem-se a proporção entre as classes alvo, de modo que treinamento e validação tenham a mesma característica do conjunto total. O método *Hold-out*, no entanto, realiza uma estimativa limitada do desempenho do classificador ao usar somente um conjunto de dados de treinamento e teste.

2.3.7.2 Cross-Validation

O método de *Cross-Validation* consiste em dividir o conjunto total de dados em um número k determinado de subconjuntos (*fold*) mutuamente exclusivos de mesmo tamanho (KOHAVI et al., 1995).

O primeiro passo a ser realizado, para garantir que uma base de testes, com registros que nunca foram vistos pelo classificador, é separar uma parte da base de dados para testes. A outra

parte da base será apresentada ao classificador usando o *Cross-Validation* seguindo os seguintes passos:

- A base de dados de treinamento é separada em k *folds*;
- Um *fold* é separado para validação, e os outros são utilizadas para treino do classificador;
- Na próxima iteração é separado outro *fold* para validação e os outros restantes utilizados para treinamento;
- Os passos são executados até o número de iterações ser igual a k.

A repetição em k vezes procura reduzir o viés, já que está sendo usado a maioria dos dados para ajuste, o que também reduz significativamente a variação, pois a maioria dos dados também está sendo usada no conjunto de testes. A figura 6 mostra o funcionamento da *Cross-Validation*.

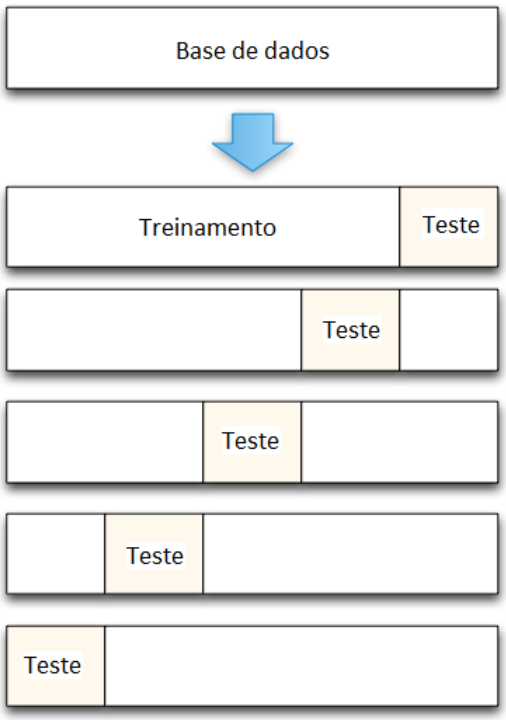

Figura 6 – Funcionamento da *Cross-Validation*. Fonte: Próprio Autor

Além de reduzir o viés, o fato da maioria dos dados de treinamento ser usado também na validação, pode gerar uma superadaptação do modelo. Buscando evitar essa situação uma estratégia de uso do *Cross-Validation* é a divisão em 2 *folds* e repetição do processo N vezes (figura 7). Como são 2 divisões da pastas dentro do ciclo isso acarreta que o treinamento e validação serão feitos em dados exclusivos no *fold* (BENGIO; GRANDVALET, 2004).

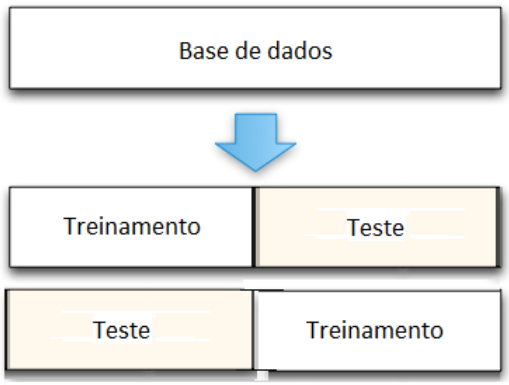

Figura 7 – Funcionamento da *Cross-Validation 10 x 2*. Fonte: Prório autor

2.3.8 Medidas de avaliação

De modo a avaliar a qualidade de classificadores, é necessário o uso de medidas para avaliação do resultado dos mesmos. Dessa forma, a escolha da medida de avaliação faz parte do processo de mineração de dados (KOHAVI; PROVOST, 1998).

2.3.8.1 Matriz de confusão

A matriz de confusão é uma tabela utilizada para análise de quanto o classificador é bom em reconhecer exemplos de diferentes classes (HAN; PEI; KAMBER, 2011). Para montar a tabela são utilizados os valores de Verdadeiros Positivos - VP e Verdadeiros Negativos - VN, que dizem o quanto o classificador acerta, Falsos Positivos - FP e Falsos Negativos - FN, que indicam o quanto o classificador erra. A tabela da Matriz de confusão é mostrada na tabela 5.

Tabela 5 – Matriz de confusão

		Preditos		
		sim	não	Total
Reais	sim	VP	FN	P
	não	FP	VN	N
Total		P’	N’	P+N

2.3.8.2 Sensibilidade

A Sensibilidade de um classificador reflete o quanto este é eficaz em identificar corretamente, dentre todos os indivíduos avaliados, aqueles que realmente apresentam a característica de interesse. O cálculo da Sensibilidade é feito com base na matriz de confusão, e sua fórmula corresponde à equação 3.

$$S = \frac{VP}{VP + FN} \tag{3}$$

2.3.8.3 Especificidade

A Especificidade de um classificador reflete o quanto ele é eficaz em identificar corretamente os indivíduos que não apresentam a condição de interesse, sua fórmula corresponde à equação 4.

$$E = \frac{VN}{VN + FP} \tag{4}$$

2.3.8.4 Acurácia

Acurácia é a proporção de predições corretas. Ou seja, mostra a porcentagem de dados que são corretamente classificados (HAN; PEI; KAMBER, 2011), sem levar em consideração o que é positivo e o que é negativo.

Em um classificador com uma alta taxa de acerto de determinada classe, quando aplicado a uma base onde essa classe possui muito mais exemplos do que outra, a acurácia média deverá ser alta. Ou seja, a acurácia é suscetível à base de dados desbalanceados e induz facilmente a uma conclusão errada sobre o desempenho do sistema (KOHAVI; PROVOST, 1998), por isso é considerada uma medida fraca para avaliação para bases desbalanceadas. Sua fórmula é mostrada na equação 5.

$$ACC = \frac{VP + VN}{P + N} \tag{5}$$

2.3.8.5 Curva característica de Operação do Receptor - ROC e Área sobre a curva - AUC

A ROC é uma representação gráfica da performance de um sistema de classificação (HANLEY; MCNEIL, 1982), sendo um gráfico de Sensibilidade versus Especificidade. Gráficos da curva ROC foram originalmente utilizados em detecção de sinais, para se avaliar a qualidade de transmissão de um sinal em um canal com ruído (EGAN, 1975), e são utilizados na área médica para avaliação de resultados em testes clínicos (ZHOU; MCCLISH; OBUCHOWSKI, 2009).

A análise da curva ROC foi introduzida na Mineração de Dados como uma ferramenta para a avaliação de modelos de classificação (BRADLEY, 1997). Ela é particularmente útil em domínios nos quais exista uma grande desproporção entre as classes ou, quando se deve levar em consideração diferentes custos/benefícios para os diferentes erros/acertos de classificação (PRATI; BATISTA; MONARD, 2008), pois mostra o quanto o classificador acerta a classe de interesse e a outra classe.

De modo a se obter um valor escalar que possa ser usado na comparação entre curvas ROC obtidas de diferentes classificadores, para uma mesma amostra de dados, é calculada a Área sobre a Curva - AUC da ROC de cada classificador, conforme figura 8 (HANLEY; MCNEIL, 1982).

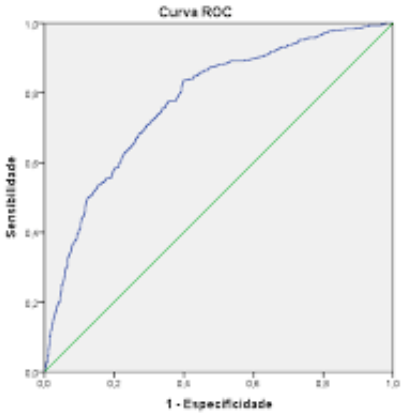

Figura 8 – Área sob a Curva (RITTA; GORLA; HEIN, 2015)

2.3.9 Testes Estatísticos

Testes estatísticos são usados para validar a comparação das medidas de avaliação dos classificadores. Pode-se usar os testes para interpretar os resultados das múltiplas execuções do classificador, a exemplo do *Cross-Validation*. Dessa forma, pode-se realizar afirmações sobre os resultados, considerando determinado classificador melhor que outro. (DEMŠAR, 2006)

Os testes estatísticos podem ser divididos em 2 grupos: Paramétricos e Não-paramétricos. Os testes paramétricos são utilizados quando o histograma de frequência dos dados apresenta um formato normal, ou seja, seu contorno segue um desenho de sino, possui apenas um ponto máximo, onde os intervalos da classe são centrados na média da distribuição.

Sendo a distribuição dos dados não normal, ou sem possibilidade de afirmação de ser uma distribuição normal, o teste estatístico utilizado fará parte do grupo de testes Não-paramétricos.

2.3.9.1 Teste t de Student - T-test

Um teste paramétrico comum é o Teste t de *Student* - T-test (DEMŠAR, 2006). Esse teste estatístico verifica se a diferença média no desempenho dos classificadores é significativamente diferente de zero. Neste contexto, o T-test sofre de três fraquezas. A primeira é mensurabilidade: o T-test só faz sentido quando as diferenças sobre os conjuntos de dados são proporcionais. Dessa forma, usar o T-test pareado para comparar um par de classificadores faz tanto sentido quanto computar as médias dos conjuntos de dados. A diferença média d é igual à diferença entre os escores médios de dois classificadores, c1 e c2. A única diferença entre o T-test e a comparação de duas médias é usar diretamente o T-test para amostras não relacionadas: o T-test pareado diminui o erro padrão pela covariância entre os classificadores (DEMŠAR, 2006).

Segundo Demšar (2006), a segunda fraqueza do T-test é a necessidade de que as diferenças entre as duas variáveis tenham que possuir uma distribuição normal, dentro do universo de problemas existentes isso limita o uso do T-test para a maioria que encontramos no mundo real. A terceira fraqueza é a influência de *outliers* no T-test, aumentando o erro estimado, assim como ocorre nas médias (DEMŠAR, 2006).

2.3.9.2 Wilcoxon signed-ranks - WSR

Dentre os testes não-paramétricos o teste de *Wilcoxon signed-ranks* - WSR (WILCOXON, 1945), se apresenta como uma alternativa não-paramétrica ao T-test, classificando o desempenho de dois classificadores, comparando-os pelas diferenças negativas e positivas. As diferenças são classificadas de acordo com seus valores absolutos. Sendo duas medidas de avaliação X_1 e Y_1, $d_1 = (X_1\text{-}Y_1)$, os valores onde $d_1 = 0$ são excluídos. O módulo de d_1 ($|d_1|$) é então ordenado em R'_1, sendo iniciado em 1 e seguindo a ordem natural (1, 2, 3). Ao existir um número grande de observações (a depender do autor esse valor varia de 10 a 25), então tem-se a fórmula representada na equação 6.

$$T = \frac{\sum R_1}{\sqrt{\sum R_1^2}} \bigcap N(0,1) \tag{6}$$

O WSR assume a mensurabilidade das diferenças, mas apenas de forma qualitativa, pois ignora as magnitudes. Como não depende de distribuições normais, é considerado mais seguro que o T-test, além de não ser tão afetado pelos *outliers* (DEMŠAR, 2006).

Quando testes não paramétricos como o WSR, são realizados em múltiplos experimentos se faz necessário calcular a correção de teste Bonferroni. A correção de Bonferroni é calculada dividindo o p-value obtido pelo número de comparações que serão executadas(HOCHBERG, 1988).

Usando esse teste, o nível de significância da família é no máximo α, para qualquer configuração das médias da população. Dessa forma, temos que o teste de Bonferroni protege a taxa de erro da família dos testes.

2.4 Trabalhos relacionados

A UTI possui um monitoramento constante de pacientes em estado grave. Com a evolução tecnológica, esse monitoramento passou a ser realizado de forma automatizada e os dados gerados armazenados para registro. O fluxo constante de informações provenientes desses registros gera uma sobrecarga para médicos e enfermeiros que precisam analisar os dados de forma rápida, para tomar a melhor decisão de tratamento para os pacientes da UTI. Nesse cenário, inúmeros estudos visando utilizar técnicas de mineração de dados de modo a classificar com maior precisão o

estado de gravidade de pacientes foram desenvolvidos (XIA et al., 2012; KIM; KIM; PARK, 2011; WONG; YOUNG, 1999; ALVES et al., 2003; NAVAZ et al., 2016)

Para descrição nesta seção, procurou-se estudos que mostrassem comparações entre técnicas de mineração de dados para predição de mortalidade em UTI e os escores de gravidade mais usados, foi utilizado nas chaves de busca os termos "APACHE", "ICU", "Data Mining","ICU SCORE", "Score Mortality ICU", "Machine Learning". Foram filtrados então os trabalhos que trouxessem informações de medidas de avaliação que pudessem ser comparadas entre si e com o presente estudo.Os trabalhos resumidos a seguir, mostram variações nas comparações entre as predições dos modelos tradicionais e as técnicas de mineração de dados estudadas, o que embasa a continuidade do estudo, seja para criar ou aperfeiçoar aplicações de técnicas de mineração de dados, melhorar o pré-processamento de dados, analisar os efeitos dos dados ausentes e do balanceamento de classe, ou para alterar a forma de cálculo dos escores atuais.

Em Kim, Kim e Park (2011), foi desenvolvido um modelo de predição de mortalidade com os dados adquiridos da *University of Kentucky Hospital* e foram comparados a performance de várias técnicas de mineração de dados, dentre elas RNA e RL. Tal estudo usou 38.474 admissões em UTI entre janeiro de 1998 e setembro de 2007, todos os atributos dos pacientes que foram aplicados ao modelo de predição fazem parte do escopo de atributos utilizado no APACHE III. A RNA utilizada foi a *Multi-Layer Perceptron*, parametrizada com 2 camadas intermediárias, os números de neurônios de cada camada intermediária foram obtidos após testes exaustivos sendo escolhidos os de melhores acurácia. A regressão logística utilizada pertence ao escopo do APACHE III e para avaliação da performance dos modelos de predição Kim, Kim e Park (2011) utilizou a AUC, para assegurar a capacidade dos modelos de distinguir entre os pacientes sobreviventes e os não sobreviventes. A RNA obteve um AUC de 0.874, enquanto a RL obteve um AUC de 0.871.

Em Wong e Young (1999), o objetivo era comparar a habilidade de predição de mortalidade de uma RNA com o escore do APACHE II. Foram usados os dados de 8.796 pacientes de 26 UTI's do Reino Unido e Irlanda, coletados entre outubro de 1987 e abril de 1989. Os atributos utilizados pertenciam ao escopo de atributos do APACHE II e valores ausentes foram substituídos por valores considerados normais. A RNA de Wong e Young (1999) foi parametrizada com 2 camadas intermediárias e a regra de aprendizado foi a *backpropagation*. Para determinar o número de neurônios das camadas intermediárias foi iniciada RNA exaustivas vezes e selecionados os parâmetros com melhor performance. A métrica de avaliação utilizada foi a ROC AUC, onde a RNA obteve um AUC de 0.84, enquanto o APACHE II obteve um AUC de 0.83.

Em Xia et al. (2012) o objetivo foi a criação de uma RNA para predição da mortalidade em UTI. Os dados utilizados foram obtidos do *PhysioNET Challenge 2012* (SILVA et al., 2012), com 12000 registros, que foram divididos em 3 conjuntos de 4000 registros cada, para treinamento, teste e validação, e continham dados fisiológicos colhidos nas primeiras 48 horas de

admissão na UTI. Dentro das técnicas de padronização dos dados foi verificado que nem todos os registros possuíam todos os dados catalogados. Dessa forma, foi colocado um valor padrão nos dados ausentes, assim, se o registro do paciente não possuía temperatura, foi colocado o valor situado entre 36 a 38.4º. Foram utilizados dois escores para avaliar o algoritmo, sendo o primeiro baseado na sensibilidade e precisão, e o segundo em *Hosmer-Lemeshow*. Devido a performance, o foco do trabalho foi no primeiro escore. Foram testadas várias topologias para a RNA, sendo que a que apresentou a melhor performance tinha 2 camadas intermediárias com 15 neurônios. O algoritmo de treinamento utilizado foi o de *back-propagation*, e o de *Levenberg-Marquad*. Os testes da RNA em Xia et al. (2012) foram repetidos 100 vezes, a predição usada foi a média das 100 probabilidades registradas e foi exposto ao treinamento mais sobreviventes do que mortos, para obter um *over sampling* dos registros positivos. Como a saída RNA é normalmente 0 ou 1 foi utilizado um limiar de *fuzzy*, ou seja, se durante o treinamento o limiar encontrado foi de 0.35, um escore de 0.34 era considerado negativo enquanto um de 0.36 era positivo. Nos dados de teste foi obtido uma sensibilidade de 0.8211.

Na pesquisa de Ghose et al. (2015) foi utilizado a RF para predição de mortalidade no desafio da *Physionet* (SILVA et al., 2012), por ser considerada capaz de prover resultados consistentes mesmo com dados com ruído e ausentes. A base de dados utilizada possuía um número de 4000 admissões na UTI e incluía 42 atributos. No estudo, a RF foi parametrizada para usar 1000 estimadores (árvores de decisão) a fim de aumentar a generalização do modelo (GHOSE et al., 2015). A validação dos testes foi realizada por *Cross-Validation* com 10 pastas e o resultado obtido foi um AUC = 0.79.

Seguindo na área de predição em saúde, a RF foi usada no estudo de Scerbo et al. (2014), para triagem de pacientes de trauma em um hospital. Neste estudo foi usado uma base de 1653 pacientes sendo que 496 foram usados para validação do modelo de RF. Foram utilizados 400 estimadores e cada estimador recebeu 7 atributos. A sensibilidade obtida foi de 89%, especificidade de 42%, predições negativas representaram 92% e valores positivos 34%. O estudo em Scerbo et al. (2014) concluiu que a RF teria potencial para auxiliar na triagem de pacientes, pois foi visto um decréscimo da *over-triage* (que representa os pacientes encaminhados para o hospital que não deveriam ter sido encaminhados) de 24%.

Em Taylor et al. (2016), foi analisada a predição de mortalidade da RF, e da RL de pacientes em uma emergência de hospital. Foram analisados 5.278 admissões, sendo separados 1056 registros para validação dos modelos. A RF utilizada contou com 500 estimadores e 25 atributos foram disponibilizados para serem distribuídos entre eles, usando o método próprio de escolha da RF (*out-of-bag*). A RL foi o modelo padrão do escore APACHE III e alcançou o AUC de 0.76, enquanto a RF alcançou um AUC de 0.86, sendo a RF considerada como uma promessa para auxiliar nas decisões realizadas pelo hospital, necessitando de mais estudos para validação dos resultados.

Awad et al. (2017), procurou destacar os principais desafios da predição de mortalidade

em pacientes, utilizando o aprendizado de máquina como abordagem para o framework chamado Previsão de Mortalidade Precoce para Unidade de Terapia Intensiva (EMPICU - *Early Mortality Prediction for Intensiva Care Unit*). No estudo foi utilizado a base do *Multiparameter Intelligent Monitoring in Intensiva Care II* - MIMIC-II, sendo utilizados 11.722 pacientes dos 25.000 disponíveis na base de dados. Os pacientes selecionados possuíam idade acima de 16 anos e eram provenientes das admissões na UTI. As técnicas utilizadas foram *Random Forest* - RF, Árvore de Decisão - AD e *Naive Bayes* - NB. O estudo de Awad et al. (2017) considerou que embora existam muitos valores ausentes nas primeiras 24 horas de admissão do paciente na UTI, essa ausência não provocava um ruído suficiente para interferir na predição de mortalidade. Para tratamento dos valores ausentes foram comparadas duas técnicas: substituição dos valores ausentes por média e utilização de modelos de regressão linear para substituir o valor ausente. Para tratar o desbalanceamento de classe foi utilizado o SMOTE, que cria exemplos sintéticos da classe minoritária, igualando a classe majoritária. Além disso, a base foi dividida por grupos de atributos para aplicação das técnicas: Sinais vitais, 5 atributos com maior ganho de informação, medido pelo software WEKA, e os 10 melhores atributos obtidos pela mesma forma, além de utilizar a base inteira. Os melhores valores obtidos foram com a *Random Forest*: AUC = 0.89 ± 0.02 para base completa com SMOTE seguido de AUC = 0.86 ± 0.02 para base original com 5 atributos.

O estudo de Fialho et al. (2012) utilizou *Data Mining* em conjunto com o modelo *fuzzy* para melhorar a predição de mortalidade em pacientes com re-admissão na UTI e foi utilizado o MIMIC-II como fonte de base de dados. Para o tratamento de dados ausentes foi utilizada a abordagem de exclusão dos pacientes com atributos ausentes. Dessa forma, do total de 25.000 pacientes presentes na base de dados do MIMIC-II, o estudo utilizou 135 registros de pacientes que atendiam aos parâmetros por possuir todos os atributos e terem sido readmitidos na UTI entre 24 e 72 horas. Após aplicação das técnicas para predição, a AUC encontrada foi de 0.72.

Em Grnarova et al. (2016) foi desenvolvida uma técnica de predição automática de mortalidade baseada no conteúdo não estruturado de notas clínicas. Os dados utilizados pertencem ao MIMIC-III, com cerca de 59.000 registros de pacientes e dessa base foi selecionado para o estudo 31.244 pacientes que atendiam aos parâmetros de seleção. Diferente dos outros estudos citados, as informações dos pacientes utilizadas para alimentar uma rede neural, foi o conteúdo das notas e registros do corpo clínico (conteúdo não estruturado), e não os atributos fisiológicos dos pacientes. Com isso, o universo de 31.244 pacientes possuía 812.518 notas associadas. O estudo construiu um vocabulário baseado nas 300.000 expressões mais frequentes do conjunto de dados. O AUC obtido pelo estudo foi de 0.682.

Schmidt et al. (2018), busca em seu estudo melhorar a qualidade de predição de mortalidade utilizando técnicas de Deep Learning, avaliando o resultado de uma Rede Neural Convolucional - RNC em comparação com o APACHE II. Os dados utilizados foram da base do MIMIC III e as métricas de avaliação do estudo foram a área sobre a curva ROC (AUC),

especificidade e sensibilidade. A base de dados foi dividida em dois conjuntos de dados: uma base D1 contendo as variáveis consideradas na versão padrão do escore APACHE II: temperatura, pressão arterial média, frequência cardíaca, frequência respiratória, oxigenação (FiO2, PaO2, PaCO2), pH arterial, sódio sérico, potássio sérico, creatinina sérica, hematócrito, leucócitos, pontos na escala de coma de Glasgow e idade, exceto problemas crônicos de saúde; a outra base D2 continha dados formados pelas variáveis presentes no primeiro conjunto, mais 11 outras variáveis: peso na admissão, glicose, SpO2, plaquetas, cloreto, hemoglobina, magnésio, pressão arterial sistólica, pressão arterial diastólica, CO2, pontos na escala de Braden. Utilizando o método Holdout o conjunto de dados D1 continha 12.919 registros de entrada, divididos em 7.751 registros para treino, 1.292 para validação e 3.876 para avaliação. O conjunto de dados D2, contendo 11.191 entradas, foi dividido em 6.715 registros para treino, 1.119 para validação e 3.858 para avaliação. No estudo de Schmidt et al. (2018) não foi descrito nenhuma metodologia de pré-processamento, como balanceamento dos conjuntos de dados de teste ou tratamento de valores ausentes das variáveis selecionadas. Após utilização de 3 versões do algoritmo de RNC proposto por Schmidt et al. (2018), o estudo considerou o desempenho de sua terceira versão melhor, obtendo um AUC de 0,85 com uma sensibilidade de 0,75 e especificidade de 0,95.

Não foi vista uma variedade de técnicas utilizadas para tratamento de atributos ausentes. Basicamente foi usado exclusão e valor padrão. Além disso, os modelos de predição mais usados foram *Random Forest* e Redes Neurais Artificiais, sendo que em Scerbo et al. (2014) e Awad et al. (2017) a *Random Forest* obteve seu maior valor. A tabela 6, mostra os valores do tamanho da base de pesquisa (Amostra), valores da AUC obtidos com os classificadores: RNA - Redes Neurais Artificias, RL - Regressão Logística e RF - *Random Forest*, abordagens de tratamento para valores ausentes e abordagens para tratamento de desbalanceamento.

Tabela 6 – Resumo das características dos estudos

Estudo	Amostra	RNA	RL	RF	Valor Ausente	Desbalanceamento
Wong e Young (1999)	8.796	0.840	0.830		Default	-
Kim, Kim e Park (2011)	38.474	0.874	0.871		Default	-
Xia et al. (2012)	12.000	0.820			Default	-
Scerbo et al. (2014)	1.653			0.890	Exclusão	-
Ghose et al. (2015)	4.000			0.790	Default	-
Taylor et al. (2016)	5.278		0.760	0.860	Exclusão	-
Grnarova et al. (2016)	31.244	0.682			Exclusão	UnderSampling
Awad et al. (2017)	11.722			0.890	Exclusão	SMOTE
Schmidt et al. (2018)	11.191	0.850			-	-

3 METODOLOGIA

Como visto no capítulo 2.4 foram encontradas diversidade nos resultados da predição de mortalidade entre técnicas de mineração de dados, durante a revisão de literatura. Além disso, a base do MIMIC-III que possui mais registros que a base MIMIC-II, não foi amplamente utilizada. Um estudo de parametrização das técnicas de mineração, do pré-processamento dos dados, da análise do desbalanceamento de classe e seu impacto na predição de mortalidade e testes com outros métodos de cálculo de escores, poderão se mostrar importantes para melhoria da predição de mortalidade, e consequentemente prover uma ferramenta confiável para uso pela equipe médica. O estudo aqui proposto busca aplicar modelos de mineração de dados que possam prover uma melhor predição de mortalidade em UTI, comparando os resultados de Redes Neurais Artificiais, *Random Forest* e Regressão Logística, a fim de identificar atributos que sejam considerados relevantes e analisar a aplicação de metodologias de tratamento do problema de desbalanceamento de classes e tratamento de dados ausentes.

Foi utilizada neste estudo a biblioteca de aprendizado de máquina *Scikit-Learn* (PEDREGOSA et al., 2011) da linguagem de programação Python. Essa ferramenta foi escolhida por possuir uma ampla base de conhecimento e documentação, fóruns de discussões e uma comunidade ativa na web, além de possuir licença aberta para uso.

3.1 Dados

Os dados utilizados para ajuste e avaliação foram obtidos da base MIMIC-III ("Medical Information Mart for Intensive Care"), uma iniciativa do MIT - Massachusetts Institute of Technology, em parceria com o *Medic Center Beth Israel Deaconess* (JOHNSON et al., 2016). Esta base de dados inclui informações relativas a pacientes internados em unidades de cuidados intensivos em um hospital de cuidados terciários de grande porte.

Os dados incluem sinais vitais, medicamentos, medidas laboratoriais, observações e notas traçadas pela equipe médica e de enfermagem, balanço de fluidos, códigos de procedimentos, códigos de diagnóstico, relatórios de imagem, período de permanência hospitalar, dados de sobrevivência e outros tópicos.

O MIMIC-III apoia aplicações, incluindo pesquisa acadêmica e industrial, iniciativas de melhoria de qualidade e cursos de educação superior. A base contém atualmente cerca de 54000 registros de pacientes internados em UTI entre 2001 e 2012.

O MIMIC-III integra dados clínicos detalhados e sem identificação de pacientes admitidos no *Medic Center Beth Israel Deaconess* em Boston, Massachusetts, e os torna acessíveis a qualquer pesquisador do mundo, sob um contrato de uso de dados. Desta forma, estudos clínicos podem ser reproduzidos e melhorados colaborativamente (JOHNSON et al., 2016). Na figura 9 é

mostrado o *overview* do processo de obtenção da base de dados.

Utilizou-se nesse estudo, as informações de 65132 admissões de pacientes em UTI.

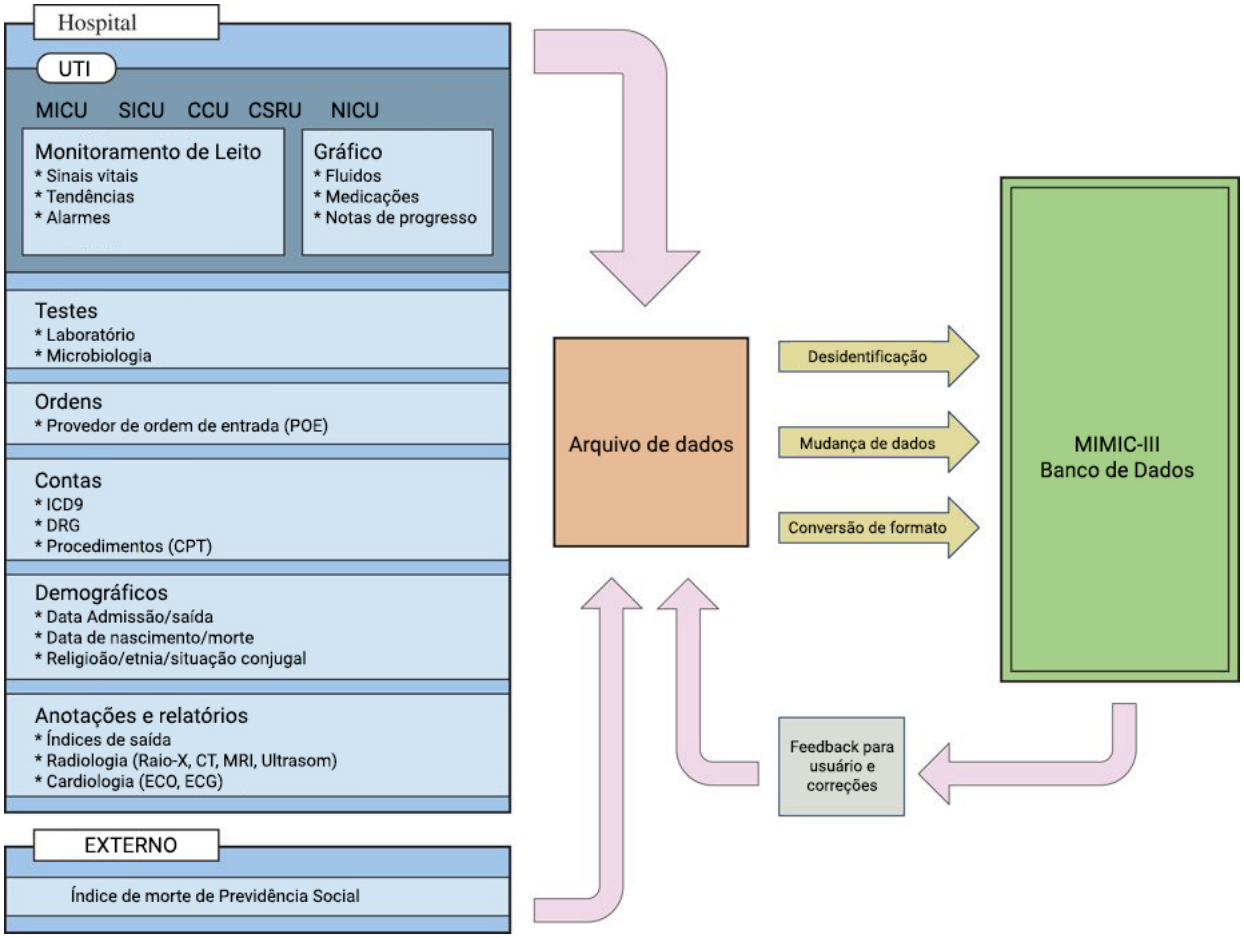

Figura 9 – Base MIMIC - III (JOHNSON et al., 2016)

Os dados dos pacientes que estão no MIMIC-III, são informações coletadas durante a rotina hospitalar, de modo a não acarretar transtorno ao fluxo da UTI, sendo as fontes de informações arquivos de sistemas de informações, dados registrados no hospital e informações provenientes da Administração de Seguro Social, nos casos de registro de falecimento.

As informações consistem em dados demográficos de pacientes e mortalidade hospitalar, resultados de exames laboratoriais, relatos de exames de imagem e eletrocardiogramas e informações relacionadas ao faturamento, como códigos da Classificação Internacional de Doenças, 9ª edição (CID-9), códigos de Grupo Relacionado com Diagnóstico (DRG) e códigos de Terminologia Processual Atual (CPT).

A data de falecimento dos pacientes foi obtida através de consulta a Administração de Seguro Social. Dados fisiológicos foram obtidos por formulários preenchidos com informações dos monitores eletrônicos da UTI.

A base possui cerca de 54.000 pacientes, sendo que é registrado cada entrada de paciente na UTI, totalizando 61532 registros. Na base de dados existem mais exemplos de pacientes

sobreviventes do que pacientes não sobreviventes: o número de pacientes sobreviventes é de 45318, enquanto o número de não sobreviventes é de 16214. Dessa forma a classe sobrevivente representa cerca de 74% da base.

Através de contribuições de pesquisadores e colaboradores do MIMIC-III, foi possível obter os registros de admissão na UTI já pré-filtrados, contendo os valores dos atributos dos pacientes dentro da janela de tempo de 24 horas de admissão. Foram retiradas as informações de recém-nascidos e pacientes que possuíam câncer, por pertencerem a uma categoria específica de tratamento. Os atributos utilizados fazem parte do escopo do APACHE-III e foi utilizado esse conjunto de atributos por, dentre os escores de severidade, possuir a maior quantidade de atributos.

Na seção 4.1, será mostrada a análise estatística dos dados e atributos utilizados, procurando identificar possíveis correlações entre os atributos, impactos dos dados ausentes e do desbalanceamento de classe.

3.2 Pré-processamento

Antes de executar os classificadores na base de dados, é necessário tratar as informações existentes de modo a evitar situações em que a classificação poderá ser prejudicada por atributos ausentes ou desbalanceamento de classes.

Devido aos atributos ausentes, foi necessário utilizar abordagens para atribuir valores, tendo em vista que alguns classificadores não trabalham com valores ausentes. Uma das abordagens utilizadas foi a atribuição de valor considerado normal na literatura médica para cada atributo ausente (Tabela 7), assim como é citado por (KNAUS et al., 1981), que utiliza essa estratégia para cálculo do APACHE. Por exemplo, se não existia um valor para medição da temperatura do paciente, foi atribuído o valor 36,5 graus célsius.

Para valores que possuem intervalo de normalidade, como por exemplo batimentos cardíacos (80 a 100) foi selecionado o valor da média, nesse caso 90. Em conjunto com essa técnica foi criado um atributo de presença ou ausência para cada atributo de modo a identificar se a ausência do atributo possa se tornar um bom preditor.

Foi também utilizado a substituição de dados ausentes por dados obtidos através de modelos de predição, sendo utilizado o modelo *Multivariate Imputation by Chained Equations* (MICE) para obtenção dos valores da cada atributo.

Modelos de predição são influenciáveis por bases com classes desbalanceadas, tendendo a atribuir a classificação dos dados para a classe dominante. Para tratar esta situação da base, foi utilizada técnica de balanceamento de classes *over sampling* SMOTE, para que a base utilizada para treinamento dos modelos, tivesse o mesmo número de sobreviventes e não sobreviventes a fim de que o classificador não sofresse influência da uma classe dominante. Também foi repetida

Tabela 7 – Valores padrões atribuídos a dados ausentes

Atributo	**Valor normal**
Batimento cardíaco	80-100
Pressão sanguínea	100-120
Temperatura em celsius	36-37
Taxa Respiratória	18-22
Pressão arterial de oxigênio	100
Diferença de tensão arterial alveolar	100
PH	7.4
Pressão arterial de gás carbônico	45
Hematócrito	40
Contagem de glóbulos brancos - Mínimo	13
Creatinina	1
Nitrogênio Uréico no Sangue	17
Sódio	140
Albumina	4.5
Bilirrubina	2
Glucose	200
Uso de ventilação extracorpórea	0
Medição do volume de urina	410
Escore Glasgow mínima	12
Escore Glasgow motor	5
Escore Glasgow verbal	4
Escore Glasgow visual	3
Entubação endotraqueal	0
Falha renal	0

a aplicação dos modelos de predição usando a técnica de *under sampling* NCR. Para efeito de comparação dos resultados, os modelos foram aplicados à base sem utilização de *under sampling* e *over sampling*.

3.3 Classificação

Foram selecionados os classificadores Regressão Logística - RL, Redes Neurais Artificias - RNA e *Random Forest* - RF, para utilização no trabalho. Estes classificadores representam diferentes abordagens de classificação, são bem documentados e podem servir de parâmetros para futuras escolhas ou alteração de paradigma para classificação de mortalidade em UTI.

Para encontrar os parâmetros de cada classificador utilizado, foram aplicados os mesmos em uma base de treino separada (não utilizada para treino dos classificadores), sendo escolhido primeiramente, parâmetros padrões, obtendo as métricas de avaliação, e executado novamente alterando um dos parâmetros, caso houvesse aumento da métrica, manteve-se o novo parâmetro e passa para o próximo, repetindo o ciclo até selecionar todos os parâmetros.

3.3.1 APACHE III

Para determinar a probabilidade da mortalidade do APACHE III (PM), na base do MIMIC, foi utilizada a equação descrita em Johnson (2014), mostrada na equação 7.

$$PM = \frac{1}{1 + e^{(-(-4.4360 + 0.04726 * (EscoreApache)))}} \tag{7}$$

3.3.2 Regressão Logística - RL

A regressão logística foi configurada para utilizar o otimizador SAGA por se tratar do melhor para aplicação em grandes bases de dados(ZHANG, 2004), dessa forma ainda, segundo Zhang (2004), o dados deverão estar na mesma escala. Para comparação de resultados e verificação do efeito do desbalanceamento de classes, o mesmo método foi realizado em uma base sem *over sampling* e com *under sampling*. Os resultados serão discutidos na seção 4.

3.3.3 Redes Neurais Artificias - RNA

Foi utilizada uma RNA do tipo *Multi-Layer Perceptron* e da mesma forma foi aplicada a base de treinamento com *over sampling* e teste sem *over sampling* e base com *under sampling*, com a diferença de terem sido normalizados os valores dos atributos para o intervalo de 0 a 1, dado a sensibilidade da RNA a dados sem normalização, que atribui maiores pesos aos dados com maiores valores. Após reinicializar a RNA 120 vezes em base de treino, alterando os parâmetros da rede até conseguir um valor com melhor desempenho de AUC, foram obtidos os valores de máximo de iteração igual a 500, números de neurônios da camada intermediaria igual a 300 e otimizador ADAM (um gradiente estocástico proposto por Kingma e Ba (2014)). A RNA utilizada possui o mesmo número de neurônios de entrada que o número de atributos, neste caso 35 e a camada de saída possui 1 neurônio devido a saída da rede ser binária (0,1 - sobrevivente ou não sobrevivente). Os resultados obtidos serão discutidos no capítulo 4.

3.3.4 *Random Forest* - RF

A RF criada possui 100 estimadores e com limite de expansão dos nós igual a 400. O número de estimadores foi definido após testes de parâmetros em uma base de treino separada, com aumento progressivo de estimadores de 10 a 300, sendo que a partir de 100 estimadores não havia acréscimo significativo da AUC, todavia havia um aumento proporcional no tempo de execução da RF. Da mesma forma foi selecionado a profundidade máxima das árvores em 400 nós e cada árvore recebeu 5 atributos.

Na tabela 8, estão descritos quais os parâmetros foram utilizados nos classificadores.

Tabela 8 – Parâmetros utilizados nos classificadores

Classificador	Parâmetros
Regressão Logística	Random state = 42, Otimizador SAGA
Random Forest	Random state = 42, Produndidade máxima = 400, Estimadores = 100, Atributos por estimador = 5
Redes Neurais Artificiais	Random state = 42, Otimizador ADAM, Máximo de iterações = 500, Camadas Intermediarias = 300

3.3.5 Métodos e medidas de avaliação

Para validação dos modelos, estimando seu desempenho, foi utilizada a *k-fold Cross-Validation* estratificada de k igual a 2, sendo repetido o *Cross-Validation* 10 vezes. Sendo estratificada, a *Cross-Validation* mantém a proporção da classe alvo.

A aplicação da *Cross-Validation* em conjunto com técnicas de balanceamento, os métodos de *Over sampling* e *Under sampling* foram aplicados só para os dados de treino, após a separação da base de teste da base de treino. Caso o balanceamento fosse feito na base completa ou nos dados de teste, poderia distorcer a real distribuição ou lançar na pasta de teste dados que foram gerados sinteticamente pelo SMOTE e possivelmente duplicados com os dados de treino, o que difere da exclusividade dos dados de validação proposto pela *Cross-Validation* (ALTINI, 2016).

As medidas de avaliação utilizadas nos modelos foram a Área sob a Curva - AUC, sensibilidade e especificidade. Para melhor visualização será apresentado na seção 4, os gráficos de curva ROC, e matriz de confusão.

3.3.6 Testes estatísticos

Será utilizado o teste *Wilcoxon signed-rank* - WSR para validar a comparação entre os classificadores. Devido aos múltiplos testes que serão executados, seguindo a literatura estatística associada (MCDONALD, 2014), será necessário utilizar a correção de Bonferroni, que consiste em simplesmente, dividir o valor encontrado do teste WSR pelo número de testes executados.

4 RESULTADOS

Neste capítulo serão apresentados os resultados obtidos pelo estudo e aplicação da metodologia descrita anteriormente. Será apresentada a análise exploratória dos dados, buscando caracterizar melhor a base, e identificar os problemas de dados ausentes e desbalanceamento de classe, para então aplicarmos as estratégias já apresentadas para minimizar os efeitos na classificação.

4.1 Análise exploratória dos dados

A base analisada possui 61.532 registros de pacientes, sendo 45.318 pacientes sobreviventes e 16.214 pacientes não sobreviventes. Os atributos analisados fazem parte do escopo de utilização do APACHE-III, e os demais escores utilizam parte desses atributos para seus cálculos, sendo o APACHE o que contém a maior quantidade de atributos. No estudo foram utilizados 35 atributos: Batimentos Cardíacos, mínimo e máximo, Pressão sanguínea, mínimo e máximo, Temperatura, mínimo e máximo, Taxa respiratória, mínimo e máximo, Pressão arterial de oxigênio, Tensão arterial alveolar, PH, Pressão de CO2, Hematócrito, mínimo e máximo, Contagem de glóbulos brancos, mínimo e máximo, Creatinina, mínimo e máximo, Nitrogênio ureico no sangue, mínimo e máximo, Sódio, mínimo e máximo, Albumina, mínimo e máximo, Bilirrubina, mínimo e máximo, Glucose, mínimo e máximo, Ventilação extracorpórea, Volume de urina, Escore Glasgow, mínimo, Escore Glasgow motor, Escore Glasgow verbal, Escore Glasgow visual. Os valores estatísticos descritivos dos atributos estão apresentados na tabela 9.

Nesse capítulo serão discutidos a interferência na classificação dos atributos ausentes e o desbalanceamento de classe.

4.1.1 Atributos Ausentes

A base de dados continha registros de atributos ausentes, alcançando 68% para alguns atributos, como mostrado na tabela 9. Conforme Knaus et al. (1985) descreve em seu estudo, a ausência do atributo segue um padrão determinado pelos valores obtidos dos sinais vitais (pressão e batimento cardíacos, taxa respiratória, hemogramas): se esses valores apresentam valores próximos a normalidade, menos exames são solicitados pelo corpo médico e consequentemente menos dados são registrados para o paciente. Os atributos pertencentes a verificação da gasometria do paciente (Pressão de arterial de oxigênio, Tensão arterial alveolar, Pressão arterial de CO2) são exemplos de exames que só serão requeridos em casos que apresentem sinais vitais irregulares, assim como a aferição da Bilirrubina e Albumina para quadros com sintomas de falha renal. Desta forma, os atributos não anotados corresponderiam a valores de normalidade, pois não representariam para o quadro de saúde do paciente alterações significativas. A Tensão

arterial alveolar possui na base original cerca de 90% de ausência e ao investigar possíveis razões para não registro da informação, encontramos na literatura que esse atributo pode ser calculado através dos valores de Pressão arterial de oxigênio e Pressão arterial de CO2. Dessa forma, foi utilizado a fórmula da equação 8, para atribuição de um valor quando temos elementos suficientes para o cálculo.

$$AaO2 = 130 - (PaCO2 + PaO2) \tag{8}$$

No histograma da figura 10, é mostrada a quantidade de atributos ausentes em relação a classe alvo. As barras de cor verde mostram os pacientes sobreviventes e as barras vermelhas os pacientes não sobreviventes. Para melhor visualização das barras, o número de pacientes foi normalizado em valores de 0 a 1.

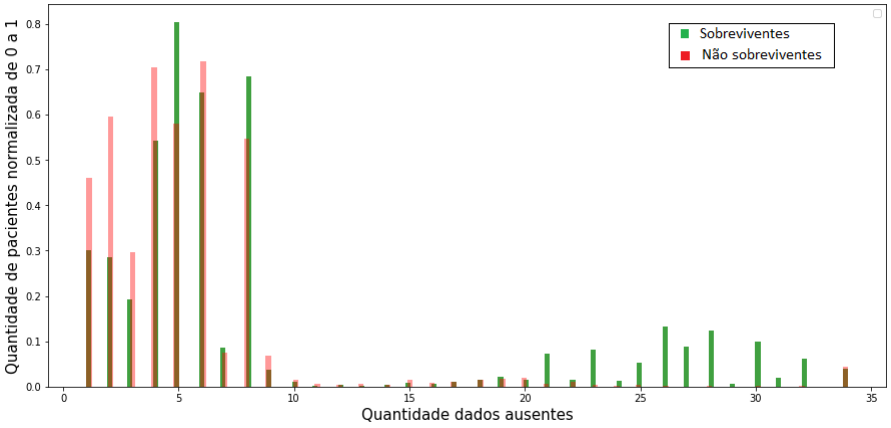

Figura 10 – Histograma de Ausência por Mortalidade

Pode ser identificado através da distribuição do histograma que a maioria dos pacientes possui de 0 a 9 atributos ausentes e que de 23 a 32 atributos ausentes, a base possui poucos registros de não sobreviventes. Isto pode evidenciar a relação de menos exames e monitoramento para pacientes que apresentam sinais vitais melhores. De 10 a 20 atributos ausentes, a distribuição das barras horizontais é semelhante entre não sobreviventes e sobreviventes.

De modo a entender melhor a distribuição dos atributos, nas tabelas 9, 10 e 11 são mostradas as características dos atributos da base, tais como percentual de ausência, média dos valores, desvio padrão, valores mínimos e máximos e valores normais de todo o conjunto de dados e de dois recortes: um contendo somente pacientes sobreviventes (tabela 10) e outro contendo somente pacientes não sobreviventes(tabela 11).

Comparando a tabela 10 e a tabela 11 identifica-se que a quantidade de atributos ausentes sofre uma diminuição para os pacientes não sobreviventes em relação aos pacientes sobreviventes, a exemplo de Bilirrubina Min e Max, sendo que para sobreviventes a ausência foi de 45,30% enquanto para pacientes não sobreviventes a ausência foi de 12,27%, Pressão arterial de oxigênio

Tabela 9 – Descrição estatística da base por atributos

Atributo	Ausência	Média	Dv. Padrão	Mínimo	Máximo	Vl.normal
Batimento cardíaco-Mín	4,03%	77,49	23,39	0,15	218	80
Batimento cardíaco-Máx	4,03%	109,72	26,58	30	286	80
Pressão sanguínea-Mín	15,02%	58,31	13,81	0,20	125	100
Pressão sanguínea-Máx	15,02%	36,00	0,79	15	40,83	100
Temperatura em célsius-Mín	16,50%	36,09	0,79	15	40,83	37
Temperatura em célsius-Máx	16,50%	37,47	0,80	30,00	46,50	37
Taxa Respiratória-Mín	15,11%	12,36	3,76	0,20	47	20
Taxa Respiratória-Máx	15,11%	27,27	6,59	8,00	69,00	20
Pressão arterial de oxigênio	68,47%	263,19	160,20	18,00	797,00	100
Tensão arterial alveolar	68,47%	417,57	155,89	70,00	652,00	100
PH	44,14%	7,36	0,11	6,40	7,89	7,4
Pressão arterial de CO2	44,14%	42,70	14,36	7,00	243,00	45
Hematócrito-Mín	3,23%	32,03	9,06	4,00	68,70	40
Hematócrito-Máx	3,23%	37,20	7,53	11,00	74,40	40
Cont. glóbulos brancos-Mín	3,83%	11,01	7,94	0,10	575,80	13
Cont. glóbulos brancos-Máx	3,83%	13,88	10,73	0,10	846,70	13
Creatinina-Mín	14,02%	1,32	1,41	0,10	28,00	1
Creatinina-Máx	14,02%	1,57	1,71	0,10	46,60	1
Nitrogênio Uréico-Mín	14,04%	24,03	19,96	1,00	254,00	17
Nitrogênio Uréico-Máx	14,04%	28,38	22,82	1,00	272,00	17
Sódio-Mín	10,93%	136,66	4,99	1,21	178,00	140
Sódio-Máx	10,93%	140,08	4,64	97,000	182,00	140
Albumina-Mín	68,59%	3,11	0,72	1,00	6,30	4,5
Albumina-Máx	68,69%	3,20	0,71	1,00	6,30	4,5
Bilirrubina-Mín	57,57%	2,04	4,15	0,10	79,00	2
Bilirrubina-Máx	57,57%	2,33	4,64	0,10	82,80	2
Glucose-Mín	11,50%	103,39	34,06	0,10	576,00	200
Glucose-Máx	11,50%	206,31	4286	9	999999	200
Ventilação extracorpórea*	0%					
Volume de urina	13,53%	1820,67	2842,27	-2600	561190	
Escore Glasgow mínima	15,75%	13,74	2,61	3,00	15,00	15
Escore Glasgow motor	15,41%	5,20	1,50	1,00	6,00	6
Escore Glasgow verbal	15,73%	3,27	2,19	0	5,00	5
Escore Glasgow visual	15,41%	3,18	1,06	1,00	4,00	4
Entubação endotraqueal*	15,2%					0
Falha renal*	0%					0
Escore APACHE III	0%	40,54	20,61	0	185,00	0

*Valores binários: 1 para ocorrência e 0 para não ocorrência

Tabela 10 – Descrição estatística da base por atributos para sobreviventes

Atributo	Ausência	Média	Dv. Padrão	Mínimo	Máximo
Batimento cardíaco-Mín	3,31%	79,55	24,98	0,15	218
Batimento cardíaco-Máx	3,31%	110,97	27,56	37,00	286
Pressão sanguínea-Mín	14,19%	59,70	13,06	0,79	125
Pressão sanguínea-Máx	14,19%	104,94	25,07	51	299,00
Temperatura em célsius-Mín	15,34%	36,15	0,73	15	39,6
Temperatura em célsius-Máx	15,34%	37,51	0,74	32,61	46,50
Taxa Respiratória-Mín	14,26%	12,07	3,56	0,20	35
Taxa Respiratória-Máx	14,26%	26,83	6,36	10,00	69,00
Pressão arterial de oxigênio	50,73%	284,75	160,95	22,00	797,00
Tensão arterial alveolar	73,54%	403,40	155,68	70,00	634,00
PH	35,02%	7,37	0,10	6,78	7,72
Pressão arterial de CO2	35,02%	42,50	13,21	7,00	209,00
Hematócrito-Mín	2,74%	33,10	9,69	4,00	68,70
Hematócrito-Máx	2,74%	38,31	7,78	11,00	74,40
Cont. glóbulos brancos-Mín	3,20%	10,92	6,05	0,10	385,80
Cont. glóbulos brancos-Máx	3,20%	13,62	7,78	0,10	572,50
Creatinina-Mín	13,43%	1,17	1,29	0,10	28,00
Creatinina-Máx	13,43%	1,40	1,60	0,10	46,60
Nitrogênio Uréico-Mín	13,44%	20,51	16,75	1,00	254,00
Nitrogênio Uréico-Máx	13,44%	24,58	19,66	1,00	272,00
Sódio-Mín	10,44%	136,76	4,69	1,21	178,00
Sódio-Máx	10,44%	140,12	4,21	108,000	182,00
Albumina-Mín	53,73%	3,23	0,70	1,00	6,30
Albumina-Máx	53,73%	3,31	0,69	1,10	6,30
Bilirrubina-Mín	45,30%	1,91	3,52	0,10	79,00
Bilirrubina-Máx	45,30%	2,17	4,03	0,10	82,80
Glucose-Mín	11,01%	101,47	30,40	0,10	482,00
Glucose-Máx	11,01%	209,10	5094	9	999999
Ventilação extracorpórea*	0%				
Volume de urina	10,74%	1912,24	3210,09	-2600	561190
Escore Glasgow mínima	14,25%	13,86	2,45	3,00	14,00
Escore Glasgow motor	14,58%	5,38	1,42	1,00	6,00
Escore Glasgow verbal	14,61%	3,40	2,10	0	5,00
Escore Glasgow visual	14,37%	3,25	1,01	1,00	4,00
Entubação endotraqueal*	14,25%				
Falha renal*	0%				
Escore APACHE III	0%	36,25	17,66	0	145,00

*Valores binários: 1 para ocorrência e 0 para não ocorrência

Tabela 11 – Descrição estatística da base por atributos para **não sobreviventes**

Atributo	Ausência	Média	Dv. Padrão	Mínimo	Máximo
Batimento cardíaco-Mín	0,72%	71,85	17,12	2	162
Batimento cardíaco-Máx	0,72%	106,28	23,36	30	280
Pressão sanguínea-Mín	0,82%	55,07	14,91	0,20	122
Pressão sanguínea-Máx	0,82%	104,09	28,32	23,00	299,00
Temperatura em célsius-Mín	1,15%	35,96	0,90	20,90	40,83
Temperatura em célsius-Máx	1,15%	37,37	0,91	30,00	42,77
Taxa Respiratória-Mín	0,85%	13,03	4,11	1,00	47,00
Taxa Respiratória-Máx	0,85%	28,29	6,99	8,00	69,00
Pressão arterial de oxigênio	17,74%	205,78	143,14	18,00	763,00
Tensão arterial alveolar	26,27%	436,36	155,77	180,00	652,00
PH	9,11%	7,34	0,13	6,40	7,89
Pressão arterial de CO2	9,11%	43,14	16,64	8,00	243,00
Hematócrito-Mín	0,48%	29,10	6,13	4,30	61,80
Hematócrito-Máx	0,48%	34,18	5,79	13,00	72,70
Cont. glóbulos brancos-Mín	0,62%	11,24	11,63	0,10	575,80
Cont. glóbulos brancos-Máx	0,62%	14,58	16,24	0,10	846,70
Creatinina-Mín	0,59%	1,67	1,60	0,10	23,60
Creatinina-Máx	0,59%	1,97	1,89	0,10	43,00
Nitrogênio Uréico-Mín	0,60%	32,28	24,03	1,00	196,00
Nitrogênio Uréico-Máx	0,60%	37,27	26,86	1,00	272,00
Sódio-Mín	0,49%	136,39	5,67	13,00	174,00
Sódio-Máx	0,49%	139,96	5,55	97,000	180,00
Albumina-Mín	14,87%	2,90	0,70	1,00	5,50
Albumina-Máx	14,87%	2,99	0,69	1,00	5,50
Bilirrubina-Mín	12,27%	2,30	5,17	0,10	54,90
Bilirrubina-Máx	12,27%	2,64	5,67	0,10	55,50
Glucose-Mín	0,49%	108,05	41,24	0,99	576,00
Glucose-Máx	0,49%	199,54	106,16	31,00	1746,00
Ventilação extracorpórea*	0%				
Volume de urina	2,79%	1576,20	1432,95	-1730	51520,00
Escore Glasgow mínima	1,00%	13,45	2,93	3,00	15,00
Escore Glasgow motor	1,17%	5,04	1,65	1,00	6,00
Escore Glasgow verbal	1,12%	2,95	2,17	0	5,00
Escore Glasgow visual	1,04%	3,02	1,15	1,00	4,00
Entubação endotraqueal*	1,00%				
Falha renal*	0%				
Escore APACHE III	0%	52,53	23,33	0	185,00

*Valores binários: 1 para ocorrência e 0 para não ocorrência

diminui de 50,73% de ausência para sobreviventes, para 17,74% em não sobreviventes. Isso confirma o que o estudo de (KNAUS et al., 1981) encontrou: que um número maior de exames e maior monitoramento é realizado para os casos que apresentam um quadro clínico mais severo. Mesmo com a diminuição na quantidade de atributos ausentes para os pacientes não sobreviventes, a distribuição dos valores dos mesmos não mostra diferença para os pacientes sobreviventes, como podemos ver nos histogramas representados nas figuras 11 e 12. Isso pode representar uma dificuldade para o classificador, pois o valor do atributo não é determinante para identificar se o paciente sobreviveu ou não.

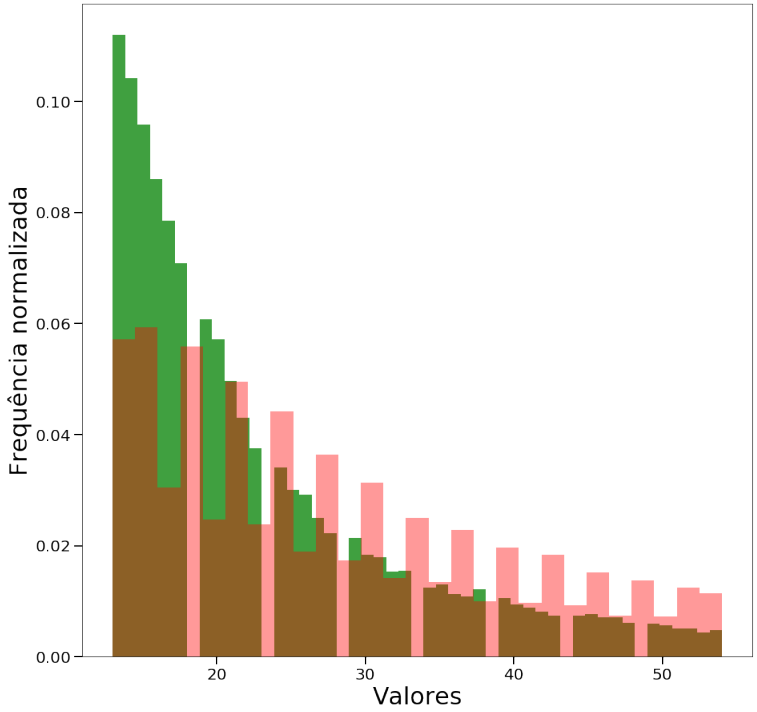

Figura 11 – Histograma do atributo "Nitrogênio no sangue - Mínimo". O rachurado em verde destaca os valores dos sobreviventes, enquanto o rachurado vermelho os não sobreviventes

Para melhor visualização e entendimento do grau de importância dos atributos, na tabela 12 é mostrado o coeficiente de correlação ponto-bisserial de cada atributo em relação a classe alvo, de modo que se possa identificar os atributos que possam estar melhor correlacionados com a mesma. O coeficiente de correlação ponto-bisserial é usado quando uma variável pode assumir dois valores distintos, sendo dicotômica, quando utilizamos uma medida contínua X e uma variável dicotômica Y. O coeficiente de correlação ponto-bisserial é matematicamente equivalente a correlação de *Pearson* (LINACRE; RASCH, 2008). Os atributos possuem baixa correlação com a classe alvo, dessa forma o classificador poderá ter dificuldades para identificar entre sobreviventes e não sobreviventes, já que não há atributos com valores determinantes, ou que só sua presença possa já identificar a classe pertencente do paciente.

Os histogramas apresentados nas figuras 11 e 12 foram obtidos de atributos que possuem maior e menor correlação com a classe alvo: Nitrogênio Ureico e Contagem de células brancas,

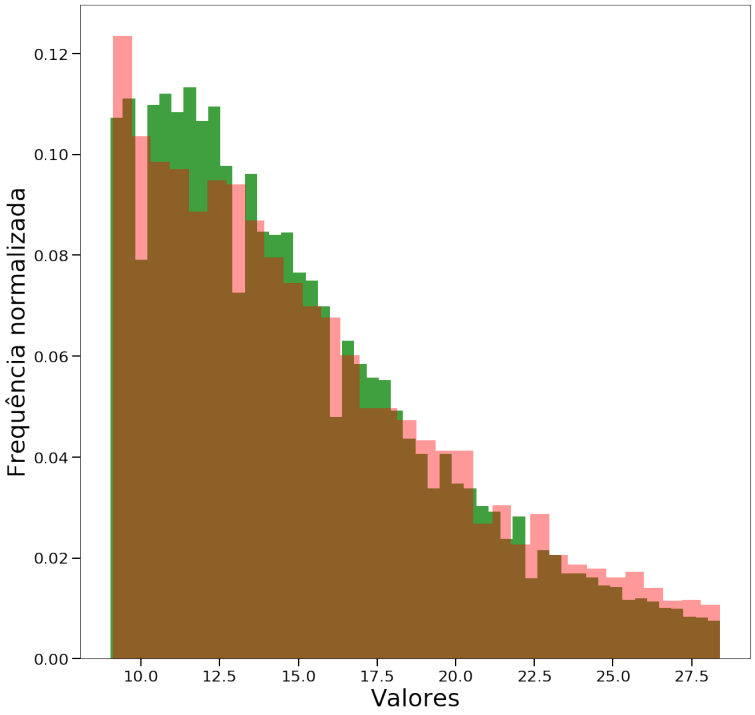

Figura 12 – Histograma do atributo "Contagem de células Brancas". Em verde destaca os valores dos sobreviventes, enquanto em vermelho os não sobreviventes

respectivamente. Mesmo com a maior correlação do Nitrogênio Ureico, o histograma não mostra uma tendência de valores que possam determinar a qual classe ele estaria relacionado.

Esse cenário onde os atributos de não sobreviventes e sobreviventes são muito próximos, pode ser representado graficamente, utilizando o *Principal Component Analyses - PCA*, reduzindo a dimensionalidade dos atributos de 34 para 2 componentes. A figura 13 identifica os pacientes sobreviventes na cor azul e pacientes não sobreviventes na cor amarela. O componente 1 representou um percentual acumulado de informação de 88,8% enquanto o componente 2 0,05%. O componente 1 foi formado pela soma dos valores dos atributos que o PCA identificou como mais relevantes multiplicados pelo seu escore de relevância (tabela 13), dentre eles hematócrito, sódio e batimento cardíaco. A formação do componente 2 pode ser vista pelos atributos e escores mostrados na tabela 14. Embora selecionados pelo PCA, isoladamente esses atributos não possuíam uma correlação alta com a mortalidade do paciente como pode ser visto na Tabela 12.

Como observado no gráfico, as classes estão muito sobrepostas, o que confirma visualmente o que foi apresentado anteriormente. Dessa forma o classificador terá dificuldades para separar as classes.

Foi utilizada a técnica *t-Distributed Stochastic Neighbor Embedding* - t-SNE, descrita em Maaten e Hinton (2008), também para realizar a visualização dos dados. O t-SNE é uma técnica de redução de dimensionalidade considerada bem adaptada para visualizar bancos de dados multidimensionais.

Tabela 12 – Tabela de correlação das bases com valores padrões e ausência e presença de atributos, com a classe de mortalidade (Coeficiente de correlação ponto-bisserial).

Atributo	Corr. Base padrão	Corr. Base Ausência e presença
Nitrogênio Uréico-Mín	0,2845	0,2845
Nitrogênio Uréico-Máx	0,27866	0,26823
Creatinina-Mín	0,17777	0,1729
Creatinina-Máx	0,1701	0,1701
Falha Renal*	0,16767	0,16767
Tensão Arterial Alveolar	0,11995	0,06064
Taxa Respiratória-Mín	0,11952	-0,03472
Taxa Respiratória-Max	0,11211	0,033643
Glucose-Mín	0,09306	-0,06364
Entubação endotraqueal	0,09106	0,09106
Permanência na UTI	0,05842	0,05752
Cont. glóbulos brancos-Máx	0,03997	0,0394
Ventilação Externa	0,03885	0,03885
Pressão arterial de CO2	0,0303	0,00389
Cont. glóbulos brancos-Mín	0,01891	0,01527
Hematócrito-Máx	-0,238268	-0,24228
Hematócrito-Mín	-0,19175	-0,19818
Escore Glasgow visual	-0,1478	-0,1478
Batimento cardíaco-Mín	-0,14561	-0,14411
Escore Glasgow verbal	-0,14561	-0,14561
Pressão sanguínea-Mín	-0,14131	0.043570
Batimento cardíaco-Max	0.032969	-0,06637
Albumina-Max	-0,13431	-0,20488
Albumina-Mín	-0,13296	-0,20497
Escore Glasgow motor	-0,13067	-0,13067
Glucose-Máx	-0,13067	-0,0009
Pressão arterial de Oxigênio	-0,12161	-0,08592
Temperatura em célsius-Min	-0,10696	-0,17757
PH	-0,09797	-0,10914
Temperatura em célsius-Max	-0,06428	-0,0251
Volume de urina	-0,04784	-0,0394
Sódio-Mín	-0,03673	-0,06957
Sódio-Máx	-0,01451	-0,01451
Pressão sanguínea-Max	-0,00053	-0,00053
Glucose-Max	-0,00038	-0,0009

Atributo com 100% de presença

Tabela 13 – Componente 1

Atributo	Escore
Hematócrito-Mín	0,31
Pressão sanguínea-Min	0,30
Escore Glasgow verbal	0,27
Batimento cardíaco-Min	0,25
Taxa Respiratória-Mín	0,25
Escore Glasgow visual	0,25
Hematócrito-Max	0,23
Entubação endotraqueal	0,23
Glucose-Mín	0,22
Ventilação Externa	0,22
Escore Glasgow motor	0,22
Temperatura em célsius-Min	0,19
Pressão arterial de Oxigênio	0,18
Tensão Arterial Alveolar	0,18
Batimento cardíaco-Max	0,16
Permanência em UTI	0,16
Albumina-Min	0,16
Albumina-Max	0,16
Taxa Respiratória-Max	0,15
Nitrogênio Uréico-Mín	0,15
Nitrogênio Uréico-Max	0,13
Creatinina-Max	0,12
Temperatura em célsius-Max	0,11
Sódio-Mín	0,11
Creatinina-Mín	0,10
Falha Renal	0,08
PH	0,07
Cont. glóbulos brancos-Min	0,06
Volume de urina	0,06
Batimento cardíaco-Max	0,04
Pressão arterial de CO2	0,04
Sódio-Max	0,02
Cont. glóbulos brancos-Máx	0,01
Glucose-Max	0,00

Tabela 14 – Componente 2

Atributo	Escore
Nitrogênio Uréico-Max	0,39
Nitrogênio Uréico-Mín	0,38
Creatinina-Mín	0,37
Creatinina-Max	0,36
Pressão arterial de Oxigênio	0,24
Ventilação Externa	0,23
Tensão Arterial Alveolar	0,22
Escore Glasgow verbal	0,20
Permanência em UTI	0,20
Entubação endotraqueal	0,20
Escore Glasgow visual	0,18
Escore Glasgow motor	0,17
Hematócrito-Max	0,16
Temperatura em célsius-Max	0,14
Albumina-Min	0,14
Albumina-Max	0,14
Falha Renal	0,14
Batimento cardíaco-Max	0,08
Volume de urina	0,07
Batimento cardíaco-Max	0,06
Temperatura em célsius-Min	0,06
Sódio-Max	0,06
Batimento cardíaco-Min	0,05
Pressão sanguínea-Min	0,05
Taxa Respiratória-Mín	0,05
Hematócrito-Mín	0,05
Cont. glóbulos brancos-Máx	0,03
Sódio-Mín	0,03
Taxa Respiratória-Max	0,02
PH	0,02
Cont. glóbulos brancos-Min	0,01
Glucose-Mín	0,01
Pressão arterial de CO2	0,00
Glucose-Max	0,00

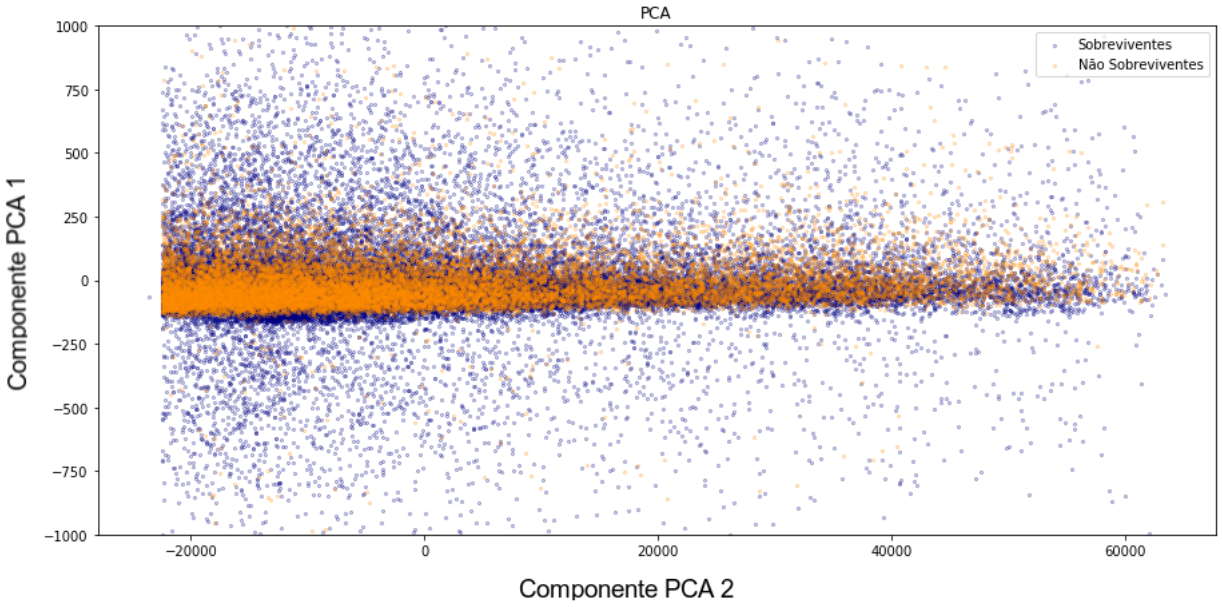

Figura 13 – Base representada pelo Principal Component Analysis - PCA

O t-SNE tem como objetivo encontrar uma representação fiel a partir de um conjunto de pontos em um espaço multidimensional de dimensão menor, frequentemente um plano 2D. O algoritmo utilizado pelo t-SNE é não-linear e se adapta aos dados, realizando diferentes transformações em diferentes regiões do espaço multidimensional. O t-SNE é capaz de capturar muito da estrutura local do espaço multidimensional enquanto também revela a estrutura global do banco de dados como a presença de conjuntos de características similares (MAATEN; HINTON, 2008).

Nas figuras 14 e 15 pode ser observado que os dados não possuem uma separação clara. Mesmo utilizando 50% da base (figura 14) e 10% da base (figura 15), não é possível observar conjuntos de dados bem definidos. O impacto dessa aproximação das classes interfere nos classificadores utilizados.

Ainda buscando identificar como tratar os valores ausentes, compara-se a seguir na tabela 15 os valores de média e desvio padrão de uma base contendo somente registros completos (4763 registros), registros com imputação de valores padrões e registros com imputação pelo MICE (ambos com 56769 registros). Será discutido os resultados obtidos no capítulo 4.

4.1.2 Desbalanceamento de classe

Como visto anteriormente na base de dados existem mais exemplos de pacientes sobreviventes do que pacientes não sobreviventes: 45.318 pacientes sobreviventes e 16.214 não sobreviventes. A classe sobrevivente representa cerca de 74% da base.

Na seção 2.4 foi visto que bases desbalanceadas tendem a enviesar os classificadores para a classe com maior quantidade de exemplos, e que é necessário executar abordagens de

Tabela 15 – Comparação entre base com atributos completos, imputação por valor padrão e imputação pelo MICE.

	Base completa	Valor padrão	MICE
Atributos	Média	Média	Média
Batimento cardíaco-Mín	73,953± 16,560	77,906± 23,352	77,810 ± 23,353
Batimento cardíaco-Máx	107,245± 21,662	108,630 ± 27,069	109,941 ± 26,375
Pressão sanguínea-Mín	57,832±14,628	65,135 ±19,851	58,335 ± 12,569
Pressão sanguínea-Máx	105,846± 27,158	103,829 ± 23,836	104,551 ± 23,800
Temperatura em célsius-Mín	35,997± 0,907	36,265 ±0,787	36,104 ± 0,709
Temperatura em célsius-Máx	37,482±0,858	37,385 ± 0,742	37,468 ± 0,720
Taxa Respiratória-Mín	12,753± 4,036	13,583 ± 4,443	12,327 ± 3,422
Taxa Respiratória-Máx	27,984± 6,917	26,023 ± 6,560	27,208 ± 5,999
Pressão arterial de oxigênio	209,876± 143,217	146,540 ± 113,881	276,217 ± 88,713
Tensão arterial alveolar	-110,143± 145,200	-0,720 ± 142,429	-179,485 ± 88,353
PH	7,355± 0,117	7,380 ± 0,085	7,362 ± 0,0841
Pressão arterial de CO2	40,214± 13,703	44,013 ± 10,461	43,113 ± 10,468
Hematócrito-Mín	28,957± 6,243	32,573 ± 9,166	32,306 ± 9,058
Hematócrito-Máx	35,851± 6,048	37,421 ± 7,519	37,329 ± 7,504
Cont. glóbulos brancos-Mín	10,601± 7,730	11,129 ± 7,804	11,048 ± 7,796
Cont. glóbulos brancos-Máx	15,092± 12,108	13,746 ± 10,372	13,778 ± 10,374
Creatinina-Mín	1,467± 1,499	1,264 ± 1,302	1,315 ± 1,298
Creatinina-Máx	1,852± 1,845	1,464 ± 1,580	1,554 ± 1,570
Nitrogênio Uréico-Mín	27,789± 23,398	22,652 ± 18,164	23,732 ± 18,024
Nitrogênio Uréico-Máx	34,234± 27,064	26,162 ± 20,875	27,869 ± 20,534
Sódio-Mín	135,701± 5,490	137,137 ± 4,754	136,753 ± 4,640
Sódio-Máx	140,369± 5,323	140,047 ± 4,294	140,051 ± 4,297
Albumina-Mín	3,107± 0,733	4,145 ± 0,705	3,118 ± 0,370
Albumina-Máx	3,216± 0,718	4,165 ± 0,674	3,195 ± 0,365
Bilirrubina-Mín	1,713± 4,269	2,043 ± 2,527	2,119 ± 2,548
Bilirrubina-Máx	2,098± 4,846	2,145 ± 2,827	2,388 ± 2,845
Glucose-Mín	101,359± 33,218	115,614 ± 45,109	103,599 ± 31,969
Glucose-Máx	217,872± 140,035	204,557 ± 4197,567	206,055 ± 4200,581
Ventilação extracorpórea*	0,414± 0,492	0,399 ± 0,489	0,399 ± 0,489
Volume de urina	1869,537± 1422,241	1609,555 ± 2765,571	1814,725 ± 2722,948
Escore Glasgow mínima	13,463± 2,855	13,975 ± 2,408	13,770 ± 2,367
Escore Glasgow motor	5,144± 1,614	5,416 ± 1,384	5,296 ± 1,360
Escore Glasgow verbal	3,157± 2,125	3,577 ± 2,053	3,283 ± 1,952
Escore Glasgow visual	3,082± 1,101	3,332 ± 1,015	3,197 ± 0,971
Entubação endotraqueal*	0,228± 0,419	0,189 ± 0,392	0,227 ± 0,383
Falha renal*	0,064± 0,244	0,025 ± 0,157	0,025 ± 0,157
Escore APACHE III	53,816± 22,185	39,430 ± 20,080	39,430 ± 20,082

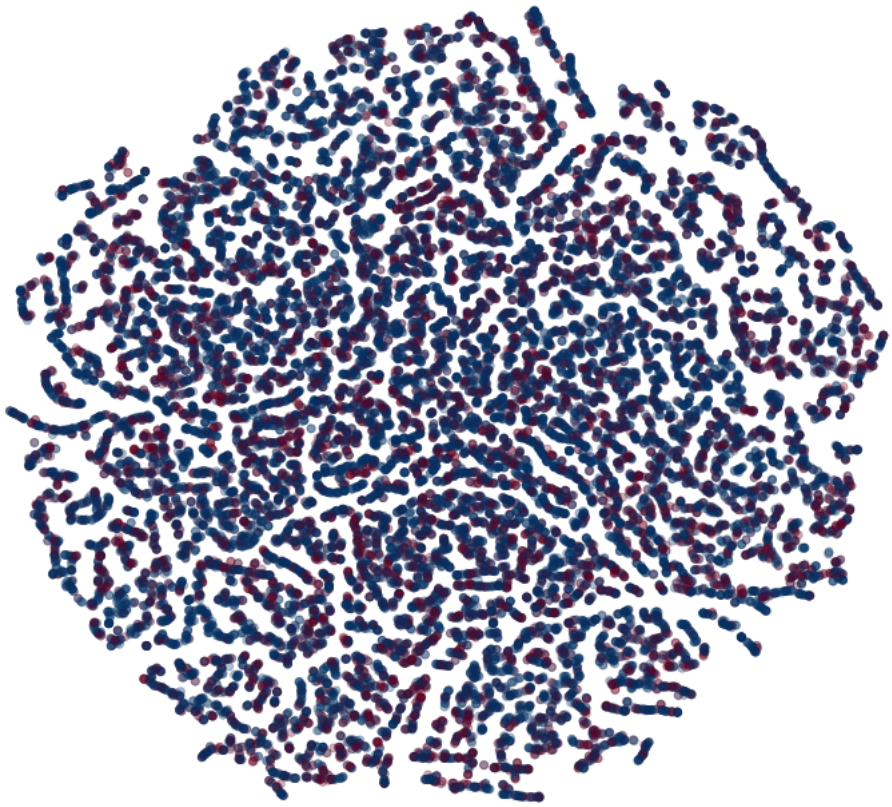

Figura 14 – Representação de 50% da base de dados utilizando o t-SNE

Under sampling e/ou *Over sampling* para treinar os modelos de classificação tentando minimizar esse efeito.

Dessa forma foi realizado o treinamento dos classificadores com duas abordagens de balanceamento de classes: *Over sampling* com SMOTE, que gerou dados sintéticos da classe minoritária até igualar ao valor da classe majoritária e *Under sampling* com NCR, que reduz os exemplos da classe majoritária até o valor da classe minoritária e uma combinação dos dois. Primeiro realizado o *Over sampling* e depois o *Under sampling*, essa estratégia é recomendada para tentar minimizar os efeitos da geração de dados sintéticos, e prevenir uma maior sobreposição de dados.

4.2 Pré-processamento

Nesta seção serão mostrados os resultados do pré-processamento da base de dados: imputação de valores ausentes e balanceamento de classe.

4.2.1 Imputação de dados ausentes

Antes da aplicação dos classificadores foi realizada a substituição dos dados ausentes por valores padrão, substituição pelo MICE e analisado a influência nos parâmetros descritivos

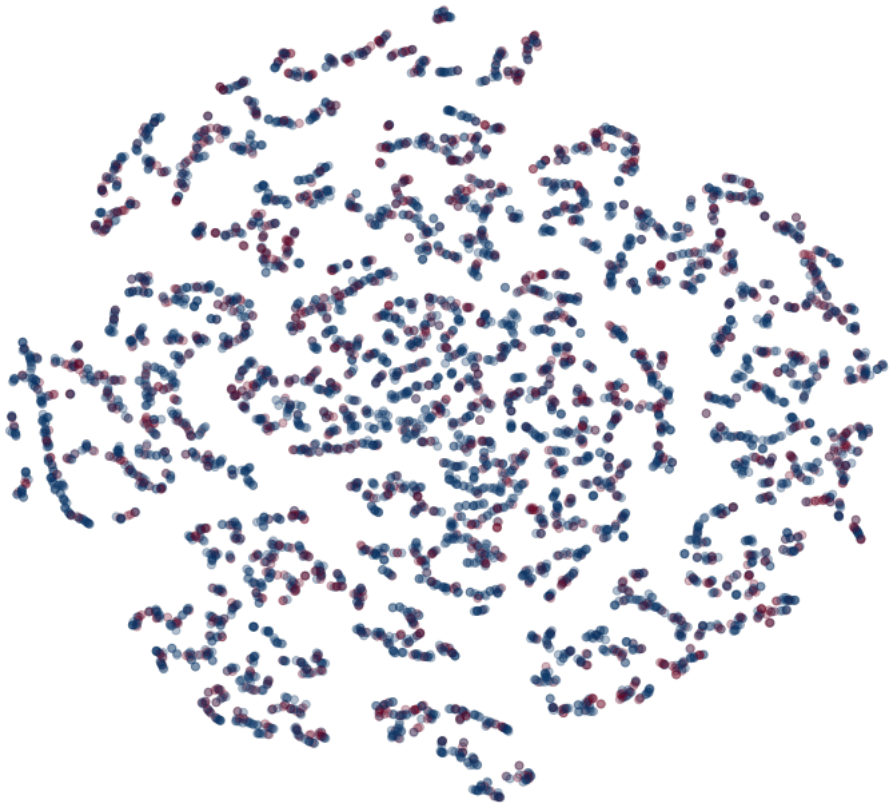

Figura 15 – Representação de 10% da base de dados utilizando o t-SNE

dos atributos (Média, desvio padrão, distribuição), conforme a seguir.

A substituição dos atributos da base com valores ausentes por valores padrões, em comparação com a base de dados com atributos completos, não alterou a apresentação dos mesmos. Nos atributos em que houve um número considerável de substituições, como no caso do atributo Pressão de Oxigênio Arterial que possui 68,47% de ausência, o histograma apresenta um pico de valores correspondente ao valor padrão do mesmo (100), como visto na figura 16. Todavia a distribuição dos valores restantes do atributo continua com curvas semelhantes se comparado com o histograma da base completa na figura 17, evidenciando a proximidade de valores de atributos de pacientes que possuem dados completos com pacientes com dados ausentes.

Com relação ao histograma da base com a substituição dos valores ausentes do atributo Pressão de Oxigênio Arterial pelo MICE, observa-se que a substituição alterou a característica da curva (figura 18), distribuindo os valores ausentes entre 200 e 300, ou seja, os valores foram distribuídos em torno da média do atributo que é de 276,21, após imputação MICE . Na base com valor padrão a maior distribuição de valores fica em torno da média (146,54, tabela 15). O valor da média e desvio padrão (276,21 e 88,91, tabela 15, respectivamente) para esse atributo com os valores imputados pelo MICE mantiveram-se mais próximos dos valores encontrados na base de

dados original (média 263,19 e desvio 160,20, tabela 9) do que os valores substituídos pelo valor padrão que foram 146,42 de média e 113,88 de desvio padrão, conforme mostrado na tabela 9. Entende-se por isso que a imputação pelo MICE aparenta preservar melhor as características da base de dados original do que a substituição por valores padrões.

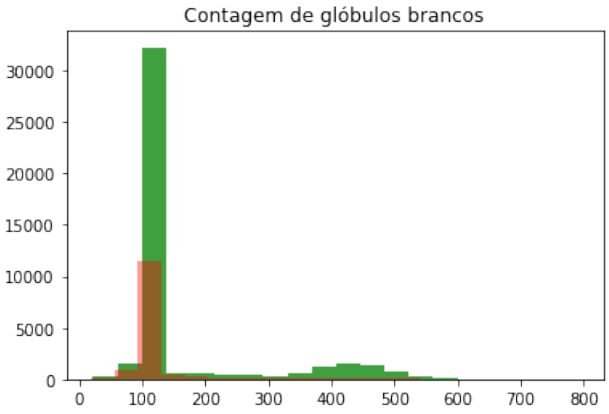

Figura 16 – Histograma do atributo Pressão Arterial de Oxigênio com a base com a imputação do valor padrão

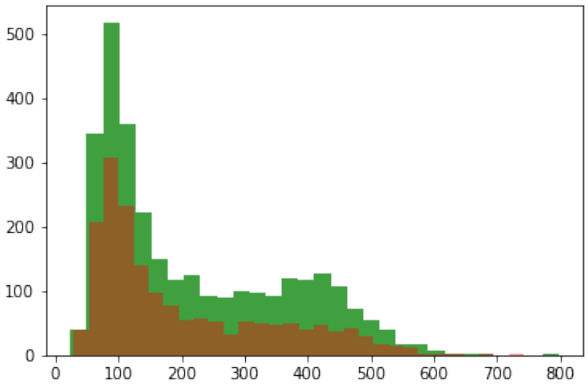

Figura 17 – Histograma do atributo Pressão Arterial de Oxigênio com a base sem valores ausentes

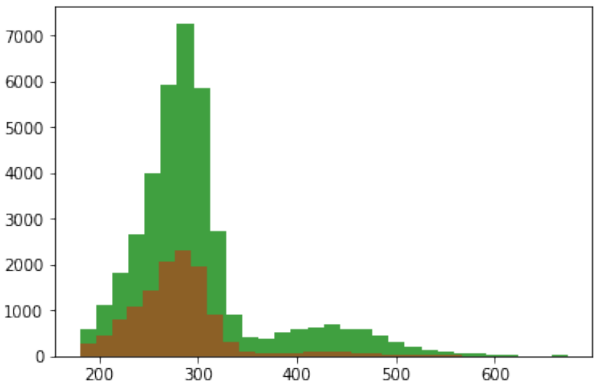

Figura 18 – Histograma do atributo Pressão Arterial de Oxigênio com a base com a imputação pelo MICE

Em alguns casos, observa-se que a imputação através do MICE, manteve os valores de média e desvio padrão próximos a base original, a exemplo de Pressão sanguínea-Min, Albumina-Min e Albumina-Max, conforme mostrado na tabela 15. Em contra partida, houve casos em que a substituição pelo MICE não manteve a proximidade, como Pressão Arterial de Oxigênio, Tensão arterial Alveolar e Hematócrito-Min. Alguns atributos substituídos pelos valores padrões aproximaram a média e desvio padrão à imputação pelo MICE, como nos casos de Batimento cardíaco-Max, Pressão arterial de CO2 e Hematócrito-Min e Hematócrito-Max (tabela 15).

.

4.2.2 Balanceamento de classes

Foram utilizadas as técnicas de balanceamento de classe para *over sampling* SMOTE e para *under sampling* NCR, além de uma combinação das duas técnicas. Essa combinação das técnicas parte da ideia que o *over sampling* da classe minoritária pode criar exemplos de borda ou sobrepostos no espaço da classe dominante e dessa forma seria necessário uma "limpeza"dos dados, que pode ser alcançada utilizando o *under sampling*. Na tabela 16 é mostrada a quantidade de exemplos obtidos pelas três técnicas.

Tabela 16 – Balanceamento da classe de interesse

	Base original completa		Base de treinamento		Base de validação	
	S	NS	S	NS	S	NS
Quantidade original	45.318	16.214	20.393	7.296	20.393	7.296
Under sampling	22.487	16.214	10.822	7.296	20.393	7.296
Over sampling	45.318	45.318	20.393	20.393	20.393	7.296
Over/Under sampling	45.318	40.792	20.393	18.969	20.393	7.296

Para visualização das alterações dos dados a seguir, nas figuras 20, 19 e 21 é mostrado o resultado através do gráfico do PCA das bases. Como pode ser observado, não há uma melhora visível na separação dos espaços das regiões das classes sobreviventes e não sobreviventes, o que evidencia a complexidade do conjunto de dados.

4.3 Classificação

Nessa seção serão apresentados os valores das medidas de avaliação obtidos pelo APACHE e pelos classificadores, RF, RNA e RL.

4.3.1 APACHE III

O APACHE III obteve os valores médios de AUC = 0,660 com desvio padrão de 0,0006, sensibilidade = 0,531 $\pm$ 0,001 e Especificidade de 0,788 $\pm$ 0,0001, aplicado a base utilizada. Na tabela 17, é mostrada a matriz de confusão referente a aplicação do APACHE III na base do MIMIC-III. O APACHE-III obteve uma especificidade alta (0,788) e sensibilidade baixa (0,531)

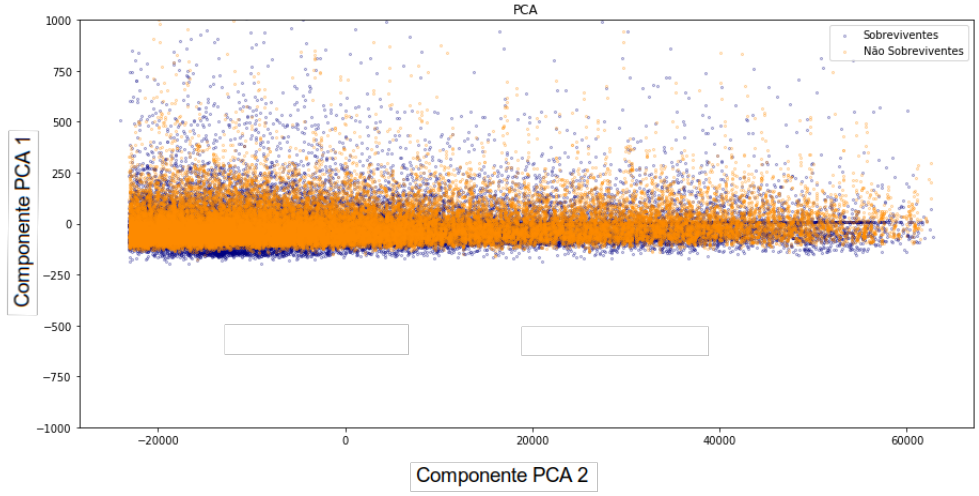

Figura 19 – Projeção da base com *under sampling* através do PCA.

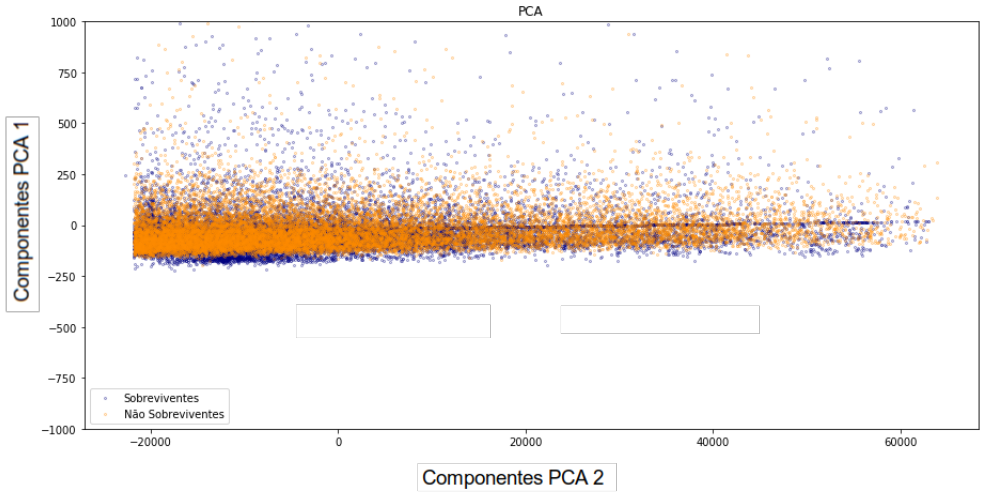

Figura 20 – Projeção da base com *over sampling* através do PCA.

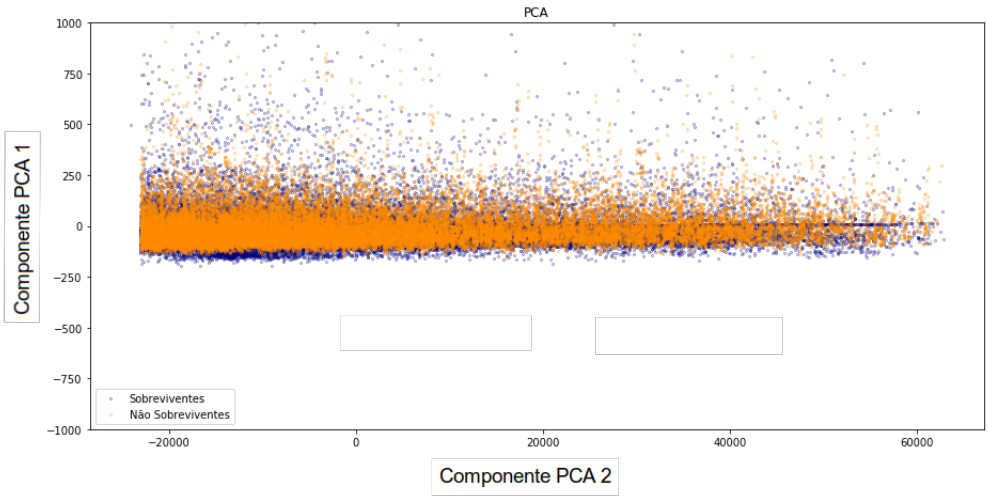

Figura 21 – Projeção da base com *over/under sampling* através do PCA.

com o limiar da fórmula de Johnson (2014) definido para 0,2, ou seja, se aplicada a fórmula mostrada em 7 valores maiores que 0,2 indicaria que o paciente seria não sobrevivente. Outros limiares foram testados, sendo que o limiar de 0,2 obteve melhores valores em relação aos outros (aplicado teste estatístico de Wilcoxon com correção de Bonferroni, α=0,05 sendo o p-value= 0,0092).

A sensibilidade representa a classe de interesse, que neste caso não conseguiu bons resultados, tendo dificuldade de identificar os não sobreviventes, seria necessária uma revisão da fórmula ou alterações constantes do limiar para definir valores personalizados para cada conjunto de dados.

O APACHE é calculado levando em conta a soma dos escores obtidos pelos atributos do paciente, sendo utilizado o valor de escore 0 para os atributos ausentes, e assim não sofre interferência com o desbalanceamento de classe. Os valores dos escores são baseados no conhecimento dos especialistas que montaram o mesmo, e necessita de ajustes para que seja adaptado em cada região que será utilizado (KLUNDERT et al., 2015).

Tabela 17 – Matriz de confusão APACHE III aplicado a tabela do MIMIC-III

		Preditos	
Reais		S	NS
	S	17.879	4.780
	NS	3.789	4.318
S - Sobreviventes			
NS - Não Sobreviventes			

4.3.2 Resultados dos classificadores por estratégia de substituição de valores ausentes

Foram aplicados os classificadores na base de dados com substituição de valores por valores padrões, substituição de valores por valores padrões com inclusão de atributo indicando ausência ou presença do mesmo e substituição dos valores pelo MICE. Foram obtidos os valores médios de AUC, sensibilidade e especificidade, descritos na tabela 18.

Tabela 18 – Resultados com as bases com valores padrões, padrões com flags de ausência e substituição de valores pelo MICE.

	Base com valores padrões.			Base com valor padrões mais *flag* de presença/ausência		
	AUC	Sens	Esp	AUC	Sens	Esp
RL	0,723±0,009	0,786±0,013	0,525±0,013	0,717±0,009	0,787±0,021	0,506±0,005
RNA	0,772±0,057	0,680±0,058	0,712±0,043	0,765±0,041	0,658±0,057	0,723±0,045
RF	0,784±0,006	0,719±0,014	0,703±0,009	0,776±0,004	0,710±0,017	0,697±0,004

	Base com valor imputado pelo MICE		
	AUC	Sens	Esp
RL	0,707±0,005	0,778±0,010	0,487±0,010
RNA	0,766±0,046	0,692±0,061	0,694±0,055
RF	0,781±0,005	0,706±0,006	0,705±0,005

A RL obteve maior AUC sendo aplicada a base com substituição por valor padrão (0,723 $\pm$0,009) ($p = 0,006$), mas não manteve uma relação equilibrada entre sensibilidade e especificidade (0,786 e 0,525, respectivamente). Esse desequilíbrio entre esses indicadores pode prejudicar a avaliação do classificador, pois embora ele seja capaz de identificar de forma correta a maioria dos casos de interesse, ele não identifica corretamente casos que não são de interesse, podendo gerar intervenções e atenção em pacientes que efetivamente não precisariam, gerando custos extras desnecessários. A RL manteve esse comportamento também nas bases com imputação pelo MICE e com flag de ausência.

A RNA, como a RL, também obteve seu maior valor de AUC na base de valor padrão ($p = 0,005$), e obteve valores mais equilibrados entre a sensibilidade e especificidade, porém demonstrou através de seu desvio padrão alto, uma variação alta da classificação, o que torna seu resultado não confiável. Assim como a RL manteve esse comportamento nas bases com imputação pelo MICE e com flag de ausência.

O valor de AUC da RF aplicado na base com valor padrão é significativamente maior que o valor da RF aplicado nas outras bases (Wilcoxon com correção de Bonferroni, α=0,05, sendo obtido o p-value de 0,0035). Diferente da RL, manteve equilibrado seus valores de sensibilidade e especificidade, o que dá consistência ao modelo, e com o desvio padrão baixo dos indicadores, possui maior confiabilidade pela possível menor variação de resultados. Seguindo o comportamento dos outros classificadores a RF obteve desempenho similar nas bases com imputação pelo MICE e com flag de ausência/presença.

Os valores de AUC obtidos pelos classificadores foram mais altos do que o obtido pelo APACHE-III ($p = 0,004$), demonstrando um melhor desempenho na classificação de mortalidade que o escore já em uso nos hospitais, o que só isso já representaria um ganho de informação para a equipe médica.

A substituição dos valores ausentes pelo MICE não fez com que os classificadores melhorassem seu desempenho. O modelo de imputação pode ser melhorado ou novas técnicas serem aplicadas, mas para esse estudo não houve melhora significativa.

Embora representasse um ganho de informação, a flag de ausência do atributo não auxiliou os classificadores a reconhecer melhor algum padrão para identificar corretamente a classe alvo.

Dessa forma, entende-se, nesse cenário, que a imputação de valor padrão possibilitou uma melhor predição por parte dos classificadores. A RNA e o RL também obtiveram valores de AUC maiores nesse modelo de imputação. A RNA apresentou um alto desvio padrão em todas as métricas, indicando que não conseguiu uma boa adaptação aos dados como os outros classificadores, que mantiveram o desvio baixo (todas as comparações foram realizadas com base no teste estatístico de Wilcoxon com correção de Bonferroni, α=0,05 sendo o p-value= 0,0083).

A justificativa para esses melhores resultados dos classificadores na base de dados com valor padrão pode ser baseada no tipo de dados ausentes que a base possui que seriam dados ausentes que tem relação indireta com os dados observados. Knaus et al. (1981) considerou em seu estudo que os dados ausentes dos pacientes se deviam, pelo quadro geral do paciente não exigir aferições e testes de determinados atributos, ou que resultados específicos de exames apontavam para outra linha de investigação que não dependeriam dos valores dos atributos que não foram registrados. Isso demonstraria que o dado ausente representaria o estado normal daquele atributo e sua substituição por um valor que representasse essa normalidade seria correto.

Por essa razão, os classificadores conseguiram se adaptar a essa base de dados com esse tipo de substituição, a inclusão da flag de ausência pode ter gerado ruído que degradou a classificação e o MICE se baseia em que os valores dos atributos ausentes são diretamente relacionados com os valores dos dados existentes.

4.3.3 Resultados dos classificadores por estratégia de balanceamento de classe

Após identificar o modelo de substituição de valores ausentes, foi aplicado os classificadores com diferentes estratégias de balanceamento de classe, *over sampling*, *under sampling* e os dois métodos combinados. Ressalta-se que o balanceamento de classe só é aplicado no conjunto de treinamento do *Cross-validation*. A tabela 19 mostra os valores médios de AUC, sensibilidade e especificidade encontrados.

Tabela 19 – Resultados dos classificadores por estratégia de balanceamento de classe.

	Over sampling			*Under sampling*		
	AUC	Sens	Esp	AUC	Sens	Esp
RL	0,727±0,008	0,752±0,014	0,567±0,009	0,723±0,009	0,786±0,012	0,525±0,013
RNA	0,780±0,058	0,732±0,054	0,676±0,045	0,776±0,056	0,680±0,057	0,712±0,044
RF	0,783±0,006	0,553±0,014	0,810±0,008	0,784±0,006	0,718±0,014	0,703±0,009

	Over/Under sampling		
	AUC	Sens	Esp
RL	0,725±0,009	0,750±0,013	0,564±0,009
RNA	0,773±0,006	0,680±0,058	0,712±0,044
RF	0,780±0,006	0,724±0,077	0,682±0,072

A exemplo da estratégia de imputação de valores ausentes, os valores de AUC obtidos pelos classificadores foram mais altos do que o obtido pelo APACHE-III ($p = 0,004$), demonstrando um melhor desempenho na classificação de mortalidade que o escore já em uso nos hospitais.

A RL obteve maior valor de AUC na base com *Over sampling*, mas continuou com o desequilíbrio entre sensibilidade e especificidade em todos os cenários, sendo obtidos valores mais altos para especificidade.

A RNA também obteve maiores valores de AUC e sensibilidade na base com *Over sampling*, e valor mais equilibrado de especificidade, mas o desvio padrão alto indica instabilidade

do modelo, o que não garante confiabilidade nos resultados.

A RF obteve maiores valores de AUC nos três cenários de balanceamento de classe, mas no *Over sampling* teve uma queda de desempenho para classificar a classe alvo, como mostrado pelo seu valor de sensibilidade menor que nos outros cenários de balanceamento (0,553±0,014).

O *Over sampling* cria exemplos não reais baseados na classe minoritária. Os exemplos nos cenários de *Over sampling Over/Under sampling* criados, foram capazes de melhorar a classificação da RNA e da RL, o que demonstra uma capacidade melhor de adaptação nesse cenário. Mesmo com essa melhora, a RL ainda continuou com um desequilíbrio entre a sensibilidade e especificidade, e esse comportamento não garante a confiabilidade do modelo. A RNA embora tenha tido equilíbrio nos valores de sensibilidade e especificidade, teve uma variação muito alta dos resultados, demonstrada pelos desvios padrões mostrados na tabela 19.

O *Under sampling* reduz a quantidade de exemplos da classe majoritária passando para o classificador quantidades similares de exemplos de cada classe, de forma a evitar que haja uma tendência de classificação para classe com maior número de exemplos. A RF obteve resultados melhores nesse cenário evidenciado pelo valor de AUC, sensibilidade e especificidade obtidos (0,783±0,006). Obteve também valores equilibrados de sensibilidade e especificidade (0,718±0,014, 0,703±0,009) e baixo desvio padrão, demonstrando confiabilidade e consistência nos resultados do modelo.

4.4 Teste final

Para teste do modelo descrito, os classificadores treinados foram aplicados nos dados de teste separados no início, de modo que o classificador seja utilizado em dados que não foram usados em nenhum momento do treinamento e da validação. Após aplicação dos classificadores, foram obtidos os dados mostrados na tabela 20. Pela comparação de valores obtidos com as médias anteriores da RF, o modelo apresentou-se estável e capaz de reproduzir sua performance em dados inéditos a ele. Na figura 22 é mostrada a curva ROC resultante dos classificadores.

Tabela 20 – Resultado em base de testes

	AUC	Sens	Esp
RL	0,7076±0,005	0,7731±0,006	0,5116±0,005
RNA	0,7647±0,015	0,6198±0,045	0,7557±0,056
RF	0,7806±0,006	0,7029±0,006	0,7116±0,008

Foi buscado identificar dentro dos quadrantes da matriz de confusão (21), padrões que sugerissem onde o classificador pudesse estar errando, como quantidade de atributos ausentes por quadrante, mas por esse atributo não foi possível estabelecer relação dos erros com particularidades da ausência. Nas figuras 23 e 24 pode ser visto o histograma de ausências dos quadrantes de erro (Não sobreviventes informados como sobreviventes e Sobreviventes informados como não sobreviventes).

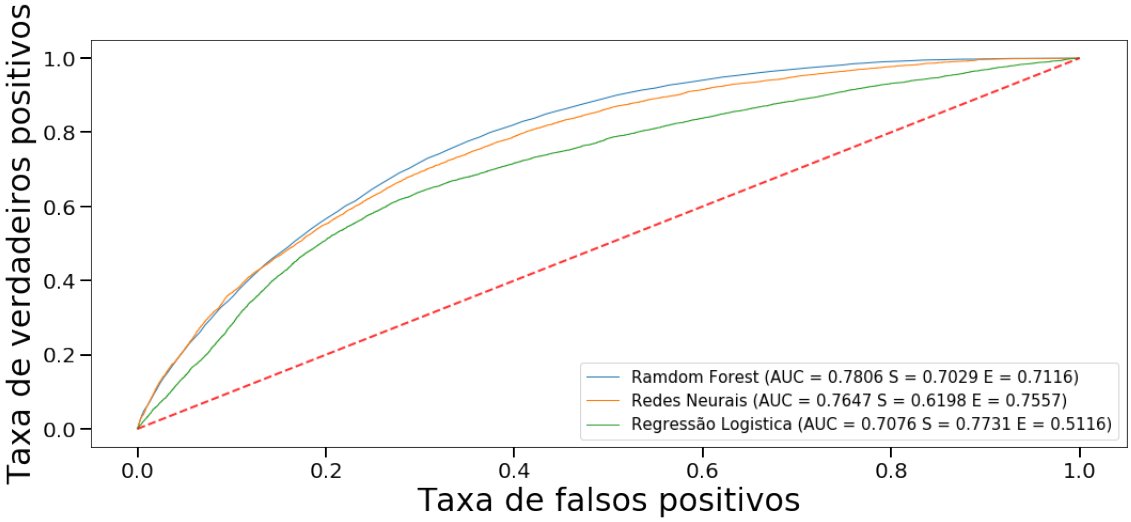

Figura 22 – Curva ROC da aplicação dos classificadores em base de testes

Tabela 21 – Matriz de confusão

Preditos	Real		
		S	NS
	S	14.488	5.905
	NS	2.161	5.193
S - Sobreviventes			
NS - Não Sobreviventes			

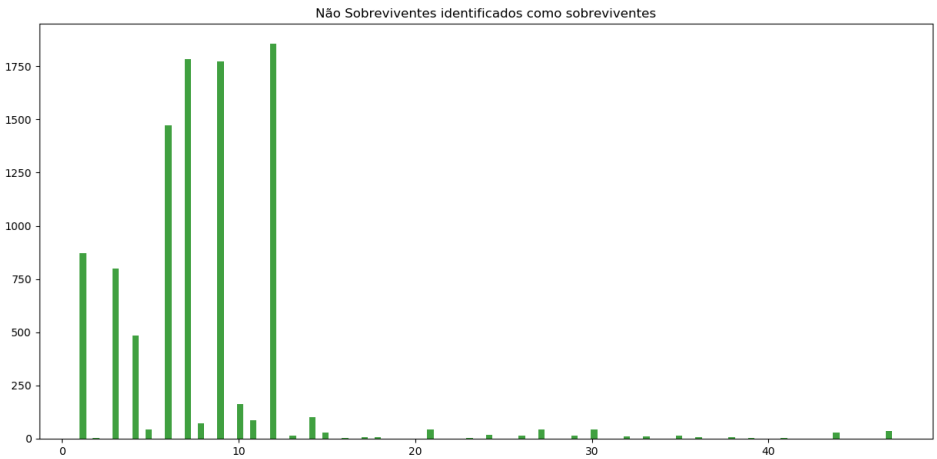

Figura 23 – Histograma quadrante de erro de identificação de Não sobreviventes como Sobreviventes

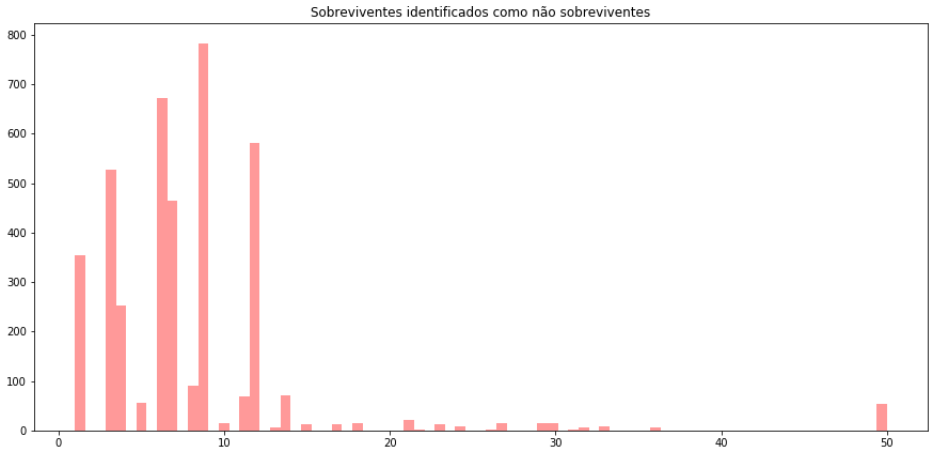

Figura 24 – Histograma quadrante de erro de identificação de Sobreviventes como Não Sobreviventes

4.4.1 Seleção de atributos

Na figura 25, é mostrado o ranqueamento dos atributos usados pela RF, baseados no índice composto com GINI e quantidade de árvores onde o atributo é utilizado. Este valor é calculado automaticamente para cada atributo após a etapa de treinamento da RF e pode-se ver que segundo esse índice, os melhores preditores, retirando o próprio valor obtido pelo escore de predição do APACHE (apsiii), seriam o nitrogênio no sangue (bun_min e max), urina expelida nas últimas 24 horas (urineoutput) e tempo de permanência na UTI (permanencia_uti). Já era esperado a posição do escore do APACHE como melhor preditor já que ele é a consolidação dos outros atributos e já usado para a predição de mortalidade nas UTI's. Com a análise da importância dos atributos podemos tentar identificar um conjunto menor de atributos mais relevantes, que possam simplificar o modelo, mantendo ou melhorando o desempenho da predição.

Comparando o *ranking* de importância dos atributos da figura 25 com a tabela 12, podemos ver que os atributos Nitrogênio Uréico-Mín e Nitrogênio Uréico-Max (bun_min e bun_max, respectivamente) aparecem melhor posicionados em ambas as análises, evidenciando que seriam bons preditores para o modelo. O Volume de urina e o tempo de Permanência na UTI (urineoutput e permanencia_uti, respectivamente), embora bem ranqueados pela *Random Forest*, não possuem uma correlação que coloque eles melhores posicionados na tabela 12.

Devido as diferenças identificadas entre a correlação dos atributos com a classe alvo com o ranqueamento da *Random Forest*, foram realizados testes com conjuntos de atributos menores (20, 15, 10 e 5) selecionados pelo grau de importância e pela correlação, e comparados nas tabelas 22 e 23, buscando identificar qual parâmetro produz melhores resultados para predição de mortalidade.

Conforme visto nas tabelas 22 e 23, a diminuição dos números de atributos reduziu

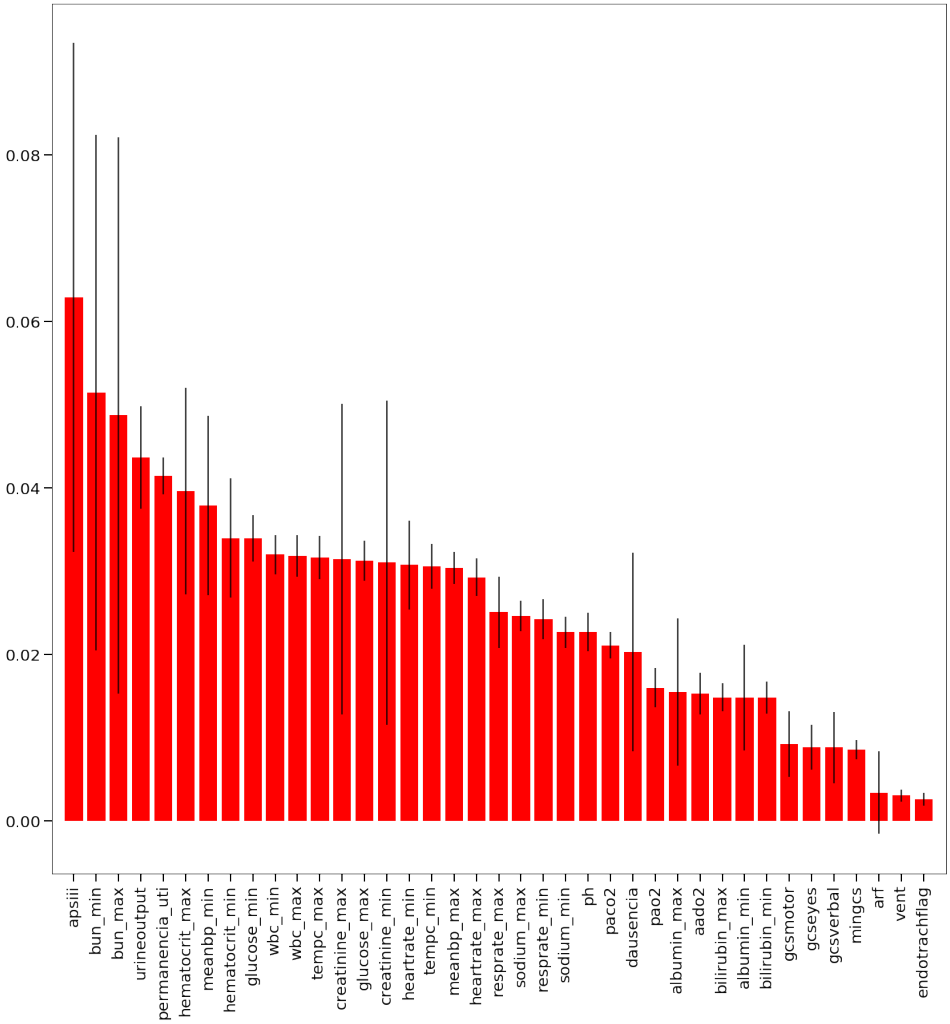

Figura 25 – Ranqueamento de atributos por importância pela RF

Tabela 22 – Resultados da *Random Forest* com balanceamento por *Under Sampling* com 20, 15, 10 e 5 atributos, selecionados pelo ranking de atributos.

	AUC	Sens	Esp
20 atributos	0,772±0,002	0,711±0,001	0,690±0,004
15 atributos	0,762±0,002	0,690±0,005	0,687±0,006
10 atributos	0,747±0,002	0,712±0,012	0,658±0,005
5 atributos	0,725±0,001	0,689±0,012	0,651±0,008

Tabela 23 – Resultados da *Random Forest* com balanceamento por *Under Sampling* com 20, 15, 10 e 5 atributos, selecionados pelo coeficiente de correlação ponto-bisserial.

	AUC	Sens	Esp
20 atributos	0,760±0,002	0,708±0,002	0,681±0,003
15 atributos	0,739±0,002	0,684±0,007	0,671±0,001
10 atributos	0,734±0,001	0,701±0,004	0,643±0,003
5 atributos	0,688±0,001	0,529±0,017	0,631±0,011

também o valor de AUC e especificidade da RF, porém isso não refletiu diretamente nos valores de sensibilidade embora não tenha ganho nenhum aumento representativo, O modelo ainda manteve equilíbrio entre sensibilidade e especificidade e baixo desvio padrão. Existe através desse resultado indicação de melhorias na seleção de atributos que possam reduzir o número de atributos coletados sem degradar o escore de severidade, mantendo a predição de mortalidade confiável.

4.5 Discussão

Dentre as técnicas de imputação de dados avaliadas, observou-se um melhor desempenho dos classificadores sobre a base com valor padrão. Dessa forma, foi entendido que o classificador consegue observar o valor padrão realmente como normalidade, pois se houvesse um aumento expressivo da sensibilidade ou da especificidade ver-se-ia a tendência ser alterada na base com essa imputação. A classificação em cima da base de dados com flag de ausência, gerou mais ruído degradando o desempenho dos classificadores, e o MICE necessita que os valores dos atributos ausentes sejam relacionados diretamente com os valores dos atributos existentes, o que não foi observado na base de dados utilizada.

O balanceamento de classe por *under sampling*, ou seja, por subamostragem da classe majoritária, também melhorou o desempenho do classificador. O balanceamento por *Over sampling*, cria exemplos não reais baseados nos exemplos existentes da classe minoritária, mas não garante que esses novos exemplos não estejam muito próximos ou similares também à classe majoritária, e por consequência, não melhora as métricas utilizadas. Isso é evidenciado por ser uma base com atributos muito próximos e de classes diferentes, como mostrado nas projeções dos gráficos de PCA (figuras 19, 20, 21) e t-SNE (figuras 15 e 14).

O APACHE-III demonstrou uma baixa especificidade, menor inclusive do que a RL. Considerando que o APACHE-III se baseia somente no valor do escore para cálculo da probabilidade enquanto as técnicas utilizadas fazem uso de todos os atributos do paciente para buscar identificar a classe do paciente, os modelos se mostraram mais capazes de adaptação. A AUC obtido pela RF (0,7806) foi significativamente maior que o AUC obtido pelo APACHE-III (0,660) (Wilcoxon com correção de Bonferroni, α=0,05, p-value = 0,0036).

A RL obteve bons resultados de AUC em todas as bases (0,786 para valores padrão, 0,787 padrão mais *flag* e 0,778 MICE), e nos fornece indícios de ser um bom classificador para a classe de interesse. Porém, sua especificidade, que identifica a outra classe, não obteve bons valores (0,525 para valores padrões, 0,506 padrão mais *flag* e 0,487 MICE), mostrando fragilidades do modelo.

A RNA teve seus valores medianos (média de AUC de 0,771, sensibilidade de 0,676 e especificidade de 0,709) mas também teve os maiores desvios padrões (tabela 18), demonstrando uma instabilidade maior que os outros classificadores.

Foi percebido, pelos resultados obtidos que que o modelo linear de classificação utilizado, a RL, e modelo não-linear, a RNA, tiveram limitações na predição da mortalidade. O modelo de classificação, RF, que por característica espera-se um desempenho melhor que modelos lineares quando há uma relação de alta não linearidade, apresentou menos limitação.

Avaliando os resultados obtidos em geral, a RF demonstrou melhor adaptação aos dados apresentados, como visto através dos melhores valores de AUC (α=0.05, p-value = 0.0053), baixo desvio padrão e valores de sensibilidade e especificidade equilibrados, isso devido a não-linearidade dos parâmetros, homogeneidade da variância e independência entre os atributos preditores. Assim, considera-se que, sendo aplicada a base com valores padrão e balanceamento de classe por *under sampling*, a *Random Forest*, nestas condições, obteve melhor desempenho se comparado com os demais modelos ou procedimentos de pré-processamento (Wilcoxon com correção de Bonferroni, α=0.05, p-value = 0.0053).

5 CONCLUSÃO

O presente trabalho apresentou como objetivo propor a aplicação de métodos de mineração de dados, no pré-processamento e na classificação utilizando Redes Neurais Artificiais - RNA, *Random Forest* - RF e Regressão Logística, para predição de mortalidade de pacientes em UTI a partir dos atributos definidos pelo APACHE III. Os resultados obtidos da AUC, sensibilidade e especificidade de cada classificador foram analisados em busca de identificar qual obteria melhor desempenho. Foi avaliado em conjunto, a influência dos dados ausentes e as técnicas de substituição desses valores, além de técnicas para tratamento de desbalanceamento de classe, de forma a permitir que os classificadores sejam treinados com dados que não enviesem seu resultado para a classe que possui mais exemplos.

Com base nos resultados é possível concluir que os modelos de classificação, em linhas gerais, são ferramentas capazes de encontrar padrões em grandes volumes de dados de UTI's, que podem identificar melhor a condição de determinado paciente e possibilitar a tomada de decisão mais embasada pela equipe médica. Pela característica de não-linearidade do conjunto de dados utilizados, a *Random Forest* obteve desempenho melhor que dos demais, baseado nos valores de AUC, sensibilidade e especificidade obtidos. Pôde ser observado também que a técnica que imputa valores padrões em substituição de valores ausentes, incrementou o desempenho do classificador, juntamente com o balanceamento de classe usando *under sampling*, resultado este, não esperado pelo que foi visto na seção 4.2, pois o MICE tinha mantido uma média e desvio padrão mais próximos à base original, do que à base imputada por valores padrões.

O escore do APACHE-III, obteve menor desempenho comparado a RF com imputação de valores padrões em substituição de dados ausentes e balanceamento com *under sampling*. Embora seja menos suscetível a desbalanceamento de classe e dados ausentes, devido a utilização do escore aplicado a equação 7, tem menos poder de adaptação aos dados. A RF mostrou ter mais poder de predição e adaptação aos dados, já que utiliza todos os atributos para classificação, podendo se tornar uma ferramenta melhor para identificar o estado de severidade de pacientes e auxiliar a equipe médica na tomada de decisão. O uso de menos atributos para RF se mostrou promissora, mas ainda carece de maior análise das estratégias para seleção de atributos. Dessa forma, será possível que a equipe médica com menos informação possa ter um escore de severidade confiável.

A metodologia utilizada neste trabalho, levando em conta o estudo das técnicas de balanceamento de classe e tratamento de dados ausentes, produziu melhores resultados que a utilização das técnicas de forma isolada, ou seja, só aplicar tratamento de dados ausentes ou somente aplicar técnicas para desbalanceamento de classes. Além disso, pode ser replicada por qualquer hospital interessado, de forma a ter um modelo de predição de mortalidade adaptado ao

seu contexto e características regionais.

Ainda existem limitações das técnicas utilizadas neste trabalho, a realidade das UTI's de cada região afeta a performance de escores padrões de gravidade (KLUNDERT et al., 2015), e portanto, novos testes utilizando bases de dados de UTI's de diferentes regiões podem confirmar se o mesmo ocorre para as técnicas de mineração de dados. Para trabalhos futuros é sugerido o uso de dados de hospitais brasileiros, além de aplicações de outras técnicas mais atuais e que mostram bons resultados em outras áreas, tais como, técnicas de *Deep Learning*.

REFERÊNCIAS

ABDI, H.; WILLIAMS, L. J. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, Wiley Online Library, v. 2, n. 4, p. 433–459, 2010. Citado na página 29.

ALTINI, M. *Dealing with imbalanced data: undersampling, oversampling and proper cross-validation*. 2016. Citado na página 44.

ALVES, C. et al. Comparação entre o modelo unicamp ii e o apache ii em uma uti geral. *RBTI*, v. 15, n. 4, p. 144–52, 2003. Citado 2 vezes nas páginas 13 e 35.

ARTS, D. et al. Quality of data collected for severity of illness scores in the dutch national intensive care evaluation (nice) registry. *Intensive care medicine*, Springer, v. 28, n. 5, p. 656–659, 2002. Citado na página 13.

AWAD, A. et al. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *International journal of medical informatics*, Elsevier, v. 108, p. 185–195, 2017. Citado 3 vezes nas páginas 36, 37 e 38.

AZUR, M. J. et al. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, Wiley Online Library, v. 20, n. 1, p. 40–49, 2011. Citado na página 20.

BATISTA, G. E.; BAZZAN, A. L.; MONARD, M. C. Balancing training data for automated annotation of keywords: a case study. In: *WOB*. [S.l.: s.n.], 2003. p. 10–18. Citado na página 22.

BATISTA, G. E.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter*, ACM, v. 6, n. 1, p. 20–29, 2004. Citado na página 13.

BENGIO, Y.; GRANDVALET, Y. No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, v. 5, n. Sep, p. 1089–1105, 2004. Citado na página 30.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, v. 13, n. Feb, p. 281–305, 2012. Citado na página 28.

BRADLEY, A. P. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, Elsevier, v. 30, n. 7, p. 1145–1159, 1997. Citado na página 32.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado 2 vezes nas páginas 25 e 26.

BREIMAN, L. Consistency for a simple model of random forests. Citeseer, 2004. Citado na página 25.

CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002. Citado na página 22.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, v. 7, n. Jan, p. 1–30, 2006. Citado 2 vezes nas páginas 33 e 34.

EGAN, J. P. Signal detection theory and {ROC} analysis. Academic Press, 1975. Citado na página 32.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, ACM, v. 39, n. 11, p. 27–34, 1996. Citado 2 vezes nas páginas 18 e 19.

FIALHO, A. S. et al. Data mining using clinical physiology at discharge to predict icu readmissions. *Expert Systems with Applications*, Elsevier, v. 39, n. 18, p. 13158–13165, 2012. Citado na página 37.

FIGUEIRA, C. V. Modelos de regressão logística. 2006. Citado na página 27.

FUJII, T. et al. Diagnosis, management, and prognosis of patients with acute kidney injury in japanese intensive care units: The jakid study. *Journal of critical care*, Elsevier, 2018. Citado na página 13.

GALL, J.-R. L.; LEMESHOW, S.; SAULNIER, F. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, American Medical Association, v. 270, n. 24, p. 2957–2963, 1993. Citado 2 vezes nas páginas 12 e 15.

GHOSE, S. et al. An improved patient-specific mortality risk prediction in icu in a random forest classification framework. In: IOS PRESS. *Driving Reform: Digital Health is Everyone's Business: Selected Papers from the 23rd Australian National Health Informatics Conference (HIC 2015)*. [S.l.], 2015. p. 56–61. Citado 2 vezes nas páginas 36 e 38.

GOODFELLOW, I. et al. Generative adversarial nets. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2014. p. 2672–2680. Citado na página 25.

GRNAROVA, P. et al. Neural document embeddings for intensive care patient mortality prediction. *arXiv preprint arXiv:1612.00467*, 2016. Citado 2 vezes nas páginas 37 e 38.

HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.]: Elsevier, 2011. Citado 4 vezes nas páginas 13, 18, 31 e 32.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, v. 143, n. 1, p. 29–36, 1982. Citado 2 vezes nas páginas 32 e 33.

HAYKIN, S. S. et al. *Neural networks and learning machines*. [S.l.]: Pearson Upper Saddle River, NJ, USA:, 2009. v. 3. Citado 2 vezes nas páginas 23 e 25.

HE, H. et al. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: IEEE. *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*. [S.l.], 2008. p. 1322–1328. Citado na página 22.

HE, H.; GARCIA, E. A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, IEEE, v. 21, n. 9, p. 1263–1284, 2009. Citado na página 21.

HOCHBERG, Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, Oxford University Press, v. 75, n. 4, p. 800–802, 1988. Citado na página 34.

JAGANNATH, V. *Random Forest Template for TIBCO Spotfire®–Wiki page*. 2017. Citado 2 vezes nas páginas 7 e 26.

JOHNSON, A. E. *Mortality prediction and acuity assessment in critical care*. Tese (Doutorado) — Department of Engineering Science, University of Oxford, 2014. Citado 3 vezes nas páginas 13, 43 e 61.

JOHNSON, A. E. et al. Mimic-iii, a freely accessible critical care database. *Scientific data*, Nature Publishing Group, v. 3, 2016. Citado 3 vezes nas páginas 7, 39 e 40.

KEEGAN, M. T.; GAJIC, O.; AFESSA, B. Comparison of apache iii, apache iv, saps 3, and mpm0iii and influence of resuscitation status on model performance. *CHEST Journal*, American College of Chest Physicians, v. 142, n. 4, p. 851–858, 2012. Citado na página 18.

KIM, S.; KIM, W.; PARK, R. W. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. *Healthcare informatics research*, v. 17, n. 4, p. 232–243, 2011. Citado 4 vezes nas páginas 12, 13, 35 e 38.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. Citado na página 43.

KLUNDERT, N. van de et al. Data resource profile: the dutch national intensive care evaluation (nice) registry of admissions to adult intensive care units. *International journal of epidemiology*, IEA, p. dyv291, 2015. Citado 3 vezes nas páginas 12, 61 e 71.

KNAUS, W. A. et al. Apache ii: a severity of disease classification system. *Critical care medicine*, LWW, v. 13, n. 10, p. 818–829, 1985. Citado 5 vezes nas páginas 8, 12, 16, 17 e 45.

KNAUS, W. A. et al. The apache iii prognostic system: risk prediction of hospital mortality for critically iii hospitalized adults. *Chest*, Elsevier, v. 100, n. 6, p. 1619–1636, 1991. Citado 3 vezes nas páginas 8, 17 e 18.

KNAUS, W. A. et al. Apache-acute physiology and chronic health evaluation: a physiologically based classification system. *Critical care medicine*, LWW, v. 9, n. 8, p. 591–597, 1981. Citado 8 vezes nas páginas 8, 12, 15, 16, 17, 41, 50 e 63.

KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: STANFORD, CA. *Ijcai*. [S.l.], 1995. v. 14, n. 2, p. 1137–1145. Citado na página 29.

KOHAVI, R.; PROVOST, F. Glossary of terms. *Machine Learning*, v. 30, n. 2-3, p. 271–274, 1998. Citado 2 vezes nas páginas 31 e 32.

KUBAT, M.; MATWIN, S. et al. Addressing the curse of imbalanced training sets: one-sided selection. In: NASHVILLE, USA. *Icml*. [S.l.], 1997. v. 97, p. 179–186. Citado na página 22.

KUZNIEWICZ, M. W. et al. Variation in icu risk-adjusted mortality: impact of methods of assessment and potential confounders. *CHEST Journal*, American College of Chest Physicians, v. 133, n. 6, p. 1319–1327, 2008. Citado na página 12.

LAURIKKALA, J. Improving identification of difficult small classes by balancing class distribution. In: SPRINGER. *Conference on Artificial Intelligence in Medicine in Europe*. [S.l.], 2001. p. 63–66. Citado na página 22.

LINACRE, J.; RASCH, G. The expected value of a point-biserial (or similar) correlation. *Rasch Meas Trans*, v. 22, n. 1, p. 1154–1157, 2008. Citado na página 50.

LIU, X.-Y.; WU, J.; ZHOU, Z.-H. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, IEEE, v. 39, n. 2, p. 539–550, 2009. Citado na página 21.

MAATEN, L. v. d.; HINTON, G. Visualizing data using t-sne. *Journal of machine learning research*, v. 9, n. Nov, p. 2579–2605, 2008. Citado 2 vezes nas páginas 51 e 54.

MCDONALD, J. *Handbook of Biological Statistics(3rd ed.)*. [S.l.]: Sparky House Publishing, 2014. 254–260 p. Citado na página 44.

NAVAZ, A. N. et al. The use of data mining techniques to predict mortality and length of stay in an icu. In: IEEE. *Innovations in Information Technology (IIT), 2016 12th International Conference on*. [S.l.], 2016. p. 1–5. Citado 2 vezes nas páginas 13 e 35.

PAULA, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP São Paulo, 2004. Citado na página 27.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 39.

PRATI, R.; BATISTA, G.; MONARD, M. Curvas roc para avaliação de classificadores. *Revista IEEE América Latina*, v. 6, n. 2, p. 215–222, 2008. Citado na página 32.

RAGHUPATHI, W.; RAGHUPATHI, V. Big data analytics in healthcare: promise and potential. *Health information science and systems*, BioMed Central, v. 2, n. 1, p. 3, 2014. Citado 2 vezes nas páginas 12 e 13.

RITTA, C. de O.; GORLA, M. C.; HEIN, N. Modelo de regressão logística para análise de risco de crédito em uma instituição de microcrédito produtivo orientado. *Iberoamerican Journal of Industrial Engineering*, v. 7, n. 13, p. 103–122, 2015. Citado 2 vezes nas páginas 7 e 33.

SCERBO, M. et al. Prehospital triage of trauma patients using the random forest computer algorithm. *journal of surgical research*, Elsevier, v. 187, n. 2, p. 371–376, 2014. Citado 2 vezes nas páginas 36 e 38.

SCHAFER, J. L.; GRAHAM, J. W. Missing data: our view of the state of the art. *Psychological methods*, American Psychological Association, v. 7, n. 2, p. 147, 2002. Citado na página 19.

SCHMIDT, D. et al. Um modelo de predição de mortalidade em unidades de terapia intensiva baseado em deep learning. In: *18º Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS 2018)*. Porto Alegre, RS, Brasil: SBC, 2018. v. 18. Disponível em: <http://portaldeconteudo.sbc.org.br/index.php/sbcas/article/view/3677>. Citado 2 vezes nas páginas 37 e 38.

SILVA, I. et al. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In: IEEE. *Computing in Cardiology (CinC), 2012*. [S.l.], 2012. p. 245–248. Citado 2 vezes nas páginas 35 e 36.

STERNE, J. A. et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, British Medical Journal Publishing Group, v. 338, p. b2393, 2009. Citado na página 13.

TAYLOR, R. A. et al. Prediction of in-hospital mortality in emergency department patients with sepsis: A local big data–driven, machine learning approach. *Academic Emergency Medicine*, Wiley Online Library, v. 23, n. 3, p. 269–278, 2016. Citado 2 vezes nas páginas 36 e 38.

WALKER, S. H.; DUNCAN, D. B. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, Oxford University Press, v. 54, n. 1-2, p. 167–179, 1967. Citado na página 27.

WILCOXON, F. Individual comparisons by ranking methods. *Biometrics bulletin*, JSTOR, v. 1, n. 6, p. 80–83, 1945. Citado na página 34.

WILSON, D. R.; MARTINEZ, T. R. Reduction techniques for instance-based learning algorithms. *Machine learning*, Springer, v. 38, n. 3, p. 257–286, 2000. Citado na página 22.

WITTEN, I. H.; FRANK, E.; HALL, M. A. Data mining: Practical machine learning tools and techniques. Morgan Kaufmann Publishers Inc., 2011. Citado na página 29.

WONG, L.; YOUNG, J. A comparison of icu mortality prediction using the apache ii scoring system and artificial neural networks. *Anaesthesia*, Wiley Online Library, v. 54, n. 11, p. 1048–1054, 1999. Citado 3 vezes nas páginas 13, 35 e 38.

XIA, H. et al. A neural network model for mortality prediction in icu. In: IEEE. *2012 Computing in Cardiology*. [S.l.], 2012. p. 261–264. Citado 3 vezes nas páginas 35, 36 e 38.

ZAKI, M. J.; MEIRA, J. W. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. [S.l.]: Cambridge University Press, 2014. ISBN 9780521766333. Citado na página 28.

ZHANG, T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: ACM. *Proceedings of the twenty-first international conference on Machine learning*. [S.l.], 2004. p. 116. Citado na página 43.

ZHOU, X.-H.; MCCLISH, D. K.; OBUCHOWSKI, N. A. *Statistical methods in diagnostic medicine*. [S.l.]: John Wiley & Sons, 2009. v. 569. Citado na página 32.

ZIMMERMAN, J. E. et al. Acute physiology and chronic health evaluation (apache) iv: hospital mortality assessment for today's critically ill patients. *Critical care medicine*, LWW, v. 34, n. 5, p. 1297–1310, 2006. Citado na página 18.