UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Estudo Comparativo de Diferentes Algoritmos de Aprendizado de Máquina na Predição de Múltiplas Doenças a Partir de Medidas Fisiológicas, Hematológicas e Bioquímicas

Relatório do trabalho para avaliação da disciplina SME0829 - Aprendizado de Máquina

Emanuel Victor da Silva Favorato, 12558151
Gabriela Scaranello Teixeira de Barros, 9284334
Renan Gabriel Fogaça, 8139718

SÃO CARLOS
2024

1 Introdução

Com a revolução digital observada nos últimos anos, a coleta e armazenamento de dados vem se tornando cada vez mais eficiente, a fim de contornar eventuais erros médicos e agilizar o diagnóstico de determinadas enfermidades, muito vem se discutindo acerca do processamento de tais informações para a obtenção de resultados mais acertivos, dentre as ferramentas utilizadas para garantir essa interação entre a tecnologia e a medicina estão os algoritmos de Aprendizado de Máquina, cuja capacidade no oferecimento de diagnósticos precisos para doenças cardiovasculares já havia sido explorada por (Oliveira et al, 2023) e (da Silva Filho et al, 2022), entretanto, sua aplicação pode ser estendida às mais diversas comorbidades.

Este projeto visa estudar a aplicabilidade de diversos modelos paramétricos e não-paramétricos de Aprendizagem de Máquina para além de problemas cardiovasculares, visando explorar, também, sua capacidade de predição em doenças como Trombose, Diabetes, Thalassemia e Anemia, bem como na diferenciação entre indivíduos saudáveis e não-saudáveis, valendo-se da análise de covariáveis obtidas através de exames médicos e sanguíneos. O objetivo é comparar a performance dos diferentes modelos de aprendizado de máquina estudados e identificar qual é o mais adequado para a predição de múltiplas doenças a partir da base de dados utilizada.

2 Metodologia

Para executar nossos objetivos neste trabalho, seguimos as etapas descritar abaixo:

1. Compreensão do problema: estudamos as variáveis presentes na base de dados escolhida a fim de compreender melhor nosso problema de pesquisa. Identificamos que esta se encaixava em um problema de aprendizado de máquina do tipo classificação multiclasse.
2. Estudamos, através das aulas da disciplina de Aprendizado de Máquina e de materiais complementares, algoritmos de aprendizado de máquina para classificação multiclasse, bem como técnicas de treinamento, seleção de modelo e avaliação de desempenho desses algoritmos.
3. Preparação dos dados: tratamos os dados para nos certificar da ausência de dados nulos e da normalização de todas as covariáveis numéricas.
4. Análise exploratório e vizualização dos dados: exploramos o balanceamento dos dados em relação à variável resposta, a correlação entre covariáveis e a distribuição de cada covariável.
5. Modelagem: implementamos 4 modelos paramétricos e 5 modelos não-paramétricos. Selecionamos os hiperparâmetros mais adequados para cada um deles.
6. Avaliação: foram computadas as medidades de desempenho acurácia, sensibilidade e precisão. Além disso, foi também computado o tempo de execução para cada modelo selecionado. Para aqueles modelos que eram compatíveis, foi feita uma análise de importância de atributos ("*feature importance*").
7. Análise dos resultados e conclusão: por fim, comparamos as medidas de desempenho e custo computacional dos modelos utilizados e selecionamos os modelos que melhor se ajustam ao nosso problema, considerando as vantegens e desvantagens de cada um.

3 A base de dados

A base de dados Multiple Disease Prediction [1] é um conjunto de dados disponibilizado no *site Kaggle* projetado para simular a avaliação o estado de saúde, determinando se uma pessoa tem uma doença específica ou se está saudável, com base em amostras de sangue e outros vários parâmetros. O conjunto de dados contém resultados simulados de 2.839 exames, onde são consideradas 24 variáveis como níveis de glicose, colesterol, pressão sanguínea e etc. para indicar 5 tipos diferentes de doenças específicas ou problemas cardíacos. Na base de dados, foi pesquisada as possíveis doenças associadas a variações desses parâmetros.

Segue abaixo, o glossário das 24 variáveis preditoras:

- Colesterol: nível de colesterol no sangue, medida em miligramas por decilitro (mg/dL).
- Hemoglobina: proteína dos glóbulos vermelhos que carrega oxigênio do pulmão para o resto do corpo.
- Plaquetas: células sanguíneas que ajudam a coagulação.
- Leucócitos (células brancas do sangue): células do sistema imunológico que combate infecções.
- Glóbulos vermelhos (células vermelhas do sangue): células que transportam oxigênio dos pulmões para o resto do corpo.
- Hematócrito: porcentagem que representa a quantidade de glóbulos vermelhos no volume total do sangue.
- Volume Corpuscular Médio: volume médio de glóbulos vermelhos.
- Hemoglobina Corpuscular Média: quantidade média de hemoglobina nos glóbulos vermelhos.
- Concentração de Hemoglobina Corpuscular Média: Concentração média de hemoglobina em glóbulo vermelho.
- Insulina: hormônio que regula o nível de glicemia no sangue.
- IMC (Índice de Massa Corporal): medida de gordura corporal com base na altura e peso.
- Pressão Arterial Sistólica: pressão arterial.
- Pressão Arterial Diastólica: pressão arterial durante a diástole ventricular.
- Triglicérides: tipo de gordura encontrada no sangue, medida em miligramas por decilitro (mg/dL).
- Hemoglobina Glicada: média do nível de glicemia nos últimos 2 ou 3 meses.
- Colesterol LDL: colesterol ruim que pode se acumular nas artérias.
- Alanina Aminotransferase: enzima encontrada principalmente no fígado.
- Colesterol HDL: colesterol bom que auxilia na remoção do colesterol LDL das artérias.

[1] https://www.kaggle.com/datasets/ehababoelnaga/multiple-disease-prediction

Doença	Frequência absoluta	Porcentagem do total
Diabetes	827	29.2%
Anemia	707	25.0%
Saudável	560	19.8%
Talassemia	556	19.7%
Trombose	139	4.9%
Doença cardíaca	39	1.4%

Tabela 1: Tabela da frequência absoluta e da porcentagem do total de cada classe da variável resposta "Doença".

- Aspartato Aminotransferase: enzima encontrada em vários tecidos incluindo fígado e coração.
- Frequência Cardíaca: quantidade de batidas por minuto (bpm).
- Creatina: resíduo produzido pelos músculos e filtrado do sangue pelos rins.
- Troponina: proteína liberada na corrente sanguínea quando há dano ao músculo cardíaco.
- Proteína C-Reativa: marcador de inflamação no corpo.

3.1 Preparação dos dados

Para preparar os dados para a análise exploratória, e posterior treinamento de modelos, concatenamos as duas bases de dados fornecidas pelo repositório do *Kaggle*. Decidimos aderir a essa estratégia porque a variável "Doença cardíaca"estava presenta em apenas uma das bases, o que impediria a classificação dessa variável.

Posteriormente, verificamos a ausência de dados nulos e a presença de dados duplicados. Os dados duplicados eram 2286 de 2837 e, por esta razão, optamos por mantê-los. A alta presença de dados duplicados pode ser explicada por duas razões. A primeira é que os dados forma baseados em resultados reais de exames e a partir deles alguns foram replicados para formar a base de dados artificial. A segunda é que os valores dos parâmetros fisiológicos, hematológicos e bioquímicos frequentemente são similares em diferentes pessoas, já que estes usualmente possuem um intervalo pequeno em que o parâmetros é considerado normal e algo fora disso corresponde a um resultado alterado.

Por fim, averiguamos se todas estavam medidas estarem entre 0 e 1. Foram encontradas 9 observações com valores abaixo negativo e, por serem incoerentes com a natureza das medidas, excluímos esses dados da base.

3.2 Análise exploratória

Nesta etapa, buscamos entender melhor a composição do banco de dados e a relação entre covariáveis. Como mostra a Tabela 1, o conjunto de dados é desbalanceado, contendo apeans 39 observações da classe "Doença Cardíaca".

3.2.1 Correlações entre variáveis

As correlação entre covariáveis, foram todas próximas de 0, variando entre -0.2 e 0.2. No entanto, ao separar os dados por classe, houve uma mudança no comportamento das covariáveis, o que indica que a alteração de algumas covariáveis podem estar mais correlacionadas a uma determinada doença. Por exemplo, no Grafico da Figura 1 podemos ver que a variável insulina está mais correlacionada com colesterol, mas a correlação destas variáveis é baixa para a classe "Talassemia"na Figura 2.

Figura 1: Gráfico de correlação das variáveis preditoras para a classe "Saudável"

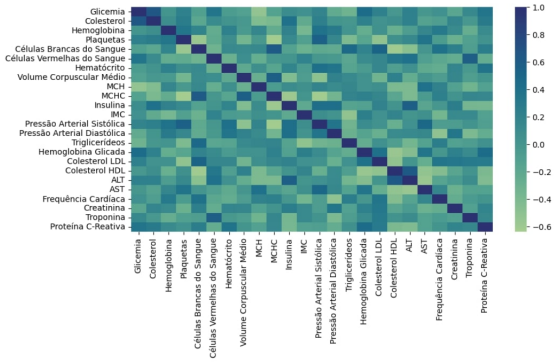
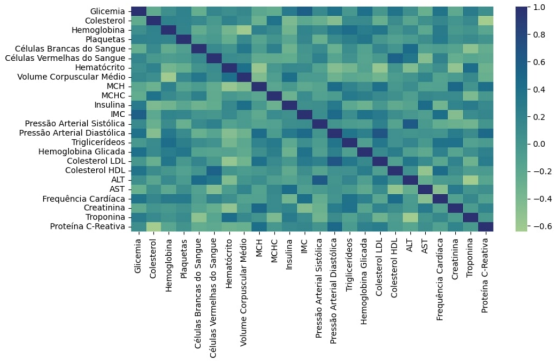


Figura 2: Gráfico de correlação das variáveis preditoras para a classe "Talassemia"



4 Treinamento dos modelos

Para o treinamento dos modelos, a base de dados foi dividida em uma base de treino, com 80% das observações da base original selecionadas aleatóriamente, e uma base de teste com o restante 20% das observações para a avaliação da performance dos modelos posteriormente. Com excesssão das Redes Neurais Artificiais, todos os modelos foram ajustados utilizando a biblioteca *scikit-learn* do *software Python*. As Redes Neurais foram treinadas por meio da biblioteca *TensorFlow*, do mesmo textitsoftware.

4.1 Modelos paramétricos

Dentre os modelos paramétricos de aprendizado de máquina para classificação, escolhemos implementar a Regressão Logística e o *Naive Bayes*. Sendo a primeira implementada de diversas variações que contemplam a classficação multiclasse: "Um contra o resto", multinomial, com e sem penalização, com e sem validação cruzada.

4.1.1 Regressão Logística Multiclasse (OvR) Sem Penalização

A Regressão Logística Multiclasse (OvR) ou "Um contra o resto"utiliza a mesma técnica da regressão logística para classificação binária. No entanto, para cada classe, é executado um algoritmo de regressão em que a variável resposta é classificado como pertencendo a esta classe ou como não pertencendo - isto é, pertencente a qualquer outra.

4.1.2 Regressão Logística Multinomial Sem Penalização

A Regressão Logística Multinomial é uma extenção da Regressão Logística Binária para classificação multiclasse.

4.1.3 Regressão Logística Multinomial com Regularização L1

Em apredizado de máquina, as penalizações adicionam um termo de regularização à função de custo a ser minimizada. No caso, da Regressão Logística Multinomial a função de custo é a função *softmax*. A regularização L1 é a mesma utilizada na Regressão Lasso, baseada no módulo dos coenficientes, portanto esse modelo é zera alguns coeficientes estimados. No entanto, para a classificação de 6 classes e 24 váriaveis preditoras, a Regressão Logística Multinomial estima 125 coeficientes, ou 5 vetores 25-dimensionais. No caso do modelo treinado neste estudo, nenhum dos coeficientes foi zerado simultaneamente em todos os 5 vetores, portanto o modelo não foi significativamente simplificado pela penalização.

4.1.4 Regressão Logística Multinomial com Regularização L2

A Regressão Logística Multinomial com Regularização L2 consite em uma Regressão Logística Multinomial que possui a mesma penalização da Regressão Ridge, baseada no quadrado dos coeficientes.

4.1.5 Regressão Logística Regularização L2 e CV

Este algoritmo é similar ao anterior, diferindo-se apenas pela aplicação do método de validação *cross-validation* ou validação cruzada com hiperparâmetro igual a 10. Assim, a validação cruzada dividiu os dados em 10 partes e avaliou o modelo 10 vezes. A intensão de aplicar esta técnica é de melhorar a acurácia do modelo.

4.1.6 *Naive Bayes*

O *Naive Bayes* consiste em um grupo de classificadores probabilísticos simples que se valem da aplicação do Teorema de Bayes a partir de uma forte suposição de independência entre as covariáveis, algo que raramente se observa em casos práticos. Mesmo com os problemas apresentados, a facilidade e agilidade de implementação, bem como seu desempenho, o tornam bastante útil em vários casos. Neste projeto, foi implementado um modelo Naive Bayes Gaussiano, a fim de avaliar seu poder preditivo comparado aos demais.

4.2 Modelos não-paramétricos

4.2.1 *XGBoosting*

O Extreme Gradient Boosting (XGBoosting) consiste em um aprimoramento do algoritmo de Gradient Boosting, projetado para melhor eficiência em termos de tempo e recursos computacionais. Este algoritmo destaca-se devido ao seu melhor desempenho e flexibilidade.

4.2.2 *K-Nearest Neighbors* (KNN)

O KNN (*K-Nearest Neighbors*) é um algoritmo que se destaca por sua simplicidade, intuitividade e facilidade de implementação, consistindo em um modelo que utiliza a distância entre os dados de treinamento e de entrada a fim de classificar as variáveis. Este projeto utiliza a distância Bray-Curtis, bastante útil para a diferenciação de grupos biológicos, considerando $K = 2$ vizinhos.

4.2.3 Floresta Aleatória (*Random Forest*)

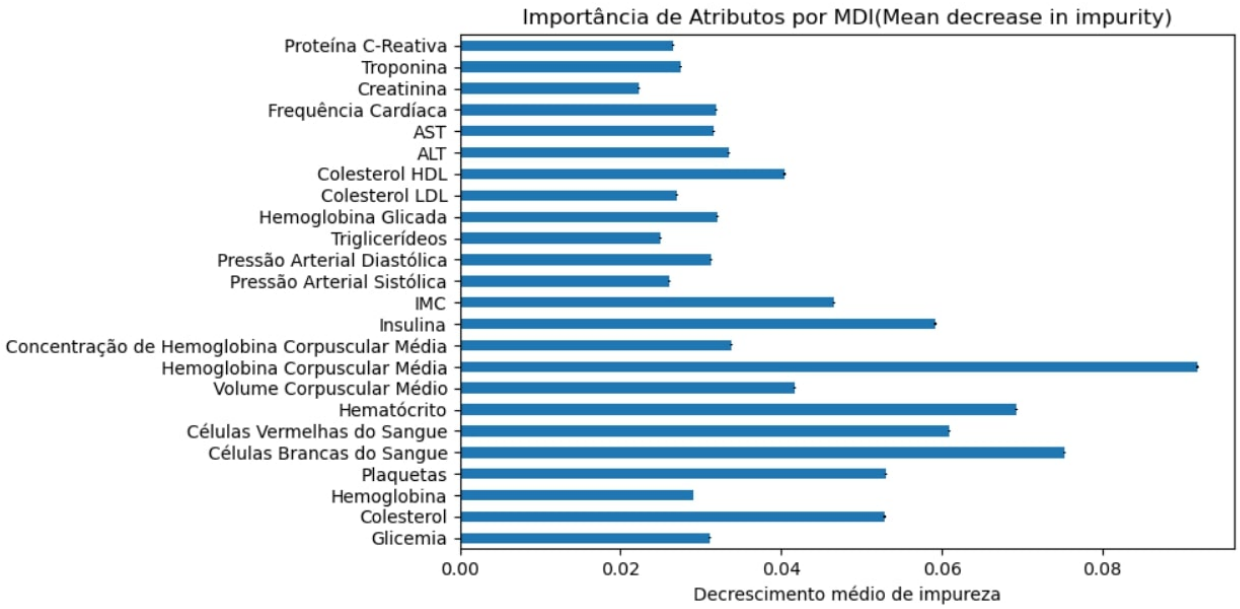
A *Random Forest* consiste em um algoritmo de aprendizado de máquina que se vale do conceito de *ensemble learning* para construir várias árvores de decisão de modo a obter a previsão final com base nas previsões de cada árvore. Este algoritmo é bastante eficaz em problemas de aprendizado supervisionado.

Este algoritmo utiliza os hiperparâmetros, número de estimadores e máxima profundidade. O número de estimadores corresponde ao número de árvores de decisão, já a profundidade máxima corresponde à profundidade de cada uma dessas árvores. Para selecionar os melhores hiperparâmetros, implementamos uma procura randomizada[2] nos intervalos (50,700) para os estimadores e (1,50) para profundidade máxima. Obtivemos como hiperparâmetros ótimos "número de estimadores = 318"e "profundidade máxima = 46".

Além disso, a Floresta Aleatória permite a análise de importância de atributos pelo método de decréscimo médio da impureza. O gráfico da Figura 2 ilustra a importância de cada variável preditora para o modelo ajustado. Destancam-se como mais relevantes a Hemoglobina Corpuscular Média, Hemócrito, Células Brancas, Insulina, Colesterol e IMC.

[2]Para executar a procura randomizada, utilizamos a função *RandomizedSearchCV* do pacote *ScikitLearn* para o *software Python*.

Figura 3: Gráfico de importância de atributos para o modelo de Floresta Aleatória

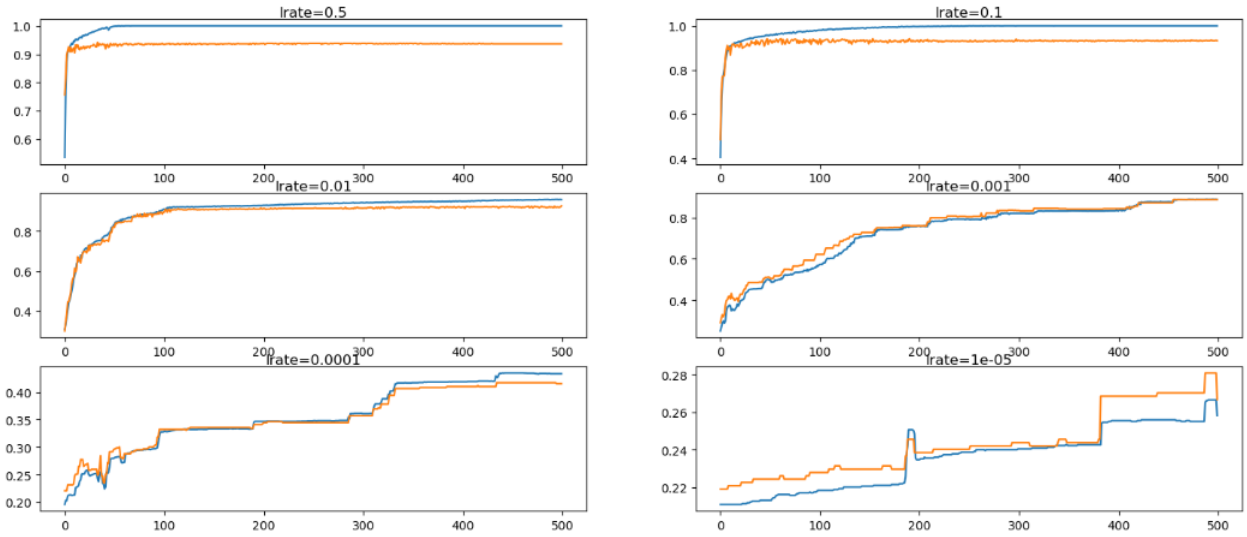


4.2.4 Redes Neurais Artificiais

As Redes Neurais Artificiais são modelos computacionais inspirados na estrutura e funcionamento das redes neurais do cérebro humano, com alta aplicabilidade e complexidade computacional. Para treinar um modelo de rede neural artificial, é necessário determinar diversos hiperparâmetros como: número de camadas, número de neurônios, número de épocas e taxa de aprendizado.

Para o treinamento das Redes Neurais Artificiais, foi necessário mapear as 6 classes da variável resposta em vetores numéricos, por meio da técnica do *one-hot-enconding*. Então, foram selecionados os hiperparâmetros taxa de aprendizado e épocas mais adequados por meio da análise do desempenho de modelos de Redes Neurais Artificiais treinados com diferentes tavas de aprendizados (0.5, 0.1, 0.01, 0.001, 0.0001, 0.00001) e 500 épocas.

Figura 4: Gráfico de acurácia por época para os cojuntos de validação (em laranja) e teste (em azul) para cada valor de taxa de aprendizado testado no ajuste dos hiperparâmetros das Redes Neurais Artificiais.



Como podemos observar no gráfico da Figura 3, o modelo com taxa de aprendizado igual a 0.5 foi o que convergiu mais rapidamente, antes daa 100 primeiras épocas. Assim, definimos os hiperparâmetros taxa de aprendizado igual a 0,5 e épocas igual a 100. As Redes Neurais Artificiais foram treinadas com 50 camadas de 24 neurônios e uma camada com 6 neurônios.

5 Avaliação e discussão

Para comparar a performace dos diversos algoritmos implementados e selecionar o mais adequado para nosso problema, utilizamos as medidas de precisão e sensibilidade para cada classe da variável resposta, a acurácia e o tempo de execução de cada modelo. Nos baseamos principalmente na medida de acurácia para a decisão do melhor modelo.

Os gráficos 3 a 8 nos fornecem a comparação dos valores de precisão e sensibilidade de cada modelo por classe da variável resposta. Com base nestes gráficos, constatamos que as classes "saudável", "diabetes", "talassemia", "anemia"e "trombose"possuem valores de sensibilidade e precisão acima de 0.7 para todos os modelos, com destaque para a melhor performance nesses casos para os modelos não-paramétricos - em especial, *XGBoosting* e Floresta Aleatória.

Por outro lado, a classe de "Doença Cardíaca"foi a que apresentou maior variabilidade entre modelos das medidas de sensibilidade e precisão. Isso se deve ao desbalanceamento da base de dados, que contém apenas 39 observações desta classe. Destaca-se neste contexto o modelo da Floresta Aleatória, que atingiu os maiores níveis de precisão e sensibilidade para esta classe simultaneamente.

Figura 5: Gráfico de Presisão e Sensibilidade para a classe "saudável"

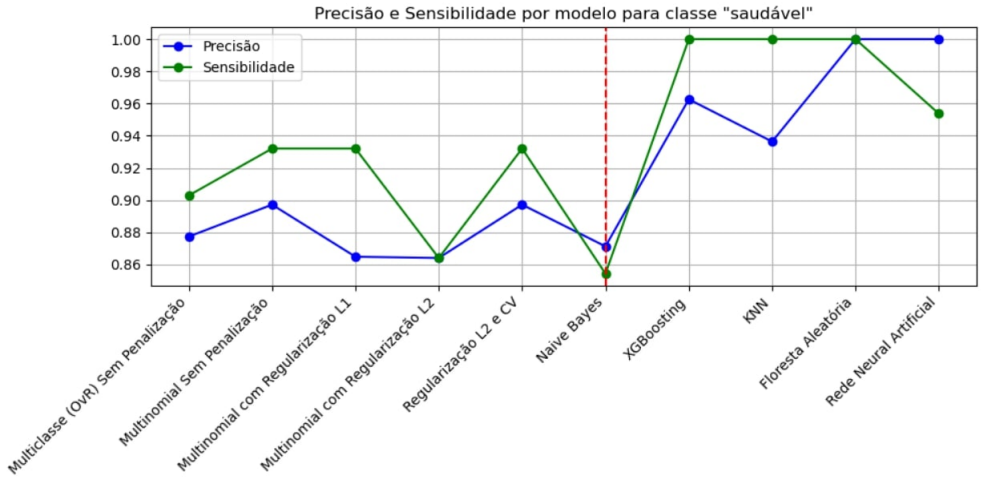


Figura 6: Gráfico de Presisão e Sensibilidade para a classe "diabetes"

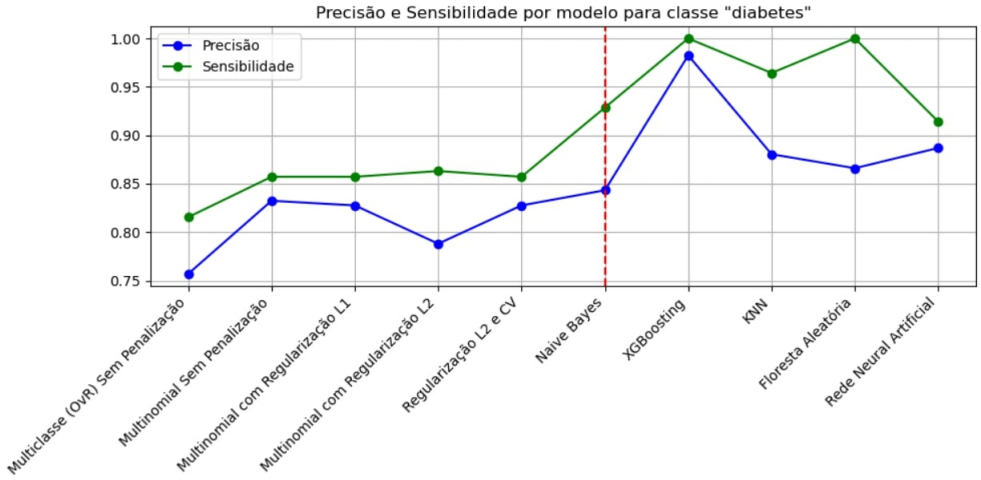


Figura 7: Gráfico de Presisão e Sensibilidade para a classe "talassemia"

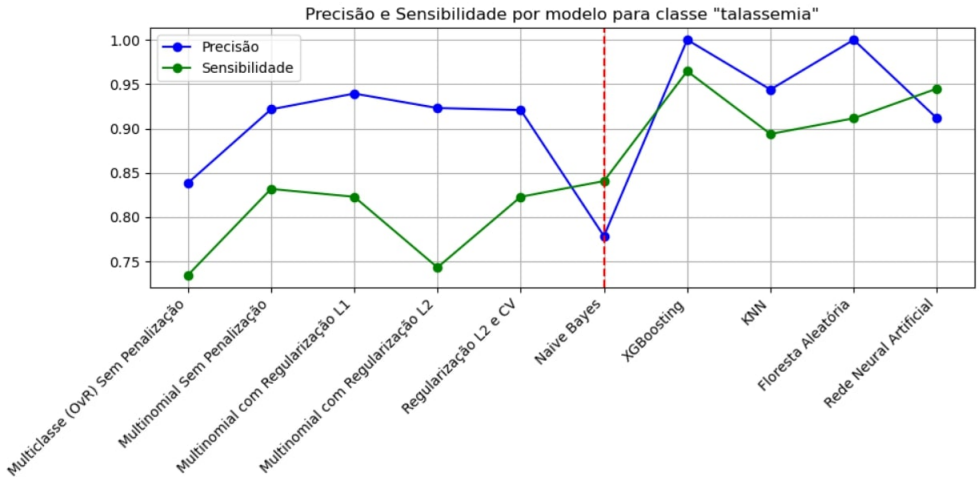


Figura 8: Gráfico de Presisão e Sensibilidade para a classe "anemia"

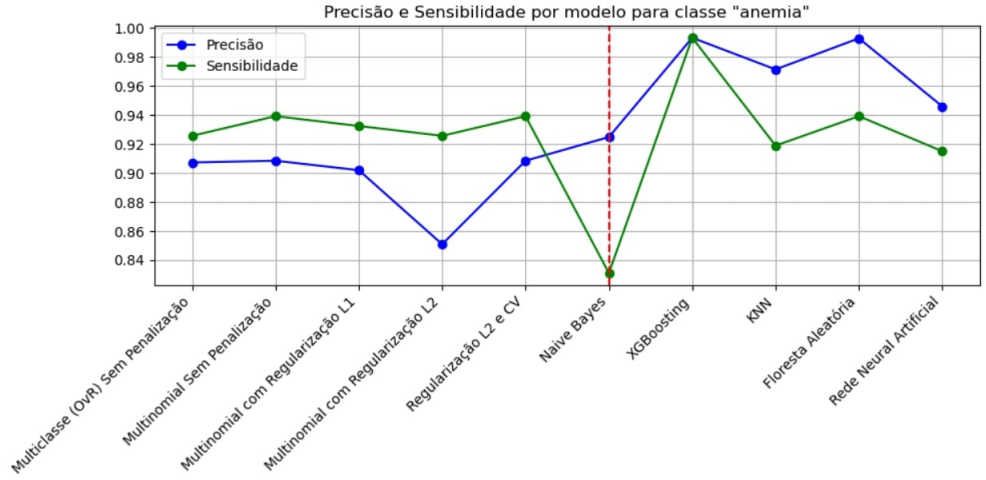


Figura 9: Gráfico de Presisão e Sensibilidade para a classe "trombose"

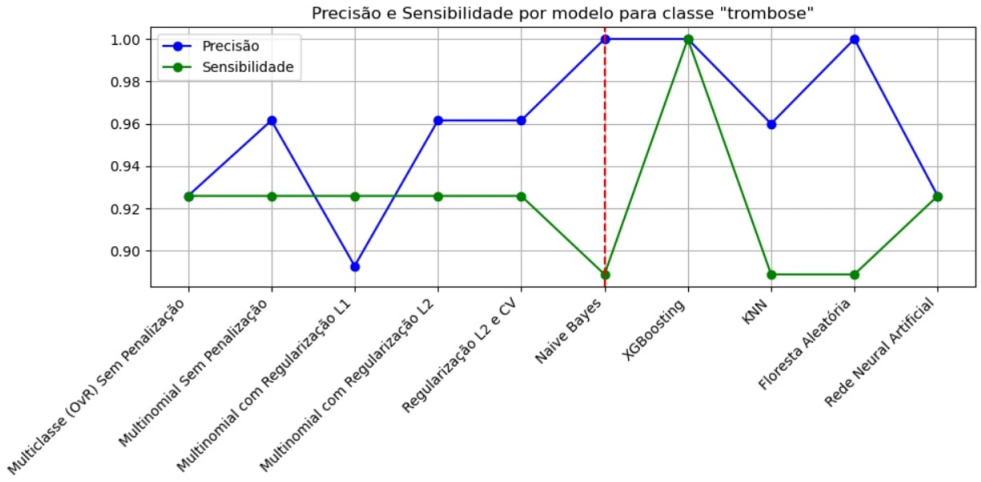
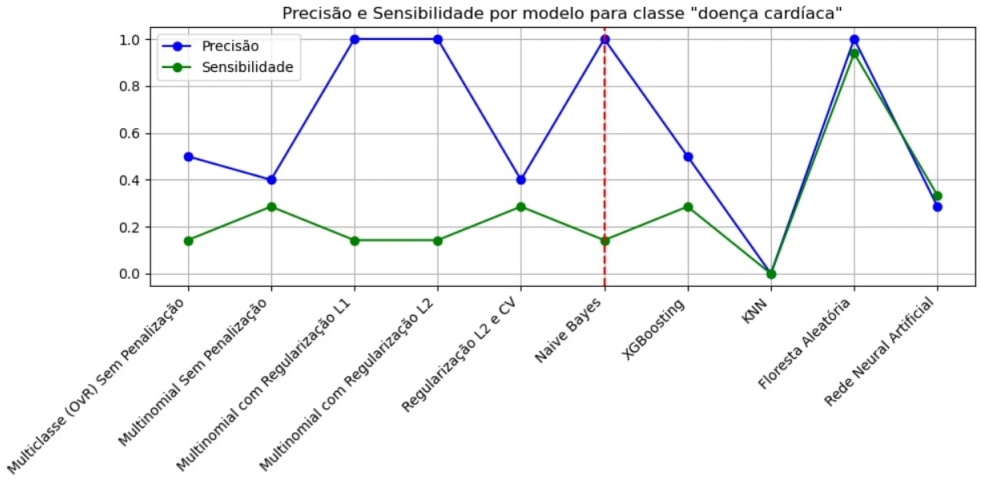


Figura 10: Gráfico de Presisão e Sensibilidade para a classe "doença cardíaca"



A Tabela 2 fornece a comparação do desempenho dos modelos com base na acurácia e no tempo de execução. Novamente, os modelos não-paramétricos superaram os paramétricos em termos de performance, obtendo maiores acurácias. Em termos de tempo de execução, a Rede Neural Artificial demorou mais de 8000% a mais que o segundo maior tempo e portanto, é mais computacionalmente eficaz que os outros algoritmos para o problema do estudo. O melhor desempenho foi do *XGBoosting*, com a maior acurácia de todas e um tempo de execução baixo. Além disso, destaca-se a eficiência dos algoritmos Naive Bayes e KNN pelos valores de acurácia relativamente altos e baixo tempo de execução.

Apesar da acurácia alta para um tempo de execução menor que os outros métodos não-paramétricos, o *KNN* não é o modelo mais adequado para a predição de doenças múltiplas com esta base de dados, já que

ele foi ineficaz em fazer a classificação da classe "Doença Cardíaca"e obteve 0 em precisão e sensibilidade. Já a Floresta Aleatória, apesar do bom desempenho de acurácia e tempo e dos maiores valores para sensibilidade e precisão da classe "Doença Cardíaca"dentre todos os modelos, é inferior ao *XGBoosting* em precisão e sensibilidade 4 da demais 5 classes. Portanto, devido à maior acurácia, bom desempenho de precisão e sensibilidade em todas as classes da variável resposta e baixo tempo de execução, identificamos que o modelo não-paramétrico *XGBoosting* é o melhor modelo ajustado para classficar doenças múltiplas a partir da base de dados utilizada.

Modelo	**Acurácia**	**Tempo (s)**
Multiclasse (OvR) Sem Penalização	0,841	0,168
Multinomial Sem Penalização	0,883	3,744
Multinomial com Regularização L1	0,878	0,527
Multinomial com Regularização L2	0,850	0,044
Regularização L2 e CV	0,881	2,917
Naive Bayes	0,860	0,006
XGBoosting	**0,982**	**0,326**
KNN	0,929	0,080
Floresta Aleatória	0,943	1,216
Rede Neural Artificial	0,922	32,386

Tabela 2: Tabela de comparação da performance dos modelos estudados através da acurácia e do tempo de execução em segundos. Destaca-se a melhor performance: *XGBoosting*.

6 Conclusão

Com o intuito de classificar multiplas comorbidades que acometem o organismo humano, este relatório apresenta modelos de Aprendizado de Máquina de natureza Paramétrica e Não-Paramétrica, a fim de obter aquele que melhor capta os nuances das 24 covariáveis análizadas, de modo a garantir maior acertividade e qualidade na predição, dentre os 10 classificadores analizados, podemos concluir que o XGBoosting é o que melhor se adequa ao problema proposto, uma vez que se destaca por possuir uma acurácia superior a 98%., relativa agilidade para sua execução, bem como por apresentar índices superiores de precisão e sensibilidade em quase todas as respostas, também apresentaram bons desempenhos os algorítmos de Floresta Aleatória e Naive Bayes, em especial pela sua precisão na detecção de trombose e doença cardíaca, as quais são mais raras e possuem menos informação a ser extraída com base nas covariáveis fornecidas, sendo que, para o último, a Floresta Aleatória também se destacou por sua sensibilidade, além de possuir a segunda maior acurácia dentre os modelos apresentados, acima de 94 %., embora apresente um tempo de rodagem não tão baixo.

Referências bibliográficas

Oliveira, Marcos Costa and Ferreira, Luís Victor Belo and de Oliveira Barreiros, Marta (2023), 'CLASSIFICAÇÃO DE DOENÇAS CARDIOVASCULARES UTILIZANDO APRENDIZADO DE MÁQUINA'.

da Silva Filho, Francisco Romes and Coutinho, Emanuel F (2022), 'Aprendizado de Máquina para Predição de Diagnósticos de Doenças Cardiovasculares'.

Silva, Tainá Caroline Beletti Valente (2023), 'Parâmetros fisiológicos, hematológicos e bioquímicos de pôneis hígidos'.

https://www.kaggle.com/datasets/ehababoelnaga/multiple-disease-prediction

https://scikit-learn.org/stable/modules/linear$_m odel.html logistic - regression$