

## RELATÓRIO FINAL DE PROJETO DE INICIAÇÃO CIENTÍFICA

**Aluno:** Gustavo Henrique Marques de Campos Pacheco

**Orientadora:** Carolina Paula de Almeida

**Período:** 01/09/2022 a 01/08/2023

**Título do Plano de Atividades:** Detecção de Doença Crônica Renal com Aprendizado de Máquina

### Resumo:

*A detecção da Doença Renal Crônica (DRC) representa um desafio, uma vez que a progressão da doença pode ocorrer de forma assintomática por um longo período de tempo. Adicionalmente, muitas vezes os sintomas da DRC só são identificados quando a doença já se encontra em estágio avançado, o que dificulta a intervenção precoce. Diante desse cenário, estratégias para a detecção nos estágios iniciais precisam ser arquitetadas. Nesse sentido, considerando as dificuldades na detecção precoce da DRC, é relevante ponderar sobre a aplicação de técnicas avançadas, tais como o aprendizado de máquina, com o intuito de aprimorar a precisão do diagnóstico e garantir intervenções eficazes. O objetivo principal deste projeto foi o estudo e aplicação de uma técnica de Aprendizado de Máquina, chamada AdaBoost, para a tarefa de classificação de dados, aplicada à previsão de doença renal crônica.*

**Palavras-Chaves:** AdaBoost, Aprendizado supervisionado, Tratamento de Dados, Modelos preditivos, Diagnóstico médico.

## 1 Introdução

Inteligência Artificial (IA) é definida como uma tecnologia que utiliza o conhecimento computacional para representar o comportamento inteligente com envolvimento humano, sendo o Aprendizado de Máquina considerado como um subconjunto das técnicas de IA, como define a IBM (2023). Normalmente, esse tipo de inteligência é comumente reconhecido como tendo começado com a inovação da robótica. Com o crescimento e desenvolvimento admiravelmente veloz do universo eletrônico, principalmente com o setor que diz respeito à área da programação, tudo indica que os computadores e máquinas futuras podem apresentar um comportamento inteligente/racional satisfatoriamente semelhante ao dos seres humanos, considerando todos os dados e avanços que indicam evolução dos quais já se tem conhecimento.

Um exemplo de uma das inúmeras evoluções que esse campo mostra ter grande desenvolvimento é a aplicação da IA no universo da medicina. Em 2016, projetos acoplados com o campos medicinal formaram mais especulações para a economia global que outros demais desenvolvimentos Buch et al. (2018). Na medicina, a IA refere-se na automação do diagnóstico e no tratamento do paciente e se a utilização da mesma for mais presente, a mesma permitirá que uma parte considerável do papel dos peritos e responsáveis da área seja automatizada. O aumento da utilização da IA na prescrição permitirá que uma parte considerável do papel seja

automatizada, abrindo o tempo dos especialistas em medicina para serem usados na execução de diferentes obrigações, que não podem ser automatizadas. Como tal, esta tecnologia promete uma utilização progressivamente significativa no domínio dos Recursos Humanos (RH).

Não obstante às vantagens supracitadas, é possível melhorar a qualidade dos dados médicos, reduzir as flutuações nas taxas de pacientes e economizar nos custos médicos. Tendo em vista tais apontamentos, esses modelos de Aprendizado de Máquina são frequentemente utilizados para investigar a análise diagnóstica quando comparados com outros métodos convencionais. Para reduzir as taxas de mortalidade por doenças crônicas (DCs), a detecção precoce e tratamentos eficazes são boas soluções. Portanto, muitos dos cientistas médicos é atraída pelas novas tecnologias de modelos preditivos na previsão de doenças. Esses novos avanços na assistência médica vêm expandindo a acessibilidade de dados eletrônicos e abrindo novas portas para suporte à decisão e melhorias de produtividade. Os métodos de Aprendizado de Máquina têm sido efetivamente utilizados na interpretação computadorizada de testes de capacidade pneumônica para análise diferencial de DCs. Ao que tudo indica, que os modelos com as maiores acurácias possam ganhar grande importância no diagnóstico médico.

## 2 Objetivos

### 2.1 Objetivo Geral

O objetivo principal deste projeto de Iniciação Científica foi o estudo e aplicação de uma técnica de Aprendizado de Máquina na tarefa de classificação de dados, aplicada à previsão de doença renal crônica.

### 2.2 Objetivos específicos

Dentre os objetivos específicos, podem ser citados:

- Pesquisa sobre o conjunto de dados a ser utilizado, bem como sua exploração e análise;
- Estudo do método de aprendizado de máquina AdaBoost a ser aplicado na tarefa de classificação;
- Implementação do classificador utilizando a biblioteca *scikit-learn* no conjunto de dados estudado;
- Análise e disseminação dos resultados;

## 3 Fundamentação Teórica

Nesta seção de fundamentação teórica, são apresentados os principais conceitos relacionados ao uso do aprendizado de máquina na detecção de doença renal crônica. Através da revisão da literatura científica e de estudos relevantes já realizados, são abordados os aspectos essenciais para a compreensão e o desenvolvimento da pesquisa.

### 3.1 Doença Crônica Renal

A doença renal crônica (DRC) é uma condição de saúde progressiva e irreversível, caracterizada pela perda gradual da função renal ao longo do tempo. Essa doença afeta os rins, órgãos responsáveis pela filtragem do sangue e pela remoção de resíduos e excesso de fluidos do organismo. A DRC pode ser causada por diversas condições, como diabetes, hipertensão arterial, doenças autoimunes, infecções recorrentes, entre outras Wibawa et al. (2017).

O diagnóstico precoce da doença renal crônica desempenha um papel fundamental na sua gestão e tratamento adequados. Uma detecção precoce permite o início de intervenções terapêuticas e mudanças no estilo de vida que podem retardar a progressão da doença, reduzir complicações e melhorar a qualidade de vida dos pacientes Wibawa et al. (2017).

A importância do diagnóstico no início da doença renal crônica reside no fato de que, muitas vezes, os sintomas da doença são sutis ou inexistentes em estágios iniciais. Os pacientes podem não apresentar sintomas evidentes até que a função renal esteja significativamente comprometida. Assim, a DRC pode passar despercebida por um longo período, levando a um diagnóstico tardio e um agravamento da condição Wibawa et al. (2017).

A detecção precoce da doença renal crônica envolve a avaliação de marcadores clínicos, como a taxa de filtração glomerular (TFG) e a presença de proteínas na urina. Além disso, exames laboratoriais, como a creatinina sérica e o clearance de creatinina, são utilizados para avaliar a função renal. No entanto, essas abordagens tradicionais podem apresentar limitações, como a falta de sensibilidade e especificidade em estágios iniciais da doença Wibawa et al. (2017).

### 3.2 Aprendizado de Máquina

O aprendizado de máquina, também conhecido como *machine learning* em inglês, é uma área da inteligência artificial que se concentra no desenvolvimento de algoritmos e técnicas que permitem que computadores “aprendam” e realizem tarefas sem serem explicitamente programados para cada uma delas. O objetivo é capacitar as máquinas a aprenderem com os dados disponíveis e a tomarem decisões ou fazerem previsões com base nesse aprendizado Bi et al. (2019).

O cerne do aprendizado de máquina está na capacidade de identificar padrões nos dados e utilizar esses padrões para realizar tarefas específicas. Em vez de programar regras e instruções específicas para cada cenário, os algoritmos de aprendizado de máquina são projetados para extrair informações relevantes dos dados e aprender com elas Bi et al. (2019).

Existem diferentes abordagens e técnicas de aprendizado de máquina, incluindo aprendizado supervisionado, não supervisionado e por reforço. No aprendizado supervisionado -técnica a qual empregamos neste trabalho, um modelo é treinado usando um conjunto de dados rotulados, ou seja, dados em que as respostas corretas são conhecidas, com o objetivo de prever respostas para novos dados. No aprendizado não supervisionado, não há rótulos e o modelo é treinado para identificar padrões e estruturas nos dados. O aprendizado por reforço envolve treinar um modelo através da interação com um ambiente, em que o modelo aprende a tomar ações que maximizem uma recompensa Géron (2019).

O aprendizado de máquina tem uma ampla gama de aplicações em diversos campos, incluindo medicina, finanças, indústria, transporte, *marketing* e muitos outros. Pode ser utilizado para classificação de dados, previsão de tendências, detecção de anomalias, recomendação de produtos, análise de sentimentos, entre outras tarefas Géron (2019).

Para que o aprendizado de máquina seja eficaz, é necessário um conjunto de dados de qualidade, que seja representativo e abranja a diversidade dos casos possíveis. Além disso, é

importante ter algoritmos robustos, que sejam capazes de lidar com diferentes desafios, como sobreajuste, generalização inadequada e interpretabilidade dos resultados Geron (2019).

Em resumo, o aprendizado de máquina é uma disciplina poderosa que permite que as máquinas aprendam com os dados e realizem tarefas complexas sem serem explicitamente programadas para cada situação. Essa abordagem revolucionária tem o potencial de impulsionar avanços significativos em diversos setores, oferecendo *insights* valiosos, automação inteligente e tomada de decisões baseada em dados.

### 3.3 *DataSet*

Um conjunto de dados, comumente conhecido como *dataSet*, representa uma compilação estruturada de informações organizadas em tabelas, onde cada linha representa uma instância ou exemplo, e cada coluna descreve uma característica específica dessas instâncias. Esse formato proporciona uma abordagem sistemática e organizada para o armazenamento e análise de dados coletados com um propósito definido.

Conjuntos de dados podem abranger uma ampla variedade de tipos de dados, incluindo números, texto, imagens e áudio, dependendo do contexto e da natureza da pesquisa ou problema em questão. Por exemplo, em um estudo relacionado à doença renal crônica, um *dataSet* pode conter informações clínicas dos pacientes, como idade, sexo, histórico médico, resultados de exames laboratoriais, dados de imagem e outros atributos relevantes.

Esses conjuntos de dados desempenham um papel fundamental na aplicação de técnicas de aprendizado de máquina, pois servem como base para treinar e avaliar modelos preditivos e classificatórios. Os algoritmos de aprendizado de máquina utilizam os dados contidos no *dataSet* para aprender a reconhecer padrões, fazer previsões ou tomar decisões com base nesses dados.

No contexto específico da detecção de doença renal crônica por meio do aprendizado de máquina, um *dataSet* consistiria em uma reunião de informações clínicas e resultados de exames laboratoriais de pacientes. Esse *dataSet* seria usado como entrada para o treinamento e validação de modelos de aprendizado de máquina, com o objetivo de identificar padrões e correlações relevantes para a condição.

É fundamental que esses conjuntos de dados sejam representativos e abrangentes, capturando a diversidade de casos possíveis e incorporando informações cruciais para o problema em análise. Além disso, é necessário realizar um pré-processamento adequado dos dados, incluindo limpeza, normalização e transformação, a fim de garantir a qualidade e a consistência dos dados utilizados nos algoritmos de aprendizado de máquina Wibawa et al. (2017).

Neste trabalho, foi utilizado o *dataSet* UCI Machine Learning <sup>1</sup>. Ele é composto por 280 instâncias e 25 atributos e apresenta informações para que o diagnóstico da doença seja feito - pertencente à classe CKD para pacientes com a doença e NOTCKD para pacientes sem a doença. Para uma visão mais detalhada dos atributos contidos no *dataSet*, a Tabela 1 lista todas as atributo disponíveis, proporcionando uma visão abrangente das informações que compõem o conjunto de dados. Essa tabela servirá como referência para uma análise mais aprofundada e a construção de modelos preditivos.

---

<sup>1</sup>Disponível em: <https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease>

Tabela 1: Atributos do *dataSet* utilizado.

Siglas	Atributos	Tipo
id	identidade	-
age	idade	Numérica
bp	pressão arterial	Numérica
sg	gravidade específica	Nominal
al	albumina	Nominal
su	açúcar	Nominal
rbc	glóbulos vermelhos	Nominal
pc	células de pus	Nominal
pcc	aglomerados de células de pus	Nominal
ba	bactérias	Nominal
bgr	glicose sanguínea aleatória	Numérica
bu	ureia sanguínea	Numérica
sc	creatinina sérica	Numérica
sod	sódio	Numérica
pot	potássio	Numérica
hemo	hemoglobina	Numérica
pcv	volume de células empacotadas	Numérica
wc	contagem de glóbulos brancos	Numérica
rc	contagem de glóbulos vermelhos	Numérica
htn	hipertensão	Nominal
dm	diabetes mellitus	Nominal
cad	doença arterial coronária	Nominal
appet	apetite	Nominal
pe	edema de membros inferiores	Nominal
ane	anemia	Nominal
class	classe	Nominal

### 3.4 Tratamento de Dados

O tratamento de dados, também conhecido como pré-processamento de dados, é uma etapa essencial no processo de análise de dados e aprendizado de máquina. Ele envolve uma série de técnicas e procedimentos para manipular, limpar e transformar os dados brutos coletados, a fim de torná-los adequados e utilizáveis para análises e modelagem.

Durante o tratamento de dados, é comum lidar com problemas comuns encontrados nos dados, como ruído, inconsistências, valores ausentes, discrepâncias e *outliers*. Esses problemas podem afetar negativamente a análise e o desempenho dos modelos de aprendizado de máquina, portanto, é necessário realizar etapas de tratamento para garantir a qualidade e a confiabilidade dos dados Géron (2019).

Uma das etapas fundamentais do tratamento de dados é a limpeza, que envolve a identificação e correção de erros nos dados. Isso pode incluir a remoção de registros incompletos, a imputação de valores faltantes ou a aplicação de técnicas estatísticas para corrigir inconsistências. Além disso, a normalização dos dados é realizada para eliminar diferenças de unidades ou magnitudes, tornando os dados comparáveis Géron (2019).

A transformação de dados é outra etapa importante, em que operações matemáticas ou

estatísticas são aplicadas para alterar a distribuição dos dados. Isso pode envolver transformações logarítmicas, exponenciais ou padronização, entre outras técnicas. A remoção de *outliers* também é comumente realizada para eliminar valores atípicos que podem distorcer a análise ou o desempenho do modelo Géron (2019).

Além disso, a seleção de atributos é realizada para identificar quais atributo dos dados são mais relevantes para a análise ou o modelo de aprendizado de máquina. Isso ajuda a reduzir a dimensionalidade dos dados, eliminando atributos desnecessários ou redundantes, melhorando a eficiência do processo de análise e reduzindo a ~~geron2019~~complexidade dos modelos Géron (2019).

Outra consideração é a amostragem dos dados, que pode ser necessária quando o conjunto de dados é muito grande. A amostragem envolve a seleção de uma subamostra representativa dos dados para análise, economizando tempo e recursos computacionais Géron (2019).

Em resumo, o tratamento de dados é uma etapa fundamental para garantir que os dados estejam limpos, consistentes, completos e prontos para análises e modelagem. As técnicas de tratamento incluem limpeza, normalização, transformação de dados, remoção de *outliers*, seleção de atributos e amostragem. Um tratamento adequado dos dados é essencial para obter resultados confiáveis e relevantes em pesquisas, análises de dados e aplicação de algoritmos de aprendizado de máquina Géron (2019).

Ao abordar o tratamento de dados, o Géron (2019) enfatiza a importância de preparar os dados adequadamente antes de aplicar técnicas de aprendizado de máquina. O pré-processamento de dados é uma etapa crítica para garantir que os dados estejam limpos, coerentes e prontos para serem utilizados em análises e modelos.

### 3.5 Acurácia

A acurácia é uma métrica fundamental no campo do aprendizado de máquina, utilizada para avaliar o desempenho de modelos de classificação. Ela representa a capacidade de um modelo em fazer previsões corretas em relação ao número total de previsões realizadas Géron (2019).

Em essência, a acurácia mede a proporção de previsões corretas feitas por um modelo em relação ao total de previsões efetuadas. Quanto maior a acurácia, mais preciso é o modelo em suas previsões. Para calcular a acurácia, divide-se o número de previsões corretas pelo número total de previsões e multiplica-se por 100 para obter a porcentagem Helmenstine (2020).

A acurácia é especialmente importante em tarefas de classificação, onde o objetivo é atribuir rótulos ou categorias a diferentes instâncias de dados. Modelos com alta acurácia são desejáveis, pois indicam a capacidade do modelo em distinguir eficazmente entre as diferentes classes ou categorias Helmenstine (2020).

No entanto, é importante destacar que a acurácia pode não ser a métrica mais apropriada em todos os cenários, principalmente quando as classes estão desbalanceadas, ou seja, quando uma classe tem muito mais exemplos do que outra. Nesses casos, a acurácia pode ser enganosa, uma vez que um modelo pode alcançar alta acurácia simplesmente prevendo a classe majoritária o tempo todo. Portanto, em situações desse tipo, outras métricas, como precisão, *recall* e *F1-score*, podem ser mais informativas para avaliar o desempenho do modelo Pádua (2020).

Em resumo, a acurácia é uma métrica-chave que mede a precisão das previsões feitas por modelos de classificação, sendo amplamente utilizada para avaliar o quão eficaz um modelo é na tarefa de classificação. No entanto, sua interpretação deve ser feita com cuidado, especialmente em situações de desbalanceamento de classes, onde outras métricas podem ser mais apropriadas para uma avaliação completa do modelo.

### 3.6 *Boosting*

O *Boosting* é uma técnica de aprendizado de máquina que visa melhorar o desempenho dos modelos de classificação ou regressão, combinando a predição de vários modelos mais fracos. Em outras palavras, o *Boosting* é um método de *ensembles* que usa vários modelos fracos para formar um modelo forte Faceli et al. (2021).

A ideia principal do *Boosting* é identificar os exemplos que são mais difíceis de classificar e fornecê-los para o modelo seguinte com mais ênfase. Com isso, o modelo seguinte dá mais atenção a esses exemplos e, assim, pode melhorar a sua capacidade de classificação Faceli et al. (2021).

O processo de *Boosting* é realizado em várias iterações, onde em cada iteração um novo modelo fraco é adicionado ao conjunto de modelos. Os exemplos que foram classificados incorretamente pelo conjunto de modelos anterior são ponderados com mais importância para a próxima iteração, para que o novo modelo seja treinado com mais atenção nesses exemplos Faceli et al. (2021).

Ao final das iterações, os resultados dos modelos fracos são combinados para produzir um modelo forte que é capaz de fazer uma previsão precisa. O modelo forte é construído com base na soma ponderada dos resultados dos modelos fracos, onde cada modelo fraco contribui com uma ponderação de acordo com a sua precisão Géron (2019).

Em resumo, o *Boosting* é uma técnica poderosa de aprendizado de máquina que ajuda a melhorar a precisão dos modelos, combinando a previsão de vários modelos mais fracos. Ele é particularmente útil quando se trabalha com conjuntos de dados complexos e grandes, onde outros métodos podem não ser eficazes.

### 3.7 AdaBoost

O Adaboost, ou *Adaptive Boosting*, é uma técnica de aprendizado de máquina que utiliza um conjunto de modelos de aprendizado fracos para gerar um modelo forte. O objetivo do Adaboost é melhorar o desempenho de classificadores fracos por meio de combinações ponderadas desses modelos Wibawa et al. (2017).

O Adaboost funciona por meio de iterações, onde cada iteração é um processo de treinamento de um novo classificador fraco. Na primeira iteração, um classificador simples e fraco é treinado no conjunto de dados de treinamento. Em seguida, os exemplos que foram classificados incorretamente pelo classificador são ponderados com mais importância e um novo classificador fraco é treinado, dando mais atenção a esses exemplos Géron (2019).

O processo de iteração continua até que um conjunto de classificadores fracos tenha sido criado. Em cada iteração, os exemplos são ponderados de forma adaptativa com base em como foram classificados nas iterações anteriores. Isso significa que, à medida que novos classificadores fracos são adicionados ao conjunto, os exemplos que foram classificados incorretamente em iterações anteriores são ponderados com mais importância, e os exemplos que foram classificados corretamente são ponderados com menos importância Géron (2019).

Ao final das iterações, o Adaboost combina os resultados dos classificadores fracos em um modelo forte por meio de uma média ponderada. A ponderação é determinada pela precisão de cada classificador fraco. Os classificadores que são mais precisos recebem uma ponderação maior, enquanto os menos precisos recebem uma ponderação menor Géron (2019).

O Adaboost é uma técnica poderosa que pode produzir modelos precisos, mesmo quando se trabalha com conjuntos de dados complexos. Ele é particularmente útil para problemas

de classificação binária, mas também pode ser adaptado para problemas de regressão Geron (2019).

### 3.8 Parâmetros do AdaBoost

O AdaBoost é um algoritmo de aprendizado de máquina que faz parte da família de métodos de conjunto (*ensemble methods*). Ele combina várias “máquinas fracas” (classificadores fracos) para formar um classificador forte. Os parâmetros do AdaBoost são essenciais para o controle e ajuste do comportamento do algoritmo. Abaixo, descrevem-se os principais parâmetros do AdaBoost<sup>2</sup> :

- **Número de Estimadores (n\_estimators):** Este parâmetro define o número de estimadores (classificadores fracos) que serão combinados para formar o classificador final. A escolha do número adequado de estimadores pode afetar significativamente o desempenho do modelo. Um valor muito baixo pode resultar em subajuste, enquanto um valor muito alto pode levar a sobreajuste.
- **Taxa de Aprendizado (learning\_rate):** A taxa de aprendizado controla a contribuição de cada estimador para o modelo final. Valores mais altos aumentam a influência de cada estimador, tornando o modelo mais complexo. Valores mais baixos reduzem a influência, tornando o modelo mais robusto, mas potencialmente mais lento para convergir.
- **Estimador Base (base\_estimator):** O AdaBoost utiliza classificadores fracos como estimadores base. O estimador base pode variar, mas a escolha comum é a Árvore de Decisão com profundidade limitada. Esse parâmetro permite personalizar o classificador fraco utilizado.
- **Semente Aleatória (random\_state):** Este parâmetro permite definir uma semente aleatória para garantir a reprodutibilidade dos resultados. A mesma semente aleatória produzirá os mesmos resultados sempre que o modelo for treinado.

A escolha adequada desses parâmetros é fundamental para o desempenho e comportamento do AdaBoost em tarefas de classificação.

---

<sup>2</sup>Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html>



## 4 DESENVOLVIMENTO

Esta seção apresenta os fundamentos teóricos, vistos como os materiais para o desenvolvimento do estudo proposto.

### 4.1 *dataSet*

O conjunto de dados utilizados para esse estudo foi retirado do repositório UCI Machine Learning<sup>3</sup> - uma coleção de bancos de dados, teorias de domínio e geradores de dados que são usados pela comunidade de Aprendizado de Máquina para a análise empírica de algoritmos de Aprendizado de Máquina. Os dados foram coletados ao longo de um período de 2 meses na Índia com 25 atributo. O alvo é a 'classificação', que é 'ckd' ou 'notckd' - ckd = doença renal crônica. Há 400 linhas. Os dados precisaram ser limpos: ou seja, eles possuíam valores NaN (*Not A Number*) e as atributo numéricas precisaram ser convertidas para o tipo de dado *floats*.

### 4.2 Tratamento dos dados

Após analisar a descrição dos dados, foram realizados cálculos para avaliar a distorção das variáveis e identificar seus tipos, como *float*, *object* e *int*. Para essa análise, utilizou-se a função `info()` do *pandas*<sup>4</sup>, que fornece informações detalhadas sobre o *dataframe* (representação do *dataSet* no *scikit-learn*), incluindo o tipo de dado de cada coluna.

Em seguida, procedeu-se ao tratamento de ruídos nos dados, como valores ausentes e informações incorretas, que poderiam comprometer a análise. Para essa finalidade, empregou-se a substituição de valores por caracteres mais apropriados, como transformar “/tyes” em “yes”. Essa operação foi realizada por meio da função `replace()` do *pandas*, que substitui valores em um *dataframe* por outros especificados.

No caso de valores NaN, que não puderam ser excluídos devido à relevância das informações que eles continham, optou-se por substituir os valores numéricos pela mediana e os valores categóricos pela moda. Essa transformação foi realizada utilizando a função `fillna()` do *pandas*, que preenche valores ausentes com valores apropriados.

Para converter todas as variáveis categóricas em numéricas, um dicionário foi criado, permitindo a substituição de “yes” por 1 e “no” por 0, transformando esses valores em inteiros. Essa conversão foi efetuada por meio da função `map()` do *pandas*, que aplica uma função a cada valor de uma série.

A fim de aprimorar ainda mais a qualidade dos dados e prepará-los para análise, foram realizadas etapas adicionais de pré-processamento. O processo iniciou-se com a remoção da coluna “id” dos dados, uma vez que essa coluna apenas indicava números de linhas e não fornecia informações relevantes para a análise. Essa operação de exclusão foi executada utilizando a função `drop()` do *pandas*.

Vale ressaltar que todo esse processo de pré-processamento foi conduzido no ambiente do *Google Colaboratory*<sup>5</sup>, que é uma plataforma baseada em Jupyter *notebooks* executada na nuvem do Google, proporcionando uma configuração simplificada e facilidade de uso.

---

<sup>3</sup><https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease>

<sup>4</sup><https://pandas.pydata.org/>

<sup>5</sup><https://colab.research.google.com/>

### 4.3 Análise de Correlação

Após a limpeza inicial dos dados, realizou-se uma análise de correlação para entender as relações entre as diferentes atributo do conjunto de dados. A correlação é uma medida estatística que indica o grau de associação entre duas variáveis. É essencial para identificar quais atributo podem ter maior influência na classificação de doença renal crônica.

Para visualizar a matriz de correlação, utilizou-se a biblioteca Seaborn<sup>6</sup> para criar um mapa de calor (*heatmap*) que destaca as relações entre as variáveis. O código abaixo mostra como gerou-se o mapa de calor:

```
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(15, 15))
data_corr = data.corr()
sns.heatmap(data_corr,
            xticklabels=data_corr.columns.values,
            yticklabels=data_corr.columns.values,
            annot=True);
```

O mapa de calor permite identificar rapidamente quais atributo estão fortemente correlacionadas e quais podem ter um impacto significativo na classificação da doença renal crônica. Isso é crucial para selecionar as atributo mais relevantes e criar modelos de aprendizado de máquina eficazes. A Figura 1 mostra o mapa de calor obtido.

---

<sup>6</sup><https://seaborn.pydata.org>

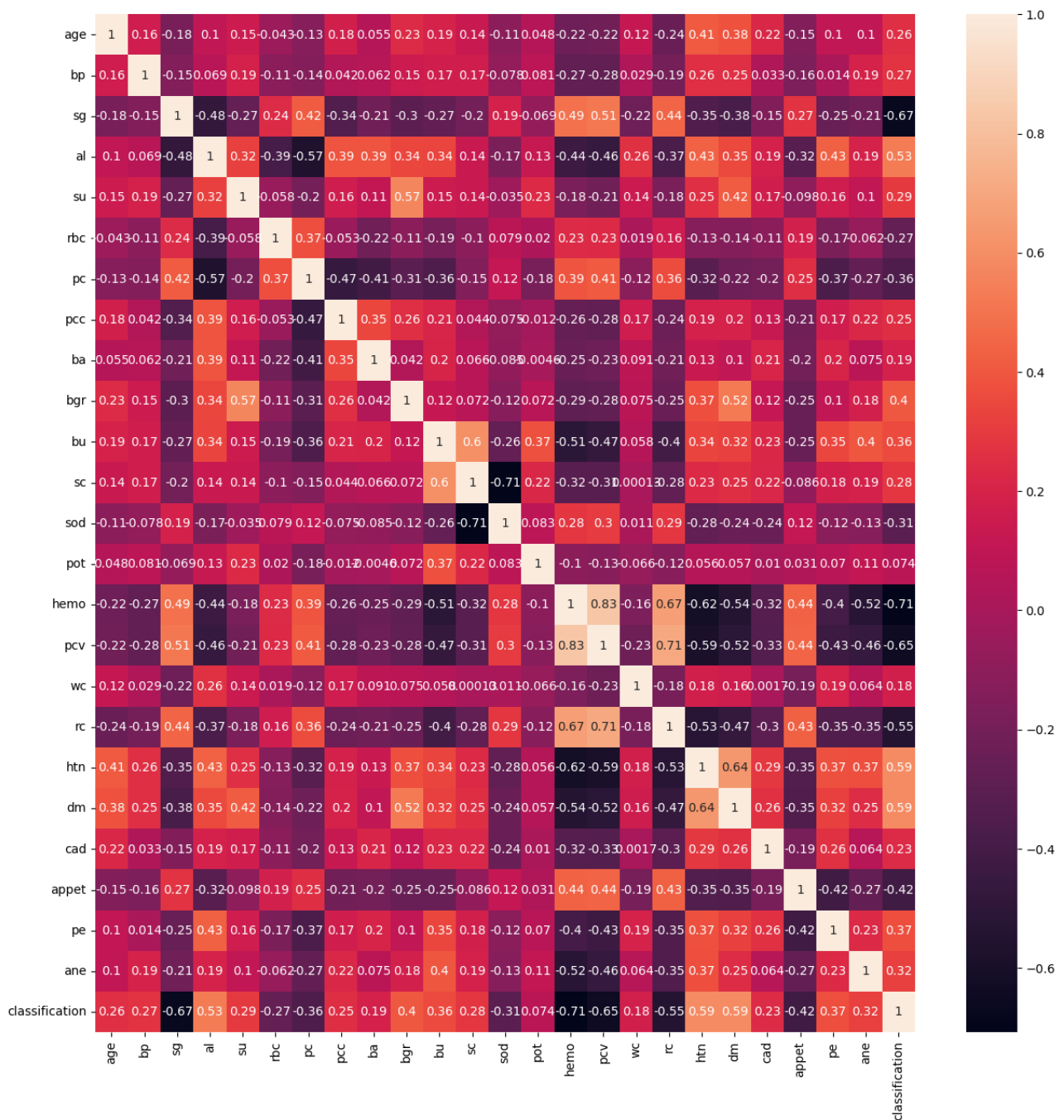


Figura 1: Mapa de Calor da Correlação dos Dados.

Nas análises de correlação realizadas, foram identificadas várias associações significativas entre atributo específicas do conjunto de dados relacionadas à detecção de doença renal crônica. Notavelmente, algumas dessas associações demonstraram correlações particularmente fortes, com coeficientes superiores a 0.6. Essas descobertas são fundamentais para compreender os fatores que podem influenciar a ocorrência ou o diagnóstico dessa condição médica.

Em primeiro lugar, destacam-se a forte correlação entre as atributo “pc” (contagem de plaquetas) e “hemo” (hemoglobina), que apresentaram um coeficiente de correlação de 0.82. Isso sugere uma relação direta entre essas duas variáveis, indicando que mudanças na contagem de plaquetas podem estar intimamente relacionadas às variações nos níveis de hemoglobina.

Além disso, observam-se correlações significativas entre outras duplas de atributo, como “rc” (contagem de glóbulos vermelhos) e “hemo” (coeficiente de correlação de 0.67), “rc” e “pcv” (hematócrito, coeficiente de correlação de 0.72), “dm” (diabetes mellitus) e “htn” (hipertensão, coeficiente de correlação de 0.64), bem como “hemo” e “htn” (coeficiente de correlação de 0.61). Essas associações podem fornecer *insights* importantes sobre as interações entre diferentes atributo que podem desempenhar um papel na detecção precoce da doença renal crônica.

Em resumo, as correlações identificadas entre essas atributo ressaltam a complexidade da relação entre os fatores médicos envolvidos na detecção da doença renal crônica. Essas descobertas podem orientar futuras investigações e modelagens, ajudando a desenvolver abordagens mais eficazes para a detecção precoce e o tratamento dessa condição, potencialmente melhorando a qualidade de vida dos pacientes afetados. É importante notar que essas conclusões são baseadas em análises estatísticas e podem fornecer um ponto de partida valioso para investigações mais aprofundadas na área médica.

## 4.4 Pré-processamento dos Dados

Após a análise de correlação, avançou-se para a preparação dos dados para o treinamento do modelo de Aprendizado de Máquina. O primeiro passo foi criar uma cópia dos dados originais para preservar a integridade do conjunto de dados original. Em seguida, realizaram-se as seguintes etapas:

1. Dividiu-se o conjunto de dados em duas partes: uma com as atributo (variáveis independentes) e outra com as classificações (variável dependente).
2. Aplicou-se a normalização dos dados usando o MinMaxScaler para garantir que todas as atributo estejam na mesma escala. Isso é importante para muitos algoritmos de aprendizado de máquina.
3. Dividiu-se os dados em conjuntos de treinamento (70%) e teste (30%) usando a função ‘train\_test\_split’ do *scikit-learn*. Isso permite avaliar o desempenho do modelo em dados não vistos.

Aqui estão as dimensões dos conjuntos resultantes:

- Tamanho do conjunto de treinamento (x treino): (196, 24) (Linha, Coluna)
- Tamanho do conjunto de teste (x teste): (84, 24) (Linha, Coluna)
- Tamanho das classificações de treinamento (y treino): (196,) (Linha, Coluna)
- Tamanho das classificações de teste (y teste): (84,) (Linha, Coluna)

Lembrando que “x” representa o conjunto de atributo sem a classificação, e “y” representa a classificação.

## 4.5 Treinamento do Modelo AdaBoost

Com os dados devidamente preparados, procedeu-se ao treinamento do modelo AdaBoost para a detecção de doença renal crônica. Utilizou-se o *scikit-learn* para implementar o AdaBoostClassifier, um algoritmo de aprendizado de máquina que combina múltiplos classificadores fracos para construir um classificador forte.

Definiu-se diferentes combinações de parâmetros que se desejava testar, incluindo o número de estimadores (`n_estimators`), taxa de aprendizado (`learning_rate`), profundidade máxima da árvore (`max_depth`) e o algoritmo (`algorithm`). Cada combinação de parâmetros foi treinada e avaliada quanto à sua precisão.

## 5 Resultados

Nossa pesquisa na área de detecção de doença renal crônica empregou técnicas avançadas de aprendizado de máquina por meio do algoritmo AdaBoost com variações nos parâmetros. Esta investigação teve como objetivo avaliar o desempenho do modelo em diferentes cenários e parâmetros, fornecendo *insights* valiosos para futuros desenvolvimentos na detecção precoce de doença renal.

Os resultados das diversas combinações de parâmetros revelaram informações cruciais sobre a capacidade preditiva do modelo. Abaixo, apresenta-se um resumo dos principais resultados:

- **Número de Estimadores (`n_estimators`):** Observou-se que o aumento no número de estimadores não necessariamente resultou em um aumento significativo na acurácia do modelo. Com 50 estimadores, o modelo alcançou uma acurácia de aproximadamente 98.8%, demonstrando uma excelente capacidade de discriminação.
- **Taxa de Aprendizado (`taxa_aprendizado`):** A variação da taxa de aprendizado revelou que valores mais baixos (0.1 e 0.5) resultaram em uma acurácia inferior em comparação com a taxa de aprendizado de 1.0. Isso sugere que, para o nosso conjunto de dados, uma taxa de aprendizado mais alta foi mais eficaz.
- **Profundidade Máxima da Árvore (`profundidade_maxima`):** Ao ajustar a profundidade máxima da árvore base, nota-se que uma profundidade de 2 proporcionou resultados superiores em comparação com profundidades menores (1 e 3), indicando que um modelo ligeiramente mais complexo era benéfico para a detecção da doença renal.
- **Semente Aleatória (`estadoestado_aleatório`):** A semente aleatória foi mantida constante para garantir a reprodutibilidade dos resultados.

Os resultados demonstraram que o modelo AdaBoost com configurações específicas mostrou ser altamente eficaz na detecção de doença renal crônica em nosso conjunto de dados. A combinação de 50 estimadores, taxa de aprendizado de 1.0 e profundidade máxima de 2 do qual resultou na melhor acurácia, atingindo aproximadamente 98.8%.

Esses resultados destacam a importância do ajuste fino de parâmetros no desenvolvimento de modelos de aprendizado de máquina para aplicações médicas. Essas descobertas têm o potencial de contribuir significativamente para o avanço na detecção precoce de doença renal crônica e, consequentemente, para melhorar a qualidade de vida dos pacientes.

Tabela 2: Resultados do AdaBoost

n_estimadores	taxa_aprendizado	profundidade_máxima	estado_aleatório	precisão
50	1,0	1	42	0,976190
50	0,5	2	0	0,988095
50	0,1	3	123	0,916667
50	1,0	1	42	0,988095
50	0,5	2	0	0,940476
50	0,1	3	123	0,976190
100	1,0	1	42	0,976190
100	0,5	2	0	0,988095
100	0,1	3	123	0,916667
100	1,0	1	42	0,988095
100	0,5	2	0	0,940476
100	0,1	3	123	0,940476
200	1,0	1	42	0,976190
200	0,5	2	0	0,940476
200	0,1	3	123	0,940476
200	1,0	1	42	0,988095
200	0,5	2	0	0,976190
200	0,1	3	123	0,916667

## 6 Conclusão

O objetivo principal desta Iniciação Científica foi explorar e aplicar uma técnica de Aprendizado de Máquina na tarefa de classificação de dados, com foco na previsão de doença renal crônica. Para alcançar esse objetivo, uma série de atividades específicas foi desenvolvida ao longo do projeto.

Inicialmente, realizou-se uma pesquisa abrangente sobre o conjunto de dados relevante para a detecção da doença renal crônica. Essa etapa foi essencial para compreender a natureza dos dados e as atributos que poderiam influenciar na construção de um modelo de classificação eficaz.

Em seguida, foram realizadas atividades de exploração e análise dos dados, visando identificar padrões, tendências e possíveis correlações entre as diferentes variáveis presentes no conjunto de dados. A limpeza e manipulação dos dados também foram parte fundamental do processo, garantindo a qualidade e consistência dos dados a serem utilizados no modelo.

Um estudo aprofundado sobre o método de Aprendizado de Máquina conhecido como AdaBoost foi conduzido. Isso incluiu a compreensão de seus princípios teóricos, funcionamento e aplicabilidade em tarefas de classificação.

Quanto aos resultados obtidos, a aplicação do AdaBoost na tarefa de classificação da doença renal crônica demonstrou um desempenho notável. O modelo alcançou uma acurácia de 98.81%, o que indica sua eficácia na previsão dessa condição médica. Essa alta taxa de acurácia é um indicativo promissor da capacidade do modelo em distinguir entre pacientes com e sem doença renal crônica com base nos dados disponíveis.

É importante ressaltar que a acurácia não é a única métrica relevante na avaliação de um modelo de Aprendizado de Máquina. No entanto, no contexto desta pesquisa, a alta acurácia alcançada sugere que o modelo AdaBoost é capaz de fazer previsões precisas e confiáveis. É

fundamental considerar que a detecção precoce e precisa da doença renal crônica pode levar a intervenções médicas mais eficazes e, conseqüentemente, à melhoria da qualidade de vida dos pacientes.

Os resultados também revelaram correlações significativas entre algumas variáveis do conjunto de dados, o que contribuiu para o desempenho do modelo. Essas correlações podem fornecer *insights* valiosos para profissionais de saúde, ajudando a identificar fatores que podem estar associados ao desenvolvimento da doença renal crônica.

Além disso, a pesquisa abriu portas para uma série de trabalhos futuros promissores. Uma possível extensão deste estudo poderia ser a exploração de outras técnicas de Aprendizado de Máquina, a fim de comparar seu desempenho com o modelo AdaBoost. A coleta de um conjunto de dados ainda mais diversificado e abrangente também pode ser considerada, com o objetivo de melhorar ainda mais a capacidade de previsão do modelo.

Outras métricas de avaliação, como sensibilidade, especificidade e F1-score<sup>7</sup> O F1-score varia de 0 a 1, onde 1 indica um desempenho perfeito do modelo em termos de precisão e *recall*. podem ser exploradas para obter uma compreensão mais completa do desempenho do modelo em diferentes aspectos. A disseminação dos resultados por meio de artigos científicos, apresentações em conferências e colaborações com profissionais de saúde é fundamental para ampliar o impacto dessa pesquisa na comunidade científica e médica.

Em suma, os resultados alcançados até o momento demonstram o potencial do modelo AdaBoost na detecção da doença renal crônica e abrem perspectivas empolgantes para futuras pesquisas e melhorias na área de diagnóstico médico assistido por máquina. A alta acurácia obtida é um indicativo promissor de que técnicas de Aprendizado de Máquina podem desempenhar um papel crucial na melhoria da saúde e bem-estar dos pacientes.

---

<sup>7</sup>O F1-score é uma métrica de avaliação amplamente utilizada em problemas de classificação, especialmente quando as classes não estão balanceadas. Ele é calculado com base na precisão e no *recall* do modelo de classificação e fornece uma medida geral de desempenho. Essa métrica é especialmente útil quando é importante encontrar um equilíbrio entre a capacidade de um modelo de classificação em identificar positivos verdadeiros (verdadeiros positivos) e evitar falsos positivos.

## Referências

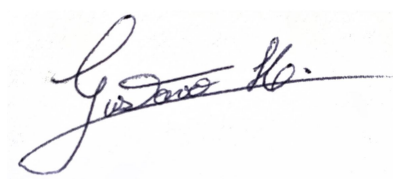
- Bi, Q., Goodman, K. E., Kaminsky, J., and Lessler, J. (2019). What is machine learning? a primer for the epidemiologist. *American Journal of Epidemiology*, 188(12):2222–2239.
- Buch, V., Ahmed, I., and Maruthappu, M. (2018). *Artificial intelligence in medicine: Current trends and future possibilities*. Br. J. Gen. Pract.
- Faceli, K., Lorena, A. C., Gama, J., Almeida, T. A. d., and Carvalho, A. C. P. d. L. F. d. (2021). *Inteligência artificial: uma abordagem de aprendizado de máquina*. LTC.
- Géron, A. (2019). *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow*. Alta Books.
- Helmenstine, Anne Marie, P. (2020). Accuracy definition in science.
- IBM (2023). <https://www.ibm.com/br-pt/cloud/learn/what-is-artificial-intelligence>.
- Pádua, M. (2020). Machine learning -métricas de avaliação: Acurácia, precisão e recall, f1-score.
- Wibawa, M. S., Maysanjaya, I. M. D., and Putra, I. M. A. W. (2017). Boosted classifier and features selection for enhancing chronic kidney disease diagnose. In *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, pages 1–6.



## 7 AVALIAÇÃO DO ORIENTADOR SOBRE O DESEMPENHO DO ORIENTADO

O acadêmico apresentou um bom desempenho durante a execução do projeto de IC, sendo assíduo e pontual às reuniões de orientação. Cumpriu com os objetivos propostos no projeto e realizou todas as tarefas propostas adequadamente, adquirindo o conhecimento técnico e científico esperado.

Guarapuava, 17 de outubro de 2023.

A handwritten signature in blue ink, appearing to read 'Gustavo Ho.', is written over a light blue rectangular background.

Aluno

A handwritten signature in blue ink, appearing to read 'Carolina Paula de Aguiar', is written over a light blue rectangular background.

Orientadora