

Análise Comparativa de Modelos de Machine Learning na Detecção da Doença Renal Crônica

Marcos Schäfer Vrielink¹, William Moraes da Silva¹

¹Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul - Campus Farroupilha

Farroupilha - RS - Brasil

marco.s.v1928@gmail.com, william.silva@farroupilha.ifrs.edu.br

RESUMO. *A Doença Renal Crônica é caracterizada pela perda progressiva da função renal. Este trabalho avalia a aplicação de algoritmos de aprendizado de máquina para detecção da patologia. Foram comparados cinco modelos supervisionados: Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine, Naive Bayes e XGBoost. Dentre eles, o modelo KNN apresentou os piores resultados, com uma acurácia média de 41%. Por outro lado, o modelo XGBoost obteve o melhor desempenho, com uma acurácia média de 92%, f1-score de 96% e sensibilidade de 98%. Os resultados indicam que o modelo XGBoost é indicado em comparação aos demais analisados para auxiliar na classificação da Doença Renal Crônica.*

ABSTRACT. *Chronic Kidney Disease is characterized by progressive loss of kidney function. This study evaluates the application of machine learning algorithms to detect the pathology. Five supervised models were compared: Random Forest, K-Nearest Neighbors, Support Vector Machine, Naive Bayes and XGBoost Classifier. Among them, the KNN model showed the worst performance, with an average accuracy of 41%. On the other hand, the XGBoost model achieved the best performance, with an average accuracy of 92%, f1-score of 96% and sensitivity of 98%. The results indicate that the XGBoost model is recommended compared to the others analyzed to assist in the classification of Chronic Kidney Disease.*

1. Introdução

A Doença Renal Crônica (DRC) é uma condição de caráter progressivo e sem cura, sendo associada a altos índices de morbidade e mortalidade. Ela é mais comum entre adultos, especialmente aqueles que apresentam diabetes ou hipertensão. Nos estágios mais avançados, conhecidos como insuficiência renal crônica terminal (IRCT), os rins deixam de regular adequadamente o equilíbrio do meio interno do organismo, podendo exigir intervenções como diálise ou transplante renal (Junior, 2004; Kalantar, 2021).

A DRC é reconhecida como um problema de saúde pública global. No Brasil, observa-se um aumento na incidência e prevalência da falência da função renal, com um prognóstico ainda desfavorável e custos elevados para o tratamento. Devido à progressão silenciosa da doença, a criação de um modelo preditivo confiável para identificar o risco de evolução para a doença renal terminal em estágios iniciais da DRC pode ter grande relevância clínica (Bai et al, 2022; Bastos et al, 2009).

Nos últimos anos, a inteligência artificial (IA) tem sido cada vez mais aplicada na área da saúde, auxiliando no diagnóstico e prognóstico de diversas doenças. Modelos preditivos baseados em aprendizado de máquina podem analisar grandes volumes de dados clínicos e identificar padrões sutis que indicam o risco de desenvolvimento de doenças. Dessa forma, essas ferramentas podem complementar a avaliação médica, permitindo intervenções mais rápidas e personalizadas (Gobbo, 2019).

Visto isso, o presente trabalho tem como objetivo desenvolver e comparar modelos preditivos baseados em aprendizado de máquina no diagnóstico de Doença Renal Crônica. Essa abordagem pode contribuir para selecionar o melhor modelo para o diagnóstico de DRC, permitindo o auxílio ao diagnóstico precoce da patologia e a implementação de algoritmos preventivos para retardar a progressão da DRC.

Este trabalho está dividido da seguinte forma: na seção 2 se encontra o referencial teórico dos temas necessários para o desenvolvimento do estudo; na seção 3 é apresentada a metodologia a ser utilizada no desenvolvimento do trabalho bem como a origem dos dados; na seção 4 são apresentados os resultados obtidos e na última seção, 5, são apresentadas as considerações finais.

2. Referencial Teórico

A fim de fundamentar a metodologia utilizada, este capítulo apresenta conceitos teóricos para melhor compreensão dos assuntos relacionados à elaboração da pesquisa.

2.1 Doença Renal Crônica

A DRC é caracterizada por uma lesão renal e pela perda progressiva e irreversível das funções dos rins, onde no seu estágio mais avançado, também conhecido como fase terminal da IRC, os rins perdem a capacidade de manter a homeostase do paciente. É uma condição de caráter progressivo e sem cura, sendo associada a altos índices de morbidade e mortalidade (Junior, 2004; Kalantar, 2021).

Tecnicamente, a DRC é caracterizada pela National Kidney Foundation (NKF) (2002) através dos seguintes critérios:

- Lesão presente por um período igual ou superior a 3 meses, definida por anormalidades estruturais ou funcionais do rim, com ou sem diminuição da taxa de filtração glomerular (TFG), manifestada por anormalidades patológicas ou marcadores de lesão renal, incluindo alterações sanguíneas ou urinárias, ou nos exames de imagem;
- TFG <60 mL/min/1,73 m², por um período maior ou igual a 3 meses, com ou sem lesão renal.

Baseado nesta definição, a NKF também classifica a doença em 5 estágios descritos no Quadro 1:

Quadro 1: Estágios da DRC

Fonte: National Kidney Foundation, 2002

Estágio	Descrição	TFG (mL/min/1,73 m²)
1	Lesão renal com TFG normal ou aumentado	≥ 90
2	Lesão renal com leve diminuição do TFG	60 - 89
3	Lesão renal com moderada diminuição do TFG	30 - 59
4	Lesão renal com aumentada diminuição do TFG	15 - 29
5	Falência renal funcional ou em Terapia Renal Substitutiva	< 15

O número de pacientes com DRC está aumentando em todo o mundo. De acordo com informações da Sociedade Brasileira de Nefrologia, as doenças renais crônicas afetam cerca de 850 milhões de pessoas no mundo. Nos países desenvolvidos estima-se que a prevalência da DRC, entre os adultos, seja em torno de 10 a 13%. No Brasil, a prevalência de pacientes por milhão de pessoas (pmp), foi de 665 pmp em 2019 e de 771 pmp em 2023 (Marinho, 2017).

Segundo Gordan (2006), quatro grupos de pacientes apresentam maior risco de desenvolver a DRC e devem ser submetidos a exames para verificar a presença de lesão renal anualmente, onde em primeiro lugar há os portadores de hipertensão arterial, sendo responsável por aproximadamente 40% dos casos. Por se tratar de um fator de risco modificável, seu tratamento adequado pode reduzir significativamente tanto a mortalidade cardiovascular quanto a velocidade de progressão da DRC (Gordan, 2006; Junior, 2004).

Os portadores de diabetes também representam um grupo expressivo, sendo a principal causa de DRC em países desenvolvidos, correspondendo a cerca de 45% dos casos. No Brasil, sua prevalência varia entre 18% na Região Norte e 27% na Região Sul, sendo a segunda principal causa de DRC. Além disso, nos últimos dez anos, a taxa de prevalência aumentou de 10% para aproximadamente 25%, o que evidencia o crescimento da doença (Gordan, 2006; Sociedade Brasileira de Nefrologia, 2006). Nesse contexto, Vieira (2019) explica:

Os altos níveis de açúcar fazem com que os rins filtrem muito sangue, sobrecarregando os órgãos e levando à perda de proteínas na urina. Com o tempo e o excesso de resíduos no sangue, a sobrecarga faz com que os rins percam a capacidade de filtração e venham a falhar.

Os pacientes com doenças cardiovasculares também apresentam um risco maior de desenvolver a DRC. Pacientes que passam por procedimentos radiológicos e cirúrgicos no sistema cardiovascular, assim como aqueles incluídos em ensaios clínicos sobre doença aterosclerótica, apresentam um risco aumentado de desenvolver DRC (Gordan, 2006; Shlipaket et al, 2002). Segundo McClellan (2004), uma parcela significativa desses pacientes, incluindo 60% dos indivíduos com Insuficiência Cardíaca e 52% daqueles que sofreram Infarto do Miocárdio e receberam alta pelo sistema Medicare nos Estados Unidos, apresentava uma taxa de filtração glomerular inferior a

60 ml/min. Esses dados sugerem que a associação entre doença cardiovascular e DRC eleva o risco de progressão para a doença.

Por fim, os familiares de pacientes com DRC também são considerados um grupo de risco. A relação entre Doença Renal Crônica Terminal (DRCT) e histórico familiar de nefropatia foi evidenciada por Ferguson, que identificou uma prevalência de Doença Renal (DR) de 26% entre familiares de primeiro e segundo grau de pacientes em Terapia Renal Substitutiva (TRS), em comparação com 11% na população em geral (Gordan, 2006; McClellan, 2004).

2.1.1 Diagnóstico da Doença Renal Crônica

A DRC pode não apresentar sintomas nas fases iniciais, fazendo com que os médicos fiquem atentos, especialmente em pacientes com maior risco. Fatores como a TFG são essenciais para identificar e classificar a doença precocemente (Bastos et al, 2011).

A TFG é a melhor medida geral da função renal e a mais facilmente compreendida pelos médicos e pacientes. Ela é definida como a capacidade dos rins de eliminar uma substância do sangue e é expressa como o volume de sangue que é completamente depurado em uma unidade de tempo. Clinicamente, o método mais utilizado para obter informações sobre a TFG é a depuração de creatinina, com coleta de urina ao longo de 24 horas, no qual a excreção de creatinina urinária em 24 horas é dividida pela concentração de creatinina sérica (Bastos et al, 2011).

Em geral, o diagnóstico da doença renal se baseia na avaliação da história clínica, exame físico e exames laboratoriais simples, como urinálise com análise microscópica, dosagem de eletrólitos, ureia, creatinina, fosfato, cálcio e hemograma. Em alguns casos, pode ser necessário realizar exames sorológicos específicos. A distinção entre lesão renal aguda e DRC é feita, principalmente, pela observação de um aumento recente e significativo da creatinina sérica ou por alterações na urinálise (Malkina, 2023).

O tratamento adequado das condições associadas pode ajudar a impedir ou retardar sua progressão para fases mais graves, reduzindo tanto a taxa de complicações quanto os custos com tratamentos mais complexos, garantindo assim uma melhor qualidade de vida aos pacientes (Luciano et al, 2012).

2.2 Machine Learning

Machine Learning (ML) é uma subárea da IA, focada em desenvolver algoritmos que permitem que sistemas aprendam a partir de dados e melhorem seu desempenho ao longo do tempo sem a necessidade de programação explícita. Diferentemente dos métodos tradicionais de programação, em que um conjunto de regras precisa ser definido por humanos, o algoritmo de ML capacita as máquinas a reconhecerem padrões e tomar decisões com base em dados históricos (Alpaydin, 2020; Mitchell, 1997).

Existem diversos tipos de algoritmos de ML, cada um com características próprias e aplicações específicas. O aprendizado de máquinas se divide em quatro grandes grupos, demonstrados na Figura 1: o aprendizado supervisionado, o aprendizado semi-supervisionado, o aprendizado não-supervisionado e o aprendizado por reforço (Lenz et al, 2020).

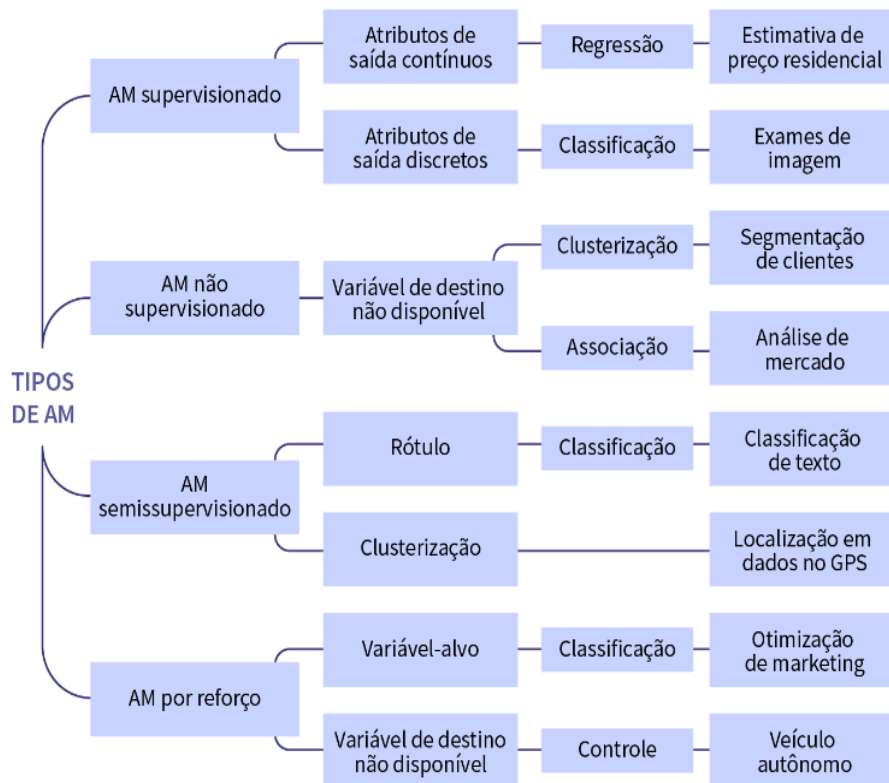


Figura 1. Tipos de aprendizagem de máquina e algumas aplicações

Fonte: Lenz et al, 2020

No aprendizado supervisionado, cada exemplo apresentado ao algoritmo deve incluir a resposta esperada, ou seja, um rótulo indicando a qual classe o exemplo pertence. Isso é comum em problemas de classificação, como a distinção entre imagens de gatos e cachorros. Cada exemplo é representado por um vetor de atributos junto com seu respectivo rótulo e o objetivo do algoritmo é desenvolver um classificador capaz de prever corretamente a classe de novos exemplos não rotulados. Quando os rótulos são discretos, trata-se de um problema de classificação, enquanto valores contínuos caracterizam a regressão (Ludermir, 2021).

No aprendizado não-supervisionado, a estrutura de dados é utilizada sem um rótulo específico. Neste caso, pode-se encontrar relações entre os valores existentes nas variáveis (ou atributos), geralmente em forma de regras, ou até agrupar objetos (ou instâncias) que possuam uma similaridade entre si. Diferentemente do algoritmo supervisionado, chama-se aprendizado não-supervisionado, pois não há associação prévia de variáveis, então fica a cargo do algoritmo agrupar os dados de uma forma lógica (Bramer, 2016; José, 2020).

José (2020) categoriza os algoritmos não-supervisionados em dois grupos, destacando suas principais características e aplicações mais comuns. O primeiro grupo corresponde a clusterização, que envolve a organização dos dados em conjuntos com características semelhantes, por exemplo, agrupar consumidores baseado no seu histórico de compras. E o segundo grupo, corresponde a associação, onde o objetivo é identificar alguma regra lógica que descreve parte dos seus dados, por exemplo, pessoas que compram X tendem também a comprar Y.

No aprendizado semi-supervisionado, diferentemente do aprendizado supervisionado, nem todos os exemplos utilizados no treinamento possuem um rótulo. Enquanto no aprendizado supervisionado cada exemplo deve incluir a resposta esperada para que o algoritmo desenvolva um classificador preciso, no aprendizado semi-supervisionado uma parte dos dados permanece sem rótulo. Esse método é útil quando a quantidade de exemplos rotulados é insuficiente para a generalização adequada do modelo, então utiliza-se informações do conjunto de testes para resolução do problema (Silva, 2008).

No aprendizado por reforço, as classes de agentes aprendem a se adaptar a um determinado estado de acordo com a construção de um modelo que diz qual ação deve

ser realizada a partir de um determinado estado. Então, compreende-se que o algoritmo aprende o que fazer de maneira a maximizar uma recompensa. O algoritmo em si não recebe uma tarefa específica e deve descobrir por tentativa e erro quais ações geram mais recompensas (Kaelbling et al, 1996; Bazzan, 2021).

Nas subseções a seguir, serão detalhados alguns conceitos utilizados para solucionar problemas que envolvem a utilização de machine learning para análises preditivas e o detalhamento dos algoritmos que serão utilizados neste estudo.

2.2.1 Validação cruzada

A validação cruzada é uma técnica de reamostragem muito utilizada para avaliação de modelos de aprendizado de máquina. Essa abordagem propõe a obtenção de uma estimativa imparcial do desempenho do modelo, a qual comparada com a técnica de *holdout* tende a gerar uma estimativa mais precisa do erro de generalização, em conjuntos de dados menores (Maleki et al, 2020).

Segundo Spezia (2024), entre as técnicas de validação cruzada mais utilizadas está a técnica K-Fold, que consiste na divisão do conjunto de dados em K partes, comumente chamadas de folds. A partir da divisão, o modelo então é treinado em K-1 partes e validado com a parte restante. Este processo é repetido K vezes, sendo que em cada iteração um fold diferente é utilizado para validação. A Figura 2 representa uma validação cruzada de 5 folds.

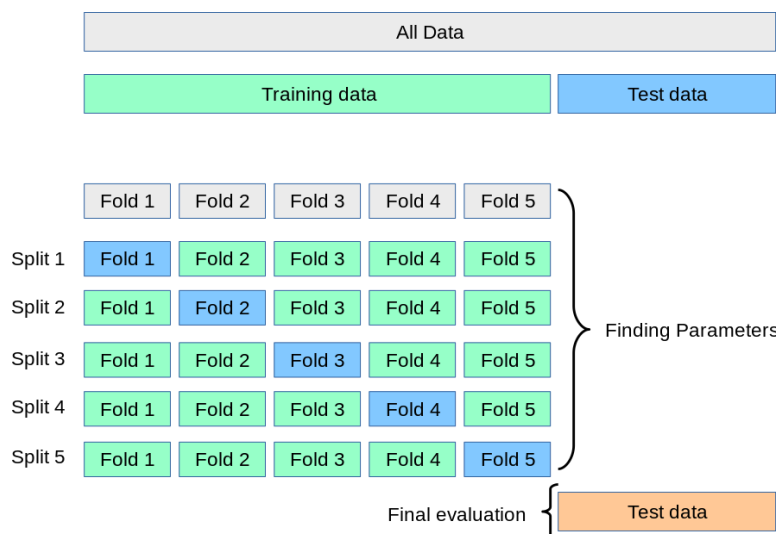


Figura 2. Representação de validação cruzada com 5 folds

Fonte: SCIKIT, 2020

2.2.2 Balanceamento de classes

Um ponto a ser considerado no treinamento de modelos preditivos é o desbalanceamento de classes. Esse desbalanceamento pode afetar negativamente a performance dos modelos, pois os algoritmos tendem a favorecer as classes majoritárias, tornando-os incapazes de realizar a classificação de novas observações (Borges et al, 2021).

Uma técnica de balanceamento disponível se chama SMOTE, a qual realiza a criação de dados sintéticos para ampliar a classe minoritária com o objetivo de balancear o número de observações entre as classes, minimizando a tendência dos algoritmos à classe majoritária. Conforme explicado por Borges et al (2021), a SMOTE utiliza como base o algoritmo k-vizinhos mais próximos para identificar, com base na distância Euclidiana, as amostras mais semelhantes de cada instância. A partir da obtenção desses vizinhos, são selecionadas aleatoriamente amostras de acordo com a proporção de dados a ser aumentada. Após a escolha, calcula-se a diferença entre cada atributo da instância original e dos vizinhos escolhidos, gerando assim um valor de *gap*. Então o novo dado sintético é obtido somando esse valor ao atributo da amostra original, resultando em um ponto que está entre a amostra e seu vizinho. Assim, o processo é repetido até que a classe minoritária esteja equilibrada com a classe majoritária.

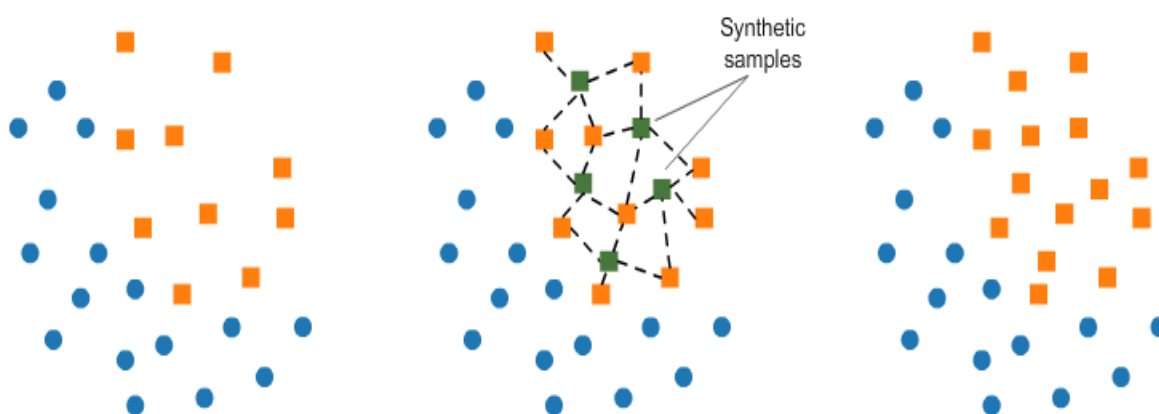


Figura 3. Representação da técnica SMOTE

Fonte: Das, 2019

2.2.3 Modelos de classificação

No presente estudo, foram selecionadas cinco técnicas de classificação. A primeira delas é a *Random Forest* (RF), ou floresta aleatória, a qual é um método de aprendizado de máquina que combina múltiplas árvores de decisão, onde cada árvore emite um voto unitário para a classe mais popular e, ao combinar esses resultados, obtém-se a classificação final. Essa abordagem se destaca por sua alta tolerância a ruídos e outliers, além de evitar o *overfitting*. Devido à sua eficácia, a RF tem sido amplamente utilizada em diversas áreas, incluindo a mineração de dados e aplicações no campo biológico (Liu et al, 2012).

Além da RF, outra técnica de classificação considerada no estudo é o *eXtreme Gradient Boosting* (XGBoost). De acordo com Chen et al. (2018), o algoritmo *XGBoost*, realiza previsões utilizando um conjunto de árvores de decisão treinadas sequencialmente. Cada nova árvore adicionada ao modelo busca corrigir os erros das árvores anteriores, um processo conhecido como *boosting*. Esse método permite a construção de um modelo de previsão mais robusto a partir da combinação de modelos individuais mais simples.

Outra técnica utilizada é a SVM, que consiste em uma técnica de aprendizagem de máquina desenvolvida por Vapnik e seus colegas em 1995, que é fundamentada na Teoria de Aprendizado Estatístico e utilizada para a classificação de dados. A proposta da SVM é criar uma margem e maximizá-la para realizar o mapeamento dos dados de entrada e do treinamento em um espaço de dimensões maiores via funções de kernel que ajudam o algoritmo a ficar mais eficiente construindo um hiperplano com uma margem de separação máxima entre os dados (Lorena, 2003, Basseto et al, 2019).

Adicionalmente, o estudo também considera o algoritmo K-Nearest Neighbor (KNN), um método de aprendizagem de máquina que identifica os k objetos mais próximos a um dado ponto de teste dentro do conjunto de treinamento, atribuindo sua classe com base na predominância de uma categoria específica entre esses vizinhos. Esse método busca lidar com a dificuldade de encontrar correspondências exatas entre os dados e com possíveis informações conflitantes provenientes dos vizinhos mais próximos. Para isso, leva em consideração elementos essenciais, como o conjunto de dados rotulados para avaliar a classe dos objetos de teste, uma métrica de similaridade para calcular a proximidade entre objetos, a escolha do valor de k, a qual determina o

número de vizinhos mais próximos a serem considerados, e a estratégia para determinar a classe do objeto-alvo (Steinbach et al, 2009).

Por fim, o estudo também inclui o método de Naive Bayes, que se baseia no teorema de Bayes para realizar classificações, assumindo de forma simplificada que as características são condicionalmente independentes entre si, dado o valor da classe (SCIKIT, [s.d]). O teorema de Bayes relaciona informações, com a probabilidade de ocorrência, para gerar uma nova probabilidade quando os fatos acontecem de maneira relacionada ou são dependentes (Silver, 2013).

2.2.4 Métricas de avaliação

A avaliação de modelos de classificação binária pode ser feita a partir da análise de diferentes métricas, auxiliando na interpretação da performance dos modelos. Uma das ferramentas mais utilizadas para essa tarefa é a matriz de confusão, a qual apresenta a distribuição das previsões corretas e incorretas e a distribuição dos rótulos da classe real verdadeira e falsa. Conforme demonstrado no Quadro 2, a matriz de confusão pode resultar nesses quatro resultados diferentes: Verdadeiro positivo, Verdadeiro negativo, Falso positivo e Falso negativo (Cechinel et al, 2019).

Quadro 2. Representação da matriz de confusão

Fonte: Cechinel et al, 2019

		Valor predito	
		Positivo	Negativo
Real	Positivo	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Negativo	Falso Positivo (FP)	Verdadeiro Negativo (VN)

A partir do quadro, pode-se observar o Verdadeiro Positivo, que refere-se aos casos que o modelo previu corretamente a classe positiva, o Verdadeiro Negativo que indica os casos que o modelo previu corretamente a classe negativa, o Falso Positivo, que indica que o modelo previu a classe positiva incorretamente e o Falso Negativo, que se refere aos casos de previsão incorreta da classe negativa (Delrue et al, 2024).

Adicionalmente, a partir da matriz de confusão é possível calcular métricas derivadas para melhor avaliação do modelo, como acurácia, f1-score, sensibilidade, especificidade, valor preditivo positivo e valor preditivo negativo, as quais estão descritas no Quadro 3.

Quadro 3. Métricas de avaliação oriundas da matriz de confusão

Fonte: Cechinel et al, 2019

Métrica	Descrição	Fórmula
Acurácia	Capacidade do modelo em acertar as predições em relação ao total de predições realizadas.	$\frac{VP + VN}{VP + VN + FP + FN}$
Sensibilidade	Capacidade do modelo de identificar corretamente os positivos.	$\frac{VP}{VP + FN}$
Especificidade	Capacidade do modelo de identificar corretamente os negativos.	$\frac{VN}{VN + FP}$
VPP	Proporção de casos classificados como positivos, dado que realmente são positivos	$\frac{VP}{VP + FP}$
VPN	Proporção de casos classificados como negativos, dado que realmente são negativos	$\frac{VN}{VN + FN}$
F1-Score	Média harmônica entre VPP e Sensibilidade	$\frac{2 \times VPP \times Sensibilidade}{VPP + Sensibilidade}$

Outra métrica comumente utilizada que pode ser gerada a partir dos valores da matriz de confusão é a curva ROC (*Receiver Operating Characteristic*), a qual consiste em uma representação gráfica da performance de um modelo, segundo a sua taxa de sensibilidade e sua taxa de especificidade (Brito et al, 2008). Segundo Hosmer e Lemeshow (1980), a área sob a curva ROC indica a capacidade de discriminação de um modelo, então definiram uma regra geral onde os valores de área sob a curva entre 0.7 e

0.8 representam discriminação aceitável, entre 0.8 e 0.9 indicam excelente discriminação e valores acima de 0.9 indicam uma discriminação excepcional.

2.2.5 Trabalhos prévios

Estudos prévios já foram realizados na área de predição da DRC, dentre eles podemos citar o trabalho de Reinheimer et al (2023), que abordou o uso de Multilayer Perceptron (MLP), Random Forest, Regressão Logística, XGBoost e Árvore de Decisão sobre 17 biomarcadores para o diagnóstico de DRC, onde os melhores modelos foram a Random Forest e a Regressão Logística, obtendo a Área Sob a Curva (AUC) de 0.77. De outra forma, Ravizza et al (2019) utilizaram o algoritmo Regressão Logística para detecção de DRC e aplicaram sobre sete variáveis: idade, índice de massa corporal, taxa de filtração glomerular e as concentrações de creatinina, albumina, glicose e hemoglobina, obtendo uma AUC de 0.79. Observa-se que a maioria dos estudos concentram-se na utilização de variáveis laboratoriais para predição da DRC, deixando uma lacuna a ser explorada, com a inserção de variáveis relacionadas ao estilo de vida e histórico do paciente, os quais podem contribuir para melhorar a acurácia preditiva da DRC.

3. Metodologia

Nesta seção é apresentada a metodologia adotada para o desenvolvimento da pesquisa, explicando as etapas seguidas. São descritos os processos de escolha da base de dados e preparação, a escolha dos modelos e respectivos parâmetros utilizados para a detecção da DRC.

3.1 Base de dados

A base de dados disponibilizada no Kaggle contém informações detalhadas sobre 1659 pessoas, incluindo pacientes diagnosticados com DRC e pessoas sem a condição. Esse conjunto de dados, descrito no Quadro 4, inclui variáveis demográficas, fatores de estilo de vida, histórico médico, medições clínicas, uso de medicamentos, sintomas, qualidade de vida, exposições ambientais e comportamentos de saúde. A base de dados foi compartilhada por Rabie El Kharoua sobre a licença CC BY 4.0, e trata-se de um conjunto de dados sintético elaborada para fins educacionais, permitindo a realização de análises e modelagem preditivas (Kharoua, 2024; Kaggle, 2024)).

Quadro 4: Variáveis presentes na base de dados

Fonte: Kharoua, 2024

Variável	Descrição
PatientID	ID do paciente
Age	Idade do paciente
Gender	Gênero do paciente
Ethnicity	Etnia do paciente
SocioeconomicStatus	Status socioeconômico do paciente
EducationLevel	Nível de educação
BMI	Índice de massa corporal
Smoking	Status de tabagismo
AlcoholConsumption	Consumo semanal de álcool
PhysicalActivity	Horas de atividade física semanal
DietQuality	Qualidade da dieta
SleepQuality	Qualidade do sono
FamilyHistoryKidneyDisease	Histórico familiar de doença renal
FamilyHistoryHypertension	Histórico familiar de hipertensão
FamilyHistoryDiabetes	Histórico familiar de diabetes
PreviousAcuteKidneyInjury	Histórico de lesão renal aguda prévia
UrinaryTractInfections	Histórico de infecção no trato urinário
SystolicBP	Pressão arterial sistólica
DiastolicBP	Pressão arterial diastólica
FastingBloodSugar	Nível de glicose em jejum
HbA1c	Nível de hemoglobina A1c
SerumCreatinine	Nível de creatinina sérica
BUNLevels	Nível de nitrogênio ureico no sangue
GFR	Taxa de filtração glomerular

ProteinInUrine	Nível de proteína na urina
ACR	Razão Albumina-Creatinina
SerumElectrolytesSodium	Nível de sódio sérico
SerumElectrolytesPotassium	Nível de potássio sérico
SerumElectrolytesCalcium	Nível de cálcio sérico
SerumElectrolytesPhosphorus	Nível de fósforo sérico
HemoglobinLevels	Nível de hemoglobina
CholesterolTotal	Nível de colesterol total
CholesterolLDL	Nível de colesterol LDL
CholesterolHDL	Nível de colesterol HDL
CholesterolTriglycerides	Nível de triglicerídeos
ACEInhibitors	Uso de inibidores de enzima conversora de angiotensina
Diuretics	Uso de diuréticos
NSAIDsUse	Frequência de uso de anti-inflamatórios
Statins	Uso de estatinas
AntidiabeticMedications	Uso de antidiabéticos
Edema	Presença de edemas
FatigueLevels	Nível de fadiga
NauseaVomiting	Frequência de náuseas e vômitos
MuscleCramps	Frequência de câibras musculares
Itching	Intensidade de coceira
QualityOfLifeScore	Qualidade de vida
HeavyMetalsExposure	Exposição a metais pesados
OccupationalExposureChemicals	Exposição a produtos químicos
WaterQuality	Qualidade da água
MedicalCheckupsFrequency	Frequência de exames médicos por ano

MedicationAdherence	Adesão à medicação
HealthLiteracy	Alfabetização em saúde
Diagnosis	Diagnóstico de Doença Renal Crônica
DoctorInCharge	Médico responsável

3.2 Tecnologias Utilizadas

Como ferramenta para desenvolvimento do projeto, foi utilizado o serviço de desenvolvimento disponibilizado pela Google, o Google Colab. Consiste em um serviço gratuito de Jupyter Notebook, o qual não necessita de configuração para uso e fornece acesso gratuito a recursos de computação, incluindo GPUs e TPUs (GOOGLE, [s. d]).

Foi utilizada a linguagem Python, versão 3.11.11, com importação das bibliotecas apresentadas na Figura 3, incluindo Pandas, Numpy e ImbLearn para realizar a importação e transformações necessárias na base de dados; Matplotlib e Seaborn para visualização dos dados; SKLearn e XGBoost para treinamento e validação dos modelos.

Devido ao fato de a base conter rótulos definidos, indicando se cada paciente possui ou não DRC, foram utilizados modelos de aprendizado de máquina supervisionado adequados para o problema de classificação.

```

# Pré-processamento
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler

# Visualização de dados
import matplotlib.pyplot as plt
import seaborn as sns

# Divisão da base
from sklearn.model_selection import train_test_split, KFold

# Balanceamento de classes
from imblearn.over_sampling import SMOTE

# Modelos
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.naive_bayes import GaussianNB
from xgboost import XGBClassifier

# Avaliação de desempenho
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix, roc_curve, auc

```

Figura 3: Bibliotecas importadas para execução do projeto

Fonte: Elaboração própria

3.3 Pré-Processamento

Primeiramente foi realizada a importação e o pré-processamento dos dados, conforme ilustrado na Figura 4. Segundo Santos (2018), as técnicas de pré-processamento são utilizadas para fazer a filtragem e transformações dos dados, mantendo assim, apenas o que é relevante para aprendizado dos algoritmos.

Nesta etapa foram removidas as colunas *PatientID*, *DoctorInCharge*, *Ethnicity*, *SocioeconomicStatus*, *EducationLevel*, *MedicalCheckupsFrequency*, *Medication Adherence*, *HealthLiteracy*, por não impactarem na detecção de DRC. Para padronização dos dados foi utilizado a técnica *StandardScaler* disponibilizada pela biblioteca *SKLearn* (*SCIKIT*, [s.d]).

```
[ ] # Remoção de colunas desnecessárias
df_drc = df_drc.drop(
    columns=[
        'PatientID',
        'DoctorInCharge',
        'Ethnicity',
        'SocioeconomicStatus',
        'EducationLevel',
        'MedicalCheckupsFrequency',
        'MedicationAdherence',
        'HealthLiteracy'
    ]
)

[ ] # Preditores
X = df_drc.drop(columns=['Diagnosis'])

# Variável alvo
y = df_drc['Diagnosis']

[ ] # Padronização dos dados
scaler = StandardScaler()
X = pd.DataFrame(scaler.fit_transform(X), columns=X.columns)
```

Figura 4: Pré-processamento dos dados

Fonte: Elaboração própria

3.4 Validação cruzada e balanceamento de classes

Após, foi realizado a etapa de validação cruzada. A base de dados foi dividida em 5 folds, o que implica na divisão de 5 subconjuntos dos dados e, conseqüentemente, o modelo foi treinado 5 vezes, sendo que em cada iteração foram utilizados 4 dos 5 fold para treinamento e o fold restante para validação. Isso garante que todos os dados sejam utilizados tanto para treinamento quanto para validação. Foram realizados testes

adicionais com diferentes quantidades de folds: 3, 5, 7 e 10, porém a divisão em 5 folds apresentou os melhores resultados.

Juntamente com a divisão da base de dados em 5 folds, foi aplicado o balanceamento de classes para lidar com conjuntos de dados desbalanceados. Para realizar esta etapa, foi utilizado a técnica SMOTE, a qual cria exemplos sintéticos para a classe minoritária, assim equilibrando a distribuição das classes.

Conforme apresentado na Figura 5, na etapa da validação cruzada, o balanceamento de classes foi realizado exclusivamente nos 4 folds de treinamento em cada iteração do processo. O fold de validação, que foi utilizado para testar o modelo, não passou pelo SMOTE, evitando que informações do conjunto de validação vazem para o modelo em questão durante o treinamento, o que poderia prejudicar a avaliação de desempenho.

```
def validacao_cruzada(modelo, X, y):
    kfold = KFold(n_splits=5)

    resultados = {
        "acuracia": [],
        "f1": [],
        "sensibilidade": [],
        "especificidade": [],
        "vpp": [],
        "vpn": []
    }

    for idx, (idx_treino, idx_validacao) in enumerate(kfold.split(X)):
        # Dados de treino
        X_split_treino = X.iloc[idx_treino, :]
        y_split_treino = y.iloc[idx_treino]

        # Balanceamento
        sm = SMOTE(random_state=SEED)
        X_split_treino, y_split_treino = sm.fit_resample(X_split_treino, y_split_treino)

        # Treinamento
        modelo.fit(X_split_treino, y_split_treino.values.flatten())

        # Dados de validação
        X_split_validacao = X.iloc[idx_validacao, :]
        y_split_validacao = y.iloc[idx_validacao]

        # Predição
        predicoes = modelo.predict(X_split_validacao)

        # Métricas
        metricas = calcular_metricas(y_split_validacao, predicoes)
        for chave in metricas:
            resultados[chave].append(metricas[chave])

    return resultados
```

Figura 5: Validação cruzada (K-Fold) e balanceamento dos dados (SMOTE)

Fonte: Elaboração própria

3.5 Aplicação dos algoritmos

Após a realização da divisão da base em folds e realizado o balanceamento dos dados, foram aplicados os algoritmos de classificação RF, KNN, SVM, Naive Bayes e XGBoost Classifier. Esses algoritmos permitem a manutenção de parâmetros para otimizar o seu desempenho. Sendo assim, foram testadas todas as combinações possíveis entre os valores definidos para cada hiperparâmetro apresentado no Quadro 5.

Quadro 5: Parâmetros e respectivos valores testados para cada algoritmo

Fonte: Elaboração própria

Algoritmo	Parâmetro	Valores
Random Forest	n_estimators	1, 5, 7, 10, 50, 90, 100 , 150, 200 e 300
	max_depth	1, 3, 5, 7, 9, 11 , 15, 20
KNN	n_neighbors	3 , 5, 7, 9, 11
SVM	C	0.1, 0.2 , 0.3, 0.5, 1, 10, 100
Naive Bayes	var_smoothing	1e-8, 1e-6, 1e-5, 1e-4, 1e-3 e 1e-2
XGBoost	n_estimators	1, 5, 7, 10, 50, 90, 100, 150, 200 e 300
	learning_rate	0.1, 0.2, 0.3 , 0.4, 0.5, 0.6 e 1
	max_depth	1, 3, 5, 7 , 9, 11, 15, 20

Os melhores resultados, com base na avaliação das métricas apresentadas na seção a seguir, foram obtidos a partir dos seguintes conjuntos de parâmetros: para o algoritmo Random Forest, os melhores resultados ocorreram com 100 estimadores e profundidade máxima de 11. Para o algoritmo KNN, os melhores resultados foram obtidos a partir de 3 vizinhos. No caso do SVM, o valor do parâmetro C que apresentou melhor desempenho foi 0.2. No algoritmo Naive Bayes, o valor de *var_smoothing* que apresentou melhores resultados foi igual a 1e-3. E por fim, para o algoritmo XGBoost, a combinação que demonstrou melhor eficiência foi de, 200 estimadores, com taxa de aprendizado de 0.3 e profundidade máxima de 7.

3.6 Avaliação de desempenho

Para realizar a avaliação do desempenho desses modelos foi utilizada a matriz de confusão, que permite visualizar o desempenho do modelo comparando os valores previstos e os reais. A partir dela, foram calculadas as seguintes métricas: acurácia, F1-score, sensibilidade, especificidade, valor preditivo positivo (VPP) e valor preditivo negativo (VPN). Além disso, foi analisada a curva ROC, que avalia a capacidade discriminativa do modelo, considerando a relação entre a taxa de verdadeiros positivos e a taxa de falsos positivos.

4. Apresentação e discussão dos resultados

Nesta seção serão apresentados os resultados obtidos após a execução de todas as etapas apresentadas na seção 3, bem como uma comparação dos resultados dos algoritmos de classificação desenvolvidos no estudo.

O Quadro 6 apresenta os resultados das métricas de desempenho dos modelos treinados com as técnicas de validação cruzada e balanceamento. Os resultados apresentados são as médias dos 5 folds, considerando o uso dos parâmetros com melhores desempenhos, apresentados na seção 3.5.

Quadro 6: Resultado das métricas de desempenho (média)

Fonte: Elaboração própria

Modelo	Acurácia	F1-Score	Sensibilidade	Especificidade	VPP	VPN
RF	0.9101	0.9515	0.9775	0.1767	0.9290	0.3977
KNN	0.4184	0.5493	0.3879	0.7961	0.9586	0.1080
SVM	0.8945	0.9426	0.9566	0.2639	0.9314	0.3130
Naive Bayes	0.7987	0.8838	0.8410	0.4181	0.9377	0.1867
XGBoost	0.9270	0.9602	0.9848	0.3028	0.9381	0.6024

Alguns resultados dos modelos de aprendizado de máquina mostraram um desempenho geral satisfatório, tendo os algoritmos XGBoost Classifier e Random Forest, ambos baseados em árvores de decisão, apresentando os melhores resultados no geral respectivamente, obtendo uma acurácia acima de 90% e sensibilidade média de 98%. O algoritmo SVM também obteve um bom desempenho no geral, com uma acurácia média de 89% e sensibilidade média de 93%, porém um pouco inferior aos

citados anteriormente. Já os algoritmos Naive Bayes e KNN obtiveram os piores resultados do estudo, com acurácias de 77% e 41% respectivamente.

Em relação às métricas de VPP e VPN, todos os algoritmos obtiveram VPP superiores a 92%, demonstrando que há uma alta precisão para identificação de casos positivos, indicando que na maioria dos casos em que os pacientes foram classificados com a patologia, realmente se enquadram nessa classe. Em contrapartida, os algoritmos aplicados apresentaram um VPN consideravelmente baixo, onde o algoritmo XGBoost obteve o valor de 60%, sendo considerado o mais eficiente ao ser comparado com os outros algoritmos, os quais apresentaram valores abaixo de 40%, fazendo com que a capacidade de descartar corretamente os casos negativos fique comprometida.

Para avaliação do modelo também é apresentado a Curva ROC na Figura 6, onde segundo a escala apresentada na seção 2.2.4, os algoritmos XGBoost e Random Forest apresentam uma excelente discriminação, e os algoritmos SVM e Naive Bayes apresentam uma discriminação aceitável.

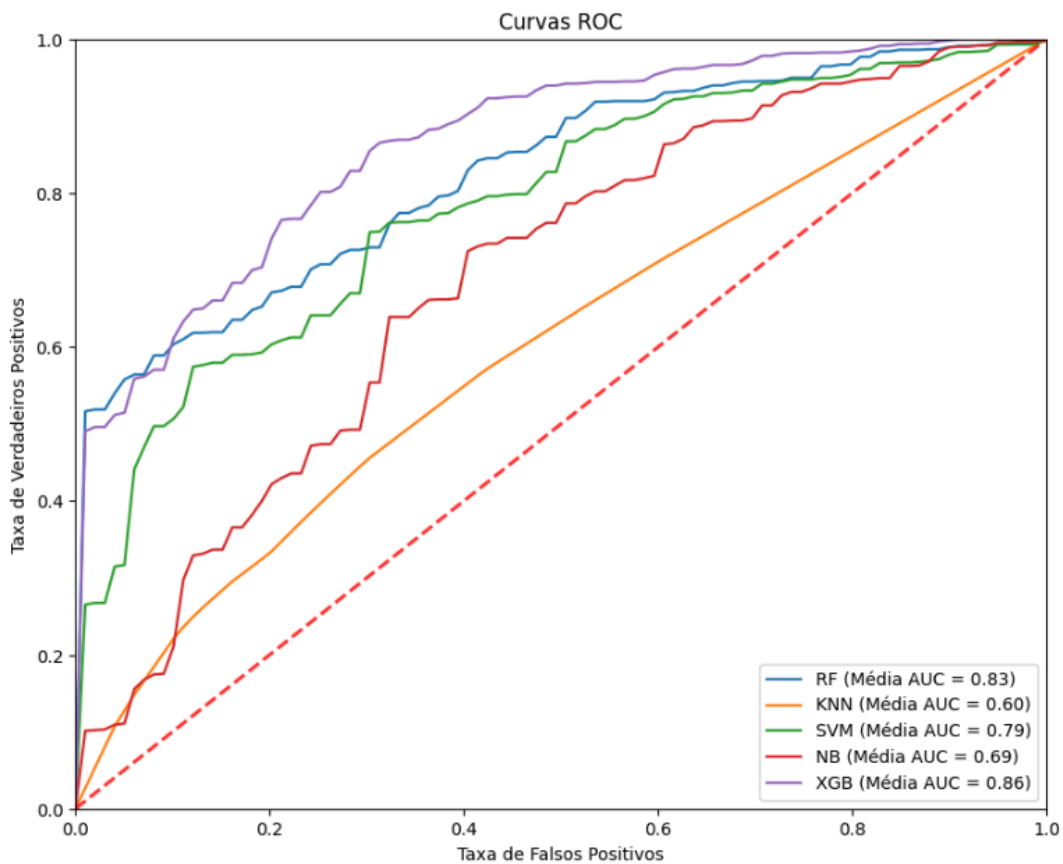


Figura 6: Curvas ROC dos modelos de classificação

Fonte: Elaboração própria

Adicionalmente, alguns algoritmos, como o XGBoost, disponibilizam funcionalidades que permitem visualizar a importância das variáveis consideradas no seu treinamento. A Figura 7 apresenta as 10 variáveis mais relevantes, ordenadas de acordo com a sua contribuição para o modelo.

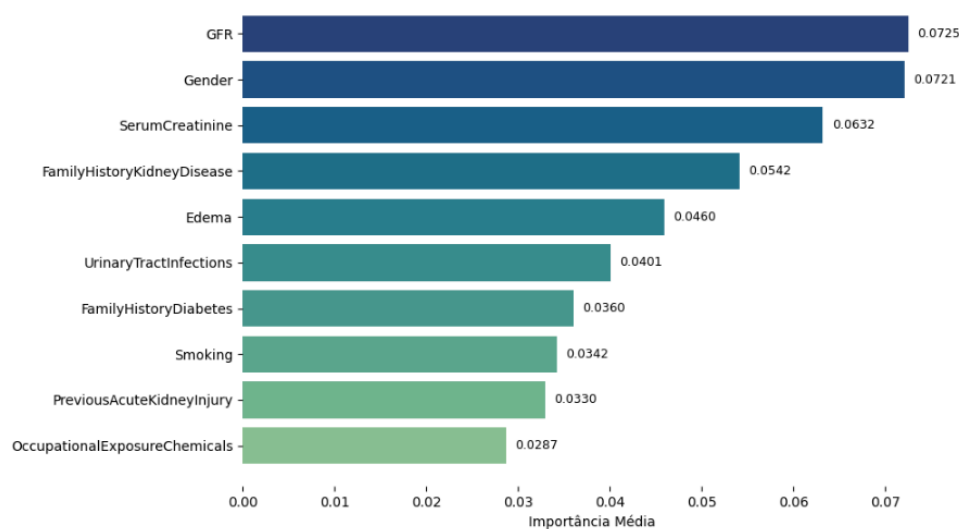


Figura 7: Gráfico das 10 variáveis mais relevantes para treinamento do XGBoost

Fonte: Elaboração própria

A variável mais relevante é a *GFR*, a qual corresponde a taxa de filtração glomerular, seguida pelo gênero, histórico de doença renal e nível de creatinina sérica. Essas variáveis demonstram ter maior impacto na predição do modelo, indicando sua relevância clínica no diagnóstico da doença. Ao ser comparada com o conteúdo trazido no referencial, percebe-se uma coerência significativa. A TFG, estabelecido pela NKF (2002), é um dos principais critérios para detecção de DRC. Seguindo esse contexto, o histórico familiar de DRC, também apontado com uma das variáveis mais importantes para detecção pelo modelo, é apontado por Gordan (2006) como um fator de risco para desenvolvimento da doença. Além disso, a creatinina sérica, também aparece como um marcador importante para detecção, visto que ela é utilizada para estimar a TFG.

Entretanto, a presença do gênero como uma das variáveis mais importantes para detecção da patologia pode representar uma contribuição adicional, onde pode haver a possibilidade de conter uma diferença entre homens e mulheres no desenvolvimento da

DRC, visto que não há viés na classe, sugerindo assim uma investigação mais profunda em análises futuras.

Apesar dos resultados satisfatórios, é importante destacar algumas limitações do estudo, as quais podem ter influenciado os resultados obtidos. A utilização de amostras sintéticas, geradas a partir das técnicas de balanceamento, como o SMOTE, pode ser importante para mitigar o desequilíbrio entre classes, porém, podem acabar aumentando a probabilidade de sobreajuste do modelo, fazendo com que o modelo aprenda os padrões dessas amostras sintéticas e defasando assim a capacidade de generalização em dados reais. Além disso, a utilização de um grande número de variáveis para treinamento dos modelos, que contém algumas redundâncias, podem ter impactado negativamente o desempenho de alguns algoritmos, como o KNN, algoritmo baseado em distância.

5. Considerações Finais

Este trabalho explorou a utilização de modelos de aprendizado de máquina para diagnóstico de Doença Renal Crônica (DRC), a partir de dados demográficos, clínicos e fatores de estilo de vida. Foram aplicados diferentes algoritmos de classificação e realizada a comparação de métricas obtidas através da validação cruzada juntamente com a utilização de uma técnica de balanceamento de classes para lidar com a desigualdade na distribuição.

Com uma acurácia de 92,70%, o modelo XGBoost Classifier destacou-se por apresentar as melhores métricas na classificação de pacientes com DRC, mostrando-se promissor a sua utilização como ferramenta de auxílio ao diagnóstico precoce de DRC, sugerindo que o seu uso pode contribuir para auxiliar nas tomadas de decisões clínicas. No entanto, um fator importante para obter essa performance foi a utilização da técnica de balanceamento SMOTE, gerando amostras sintéticas, o que pode causar um viés no conjunto de dados.

Para trabalhos futuros, sugere-se utilizar uma base de dados maior e melhor distribuída, com o intuito de melhorar a generalização dos modelos e reduzir possíveis ruídos causados por desbalanceamento. Além disso, sugere-se a aplicação de algoritmos baseados em redes neurais, os quais podem explorar melhor as relações entre as variáveis clínicas, bem como a utilização de técnicas de balanceamento mais avançadas.

Por fim, este trabalho pode contribuir para a comunidade acadêmica ao realizar uma análise comparativa de modelos preditivos em um conjunto de dados com variáveis relacionadas ao histórico e estilo de vida, além das variáveis clínicas de pessoas com DRC, o que pode auxiliar em pesquisas futuras.

6. Referências

Alpaydin; Ethem. **Introduction to machine learning**. MIT press, 2020;

ALURA. **Python: Origem do nome**. Disponível em <https://www.alura.com.br/artigos/python-origem-do-nome?srsId=AfmBOopMZCX8lnZfLkIQoO-UDPoPiE9ovwgJeRjuCSBebvbBv-FBNzbn>. Acesso em 10/03/2025;

Bai, Q., Su, C., Tang, W. et al. **Machine learning to predict end stage kidney disease in chronic kidney disease**. Sci Rep 12, 8377, 2022;

Basseto, Edson Luis; Dal Pai, Alexandre; De Souza Marques, Adriano. **Máquinas de Vetores de Suporte (SVM) na estimativa da fração difusa da irradiação solar global**. Energia na Agricultura, v. 34, n. 4, p. 532-543, 2019;

Bastos, Marcus Gomes et al. **Doença Renal Crônica: frequente e grave, mas também prevenível e tratável**. Universidade Federal de Juiz de Fora - MG, 2009;

Bastos, Marcus Gomes; Kirszrajnk, Gianna Mastroianni. **Doença renal crônica: importância do diagnóstico precoce, encaminhamento imediato e abordagem interdisciplinar estruturada para melhora do desfecho em pacientes ainda não submetidos à diálise**. Brazilian Journal of Nephrology, v. 33, p. 93-108, 2011;

Bazzan, Ana L. C.. **Contribuições de aprendizado por reforço em escolha de rota e controle semafórico**. Porto Alegre – RS, 2021;

Borges, Fernando Elias de Melo et al. **Classificação de Dados do Cadastro Ambiental Rural com uso de Algoritmos de Aprendizagem de Máquina**. SBAI, 2021;

Bramer, M.. **Principles of Data Mining**. 3 ed. Londres: Springer London, 2016;

Brito, Giovani Antonio Silva, et al. **Modelo de classificação de risco de crédito de empresas**. Revista Contabilidade & Finanças, 2008;

Cechinel, Cristian; Da Silva Camargo, Sandro. **Mineração de dados educacionais: avaliação e interpretação de modelos de classificação**. Porto Alegre: SBC, 2020;

Chen, Z. et al. **Xgboost classifier for ddos attack detection and analysis in sdn-based cloud**. IEEE., 2018;

Das, Ashesh. **Oversampling to remove class imbalance using SMOTE**. Disponível em <<https://medium.com/@asheshdas.ds/oversampling-to-remove-class-imbalance-using-smote-94d5648e7d35>>. Acesso em 17/06/2025;

Delrue, Charlotte. et al. **Application of machine learning in chronic kidney disease: current status and future prospects**, 2024;

Ferguson R, Grim CE, Opgenorth TJ. **A familial risk of chronic renal failure among blacks on dialysis?**. J Clin Epidemiol. 1988;

Gobbo, M. P. A. et al. **Inteligência artificial como ferramenta de auxílio no diagnóstico de alterações vocais e patologias de laringe**. Anais Científicos, 2019;

GOOGLE. **Colaboratory: Frequently Asked Questions**. Disponível em <<https://research.google.com/colaboratory/faq.html#resource-limits>>. Acesso em 10/03/2025;

Gordan, Pedro A.. **Grupos de risco para doença renal crônica**. Braz. Nephrol. 2006;

Hosmer, D. W; Lemeshow, S. **A goodness-of-fit test for the multiple logistic regression model.** Communications in Statistics, 1980;

IBM. **O que é regressão logística?** Disponível em <<https://www.ibm.com/br-pt/topics/logistic-regression>>. Acesso em 09/03/2025;

Junior, João Egidio Romão. **Doença Renal Crônica: Definição, Epidemiologia e Classificação.** Brazilian Journal of Nephrology, Volume 47, 2004;

Kaelbling, Leslie Pack; Littman, Michael L.; Moore, Andrew W. **Reinforcement learning: A survey.** Journal of artificial intelligence research, p. 237-285, 1996;

KAGGLE. **Chronic Kidney Disease Dataset.** Disponível em <<https://www.kaggle.com/datasets/rabieelkharoua/chronic-kidney-disease-dataset-analysis>>. Acesso em 15/02/2025;

Kalantar-Zadeh, Kamyar et al. **Chronic kidney disease.** The Lancet, Volume 398, Issue 10302, 786 - 802;

Kharoua, Rabie El. **Chronic Kidney Disease Dataset.** Disponível em <<https://doi.org/10.34740/KAGGLE/DSV/8658224>>. Acesso em 15/02/2025;

Lenz, Maikon Lucian et al. **Fundamentos de aprendizagem de máquina.** Porto Alegre: Sagah, 2020;

Liu, Yanli; Wang, Yourong; Zhang, Jian. **New machine learning algorithm: Random forest.** In: **International conference on information computing and applications.** Berlin, 2012;

Lorena, Ana Carolina; CARVALHO, André Carlos Ponce de Leon Ferreira de. **Revisão de técnicas para geração de classificadores de margens largas multiclases.** 2003;

Luciano, Eduardo de Paiva et al. **Prospective study of 2151 patients with chronic kidney disease under conservative treatment with multidisciplinary care in the Vale do Paraíba, SP.** Brazilian Journal of Nephrology, v. 34, p. 226-234, 2012;

Ludermir, Teresa Bernarda. **Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências.** Estudos Avançados, v. 35, p. 85-94, 2021;

Maleki, Farhad et al. **Machine learning algorithm validation: from essentials to advanced applications and implications for regulatory certification and deployment.** Neuroimaging Clinics of North America, 2020;

Malkina, Anna. **Doença Renal Crônica.** Universidade da Califórnia, 2023. Disponível em <https://www.msmanuals.com/pt/profissional/dist%C3%BArbios-geniturin%C3%A1rios/doen%C3%A7a-renal-cr%C3%B4nica/doen%C3%A7a-renal-cr%C3%B4nica>. Acessado em 24 de fevereiro de 2025;

Marinho, Ana Wanda Guerra Barreto et al. **Prevalência de doença renal crônica em adultos no Brasil: revisão sistemática da literatura.** Cadernos Saúde Coletiva, v. 25, p. 379-388, 2017;

McClellan, WM, Langston RD, Presley R, et al. **Medicare patients with cardiovascular disease have a high prevalence of chronic kidney disease and a high rate of progression to end-stage renal disease.** J Am Soc Nephrol. 2004;

Mitchell, Tom M.. **Machine Learning.** Portland: McGraw-Hill Science, 1997;

NATIONAL KIDNEY FOUNDATION. **K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification.** Am J Kidney Dis. 2002;

Reinheimer, Isabel Cristina et al. **Avaliação de Tecnologia em Saúde na Doença Renal Crônica: Análise de Biomarcadores por Inteligência Artificial.** Jornal de Assistência Farmacêutica e Farmacoeconomia, v. 1, n. s. 2, 2023;

SCIKIT LEARN. **Naive Bayes.** Disponível em https://scikit-learn.org/stable/modules/naive_bayes.html. Acesso em 09/03/2025;

Segatto, Vinicius. **SVM, ou Support Vector Machine.** Disponível em <https://medium.com/liga-mackenzie-de-ia-ci%C3%A9ncia-de-dados/svm-ou-support-vector-machine-7efcabdcc7be>. Acesso em 04/03/2025;

Shlipak, MG, Heidenreich PA, Noguchi H, et al. **Association of renal insufficiency with treatment and outcomes after myocardial infarction in elderly patients.** 2002;

Silva, Marcelo Mourão. **Uma abordagem evolucionária para aprendizado semi-supervisionado em máquinas de vetores de suporte.** 2008;

Silva, Rafael Rodrigues da. **Áreas em que Python vem sendo utilizado no aprendizado de máquina: um mapeamento sistemático.** 2023;

Silver, Nate. **O sinal e o ruído: porque tantas previsões falham e outras não.** Rio de Janeiro. Intrínseca, 2013;

SOCIEDADE BRASILEIRA DE NEFROLOGIA. **Censo Brasileiro de Diálise 2023.** Disponível em https://www.bjnephrology.org/wp-content/uploads/articles_xml/2175-8239-jbn-47-1-e20240081/2175-8239-jbn-47-1-e20240081-pt.pdf. Acesso em 16/02/2025;

Steinbach, Michael; TAN, Pang-Ning. **The Top Ten Algorithms in Data Mining**. p. 12. Chapman and Hall/CRC, 2009;

Vieira, Marcos. **Doença renal é mais comum em pessoas com diabetes**. Disponível em
<<https://prorim.org.br/doenca-renal-e-mais-comum-em-pessoas-com-diabetes/#:~:text=A%20diabetes%20pode%20trazer%20danos,Marcos%20Vieira.>>. Acesso em 16/02/2025;

Xue, JL, Ma JZ, Louis TA, Collins AJ. **Forecast of the number of patients with end-stage renal disease in the United States to the year 2010**. J Am Soc Nephrol 2001.