Predição de doenças crônicas nos rins

*Note: Sub-titles are not captured for https://ieeexplore.ieee.org and should not be used

Alex Oliveira
Inteligência Computacional em Saúde
Universidade Federal do Espírito Santo
Vitória, Brasil
aleks.vix@outlook.com

Icaro Madalena do Nascimento
Inteligência Computacional em Saúde
Universidade Federal do Espírito Santo
Vitória, Brasil
icaro.nascimento@edu.ufes.br

***Abstract*—A Doença Renal Crônica (DRC) representa um grave desafio global de saúde, com milhões de pessoas afetadas e uma alta taxa de mortalidade anual. A detecção precoce da DRC é fundamental para impedir sua progressão e otimizar a eficácia do tratamento. Nesse contexto, as técnicas de aprendizado de máquina (Machine Learning - ML) têm se mostrado ferramentas valiosas para prever e classificar a DRC de forma mais rápida e precisa do que os métodos tradicionais. Este trabalho aborda a classificação da DRC empregando algoritmos de ML, com foco particular nos modelos Support Vector Machine (SVM) e Random Forest (RF). Modelos como o SVM e o RF, apresentaram consistentemente altos desempenhos, com acurácias superiores a 96% em cenários onde a imputação de valores ausentes foram aplicadas, e mesmo sem algumas técnicas comuns de tratamento de dados o desempenho se manteve alto, destacando sua eficácia no diagnóstico da DRC.**

***Index Terms*—machine learning, ckd, SVM, Random Forest, health, ai.**

I. Introdução

A Doença Renal Crônica (DRC) é um desafio de saúde global, afetando mais de 10% da população mundial e sendo responsável por milhões de mortes anualmente [2]. Em 2016, a DRC atingiu 753 milhões de pessoas globalmente [2]. Em 2017, aproximadamente 843,6 milhões de pessoas foram diagnosticadas com a doença, que se tornou a 11ª causa mais letal de mortalidade em todo o mundo, com 1,2 milhão de óbitos anuais [1]. Atualmente, é a sexta causa de morte que mais cresce globalmente [1].

A detecção precoce da DRC é de importância vital para otimizar a eficácia do tratamento e mitigar o impacto financeiro, especialmente porque o tratamento e a medicação nem sempre são acessíveis ou economicamente viáveis na maioria dos países em desenvolvimento [1]. A identificação da doença em estágio inicial e a observação dos fatores de risco podem prevenir o avanço da doença e limitar as complicações para a saúde do paciente [2].

Machine Learning (ML) é uma área da inteligência artificial que tem se destacado por sua capacidade de conferir inteligência a sistemas por meio de experiências passadas, permitindo que eles realizem tarefas de tomada de decisão de forma autônoma, sem a necessidade de programação explícita [4]. Essencialmente, o cerne da ML reside no aprendizado automático de um sistema computacional a partir de dados fornecidos, sejam eles dados brutos ou conjuntos de dados [4]. O objetivo primordial da ML é capacitar computadores a aprender a partir de dados de treinamento para extrair informações e, consequentemente, desempenhar tarefas em dados futuros, comumente denominados dados de teste [1] [4].

O processo de aprendizado de máquina geralmente se desdobra em duas etapas fundamentais: o treinamento do modelo, onde os dados de treinamento são usados para ensinar o sistema, e a tomada de decisão ou teste do modelo, onde o modelo treinado é aplicado a conjuntos de dados desconhecidos para determinar resultados [4]. A acurácia do modelo pode ser obtida nesta última etapa, comparando os resultados previstos com os resultados reais já conhecidos [4]. Esta abordagem tem encontrado um vasto leque de aplicações em diversas áreas, como reconhecimento de padrões, classificação de imagens, previsão de modelos, mineração de dados, mecanismos de busca, análise de sentimentos, previsão de séries temporais, monitoramento da saúde estrutural e assistentes pessoais virtuais como Siri, Alexa e Google Now [4]. Além disso, a ML é uma área de pesquisa e inovação intensamente popular em setores cruciais como saúde, finanças, segurança, gestão de dados e análise e previsão de tendências [4].

Atualmente, os métodos de ML são habitualmente categorizados em três paradigmas de aprendizado: supervisionado, não supervisionado e por reforço [4]. Este trabalho será realizado apenas com aprendizado supervisionado.

O aprendizado supervisionado exige supervisão para a previsão ou tomada de decisão, com o conjunto de dados de entrada dividido em dados de treinamento e teste, e valores de saída ou alvo já atribuídos aos dados de treinamento. É frequentemente empregado para resolver problemas de classificação (valores discretos) e regressão (valores contínuos) [4].

Nesse cenário, as técnicas de aprendizado de máquina emergem como ferramentas promissoras para aprimorar a acurácia e a velocidade do diagnóstico de doenças [1] [3].

II. Trabalhos relacionados

Diversas abordagens têm sido propostas para a classificação da DRC utilizando ML e DL:

El Sherbiny et al. (2023) propõem um método para predizer a infecção por DRC utilizando nove algoritmos distintos de

ML: Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Regressão Logística (LR), Extreme Gradient Boosting (XGB), Adaptive Boosting (ADB ou AdaBoost), K-Nearest Neighbors (KNN) e Rede Neural (NN). Os resultados demonstraram que o AdaBoost superou as outras técnicas, atingindo uma acurácia de 99,17%. O estudo também focou na redução da complexidade do dataset através da seleção das características mais relevantes e empregou diferentes técnicas de imputação de valores ausentes, incluindo KNN, o que foi considerado uma novidade para o dataset de DRC.

Khamparia et al. (2019) introduzem uma estrutura inovadora de aprendizado profundo (KDSAE) para a classificação da DRC, utilizando um modelo de autoencoder empilhado (Stacked Autoencoder - SAE) com um classificador Softmax. O SAE é empregado para extrair características úteis do dataset, e o classificador Softmax prevê a classe final. Este modelo demonstrou um desempenho excepcional, atingindo 100% de acurácia, especificidade, precisão, recall e F1-score no dataset UCI de 400 pacientes.

Aswathy et al. (2022) apresentam um modelo de diagnóstico de DRC baseado em IoT e computação em nuvem, denominado FPA-DNN (Flower Pollination Algorithm-based Deep Neural Network). Este modelo incorpora o algoritmo Oppositional Crow Search (OCS) para seleção de características, otimizando o subconjunto de características dos dados pré-processados. O Algoritmo de Polinização de Flores (FPA) é aplicado para ajustar os parâmetros da Rede Neural Profunda (DNN), visando um melhor desempenho de classificação. O FPA-DNN alcançou uma sensibilidade de 98,80%, especificidade de 98,66%, acurácia de 98,75%, F-score de 99% e kappa de 97,33%.

Em suma, a aplicação de técnicas avançadas de ML e DL, incluindo autoencoders, algoritmos de otimização para ajuste de parâmetros e seleção de características, e a integração com IoT e computação em nuvem, tem mostrado um potencial significativo para melhorar a precisão e a eficiência no diagnóstico precoce da DRC.

As próxima seções abordarão os seguintes tópicos: a metodologia aplicada, o desenvolvimento e a análise dos resultados.

III. Metodologia

A. Base de dados

Para este trabalho foi utilizado uma base com amostras de pacientes coletadas durante 2 meses em um hospital [5]. A base contém 400 registros de pacientes identificados apenas pelo id. São 24 features presentes com 11 variáveis númericas e 13 variáveis nominais. Além disso a base possui a classe alvo. Existem dados faltantes na base, além de alguns erros de inserção nas variáveis nominais com caracteres estranhos. As amostras estão divididas em duas classes com proporções de 60% e 40%.

As características numéricas incluem: idade (em anos, variando de 2 a 90); pressão arterial (Bp) (com valores de 50 a 180, inferido em mmHg); gravidade específica (sg) (de 1.005 a 1.025); albumina (al) (de 0 a 5); açúcar (su) (de 0 a 5); glicose sanguínea aleatória (bgr) (de 22 a 490); ureia sanguínea (bu) (de 1.5 a 322); creatinina sérica (sc) (de 0.3 a 76); sódio (sod) (de 4.5 a 150); potássio (pot) (de 2.4 a 47); hemoglobina (hemo) (de 3.1 a 17.8); volume de células compactadas (pcv) (de 9 a 53); contagem de glóbulos brancos (wc) (de 65 a 26.400); e contagem de glóbulos vermelhos (rc) (de 2.1 a 8.5). As características categóricas, que geralmente indicam a presença ou ausência de uma condição, são: glóbulos vermelhos (rbc) (normal/anormal); células de pus (pc) (normal/anormal); aglomerados de células de pus (pcc) (presente/ausente); bactérias (ba) (presente/ausente); hipertensão (htn) (sim/não); diabetes mellitus (dm) (sim/não); doença arterial coronariana (cad) (sim/não); apetite (appet) (bom/ruim); edema pedial (pe) (sim/não); e anemia (ane) (sim/não)

B. pré processamento

O primeiro passo consistiu na correção da classificação das colunas numéricas com dados faltantes. Inicialmente, essas colunas estavam identificadas como categóricas no pandas, sendo necessário convertê-las para o tipo numérico para as etapas seguintes.

Em seguida, os valores faltantes foram preenchidos utilizando estratégias diferentes de acordo com o tipo de dado. Para colunas numéricas, adotou-se a imputação pela média, enquanto, para colunas categóricas, utilizou-se a moda.

Com o dataset completo, aplicou-se o *StandardScaler* para normalização dos dados em uma escala comum. Essa etapa é fundamental para algoritmos como SVM e KNN, que dependem de métricas de distância para classificação.

Quanto ao *encoding* das variáveis categóricas, optou-se pelo *Ordinal Encoding*, uma vez que o conjunto de dados incluía features nominais ordinais. Porém, esse método pode introduzir relações de ordem indesejadas em variáveis binárias (como ”sexo”), o que poderia prejudicar o desempenho do modelo. No entanto, os classificadores utilizados apresentaram resultados satisfatórios com a estratégia adotada, desencorajando a reavaliação do pipeline para outros métodos de *encoding*.

C. Algoritmos de aprendizado

Para a predição dos pacientes dois algoritimos de aprendizado supervisionado foram utilizados. O primeiro, o SVM que é baseado em distâncias e muito bom para o problema de predição de DRC.

Outro algoritmo utilizado foi o Random Forest, como haviam muitas variáveis nomimais o Random Forest poderia se beneficiar desse fato.

D. Avaliação dos algoritmos

Para avaliar o desempenho dos modelos, foram utilizadas as seguintes métricas: acurácia, F1-score, precisão e recall, que fornecem uma análise abrangente do poder de generalização dos algoritmos. Além disso, em casos específicos, matrizes de confusão foram geradas para permitir uma interpretação visual dos acertos e erros por classe.

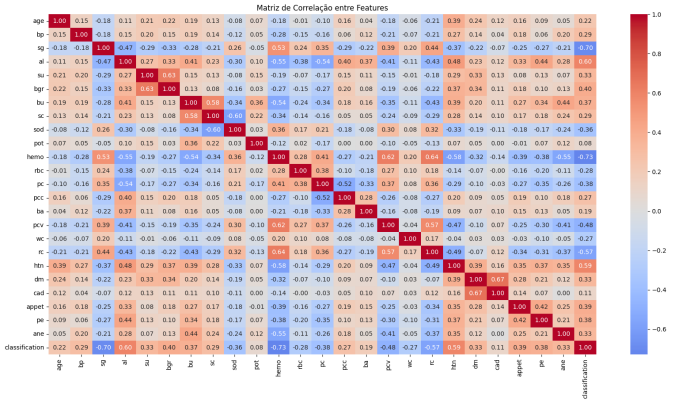

Fig. 1. Matriz de correlação entre features

A fim de garantir a robustez dos resultados, todas as métricas foram calculadas por meio de validação cruzada, técnica que reduz o viés de avaliação e aumenta a confiabilidade das medições reportadas.

IV. Desenvolvimento

A. Matriz de correlação

Para identificar se não havia nenhum vazamento de dados a partir das características, uma matriz de correlação (figura 1) foi gerada para essa análise e algumas características foram identificadas com forte correlação.

As características que demonstraram alta correlação com a Doença Renal Crônica (DRC) são a gravidade específica (sg) com uma correlação negativa de -0.70, a hemoglobina (hemo) com uma correlação negativa de -0.73 (explicada pela relação entre doença renal e anemia, que diminui a hemoglobina e as células vermelhas no sangue), e a contagem de glóbulos vermelhos (rc) com uma correlação positiva de 0.74 com a hemoglobina, devido à presença da hemoglobina nessas células. Além disso, a albumina (al) mostrou uma alta correlação positiva de 0.60, indicando que a alta presença dessa proteína na urina é um forte sinal de disfunção renal. Por fim, a hipertensão (htn) apresentou uma correlação de 0.59, atuando tanto como causa quanto como consequência de problemas renais.

B. Support Vector Machine

Para avaliar os algoritmos algumas abordagens foram utilizadas. No SVM uma matriz de confusão , que pode ser vista na figura 2, foi utilizada após o treinamento do modelo sem validação cruzada.

A fim de obter uma maior confiança nas métricas geradas utilizou-se um *Stratified K-Fold* com 10 *folds* e as métricas de f1-score, acurácia, etc podem ser vistas na figura 3.

Ao comparar os resultados obtidos com outros trabalhos correlacionados, optou-se pelo uso de *Grid Search* com validação cruzada de 10 *folds* para melhorar o desempenho do classificador e os resultados podem ser vistos na figura 4. Uma matriz de confusão pode ser visualizada na figura 5

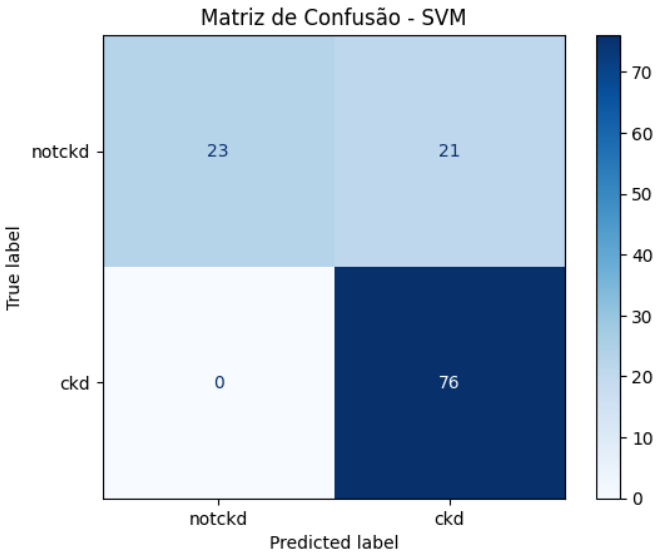

Fig. 2. Matriz de confusão do algoritmo SVM

```
Acurácias por fold: [0.89285714 0.92857143 0.85714286 0.89285714 0.82142857 0.92857143
 0.82142857 0.92857143 0.96428571 0.85714286]
Acurácia média: 0.8893

Métricas no conjunto de teste:
Acurácia: 0.8250
F1-score (macro): 0.7826
Recall (macro): 0.7614
```

Fig. 3. Resultados do SVM com Stratified K-Fold

C. Random Forest

A avaliação do algoritmo *Random Forest* foi conduzida por meio de validação cruzada estratificada com 10 *folds*. Após o treinamento, foram obtidas as métricas de acurácia, *F1-score* e *recall*, apresentadas na Figura 7. Além disso, foi gerada uma matriz de confusão (Figura 6) com o objetivo de facilitar a interpretação dos resultados da classificação.

V. Resultados e Discussão

Os classificadores demonstraram alta eficácia na predição, com excelentes métricas de recall (97,7% para Random Forest e 96,6% para SVM) nos dados de teste, identificando corretamente todos os casos positivos (zero falsos negativos) e cometendo apenas 2-3 falsos positivos. Esse desempenho é particularmente valioso em diagnósticos médicos, onde é preferível priorizar a detecção de todos os casos de doença,

```
Melhores parâmetros: {'C': 0.1, 'gamma': 'scale', 'kernel': 'linear'}
Melhor F1-score (refit): 0.9857142857142858
              precision    recall  f1-score   support

           0      1.000     0.932     0.965        44
           1      0.962     1.000     0.981        76

    accuracy                          0.975       120
   macro avg      0.981     0.966     0.973       120
weighted avg      0.976     0.975     0.975       120
```

Fig. 4. Métricas do SVM após o GridSearchCV

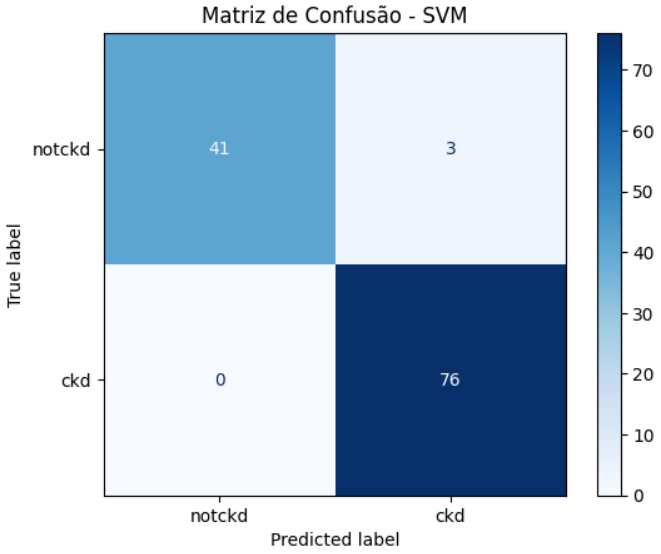

Fig. 5. Matriz de confusão do SVM após o grid search

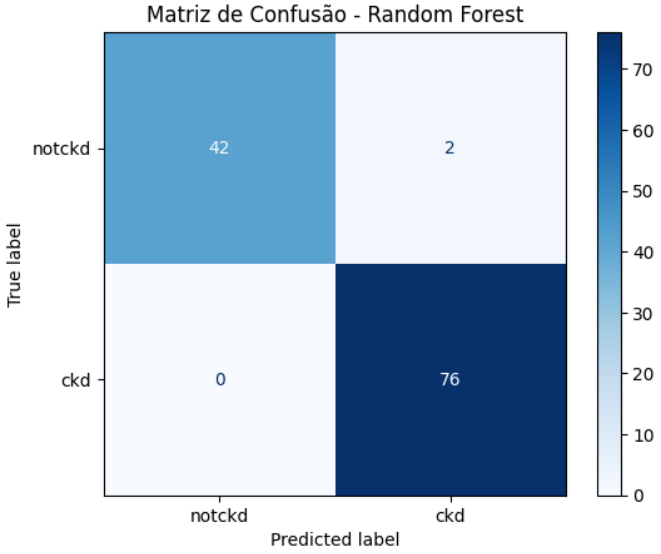

Fig. 6. Matriz de confusão do RF

```
Acurácias por fold: [0.96428571 1.         1.         0.96428571 0.96428571 0.96428571
 1.         1.         0.89285714 0.96428571]
Acurácia média: 0.9714

Métricas no conjunto de teste:
Acurácia: 0.9833
F1-score (macro): 0.9819
Recall (macro): 0.9773
```

Fig. 7. Métricas do RF

mesmo que isso resulte em alguns diagnósticos incorretos de pacientes saudáveis. A alta acurácia dos modelos provavelmente decorre da forte correlação entre as características fisiológicas analisadas e as complicações renais.

Ao analisar a matriz de correlação é possível destacar algumas características que tem forte relação preditiva com o diagnóstico de DRC.

A gravidade específica (sg) da urina é um dos atributos importantes no conjunto de dados para a classificação da DRC [2]. Sabe-se que os rins são órgãos essenciais que atuam na excreção de resíduos e excesso de fluidos do sangue [1]. No contexto da DRC, a funcionalidade dos rins é gradualmente perdida ao longo do tempo [1] [2]. A alteração na gravidade específica da urina pode indicar uma comprometimento na capacidade renal de concentrar ou diluir a urina, um sinal claro de disfunção. A inclusão dessa característica no conjunto de dados e sua utilização por algoritmos de ML sublinha sua importância como preditora da condição renal, permitindo que os modelos aprendam a correlação entre sua faixa de valores e a presença da doença [2].

Os níveis de hemoglobina (hemo) e a contagem de glóbulos vermelhos (rc) são marcadores cruciais para a detecção da DRC [2]. A doença renal crônica pode levar a uma diminuição no nível de contagem de células sanguíneas, resultando em anemia, que é uma complicação comum da DRC [2]. Como a hemoglobina está intrinsecamente ligada aos glóbulos vermelhos do sangue, uma baixa contagem de glóbulos vermelhos leva diretamente a baixos níveis de hemoglobina, explicando a alta correlação entre essas duas características [2]. Esses indicadores fornecem informações valiosas para os algoritmos de ML na identificação de pacientes com DRC, uma vez que a anemia é uma das consequências fisiológicas da progressão da doença [2].

A presença de albumina (al) na urina é um forte indicativo de doença renal [2]. A albumina é uma proteína essencial normalmente retida no sangue por rins saudáveis que atuam como filtros eficazes para excretar resíduos e excesso de fluidos [1]. Quando os rins estão danificados e não conseguem filtrar o sangue adequadamente, quantidades elevadas de albumina podem vazar para a urina, processo conhecido como albuminúria [1] [2]. Assim, uma alta quantidade de albumina na urina serve como um forte indicativo de disfunção renal, sendo uma característica de alta correlação e valor preditivo para os modelos de ML distinguirem pacientes com DRC de indivíduos saudáveis.

Por fim, a hipertensão (htn) é um fator de grande relevância no contexto da DRC, sendo listada como um sintoma associado à doença renal [2]. A relação entre hipertensão e DRC é complexa e bidirecional: a pressão arterial elevada pode causar danos aos rins, e, inversamente, rins danificados podem contribuir para o aumento da pressão arterial, pois perdem a capacidade de regular adequadamente os fluidos e eletrólitos do corpo. Essa interconexão faz da hipertensão uma característica vital para a classificação da DRC, auxiliando os algoritmos de ML a identificar padrões que levam à detecção precoce da doença.

A base de dados tem apenas 400 amostras, apesar de terem sido coletadas ao longo de 2 meses em pacientes que deram entrada no hospital, o número baixo implica em pouca diversidade dos dados. Esse fato pode contribuir para o bom desempenho dos classificadores nesse cenário.

Como possíveis melhorias futuras, sugere-se: (1) aplicar técnicas adicionais de pré-processamento, como *One-Hot Encoding* para variáveis categóricas binárias e remoção de outliers mediante análise de intervalos interquartis; (2) implementar métodos de seleção de características (*feature selection*) para otimizar o desempenho dos modelos, especialmente se houver expansão do volume de dados disponíveis. Contudo, o excelente desempenho atual dos classificadores (evidenciado pelas métricas reportadas) indica que as estratégias de limpeza adotadas foram adequadas, desencorajando a reavaliação da limpeza dos dados para buscar eficiência ou melhoria nas métricas.

VI. Conclusão

Este trabalho demonstrou a eficácia de algoritmos como Random Forest e SVM na predição de complicações renais, atingindo recalls superiores a 96% e minimizando falsos negativos — um critério essencial para aplicações médicas. A robustez dos resultados sugere que as características fisiológicas selecionadas possuem alta correlação com o diagnóstico, validando a abordagem adotada no pré-processamento e modelagem. O desempenho do Random Forest e do SVM sem a aplicação de algumas estratégias de pré-processamento indicam que nem sempre um alto refinamento dos dados implica em ótimo desempenho, vale ressaltar que nesse contexto específico, poucos dados estavam disponíveis para análise. Por fim, os resultados reforçam o potencial do machine learning como ferramenta auxiliar no diagnóstico, desde que aliado ao conhecimento médico especializado.

References

[1] El Sherbiny, M. M., Abdelhalim, E., Mostafa, H. E.-D., & El-Seddik, M. M. (2023). **Classification of chronic kidney disease based on machine learning techniques**. *Indonesian Journal of Electrical Engineering and Computer Science*, *32*(2), 945-955. DOI: 10.11591/ijeecs.v32.i2.pp945-955.

[2] Khamparia, A., Saini, G., Pandey, B., Tiwari, S., Gupta, D., & Khanna, A. (2019). **KDSAE: Chronic kidney disease classification with multimedia data learning using deep stacked autoencoder network**. *Multimedia Tools and Applications*. DOI: 10.1007/s11042-019-07839-z.

[3] Aswathy, R. H., Suresh, P., Sikkandar, M. Y., Abdel-Khalek, S., Alhumyani, H., Saeed, R. A., & Mansour, R. F. (2022). **Optimized Tuned Deep Learning Model for Chronic Kidney Disease Classification**. *Computers, Materials & Continua*, *70*(2), 2098-2111. DOI: 10.32604/cmc.2022.019790.

[4] Pandey, D., Niwaria, K., & Chourasia, B. (2019). **Machine Learning Algorithms: A Review**. *International Research Journal of Engineering and Technology (IRJET)*, *06*(02), 916–921.

[5] UCI Irvine, "Chronic KIdney Disease dataset", 2015. [Online]. Disponível em: https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease. Acessado em: 22 Jul. 2025.