AI in Stroke Prediction

Eduardo Motta
Bacharelando em Ciência da Computação
Universidade Federal de São Paulo
São José dos Campos, SP - Brasil
motta.eduardo@unifesp.br

***Abstract*—This study examines the application of artificial intelligence (AI) algorithms in predicting stroke, a major public health challenge globally. Utilizing the Stroke Prediction Dataset, we implement and compare three distinct AI classification algorithms: K-Nearest Neighbors (KNN), Gaussian Naive Bayes, and Decision Tree Classifier. Our objective is to evaluate the effectiveness of these algorithms in identifying individuals at risk of stroke, contributing to advancements in predictive medicine and public health strategies for stroke prevention.)**

Keywords—Stroke Prediction, Artificial Intelligence, Machine Learning, K-Nearest Neighbors, Gaussian Naive Bayes, Decision Tree Classifier, Healthcare.

I. INTRODUÇÃO

O Acidente Vascular Cerebral (AVC), também conhecido como derrame cerebral, é uma das maiores causas de mortalidade e incapacidade em todo o mundo, impactando milhões de pessoas anualmente. Esta emergência médica ocorre quando o fluxo de sangue para uma parte do cérebro é abruptamente interrompido, resultando em danos cerebrais graves. As implicações do AVC são profundas, variando de problemas de mobilidade a dificuldades de comunicação e perda de independência, afetando não apenas os pacientes, mas também suas famílias e sistemas de saúde.

Neste cenário, a detecção precoce e a prevenção do AVC tornam-se fundamentais. Com o advento da era dos dados e o desenvolvimento de tecnologias avançadas de IA, surge uma nova era na medicina preditiva. A IA oferece ferramentas poderosas para analisar grandes conjuntos de dados clínicos, identificando padrões e correlações que não são facilmente perceptíveis. Este trabalho se concentra na aplicação de algoritmos de aprendizado de máquina - K-Nearest Neighbors, Gaussian Naive Bayes e Decision Tree Classifier - para analisar dados clínicos e prever o risco de AVC. Estes algoritmos foram escolhidos devido à sua capacidade de lidar com grandes volumes de dados e complexidades variadas, oferecendo uma visão valiosa para a tomada de decisão clínica.

Ao empregar essas técnicas avançadas de IA no "Stroke Prediction Dataset", buscamos fornecer insights significativos para a prevenção do AVC, visando melhorar a qualidade de vida dos pacientes e reduzir a carga sobre os sistemas de saúde. Este estudo não só se alinha com os esforços globais para combater o AVC mas também contribui para o campo emergente da IA na saúde, propondo soluções inovadoras e orientadas por dados para um dos desafios mais prementes da saúde pública contemporânea.

II. TRABALHOS RELACIONADOS

A aplicação de Inteligência Artificial (IA) para a previsão de AVC é um campo de pesquisa emergente, com vários estudos destacando a sua eficácia. Um estudo notável de Kansadub et al. (2020) [1] ilustra a aplicação bem-sucedida do K-Nearest Neighbors (KNN) na análise de dados de saúde cardíaca, estabelecendo um precedente para o uso de algoritmos de classificação na previsão de condições médicas. Além disso, Dhilsath et al. (2021) [2] explorara a eficácia do Gaussian Naive Bayes na previsão de doenças crônicas, um trabalho que demonstra a aplicabilidade desses algoritmos em contextos clínicos variados.

Estudos adicionais, como os conduzidos por M. Ramla (2018) [3], enfatizam a importância das Árvores de Decisão na modelagem de dados de saúde complexos. Seu trabalho destaca a habilidade das Árvores de Decisão em lidar com múltiplas variáveis clínicas, com foco na saúde de fetos, um aspecto crucial para a análise de dados em previsões de AVC. Esses estudos formam a base teórica para a abordagem deste trabalho, destacando a relevância da seleção de múltiplos algoritmos para análise comparativa e resultados mais precisos.

Além disso, pesquisas sobre a integração de dados clínicos com algoritmos de IA fornecem insights valiosos para este estudo. Por exemplo, o trabalho de van Hartskam et al. (2019) [4] O estudo aborda o contexto da inteligência artificial (IA) na área biomédica, destacando os desafios específicos associados à aplicação da IA em saúde clínica. O artigo propõe seis recomendações, conhecidas como "6Rs", com o objetivo de aprimorar os projetos de IA nesse domínio e facilitar a comunicação entre cientistas de IA e profissionais de saúde. Essas recomendações incluem a importância de começar com questões clínicas relevantes, o uso de dados adequados e de alta qualidade, a adaptação do método de IA ao número de pacientes e variáveis, a criação de relações diretas entre dados e resultados, o preparo para requisitos regulatórios e a escolha do método de IA apropriado para tarefas específicas na área biomédica. Essas diretrizes visam melhorar a eficácia e a aplicabilidade da IA na saúde, com foco na saúde clínica. pesquisa ressalta a importância de combinar dados clínicos e O que se relaciona diretamente com o Stroke Prediction Dataset", para desenvolver modelos preditivos mais precisos e personalizados.

Finalmente, a literatura sobre os desafios e as limitações do uso de IA na medicina, como apresentado por Jessica Morley (2020) [5], oferece uma perspectiva crítica sobre a ética e a precisão na aplicação de IA na saúde. Enfatizando a necessidade de abordagens cuidadosas e bem pensadas, especialmente em condições médicas complexas como o AVC, garantindo que os modelos de IA sejam tanto eficazes quanto éticos.

III. JUSTIFICATIVA

Os Objetivos de Desenvolvimento Sustentável (ODS) 2030 das Nações Unidas representam um plano global para a promoção da saúde, bem-estar e desenvolvimento sustentável. O compromisso com tais Objetivos de Desenvolvimento Sustentável das Nações Unidas, especialmente o Tema 3: Saúde e Bem-Estar, é o alicerce deste trabalho.

O AVC, sendo uma das principais causas de morbidade e mortalidade global, requer uma atenção especial na saúde pública. A crescente integração de tecnologias avançadas como a IA na medicina oferece um caminho promissor para abordar essa questão premente. Ao empregar algoritmos de IA para a previsão de AVC, este estudo visa aprimorar as estratégias de prevenção e diagnóstico, melhorando assim a qualidade e a eficiência dos cuidados de saúde.

No Brasil, a incidência do AVC e suas repercussões são particularmente preocupantes. Segundo o Ministério da Saúde, a alta taxa de mortalidade associada ao AVC reflete desafios únicos relacionados ao estilo de vida e à demografia do país. A necessidade de soluções inovadoras é evidenciada não apenas pelos dados alarmantes de mortalidade, mas também pela crescente prevalência de fatores de risco como hipertensão, obesidade e diabetes. A aplicação de IA na previsão de AVC se apresenta como uma solução potencial, oferecendo uma ferramenta valiosa para os profissionais de saúde na identificação precoce de indivíduos em risco, permitindo intervenções preventivas mais eficazes.

Além disso, este estudo reforça a importância da medicina baseada em dados no contexto da saúde global. Ao alavancar o poder dos dados e da análise preditiva, podemos ultrapassar os limites da medicina tradicional, movendo-nos em direção a um modelo de saúde mais proativo e personalizado. Este avanço é essencial para atingir os objetivos globais de saúde e bem-estar, contribuindo para a criação de sistemas de saúde mais resilientes e adaptáveis às necessidades da população.

IV. Objetivo Geral

Este estudo tem como objetivo geral explorar a contribuição da Inteligência Artificial (IA) na saúde, com foco na prevenção e previsão de doenças, especialmente o AVC. Procuramos entender como diferentes metodologias e algoritmos de IA podem ser utilizados para analisar e interpretar dados de saúde complexos. O escopo deste objetivo abrange não apenas a aplicação prática desses algoritmos, mas também a avaliação de seu impacto na melhoria das estratégias de prevenção e no aprimoramento dos cuidados de saúde. Além disso, o estudo visa contribuir para a discussão sobre como a IA pode ser integrada de forma ética e eficaz no contexto da saúde pública, potencializando a medicina preditiva e a gestão de doenças crônicas.

V. Objetivos específicos

O estudo se concentra na implementação detalhada e na análise criteriosa de três algoritmos de classificação em IA: K-Nearest Neighbors, Gaussian Naive Bayes e Decision Tree Classifier. O objetivo é avaliar profundamente cada algoritmo em um contexto real de dados de saúde, focando em métricas chave para determinar sua precisão e confiabilidade na previsão de AVC. Esta análise permitirá identificar o algoritmo mais eficaz, proporcionando insights valiosos sobre sua aplicabilidade em situações clínicas reais.

Uma parte significativa do estudo está na exploração dos padrões e correlações dentro dos dados clínicos e de estilo de vida. O objetivo é ir além da mera aplicação de algoritmos, buscando compreender as nuances e características subjacentes nos dados que podem indicar um risco elevado de AVC. Esta análise detalhada tem o potencial de revelar novos fatores de risco ou padrões até então não identificados, contribuindo para a literatura existente e para o conhecimento médico.

Finalmente, o trabalho visa contribuir para a medicina preditiva, uma área em rápida expansão que combina dados, tecnologia e insights clínicos para prever e prevenir doenças. Ao aplicar e comparar esses algoritmos avançados de IA, busco fornecer recomendações concretas sobre como a tecnologia pode ser utilizada para melhorar a detecção precoce do AVC. Essa abordagem não só tem o potencial de salvar vidas, mas também de otimizar recursos nos sistemas de saúde, movendo-se em direção a um modelo de cuidado mais proativo e personalizado.

VI. Conjunto de Dados Utilizado

Primeiramente, é necessário destacar que a autoria do *dataset* [6] pertence ao senhor Federico Solianos Palacios, um cientista de dados natural de Madrid, Espanha, que possui dois mestrados em Data Science. Sua expertise no campo da ciência de dados é evidente na qualidade e na relevância dos dados coletados.

O contexto do *dataset* segundo o próprio é de que:

> "De acordo com a Organização Mundial da Saúde (OMS), o AVC é a 2ª principal causa de morte no mundo, responsável por aproximadamente 11% do total de mortes." – Palacios, Federico

Quanto ao objetivo do mesmo, ainda segundo o próprio autor:

> "Este conjunto de dados é usado para prever se um paciente tem probabilidade de sofrer acidente vascular cerebral com base em parâmetros de entrada como sexo, idade, várias doenças e tabagismo. Cada linha dos dados fornece informações relevantes sobre o paciente." – Palacios, Federico

Ademais, algumas considerações éticas e de privacidade informadas pelo autor são, primeiramente, de que fonte dos dados coletados para a montagem do conjunto é confidencial, e autorização para uso está feita apenas para fins educacionais, desde que o mesmo devidamente seja creditado, assim como realizado nesse *paper*.

O *dataset* contém um total de 5110 amostras, organizadas por linhas em uma tabela de formato '.csv'.

É de extrema importância destacar que mediante análise das amostrar, foi possível constatar que o *dataset* não está equilibrado, visto que há quantidade de indivíduos que não tiveram AVC é majoritariamente superior a aquelas de caso contrário. Esse fato influenciará diretamente o desempenho de qualquer algoritmo de Inteligência Artificial e Aprendizado de Máquina que trabalhe tendo esse conjunto de dados como base.

A. Features do Dataset

O *dataset* tem um total de doze (12) atributos – ou *features*- sendo estes:

1) id: Trata-se do identificador único da amostra, atributo do tipo categórico nominal.

2) gender: Trata-se do gênero do paciente, atributo do tipo categórico nominal. Pode assumir os valores de "Male", "Female" e "Other"

3) age: Trata-se da idade do paciente, atributo do tipo númerico contínuo.

4) hypertension: Trata-se de se o paciente possui hipertensão. Tem valor binário, sendo '1' caso o paciente possua a doença e '0' em caso negativo.

5) heart_disease: Trata-se de se o paciente possui alguma doença cardiovascular. Tem valor binário, sendo '1' caso o paciente possua a doença e '0' em caso negativo.

6) ever_married: Trata-se do estado civíl do paciente, atributo do tipo categórico nominal. Pode assumir or valores de "No" para caso o paciente não seja casado e "Yes" em caso positivo.

7) work_type: Trata-se do tipo de trabalho exercido pelo paciente, atributo do tipo categórico nomial. Pode assumir os valores de "children" caso o paciente apenas cuide dos filhos, "Govt_jov" caso seja um funcionário público", "Never_worked" caso o paciente nunca tenha realizado nenhum tipo de atividade profissional remunerada, "Private" caso o paciente esteja empregado no setor privado e "Self-employed" caso o paciente seja um trabalhador independente.

8) Residence_type: Trata-se do contexto geográfico em que o paceinte mora, atributo do tipo categórico nominal. Pode assumir o valor de "Rural" caso o paciente more em uma região rural ou "Urban" caso o paciente more em uma região urbana.

9) avg_glucose_level: Trata-se do nível médio de glicose no sangue, atributo do tipo número contínuo.

10) bmi: Trata-se do índice de massa corporal do paciente, atributo do tipo numérico contínuo.

11) smoking_status: Trata-se do status de tabagismo do paciente, atributo dotipo categórico nominal. Pode assumir os valores de "formely smoked" caso o paciente tenha fumado no passado, "never smoked" caso o paciente nunca tenha praticado o fumo, "smokes" caso o paciente tenha a prática do fumo ativa ou "Unknown" caso essa informação não esteja disponível para o paciente em questão.

12) stroke: Trata-se de se o paciente já teve um episódio de Acidente Vascular Cerebral (AVC). Tem valor binário, sendo '1' caso o paciente já tenha tido um AVC e '0' em caso negativo.

VII. METODOLOGIA

A. Linguagem de Programação e Ambiente de Desenvolvimento

Para a implementação dos algoritmos, o trabalho foi inteiramente realizado na linguagem de programação Python, como é de praxe em aplicações dessa natureza visto os recursos, bibliotecas e funções para essa linguagem que auxiliam muito o desenvolvedor e assim produzem resultados melhores.

Além disso, os algoritmos foram montados em um Notebook Jupyter, que permite escrever, executar e depurar códigos interativamente, possibilitando organizar e testar os códigos em células individuais, visualizando os resultados à medida que eles avançam.

Por fim, o ambiente de desenvolvimento foi o Google Colabotory (Google Colab), devido ao fato de ser online e armazenado no Google Cloud, o que permite o acesso de qualquer dispositivo com acesso à internet, assim como também é executado com os recursos (CPU e GPU) do próprio servidor, o que poupa os recursos da máquina pessoal que estará apenas acessando o Notebook no Google Colab.

B. Pré-Processamento de Dados

Devido ao fato de que todos os 3 algoritmos selecionados para implementação irão atuar no mesmo conjunto de dados, o pré-processamento de dados será de extrema valia para o desempenho final de todos.

A preparação e o pré-processamento dos dados são etapas críticas na construção de modelos de aprendizado de máquina. Para garantir a integridade e a qualidade dos dados, as seguintes etapas foram realizadas:

1) Limpeza de Dados: Inicialmente, a limpeza dos dados foi realizada, tendo como objetivo remover todas as amostras do dataset que estavam ou incompletas ou inconsistentess. Isso incluiu tanto o tratamento de valores ausentes quanto a detecção de outliers que poderiam afetar negativamente o desempenho final dos modelos implementados.Essa etapa só foi possível pelo da função 'dropna()'[7] da biblioteca Pandas.

2) Transformação de Variáveis Categóricas: O dataset "Stroke Prediction Dataset" continha diversas features do tipo categórico, sendo estas "gender", "ever_married", "work_type", "Residence_type" e "smoking_status", assim como listadas anteriormente. No entanto, afim de que estas pudessem ser utilizadas nos modelos de aprendizagem de máquina, o método "LabelEncoder()"[8], proveniente da biblioteca sklearn, foi aplicado com o objetivo de converter todas as variáveis categóricas listadas em valores númericos. Tal transformação foi crucial para o funcionamento correto dos algoritmos por garantir a total compatibilidade dos dados com os mesmos.

C. Divisão do Conjunto de Dados

A divisão apropriada do conjunto de dados em conjunto de treinamento e conjunto de teste é uma etapa crítica na construção e avaliação de modelos de aprendizado de máquina. Portanto, o conjunto original consistente de um total de 5110 amostras, como já mencionado, foi divido de forma que 80% deste, cerca de 4088 amostras, foram designadas como conjunto de treinamento, utilizado exclusivamente para treinamento dos modelos de IA. Enquanto os 20% restantes, um total de 1022 amostras, foram designadas como conjunto de teste. Este, por sua vez, mantido separado durante todo o processo de treinamento e ajuste de parâmetros, desempenhou um papel fundamental na avaliação imparcial dos modelos.

A razão por trás de tal divisão foi garantir que os modelos fossem avaliados em um conjunto de dados que não havia sido "visto" durante o treinamento. Essa abordagem, portanto, ajuda a minimizar e evitar problemas com *overfitting*, ou superajustamento, dos modelos aos dados de treinamento, fornecendo uma estimativa bem mais próxima da realidade em termos de desempenho cenários reais.

D. Implementação dos Algoritmoss

A escolha dos algoritmos a serem usados neste estudo desempenha um papel fundamental na construção de modelos de IA eficazes.

a) K-Nearest Neighbors Algorithm:O algoritmo K-Nearest Neighbors (KNN) é u m método de aprendizado supervisionado utilizado para classificação e regressão. Sua lógica é bastante simples. Primeiramente, o algoritmo verifica a classe da maioria dos vizinhos mais próximos desse ponto no espaço. Os vizinhos serão determinados com base na medida de distância adotada e são os "K" pontos mais próximos do novo ponto. O valor de K (número de vizinhos mais próximos considerados) é um hiperparâmetro do algoritmo. A escolha do K pode influenciar a eficácia do modelo.
Quanto a métrica de distância, a Distancia Euclidiana foi adotada, e para a seleção do parâmetro K, foi realizado um processo de ajuste para determinar o valor mais adequado afim de maximizar o desempenho do modelo, este sendo K = 5.
O modelo foi treinado com o conjunto de treinamento mencionado anteriormente. [9]

b) Gaussian Naive Bayes: O Gaussian Naive Bayes (GNB) é um classificador probabilístico que se baseia no Teorema de Bayes com a suposição de independência condicional entre os atributos do conjunto de dados. O algoritmo assume que os valores dos atributos são independentes entre si, dado o valor da classe. Ele funciona por meio do cálculo da probabilidade de um determinado ponto de dados pertencer a uma classe específica com base na distribuição normal (gaussiana) dos valores dos atributos para cada classe.
O modelo foi treinado com o conjunto de treinamento mencionado anteriormente. [10]

c) Decision Tree Classifier: O Decision Tree Classifier (DTC), ou Classificador de Árvore de Decisão, é um algoritmo de aprendizado supervisionado utilizado para tarefas de classificação e regressão. Ele constrói uma estrutura de árvore hierárquica na qual cada nó interno representa um atributo, cada ramo representa uma decisão baseada nesse atributo, e cada nó folha representa um rótulo de classe ou um valor de regressão.
O modelo foi treinado com o conjunto de treinamento mencionado anteriormente. [11]

Cada um dos três (3) algoritmos acima mencionados foi implementado com o objetivo de prever se um paciente tinha probabilidade de sofrer um AVC com base em parâmetros de entrada, como sexo, idade, histórico médico e hábitos de vida. O ajuste dos parâmetros e o treinamento dos modelos foram realizados com rigor para garantir que os resultados fossem confiáveis.

Esses modelos foram posteriormente avaliados quanto à sua eficácia e desempenho, como será discutido na seção "Avaliação dos Modelos".

E. Avaliação dos Modelos

A avaliação dos modelos desempenha um papel crítico na determinação da eficácia e confiabilidade das técnicas de IA aplicadas à previsão de AVC. Nesta seção, será descrito em detalhes como os modelos foram avaliados e as métricas utilizadas para medição de desempenho.

a) Accuracy: A acurácia é uma métrica que mede a proporção de previsões corretas em relação ao total de previsões. É uma medida geral da qualidade do modelo, mas pode ser enganosa em conjuntos de dados desequilibrados, como é o caso deste estudo. A acurácia não leva em consideração a distribuição das classes (AVC positivo e negativo).

b) Precision: A precisão mede a proporção de previsões positivas corretas (casos de AVC previstos corretamente) em relação ao total de previsões positivas. É uma métrica relevante quando se deseja evitar falsos positivos, ou seja, classificar erroneamente um paciente como tendo AVC quando na verdade não tem.

c) Recall: O recall, também conhecido como sensibilidade, mede a proporção de casos de AVC previstos corretamente em relação ao total de casos de AVC reais. É uma métrica importante quando o objetivo é identificar todos os casos de AVC, minimizando os falsos negativos.

d) F1-Score: O F1-score é uma métrica que combina precisão e recall em uma única medida. É útil quando se deseja um equilíbrio entre a capacidade de identificar casos de AVC (recall) e evitar falsos positivos (precisão). É calculado como a média harmônica entre precisão e recall

Os modelos foram avaliados usando as métricas mencionadas acima, e os resultados foram registrados e analisados. As métricas de avaliação proporcionaram informações valiosas sobre o desempenho de cada modelo em relação à previsão de AVC. É importante também a medição de desempenho só foi possível pelo uso da função “classification_report()” [12] da biblioteca sklearn que já realiza essa medição e retorna os dados em um *report*.

VIII. RESULTADOS OBTIDOS

A. K-Nearest Neighbors:

O modelo K-Nearest Neighbors (KNN) foi avaliado e apresentou os seguintes resultados:

1) *Acurácia do modelo KNN: 0.939 ou 93.9%.*
2) *Tempo de execução: 0.147 segundos.*
3) *Precisão para a classe 0: 0.95*
4) *Precisão para a classe 1: 0.00*
5) *Recall para a classe 0: 0.99*
6) *Recall para a classe 1: 0.00*
7) *F1-Score para a classe 0: 0.97*
8) *F1-Score para a classe 1: 0.00*
9) *Acurácia geral: 0.94*
10) *Precision médio: 0.47.*
11) *Recall médio: 0.50*
12) *F1-Score médio: 0.60*
13) *Support Total: 982*

B. Gaussian Naive Bayes::

O modelo Gaussian Naive Bayes (GNB) foi avaliado e apresentou os seguintes resultados:

1) *Acurácia do modelo GNB: 0.877 ou 87.7%.*
2) *Tempo de execução: 0.047 segundos.*
3) *Precisão para a classe 0: 0.96*
4) *Precisão para a classe 1: 0.20*

5) *Recall para a classe 0: 0.90*
6) *Recall para a classe 1: 0.42*
7) *F1-Score para a classe 0: 0.93*
8) *F1-Score para a classe 1: 0.27*
9) *Acurácia geral: 0.88*
10) *Precision médio: 0.58*
11) *Recall médio: 0.66*
12) *F1-Score médio: 0.60*
13) *Support Total: 982*

C. Decision Tree Classifier:

O modelo Decision Tree Classifier (DTC) foi avaliado e apresentou os seguintes resultados:

1) *Acurácia do modelo GNB: 0.925 ou 92.5%.*
2) *Tempo de execução: 0.047 segundos.*
3) *Precisão para a classe 0: 0.95*
4) *Precisão para a classe 1: 0.22*
5) *Recall para a classe 0: 0.97*
6) *Recall para a classe 1: 0.15*
7) *F1-Score para a classe 0: 0.96*
8) *F1-Score para a classe 1: 0.18*
9) *Acurácia geral: 0.92*
10) *Precision médio: 0.58*
11) *Recall médio: 0.56*
12) *F1-Score médio: 0.57*
13) *Support Total: 982*

IX. Análise e Comparação de Desempenho

Afim de atingir os objetivos propostos na realização do trabalho é necessário analisar individualmente os resultados obtidos e expostos anteriormente, assim como comparar o desempenho dos 3 algoritmos implementados a fim de determinar qual foi mais eficaz e, portanto, seria mais apropriado para utilização.

O modelo KNN apresentou uma acurácia geral de aproximadamente 93.89%, o que indica que a maioria das previsões está correta. No entanto, ao analisar as métricas de classificação para a classe 1 (AVC), observamos que o recall, a precision e o F1-score são todos iguais a zero. Isso significa que o modelo não conseguiu identificar nenhum caso positivo de AVC. Possíveis motivos para esse desempenho incluem principalmente a falta de balanceamento de classes e sensibilidade ao ruído nos dados.

Por fim. o modelo Decision Tree Classifier (DTC) obteve uma acurácia de aproximadamente 92.46%, que está entre os modelos anteriores em termos de desempenho geral. No entanto, ao analisar as métricas de classificação para a classe 1 (AVC), observamos um recall e um F1-score ainda mais baixos em comparação com o Gaussian Naive Bayes. Isso indica que o DTC teve dificuldades em identificar casos positivos de AVC. Possíveis motivos incluem a complexidade das decisões tomadas pela árvore de decisão e a necessidade de maior ajuste de hiperparâmetros para evitar *overfitting*.

Ao comparar o desempenho dos três algoritmos, é evidente que cada um deles possui suas vantagens e desvantagens. O KNN obteve a maior acurácia geral, mas falhou na identificação de casos de AVC. O Gaussian Naive Bayes teve um desempenho razoável na identificação de casos de AVC, mas sua acurácia geral foi menor que a do KNN. O DTC se encontra em uma posição intermediária em termos de acurácia geral, mas também teve dificuldades em identificar casos positivos de AVC.

Com base na análise, o desempenho de todos os algoritmos na identificação de casos positivos de AVC (classe 1) não foi satisfatório. Portanto, é necessário um aprimoramento significativo dos modelos para torná-los clinicamente úteis em cenários reais.

A escolha do melhor algoritmo depende dos objetivos específicos do projeto. Se o foco for na acurácia geral, o KNN é a escolha, embora precise de melhorias na identificação de casos de AVC. Se a prioridade for identificar casos de AVC, o Gaussian Naive Bayes mostrou um desempenho relativamente melhor nesse aspecto. O DTC se encontra em uma posição intermediária.

X. Conclusão

Neste estudo, foi explorada a aplicação de técnicas de Inteligência Artificial (IA) na previsão de AVC utilizando o "Stroke Prediction Dataset". Ao longo deste trabalho, foram realizadas diversas etapas, desde a preparação e pré-processamento dos dados até a implementação de três diferentes algoritmos de aprendizado de máquina: K-Nearest Neighbors (KNN), Gaussian Naive Bayes e Decision Tree Classifier (DTC). Além disso, conduzimos uma análise detalhada dos resultados obtidos por cada algoritmo e comparamos seus desempenhos.

Os resultados revelaram que, embora todos os modelos tenham obtido resultados insatisfatórios na identificação de casos positivos de AVC, cada um deles possui suas próprias vantagens e desvantagens. O KNN obteve a maior acurácia geral, mas falhou na identificação de casos de AVC. O Gaussian Naive Bayes mostrou um desempenho relativamente melhor na identificação de casos de AVC, mas teve uma acurácia geral menor que a do KNN. O DTC se encontra em uma posição intermediária em termos de acurácia geral e identificação de casos de AVC.

A escolha do melhor algoritmo depende dos objetivos específicos do projeto, destacando a importância de considerar tanto a acurácia geral quanto a capacidade de identificar casos de AVC. É fundamental ressaltar que este é um trabalho em andamento, e futuras iterações podem envolver aprimoramentos nos modelos, além da exploração de outros algoritmos e técnicas de aprendizado de máquina.

Além disso, este estudo se justifica pela urgente necessidade de desenvolver métodos mais eficientes e precisos para a prevenção e diagnóstico precoce do AVC. O AVC é uma das principais causas de morte e incapacidade em todo o mundo, representando um desafio significativo na área da saúde pública. Com o avanço da tecnologia e o crescente acesso a grandes volumes de dados de saúde, a aplicação da IA na previsão de AVC oferece uma oportunidade sem precedentes para melhorar a qualidade de vida dos pacientes e reduzir o fardo nos sistemas de saúde.

Ademais, este trabalho está alinhado aos Objetivos de Desenvolvimento Sustentável das Nações Unidas, em particular ao objetivo de garantir uma vida saudável e promover o bem-estar para todos. Ao aplicar a IA na previsão do AVC, contribuímos para o avanço da medicina preditiva e para a construção de sistemas de saúde mais resilientes e eficazes.

Em suma, este estudo representa um passo inicial na direção de soluções mais eficazes na previsão de AVC, com o potencial de impactar positivamente a saúde pública e a qualidade de vida dos pacientes. O trabalho continua em progresso, com o compromisso de buscar melhorias contínuas e avanços na prevenção e diagnóstico precoce do AVC.

REFERÊNCIAS

[1] Tahia Tazin, Md Nur Alam, Nahian Nakiba Dola, Mohammad Sajibul Bari, Sami Bourouis, Mohammad Monirujjaman Khanc "Stroke Disease Detection and Prediction Using Robust Learning Approaches" Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8641997/. Acesso em 22 de novembro de 2023

[2] Dhilsath Mohideen, Justin Raj, Raja Raj. "Regression Imputation and Optimized Gaussian Naïve Bayes Algorithm for an Enhanced Diabetes Mellitus Prediction Model". Disponível em: https://www.scielo.br/j/babt/a/3HDJLgqSPVYcSDT4RFY9wzp/?lang=en#. Acesso em 22 de novembro de 2023

[3] M. Ramla; S. Sangeetha; S. Nickolas. "Fetal Health State Monitoring Using Decision Tree Classifier from Cardiotocography Measurements". Disponível em: https://ieeexplore.ieee.org/abstract/document/8663047?casa_token=mDabxWPu6BoAAAAA:PIhP6dB0KRRvA4_f_xjJq1y3Abx9JcypXse3Kt0UI0YTP0E6p9oG6s0lDge-pgQxl0spB4Es-PlI4Q. Acesso em: 22 de Novembro de 2023.

[4] Michael van Hartskamp, Sergio Consoli, Wim Verhaegh, Milan Petkovic, Anja van de Stolpe. "Artificial Intelligence in Clinical Health Care Applications: Viewpoint". Disponível em: https://i-jmr.org/2019/2/e12100. Acesso em 22 de novembro de 2023

[5] Jessica Morley, Caio C.V. Machado, Christopher Burr, Josh Cowls, Indra Joshi, Mariarosaria Taddeo, Luciano Floridi. "The ethics of AI in health care: A mapping review". Disponível em: https://www.sciencedirect.com/science/article/pii/S0277953620303919?casa_token=Ao78ovm39R0AAAAA:Y3dWau5Ec_GQzCzgff-QbBgsHp2NjBoHtCkk6xTDuvsbYkd-EShPVHy_QQNwT0gLJr_8hCMtf-A . Acesso em 23 de novembro de 2023

[6] Soriano, Federico. "Stroke Prediction Dataset". Disponível em: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data. Acesso em: 28 de Novembro de 2023.

[7] Pandas. "pandas.DataFrame.dropna". Disponível em: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.dropna.html.. Acesso em: 28 de Novembro de 2023.

[8] Scikit Learn. "sklearn.preprocessing.LabelEncoder". Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html#sklearn.preprocessing.LabelEncoder. Acesso em: 28 de Novembro de 2023.

[9] Scikit Learn, "sklearn.neighbors.KNeighborsClassifier". Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html#sklearn.neighbors.KNeighborsClassifier . Acesso em: 28 de Novembro de 2023.

[10] Scikit Learn, "sklearn.naive_bayes.GaussianNB". Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html#sklearn.naive_bayes.GaussianNB. Acesso em: 28 de Novembro de 2023

[11] Scikit Learn, "sklearn.tree.DecisionTreeClassifier". Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier. Acesso em: 28 de Novembro de 2023

[12] Scikit Learn, "sklearn.metrics.classification_report". Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html#sklearn.metrics.classification_report. Acesso em: 28 de Novembro de 2023