10º Encontro de Ensino Pesquisa e Extensão

Patrocínio, MG, outubro de 2023

ANÁLISE COMPARATIVA DE TÉCNICAS DE APRENDIZADO DE MÁQUINAS PARA AUXÍLIO NA PREDIÇÃO DE DIABETES MELLITUS

Daniel Gomes Januario Júnior, Danielli Araújo Lima
Instituto Federal do Triangulo Mineiro (IFTM) Campus Patrocínio
daniel.junior@estudante.iftm.edu.br, danielli@iftm.edu.br
Modalidade: Pesquisa
Formato: Artigo Completo

Resumo: A diabetes mellitus é um problema global. Muitas pessoas vivem com a doença sem saber, até que os sintomas se tornem graves, prejudicando sua saúde. O foco da pesquisa é avaliar algoritmos de classificação de diabetes usando dados públicos de pessoas com e sem a doença. Este estudo usa sete algoritmos de aprendizado de máquina para auxiliar no diagnóstico de diabetes mellitus. A pesquisa avalia o desempenho de cada algoritmo e destaca a importância de escolher o algoritmo certo com base nos dados e nos objetivos de predição. O Support Vector Classifier demonstrou uma acurácia geral notável, mantendo uma Precision elevada, o que o torna uma escolha sólida para cenários onde é fundamental evitar falsos positivos, assim como o algoritmo Logistic Regression. Isso é relevante para a área de previsão de diabetes e abre oportunidades para mais pesquisas em inteligência artificial e saúde pública.
Palavras-chaves: Mineração de Dados, Aprendizado de Máquina, Inteligência Artificial, Diabetes Mellitus, Support Vector Classifier, Logistic Regression.

Introdução

A diabetes mellitus é uma doença crônica que afeta a forma como o corpo usa a glicose, o açúcar no sangue, para energia. É uma condição médica com grande prevalência global e impacto na saúde pública. Segundo a Federação Internacional de Diabetes[1] (IDF)

[1]"Diabéticos podem chegar a 784 milhões no mundo em 2045, estima IDF". Link de acesso 24 de setembro de 2023: https://agenciabrasil.ebc.com.br/saude/noticia/2021-11/diabeticos-podem-chegar-784-milhoes-no-mundo-em-2045-estima-idf.

em 2021, cerca de 537 milhões de adultos (20-79 anos) têm a doença, o que equivale a 1 em 10 pessoas. Espera-se que esses números cresçam para 643 milhões até 2030 e 783 milhões até 2045, desafiando os sistemas de saúde mundiais. Há vários tipos de diabetes: Tipo 1, Tipo 2 e Gestacional (GARCÍA, 2008). O Tipo 1 afeta pessoas de todas as idades, é autoimune, requer insulina diária. O Tipo 2, mais comum, geralmente afeta adultos, mas também jovens devido à obesidade. Pode exigir medicamentos e mudanças no estilo de vida. A Gestacional ocorre em grávidas, normalmente desaparece após o parto, mas aumenta o risco de Tipo 2 mais tarde (BASSO et al., 2007).

A diabetes é uma doença que pode causar sérias consequências para a saúde já que está intimamente ligada a um aumento do risco de doenças cardiovasculares, acidentes vasculares cerebrais, insuficiência renal e neuropatia, entre outras condições graves (SCHEFFEL et al., 2004). A diabetes foi responsável por uma quantidade alarmante de mortes em 2021, com 6,7 milhões de óbitos registrados[2], o que equivale a uma morte a cada 5 segundos. Essas estatísticas enfatizam a importância crítica da prevenção, do diagnóstico precoce e do tratamento eficaz da diabetes como medidas para salvar vidas.

Atualmente, uma variedade de métodos estão à disposição para identificar a diabetes em suas fases iniciais, porém, é imperativo intensificar as investigações com o intuito de aprimorar tanto a detecção quanto o tratamento da doença. Este artigo tem como objetivo analisar o uso de sete algoritmos de aprendizado de máquina para prever diabetes. Para realizar essa análise será utilizado um conjunto de dados obtidos no Repositório de Aprendizado de Máquinas do Kaggle[3], ampliando nosso entendimento sobre a aplicação eficaz dessas ferramentas na área médica.

Fundamentação Teórica

Nesta seção, resumiremos o conhecimento teórico essencial para entender o modelo de predição de dados. Começaremos com os conceitos fundamentais da diabetes e suas características. Em seguida, apresentaremos os algoritmos de predição usados neste

[2]Site UOL Viva Bem "Diabetes mata 1 pessoa a cada 5 segundos no mundo, mostram dados inéditos" `https://www.uol.com.br/vivabem/noticias/redacao/2021/11/05/diabetes-mata-1-pessoa-a-cada-cinco-segundos-no-mundo-mostra-pesquisa.htm`.

[3]Diabetes prediction dataset A Comprehensive Dataset for Predicting Diabetes with Medical & Demographic Data. Link disponível em `https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset`.

estudo. Por fim, destacaremos as ferramentas de análise de dados.

A diabetes mellitus é uma doença metabólica crônica de grande relevância global, afetando o uso da glicose no corpo para produzir energia. É amplamente disseminada, afetando milhões de pessoas em todo o mundo, incluindo aquelas com risco aumentado, como tolerância diminuída à glicose (TDG) e glicemia de jejum alterada (GJA), que requerem intervenções para minimizar o risco (LYRA et al., 2023). Estudos em todo o mundo confirmam a eficácia de abordagens comportamentais, como mudanças na dieta e exercícios, e do uso de terapias farmacológicas, como a metformina, para controlar o diabetes Tipo 2. Outras terapias incluem sulfonilureas, tiazolidinedionas, inibidores DPP-4, inibidores SGLT-2, insulina e agonistas GLP-1. O diagnóstico precoce e o tratamento adequado são cruciais para o sucesso no tratamento da diabetes, melhorando o controle glicêmico e reduzindo complicações futuras nos pacientes (VARGAS et al., 2020). Estudos nesta área abordam modelos de previsão de diabetes Tipos 1 e 2, incluindo gestantes, uma classe que requer atenção devido aos riscos para mãe e bebê (MOREIRA et al., 2021). A detecção precoce é crucial para intervenções eficazes e para reduzir os impactos negativos da diabetes.

A área do aprendizado de máquina é promissora na previsão de diabetes, usando modelos que aprendem com experiências passadas (FACELI et al., 2011). Esses modelos, baseados em aprendizado indutivo, supervisionado e não supervisionado (COSTA-FILHO et al., 2019; SOUZA; LIMA, 2023), visam melhorar a detecção precoce e o tratamento da doença, prevendo sua ocorrência em indivíduos. O Aprendizado de Máquina (AM) visa melhorar o desempenho através de exemplos (JORDAN; MITCHELL, 2015), sendo orientado por dados e métodos indutivos (DORNELAS; LIMA, 2023). A qualidade dos dados influencia a precisão das generalizações.

O AM teve raízes no século XX e cresceu nas décadas de 1950 e 1960, com o modelo Perceptron (HAYKIN, 1994). Nos anos 80, surgiram redes neurais mais profundas, como o Multilayer Perceptron (MLP) (HAYKIN, 1994), embora enfrentassem desafios de treinamento devido à falta de dados e poder computacional (HAYKIN, 1994; LIMA; FERREIRA; SILVA, 2021). À medida que Redes Neurais Artificiais (RNAs) complexas surgiram, outros métodos ganharam destaque, como SVC, Regressão Logística e Árvores de Decisão (DANTAS; DONADIA, 2013; ALVES; LIMA, 2018; ABREU; SIQUEIRA; CAIAFFA, 2009; SILVA; SILVA FILHO, 2010). Além disso, a

combinação de algoritmos levou a abordagens como Random Forest, boosting e stacking (FRIEDMAN, 2001). A evolução recente inclui o Aprendizado Profundo (Deep Learning), que emprega múltiplas estratégias de aprendizado para predições mais precisas. No contexto deste estudo, serão investigadas diversas técnicas de predição de diabetes, incluindo o uso de algoritmos como o Support Vector Classifier (SVC), Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbors (KNN), Gaussian Naive Bayes e Multilayer Perceptron (MLP) Neural Network.

1. Support Vector Classifier (SVC): É um algoritmo que ajuda a separar diferentes grupos de dados traçando uma linha, chamada hiperplano, entre eles. Ele busca maximizar a distância entre os pontos de dados mais próximos de cada grupo.

2. Logistic Regression Learner (LRL): Não é exatamente uma regressão no sentido tradicional, mas um algoritmo usado para prever resultados binários, como `SIM` ou `NÃO`. Ele cria uma curva em forma de S (curva logística) para estimar a probabilidade de um resultado pertencer a uma categoria específica.

3. Random Forest Learner (RFL): É um conjunto de árvores de decisão, onde várias árvores são criadas e suas previsões são combinadas para obter resultados mais robustos e precisos. É usado para classificação e regressão.

4. Decision Tree Learner (DTL): É uma estrutura em forma de árvore que ajuda a tomar decisões com base em uma série de condições. Cada nó representa uma escolha e cada folha representa um resultado.

5. K-Nearest Neighbors (KNN): Este algoritmo classifica um ponto de dados com base na maioria dos pontos vizinhos a ele. O K representa o número de vizinhos mais próximos a serem considerados para a classificação.

6. Gaussian Naive Bayes (GNB): Utiliza o Teorema de Bayes para prever a probabilidade de um ponto de dados pertencer a uma categoria específica. "Naive" significa que ele assume independência entre as características, o que pode simplificar o cálculo.

7. Multilayer Perceptron (MLP) Neural Network: É um tipo de rede neural artificial com múltiplas camadas de neurônios interconectados. É usado para problemas

complexos de aprendizado de máquina, como reconhecimento de padrões e processamento de linguagem natural.

Cada um dos algoritmos de mineração de dados aqui estudados possui suas vantagens e desvantagens. A aplicação de cada um deles depende da natureza e tamanho do conjunto de dados.

Materiais e métodos

Esta pesquisa é classificada como uma pesquisa quantitativa e é aplicada à área de saúde, uma vez que se baseia na coleta e análise de dados numéricos para responder a perguntas de pesquisa específicas. A coleta de dados foi realizada por meio de questionários estruturados, aplicados a uma amostra representativa da população de estudo, neste caso, coletamos os dados de uma base no Kaggle Repository. As fontes de informação incluem respostas diretas dos participantes aos questionários. Quanto à análise de dados, será empregada uma abordagem estatística, com a utilização de software estatístico para processar e interpretar os dados coletados, permitindo a identificação de padrões, tendências e relações estatisticamente significativas. Mais precisamente, usaremos técnicas de AM para a interpretação de resultados.

Para o processo de normalização dos dados, conforme (CONTINI; CORRAR; FILHO, 2002), amostragem é a seleção representativa de uma parcela da população em pesquisas científicas. A qualidade da amostra é crucial, exigindo critérios de amostragem que variam com os dados. A saturação é um critério comum quando não há novidades ou relevância nos dados (FONTANELLA et al., 2011). Além disso, a preparação de dados organiza e melhora sua qualidade. A padronização e normalização são técnicas usadas para tornar as variáveis comparáveis. A normalização, usada neste trabalho, ajusta os valores para uma escala de 0 a 1, usando a fórmula Min-Max, conforme é apresentado na Equação 1.

$$z' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

Já a padronização, normalmente é feita usando a fórmula z-score apresentada na Equação 2.

$$z' = \frac{x - \mu}{\sigma} \tag{2}$$

Algumas métricas são comumente usadas para avaliar o desempenho de modelos de classificação em aprendizado de máquina. Para entender cada uma dessas métricas, inicialmente é necessário compreender os seguintes conceitos:

TP True Positives: Representa o número de exemplos positivos que foram corretamente classificados como positivos pelo modelo.

FP False Positives: Indica o número de exemplos negativos que foram incorretamente classificados como positivos pelo modelo.

TN True Negatives: É o número de exemplos negativos que foram corretamente classificados como negativos pelo modelo.

FN False Negatives: Representa o número de exemplos positivos que foram incorretamente classificados como negativos pelo modelo.

Métricas essenciais para avaliar modelos de classificação em aprendizado de máquina incluem o "Recall" (ou "Recuperação" ou "Sensibilidade"), que indica a habilidade do modelo em identificar exemplos positivos corretamente, calculado pela proporção de exemplos positivos corretamente classificados em relação ao total de positivos reais. A "Precision" avalia a precisão das previsões positivas, representando a proporção de exemplos positivos corretamente classificados em relação ao total classificado como positivos pelo modelo. A "Specificity" mensura a capacidade do modelo em identificar negativos corretamente, calculada pela proporção de exemplos negativos corretamente classificados em relação ao total de negativos reais. A "F-measure" combina Recall e Precision, útil para equilibrar precisão e capacidade de identificar positivos. A "Accuracy" avalia a precisão geral do modelo, representando a proporção de previsões corretas em relação ao total de previsões. Finalmente, o "Cohen's Kappa" mede a concordância considerando a concordância esperada por acaso entre previsões e valores reais, sendo particularmente útil em desbalanceamento de classes. Essas métricas são cruciais para analisar o desempenho de modelos de classificação em diversos aspectos.

Para esta pesquisa, selecionamos rigorosamente uma base de dados no Kaggle

com informações relevantes para prever diabetes, considerando critérios como relevância, tamanho da amostra, qualidade e disponibilidade. A base contém cerca de 100 mil registros de pacientes, incluindo dados médicos, demográficos e status de diabetes `YES` (positivo) ou `NO` (negativo). Ela inclui variáveis como idade, gênero, IMC, hipertensão, doença cardíaca, histórico de tabagismo, HbA1c e níveis de glicose. Usaremos esses dados para construir modelos de aprendizado de máquina que auxiliam profissionais de saúde na identificação de riscos e tratamentos personalizados, bem como permitem que pesquisadores explorem relações entre fatores médicos e o desenvolvimento de diabetes.

Para analisar a base de dados, utilizamos diversas ferramentas em Python. Primeiramente, examinamos os valores únicos em cada coluna, orientando as estratégias de transformação de dados. Identificamos um desequilíbrio na classe dependente (diabetes), onde a classe sem diabetes representava a maioria esmagadora dos registros (91,5%), enquanto a classe com diabetes era minoritária. Uma análise de dispersão revelou sobreposição entre os valores das duas classes. Em seguida, treinamos um modelo de regressão logística após aplicar técnicas de normalização e transformação de dados, como substituir valores não numéricos por numéricos nas colunas "gênero"e "histórico de fumante". O modelo inicial alcançou uma promissora acurácia geral de 96%, mas apresentou baixa acurácia para a classe com diabetes (72%), em contraste com 98% para a classe sem diabetes, destacando a influência do desbalanceamento de classes.

Para enfrentar esse desafio, foi aplicada uma técnica de reamostragem usando o IHT (Instance Hardness Threshold) para equilibrar as classes. Isso resultou em uma melhora significativa na acurácia geral do modelo, atingindo 99% de acurácia para ambas as classes da base de dados. A aplicação do IHT equilibrou as classes, mantendo os dados da classe majoritária que eram mais adequados ao modelo. A análise mostrou que o processo de reamostragem teve um impacto positivo no desempenho dos modelos, melhorando a capacidade de prever ambas as classes de forma equitativa.

Resultados e discussão

Nesta seção apresentaremos os resultados da nossa análise de dados. Primeiramente, a Figura 1 apresenta o conjunto de dados por meio de dois parâmetros o nível de glicose no sangue e o parâmetro gênero. Neste caso, a maioria é do gênero masculino e no

caso do nível de glicose, existem muitas pessoas com taxas na casa de 150 (glicemia). A Tabela 1 apresentada fornece uma visão abrangente das medidas de desempenho de

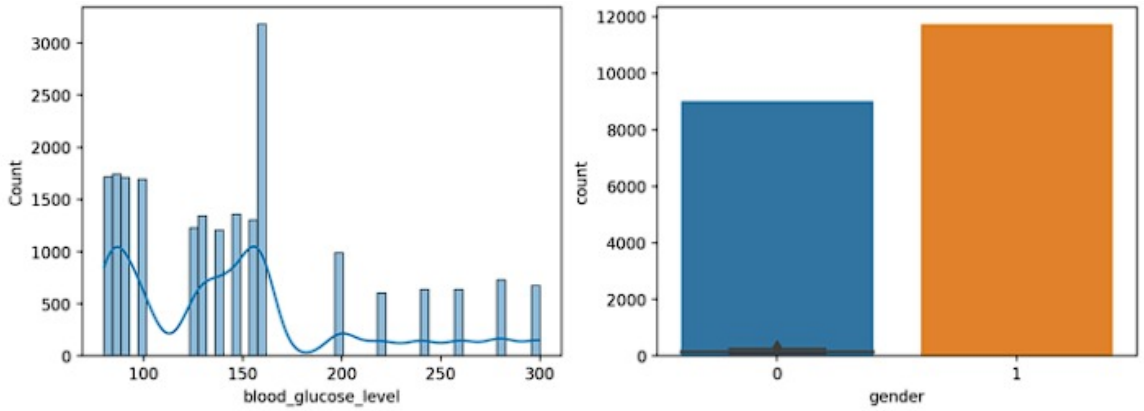

Figura 1: Atributos de índices de glicose no sangue e genero.

diferentes algoritmos de mineração de dados na tarefa de prever diabetes, com uma análise específica que exclui o parâmetro Hba1c. A análise dos resultados obtidos com

Tabela 1: Medidas de desempenho dos algoritmos de mineração de dados para a previsão da diabetes sem o parâmetro Hba1c.

Classifier	TP	FP	FN	TN	Specificity	Sensitivity	Recall	Precision	F1-Score	F-measure	Cohen's Kappa	Accuracy	Overall Accuracy
SVC	12193	50	84	8416	0.626865672	0.591634723	0.990117647	0.994094023	0.992101851	0.992101851	0.986636961	0.993539989	0.993539989
LRL	12202	41	96	8404	0.700729927	0.592157624	0.988705882	0.995145056	0.991915019	0.991915019	0.986332624	0.993395362	0.993395362
RFL	12243	0	0	8500	0	0.590223208	1	1	1	1	1	1	1
DTL	12243	0	0	8500	0	0.590223208	1	1	1	1	1	1	1
KNN	12241	2	126	8374	0.984375	0.593790929	0.985176471	0.999761223	0.992415264	0.992415264	0.9872146	0.993829244	0.993829244
GNB	12243	0	2676	5824	1	0.677644324	0.685176471	1	0.813180676	0.813180676	0.719817565	0.870992624	0.870992624
MLP	12237	6	4	8496	0.4	0.590218492	0.999529412	0.999294284	0.999411834	0.999411834	0.999003404	0.99951791	0.99951791

diferentes algoritmos de classificação, desconsiderando o atributo Hba1c da diabetes, revela insights importantes sobre o desempenho dos modelos nesse cenário específico. Os algoritmos Random Forest Learner (RFL) e Decision Tree Learner (DTL) alcançaram uma acurácia perfeita de 100%, o que inicialmente pode parecer impressionante, mas também suscita preocupações sobre possíveis overfitting, uma vez que eles não identificaram nenhum falso negativo. Já os demais algoritmos como o Support Vector Classifier (SVC) e o Logistic Regression Learner (LRL), por exemplo, mantiveram altos valores de Precision e Recall, indicando que suas previsões positivas tendem a ser precisas. Já o Gaussian Naive Bayes (GNB) obteve um valor de Recall relativamente baixo. Isso sugere que esse modelo pode estar sendo excessivamente conservador na classificação de casos positivos, o que é justificado analisando o valor geral de precisão do mesmo, que dentre os analisados, foi o pior. Já o MLP Neural Network obteve um desempenho notável, com alta acurácia, Recall e Precision, sugerindo sua utilidade em

cenários onde a previsão precisa de diabetes é essencial.

A Tabela 2 apresenta as medidas de desempenho de diferentes algoritmos de mineração de dados para a previsão de diabetes com base no parâmetro Hba1c. A análise

Tabela 2: Medidas de desempenho dos algoritmos de mineração de dados para a previsão da diabetes com o parâmetro Hba1c.

Classifier	TP	FP	FN	TN	Specificity	Sensitivity	Recall	Precision	F1-Score	F-measure	Cohen's Kappa	Accuracy	Overall Accuracy
SVC	12193	50	84	8416	0.626865672	0.591634723	0.990117647	0.994094023	0.992101851	0.992101851	0.986636961	0.993539989	0.993539989
LRL	12202	41	96	8404	0.700729927	0.592157624	0.988705882	0.995145056	0.991915019	0.991915019	0.986332624	0.993395362	0.993395362
RFL	12243	0	0	8500	0	0.590223208	1	1	1	1	1	1	1
DTL	12243	0	0	8500	0	0.590223208	1	1	1	1	1	1	1
KNN	12241	2	126	8374	0.984375	0.593790929	0.985176471	0.999761223	0.992415264	0.992415264	0.9872146	0.993829244	0.993829244
GNB	12243	0	2676	5824	1	0.677644324	0.685176471	1	0.813180676	0.813180676	0.719817565	0.870992624	0.870992624
MLP	12220	23	5	8495	0.178571429	0.589910693	0.999411765	0.997299836	0.998354683	0.998354683	0.997210334	0.998650147	0.998650147

dos resultados agora com a variável de Hba1c (valor do exame de hemoglobina glicada) ressalta a capacidade preditiva do MLP Neural Network de se adaptar aos valores da base e a sensitividade do mesmo em se ajustar dependendo de cada situação presente nos modelos trabalhados. Nessa nova análise, novamente, os algoritmos SVC e LRL alcançaram altas taxas de acurácia global, ambas superiores a 99%, o que indica uma capacidade impressionante de fazer previsões precisas em geral, porém ainda mantendo uma pequena diferença nas métricas de Recall, que medem a capacidade de identificar positivos corretamente, nesse cenário, o SVC apresentou um Recall de 99%, enquanto o LRL teve um Recall de 98,8% o que sugere que o SVC é mais eficaz em identificar pacientes com diabetes, o que é fundamental em contextos clínicos para evitar falsos negativos. Essa diferenciação entre Recall e Precision destaca a importância de considerar a aplicação específica ao escolher um modelo, pois diferentes contextos podem exigir ênfases distintas em precisão e capacidade de identificar positivos.

Cada algoritmo tem suas vantagens e desvantagens, e a escolha depende dos objetivos específicos do problema. O SVC é adequado quando a ênfase está na identificação precisa de casos positivos, enquanto o KNN pode ser preferível quando se busca um equilíbrio entre positivos e negativos. Já o MLP procura equalizar a precisão e capacidade de identificar positivos, o que aliado com a sua natureza de se adaptar às nuances dos dados,no torna uma opção versátil em muitos cenários A decisão entre esses algoritmos deve ser baseada nas necessidades e requisitos da aplicação clínica ou de pesquisa em questão. Portanto, é aconselhável considerar essas métricas em conjunto com outros fatores, como eficiência computacional, interpretabilidade do modelo e os trade-offs entre Recall, Precision e acurácia geral, ao selecionar o algoritmo mais

adequado para a tarefa do problema clínico de previsão de diabetes.

Conclusões

Este estudo abordou a previsão da diabetes por meio da análise de dados médicos e demográficos com a aplicação de várias técnicas de aprendizado de máquina. Inicialmente, realizamos uma exploração detalhada da base de dados, identificando desequilíbrios nas classes e a necessidade de transformações nos dados. Durante o treinamento dos modelos, avaliamos o desempenho da regressão logística como ponto de partida. No entanto, ficou claro que o desbalanceamento de classes impactava negativamente a capacidade de prever a classe minoritária. Para abordar esse desafio, implementamos uma técnica de reamostragem com o IHT, resultando em uma melhoria significativa na acurácia global e na capacidade de prever ambas as classes de forma equilibrada. A partir dessa abordagem, alcançamos uma acurácia geral de 100% para os algoritmos Random Forest e Decision Tree, embora esses resultados impressionantes tenham sido acompanhados de um comprometimento na especificidade. No entanto, o Support Vector Classifier (SVC) demonstrou uma acurácia geral notável, mantendo uma alta precisão, o que o torna uma escolha sólida quando a prevenção de falsos positivos é crucial.

Além disso, a Logistic Regression apresentou resultados similares, e o MLP Neural Network se destacou com um desempenho excepcional, equilibrando efetivamente Recall e Precision. Esses resultados enfatizam a importância da escolha cuidadosa das técnicas de pré-processamento de dados e da atenção ao desbalanceamento de classes ao criar modelos preditivos para problemas de saúde, contribuindo para avançar a previsão da diabetes e fornecendo uma sólida base para futuras pesquisas na área de saúde e aprendizado de máquina. Como trabalhos futuros, recomendamos explorar abordagens para melhorar a especificidade dos modelos mantendo altas taxas de Recall e Precision, bem como investigar a interpretabilidade dos modelos em um contexto clínico.

Referências

ABREU, M. N. S.; SIQUEIRA, A. L.; CAIAFFA, W. T. Regressao logistica ordinal em estudos epidemiologicos. **Revista de Saude Publica**, v. 43, n. 1, p. 183–194, 2009.

ALVES, F. B.; LIMA, D. A. Uso de la clasificación para el análisis y la minería de datos en la herramienta de enseñanza-aprendizaje Google Classroom. **Nuevas Ideas en Informática Educativa**, v. 4, p. 589–594, 2018.

BASSO, N. A. d. S. et al. Insulinoterapia, controle glicêmico materno e prognóstico perinatal: diferença entre o diabetes gestacional e o clínico. **Revista Brasileira de Ginecologia e Obstetrícia**, SciELO Brasil, v. 29, p. 253–259, 2007.

CONTINI, E.; CORRAR, L. J.; FILHO, E. O. Técnicas de amostragem e análise de dados para a avaliação da qualidade de serviços bancários. **Revista Contabilidade &; Finanças**, SciELO Brasil, v. 13, n. 29, p. 7–18, 2002.

COSTA-FILHO, S. V. S. da et al. Configuração de algoritmos de aprendizado de máquina na modelagem florestal: um estudo de caso na modelagem da relação hipsométrica. **Ciência Florestal**, v. 29, n. 4, 2019.

DANTAS, D.; DONADIA, E. A. Comparação entre as técnicas de regressão logística, árvore de decisão, bagging e random forest aplicadas a um estudo de concessão de crédito. **Revista Ciências Exatas e Naturais**, v. 12, n. 2, p. 269–293, 2013.

DORNELAS, R. S.; LIMA, D. A. Correlation Filters in Machine Learning Algorithms to Select Demographic and Individual Features for Autism Spectrum Disorder Diagnosis. **Journal of Data Science and Intelligent Systems**, 2023.

FACELI, K. et al. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina**. Rio de Janeiro: LTC, 2011. P. 378. ISBN ISBN do livro, se disponível.

FONTANELLA, B. J. B. et al. Amostragem em pesquisas qualitativas: proposta de procedimentos para constatar saturação teórica. **Cadernos de saude publica**, SciELO Brasil, v. 27, n. 2, p. 388–394, 2011.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, p. 1189–1232, 2001.

GARCÍA, C. G. Diabetes mellitus gestacional. **Medicina interna de México**, v. 24, n. 2, p. 148–156, 2008.

HAYKIN, S. Artificial Neural Networks: A Review of Applications. **Proceedings of the IEEE**, v. 82, n. 6, p. 956–985, 1994.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, American Association for the Advancement of Science, v. 349, n. 6245, p. 255–260, 2015.

LIMA, D. A.; FERREIRA, M. E. A.; SILVA, A. F. F. Machine learning and data visualization to evaluate a robotics and programming project targeted for women. **Journal of Intelligent & Robotic Systems**, Springer, v. 103, n. 1, p. 4, 2021.

LYRA, R. et al. Prevention of Type 2 Diabetes Mellitus. **Arquivos Brasileiros de Endocrinologia & Metabologia**, Editora da Sociedade Brasileira de Endocrinologia e Metabologia, São Paulo, Brazil, v. 50, p. 239–249, 2023.

MOREIRA, J. et al. Modelos de Aprendizado de Máquina na Predição de Diabetes Tipo 1 na Gestação usando Dados do Sistema Único de Saúde. In: ANAIS do XXI Simpósio Brasileiro de Computação Aplicada à Saúde. Evento Online: SBC, 2021. P. 392–403.

SCHEFFEL, R. S. et al. Prevalência de complicações micro e macrovasculares e de seus fatores de risco em pacientes com diabetes melito do tipo 2 em atendimento ambulatorial. **Revista da Associação Médica Brasileira**, SciELO Brasil, v. 50, p. 263–267, 2004.

SILVA, F. J. d.; SILVA FILHO, J. G. d. Aplicações da árvore de decisão na análise e pós-análise do desempenho acadêmico dos alunos do curso técnico em informática do IFPB-campus Cajazeiras. **Revista Eletrônica do Mestrado em Educação Ambiental**, v. 23, 2010.

SOUZA, V. S.; LIMA, D. A. Identifying Risk Factors for Heart Failure: A Case Study Employing Data Mining Algorithms. **Journal of Data Science and Intelligent Systems**, 2023.

VARGAS, J. M. L. et al. Efecto de terapias farmacológicas para el control glicémico en pacientes con diabetes mellitus tipo 2 en los desenlaces vasculares. **Revista Colombiana de Nefrología**, v. 7, n. 1, p. 44–59, 2020.