Predição de Diabetes Utilizando Modelos de Machine Learning

Murilo Batista da Silva
Pontifícia Universidade Católica de Campinas
Campinas, São Paulo, Brasil
murilo.bs3@puccampinas.edu.br

Lucas Suzano Guerino
Pontifícia Universidade Católica de Campinas
Campinas, São Paulo, Brasil
lucas.sg2@puccampinas.edu.br

Rodrigo de Faria Perico
Pontifícia Universidade Católica de Campinas
Campinas, São Paulo, Brasil
rodrigo.fp6@puccampinas.edu.br

Pedro Augusto Vermude de Carvalho
Pontifícia Universidade Católica de Campinas
Campinas, São Paulo, Brasil
pedro.avc@puccampinas.edu.br

Vinícius de Castro Vicente Dourado
Pontifícia Universidade Católica de Campinas
Campinas, São Paulo, Brasil
vinicius.cvd@puccampinas.edu.br

Saullo Haniell Galvão de Oliveira
Pontifícia Universidade Católica de Campinas
Campinas, São Paulo, Brasil
saullo.haniell@puc-campinas.edu.br

Abstract

This study investigates the application of supervised learning algorithms to predict the incidence of diabetes, leveraging easily accessible data such as demographic information and health indicators to identify the most effective approaches for early disease detection. The primary goal is to compare the performance of different classification models in this task.

The experimental results considered three classification methods, namely: K-Nearest Neighbors (KNN), Logistic Regression, and Support Vector Machine (SVM). Due to the imbalanced distribution of the classes (presence or absence of diabetes), besides obtaining decent accuracy values, the recall of all methods was highly impacted. The continuation of this work will include: i) adding more classification methods to the experiment, such as neural networks and ensemble-based methods; ii) compare the obtained results with the literature; and iii) consider the impact of data pre-processing steps to mitigate the class imbalance.

Keywords

Machine Learning, Diabetes, Classificação.

1 Introdução

Diabetes é uma doença crônica que atinge milhões de pessoas ao redor do mundo, caracterizada pela dificuldade do organismo em regular a glicose no sangue. Seus riscos, quando não tratados, podem gerar: i) doenças cardiovasculares; ii) danos ao sistema nervoso; e até iii) insuficiência renal [1, 2]. Com um diagnóstico precoce, é possível realizar mudanças no estilo de vida para retardar o aparecimento ou as complicações da diabetes [3].

Métodos de aprendizado de máquina (do inglês, *Machine Learning* - ML) têm demonstrado grande potencial de aplicações no campo da saúde, podendo servir de auxílio ao diagnóstico médico ou a outros processos de tomada de decisão. Diversos modelos de ML apresentam um desempenho eficaz e alta precisão para prever doenças, incluindo a diabetes [4].

Neste contexto, este trabalho explora o uso de algoritmos de aprendizado supervisionado para prever a incidência de diabetes. O objetivo desse trabalho é comparar o desempenho de diferentes modelos de classificação na predição de diabetes, utilizando dados de fácil coleta, como dados demográficos e indicadores de saúde.

2 Solução Proposta

A solução proposta nesse trabalho envolve o treinamento de modelos de classificação utilizando a base de dados "*Diabetes Health Indicators*" [5], bem como a análise dos resultados obtidos. O processo é detalhado nas subseções a seguir.

2.1 Base de Dados e Pré-Processamento

O "*Diabetes Health Indicators*" é um conjunto de dados de 253.680 respostas de pesquisa para o CDC (Centros de Controle e Prevenção de Doenças dos EUA) [5], realizada por telefone anualmente pelo CDC. A base de dados contém 21 atributos obtidos tanto por meio de perguntas feitas diretamente aos participantes, quanto por meio de variáveis calculadas com base em suas respostas individuais. Contém 35.346 amostras de pessoas diabéticas e 218.334 não diabéticas, apresentando grande desbalanceamento entre as classes.

A Tab. 1 exibe todos os atributos da base de dados, enquanto a Fig. 1 apresenta a correlação de Pearson entre todos os pares de variáveis. Nota-se que nenhum dos pares apresenta uma forte correlação.

A divisão nos conjuntos de treino, validação e teste, foi estratificada para considerar o desbalanceamento das classes, sem *undersampling* ou *oversampling*, para evitar ruídos e perda de informação. As variáveis numéricas foram normalizadas para que fiquem dispostas na mesma escala, enquanto as variáveis categóricas foram codificadas utilizando o método *one-hot encoding*. Além disso, foi realizada a redução de dimensionalidade por meio da aplicação da técnica *Principal Component Analysis* (PCA)[6]. Todas as etapas de pré-processamento exigem o cálculo de parâmetros específicos. Nesse caso, foram utilizados os conjuntos de treino e validação para a configuração das etapas de pré-processamento, para em seguida avaliar o desempenho dos modelos no conjunto de teste.

2.2 Métodos de Classificação

Os métodos de classificação selecionados para o estudo de desempenho foram: K-Nearest Neighbors (KNN), *Logistic Regression* (Regressão Logística) e Support Vector Machine (SVM). A Regressão Logística foi escolhida por suas propriedades interpretativas. Já o KNN e o SVM pela não-linearidade dos parâmetros.

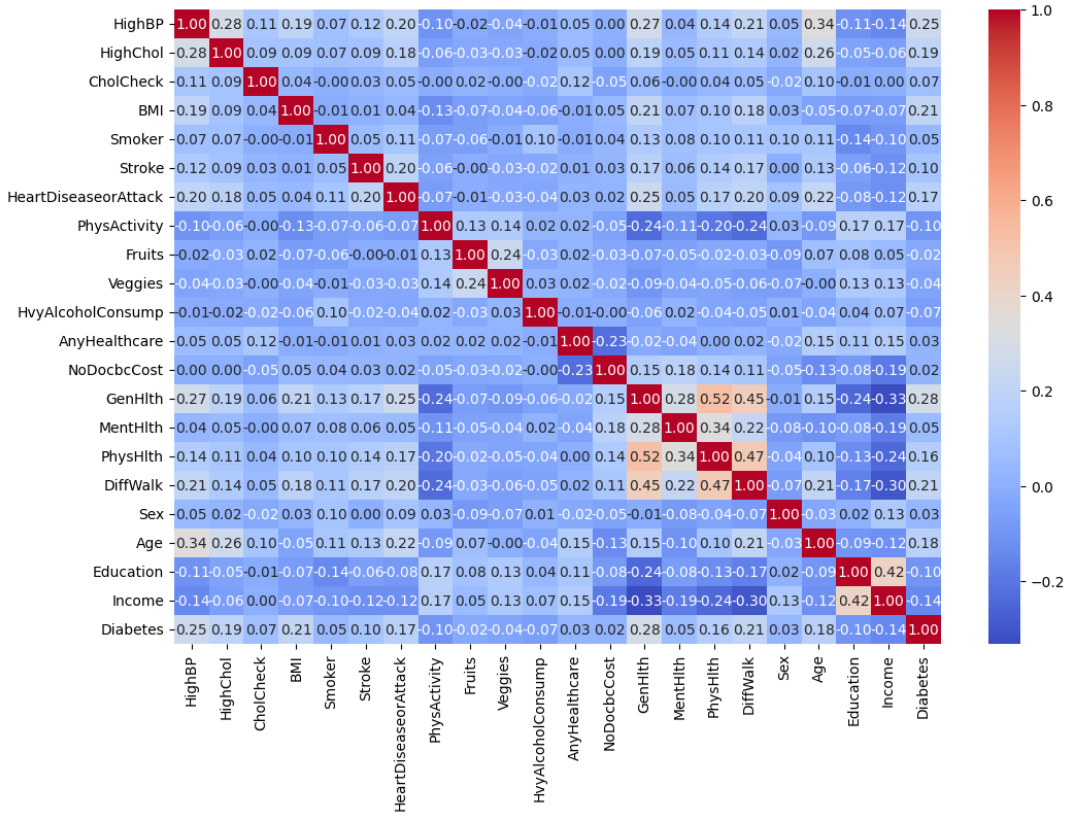

Figura 1: Atributos do conjunto de dados e suas respectivas correlações dispostas em um mapa de calor.

Tabela 1: Atributos da base de dados

Atributo	Papel	Tipo
ID	ID	Inteiro
Diabetes_binary	Classe	Binário
HighBP	*Feature*	Binário
HighChol	*Feature*	Binário
CholCheck	*Feature*	Binário
BMI	*Feature*	Inteiro
Smoker	*Feature*	Binário
Stroke	*Feature*	Binário
HeartDiseaseorAttack	*Feature*	Binário
PhysActivity	*Feature*	Binário
Fruits	*Feature*	Binário
Veggies	*Feature*	Binário
HvyAlcoholConsump	*Feature*	Binário
AnyHealthcare	*Feature*	Binário
NoDocbcCost	*Feature*	Binário
GenHlth	*Feature*	Inteiro
MentHlth	*Feature*	Inteiro
PhysHlth	*Feature*	Inteiro
DiffWalk	*Feature*	Binário
Sex	*Feature*	Binário
Age	*Feature*	Inteiro
Education	*Feature*	Inteiro
Income	*Feature*	Inteiro

2.3 Otimização de Hiperparâmetros

A melhor configuração dos hiperparâmetros dos modelos foi determinada por meio de uma busca bayesiana utilizando o *Tree-structured Parzen Estimator* (TPE) [7], com o auxílio da biblioteca Optuna [8]. Devido ao desbalanceamento de classes na base de dados, a métrica escolhida para avaliar as combinações sugeridas foi o *F1-Score*.

Para cada modelo, foram testadas as seguintes configurações de hiperparâmetros:

- `KNN`: número de vizinhos (2 a 25), peso das amostras (uniforme e baseado em distância) e distância utilizada (euclideana, manhattan e minkowski).
- `Logistic Regression`: regularização (l2 e l1), e C (0,01 a 10).
- `SVM`: C (0,001 a 1.000), kernel (linear, RBF (*radial basis function*), polinomial e sigmoide) e gamma (0.0001 a 10).

Configuramos a busca para utilizar 100 execuções para cada modelo, retornando a melhor configuração de hiperparâmetros encontrada.

3 Resultados Obtidos

Após a seleção de hiperparâmetros, o treinamento e cálculo de métricas de cada modelo foi executado 30 vezes. Em cada execução:

Tabela 2: Métricas gerais para treino e teste

	Treino				Teste			
Modelo	**Acurácia**	**Precisão**	**Recall**	**F1**	**Acurácia**	**Precisão**	**Recall**	**F1**
SVM	$0,78 \pm 2,22^{-16}$	$0,27 \pm 5,55^{-17}$	$0,26 \pm 5,55^{-17}$	$0,26 \pm 5,55^{-17}$	$0,78 \pm 2,22^{-16}$	$0,28 \pm 5,55^{-17}$	$0,26 \pm 5,55^{-17}$	$\mathbf{0,27 \pm 5,55^{-17}}$
KNN	$\mathbf{0,99 \pm 1,11^{-16}}$	$\mathbf{1,00 \pm 0,00}$	$\mathbf{0,96 \pm 0,00}$	$\mathbf{0,98 \pm 5,55^{-17}}$	$0,77 \pm 1,11^{-16}$	$0,27 \pm 5,55^{-17}$	$\mathbf{0,27 \pm 0,00}$	$\mathbf{0,27 \pm 5,55^{-17}}$
Logistic Reg.	$0,85 \pm 2,22^{-16}$	$0,55 \pm 1,11^{-16}$	$0,14 \pm 2,77^{-17}$	$0,23 \pm 8,32^{-17}$	$\mathbf{0,85 \pm 2,22^{-16}}$	$\mathbf{0,57 \pm 1,11^{-16}}$	$0,14 \pm 2,77^{-17}$	$0,23 \pm 8,32^{-17}$

i) base de dados foi embaralhada; ii) os hiper-parâmetros dos modelos foram ajustados (utilizando a união dos conjuntos de treino e validação); e iii) as métricas calculadas nos conjuntos de: i) treino e validação (treino) e ii) teste. Consideramos as métricas F1-Score (Eq. 1), acurácia (Eq. 2), precisão (Eq. 3) e recall (Eq. 4) de cada modelo. Os resultados reportados na tabela 2 consideram a média e o desvio padrão das 30 execuções.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{1}$$

$$\text{Acurácia} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{2}$$

$$\text{Precisão} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{4}$$

Os resultados obtidos demonstram que os modelos SVM e KNN apresentaram um F1-Score de 27% no conjunto de teste, superando a Regressão Logística. Apresentam também um desvio padrão baixo, o que demonstra consistência nos resultados entre diferentes execuções. Note que todos os modelos apresentam uma acurácia razoável (com valores entre 0.77 e 0.85), mas um F1 considerado baixo. Enquanto a Regressão Logística tem seu F1 prejudicado principalmente pelo baixo *recall* (0.14), SVM e KNN também obtiveram baixos valores para a precisão. O KNN também apresentou um pouco de *overfitting*, obtendo desempenho excelente no conjunto de treino, e baixo no conjunto de teste. Tais resultados, contudo, são motivados pelo desbalanceamento das classes no conjunto de dados.

4 Considerações Finais

O objetivo do projeto foi analisar a aplicabilidade de algoritmos de classificação na predição da presença de diabetes, utilizando como entrada dados de fácil coleta, relacionados com histórico de saúde, hábitos de vida e características demográficas dos pacientes. No experimento realizado, avaliamos o desempenho de diferentes métodos de classificação nessa tarefa, utilizando uma base de dados disponibilizada pelo CDC, que contém informações de fácil coleta dos pacientes. Entre os modelos testados, o Support Vector Machine (SVM) destacou-se como a solução mais eficaz, apresentando o melhor desempenho nas métricas avaliadas. No entanto, todos os modelos apresentaram um resultado que precisa ser melhorado para que a predição de diabetes a partir de dados como esses seja confiável para a prática clínica.

A continuação deste trabalho irá: i) incluir mais métodos de classificação, como redes neurais e métodos baseados em *ensemble*; ii) comparar os resultados obtidos com a literatura; e iii) analisar possíveis ganhos com a implementação de técnicas de balanceamento de dados.

Referências

[1] Muhammad Shariq Usman, Muhammad Shahzeb Khan, and Javed Butler. The interplay between diabetes, cardiovascular disease, and kidney disease. *ADA Clinical Compendia*, 2021(1):13–18, 2021. doi: 10.2337/db20211-13. URL https://doi.org/10.2337/db20211-13.

[2] Rayne Gomes Amorim, Glaucevane da Silva Guedes, Sandra Mary Lima Vasconcelos, and Juliana Célia de Farias Santos. Kidney disease in diabetes mellitus: Cross-linking between hyperglycemia, redox imbalance and inflammation. *Arquivos Brasileiros de Cardiologia*, 112(5):577–587, 2019. doi: 10.5935/abc.20190077.

[3] Jay S. Skyler, George L. Bakris, Ezio Bonifacio, Tamara Darsow, Robert H. Eckel, Leif Groop, Per-Henrik Groop, Yehuda Handelsman, Richard A. Insel, Chantal Mathieu, Allison T. McElvaine, Jerry P. Palmer, Alberto Pugliese, Desmond A. Schatz, Jay M. Sosenko, John P.H. Wilding, and Robert E. Ratner. Differentiation of diabetes by pathophysiology, natural history, and prognosis. *Diabetes*, 66(2): 241–255, 12 2016. ISSN 0012-1797. doi: 10.2337/db16-0806. URL https://doi.org/10.2337/db16-0806.

[4] Tadesse Melaku Abegaz, Muktar Beshir Ahmed, Fatimah Sherbeny, Vakaramoko Diaby, Hongmei Chi, and Askal Ayalew Ali. Application of machine learning algorithms to predict uncontrolled diabetes using the all of us research program data. *Healthcare*, 11(8), 2023. ISSN 2227-9032. doi: 10.3390/healthcare11081138.

[5] Centers for Disease Control and Prevention (CDC). Cdc diabetes health indicators dataset. https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators, n.d. Accessed: 2024-12-08.

[6] Jonathon Shlens. A tutorial on principal component analysis. *CoRR*, abs/1404.1100, 2014. URL https://arxiv.org/abs/1404.1100.

[7] Shuhei Watanabe. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance, 2023. URL https://arxiv.org/abs/2304.11127.

[8] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.