Análise comparativa de modelos de Machine Learning para predição de diabetes por meio dos modelos SVM e K-means

Ester Miriã de Sousa Freire[1], Paulo Anderson Gonçalves de Lima[1], Ângelo Carlos Avelino Feitosa[1], Lucas Fontineli Alves Sousa[1], Simone Azevedo Bandeira de Melo Aquino[2], Daniel Duarte Costa[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Maranhão – Campus Imperatriz (IFMA)

CEP 65.906-335 – Imperatriz – MA – Brasil

[2]Departamento de Ensino Superior e Tecnologia - Instituto Federal do Maranhão, IFMA

[2]UFMA: Centro de Ciências de Imperatriz - Universidade Federal do Maranhão.

{paulo.anderson, angelo.a, emiria, sousafontineli}@acad.ifma.edu.br,
simonebandeira@ifma.edu.br, daniel.dc@ufma.br

Abstract. *This paper compares machine learning techniques for diabetes prediction using the Pima Indians Diabetes Database. Four supervised models (logistic regression, MLP, SVM with and without PCA) and an unsupervised method (K-means) were evaluated. Performance was measured by accuracy, F1-score, sensitivity, specificity and AUC-ROC, and the impact of dimensionality reduction (PCA) was analyzed. SVM with PCA achieved the best discrimination (AUC = 0.84). K-means revealed clinically meaningful clusters without labels.*

Resumo. *Este trabalho compara técnicas de aprendizado de máquina para predição de diabetes utilizando o Pima Indians Diabetes Database. Foram avaliados quatro modelos supervisionados (regressão logística, MLP, SVM com e sem PCA) e uma abordagem não supervisionada (K-means). A avaliação envolve métricas de performance (acurácia, F1-score, sensibilidade, especificidade e AUC-ROC) e análise da robustez dos modelos frente à redução de dimensionalidade via PCA. Os resultados mostram que o SVM com PCA obteve a melhor discriminação (AUC = 0,84), enquanto o K-means ajudou a revelar padrões clínicos relevantes sem uso de rótulos.*

1. Introdução

O diabetes é uma doença crônica caracterizada por altos níveis de glicose no sangue, sendo o tipo 2 responsável por 90% dos casos globais e considerado um dos principais desafios de saúde pública. O diagnóstico precoce é fundamental para prevenir complicações graves [Antunes et al apud Medeiros et al, 2021]. Avanços em tecnologias computacionais, especialmente aprendizado de máquina, têm apoiado a detecção e o monitoramento da doença. Modelos supervisionados, como regressão logística, redes neurais e SVM, são

amplamente usados em classificações binárias, enquanto técnicas como PCA ajudam a reduzir ruídos e métodos não supervisionados, como K-means, revelam padrões ocultos [Ferreira, 2014; Fleck et al, 2016; Aithal, 2023].

Este artigo compara diferentes modelos supervisionados (regressão logística, redes neurais MLP e SVM com e sem PCA) e o K-means, aplicados ao Pima Indians Diabetes Database. O objetivo é avaliar métricas como acurácia, F1-score e AUC-ROC e explorar padrões ocultos, destacando a relevância da inteligência artificial na saúde pública [Witten et al, 2016]. O trabalho organiza-se em metodologia, resultados e considerações finais.

2. Metodologia

Para este estudo, utilizou-se o Pima Indians Diabetes Database, contendo 768 registros de mulheres Pima acima de 21 anos, com oito atributos clínicos: gestações, glicose, pressão arterial, espessura da pele, insulina, IMC, função hereditária e idade. A variável-alvo é binária, indicando presença ou ausência de diabetes [Aithal, 2023]. A preparação dos dados envolveu tratamento de valores ausentes, substituindo zeros pela mediana, análise estatística e visual com histogramas e boxplots, padronização via z-score e divisão estratificada em 70% treino e 30% teste, mantendo o equilíbrio entre classes [Ferreira, 2014; Aithal, 2023; Fleck et al, 2016; Pedregosa et al, 2011].

Foram aplicados quatro modelos supervisionados: regressão logística (baseline com regularização L2), rede neural MLP (duas camadas ocultas, ReLU, otimização Adam, 200 épocas com early stopping), SVM com kernel RBF e SVM com PCA, este último visando redução de ruído e maior velocidade de convergência [Miranda, 2023; Furtado, 2019; Rauber, 2005; Ferreira, 2014]. Como abordagem não supervisionada, aplicou-se o K-means com k = 2, com e sem PCA, sendo o agrupamento avaliado pelo Silhouette Score. Os modelos supervisionados foram avaliados por acurácia, F1-score, ROC e AUC, enquanto o K-means também considerou a distribuição de classes por cluster [Witten et al, 2016; Bishop, 2006].

3. Resultados

O desempenho dos modelos supervisionados apresentou diferenças sutis, mas suficientes para avaliar sua qualidade. O SVM com PCA obteve os melhores resultados, com acurácia de 0,78 e AUC-ROC de 0,84, enquanto sem PCA apresentou 0,75 e 0,81. A rede neural registrou 0,777 e 0,83, e a regressão logística 0,76 e 0,82. Nas análises com K-means, conforme a figura 1, o Silhouette Score foi 0,42 sem PCA e 0,48 com PCA. Observou-se sobreposição entre clusters e classes reais: o cluster 0 concentrou 72% de não diabéticos e 28% de diabéticos, enquanto o cluster 1 reuniu 58% de diabéticos e 42% de não diabéticos.

Quanto à importância das features, a rede neural destacou-se, com glucose em 0,32, BMI em 0,25 e idade em 0,18. Na aplicação do K-means, os clusters com PCA usaram os principais componentes 1 e 2, enquanto os sem PCA utilizaram glicose e pressão arterial. Conforme as figuras 1 e 2, o gráfico com PCA apresentou melhor separação dos clusters, condensando informações relevantes em poucas dimensões, enquanto a clusterização baseada apenas em glicose e pressão arterial foi menos eficiente.

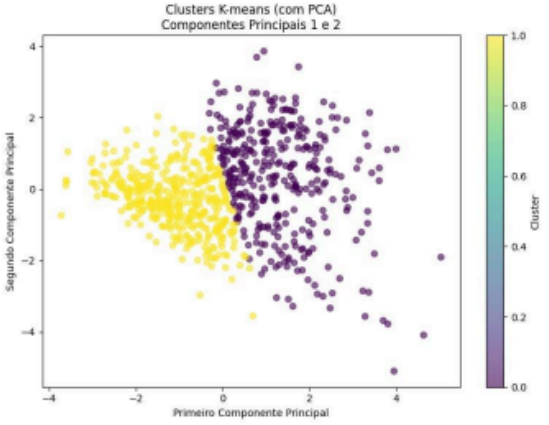

Figura 1. Clusterização com PCA.

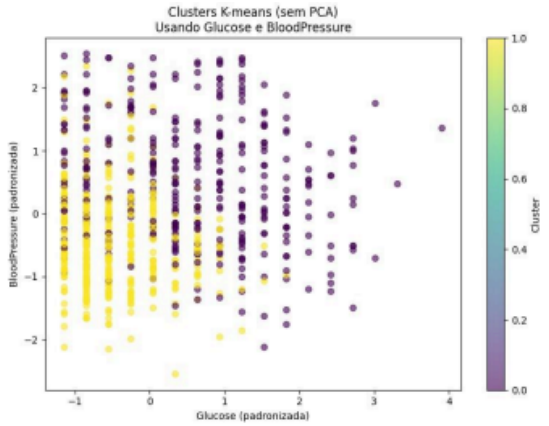

Figura 2. Clusterização sem PCA.

As matrizes de confusão e curvas ROC do SVM com e sem PCA mostraram pequenas diferenças. Na figura, a variação nas predições foi mínima, com apenas uma diferença entre labels 0 e 1, indicando eficiência similar, ligeiramente superior com PCA.

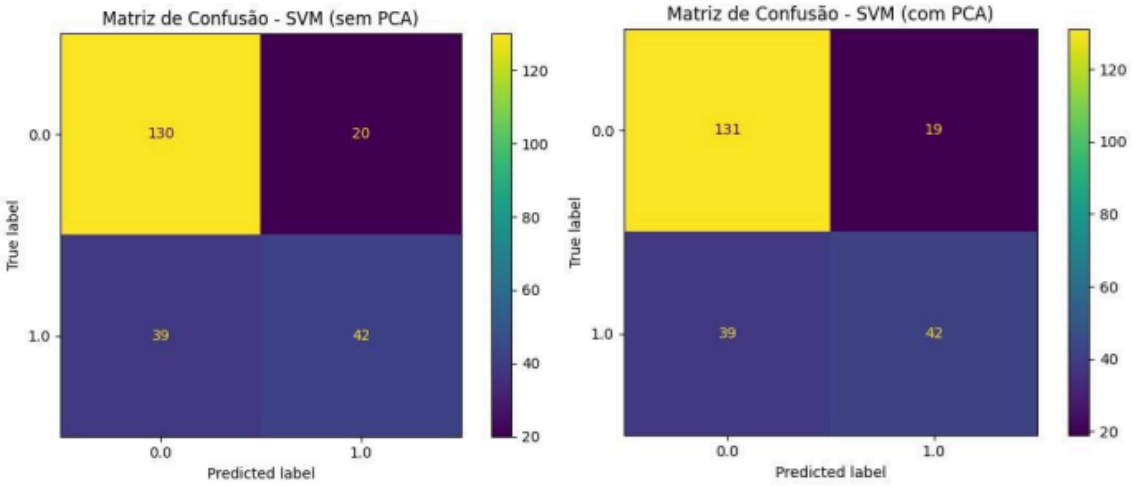

Figura 3. Matriz de confusão sem PCA e com PCA, respectivamente.

A curva AUC-ROC também apresentou resultado semelhante, com apenas 0.01 de diferença na AUC, na qual a curva com PCA obteve 0.83 de AUC e a sem PCA, 0.82.

4. Considerações finais

Este trabalho realizou uma análise comparativa entre diferentes modelos de aprendizado de máquina aplicados à predição de diabetes, utilizando o Pima Indians Diabetes Database. Entre os modelos testados, o SVM com PCA apresentou o melhor desempenho, beneficiado pela redução da dimensionalidade e menor risco de sobreajuste. As técnicas não supervisionadas, como o K-means, também demonstraram utilidade ao identificar padrões clínicos relevantes sem a necessidade de rótulos.

Comparando com estudos recentes que empregaram o mesmo conjunto de dados, os resultados obtidos neste trabalho mostraram-se consistentes com o estado da arte. Kaur e Kumari (2022) reportaram uma AUC de 0,85 para o modelo SVM com otimização de parâmetros, enquanto Rahman et al. (2023) alcançaram 0,82 utilizando Random Forest. O desempenho de 0,84 obtido pelo SVM com PCA neste estudo confirma que a aplicação de técnicas de redução de dimensionalidade contribui para resultados competitivos, especialmente quando comparada a abordagens sem pré-processamento dimensional.

Referências:

Antunes, Y. R., de Oliveira, E. M., Pereira, L. A., & Picanço, M. F. P. (2021). *Diabetes Mellitus Tipo 2: A importância do diagnóstico precoce*. Brazilian Journal of Development, 7(12), 116526-116551.

Aithal, P. S. (2023). *Visualization and Explorative Data Analysis*. IJERSTE, 12(3), 11-21.
Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Ferreira, A. A. (2014). *Fundamentos de Aprendizado de Máquina*. LTC.

Fleck, L. et al. (2016). *Redes neurais artificiais: Princípios básicos*. Revista Eletrônica Científica Inovação e Tecnologia, 1(13), 47–57.

Furtado, M. I. V. (2019). *Redes Neurais Artificiais: uma abordagem para sala de aula*. Atena Editora.

Miranda, C. E. B. (2023). *Aplicação da regressão logística binária para manutenção preditiva em máquinas de ressonância magnética*. UTFPR.

Pedregosa, F. et al. (2011). *Scikit-learn: Machine Learning in Python*. JMLR, 12, 2825–2830.

Rauber, T. W. (2005). *Redes Neurais Artificiais*. Universidade Federal do Espírito Santo.

Witten, I. H., Frank, E., & Hall, M. A. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Tiago, C. M., & Leitão, V. M. (2002). Utilização de funções de base radial em problemas unidimensionais de análise estrutural. *Métodos Numéricos en Ingeniería V, CD. SEMNI*.

Witten, I. H., Frank, E., & Hall, M. A. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.