# Título em Português

Avaliação de Desempenho do Classificador de Máquina de Vetores

de Suporte com Diferentes Estratégias de Validação Cruzada.

# Title in English

Performance Evaluation of the Support Vector Machine Classifier with

Different Cross-Validation Strategies.

# Título en Español

Evaluación de Desempeño del Clasificador de Máquinas de Vectores

de Soporte con Diferentes Estrategias de Validación Cruzada..

## Introdução

O Diabetes Mellitus representa um dos maiores desafios de saúde pública do século XXI, caracterizando-se como uma doença metabólica crônica de prevalência crescente em escala global. Estimativas da Organização Mundial da Saúde e da Federação Internacional de Diabetes indicam que mais de 530 milhões de adultos vivem com a doença atualmente, com projeções que apontam para um aumento significativo nas próximas décadas¹. No Brasil, o cenário é igualmente preocupante, com mais de 16 milhões de casos diagnosticados, posicionando o país entre os mais afetados mundialmente². As complicações associadas ao diabetes, como doenças cardiovasculares, insuficiência renal e neuropatias, geram um impacto substancial na qualidade de vida dos indivíduos e nos sistemas de saúde, tornando o diagnóstico precoce e preciso uma prioridade fundamental para a prevenção e manejo eficaz da doença.

Nesse contexto, o avanço da Inteligência Artificial tem aberto novas fronteiras para o apoio à decisão clínica, e o Aprendizado de Máquina (*Machine Learning*), em particular, surge como uma ferramenta promissora para a criação de modelos preditivos a partir de dados clínicos³. Diversos algoritmos de classificação supervisionada têm sido aplicados para a predição de diabetes, contudo, existe uma lacuna no conhecimento quanto à comparação sistemática e otimização desses modelos para conjuntos de dados específicos, visando maximizar a acurácia e identificar os fatores de risco mais determinantes. A avaliação de diferentes abordagens é, portanto, essencial para desenvolver ferramentas que possam efetivamente auxiliar os profissionais de saúde na identificação de indivíduos em risco.

O presente estudo tem como objetivo comparar o desempenho do classificador SVM utilizando dois esquemas de validação cruzada: *KFold-10* e *ShuffleSplit* (30%, N = 10), analisando as métricas de acurácia e matriz de confusão.

## Material e Métodos

Trata-se de um estudo quantitativo, com delineamento preditivo, que utilizou dados secundários de domínio público. A fonte de dados foi o conjunto *Pima Indians Diabetes Database*, obtido do repositório da *University of California, Irvine* (UCI). O período de realização do estudo computacional foi setembro de 2025. A amostra foi composta por 768 registros, contendo oito variáveis preditoras (número de gestações, concentração de glicose, pressão arterial, espessura da dobra cutânea do tríceps, insulina sérica, índice de massa corporal, função de pedigree de diabetes e idade) e uma variável alvo binária (diagnóstico de diabetes). O tratamento dos dados, realizado em ambiente *Google Colaboratory* com a biblioteca *scikit-learn*, envolveu a substituição de valores fisiologicamente implausíveis (zero) por valores ausentes, que foram subsequentemente tratados com a técnica de imputação pela média (*SimpleImputer*). As variáveis foram padronizadas com *StandardScaler*. A análise dos dados consistiu na implementação de quatro modelos de classificação (Pipeline): SVM, Regressão Logística, KNN e *Gaussian Naive Bayes*. A otimização de hiperparâmetros foi realizada via *GridSearchCV* com validação cruzada *K-Fold* (k=10). A importância dos atributos do melhor modelo foi avaliada pela técnica de Importância por Permutação. Por se tratar de um banco de dados público e anonimizado, a pesquisa dispensa aprovação de Comitê de Ética em Pesquisa..

## Resultados e Discussão

Ambos os métodos apresentaram desempenho razoável, com acurácia média próxima de 77%. O *KFold*-10 obteve acurácia média de 76,56%, enquanto o *ShuffleSplit* 30% / 10 *splits* alcançou 77,40% (±2,6). Apesar da similaridade nos valores globais, a matriz de confusão revelou diferenças interessantes. No *KFold*-10, observou-se uma leve tendência a mais falsos negativos (pacientes com diabetes classificados como não-diabetes), enquanto no *ShuffleSplit* a distribuição de falsos positivos e falsos negativos foi mais variável entre os splits.

Isso sugere que, embora ambos os métodos forneçam estimativas próximas da acurácia, a escolha do esquema de validação pode influenciar a consistência da classificação de casos minoritários.

## Conclusão

## Os resultados demonstram que, embora as métricas globais de acurácia sejam semelhantes, a escolha do tipo de validação cruzada pode impactar a distribuição dos erros do modelo, evidenciada pela matriz de confusão.

## No caso do *ShuffleSplit*, a aleatoriedade na seleção dos conjuntos de teste gerou variação nos falsos positivos e falsos negativos entre os *splits*, indicando que este método pode produzir estimativas ligeiramente menos consistentes para classes minoritárias em *datasets* desbalanceados. Esse achado reforça a importância de explicitar e justificar a escolha do método de validação em estudos que aplicam aprendizado de máquina em contextos biomédicos, especialmente quando se deseja avaliar de forma confiável o desempenho do classificador SVM.

## Conflito de Interesses

Os autores declaram não haver conflito de interesses.

## Agradecimentos e Financiamentos

A presente pesquisa não recebeu qualquer tipo de financiamento para sua realização.

## Disponibilidade dos Dados

O conjunto de dados utilizado neste estudo é de domínio público e pode ser acessado através do repositório de Aprendizado de Máquina da UCI (*Machine Learning Repository*) ou pelo link fornecido no código-fonte do projeto, disponível em https://docs.github.com/pt/get-started/exploring-projects-on-github/contributing-to-a-project.

## Referências

*International Diabetes Federation. IDF Diabetes Atlas, 11th edn. Brussels, Belgium: International Diabetes Federation; 2024.* Disponível em: https://www.diabetesatlas.org

*World Health Organization. Diabetes. [Internet]. Geneva: WHO; 2024 [*citado em 26 de setembro de 2025*].* Disponível em: https://www.who.int/news-room/fact-sheets/detail/diabetes

*Deo RC. Machine Learning in Medicine. Circulation. 2015;132(20):1920-30.* Disponível em: https://www.ahajournals.org/doi/10.1161/circulationaha.115.001593

## Figuras e Gráficos

## Tabelas e Quadros