Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, 2025.

Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Diabetes in Women: Application of Support Vector Machine for Modeling and Prediction

Natália França dos Reis[1], João Batista Florindo[2], Laércio Luís Vendite[3]
UNICAMP, Campinas, SP

Gestational Diabetes Mellitus is a common complication in pregnancy and has a worldwide average incidence of 16.2% among pregnant women [1]. It arises from metabolic changes and aids in the transportation of glucose from the mother to the fetus through the placenta. Modern lifestyle factors such as high-carb diets and obesity can lead to gestational diabetes, marked by high blood glucose levels due to insulin problems [3]. Recognizing its specificities is vital to identify risks, take preventive measures and maintain quality of life. Furthermore, integration of machine learning in medicine allows analyzing complex medical data, leading to appropriate treatments and more personalized results with reduced healthcare costs.

In this study, we utilized a dataset from *National Institute of Diabetes and Digestive and Kidney Diseases* [2], containing 768 instances, each representing a female patient. The data encompasses 9 attributes describing the development of diabetes. The response variable indicates the presence or absence of the disease. We applied the K-Nearest Neighbors (K-NN) to deal with missing data. The StandardScaler was used to normalize the data and we used Synthetic Minority Oversampling Technique (SMOTE) to balance the classes.

The SVM (Support Vector Machine) method [3] was used to classify data due to its efficiency in high-dimensional spaces. The choice of this algorithm was motivated and its easy interpretability, which facilitates a clear analysis of the results in identifying patterns and understanding the distribution of classes. Additionally, we applied PCA (Principal Component Analysis) to reduce dimensionality, optimizing SVM effectiveness and simplifying data interpretation.

We applied the SVM algorithm to a dataset divided into training (80%) and testing (20%). Below are the results, where 0 indicates absence and 1 presence of diabetes (Table 1). In Table 2, the Confusion Matrix is presented.

Table 1: SVM model metrics.

	Precision	*Recall*	*F1-score*	*Support*
0	0.83	0.73	0.78	79
1	0.77	0.85	0.81	82

Table 2: Confusion Matrix

TN	FP
58	21
FN	TP
12	70

Analysis of the metrics indicates promising results: the model achieves an accuracy of approximately 80%. Precision and recall present high values, demonstrating the model's robustness in accurately identifying positive and negative cases. This alignment is reflected in the F1 score, which demonstrates the balance of the model in detecting diabetes. The confusion matrix shows

[1] n262894@dac.unicamp.br
[2] florindo@unicamp.br
[3] lvendite@unicamp.br

high values on the main diagonal, indicating the model's strong ability to correctly classify both positive and negative cases. The Figure 1 shows the distribution of data on the axes of the two principal components.

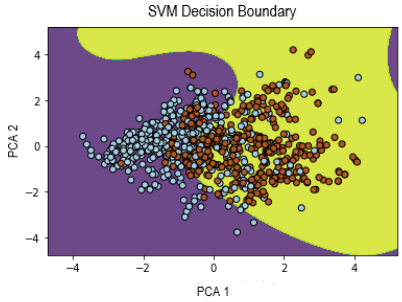

Figure 1: Each dot on the graph represents a patient, whose blue dots represent patients without diabetes and the red dots, patients with the disease. The hyperplane was generated to classify the points into the purple (no diabetes) and yellow (diabetes) regions. Source: Authors.

So far, we have noticed that there are areas in the plan where the data remains mixed, with an unclear decision boundary. In a medical context, where diagnostic accuracy is crucial, attention turns to cases diagnosed as negative and those that are false negatives and false positives. The analysis of these specific cases is crucial to improve the sensitivity and specificity of the SVM model, thereby improving the reliability of decision-making in diabetes detection.

Analysis of the confusion matrix reveals that 21 out of 79 cases are classified as False Negatives (27.6% of the total). These 21 cases represent situations in which the model was unable to correctly identify the presence of the condition. This error could result in patients who require treatment are not correctly identified by the algorithm. Therefore, a detailed analysis of these cases is essential to identify any patterns that may improve the model's accuracy.

In subsequent analyses, the distribution of points classified as negative reveals a distinct separation between the True Negative and False Negative classes. Accordingly, the SVM will be employed in the next phases of this research with the aim of enhancing the model's ability to accurately identify the absence of the disease, thereby optimizing its clinical utility.

References

[1] M. Hod, A. Kapur, D. A. Sacks, E. Hadar, M. Agarwal, G. C. Di Renzo, L. C. Roura, H. D. McIntyre, J. L. Morris, and H. Divakar. "The International Federation of Gynecology and Obstetrics (FIGO) Initiative on gestational diabetes mellitus: A pragmatic guide for diagnosis, management, and care". In: **International Journal of Gynecology and Obstetrics** 131 (2015), pp. 173–211. DOI: 10.1016/S0020-7292(15)30033-3.

[2] NIH. **Diabetes Dataset**. Online. Acessado em 16 de julho de 2023, https://www.kaggle.com/datasets/mathchi/diabetes-data-set. 2021.

[3] P. Pujari. "Classification of Pima Indian diabetes dataset using support vector machine with polynomial kernel". In: **Deep Learning, Machine Learning and IoT in Biomedical and Health Informatics**. CRC Press, 2022, pp. 55–67. ISBN: 9780367548445.