International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Predictive Modeling for Diabetes Using Support Vector Machines: A Case Study of the Pima Indian Diabetes Dataset

Rishabh Patel[1], *Dwipraj Nath*[2], *Rahul Anjana*[3]

[1] CSE Galgotias University Uttar Pradesh, India
[2] Cse Galgotias University Uttar Pradesh, India
[3] Cse Galgotias University Uttar Pradesh, India

ABSTRACT—

Diabetes mellitus is a chronic metabolic disorder with a worldwide prevalence, and early detection is critical for effective management and treatment. Machine learning models have increasingly become valuable tools in medical diagnostics, offering predictive capabilities for disease onset. This paper presents a comprehensive study on diabetes prediction using a Support Vector Machine (SVM) with a linear kernel, trained on the Pima Indian Diabetes dataset. Data preprocessing techniques, such as standardization, were applied, and the model was evaluated using accuracy metrics. Visualization techniques such as scatter plots, heatmaps, and histograms were used to investigate feature relationships. Our model achieved an accuracy of 77.92%, indicating that SVMs can effectively assist in the prediction of diabetes. The study identifies glucose as a primary predictive factor and highlights the potential for improving model accuracy by addressing class imbalance and exploring non-linear kernels.

Introduction :

Diabetes is one of the most pervasive health conditions in the world, affecting an estimated 463 million people according to the International Diabetes Federation. It is a metabolic disorder characterized by elevated blood glucose levels, which, if left unmanaged, can lead to severe complications such as cardiovascular diseases, kidney failure, and neuropathy. Early diagnosis of diabetes plays a crucial role in mitigating these risks and enabling timely medical intervention.

In recent years, machine learning (ML) has shown great promise in diagnosing and predicting the onset of diseases. ML models can analyze large volumes of patient data, identifying patterns that are often invisible to the naked eye. In this paper, we investigate the application of a **Support Vector Machine (SVM)** for predicting diabetes based on the **Pima Indian Diabetes dataset**. Our aim is to develop a model that can provide reliable predictions, thus assisting healthcare professionals in identifying at-risk individuals.

Literature review :

Numerous machine learning algorithms have been utilized for the prediction of diabetes, including **Logistic Regression**, **Decision Trees**, **Random Forests**, **K-Nearest Neighbors (KNN)**, and **Neural Networks**. Each model has its own strengths and limitations when applied to medical datasets.

Logistic Regression has been widely adopted for its simplicity and ease of interpretation. However, it is limited in its ability to model complex, non-linear relationships between features. In contrast, **Decision Trees** and **Random Forests** have shown better performance in handling non- linear data, though they can be prone to overfitting if not properly regularized.

More recent studies have explored **Deep Learning** techniques, particularly **Convolutional Neural Networks (CNNs)** and **Recurrent Neural Networks (RNNs)**. While these models tend to offer higher accuracy in large datasets, they are computationally expensive and require significant amounts of labeled data. **Support Vector Machines (SVMs)**, originally introduced by Cortes and Vapnik in 1995, have shown great efficacy in small to medium-sized datasets, particularly in binary classification problems like medical diagnosis.

Research by Jothi and Rashmi (2020) demonstrated the superior performance of SVMs in medical diagnostics due to their robustness against high-dimensional spaces and outliers. However, few studies have examined the specific application of linear kernel SVMs to the Pima Indian Diabetes dataset, leaving a gap in understanding how well this algorithm performs on this particular medical dataset.

Research Gap :

While much research has focused on the use of ML algorithms for diabetes prediction, there remains a significant gap in understanding the performance of different SVM kernels, particularly in datasets with imbalanced class distributions. Additionally, many studies lack comprehensive feature analysis through visualizations, which can provide important insights into the relationships between features and disease outcomes. Furthermore, there is a lack of exploration into techniques that address class imbalance, such as **Synthetic Minority Oversampling Technique (SMOTE)** or **Adaptive Synthetic Sampling (ADASYN)**. This paper seeks to fill these gaps by:

1. Investigating the performance of a linear kernel SVM on the Pima Indian Diabetes dataset.
2. Utilizing extensive exploratory data analysis (EDA) to interpret feature relationships.
3. Addressing class imbalance and exploring future enhancements to improve model accuracy.

Methodology :

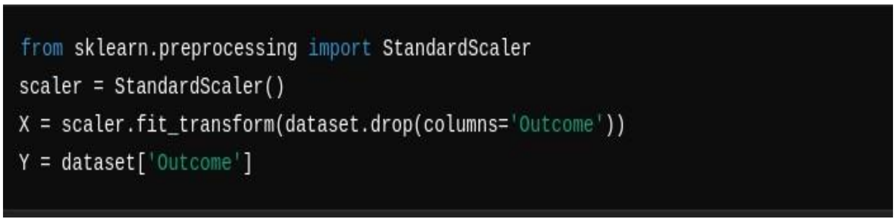
a) Dataset Description

- The **Pima Indian Diabetes dataset** is a widely used dataset for diabetes prediction. It contains data on 768 female patients of Pima Indian heritage, with 8 predictive features and a binary target variable (**Outcome**), where 1 denotes a positive diabetes diagnosis and 0 indicates non-diabetic. The features include:
  - **Pregnancies**: Number of times pregnant.
  - **Glucose**: Plasma glucose concentration after a 2-hour oral glucose tolerance test.
  - **BloodPressure**: Diastolic blood pressure (mm Hg).
  - **SkinThickness**: Triceps skinfold thickness (mm).
  - **Insulin**: 2-hour serum insulin (mu U/ml).
  - **BMI**: Body mass index (weight in kg/(height in m)^2).
  - **DiabetesPedigreeFunction**: A function that represents a score for diabetes susceptibility based on family history.
  - **Age**: Age of the patient.

b) Data Preprocessing

Data preprocessing is an essential step to ensure that the model performs well and is not biased by any irregularities in the dataset. The dataset does not contain missing values but suffers from class imbalance, with fewer diabetic cases than non-diabetic ones.

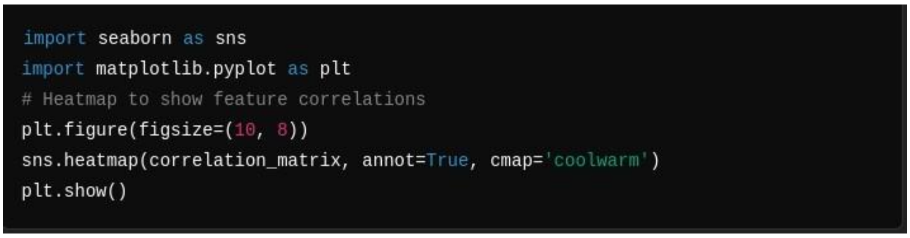
We standardized the dataset using **StandardScaler** from scikit-learn to ensure that each feature had a mean of 0 and a standard deviation of 1. This was necessary since SVMs are sensitive to feature scaling due to their reliance on calculating distances between data points.



```python
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X = scaler.fit_transform(dataset.drop(columns='Outcome'))
Y = dataset['Outcome']
```

c) Exploratory Data Analysis (EDA)

- Before model training, we conducted an extensive EDA to understand the relationships between various features and the target outcome. We utilized several visualization techniques to accomplish this:
  - **Scatter Plot**: Showed the relationship between glucose levels and pregnancies in diabetic and non- diabetic individuals.
  - **Heatmap**: Illustrated the correlation between features, with glucose showing the strongest correlation with the outcome.
  - **Histograms**: Provided insight into the distribution of glucose levels across the two outcome classes.



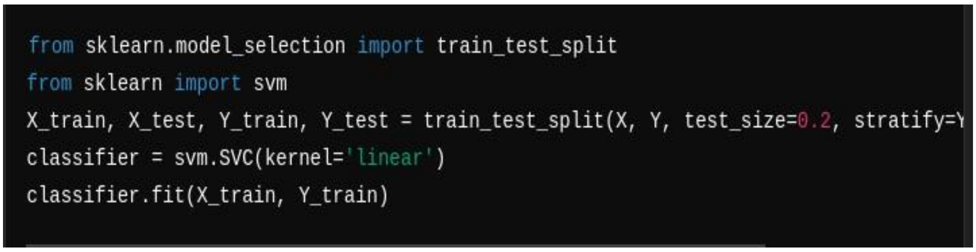
```python
import seaborn as sns
import matplotlib.pyplot as plt
# Heatmap to show feature correlations
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.show()
```

Model Training and Implementation :

a) Support Vector Machine (SVM) Model

The model chosen for this study was an SVM with a **linear kernel**. SVMs work by finding the hyperplane that best separates the two classes (diabetic vs. non-diabetic) by maximizing the margin between the classes. The choice of a linear kernel was motivated by its effectiveness in high- dimensional spaces and its relative simplicity.

The dataset was split into an 80-20 training-test ratio, with stratification to preserve the class distribution across both sets. The model was then trained using the **SVC** function from the scikit-learn library.



```
from sklearn.model_selection import train_test_split
from sklearn import svm
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, stratify=Y
classifier = svm.SVC(kernel='linear')
classifier.fit(X_train, Y_train)
```

b) **Cross-Validation**

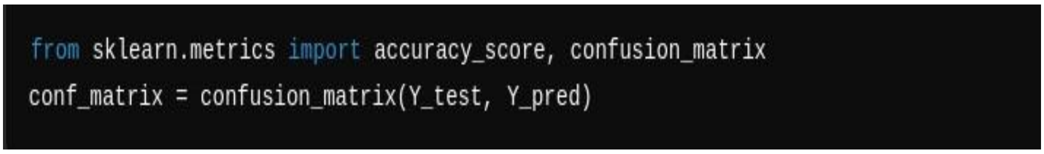
To ensure the robustness of the model, **5-fold cross- validation** was employed, providing a better understanding of how the model generalizes to unseen data. Cross- validation helps mitigate the risk of overfitting by evaluating the model across multiple data splits.

Evaluation and Results :

a) After training, the model was evaluated using the **accuracy score** metric. Both the training and test accuracies were computed to assess model performance:

- **Training Accuracy**: 79.87%
- **Test Accuracy:** 77.92%

These results indicate that the model generalizes well, as the gap between training and test accuracy is minimal. A **confusion matrix** was used to evaluate the true positive, true negative, false positive, and false negative rates.



```
from sklearn.metrics import accuracy_score, confusion_matrix
conf_matrix = confusion_matrix(Y_test, Y_pred)
```

b) **Feature Importance**

The visualization of feature correlations revealed that glucose had the strongest positive correlation with diabetes. This is consistent with medical knowledge that glucose levels are a primary indicator of diabetes.

c) **Limitations**

One major limitation observed in this study was the class imbalance in the dataset. As the number of diabetic patients was significantly smaller than non-diabetic patients, the model was somewhat biased toward predicting the majority class. Techniques like **SMOTE** could be employed in future work to mitigate this issue.

Discussion :

The linear SVM model demonstrates its capability to predict diabetes with a reasonable degree of accuracy. Glucose and BMI were identified as the most significant features influencing the outcome. The results also suggest that the model's performance could be enhanced by addressing the class imbalance, either through resampling methods or adjusting the decision threshold.

Conclusion :

In this study, we applied a Support Vector Machine (SVM) with a linear kernel to predict diabetes using the Pima Indian Diabetes dataset. Our results indicate that the model performs well, achieving an accuracy of 77.92%.

Through exploratory data analysis (EDA), we identified glucose and BMI as the most influential features for predicting diabetes, aligning with established medical knowledge. Additionally, the use of standardization helped improve the model's performance by ensuring that all features contributed equally during training.

Despite these positive results, the study also highlighted certain limitations, particularly the class imbalance in the dataset, which likely affected the model's ability to generalize to minority class examples. Future work could explore techniques such as **SMOTE or ADASYN** to address this issue, as well as the use of non-linear kernels or ensemble models to enhance predictive performance. Moreover, applying more advanced techniques such as deep learning on larger datasets may provide further insights into improving diabetes prediction.

IX Future Work :

To build upon this work, future research could focus on several key areas:

- **Class Imbalance**: Investigating different strategies for addressing class imbalance, such as SMOTE, oversampling, or undersampling techniques.
- **Non-Linear Kernels**: Exploring non-linear kernels (e.g., radial basis function (RBF)) that may capture complex relationships between features more effectively than a linear kernel.

- **Feature Engineering**: Creating new features from existing ones, such as interaction terms, could help capture more nuanced patterns in the data.
- **Deep Learning**: Implementing neural network- based models such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs) on larger datasets to evaluate their performance in diabetes prediction.

REFERENCES :

1. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. https://doi.org/10.1007/BF00994018
2. Jothi, R. and Rashmi, K. (2020). A Comparative Analysis of Machine Learning Models for Diabetes Prediction. *International Journal of Computer Applications*, 175(9), pp.12-18.
3. Kaur, H., & Kumari, V. (2018). Predictive Modelling and Analytics for Diabetes using a Machine Learning Approach. *Applied Computing and Informatics*, 14(2),179–185. https://doi.org/10.1016/j.aci.2018.05.001
4. International Diabetes Federation. (2019). IDF Diabetes Atlas, 9th edition. https://diabetesatlas.org/
5. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. https://doi.org/10.1038/nature14539
6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825– 2830.
7. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357. https://doi.org/10.1613/jair.953