IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 5, May 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205158 |

A Comparative Analysis of Random Forest and Support Vector Machine Models for Enhancing Diabetes Prediction Using Machine Learning

Deepali Gavhane, Kuldip Parbat, Rutvik Raut, Saisharan Pothnedi

Assistant Professor, Department of M.C.A, G.H. Raisoni College of Engineering and MGMT, Wagholi, Pune, India

P.G. Student, Department of M.C.A,G.H. Raisoni College of Engineering and MGMT, Wagholi, Pune, India

P.G. Student, Department of M.C.A, G.H. Raisoni College of Engineering and MGMT, Wagholi, Pune, India

P.G. Student, Department of M.C.A, G.H. Raisoni College of Engineering and MGMT, Wagholi, Pune, India

Abstract: Diabetes is a common and long-lasting health condition. Detecting it early can help improve how it's managed. In our study, we used data analysis techniques to predict diabetes and understand how different factors relate to it. We focused on finding the most important factors linked to diabetes. We used a method called principal component analysis to select these key factors. Our study found that body mass index (BMI) and glucose levels are strongly connected to diabetes. Machine learning algorithms are used to create predictive models for many applications, including illness diagnosis, recommendation systems, stock price prediction, and object identification. This research evaluates the performance of machine learning models: Support Vector Machine (SVM)- using different kernels (linear, poly, rbf), and Random Forest (RF) using predictive performance accuracy.

Among these, Random Forest Classifier gave us the highest accuracy of 81%. This means it was the best at predicting diabetes in our study. Our findings suggest that tools like Random Forest Classifier could be valuable for healthcare professionals in making decisions about diabetes treatment. This information could potentially improve how diabetes is managed in the future.

KEYWORDS: Machine learning, SVM, Random Forest, Clustering, Accuracy, Confusion Matrix.

I. INTRODUCTION

The research aims to develop and evaluate machine learning models, specifically SVM and Random Forest, for predicting diabetes using data from the Kaggle Diabetes Data Repository. These models utilize various patient attributes like age, BMI, blood pressure, insulin level, and plasma glucose concentration (PGC) to classify diabetes. While several machine learning algorithms exist for disease classification, SVM and Random Forest are chosen for their simplicity in training and effectiveness.

The study addresses the following research questions:

1. Can SVM and Random Forest accurately predict diabetes?
2. How do SVM and Random Forest perform in diabetes prediction?
3. How can SVM and Random Forest algorithms improve diabetes prediction accuracy?

By exploring these questions, the research aims to contribute to the development of efficient tools for early diabetes detection, potentially aiding clinicians in decision-making and preventive care.

II. LITERATURE SURVEY

[1]Shafiqul et al. applied eight ML strategies to predict type 2 diabetes development using the San Antonio Heart Study data. Ensemble Naïve Bayes achieved 95.94% accuracy, while Random Forest and Support Vector Machine had lower sensitivity.
[2]Wang et al. proposed a diabetes prediction algorithm to handle missing values in the Pima Indians dataset. Random Forest achieved 87.10% precision and an AUC score of 0.928.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205158 |

[3]Chen et al. investigated boosting algorithms for diabetes classification, achieving 95.30% accuracy with LogitBoost using 10-fold cross-validation.
[4]Roshan and Ashish et al. compared Logistic Regression, Gradient Boosting, and Naive Bayes on the Pima Indians dataset, with Gradient Boosting outperforming the others with an 86% accuracy.
[5]Yukai Li et al. used various ML techniques on data from the Urumqi population to predict diabetes, achieving a g-mean of 94.65% with AdaBoost.
[6]Deepti et al. proposed a model using Support Vector Machine, Decision Tree, and Naïve Bayes on the Pima Indians Diabetes Database for early diabetes detection.

III. METHODOLOGY

3.1. Dataset:
The Pima Indian Diabetes Database is a well-known dataset used for predicting diabetes. It includes data on 768 female patients, with 500 not diagnosed and 268 diagnosed with diabetes.

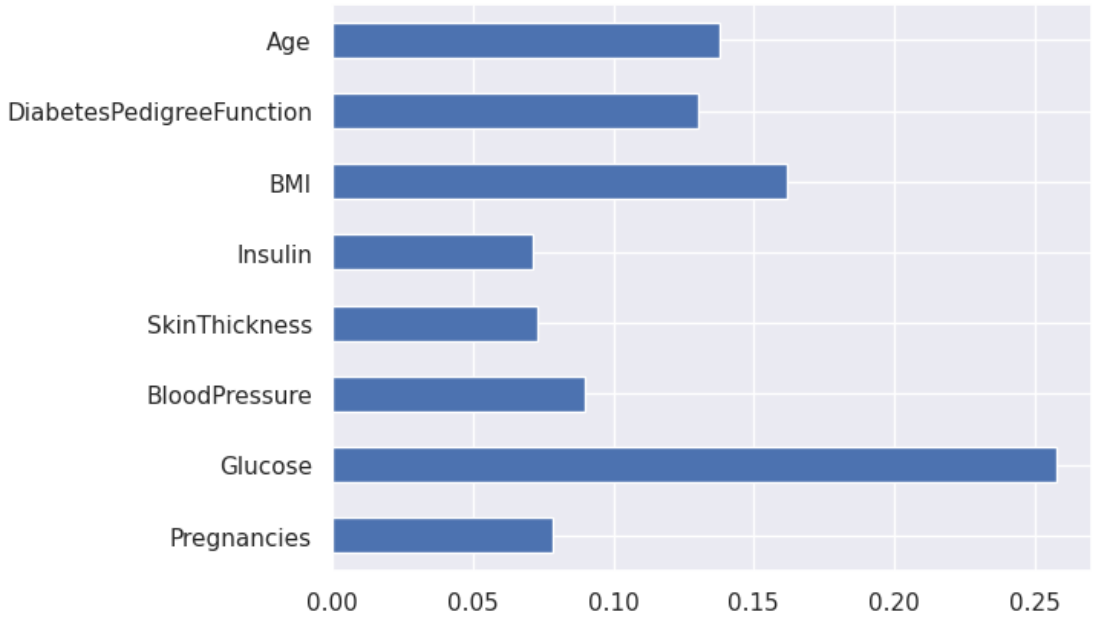
3.2. Data Preprocessing:
To ensure reliable results, we addressed missing or inconsistent data by filling missing values and removing noise. For example, we replaced missing values with attribute medians.

3.2.1. Data Cleaning:
We removed noisy data and handled missing values by replacing zeros with attribute medians, ensuring consistency.

3.2.2. Data Reduction:
Principal component analysis (PCA) was used to extract significant attributes, focusing on key predictors like glucose, BMI, blood pressure, and age.

3.2.3. Data Transformation:
We simplified the dataset using techniques like binning and categorization. Age was categorized into groups, and glucose, blood pressure, and BMI were categorized based on their relationship with diabetes.

3.3. Association Rule Mining:
We applied association rule mining to uncover patterns between health factors and diabetes. By setting support and confidence levels, we identified meaningful rules, revealing the complex relationship between attributes and diabetes diagnosis.



| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205158 |

3.4. Modelling

Three models were used for early prediction of diabetes, following.

3.4.1. Random Forest (RF)

Random Forest (RF) is a versatile machine learning algorithm known for its flexibility, speed, and ability to handle various data types. It combines multiple decision trees to enhance prediction accuracy by aggregating their outputs. RF excels in classification and regression tasks, thanks to its ensemble approach, which mitigates individual tree weaknesses and improves overall performance.

RF utilizes overlapping random trees, where each tree is built using a random subset of attributes. This promotes model diversity and robustness. During prediction, if multiple trees offer insights on a variable, their outputs are combined through majority voting to make a collective decision.

To combat overfitting, RF generates random attribute subsets and thresholds, reducing the risk and enhancing generalization performance. In summary, RF is a powerful algorithm capable of accurate predictions across diverse datasets, making it invaluable in predictive modeling and data analysis.

	precision	recall	f1-score	Support
0	0.81	0.91	0.86	146
1	0.80	0.62	0.70	85
accuracy			0.81	231
macro avg	0.80	0.77	0.78	231
weighted avg	0.80	0.81	0.80	231

3.4.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a effective supervised getting to know set of rules used for category and regression tasks. It works by finding the optimal hyperplane that best separates the different classes in the dataset. SVM can handle both linearly separable and non-linearly separable data through the use of different kernel functions such as linear, polynomial, and radial basis function (RBF).

- **Linear SVM:** Linear SVM works by finding the hyperplane that best separates the classes in the input space. It is particularly effective for linearly separable data and is computationally efficient.

	precision	recall	f1-score	Support
0	0.78	0.92	0.84	146
1	0.80	0.55	0.65	85
accuracy			0.78	231
macro avg	0.79	0.74	0.75	231
weighted avg	0.79	0.78	0.77	231

- **Polynomial SVM:** Polynomial SVM extends the linear SVM by transforming the input space into a higher-dimensional space using polynomial functions. This allows it to handle non-linearly separable data by finding polynomial decision boundaries.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205158 |

	precision	recall	f1-score	Support
0	0.78	0.92	0.85	146
1	0.81	0.56	0.67	85
accuracy			0.79	231
macro avg	0.80	0.74	0.76	231
weighted avg	0.80	0.79	0.78	231

- **RBF SVM:** RBF SVM is a versatile kernel method that maps the input space into an infinite-dimensional space using radial basis functions. This allows it to capture complex non-linear relationships between the input variables, making it suitable for a wide range of datasets.

	precision	recall	f1-score	Support
0	0.78	0.93	0.85	146
1	0.82	0.55	0.66	85
accuracy			0.79	231
macro avg	0.80	0.74	0.76	231
weighted avg	0.80	0.79	0.78	231

SVM is known for its ability to find the optimal decision boundary with a margin of maximum width, known as the maximum margin hyperplane. This property makes SVM robust to outliers and noise in the data. Additionally, SVM allows for the use of regularization parameters to control model complexity and prevent overfitting.
Overall, SVM is a versatile and powerful algorithm that can handle both linear and non-linear classification tasks with high accuracy and robustness.

3.5. Analysis of Results Using Different ML Techniques.
In our study, we constructed four classifier models: Random Forest (RF), Support Vector Machine (SVM) with three kernel functions (linear, polynomial, and radial basis function or RBF). Before training the data, we removed any outliers present. We compared the performance of these models using accuracy and F1-score metrics.

- Random Forest (RF): Achieved an accuracy of 81% with an F1-score of 0.7964. Random Forest is a versatile and powerful algorithm that combines the predictions of multiple decision trees to improve accuracy. It handles different types of data and is robust to overfitting.

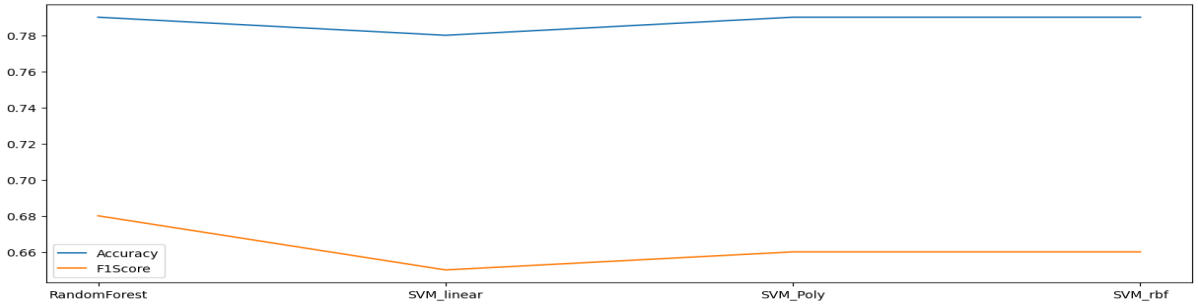
- Support Vector Machine (SVM):
- Linear Kernel: Achieved an accuracy of 78% with an F1-score of 0.7700. - Polynomial Kernel: Achieved an accuracy of 79% with an F1-score of 0.7837. - RBF Kernel: Also achieved an accuracy of 79% with an F1-score of 0.7800.
SVM is a famous set of rules used for class tasks. It finds the optimal hyperplane to separate different classes in the data. SVM with different kernels can handle linear and non-linear data patterns.

We noticed that both Random Forest and SVM achieved high accuracy levels of 79%. Additionally, after applying hyperparameters to Linear Regression (LR), we were able to improve the accuracy by 1%.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205158 |



Preprocessing Insights:

- We observed a strong positive correlation between BMI and the number of pregnancies, indicating that diabetic individuals tend to have higher BMIs.
- The presence of outliers in clinical test reports, particularly in those with high pedigree function, suggests that diabetes may have a hereditary component.
- Women with more pregnancies had higher BMI, indicating a potential risk factor for diabetes.
- Additionally, women over 31 years of age were at a higher risk of diabetes diagnosis compared to younger women.

Performance Evaluation:

- The confusion matrix provided insights into the classification performance, with the diagonal elements representing correctly classified instances.
- Random Forest showed promising results in disease prediction with high accuracy.

In conclusion, our analysis suggests that Random Forest is a suitable model for predicting disease outcomes with high accuracy. Additionally, our insights into preprocessing and performance evaluation provide valuable information for further research and application in healthcare settings.

IV. SUMMARY AND CONCLUSION

In our study, we examined the performance of different machine learning models in predicting diabetes. We tested Random Forest, Support Vector Machine with various kernels. After careful analysis, we found that Random Forest and SVM with polynomial and RBF kernels achieved the highest accuracy of 81%.

Our preprocessing steps revealed interesting insights, such as a strong correlation between BMI and the number of pregnancies, indicating a potential risk factor for diabetes. Additionally, the presence of outliers in clinical test reports suggested a hereditary component in diabetes.

Random Forest emerged as the most promising model for disease prediction, with a high accuracy rate.

In conclusion, our study demonstrates the effectiveness of machine learning in predicting diabetes outcomes. By leveraging advanced algorithms and preprocessing techniques, we can improve disease diagnosis and management. Our findings contribute valuable insights to healthcare research and highlight the potential of machine learning in improving patient outcomes.

REFERENCES

1. T. M. Alama, M. A. Iqbala, Y. Ali et al., “A Model for Early Prediction of Diabetes,” Informatics in Medicine Unlocked, vol. 16, Article ID 100204,2019.
2. S. Kapoor and K. Priya, “Optimizing hyper parameters for improved diabetes prediction,” International Research Journal of Engineering and Technology, vol. 5, 2018.
3. S. Srivastava, L. Sharma, V. Sharma, and A. Kumar, “Prediction of diabetes using artificial neural network approach,” in Engineering Vibration, Communication and Information ProcessingVol. 29, Springer, Berlin/Heidelberg, Germany, 2020.
4. T. Santhanam and M. S. Padmavathi, “application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis,” Procedia Computer Science, vol. 47, 2015.
5. N. Nai-aruna and R. Moungmaia, “Comparison of classifiers for the risk of diabetes prediction,” Procedia Computer Science, vol. 69, 2015.
6. A. Mujumdara, V. Vaidehi, Diabetes prediction using machine learning algorithms,” Procedia Computer Science, vol. 165, 2019.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details