SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 7 | July 2025 - Peer Reviewed Journal

DIABETES MELLITUS PREDICTION AND DIAGNOSIS USING MACHINE LEARNING

Zara Hussain[1], Vikas Garg[2], Dr. Tapsi Nagpal[3]
[1]*Amity University, Noida, Uttar Pradesh, India*
[2]*Lingayas Vidyapeeth, Faridabad, Haryana, India*
[3]*Lingayas Vidyapeeth, Faridabad, Haryana, India*

Article DOI: https://doi.org/10.36713/epra22950
DOI No: 10.36713/epra22950

ABSTRACT

One chronic metabolic disease, which is now a major global public health issue is diabetes mellitus. It is caused by either insufficient insulin synthesis or poor insulin use by the body, and elevated blood sugar levels mostly mark it. Numerous factors, such as ageing, obesity, poor eating habits, sedentary lifestyles, and genetic susceptibility, affect its development. In order to reduce long-term effects like cardiovascular illnesses, kidney problems, and nerve damage, early detection of diabetes is essential.

In capacity various machine learning methods are used to analyse complex medical information and aid in early diagnosis, including Pima Indians Diabetes, has garnered significant attention due to the rapid breakthroughs in data science. Clinical factors include blood pressure, blood sugar, age, BMI, and family history. Finding patterns and correlations in with conventional statistical methods may prove difficult.[5] Examples of these models are XGBoost, K-Nearest Neighbours (KNN), Random Forest, and Support Vector Machine (SVM).[9],[10],[11] The objective of the previously described project is to develop machine learning models for the identification of diabetes. Using open-source datasets by applying supervised machine learning methods.

KEYWORDS- *Diabetes prediction, Machine learning, Classification, confusion matrix, healthcare, ROC curve and AUC, predictive modelling*

I.INTRODUCTION

Diabetes mellitus, especially Type 2, has become one of global health concerns of the twenty-first century. More over 537 million people globally had diabetes in 2021, according to the International Diabetes Federation (IDF); estimates suggest that number would increase to 643 million by 2030 and 783 million by 2045. [1] In addition to endangering people's quality of life, this startling rise puts a heavy financial burden on healthcare systems, particularly in developing nations [1]. If the aforementioned problem is not identified or is not adequately treated, it may instead result in serious complications. In response to this growing concern, AI is rapidly transforming the healthcare sector by utilizing advanced computational capabilities and data-centric approaches. It is significantly reshaping fields such as diagnostics, patient monitoring, personalized care planning, and predictive modeling [2]. Within the broader scope of AI, ML has demonstrated considerable potential in enabling early disease identification and enhancing medical outcomes by recognizing patterns and aiding clinical decisions [3].

In today's healthcare ecosystem, ML models are increasingly being adopted to assist medical professionals to make faster, more accurate, and data-backed diagnoses. These systems can analyse vast amounts of patient data far beyond human capacity, uncovering subtle correlations and trends that are critical for preventive care [2]. In the context of diabetes prediction, ML algorithms are proven effective in recognizing those who are at risk before symptoms appear, thereby enabling timely intervention and lifestyle modifications [4].

Important health parameters and features, including age, blood pressure, glucose, diabetes, skinfold thickness, BMI, diabetes pedigree function, and insulin concentration levels, are included in this study's popular open-source dataset of female patients.[5] These characteristics are essential training inputs for supervised ML models capable of reliably detecting diabetes. This study intends to demonstrate how incorporating ML into clinical practice can improve diagnostic efficiency and decrease human error, which will ultimately lead to improved disease management and resource allocation.

II. LITERATURE SURVEY

Sharma et al. [5] emphasized the growing role of machine learning (ML) in early diabetes prediction using the Pima Indians Diabetes Dataset (PIDD), where features like blood pressure, glucose, BMI, and age are key indicators. Patil et al. [9] showed K-Nearest Neighbors (KNN) could detect hidden patterns, while Breiman [10] noted Random Forest's strength in capturing complex relationships. Cortes and Vapnik [13] highlighted SVM's efficiency with high-dimensional data, though sensitive to hyperparameters, and Chen & Guestrin [11] demonstrated XGBoost's superior performance through regularization and parallelism. Medical datasets often contain zero values in critical features, requiring imputation for reliable modeling (Kumar et al. [14]). Feature scaling, especially standardization, ensures fair contribution in KNN and SVM (Jain et al. [15]). Splitting data into 80% training and 20% testing is key for evaluating generalization (Mishra et al. [16]; Verma et al. [17]). Tuning hyperparameters—like K in KNN

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 7 | July 2025 - Peer Reviewed Journal

(Pandey et al. [18]), tree depth in Random Forest (Breiman [19]), kernels in SVM (Cortes & Vapnik [20]), and XGBoost settings (Chen & Guestrin [11])—is crucial, often done via cross-validation (Verma et al. [17]). Metrics such as accuracy, precision, recall, and F1-score (Pandey et al. [21]), alongside confusion matrices and heatmaps (Gupta et al. [22]; Yadav et al. [23]), guide performance evaluation. ROC-AUC analysis showed Random Forest (0.83) outperformed XGBoost (0.82), SVM (0.81), and KNN (0.79) (Pandey et al. [25]). SMOTE was applied to balance the dataset, improving fairness in learning (Chawla et al. [27]; Singh et al. [26]). This workflow enabled a robust comparison of models for diabetes prediction.

A. Proposed Methodology

This study is configured and uses a comparative methodology to predict diabetes at the initial stage using various machine learning algorithms. From the purification and preparation of a number of data sets for education and evaluation models, each stage is carried out for reliable and effective implementation [6] [7]. Most algorithms can balance understanding of accuracy, calculation efficiency and analysis damage by considering actual placement under healthy conditions [2]. In order to ascertain if the clinical test is a diabetes risk factor., this study seeks to ensure dependable support for machine learning and identify the most effective model [3]. This study's primary objective is to assess and contrast different machine learning models. designed to predict diabetes in individuals, using a range of clinical and physiological features [5].This methodology is carefully configured to ensure a systematic approach from preliminary data processing to model. This section describes a brief description of sequential steps [6], data set, preparation and analysis of selected algorithms and data.

B. Dataset Description

The study's data set originates from a reputable PIMA diabetes set that can be publicly used through the UCI learning repository [5]. This data set has become a standard for mechanical training models in the medical field, especially binary classifications such as diabetes [5].

The data set consists of 768 samples, which are all 21 years old or older than the Indian community of PIMA. This emphasis on a particular population can limit the generalization of wider demographic statistics, but adds consistency level of data collection [5].

Each data record includes eight clinical features, which are recognized as important diabetes Skin thickness, insulin, blood pressure, blood sugar, age, body mass index, and diabetes at home [8]. The target variables have binary results: 0 denotes a member with diabetes, and 1 denotes the presence of diabetes.. These features are biologically and clinically related, as a rule, are part of the diagnosis evaluation of diabetes, and this data set is suitable for research [8].

III. OVERVIEW OF SELECTED MACHINE LEARNING ALGORITHMS

Four different types of ML classifications and properly mentioned effective algorithms have been implemented to create a comparative structure of prediction. Each algorithm shows a unique view of binary classification and provides a variety of opportunities related to accuracy, interpretation and calculation efficiency [6].

A. K-Nearest Neighbors (KNN)

Using supervised learning, the KNN method determines identifies the 'K' closest data points to a new input into the most common category. In this study, KNN is applied to a diabetes dataset using features like Blood pressure, age, BMI, and glucose levels are used to determine if a person has diabetes.. It helps uncover hidden patterns in clinical data that traditional methods might miss.[9]

B. Random Forest (RF)

To increase prediction accuracy and reduce overfitting, several DT are created with the RF ensemble learning method, and the outcomes are subsequently combined.. Random subsets of characteristics make up every DT in the woodland. a majority vote is typically used to determine the outcome. Random Forest is a strong contender for medical classification problems because of its robustness and ability to manage nonlinear correlations and dynamic relationships between variables [10].

C. Support Vector Machine (SVM)

The supervised learning algorithm The main objective of SVM is to identify the best hyperplane in a high-dimensional feature space for efficiently dividing two classes.It works especially well with high-dimensional datasets and can simulate intricate decision limits using kernel functions. It can, however, be computationally demanding and very sensitive to hyperparameters such as gamma, regularisation, and kernel selection [6][9][13].

D. Extreme Gradient Boosting (XGBoost)

The performance of XGBoost, an optimised gradient boosting solution, in competitive machine learning tasks has made it popular. It builds models incrementally, with every new model seeking to rectify the flaws of the one before it.. For structured data issues like diabetes classification, XGBoost is incredibly effective and potent since it incorporates regularisation techniques, parallel processing, and automatic handling of missing data [11].

IV. METHODOLOGICAL WORKFLOW

The methodology is executed in multiple well-defined stages to ensure a clean, consistent, and unbiased evaluation of each algorithm. The process flow is detailed below [6][7].

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 7 | July 2025 - Peer Reviewed Journal

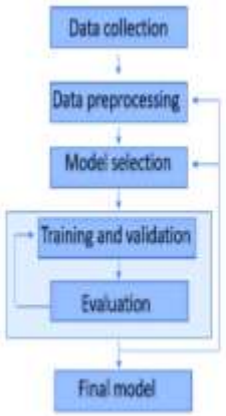


Figure 1: Implementation Workflow

A. Data Cleaning and Preprocessing

Preprocessing is crucial in ensuring it is clean and appropriate for ML model training. BP, skin thickness, insulin, glucose, & BMI are among the features in this dataset which have zero values; these are not medically acceptable & most likely reflect missing or unrecorded data. [14].Depending on context and distribution analysis, these values are eliminated or imputed using suitable statistical methods, like median imputation[14]. Once missing values are handled, all of the input variables' ranges are aligned by feature scaling. This is especially important for models that rely on optimization, such as SVM, and those that use distance calculations, like KNN [15].Standardization (z-score normalization) is typically used to center the data around zero with unit variance [15].

B. Splitting the Dataset

A typical approach to accurately assess the model's performance is to divide the dataset into training and testing sets in an 80:20 ratio [16]. Models are constructed using the training subset, and the test subset is saved for the last assessment. This method aids in assessing how effectively the models extrapolate to fresh, untested data. [16][17].

C. Model Training and Hyperparameter Tuning

All four selected models are trained on the training dataset. Tuning of hyperparameters is carried out in this stage to identify the ideal configurations that optimise performance. For example: KNN relies on selecting an optimal value for k [18].

Effective use of Random Forest requires hyperparameter tuning, notably the quantity of trees and each tree's depth. [19].Proper kernel selection and regularization parameter tuning are crucial for the success of SVM models [20].XGBoost tuning involves setting variables include the maximum tree depth, learning rate, and quantity of trees (estimators) [21]. In order to minimize variance and avoid overfitting, hyperparameter optimization is often performed using techniques like k-fold cross-validation [17][21].

D. Model Evaluation

The reserved test dataset is utilized to evaluate the models following training. Their effectiveness is measured with commonly used performance metrics including:

Model	Accuracy	Precision	Recall	F1 Score
SVM	0.7468	0.6667	0.5185	0.5833
KNN	0.7468	0.6400	0.5926	0.6154
Random Forest	0.7722	0.6957	0.5926	0.6400
XG Boost	0.7342	0.6250	0.5556	0.5882

Table 1: Comparison table

Accuracy - Percentage of all predictions made that were accurate [21].

$$\text{“}Accuracy = \frac{TP+TN}{TP+TN+FP+FN}\text{ “}$$

Precision - Proportion of true positives among predicted positives (helps assess false positives) [21]

$$\text{“}Precision = \frac{TP}{TP+FP}\text{”}$$

Recall (Sensitivity): How successfully a model detects positive situations while reducing false negatives depends on its recall, this represents the ratio of genuine benefits to total benefits. [21].

$$\text{“}Recall = \frac{TP}{TP+FN}\text{”}$$

The F1-Score is the harmonic mean of precision and recall. has applications in unbalanced datasets. [21].

$$\text{“}f1\text{-}score=2 \times \frac{precision \times Recall}{Precision+Recall}\text{,,}$$

Four main outcomes impact the model's success in binary classification issues, especially in medical diagnostics. When the model accurately detects a positive case, such as forecasting that a patient has diabetes when they actually do, this is known as a TP [21]. A TN indicates that a patient has been correctly classified as not having diabetes by the model. Conversely, a FN occurs when the model does not detect a person with diabetes, while a FP occurs when Diabetes is mistakenly diagnosed in a patient who does not have it. [21]. These results are particularly crucial in the medical field, where both missed diagnoses and inaccurate notifications can greatly impact patient care. These measures provide a comprehensive analysis of each model's functionality, encompassing not just its overall accuracy but furthermore, its capacity to differentiate between positive and negative types. [21].

V. CONFUSION MATRIX AND VISUALIZATION

Confusion matrices are created for every classifier to examine the model's behavior more comprehensively. They provide a detailed explanation of FP, FN, TP, and TN. [22].Visualization tools such as heatmaps display these matrices intuitively, making it easier to interpret misclassification trends [23].
In addition, visual comparisons of the performance metrics across all models are created using bar plots or grouped line

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 7 | July 2025 - Peer Reviewed Journal

graphs. This visual analysis aims to clearly identify which model performs the best across various different evaluation dimensions [23].

VI. MODEL COMPARISON

Four ML methods are compared in this section using Random Forest, KNN, SVM, & XGBoost on the Pima Indians Diabetes Dataset. Among the metrics utilized to assess the models were confusion matrix, recall, accuracy, precision, or F1 score. [14][22]. KNN came in second place, while Random Forest outperformed SVM and XGBoost overall. This analysis sheds light on the advantages and disadvantages of each diabetes prediction method. [19][20].

To better analyse the efficacy of the classification models, confusion matrices were generated using the sklearn.metrics package's confusion_matrix function and displayed using matplotlib [22][23]. These matrices give a detailed picture of the model's predictions and show the counts of TP, TN, FP, and FN [22]. A more complete grasp of the benefits and drawbacks of each model is possible by comparing these images to standard parameters such as F1-score, recall, accuracy, and precision. [22][23][27].

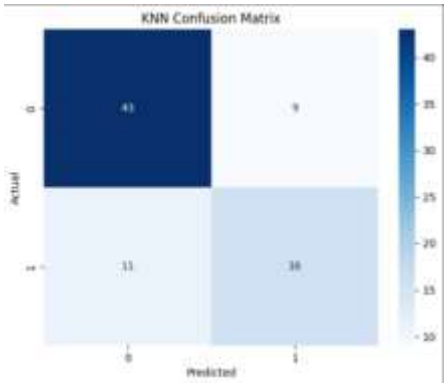


Figure 2: KNN Matrix

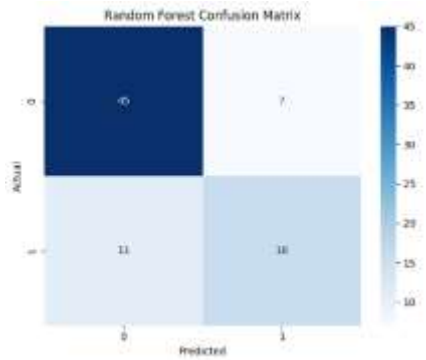


Figure 3: Random Forest Matrix

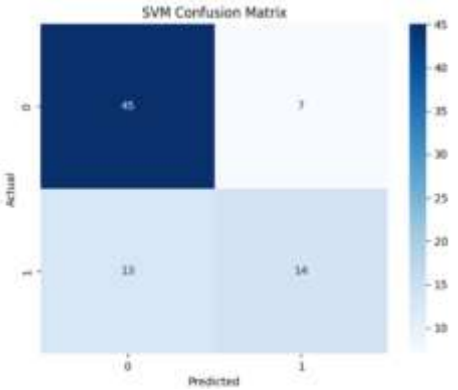


Figure 4: SVM Matrix

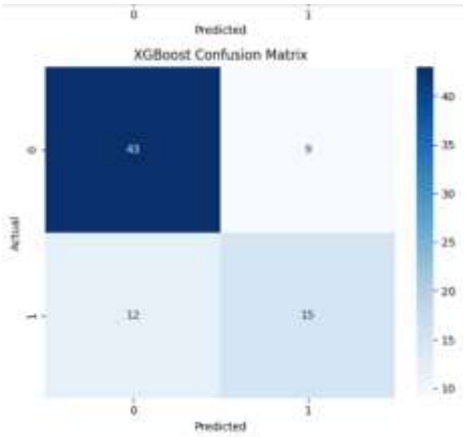


Figure 5: XGBoost Matrix

VII. ROC CURVE

We displayed confusion matrix heatmaps and ROC curves to evaluate the effectiveness of any ML model. The pseudo positive rate (one minus specificity) and the ROC curve show the trade-off between along with the true positive rate (recall) for different threshold values [24]. For a more succinct summary of the performance of the model was further assessed by looking at the area under the ROC curve (AUC).; a greater AUC value denotes a better ability to distinguish between cases both with and without diabetes. [25]. To find the best diabetes prediction model, we analyzed the ROC curves of KNN, Random Forest, SVM, and XGBoost. Ultimately, the model with the best AUC score was selected.

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 7 | July 2025 - Peer Reviewed Journal

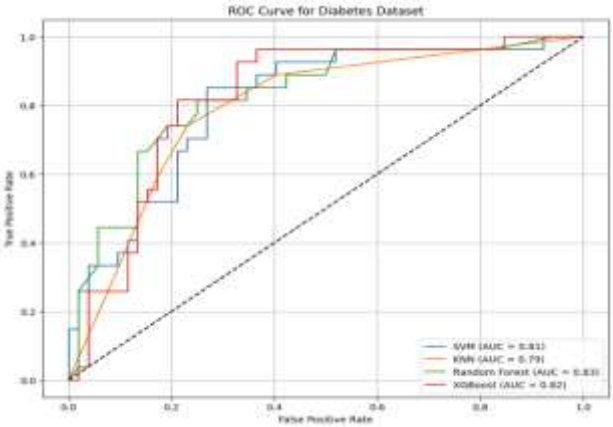


Figure 6: ROC Curve for Diabetes Dataset

A. Interpretation

All four models demonstrate solid classification performance, with AUC values at or near 0.80. This indicates that each model is fairly capable of differentiating between positive (diabetic) and negative (non-diabetic) cases [24].

1. Random Forest leads with the highest AUC of 0.83, highlighting it as the most effective model for this classification task [25].
2. XGBoost is a close second with an AUC of 0.82, making it a highly competitive alternative [25].
3. SVM also performs well, achieving an AUC of 0.81, only slightly behind XGBoost [24].
4. While KNN has the lowest AUC of 0.79 among the evaluated models, it still performs within an acceptable range and can be considered a reasonably good classifier [25].
5.

According to the ROC curve analysis, the RF model had the highest AUC (0.83), followed by XGBoost (0.82), which showed that they were better at differentiating between cases with and without diabetes [25]. The SVM and KNN models also performed reasonably well, with AUCs of 0.81 and 0.79, respectively [24]. Overall, the ROC curve demonstrated that RF is the best classifier among those assessed, which makes it the best option for this study's precise diabetes prediction [25].

Table 2: AUC values

Index	Algorithm	AUC values
0	Random forest	0.83
1	XGBoost	0.82
2	SVM	0.81
3	KNN	0.79

VIII. OVERSAMPLING TECHNIQUE: ADDRESSING CLASS IMBALANCE WITH SMOTE

When creating trustworthy ML models for class imbalance & diabetes prediction presents a significant obstacle. With more non-diabetic patients than diabetic ones, the PIDD utilized in this work shows a notable imbalance [26]. Because of this disparity, models may Favor the majority class, thereby resulting in poor forecasts for the minority group., in this instance the diabetic population [26, 29]. SMOTE was used to solve this problem. By interpolating between current cases and their closest neighbors, SMOTE is a successful oversampling technique that creates synthetic minority class samples. [29]. Compared to random oversampling, SMOTE improves minority class representation during training and reduces the risk of overfitting [29, 28].

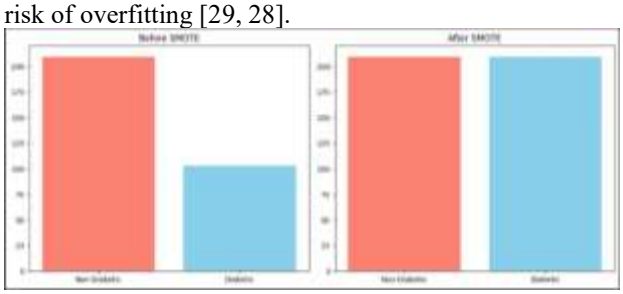


Figure 7: SMOTE Analysis

Post-SMOTE, the visual comparison shows that each class was brought to an equal level, ensuring a more balanced dataset. Initially, the dataset consisted of 209 Non-Diabetic patients and only 104 Diabetic patients, highlighting a major significant class imbalance [26,29]. After applying SMOTE, the number of diabetic instances was synthetically increased to 209, thereby matching the count of the non-diabetic cases. This balancing is very crucial for training unbiased machine learning models that can perform equitably across both classes [28,29].

IX. CONCLUSION

This study employed supervised ML techniques with the PIDD to detect and predict diabetes mellitus. Four widely used classification algorithms—KNN, RF, SVM, and XGBoost—were implemented to determine which model most effectively identifies diabetic individuals using key clinical features.

The RF classifier consistently performed best after extensive data preprocessing, which included using SMOTE to address class imbalance, measures like ROC-AUC, F1-score, confusion matrix, recall, accuracy, and precision are utilized in model training, hyperparameter tuning, or assessment. At 77.22%, the highest accuracy and an AUC of 0.83, it showed a significant capacity to distinguish between cases with and without diabetes. Furthermore, XGBoost produced accurate and dependable findings.

These findings show how machine learning could aid in diabetes early detection, which is essential for prompt medical attention and lowering chronic health problems.. RF had been the most reliable and accurate model among all those examined, suggesting that it may find practical application in clinical decision support systems. These findings could be expanded upon in future studies by using larger and more varied datasets and looking into deep learning techniques for even greater improvement.

In conclusion, incorporating and implementing intelligent predictive models into healthcare workflows—especially when combined with techniques like SMOTE to handle data imbalance—can empower medical professionals to make

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 7 | July 2025 - Peer Reviewed Journal

quicker, more accurate decisions, eventually enhancing patient care and making the best use of medical resources.

REFERENCES

1. *[1] International Diabetes Federation. (2021). IDF Diabetes Atlas, 10th edition. https://diabetesatlas.org*
2. *[2] World Economic Forum. (2024). How AI is improving diagnostics and health outcomes. https://www.weforum.org/stories/2024/09/ai-diagnostics-health-outcomes*
3. *[3] Goyal, M., & Kadam, S. (2023). AI in Healthcare: Revolutionizing Early Diagnosis. Journal of Artificial Intelligence in Healthcare, 5(2), 123–132.*
4. *[4] Analytics Vidhya. (2022). Diabetes Prediction Using Machine Learning. https://www.analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machine-learning*
5. *[5] UCI Machine Learning Repository. (n.d.). Pima Indians Diabetes Dataset. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database*
6. *[6] Raschka, S., & Mirjalili, V. (2022). Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2 (3rd ed.). Packt Publishing.*
7. *[7] Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2019). A primer on deep learning in genomics. Nature Genetics, 51(1), 12–18. https://doi.org/10.1038/s41588-018-0295-5*
8. *[8] American Diabetes Association. (2022). Standards of Medical Care in Diabetes – 2022. Diabetes Care, 45(Supplement_1), S1–S264. https://doi.org/10.2337/dc22-S001*
9. *[9] Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician, 46(3), 175–185. https://doi.org/10.1080/00031305.1992.10475879*
10. *[10] Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3), 18–22. https://CRAN.R-project.org/doc/Rnews/*
11. *[11] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785*
12. *[12] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, 15, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005*
13. *[13] Singh, A., & Misra, N. (2017). Detection of Diabetes Using Support Vector Machine. International Journal of Advanced Research in Computer Science and Software Engineering, 7(5), 92–97. https://doi.org/10.23956/ijarcsse.v7i5.504*
14. *[14] Brownlee, J. (2020). Data Cleaning for Machine Learning. Machine Learning Mastery. https://machinelearningmastery.com/data-cleaning-for-machine-learning/*
15. *[15] Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.*
16. *[16] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.*
17. *[17] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence.*
18. *[18] Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician, 46(3), 175-185.*
19. *[19] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.*
20. *[20] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.*
21. *[21] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD.*
22. *[22] Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies, 2(1), 37-63.*
23. *[23] Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95.*
24. *[24] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010*
25. *[25] Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2*
26. *[26] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035*
27. *[27]https://link.springer.com/article/10.1007/s40009-025-01671-w*
28. *[28] 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*
29. *[29] Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from Imbalanced Data Sets. Springer. https://doi.org/10.1007/978-3-319-98074-4*

ACKNOWLEDGEMENT

This paper used the Pima Indians Diabetes Dataset from Kaggle, which includes 768 records with 8 features and a binary outcome for diabetes. We handled missing or zero values by replacing them with median values and applied standard scaling. The data was split into 80% training and 20% testing. The models used were KNN (k=5), Random Forest (100 trees), SVM (RBF kernel, C=1.0), and XGBoost (100 trees, learning rate 0.1). All work was done in Python using libraries like scikit-learn, XGBoost, NumPy, and Pandas. The models were evaluated using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC.