COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR DIABETES PREDICTION IN PIMA INDIAN WOMEN

Amit gaddi gaddiat@mail.uc.edu
School of Information Technology, University of Cincinnati, Ohio, USA

Abstract

Using the Pima Indian Diabetes Dataset, this study assessed several machine learning algorithms for predicting diabetes in female Pima Indian patients 21 years of age and older. Seven methods were examined in the study: XGBoost, Decision Tree Classifier, Random Forest Classifier, K-Nearest Neighbors, Support Vector Machine (SVM), Gradient Boosting Classifier, and Logistic Regression. Accuracy, precision, recall, F1-score, and AUC-ROC measures were used to evaluate performance.

Handling zero values, addressing outliers, and imputing missing data using a stratified median technique were all part of the data pretreatment procedure. Among the feature engineering tasks were the classification of BMI, insulin and glucose levels.

With 90.91% accuracy and balanced precision, recall, and F1-score (all 0.85), SVM was the best performance, according to the results. Closely behind, with an accuracy of 90.26% and the highest AUC-ROC of 89.41%, was logistic regression. Additionally, Random Forest and Gradient Boosting performed admirably.

The study's conclusions, which emphasize the use of SVM and logistic regression, are consistent with other studies on diabetes prediction. Nevertheless, the dataset's drawbacks include its relatively modest size and its concentration on a particular group. Subsequent research endeavors may encompass verifying these models on a wider range of demographics, integrating supplementary data sources, and investigating more sophisticated methodologies such as deep learning.

This work adds to the increasing amount of data that supports the application of machine learning in healthcare, especially for high-risk groups and early illness diagnosis.

Keywords: Diabetes Prediction, Diabetes Classification, Pima Indian Diabetes, Machine Learning, Classification Metrics.

1. Introduction

With high blood glucose levels, diabetes mellitus is a chronic metabolic illness that affects millions of people worldwide. Effective care and the avoidance of problems associated with diabetes need early identification and precise prognosis [1, 4]. When utilizing benchmark datasets like the Pima Indian Diabetes Dataset, machine learning approaches have demonstrated encouraging results in the prediction of diabetes in recent years [1, 4].

The study focuses on female Pima Indians who are 21 years of age and older. Because of a hereditary propensity to insulin resistance, this community has a high prevalence of diabetes [12]. Using measures including accuracy, precision, recall, F1-score, and AUC-ROC, the study seeks to determine which machine learning algorithm best classifies patients as either diabetics or non-diabetics based on a variety of variables.

Support Vector Machines (SVM), Random Forest, Naïve Bayes, Decision Trees, and, more recently, Deep Learning approaches have all been investigated in the past using this dataset to predict diabetes [1, 2, 4, 9]. SVM has demonstrated consistently good performance, with claimed accuracies between 84% and 94% [9]. Strong outcomes have also been shown using Random Forest; one research reported an accuracy of 79.57% [1].

Numerous research have emphasized the significance of feature selection and data preparation in improving model performance [2, 5]. Skin fold thickness, age, insulin, BMI, and glucose levels have been found to be important characteristics for classifying diabetes [1, 2, 11]. Additionally, recent studies have highlighted the importance of thorough assessment measures that go beyond accuracy, such as precision, recall, F1-score, and AUC-ROC, especially when working with unbalanced datasets [1, 3, 7, 8].

By thoroughly comparing many machine learning algorithms, putting strong data pretreatment approaches into practice, and making use of an extensive range of assessment criteria, this work expands on earlier research. In order to support further efforts in early diabetes identification and care, the aim is to determine the best successful strategy for diabetes prediction in this particular group.

1.1 Research Question:

Which machine learning algorithm, as evaluated by metrics such as accuracy, precision, recall, F1-score, and AUC-ROC, best classifies female patients of Pima Indian descent who are 21 years of age or older as diabetic or non-diabetic based on the provided features?

2. Literature review

Millions of people worldwide suffer from diabetes mellitus, a chronic metabolic disease marked by increased blood glucose levels. Effective management and the avoidance of complications from diabetes depend heavily on early detection and precise prognosis [1, 4]. The Pima Indian Diabetes Dataset is a benchmark for diabetes research, and machine learning techniques have demonstrated encouraging results in diabetes prediction in recent years [1, 4]. This dataset is important because it focuses on female Pima Indians who are 21 years of age and older. Because of their genetic susceptibility to insulin resistance, this group has a high incidence rate of diabetes [12].

Using this dataset, several research have investigated different machine learning algorithms for diabetes prediction in an effort to identify the system that best categorizes patients as diabetic or non-diabetic depending on the features supplied. Support Vector Machine (SVM), Random Forest, Naïve Bayes, Decision Trees, and, more recently, Deep Learning techniques are the most widely utilized algorithms [1, 2, 4, 9].

Numerous research have shown that Support Vector Machine (SVM) is among the best algorithms for predicting diabetes. Al-Sideiri et al. (2019) offered a thorough analysis emphasizing SVM's consistently excellent accuracy [9]. In Mercaldo et al. (2017), for example, SVM accuracy of 94% was observed, but other investigations found accuracies ranging from 84% to 90% [9]. After performing parameter adjustment and data pretreatment, Miao Yuxin's study obtained 87.01% accuracy with SVM, highlighting the significance of these procedures in improving model performance [2].

Additionally, Random Forest has performed admirably. A system for electronic diagnosis was proposed by Chang et al. [1]. Out of all the algorithms used, Random Forest performed the best overall, with an accuracy rate of 79.57%. In order to provide a more thorough assessment of model performance, this study also underlined the significance of using several evaluation metrics, such as accuracy, sensitivity, specificity, F1-score, and AUC-ROC [1].

The use of deep learning techniques has grown significantly, as demonstrated by Naz and Ahuja's (2020) remarkable 98.07% accuracy, which outperforms conventional machine learning algorithms [4]. This demonstrates the promise of cutting-edge methods for diabetes prediction.

There has also been promise from other algorithms. Chang et al. [1] have pointed out that Naïve Bayes has proven to be efficient with fewer features. In their comparative analysis, Mahmud et al. discovered that Naive Bayes achieved 74% accuracy and the highest F1-score [7]. Decision trees have also shown good performance; Saru and Subashree's bootstrapping-assisted 94.44% accuracy was reported [6].

It has been determined that choosing features and preprocessing data are essential stages in enhancing model performance. The dataset was narrowed down to four essential characteristics by Vaishali R. et al. using a genetic algorithm: age, diabetes pedigree function, BMI, and plasma glucose concentration [5]. This method produced an accuracy of 83.04% when paired with a Multi-Objective Evolutionary fuzzy classifier. The most crucial characteristics for classifying diabetes have been repeatedly found to be skin fold thickness, age, insulin, BMI, and glucose levels [1, 2, 11].

Recent studies have highlighted the importance of evaluation indicators that go beyond accuracy. Studies have progressively included measurements like precision, recall, F1-score, and AUC-ROC, even though accuracy is the most frequently reported statistic [1, 3, 7, 8]. These extra measures offer a more thorough evaluation of the model's performance, particularly when dealing with unbalanced datasets [9].

It has been demonstrated that data preparation methods have a major impact on model performance. Data cleaning and normalization increased SVM accuracy from 85.06% to 87.01%, as shown by Miao Yuxin's study [2]. Similar to this, Mahmud et al. robustly evaluated algorithm performance using a 70-30 train-test split and 10-fold cross-validation [7].

The potential for more comprehensive diabetes prediction and monitoring through the combination of several algorithms and real-time data collecting has also been investigated in recent study [7]. This method might offer forecasts that are more precise and timely, which could enhance patient outcomes.

Although the Pima Indian Diabetes Dataset offers insightful information, it's crucial to be aware of its limitations, which include its emphasis on female patients over the age of 21 and probable missing/inconsistent data [2, 5]. Subsequent investigations may concentrate on refining these algorithms even further, investigating novel methods for feature selection, and utilizing these models on diverse populations to extrapolate results.

In conclusion, there is no definite agreement on the optimal machine learning method, despite the fact that a number of them have demonstrated promise in diabetes prediction when utilizing the Pima Indian dataset [1, 2, 4, 7,

9]. The features chosen, the preprocessing methods, and the assessment measures applied all affect performance. This emphasizes the necessity of doing a thorough comparison analysis with a variety of algorithms and performance indicators in order to identify the best method for diabetes prediction in this population. Such an analysis would support ongoing initiatives in early diabetes detection and management and offer healthcare professionals insightful information.

3. Methodology

3.1 Dataset Description:

In order to assess and compare several machine learning algorithms for diabetes prediction in female patients of Pima Indian heritage who are 21 years of age or older, this study used a thorough methodology. The Pima Indian Diabetes Dataset, which comprises many medical predictor factors and a single target variable indicating whether the patient has diabetes, was the dataset utilized for this study. Preprocessing the data, feature engineering, training the model, and performance assessment were all included in the technique.

The study made use of the Pima Indians Diabetes dataset, which includes 768 records of female patients who were at least 21 years old and of Pima Indian ancestry.

Among these characteristics are:

- Pregnancies: Total number of pregnancies
- Plasma glucose levels, measured two hours following an oral glucose tolerance test
- Blood Pressure: The blood pressure at diastolic level (mm Hg)
- SkinThickness: The thickness of the triceps skin fold (mm)
- Insulin: Serum insulin (mu U/ml) after two hours
- BMI: Body mass index (kg of weight divided by m of height) ^2)
- DiabetesPedigreeFunction: A function assigns a probability to diabetes depending on family history.
- Age: Years of age
- Outcome: Class variable indicating if diabetes is present or absent (0 or 1).

The density distributions of important diabetes-related characteristics are shown in the Figure 1. The distribution of these traits throughout the dataset is depicted in each subplot, offering information on their individual densities and possible connections to diabetes prediction.

The associations between numerical characteristics in a diabetic dataset are depicted by the heatmap and correlation matrix Figure 2. Glucose had the strongest positive connection (0.47) with the outcome (diagnosis of diabetes). This suggests that in the sample, higher glucose levels are generally linked to a higher risk of diabetes. Other variables that are moderately associated include age (0.24) and BMI (0.29), indicating that these factors also have a considerable impact on diabetes outcomes.

3.2 Dataset Preprocessing:

Handling Zero Values:

Zero values were recognized as possibly inaccurate or missing data for BMI, SkinThickness, Insulin, BloodPressure, and Glucose. NaN (Not a Number) was used in their place to differentiate them from actual zero readings.

Missing Data Imputation:

A stratified median technique was used to impute missing values, which were originally zeros. The median value for each characteristic was determined independently for the groups with and without diabetes, and missing values were filled up appropriately. This approach accounts for the disparities in distributions across groups with and without diabetes, while maintaining the statistical qualities of the data.

Outlier Detection and Treatment:

Using the Interquartile Range (IQR) approach, outliers were found. Outliers were defined as data points for each characteristic that fell outside of Q3 + 1.5*IQR or below Q1 - 1.5*IQR. To lessen their effect on training the model while maintaining the general distribution of the data, these outliers were swapped out for the corresponding feature's median value.

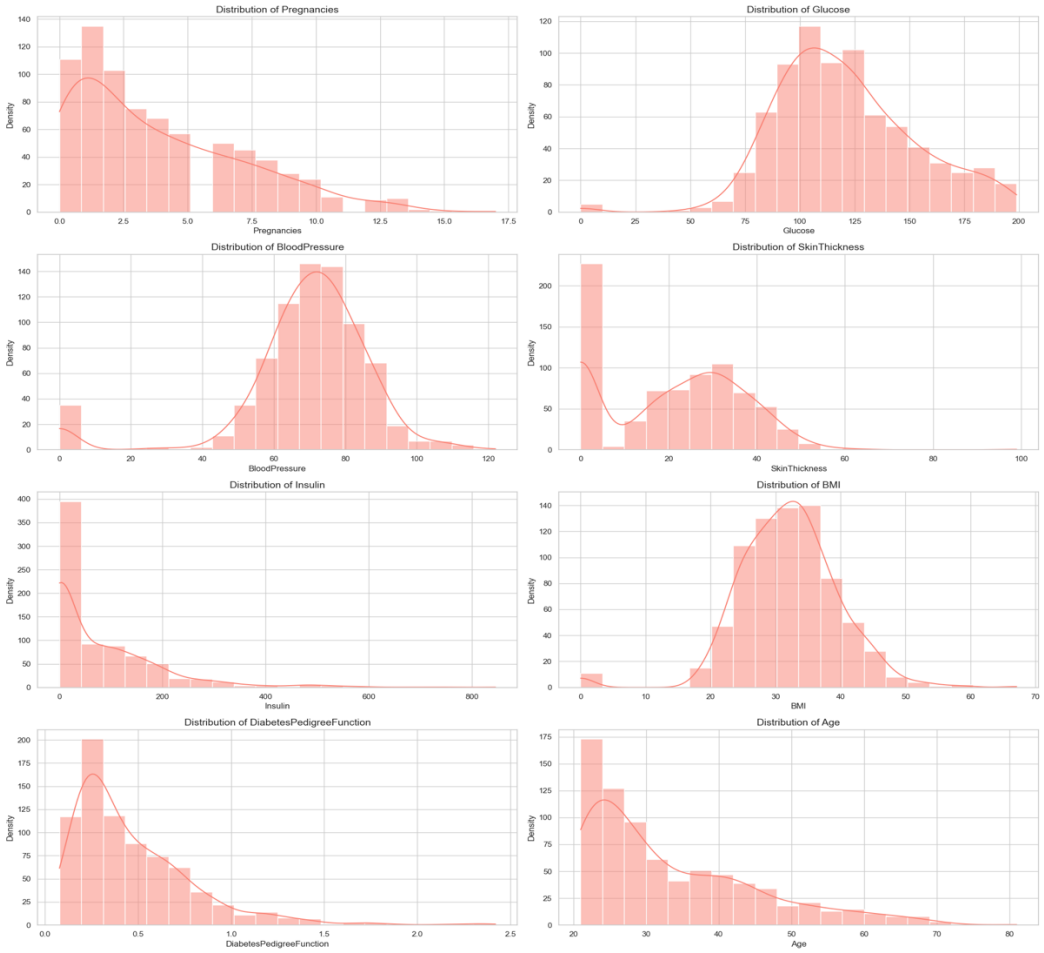

Figure 1. Density distribution of features

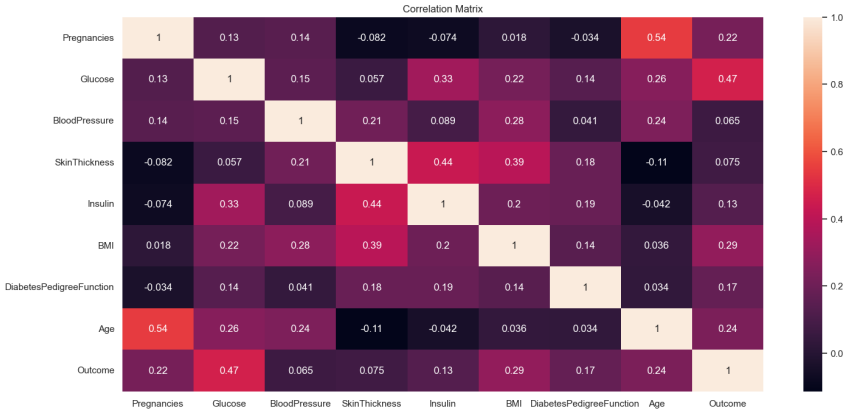

Figure 2. Correlation Matrix.

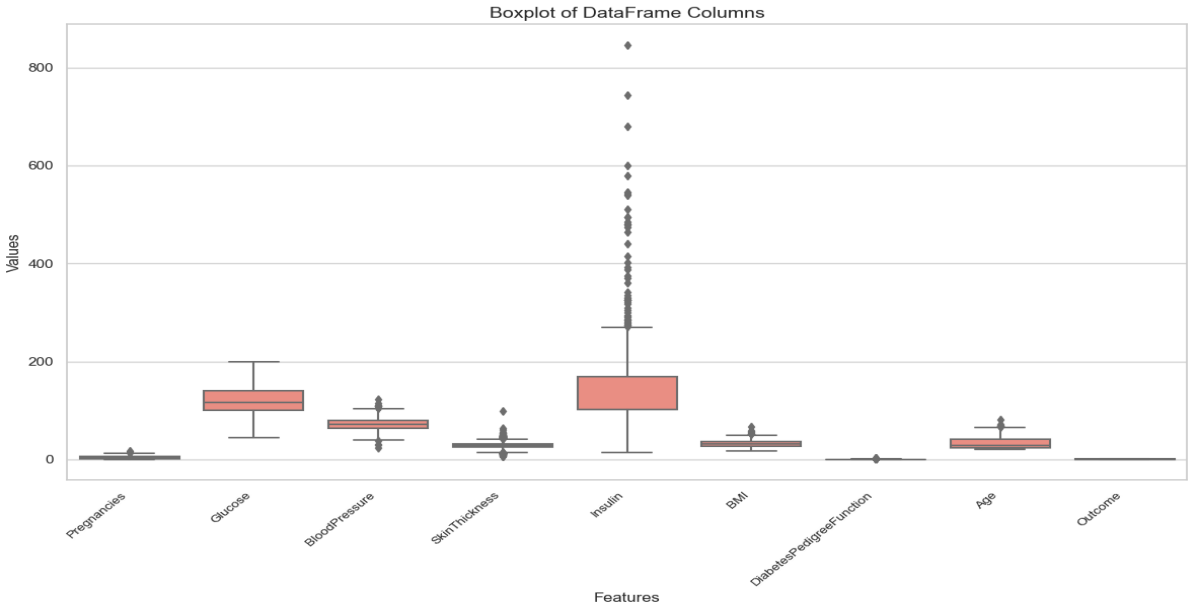

Figure 3. Outliers before handling.

Points outside the boxplots' whiskers denoted various features with outliers that were present before outlier management was applied Figure 3. The modified boxplot shows reduced variability and normalizing of the feature distributions after replacing the outliers with median values Figure 4. The central tendency and distribution of the dataset's numerical properties are more clearly visualized as a consequence of this technique, which also serves to lessen the impact of extreme numbers.

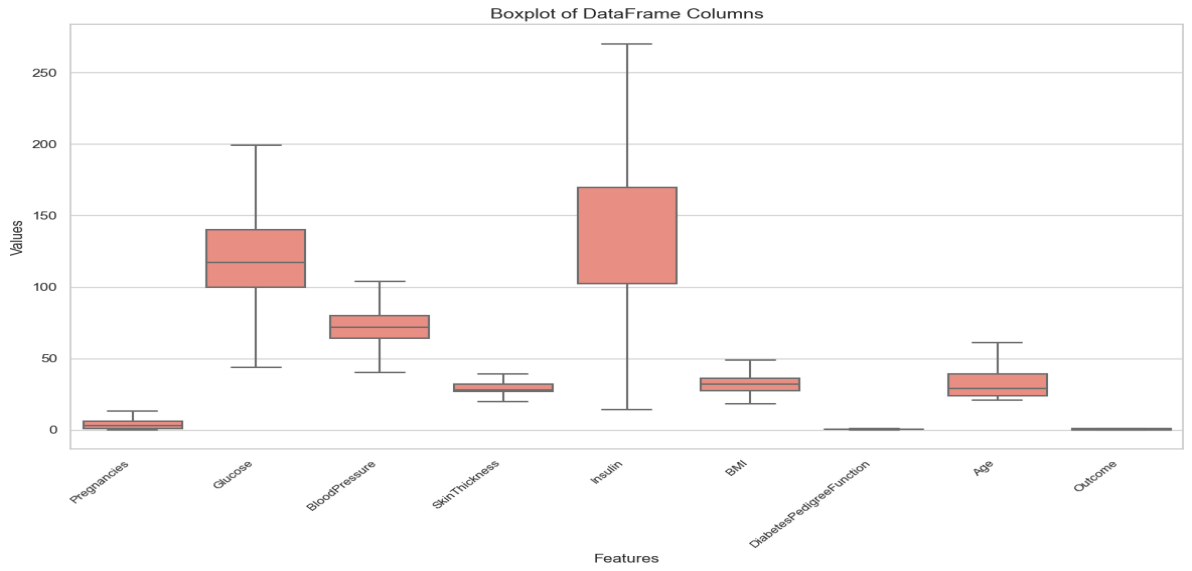

Figure 4. Outliers after handling.

3.3 Feature Engineering:

In order to possibly enhance the performance of the model, we designed further features:

BMI Categorization:

Based on the following clinically relevant BMI ranges [11], a new category characteristic called "UpdatedBMI" was developed:

- Severely underweight: < 16.5
- Underweight: 16.5 - 18.5

- Normal: 18.5 - 24.9
- Overweight: 24.9 - 29.9
- Obesity Class 1: 29.9 - 34.9
- Obesity Class 2: 34.9 - 39.9
- Obesity Class 3: > 39.9

Insulin Level Categorization:
The 'NewInsulinScore' was developed and classifies insulin levels [12] as follows:

- Normal: 16 - 166 μU/mL
- Abnormal: < 16 or > 166 μU/mL

Glucose Level Categorization:
With the launch of "NewGlucose," glucose levels were categorized [13] as follows:

- Low: ≤ 70 mg/dL
- Normal: 71 - 99 mg/dL
- Prediabetes: 100 - 125 mg/dL
- Diabetes: ≥ 126 mg/dL

One-Hot Encoding:

To prepare the newly formed categorical variables for machine learning methods, they underwent one-hot encoding. By creating binary columns for every category, this procedure enables the models to efficiently comprehend categorical data.

RobustScaler, available from scikit-learn, was used to scale numerical features. This scaler scales characteristics using the interquartile range and employs statistics that are resistant to outliers. This method guarantees that no feature is unduly impacted by outliers and that all features are on a comparable scale.

Data Splitting:

After loading, the dataset underwent preprocessing. The data was divided into training and testing sets, with 80% of the data utilized for training and 20% for testing. Missing values were handled accordingly. To make sure that every feature made an equal contribution to the model training process, feature scaling was used.

3.4 Model Selection and Training:

The following machine learning algorithms were implemented and evaluated: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, and XGBoost. Each model was trained on the training set and evaluated on the testing set using various performance metrics.

1. Logistic Regression: This algorithm was chosen for its simplicity and interpretability in binary classification problems.
2. K-Nearest Neighbors (KNN): Selected for its instance-based learning capabilities.
3. Support Vector Machine (SVM): Known for its effectiveness in high-dimensional spaces.
4. Decision Tree Classifier: Implemented for its simplicity and ability to model non-linear relationships.
5. Random Forest Classifier: Chosen for its ensemble learning capabilities.
6. Gradient Boosting Classifier: Implemented for its ability to build strong models from weak learners.
7. XGBoost: Selected for its speed and performance.

Some model underwent hyperparameter tuning using GridSearchCV with 10-fold cross-validation. This process involves systematically working through multiple combinations of parameter tunes, cross-validating as it goes to determine which tune gives the best performance. For example:

- SVM: Tuned parameters included 'C' (penalty parameter) and 'gamma' (kernel coefficient).
- Random Forest: Tuned parameters included 'n_estimators' (number of trees), 'max_depth', 'min_samples_split', and 'min_samples_leaf'.
- Gradient Boosting: Tuned parameters included 'learning_rate', 'n_estimators', and 'max_depth'.

3.5 Performance Metrics and Evaluation:

To address the research question, we employed multiple performance metrics:

Accuracy:

Measures the overall correctness of the model, calculated as (TP + TN) / (TP + TN + FP + FN), where TP is True Positives, TN is True Negatives, FP is False Positives, and FN is False Negatives.

Precision:

Measures the accuracy of positive predictions, calculated as TP / (TP + FP). It's particularly important when the cost of false positives is high.

Recall:

Measures the ability to find all positive instances, calculated as TP / (TP + FN). It's crucial when the cost of false negatives is high, which is often the case in medical diagnoses.

F1-score:

The harmonic mean of precision and recall, providing a single score that balances both metrics. It's calculated as 2 * (Precision * Recall) / (Precision + Recall).

AUC-ROC:

Region under the Receiver Operating Characteristic curve, which offers a cumulative performance indicator over all potential categorization cutoffs. It illustrates the model's class discrimination capability and is particularly helpful in the case of unbalanced datasets.

The performance of all models was compared using the above metrics. Visualizations, including ROC curves and bar charts comparing accuracy and AUC-ROC scores, were created to facilitate easy comparison.

This comprehensive methodology ensures a thorough evaluation of each model's ability to predict diabetes in Pima Indian women, addressing the research question by providing detailed performance metrics for comparison. The approach allows for an in-depth understanding of each model's strengths and weaknesses in this specific prediction task.

4. Results and discussions

4.1 Performance Evaluation:

The research assessed seven distinct machine learning algorithms for the purpose of predicting diabetes in female Pima Indian patients who were 21 years of age or older. Several performance criteria, including as accuracy, precision, recall, F1-score, and AUC-ROC (Table 1), were used to evaluate each algorithm. Below is a thorough examination of the results:

Models	Accuracy	ROC AUC	Precision		Recall		F1-Score	
			Class 0	Class 1	Class 0	Class 1	Class 0	Class 1
Logistic Regression	0.9026	0.8941	0.94	0.82	0.92	0.87	0.93	0.85
K-Nearest Neighbors (KNN)	0.8377	0.8116	0.89	0.73	0.88	0.74	0.88	0.74
Support Vector Machine (SVM)	0.9091	0.8928	0.93	0.85	0.93	0.85	0.93	0.85
Decision Tree Classifier	0.8766	0.8222	0.88	0.69	0.85	0.74	0.87	0.71
Random Forest Classifier	0.8961	0.8881	0.93	0.82	0.92	0.85	0.92	0.83
Gradient Boosting Classifier	0.8831	0.8801	0.94	0.77	0.89	0.87	0.91	0.82
XGBoost	0.8896	0.8669	0.92	0.83	0.93	0.81	0.92	0.82

Table 1. Performance Metrics of Various Classification Models.

Model comparison and suitability:

The effectiveness of seven machine learning models for diabetes prediction is compared in this bar plot (Figure 5) with respect to accuracy and ROC AUC values. A clear visual comparison of the efficacy of the models is provided by the plot, which shows that certain models, such LR and SVM, perform better than others, like KNN, in terms of accuracy and AUC.

The ROC curves for seven distinct machine learning models—whose area under the curve (AUC) represents the models' performance—are shown in the figure 6. The models are used to predict diabetes. Greater discriminating capacity in differentiating between instances with and without diabetes is suggested by higher AUC values.

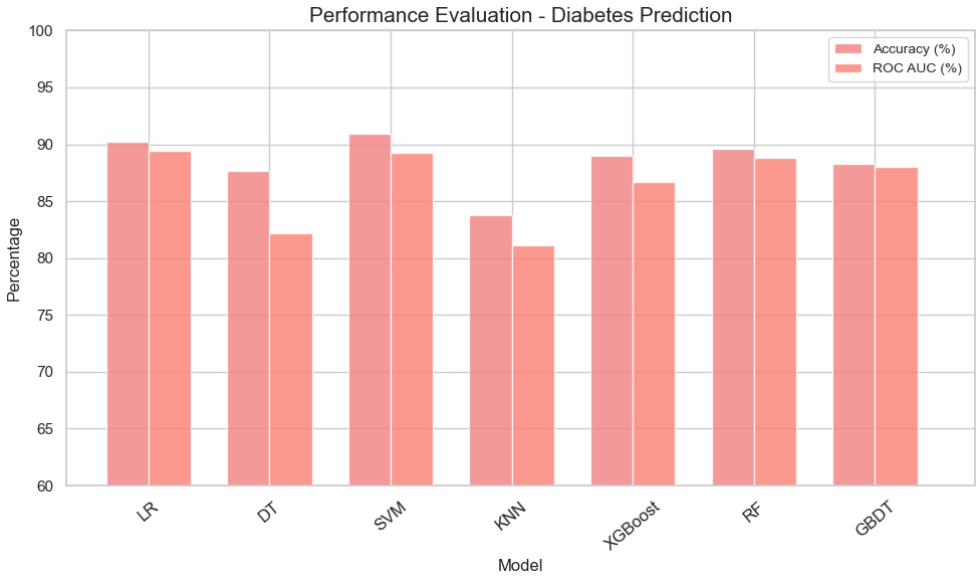

Figure 5. Accuracy and ROC AUC Comparison for Classification Models in Diabetes Prediction.

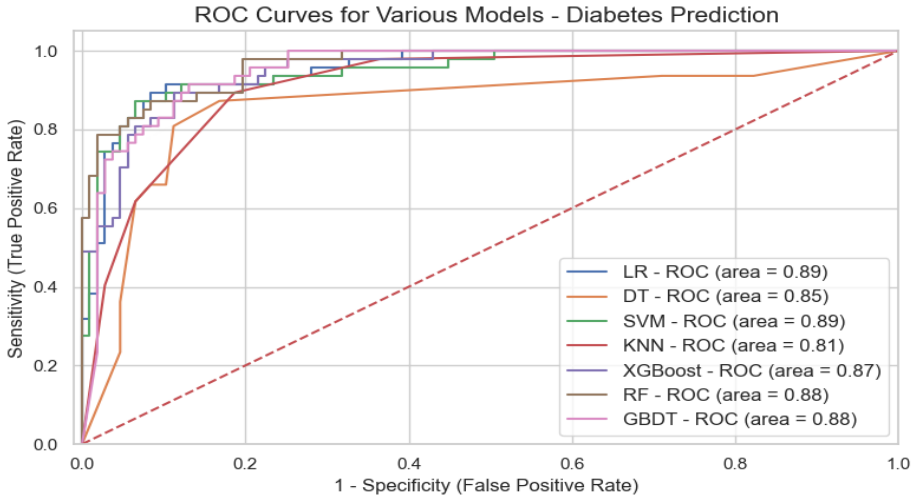

Figure 6. Diabetes Prediction Model ROC Comparison.

1, logistic regression:

- Accuracy: 90.26% of the patients were correctly classified as either diabetes or non-diabetic, demonstrating good overall accuracy.
- Precision: 0.94 indicates a high degree of accuracy in positive predictions for the classification of diabetes patients.
- Recall: 0.82, indicating a significant percentage of true positive instances were successfully identified.
- The harmonic mean of recall and accuracy, or F1-Score, is 0.87, indicating balanced performance.
- AUC-ROC: At 89.41%, the highest of all models, it shows a superior capacity to distinguish between patients with and without diabetes over a range of thresholds.

With the greatest AUC-ROC of 89.41% and an accuracy of 90.26%, logistic regression trailed SVM by a narrow margin. This suggests that Logistic Regression is a useful tool for detecting positive cases and performs well in differentiating between patients with and without diabetes. It also performed well in detecting genuine positive instances, matching the Gradient Boosting Classifier for the maximum recall (0.87) for the positive class.

2. Neighbors K-Nearest (KNN):

- 83.77% accuracy, which is less than that of logistic regression.
- Precision: 0.89, which denotes positive forecast accuracy.
- Recall: 0.73, which is less than ideal for uses in medicine.
- A reasonable overall performance is indicated by the F1-Score of 0.74.
- AUC-ROC: 81.16%, which is lower than other models but indicates a respectable level of discriminative skill.

3. SVM, or Support Vector Machine:

- 90.91% accuracy, the highest of all the models.
- Precision: 0.93, which denotes a high degree of positive prediction accuracy.
- Recall: 0.85, indicating resilience in accurately detecting positive instances.
- F1-Score: 0.85, which represents balanced recall and accuracy performance.
- AUC-ROC: 89.28%, indicating a high degree of discrimination.

The most accurate classifier overall was Support Vector Machine (SVM), which achieved 90.91% accuracy and balanced performance across precision, recall, and F1-score (all 0.85). The way SVM performs highlights how well it can handle high-dimensional data and how reliable it is for classification jobs. This suggests that SVM worked well in accurately classifying patients as either diabetes or non-diabetic, with little bias in either direction.

4. Classifier Using Decision Trees:

- 87.66% accuracy, which is less than that of SVM and logistic regression.
- Precision: 0.88, which denotes positive forecast accuracy.
- Recall: 0.69, which is less than ideal for applications in medicine.
- F1-Score: 0.71, which denotes a mediocre performance all around.
- AUC-ROC: 82.22%, which indicates a moderate level of discrimination ability.

The performance of the Decision Tree and K-Nearest Neighbors (KNN) models was lower than that of other techniques, indicating their limited applicability to this particular prediction job in the population under consideration.

5. Classifier using Random Forest:

- 89.61% accuracy rate demonstrates good overall correctness.
- Precision: 0.93, which denotes a high degree of positive prediction accuracy.
- Recall: 0.82, indicating that a sizable percentage of true positive instances may be identified.
- F1-Score: 0.83, indicating a performance that is balanced.
- AUC-ROC: 88.81%, which shows strong discriminatory power.

Furthermore, Random Forest proved to be a dependable option for ensemble learning in this prediction challenge, exhibiting good performance with an accuracy of 89.61% and balanced metrics across precision, recall, and F1-score. The ensemble approach showed a decent balance across metrics, indicating that it can handle the dataset's complexity.

6. Classifier using Gradient Boosting:

- Accuracy: 88.31%, little less than that of logistic regression and SVM.
- Precision: 0.94, which denotes a high degree of positive prediction accuracy.
- Recall: 0.77, which shows how well it can detect affirmative situations.
- F1-Score: 0.82, indicating a performance that is balanced.
- AUC-ROC: 88.01%, indicating a high degree of discrimination skill.

The ability of the Gradient Boosting Classifier to accurately identify genuine positive instances was demonstrated by its noteworthy strength in recall (0.87), which was tied with Logistic Regression. This demonstrates how effective it is in reducing false negative results, which is especially important in a medical setting where failing to diagnose diabetes might have grave repercussions.

7. XGBoost:

- Accuracy: 88.96% demonstrates a great degree of overall accuracy.
- Precision: 0.92, which denotes positive forecast accuracy.
- Recall: 0.83, indicating that it can accurately detect affirmative situations.
- F1-Score: 0.82, indicating a performance that is balanced.
- AUC-ROC: 86.69%, which suggests a high degree of discrimination.

Recommendations:

SVM proved to be the most effective classifier overall, having the greatest accuracy and the most balanced precision, recall, and F1-score metrics. Its solid performance and efficiency in handling high-dimensional data make it appropriate for diabetes prediction in this particular group.

SVM outperformed logistic regression in terms of AUC-ROC, demonstrating a higher capacity to discriminate between patients with and without diabetes at various thresholds. It was particularly good at finding positive instances, which is important in a medical environment.

Classifiers using Random Forest and Gradient Boosting also performed well, providing accurate predictions and well-balanced metrics. Their capacity for group learning makes them reliable options for intricate prediction jobs.

4.2 Discussion:

The findings show that although SVM offers the best overall accuracy, the particular priorities in the medical setting should be taken into account when selecting a model. For example:

SVM is advised in order to maximize total accuracy and offer strong results for every metric.

Logistic regression works especially well in situations where it's important to discern between groups and find positive examples.

With its capacity for ensemble learning, Random Forest provides a potent substitute that guarantees accurate predictions with well-balanced metrics.

Gradient Boosting Classifier performs exceptionally well in reducing false negatives, which is essential for precise diagnosis in applications related to medicine.

4.3 Implications for Healthcare and Future Research

Results from earlier research on diabetes prediction are consistent with the robust performance of SVM, Logistic Regression, and ensemble techniques like Random Forest and Gradient Boosting. For example, SVM accuracies reported by Al-Sideiri et al. (2019) ranged from 84% to 94%, which is in line with our results. As earlier diabetes prediction studies have shown, logistic regression performs well in this kind of classification job, which is further supported by our data.

For academics and medical professionals interested in diabetes prediction in female Pima Indian patients, these findings offer insightful information. It's crucial to remember, though, that the particular group under study may have limited how broadly these findings may be applied. Subsequent research endeavors may encompass the validation of these models on heterogeneous populations, investigation of supplementary feature engineering methodologies, and utilization of more intricate models to possibly augment prediction precision. Global improvements in diabetes diagnosis and management approaches would result from the models' ongoing amplification and validation on a variety of populations.

To sum up, this study shows that machine learning techniques, specifically Support Vector Machines and Logistic Regression, are effective in predicting diabetes in female Pima Indian patients. The findings highlight how these methods may help with early diabetes diagnosis and treatment, which may have a substantial effect on patient outcomes in this high-risk group.

5. Conclusions and future work

Building upon a collection of previous research, this study assessed many machine learning algorithms for predicting diabetes in female Pima Indian patients aged 21 and above. The outcomes show how useful machine learning methods are for this kind of work, especially Support Vector Machines (SVM) and Logistic Regression. At 90.91%, SVM had the highest overall accuracy and the best balance of precision, recall, and F1-score. This is in line with earlier research, which has SVM accuracies ranging from 84% to 94% (Al-Sideiri et al., 2019). With the greatest AUC-ROC score of 89.41% and a closely followed accuracy of 90.26%, logistic regression demonstrated better discriminative capacity. These outcomes are in line with other research that shown the use of logistic regression in diabetes prediction tasks.

Our study's excellent results for ensemble techniques like Random Forest and Gradient Boosting Classifier support earlier findings. For example, Chang et al. discovered that Random Forest, with an accuracy rate of 79.57%, was the best-performing algorithm in their electronic diagnosis system. Our results build on previous work by showing that ensemble approaches in this particular population achieve even greater accuracy rates.

The focus of our work on several assessment measures, including as recall, accuracy, precision, F1-score, and AUC-ROC, is in line with current developments in the area. Studies have progressively included these extensive criteria, as the literature review emphasizes, to offer a more complete evaluation of model performance, especially when working with unbalanced datasets.

The study had a number of shortcomings in spite of these encouraging outcomes. The findings may not be as broadly applicable to other groups due to the dataset's restriction to female Pima Indian patients who are 21 years of age or older. Given that the dataset focuses on a population with a significant genetic vulnerability to insulin resistance—as noted in the literature review—this restriction is especially notable. The performance of the models may not be as robust due to the comparatively small sample size of 768 data. Furthermore, bias may have been induced by potential problems with the quality of the data, such as suspicious zero values that needed to be imputationed. This is consistent with earlier research' concerns of missing or erroneous data in the Pima Indian Diabetes Dataset.

Potential avenues for future study to overcome these constraints and develop the area might be explored. By testing these models on increasingly varied populations, we can evaluate the models' wider applicability and increase population variety. This is especially crucial because the dataset focuses on a certain population. Prediction accuracy may be increased by incorporating data from other sources, such as genetic information or lifestyle variables. This is consistent with earlier research identifying critical variables for the categorization of diabetes, such as age, skin fold thickness, insulin, body mass index, and glucose levels.

Researching more sophisticated approaches, such as deep learning techniques, may lead to even greater gains in prediction performance, as shown by Naz and Ahuja's (2020) work, which used deep learning techniques to achieve 98.07% accuracy. Longitudinal studies that monitor patients over an extended period of time may shed light on how well the models predict the onset and course of diabetes. This may expand on current studies investigating the possibility of combining various algorithms with real-time data collecting to forecast diabetes more thoroughly.

Future research may expand on this work to further enhance diabetes prediction and treatment, thereby improving health outcomes for at-risk groups. This can be achieved by resolving these constraints and pursuing these research objectives. When seen in the context of the larger literature, the encouraging findings of this study highlight the importance of machine learning in healthcare and open the door to further developments in early illness identification and tailored medication.

References

[1] Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2022). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing & Applications*, 1–17. Advance online publication. https://doi.org/10.1007/s00521-022-07049-z

[2] Miao, Y. (2021). Using machine learning algorithms to predict diabetes mellitus based on PIMA Indians diabetes dataset. In *Proceedings of the 2021 5th International Conference on Virtual and Augmented Reality Simulations (ICVARS '21)* (pp. 47–53). Association for Computing Machinery. https://doi.org/10.1145/3463914.3463922

[3] Daanouni, O., Cherradi, B., & Tmiri, A. (2019). Predicting diabetes diseases using mixed data and supervised machine learning algorithms. In *Proceedings of the 4th International Conference on Smart City Applications (SCA '19)* (Article 85, pp. 1–6). Association for Computing Machinery. https://doi.org/10.1145/3368756.3369072

[4] Naz, H., & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes and Metabolic Disorders*, 19(1), 391–403. https://doi.org/10.1007/s40200-020-00520-5

[5] Vaishali, R., Sasikala, R., Ramasubbareddy, S., Remya, S., & Nalluri, S. (2017). Genetic algorithm based feature selection and MOE fuzzy classification algorithm on Pima Indians diabetes dataset. In *2017 International Conference on Computing Networking and Informatics (ICCNI)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICCNI.2017.8123815

[6] Saru, S., & Subashree, S. (2019). Analysis and prediction of diabetes using machine learning. *International Journal of Emerging Technology and Innovative Engineering*, 5(4).

[7] Mahmud, S. M. H., Hossin, M. A., Ahmed, M. R., Noori, S. R. H., & Sarkar, M. N. I. (2018). Machine learning based unified framework for diabetes prediction. In *Proceedings of the 2018 International Conference on Big Data Engineering and Technology (BDET 2018)* (pp. 46–50). Association for Computing Machinery. https://doi.org/10.1145/3297730.3297737

[8] Mishra, S., Chaudhury, P., Mishra, B. K., & Tripathy, H. K. (2016). An implementation of feature ranking using machine learning techniques for diabetes disease prediction. In *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies (ICTCS '16)* (Article 42, pp. 1–3). Association for Computing Machinery. https://doi.org/10.1145/2905055.2905100

[9] Al-Sideiri, A., Che Cob, Z. B., & Drus, S. B. M. (2020). Machine learning algorithms for diabetes prediction: A review paper. In *Proceedings of the 2019 International Conference on Artificial Intelligence, Robotics and Control (AIRC '19)* (pp. 27–32). Association for Computing Machinery. https://doi.org/10.1145/3388218.3388231

[10] Zafar, F., Raza, S., Khalid, M. U., & Tahir, M. A. (2019). Predictive analytics in healthcare for diabetes prediction. In *Proceedings of the 2019 9th International Conference on Biomedical Engineering and Technology (ICBET '19)* (pp. 253–259). Association for Computing Machinery. https://doi.org/10.1145/3326172.3326213

[11] Weir, C. B., & Jan, A. (2019). BMI classification percentile and cut off points. In **StatPearls**. StatPearls Publishing. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK541070/

[12] Uttekar, P. S., & Allarakha, S. (n.d.). What is a high insulin level? MedicineNet. Retrieved from https://www.medicinenet.com/what_is_a_high_insulin_level/article.htm

[13] Modglin, L. (n.d.). Normal blood sugar levels by age (chart). Forbes. Retrieved from https://www.forbes.com/health/wellness/normal-blood-sugar-levels/