**Comparative Analysis of Classification Algorithms for Diabetes Detection**

Dr.Siddique Ibrahim ,a)V.Krishna Chiatanya,b)N.Bhuvanesh,c)K.V.Pavan Kumar,d)A.V.N.S.S .Samanvi,e)J.Yaswanth

*School of Computer Science and Engineering VIT-AP University, Amaravati, Andhra Pradesh, India*

1. Dr. Siddique Ibrahim S : [siddique.ibrahim@vitap.ac.in](mailto:siddique.ibrahim@vitap.ac.in)
2. krishnavejendla34@gmail.com
3. bhuvaneshnadellal7@gmail.com d)vpkr2004@gmail.com e)samanvikrishna1999@gmail.com

f)jaladiyaswanth2005@gmail.com

Abstract--Diabetes diagnosis has been essentially upgraded by the application of classification mining strategies. This ponder investigates the adequacy of different machine learning findings, such as Choice Trees, Support Vector Machines (SVM), and Neural Systems, in foreseeing diabetes onset based on understanding information. Utilizing datasets from sources like the PIMA Indian Diabetes dataset, we assess the execution of these classifiers in terms of exactness, exactness, review, and F1-score. Our discoveries illustrate that whereas all strategies offer important bits of knowledge, certain calculations like SVM and Neural Systems display prevalent prescient execution. This inquire about underscores the potential of classification mining to progress demonstrative exactness and bolsters its integration into clinical hone for early and solid diabetes location. Random Forest Algorithm has highest accuracy of 96.07% among all models. Cross Validation is also performed ,the box plot(Fig-4)shows the accuracy scores on each model .[1]

Keywords-- Diabetes Prediction, Machine Learning Classification, Supervised Learning, Logistic Regression, Support Vector Machine , Random Forest , Gradient Boosting, Diabetes Diagnosis, Feature Selection, Data Preprocessing, Cross-Validation, Model Evaluation Metrics, Accuracy, Precision, Recall,F1-Score, ROC Curve, Confusion Matrix, imbalanced dataset, Health Informatics [1] [4] [5] [6] [8] [9] [15]

# **INTRODUCTION**

Diabetes mellitus is a long-term metabolic disorder that causes high blood sugar levels because the body either does not make enough insulin or does not use it properly. It is a major global health concern, affecting millions of people and resulting in serious complications such as cardiovascular disease, renal damage, and neurological problems.Early detection and treatment are critical for managing diabetes and preventing complications. Traditional diagnostic approaches, on the other hand, are often invasive, time-consuming, and costly. As a result, classification mining approaches are gaining popularity as a more accurate way to diagnose diabetes. These approaches provide a more efficient, accurate, and non-invasive approach to detecting the condition.

Classification mining, a branch of data mining, involves the analysis of large datasets to discover meaningful patterns and relationships. By applying various machine learning algorithms such as Decision Trees, Naïve Bayes and Random Forest, researchers can develop models that classify patients as diabetic or non-diabetic based on their medical data. These models learn from historical datasets, where known instances of diabetes and non-diabetes are labeled, to predict outcomes for new data points. The accuracy and effectiveness of these models depend on various factors, including the quality of the dataset, the preprocessing techniques applied, and the algorithm used. [3] [9][8]

Recent advances in processing power and machine learning approaches have enabled the creation of more complex categorization models. These models not only increase accuracy but also provide information about the major risk factors for diabetes. Furthermore, feature selection approaches such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) are often utilized to improve classification model performance by lowering data dimensionality. [5]

This study aims to investigate and evaluate multiple classification mining approaches for diabetes diagnosis. By comparing the performance of various algorithms and reviewing their strengths and limits, the study aims to lead to the creation of more accurate and efficient diagnostic tools, which could potentially transform diabetes early detection and management.

This study aims to examine and assess multiple classification mining approaches for diabetes diagnosis. By comparing the performance different algorithms and reviewing their strengths and limits, the study seeks to contribute to the creation of more accurate and efficient diagnostic tools, which could potentially transform diabetes early detection and management.

# **OBJECTIVES**

* Compare machine learning strategies for predicting diabetes using specifications such as accuracy, precision, recall, and F1-scores.
* To investigate how feature selection and data pre-processing affect the performance of machine learning models for diabetes detection.
* Evaluating cross-validation and hyperparameter changes to enhance the accuracy of models for forecasting diabetes.

# **ACRONYMS**

* DM: Diabetes Mellitus
* ML: Machine Learning
* SVM: Support Vector Machine
* RF: Random Forest
* PCA: Principal Component Analysis
* ROC: Receiver Operating Characteristic
* F1: F1 Score

# **OVERVIEW**

Diabetes mellitus (DM) is a prevalent chronic health condition characterized by elevated sugar levels in the bloodstream due to inadequate insulin production or ineffective insulin utilization. Early detection of diabetes is crucial for effective management and prevention of serious complications. While traditional diagnostic methods may be time-consuming, data-driven techniques, especially classification mining approaches, provide quicker and more precise alternatives.

Classification mining entails using machine learning (ML) algorithms to predict whether a patient has diabetes based on clinical data such as glucose levels, BMI, age, and family history. Common algorithms include Decision Trees (DT), Support Vector Machines (SVM), Naïve Bayes (NB) and Artificial Neural Networks (ANN). These models are evaluated using metrics such as accuracy, precision, recall, and F1-score to gauge their diagnostic efficacy.

Advanced techniques, such as Random Forest (RF) and ensemble methods like XGBoost, can improve diagnostic accuracy by combining multiple classifiers. Data preprocessing, including normalization and feature selection methods like Principal Component Analysis (PCA), further enhances model effectiveness. [11] [5]

The aim of this research is to explore and compare various classification techniques for diabetes diagnosis. The study seeks to develop a robust, automated model that offers improved accuracy and efficiency in clinical settings, aiding in the early detection and management of diabetes. [1]

**DIABETES TYPES**  
1. Type 1 Diabetes: An immune system disorder in which the body is unable to manufacture insulin. It typically affects kids and young adults and necessitates ongoing insulin therapy.   
2. Type 2 Diabetes: The most prevalent type, where the body becomes resistant to insulin or doesn't create enough. It can be controlled with food, exercise, and medication and is frequently linked to obesity and lifestyle factors.   
3. Gestational diabetes: Usually disappears after delivery and develops during pregnancy, although it raises the chance of developing Type 2 diabetes in the future.   
4. MODY: A uncommon genetic type of diabetes caused by a gene mutation that affects younger people.

5.Secondary Diabetes: Diabetes caused by another condition or medication, like pancreatitis or long-term steroid use.

Most Important Type: Type 2 Diabetes

Why It's Most Important: Ninety to ninety-five percent of diabetes cases worldwide are type 2 diabetes, which is the most common type of the disease. Particularly in industrialized and emerging countries, Type 2 diabetes is becoming a major public health concern due to the rise in obesity, poor food, and sedentary lifestyles. It has a huge influence on the world's health systems because of the serious morbidity and mortality that result from complications like neuropathy, kidney failure, and cardiovascular disease. Nowadays, a primary emphasis of diabetes research and public health initiatives is the effective management of Type 2 diabetes, which is essential in preventing severe consequences.

# **SYMPTOMS, DIAGNOSIS AND TREATMENT**

Depending on the kind, diabetes can cause different symptoms. (Type 1 or Type 2) and the severity, but common signs include:

1. Greater thirst

2 recurring urination

3. Unexplained weight loss

4. Increased hunger

5. Fatigue or feeling tired easily

6. hazy vision

7. Sluggish healing of lesions or wounds

8.Frequent infections, such as gum, skin, or infections in the vagina

9.Numbness or tingling in hands and feet

10.Dry skin and frequent itching

Diabetes can cause major side effects like heart disease, renal damage, and nerve damage if it is not addressed,so recognizing these symptoms early is important for timely diagnosis and treatment.

The most frequent symptoms of diabetes include:

1. Greater thirst

2. recurring urination

3.Increased hunger

4.Fatigue or tiredness

5.Unexplained weight loss

6. hazy vision

These are the classic symptoms of both Type 1 and Type 2 diabetes that are often seen, and they are frequently the initial indications that lead people to seek medical assistance.   
For various kinds of diabetes, the following methods comprise the most frequent diabetic therapy.

Type 1 Diabetes: Insulin therapy and continuous blood sugar monitoring.

Type 2 Diabetes: Lifestyle changes like dieting, exercise, oral medications recommended by the doctor, and sometimes using insulin.

Gestational Diabetes: Diet, exercise, and occasionally insulin during pregnancy care mandatory

All the above-mentioned types require regular blood sugar monitoring.

# **RELATED WORK**

he determination of diabetes utilizing prescient models has been a noticeable investigate range for over a decade. The tremendous lion's share of ponders centre on leveraging clustering calculations and fake neural systems (ANNs) to progress symptomatic precision. In any case, there has been a persistent thrust toward creating more modern, quicker, and user-friendly models.  
  
One think about utilized three techniques—EM calculation, H-means+ clustering, and Hereditary Calculation (GA)—for classifying diabetic patients. H-means+ clustering beated other approaches when gathering comparative indications into clusters, demonstrating successful for identifying inconspicuous designs in quiet data.  
  
Fuzzy Insect Colony Optimization (ACO) has too been utilized for diabetes determination, particularly utilizing the Pima Indian Diabetes dataset. This method was fruitful in producing a set of rules that may offer assistance recognize diabetic patients with higher precision.  
  
  
Another neural network-based ponder pointed to anticipate diabetes by analysing 13 early side effects, illustrating the esteem of combining progressed calculations with real-time persistent information. The usage in MATLAB demonstrated fruitful in making a demonstrate that may distinguish diabetes early on with a tall degree of accuracy.  
  
In expansion to these approaches, other machine learning procedures like Random forest, and Calculated Relapse have moreover been connected to diabetes determination. These strategies, whereas compelling, regularly require seriously parameter tuning and long preparing times, which can constrain their commonsense utilize in real-time therapeutic settings.  
  
Despite the advance made in these regions, there remains a crevice in creating models that combine probability-based classification with highlight determination to upgrade both exactness and computational effectiveness. Earlier thinks about have utilized complex, time-consuming methods, regularly utilizing weighted approaches that require broad handling control and are not appropriate for real-time diagnosis.

Furthermore, this show has the potential to be scaled universally, with future investigate pointing to consolidate datasets from assorted populaces. As unused information gets to be accessible, particularly from diverse geographic districts, the precision and generalization of the prescient show may be improved, making it a more vigorous apparatus for worldwide healthcare frameworks. Furthermore, future work may coordinate other advancing methods, such as profound learning, to encourage make strides expectation exactness and decrease conclusion time. [9]

# **METHODOLOGIES**

A)Dataset Description and Data Modelling

The Pima Indians dataset is a medical dataset used for predicting diabetes in women of Pima Indian heritage. It includes 768 samples with 8 features, such as the number of pregnancies, glucose levels, BMI, and age. The target variable indicates whether a patient has diabetes. The dataset is widely used for binary classification tasks in machine learning, particularly for testing algorithms like logistic regression and decision trees. It also includes some missing or invalid values, making it useful for data preprocessing exercises.

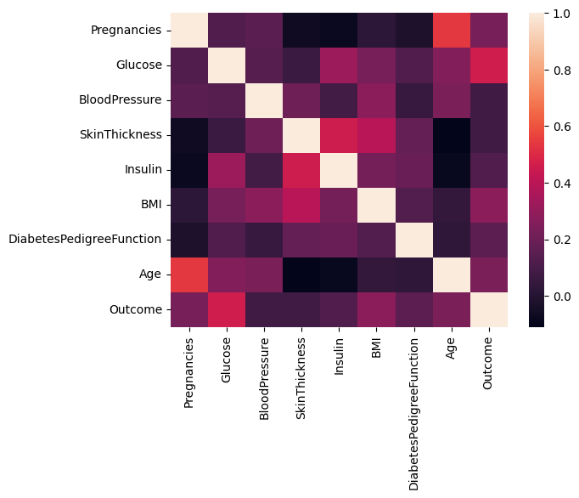
The study investigates the use of Random Forest, Gradient Boosting, and Decision Trees as data mining tools to identify diabetes. The major goal is to use data mining techniques and the available medical data to predict if the patient has been impacted by diabetes. The Pima Indians Diabetesdataset has undergone data mining using the classification type. Table – 1 gives a description of dataset we have taken into consideration. [4] [7]

|  |  |  |
| --- | --- | --- |
| Dataset | Number of Attributes | Number of Instances |
| Pima Indian Diabetes | 9 | 2000 |

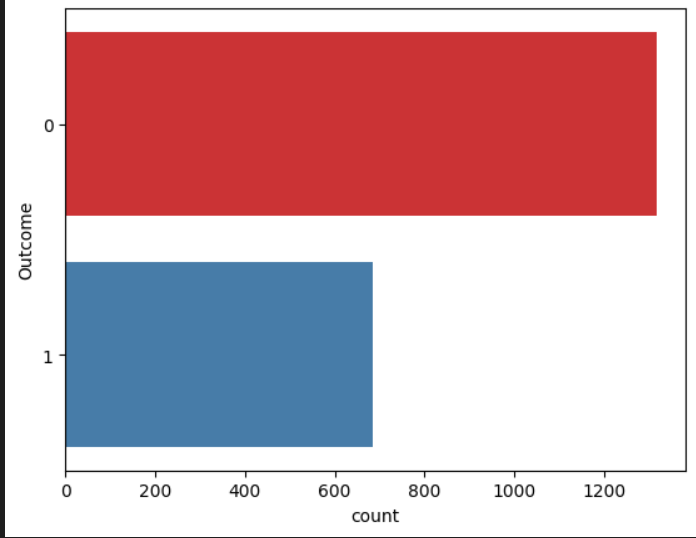
Table 1. Dataset Information and Description

Table 2. Description Of Each Attribute

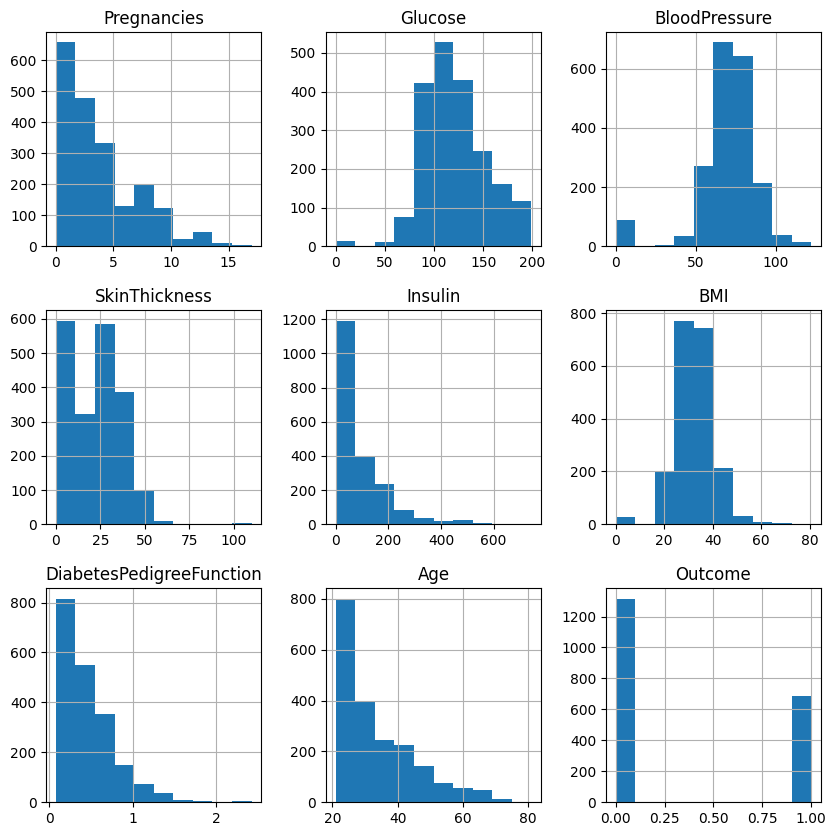
|  |  |  |
| --- | --- | --- |
| S.No | Attribute | Description[datatype] |
| 1 | Pregnancies | 2000 non – null int 64 |
| 2 | Glucose | 2000 non – null int 64 |
| 3 | Blood Pressure | 2000 non – null int 64 |
| 4 | Skin Thickness | 2000 non – null int 64 |
| 5 | Insulin | 2000 non – null int 64 |
| 6 | BMI | 2000 non – null float 64 |
| 7 | Diabetes Pedigree Function | 2000 non – null float64 |
| 8 | Age | 2000 non – null int 64 |
| 9 | Outcome | 2000 non – null int 64 |



*Fig-1: Co-relation between attributes*



*Fig-2: Class Attribute Detection (0-means tested -ve ,1-means tested +ve )*



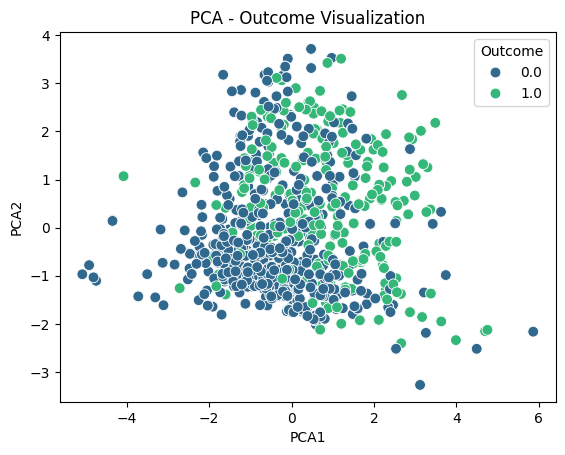
*Fig-3: Histograms of the different attributes*

B)Data Pre-Processing

Dataset is clean, consistent, and ready for analysis. In the context of diabetes detection, preprocessing begins by handling missing data, which is common in real-world datasets. This can be done by either removing records with missing values or imputing them with the mean, median, or mode of the available data. Feature scaling is then applied to standardize the range of numerical features like glucose levels and BMI, so that no single feature dominates the model. Normalization or standardization is typically used to bring all features to a common scale. Additionally, categorical variables, such as gender, need to be encoded into numerical form using methods like one-hot encoding or label encoding to make the data compatible with machine learning algorithms.

Outliers, or extreme values, are another issue that can distort the model’s performance. Detecting and removing outliers using methods like the interquartile range (IQR) or z-scores can prevent these anomalies from skewing the results. Finally, feature selection and dimensionality reduction techniques are applied to ensure the model focuses on the most relevant features. By eliminating irrelevant or redundant features, algorithms can perform more efficiently and with greater accuracy. Dimensionality reduction techniques, such as principal component analysis (PCA), can further simplify the dataset while retaining essential patterns. These preprocessing steps collectively enhance the quality of the data and ensure that the model produces reliable predictions in diabetes detection. [5]

The process involves handling missing data using mean imputation, scaling features with StandardScaler, and reducing dimensionality via PCA, ensuring the dataset is optimized for improved performance in machine learning models.



*Fig 4: Missing Data Handling*

C) Cross-Validation

Cross-validation is a model evaluation technique where data is split into multiple subsets (folds), and the model is trained and validated on different folds, improving performance reliability and preventing overfitting.

We used 10-fold cross-validation to evaluate the classifiers' performance. Cross-validation is a powerful model evaluation approach that is especially effective for minimizing overfitting and verifying that the model generalizes well to new data. StratifiedKFold was used to divide the dataset into ten equal subsets, ensuring that the percentage of diabetes and non-diabetic occurrences stays consistent in each fold. Each model, including Random Forest, Gradient Boosting, and Support Vector Machine, was trained on 9 folds and verified on the final fold. This process was done 10 times, with each fold serving as a test set once, allowing for a more trustworthy assessment of the model's accuracy.

StandardScaler was used to standardize the data for the SVM model, which is sensitive to feature magnitudes. After performing cross-validation, we calculated the mean accuracy and standard deviation to evaluate the model's overall performance and variability between folds. This strategy enabled us to derive a more generalized measure of performance while reducing the bias and variance that can arise from a single train-test split.

D) Hyper Parameter Tuning and its effect on Reliabilty

Hyperparameter tuning is an important part of machine learning model construction. Unlike model parameters, which are learned from training data, hyperparameters are established before the learning process and control how the model is taught. Hyperparameters include learning rates, regularization parameters, the number of trees in a random forest, and kernel functions in support vector machines. Optimizing these hyperparameters has the potential to greatly increase model performance.

Effective hyperparameter tweaking can increase a model's accuracy, precision, recall, and other performance measures by achieving the best balance between underfitting and overfitting. The tuning procedure improves model generalization to unknown data, enhancing reliability and robustness, particularly in real-world applications.

Methods for Hyperparameter Tuning

1.Grid Search: Grid search tests all potential hyperparameter combinations within a given range. Even though it ensures to discover the greatest combination, it can be computationally intensive.

2. Randomized Search: This method selects and tests random combinations of hyperparameters. While it may not test all possible outcomes, it can be more efficient in practice, particularly when the hyperparameter space is huge.

3.Bayesian Optimization: This method uses the relationship between hyperparameters and the evaluation metric to intelligently explore the hyperparameter space, making it more efficient than grid or randomized search.

Evaluation Metrics in Hyperparameter Tuning

The efficiency of hyperparameter tuning is frequently assessed using approaches such as cross-validation, which involves training and validating the model numerous times on different splits of the dataset. This ensures that the hyperparameters generalize effectively to new data and do not overfit a specific training set.

Example: Tuning Hyperparameters in SVM.

Hyperparameters for an SVM include the regularization parameter 'C', kernel type, and 'gamma' (in the case of an RBF kernel). The optimal mix of these hyperparameters will affect the SVM's decision boundary and, ultimately, classification results.

Hyperparameter adjustment is essential for optimizing machine learning models. By carefully selecting the appropriate strategy (grid search, random search, Bayesian optimization) and analyzing the results with cross-validation, one can considerably improve model dependability, making it suitable for deployment in real-world scenarios. Advanced visualization approaches aid in analyzing and fine-tuning these models, making the tuning process more informative and efficient.

Tuning Random Forest...

Best parameters for Random Forest: {'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 100}

Best accuracy for Random Forest: 0.9765

Tuning Gradient Boosting...

Best parameters for Gradient Boosting: {'learning\_rate': 0.2, 'max\_depth': 5, 'n\_estimators': 100, 'subsample': 0.8}

Best accuracy for Gradient Boosting: 0.9795

Tuning SVM...

Best parameters for SVM: {'C': 10, 'gamma': 'auto', 'kernel': 'rbf'}

Best accuracy for SVM: 0.8760000000000001

Reliability can be enhanced by Hyper Parameter Tuning. Hyperparameter tuning, such as Grid Search or Randomized Search, allows you to experiment with different hyperparameter configurations to find the best combination that yields the best performance (accuracy, precision, recall, and so on). This decreases the danger of underfitting and overfitting, making the model more dependable in a variety of real-world settings.

E)Results:

Random Forest Accuracy Scores: [0.99 0.99 0.975 0.99 0.975 0.99 1. 0.995 1. 0.985]

Random Forest Mean Accuracy: 0.9889999999999999

Random Forest Standard Deviation of Accuracy: 0.008306623862918082

Gradient Boosting Accuracy Scores: [0.865 0.845 0.835 0.86 0.85 0.9 0.89 0.88 0.9 0.85 ]

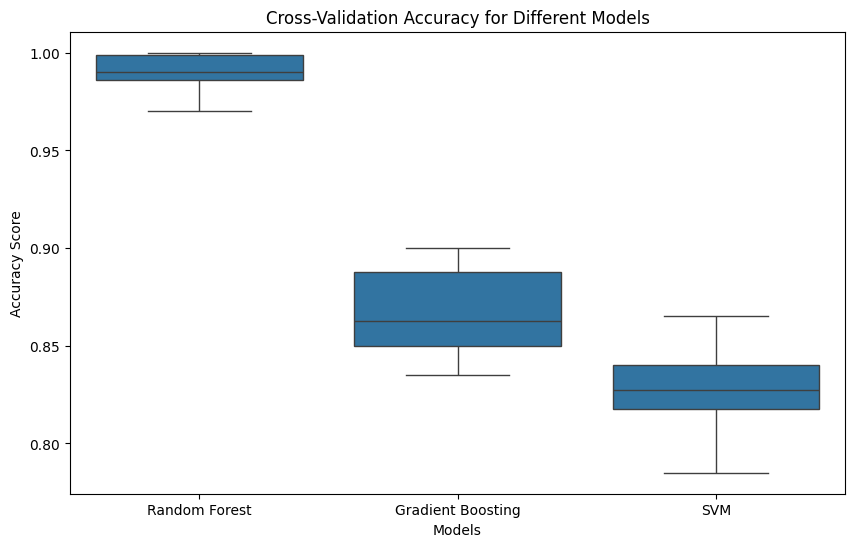
Gradient Boosting Mean Accuracy: 0.8675

Gradient Boosting Standard Deviation of Accuracy: 0.022388613177238132

SVM Accuracy Scores: [0.84 0.8 0.785 0.83 0.84 0.825 0.865 0.855 0.825 0.815]

SVM Mean Accuracy: 0.828

SVM Standard Deviation of Accuracy: 0.02282542442102664



*Fig 5: Cross Validation Box Plot*

# **RESULTS AND ANALYSIS**

# **SVM**

Designed to determine the optimal hyperplane that splits many classes within the feature space, the Support Vector Machine (SVM) is an efficient classification tool. It functions by maximizing the separation between each class's nearest data points, often known as support vectors. SVM transforms the input data using kernel functions, allowing it to handle both linear and non-linear datasets with ease. Because of this feature, SVM is particularly well-suited for complicated datasets, like those used to identify diabetes, where precise class separation is essential for producing predictions that can be trusted.

Support Vector Machine (SVM) is most efficient at processing both linear and nonlinear datasets, and use kernel methods to improve class separation. This makes SVM particularly useful in fields like diabetes diagnosis, which need complex data and precise categorization.

SVM often outperforms alternative diabetes diagnosis techniques such as logistic regression, decision trees, and K-nearest neighbors. The ability to examine multidimensional data without overfitting improves accuracy. Despite being more computationally expensive than simpler models like Naive Bayes, SVM's durability and accuracy make it one of the most reliable medical prediction algorithms available.

# **GRADIENT BOOSTING**

A potent ensemble learning method used for both classification and regression applications is gradient boosting.It constructs models incrementally by combining weak learners, usually decision trees, to develop a robust predictive model. The process involves training each new tree to address the errors made by the previous models, focusing on the residuals from earlier iterations to minimize a specific loss function. This iterative approach allows Gradient Boosting to capture complex patterns in high-dimensional data effectively. Additionally, it incorporates regularization techniques to reduce the risk of overfitting, making it an ideal choice for applications like diabetes detection, where accurate predictions depend on understanding intricate relationships within the data. [4]

To increase the accuracy of diabetes prediction, this study applies gradient boosting techniques to the Pima Indians Diabetes Database. In order to correct the mistakes of its predecessors, gradient boosting builds an ensemble of decision trees that are trained one after the other. This approach improves classification performance by efficiently capturing complex relationships between features, such as BMI and glucose levels. The findings show that gradient boosting performs better than conventional classifiers, offering insightful information about important predictive variables and confirming its suitability for use in medical decision-making for the treatment of diabetes.

# RANDOM FOREST

During the training phase, the Random Forest ensemble learning approach generates many decision trees.

Combining their results for enhanced predictive performance. Each tree is constructed using a random sample of the dataset and a subset of features, which helps minimize overfitting and improves generalization. By aggregating predictions through majority voting, Random Forest produces a more consistent model compared to individual trees. Its capacity to handle large, complex datasets makes it particularly effective in challenging tasks like diabetes detection, where various interacting features must be considered.

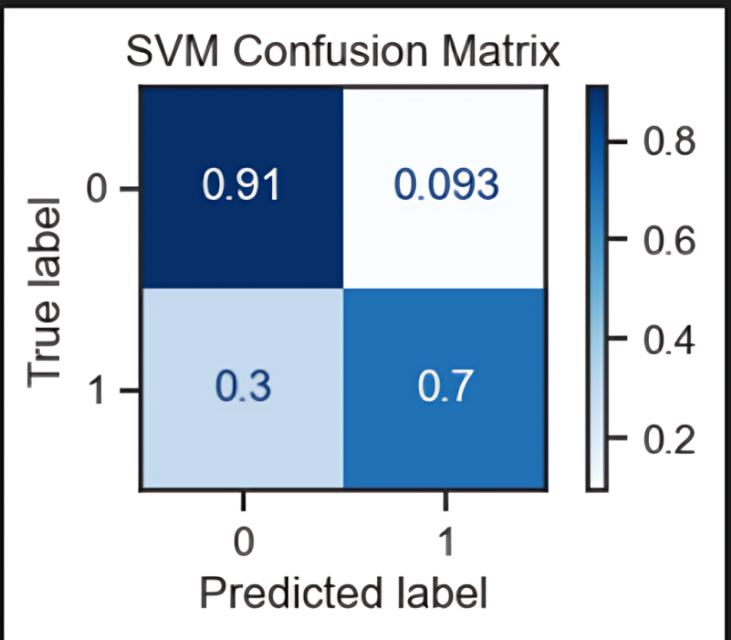
The Pima Indians Diabetes Dataset must be loaded, features and the target variable separated, and then divided into training and testing sets in order to use Random Forest. Make predictions on the test set after training the Random Forest model on the training data, and assess its performance using metrics such as the confusion matrix and accuracy.

# ANALYSIS

Table 3. F1-Score Analysis of used models for eavluation

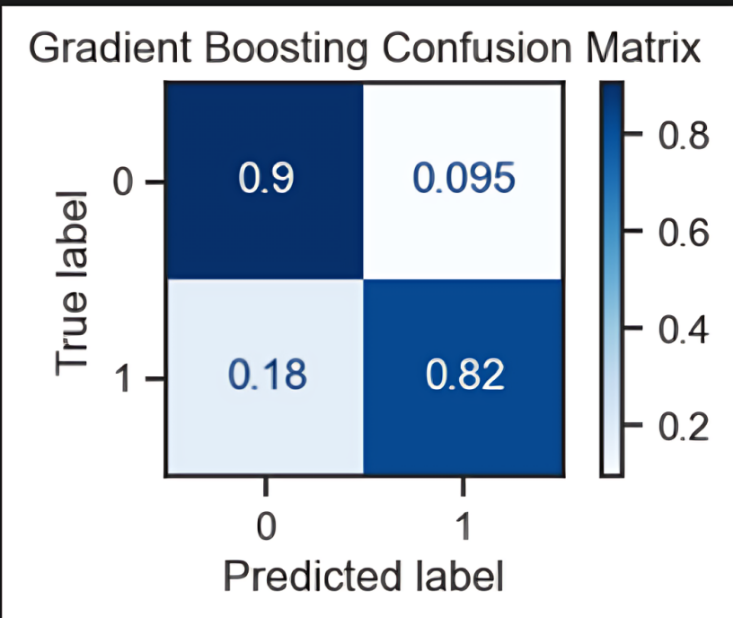
|  |  |
| --- | --- |
| Model | F1-Score |
| SVM | 0.9441 |
| Gradient Boosting | 0.8264 |
| Random Forest | 0.7493 |

Starting working with the logistic regression SVM, the trained model acquired 77.94% accuracy on the unseen dataset. Figure 5 depicts the confusion matrix for this model.



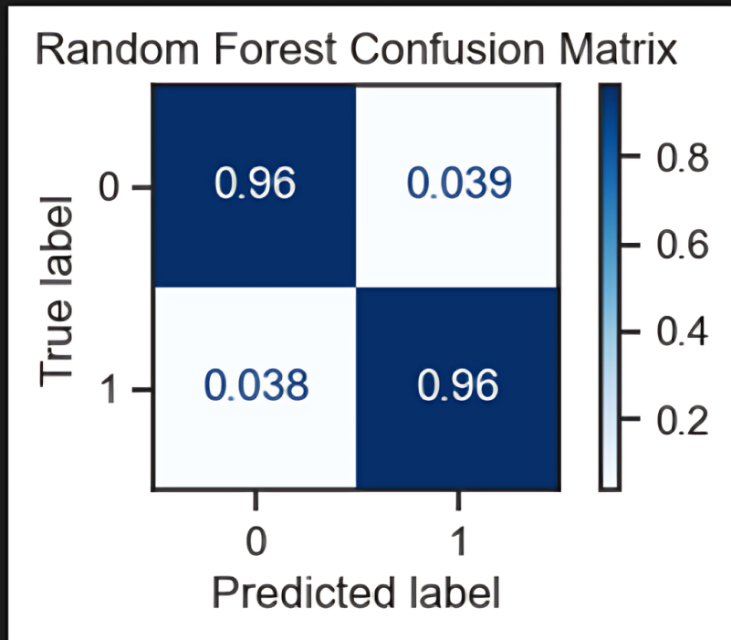
*Fig-6: SVM Confusion Matrix*

The second model was trained using the Gradient Boosting. The accuracy of the model, when run on the unseen test set, was 88.21%. Figure 6 shows the confusion matrix obtained from this model. [4]



*Fig-7: Gradient Boosting Confusion Matrix* [4]

The model was trained using the Random Forest approach to reach the best-scoring classifier and obtained the maximum accuracy among the classifiers given. When tested on hypothetical data, the accuracy was 96.07%. The confusion matrix of its findings is displayed in Figure 7.



*Fig8: Random Forest Confusion Matrix*

# **CONCLUSION**

This research examined three well-known classification algorithms—Support Vector Machine (SVM), Gradient Boosting, and Random Forest—for detecting diabetes. Each model has distinct strengths and approaches that contribute to effective classification. SVM is proficient at identifying optimal separation boundaries between classes, making it ideal for intricate datasets. Gradient boosting is more scalable to large, complicated datasets, automatically detects non-linear correlations, and is more robust against noise and outliers than Support Vector Machines (SVM), which necessitates meticulous feature scaling and tuning. On the other hand, and Random Forest adopts an ensemble method,

enhancing prediction accuracy and consistency by aggregating results from multiple decision trees. Among the models assessed, Random Forest proved to be the most reliable classifier for diabetes detection, showing greater accuracy and robustness when managing complex, high-dimensional data compared to the others. [3] [4]

This comparative analysis highlights that each algorithm brings its own advantages, and the selection of the appropriate model can significantly impact the accuracy of diabetes predictions. Overall, this study emphasizes the necessity of choosing suitable machine learning techniques tailored to the specific features of the dataset. Future research could focus on further refining these models and investigating hybrid approaches to improve classification effectiveness in diabetes detection. [13]

According to the comparative analysis of classification algorithms, the Random Forest model performed the best in terms of diabetes detection, with an accuracy of 98.9% and a minimal standard deviation of 0.0083 across folds. This consistency illustrates the model's robustness and ability to be generalized across a variety of data subsets. Gradient boosting performed consistently on more complex patterns, with an accuracy of 86.75% and a standard deviation of 0.0223, despite capturing intricate non-linear interactions. The Support Vector Machine (SVM) had the lowest accuracy at 82.8% with a standard deviation of 0.0228, demonstrating more sensitivity to feature scaling and tuning, although being generally effective for high-dimensional data. According to these findings, Random Forest is the most reliable and consistent model for diabetes detection, exceeding others methods. Gradient Boosting and SVM, on the other hand, show promise for additional improvement.

# **REFERENCES**

1. A. D. Association, “Classification and diagnosis of diabetes: standards of medical care in diabetes,” Diabetes Care, vol. 43, no. Supplement 1, pp. S14-S31, 2020.

2. L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

3. V. Vapnik, “The nature of statistical learning theory,” IEEE Transactions on Neural Networks, vol. 8, no. 6, pp. 1564-1565, 1995.

4. J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.

5. I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” Journal of Machine Learning Research, vol. 3, pp. 1157-1182, 2003.

6. R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 1995, pp. 1137-1145.

7. T. Hastie, R. Tibshirani, and J. Friedman, “The elements of statistical learning: data mining, inference, and prediction,” Springer Series in Statistics, New York, NY: Springer, 2009.

8. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

9. C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

10. A. Y. Ng, “Feature selection, L1 vs. L2 regularization, and rotational invariance,” in Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004.

11. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785-794.

12. J. Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” in Advances in Large Margin Classifiers, Cambridge, MA: MIT Press, 1999, pp. 61-74.

13. P. Domingos, “A few useful things to know about machine learning,” Communications of the ACM, vol. 55, no. 10, pp. 78-87, 2012.

14. I. Kononenko, “Machine learning for medical diagnosis: history, state of the art, and perspective,” Artificial Intelligence in Medicine, vol. 23, no. 1, pp. 89-109, 2001.

15. T. Fawcett, “An introduction to ROC analysis,” Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006.

16. S.P. Siddique Ibrahim, Dr.M. Sivabalakrishnan., “Lazy Learning Associative Classification in Map Reduce Framework” International Journal of Recent Technology and Engineering (IJRTE) **(Scopus),** Vol. 7, Issue 4S, pp. 168-172, Nov 2018.

17. Siddique Ibrahim S P., "Rare Lazy Learning Associative Classification using Cogency Measure for Heart Disease Prediction" International conference on Green Technologies for power generation, communication & Instrumentation (ICGPC’19), April 2019.