Predictive Modeling for Diabetes Diagnosis Using Machine Learning

Rudraksh Yadav[1], Aditi Sharma[2]

[1]Centre for Advanced Studies, Dr. A.P.J. Abdul Kalam Technical University, Lucknow, India
[2]Institute of Engineering and Technology, Lucknow, India

Abstract

Diabetes mellitus is one of the chronic conditions that leads to a global health crisis. Numerous factors, including obesity and elevated blood glucose levels, contribute to the growth of diabetes. Amputations, heart failure, stroke, blindness, renal failure, etc., are all primarily brought on by diabetes. Our body converts food into glucose and sugars when we eat. Currently there are around 830 million people in the world having diabetes mellitus. There are two most prevalent forms of diabetes, which are type 1 diabetes and type 2 diabetes, and one more, which includes gestational diabetes mellitus that develops during pregnancy. The project's goal is to develop a structure that could project a patient's risk of developing diabetes mellitus by working more accurately and integrating the findings of several different types of algorithms. The dataset used in this study is the Pima Indians Diabetes Dataset, which is taken from Kaggle.

KEYWORDS: Diabetes Mellitus, Chronic Condition, Global Health Crisis, Obesity, Elevated Blood Glucose , Metabolism , Gestational Diabetes

1. Introduction

Diabetes mellitus is a condition that is rapidly spreading throughout society, including among children. In order to comprehend diabetes mellitus and how it arises, we must first comprehend what occurs in a person's body in the absence of diabetes. Our diets contain sugar (glucose), particularly something that is high in carbohydrates. Everyone needs carbohydrates because carbohydrates are the main energy source for the body. Foods that are high in carbohydrates include bread, cereal, rice, pasta, fruit, dairy products, and vegetables, particularly those that are starchy.

The body converts these meals into glucose when we consume them, and then through the bloodstream, the glucose travels throughout the body. A small section of the glucose is absorbed by the brain to facilitate optimal function as well as clear thinking. It is also carried to our liver, from where it is stored for later use. The cells in our body use the remaining glucose as fuel, which acts like a source of energy. The pancreatic beta cells produce the hormone insulin, which works similarly to a door key. Cell doors are opened by insulin binding to them, allowing bloodstream glucose to enter the cells. Blood glucose levels rise.
which results in developing diabetes mellitus if the pancreas is unable to generate the required amount of insulin. This results in high blood sugar (glucose) and urine sugar levels, which are indicative of diabetes mellitus.

1.1 Types of Diabetes

Type 1 diabetes is characterized by both inadequate cell synthesis of insulin and a compromised immune system.. There are currently no preventative measures that are effective for type 1 diabetes, and there is also no compelling research that demonstrates its causes.
Type 2 diabetes occurs if the body is not able to use insulin or when the cells create insufficient amounts of it. Since it is one of the most common types of diabetes, 90% of those who have been diagnosed with the disease are affected. It results from a combination of lifestyle choices as well as hereditary factors.

1.2 Symptoms of Diabetes

- Increased thirst
- Frequent urination
- Tired and Sleepiness
- Blurred vision
- Weight loss
- Frequent Infections
- Mood swings

1.3 Causes of Diabetes

Different genetic factors can lead to the cause of diabetes mellitus. It is mainly originated by not less than two mutant genes in chromosome 6. Viral infections can have an impact on the occurrence of type 1 and type 2 diabetes. Research has indicated that certain viral infections, including cytomegalovirus, hepatitis B virus, mumps, rubella, and Coxsackievirus, raise the risk of diabetes.

2. Literature Review

Yasodha et al. [1] use sorting on a range of dataset types to detect if an individual has diabetes. Information from the depository of hospitals, which contains about two hundred occurrences with at least nine characteristics, is compiled to generate the data set for diabetes patients. Urine and blood tests are the two classifications to which these dataset instances belong. After that, the results may be different. They use J48, Random Tree, Naïve Bayes, and REP Tree. The J48 was found to have the best performance out of all of them, with an accuracy rate of 60.2%. To determine several methods for diagnosing diabetes, Aiswarya et al.

Aiswarya et al. [2] used categorization analysis using different algorithms like naïve Bayes as well as decision trees so as to examine and evaluate the patterns that emerge in the data in order to discover various methods for identifying diabetes mellitus. The main aim of this learning was to build a more rapid as well as efficient method of detecting the condition, which would help patients recover more quickly.

Gupta et al. [3] focuses on observing and computing the sensitivity and accuracy as well as the explicitness percentage of quite a few classification methods and tries to differentiate and look over the results. The learning differentiates the performances of the same classifiers when

implemented on various outer tools like RapidMiner and MATLAB while using the same specifications. Different models like BayesNet, JRIP, and Jgrapht were used. In accordance with the findings of this learning , JGraphT has the highest accuracy, which was 81.3%, followed by sensitivity, which was 59.7%, and particularity 81.4% at the last.

Lee et al. [4] concentrates on using the CART decision tree method on the diabetes mellitus dataset with the application of a resample filter. Before implementing any technique to improve accuracy rates, the author emphasizes the importance of addressing the issue of class imbalance. A dataset that consists of dichotomous values is more likely to exhibit class imbalance. This suggests that the class variable has two possible outcomes, both of which may be easily controlled as long as identified early while in the preparation phase. That would increase the model's accuracy.

Study	Dataset Characteristics	Algorithms Used	Tools/Platform	Main Focus / Objective	Best Accuracy Achieved
Yasodha et al. [1]	~200 instances, 9+ features, hospital data (urine & blood tests)	J48, Random Tree, Naïve Bayes, REP Tree	Not specified	Compare basic classification algorithms on small-scale dataset	60.2% (J48)
Aiswarya et al. [2]	Not specified clearly; exploratory data analysis	Naïve Bayes, Decision Tree	Not specified	Discover patterns in the data to improve rapid diabetes detection	Not reported
Gupta et al. [3]	Used same data across different platforms	BayesNet, JRIP, JGraphT	RapidMiner & MATLAB	Compare classifier performance across platforms/tools	81.3% (JGraphT)
Lee et al. [4]	Binary/dichotomous diabetes dataset	CART Decision Tree with resample filter	Not specified	Handle class imbalance to improve model performance	Not clearly reported; focus was on balancing data

3. Methodology

This section will cover the different classifiers that are used to predict diabetes mellitus. In order to increase the accuracy, we will additionally provide our suggested methodology. This study employs a variety of approaches. A description of the various techniques and models is given below. The accuracy of the measurements are the outcomes that could be used for diabetes mellitus prediction.

The diabetes dataset used in this paper consists of 769 data points. The principal purpose of this learning is to project whether a patient is affected by diabetes mellitus or not based on the outcomes.

3.1 Data Selection

The dataset used in this study is the Pima Indians Diabetes Dataset, which is taken from Kaggle. Then the preprocessing of data is done, in which we have handled the missing values, for example, replacing zero values with mean and median.

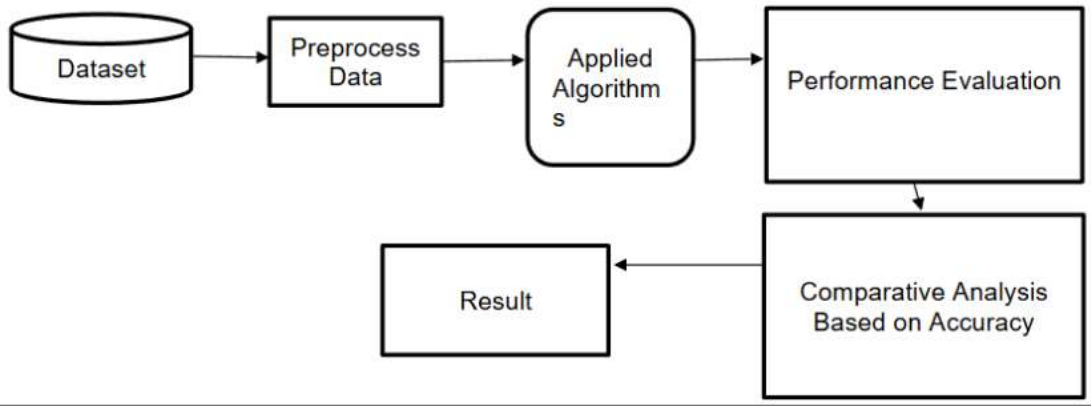
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

- The diabetes mellitus dataset comprises 769 data along with 9 characteristics each.
- The characteristics which would be forecasted are called outcomes where 0 denotes not diabetic while 1 denotes diabetic.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               768 non-null    int64  
 1   Glucose                   768 non-null    int64  
 2   BloodPressure             768 non-null    int64  
 3   SkinThickness             768 non-null    int64  
 4   Insulin                   768 non-null    int64  
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64  
 8   Outcome                   768 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

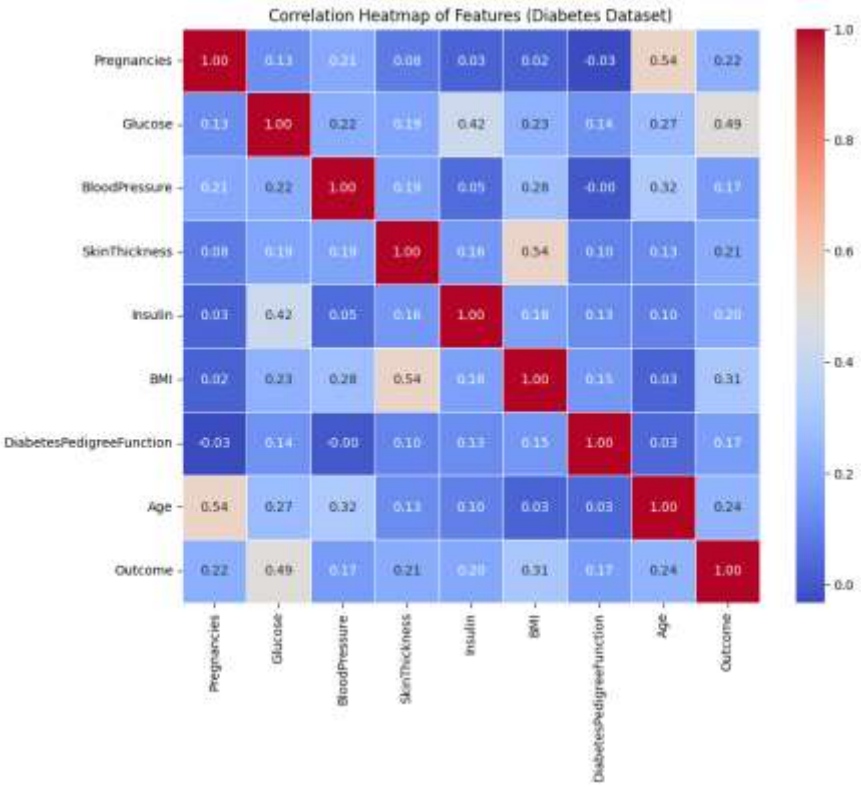
- There is not a single null value in the dataset used.



Proposed Model Diagram

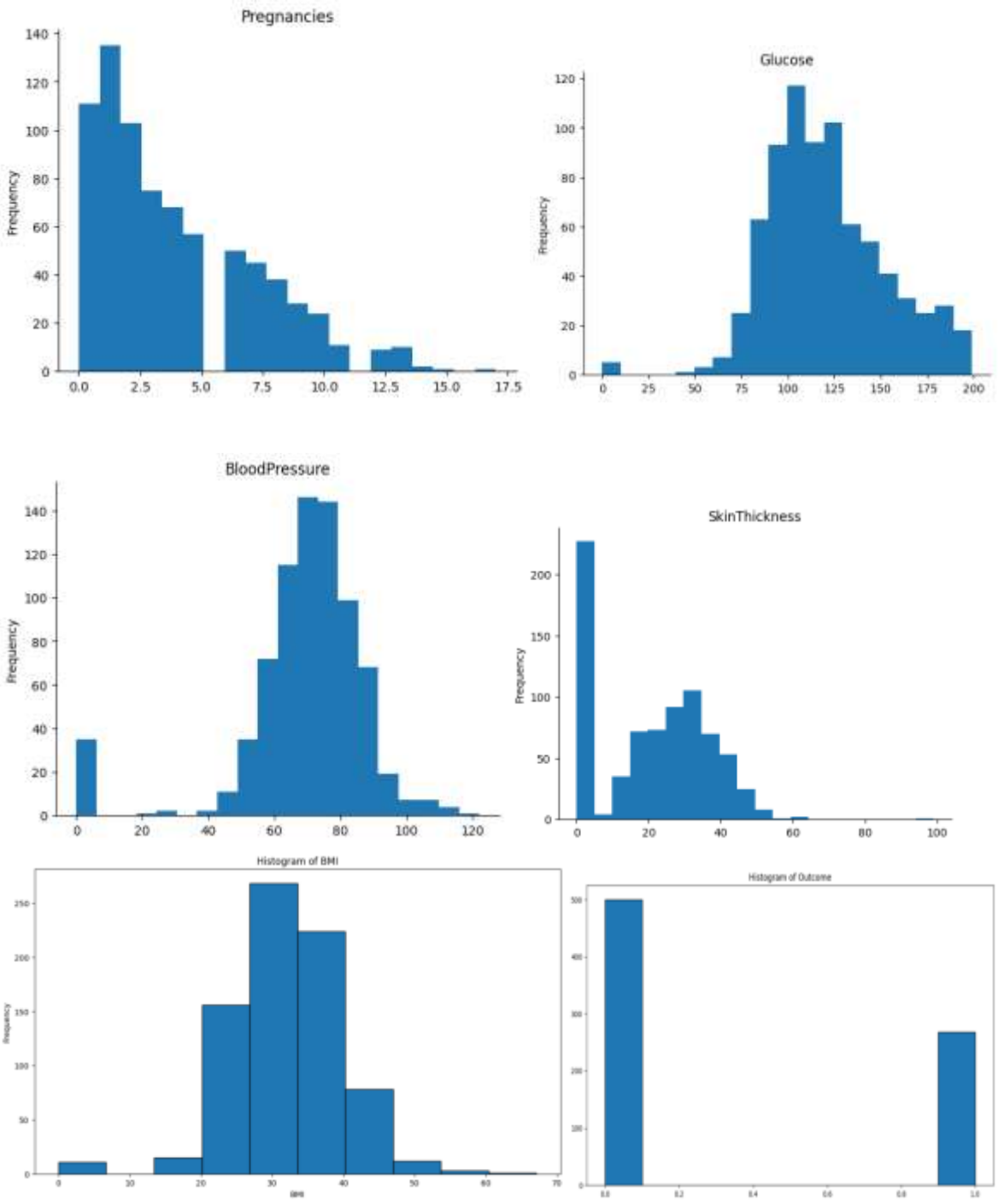
- The dataset is first preprocessed where we separate the features and labels. Scaling is also performed because SVM benefits from scaled data.
- Then we apply the classification algorithm based on the concept of maximizing the margin between the two classes-
- **Positive class :** Patient has diabetes(outcome = 1)
- **Negative class :** Patient does not have diabetes(outcome = 0)
- Then the performance of the model is evaluated.
- It is followed by comparative analysis based on accuracy.
- Then we get the final results.

3.2 Correlation Heatmap



- It is evident that our result value does not have a strong association with any one attribute.
- It can be seen that some of the traits are positively correlated with the outcome value, while others are negatively correlated.

3.3 Histogram



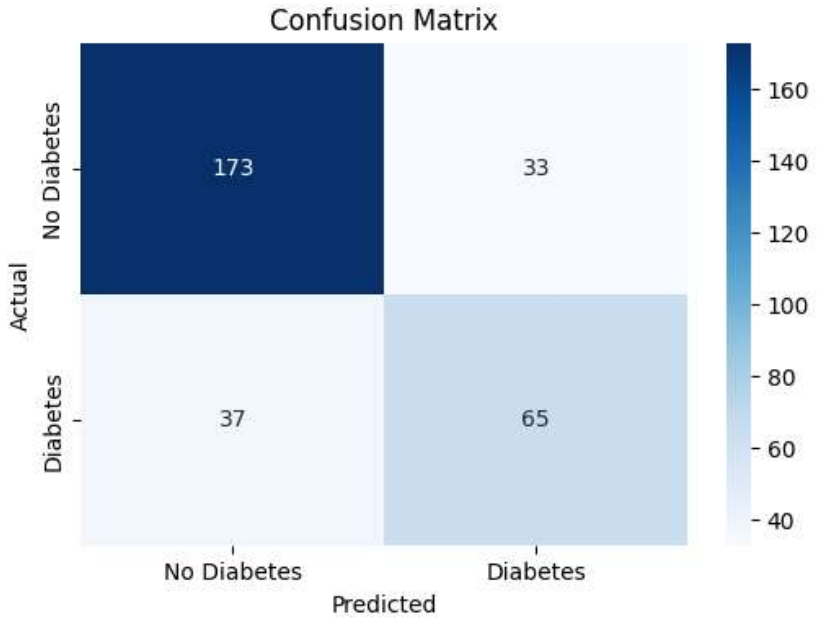
Let's look at the plots that demonstrate the distribution of each characteristic and label across different ranges and validate the necessity of scaling. Discrete bars show that each of these is

a categorical variable that needs to be addressed before machine learning is used. Additionally, we categorize our outcome labels into two classes: 0 for no disease and 1 for disease.

3.4 Support Vector Machine Classifier

This supervised machine learning approach is applied to tasks involving regression and classification. Support vector machines can handle regression issues and are particularly well-suited for classification jobs. The support vector machine classifier finds the best hyperplane in an N-dimensional space to divide the input points into distinct groups. The algorithm maximizes the distance between the closest points of different classes.

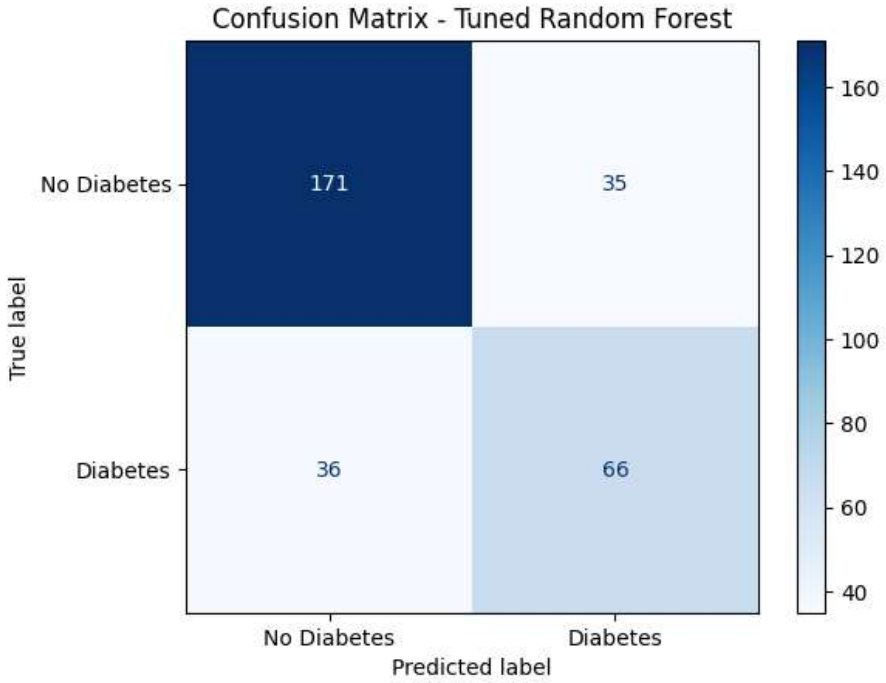
- **Training the Support Vector Machine Classifier**
- The support vector machine classification splits the dataset into training and test sets and standardizes features by removing the mean and scaling to unit variance.
- 20% of the data is reserved for testing while 80% for training.
- We have used a linear kernel, which works well for linearly separable data.
- Other options include ‘rbf’,’poly’ and ‘sigmoid’.
- During training, SVM finds the ‘best’ hyperplane that maximally separates the classes in our training data.
- During prediction, it uses this hyperplane to classify new points.
- Then we evaluate the accuracy of the model. In this, we compare the predicted labels to the true labels from the test set.
- Accuracy = (Number of correct predictions)/(Total number of predictions). It’s a measure of how well is the proposed model performing on the unseen data.
- High accuracy indicates that the SVM model has generalized well.
- Low accuracy could indicate that the features are not scaled properly.



3.5 Random Forest Classifier

An ensemble of decision trees is used by a machine learning method called a Random Forest Classifier to classify data. It reduces overfitting and increases accuracy by combining the predictions of several decision trees. To identify a test object's final class, the votes from various decision trees are combined.

- **Training the Random forest classifier**
- Divide the data into training and testing sets.
- Training set is used to fit the model.
- Test set is used to evaluate its performance on unseen data.
- Use the trained model to predict the outcomes on test data.
- Hyperparameter tuning is also performed using GridSearchCV.

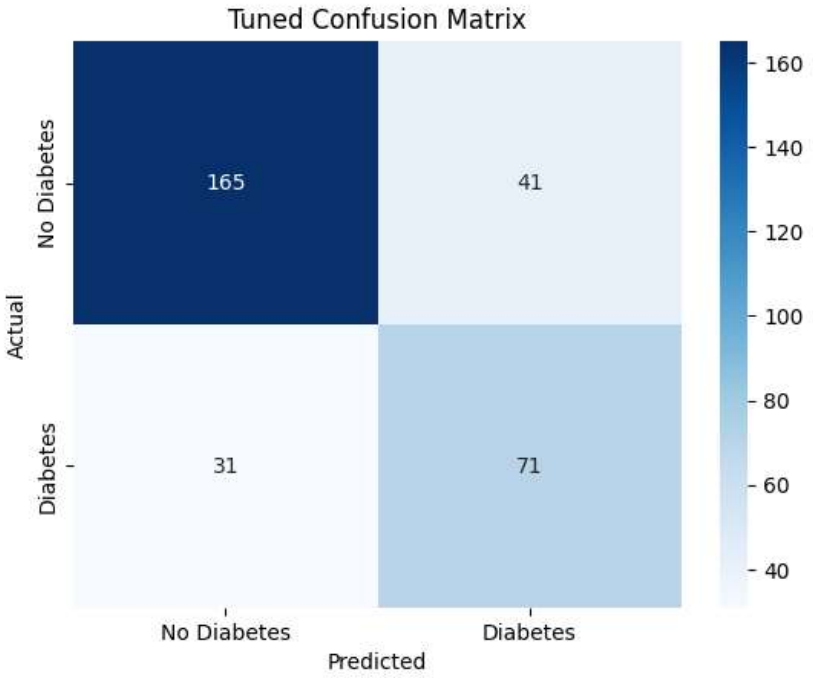


3.7 XGBoost

Extreme Gradient Boosting, or XGBoost, is a distributed gradient-boosted decision tree (GBDT) machine learning toolkit that is scalable. It is the top machine learning package for tasks including regression, classification, and ranking and offers parallel tree boosting.
Understanding the machine learning principles and methods that XGBoost is based on—supervised machine learning, decision trees, ensemble learning, and gradient boosting—is essential to comprehending XGBoost.

- **Training the XGBoost Classifier**
- Separate the features (X) and labels (y), and split into training and testing sets.
- XGBoost starts with an initial prediction — typically the mean of the target variable (e.g., the average probability of diabetes).
- It uses a loss function — usually log loss for binary classification (diabetes yes/no).
- This tells the model how far off its prediction is from the actual outcome.
- The tree’s predictions are added to the previous predictions (not replacing them), scaled by a learning rate.

- This gradually improves the model's accuracy.
- XGBoost continues building many small trees, each one correcting the previous one.
- It stops when a fixed number of trees is reached.
- Or improvement is no longer significant.



4 Result and Discussion

Model	Train ,Test Ratio	Precision	Recall	F1 Score	Accuracy
Support Vector Machine	80:20	76%	76%	76%	75.9%
	70:30	75%	75%	75%	74.9%
	60:40	77%	77%	77%	85%

Model	Train,Test Ratio	Precision	Recall	F1 Score	Accuracy
Random	80:20	72%	72%	72%	72%

Forest	70:30	76%	75%	75%	75%
	60:40	77%	77%	77%	76.9%

Model	**Train,Test Ratio**	**Precision**	**F1 Score**	**Recall**	**Accuracy**
XGBoost	80:20	75%	75%	75%	74.6%
	70:30	75%	75%	75%	74.8%
	60:40	77%	77%	77%	76.6%

- SVM performs well achieving high accuracy after performing hyperparameter tuning .
- Good precision indicates that when the model predicts diabetes, it's usually correct.
- The SVM model correctly identifies most non-diabetic individuals.
- The formula to calculate accuracy is :

Accuracy = TP+TN/TP+TN+FP+FN

- The formula to calculate precision is :

Precision = TP/TP+FP

- The formula for Recall is:

Recall = TP/TP+FN

- The formula for F1 Score is:

F1 Score = 2 x [(Precision x Recall)/(Precision + Recall)]

4.1 Conclusion and Future work

Early identification of diabetes is one of the major real-world medical challenges. This study aims to forecast diabetes by means of a system that is designed with methodical efforts. Throughout this project, the SVM, Random Forest and XGBoost classification methods are examined and assessed using a range of metrics. Research is conducted using the dataset based on diabetes. Based on experimental data, the Support Vector Machine algorithm achieves 85% accuracy in determining the suitability of the proposed system.

Future Scope - Other diseases may be predicted or diagnosed in the future using the system that was created and the machine learning classification method that was employed. The work can be enhanced and expanded to include other machine learning methods for the automation of diabetes analysis.

References

[1]. Yasodha, P., & Ramesh, V. (2012). *Analysis of a population of diabetic patients databases in Weka tool*. International Journal of Scientific & Engineering Research, 3(3), 1–4.
[2]. Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). *Diagnosis of diabetes using classification mining techniques* (International Journal of Data Mining & Knowledge Management Process, Vol. 5, No. 1).
[3]. Gupta, S., & Ranjan, P. (2014). *Diabetes prediction using data mining techniques*. International Journal of Advanced Research in Computer Science and Software Engineering, 4(7), 559–564.
[4]. Lee, S. (2012). *Improving diabetes classification with resampling techniques and decision trees*. International Journal of Advanced Computer Science and Applications, 3(8), 107–111.
[5].Dhomse Kanchan B., M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE. pp. 5–10.
[6]. Sharief, A.A., Sheta, A., 2014. Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming. International Journal of Advanced Research in Artificial Intelligence (IJARAI) 3, 54–59.doi:doi:10.14569/IJARAI.2014.031007.
[7]. Sisodia, D., Shrivastava, S.K., Jain, R.C., 2010.ISVM for face recognition. Proceedings - 2010 International Conference on Computational Intelligence and Communication Networks, CICN 2010 , 554–559doi:10.1109/CICN.2010.109.
[8]. Khokhar, P. B., Pentangelo, V., Palomba, F., & Gravino, C. (2025, January 30). *Towards [9]Transparent and Accurate Diabetes Prediction Using Machine Learning and Explainable Artificial Intelligence*. arXiv.
[10]Alzboon, M. S., Al-Batah, M., Alqaraleh, M., Abuashour, A., & Bader, A. F. (2025, June 11). *A comparative study of machine learning techniques for early prediction of diabetes*.
Yang, Z., Wang, F., Huang, X., Li, X., Liu, S., & Zhang, H. (2024, October 16). *Optimization and application of cloud-based deep learning architecture for multi-source data prediction*.
[11]. Dataset - https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.