Diabetes Destiny: Unraveling the Future with Precision Prediction

Maram Bakini[a], Mrs Monica Subashini[a]

[a] *Vellore Institute of Technology, , Vellore, , Tamil Nadu, India*

Keywords: KNN, Decision Tree, GBN, SVM, machine learning, xgboost

1. Abstract

Diabetes illness prediction remains a crucial problem in healthcare, and the literature has offered a variety of complex procedures and elaborate prepossessing techniques. In this study, we investigated the efficacy of a simplified preprocessing strategy for diabetes prediction, seeking to show that simplicity can produce accuracy that is competitive with complicated approaches.

We rigorously tested our preprocessing pipeline using a well-known dataset called the Indian Pima Dataset and our simple preprocessing workflow. Our method utilizes a more easy and intuitive preprocessing strategy, greatly lessening the preprocessing workload compared to previous approaches that involve complex feature engineering and data manipulation.

2. Introduction

Figure 1: The Procedure on Blood Test

Understanding diabetes' multiple character is crucial to understanding how complex the disease is. There are several varieties of diabetes, including Type 1, Type 2, gestational, and other uncommon variants, each having its own underlying causes and pathophysiological mechanisms. Diabetes has historically been diagnosed primarily through clinical signs, genetic predispositions, and glucose tolerance testing. However, these techniques frequently fall short in terms of making precise forecasts and tailored responses.

This study aims to close this gap by revolutionizing diabetes prediction by leveraging the capabilities of ML algorithms. Large-scale patient data can be used to help ML models find complex patterns, obscure relationships, and undetectable biomarkers that can greatly increase prediction accuracy. Additionally, ML algorithms have the capacity to continuously learn from and adapt to new data, producing predictions that are dynamic and up-to-date. By creating a cutting-edge ML-based model that incorporates numerous patient-specific elements such genetic profiles, lifestyle habits, medical histories, and environmental impacts, we want to advance the area of diabetes prediction through this work. With the use of our suggested approach, we hope to improve the accuracy of diabetes prediction while also offering insightful information about specific risk factors and relevant treatment options.

Wide-ranging effects of this research could alter tailored healthcare for people who are at high risk of acquiring diabetes. Healthcare providers can better target preventative measures, lifestyle interventions, and therapeutic approaches to slow the incidence and course of diabetes by accurately identifying high-risk individuals. Furthermore, ML-driven diabetes prediction models can maximize the use of healthcare resources, lower expenses, and lessen the strain on healthcare systems.

In conclusion, the application of ML approaches holds enormous promise for the field of diabetes prediction, enabling healthcare professionals to give individualized interventions and more informed judgments. In order to advance patient outcomes, preventative treatment, and diabetes patients' quality of life, this research study intends to make a contribution to this interesting field by creating a novel ML-based model.

3. Literature review

Aishwarya Mujumdar [MV19] has implemented various machine learning algorithms such as Support Vector Classifier, Random Forest Classifier, Decision Tree Classifier, Extra Tree Classifier, Ada Boost algorithm, Perceptron, Linear Discriminant Analysis algorithm, Logistic Regression, K-Nearest Neighbour, Gaussian Naive Bayes, Bagging algorithm, and Gradient Boost Classifier.Clustering is performed also using the K-means clustering algorithm on highly correlated attributes, such as Glucose and Age, to find patterns and group similar patients together based on these attributes.

MD. KAMRUL HASAN and all [Has+20] have started with preprocessing techniques. This includes outlier rejection, filling missing values, data standardization, feature selection, and K-fold cross-validation. These preprocessing steps are crucial for achieving the state-of-the-art results in diabetes prediction. Next, different machine learning classifiers are applied. The classifiers used in the experiments include k-nearest Neighbour, Decision Trees, Random Forest, AdaBoost, Naive Bayes, XGBoost, and Multilayer Perceptron (MLP). Extensive experiments are conducted using different combinations of preprocessing techniques and classifiers to maximize the area under the curve (AUC) of diabetes prediction. The performance of each classifier is evaluated using various evaluation metrics such as sensitivity, specificity, false omission rate, diagnostic odds ratio, and AUC. The best performing classifier is then selected as the baseline model for further evaluation. In addition, an ensembling classifier is proposed by combining multiple machine learning models. This ensemble classifier, which combines boosting classifiers (adaptive and gradient boosting), outperforms standalone classifiers and other combinations of classifiers. It achieves a higher AUC and provides more accurate predictions for diabetes. Overall, the methodology involves a combination of preprocessing techniques, extensive experimentation with different classifiers, evaluation using various metrics, and ensembling of classifiers to improve the accuracy of diabetes prediction.

The methodology used by Mitushi Soni [SV20] to predict diabetes involves the following steps: 1. Data Preparation: The author uses the Pima Indian Diabetes Dataset, which contains various attributes related to diabetes. The dataset is prepared by cleaning and preprocessing the data, ensuring that it is in a suitable format for analysis. 2. Machine Learning Techniques: The author applies various classification and ensemble techniques to the prepared dataset. These techniques include K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB), and Random Forest (RF). 3. Model Evaluation: The performance and accuracy of the applied algorithms are evaluated and compared. The author compares the results of different machine learning techniques to determine which algorithm is best suited for predicting diabetes. The goal of the methodology is to develop a model that can accurately predict the onset of diabetes. The author aims to achieve higher accuracy in predicting diabetes by comparing and selecting the most suitable machine learning algorithm for the task.

The methodology used to predict diabetes in this research study by Jobeda Jamal Khanam, Simon Y. Foo [KF21] involved the use of data mining and machine learning algorithms. The researchers used the Pima Indian Diabetes (PID) dataset, which contains information about 768 patients and their corresponding nine unique attributes. They applied data mining techniques to preprocess and select relevant features from the healthcare data. Then, they used seven different machine learning classification algorithms, including Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Adaboost, Random Forest, K Nearest Neighbor (KNN), Decision Tree (DT), and Neural Network (NN). These algorithms were applied to the dataset to predict whether a patient has diabetes or not. The performance of each algorithm was evaluated using various measures such as accuracy, precision, recall, and F-measure. The researchers found that the model with Logistic Regression (LR) and Support Vector Machine (SVM) yielded good results in terms of accuracy for predicting diabetes. Overall, the combination of data mining techniques and machine learning algorithms was used in this research to automate the prediction of diabetes and improve accuracy.

Safial Islam Ayon, Md. Milon [AI19] used a deep neural network as the methodology to predict diabetes. They trained the dataset before predicting diabetes to ensure accurate results. The deep neural network is a popular method in machine learning and has been shown to be effective in medical prognoses.

USAMA AHMED and all [Ahm+22] used two types of models, namely Support Vector Machine (SVM) and Artificial Neural Network (ANN), to predict diabetes. These models were trained and tested using a dataset obtained from the UCI Machine Learning Repository. The dataset was preprocessed by cleaning, normalizing, and dividing it into training and testing datasets. The SVM classifier achieved a higher accuracy rate of 79.13 compared to other machine learning algorithms used in previous studies. The ANN model with MMS achieved an accuracy of 82.35, which was higher than the other four algorithms. The testing layer of the system acquired data from a medical database and loaded a preprocessed training model from the cloud. The fused model, which used fuzzy logic, was used to predict whether a diabetes diagnosis is positive or negative. The prediction accuracy was calculated by comparing the required output with the actual output. The FMDP system, based on the fused model, made predictions based on whether both the ANN and SVM models predicted yes or no for diabetes.

Xue, Jingyu and Min [XMM20] uses supervised machine-learning algorithms, specifically Support Vector Machine (SVM), Naive Bayes classifier, and LightGBM, to train on actual data of 520 diabetic patients and potential diabetic patients aged 16 to 90. The data set used in the study is obtained from the UCI open source standard test data set website. The algorithm process involves using the data set as input for the prediction algorithm and then evaluating the classification accuracy of the algorithm using a confusion matrix. The aim is to identify the algorithm with the highest accuracy in predicting diabetes. The study concludes that SVM has the highest accuracy compared to the other algorithms tested.

The methodology used by Aishwariya Dutta and all [Dut+22] to predict diabetes is machine learning (ML). They employ various ML algorithms and techniques, such as decision trees (DT), naive Bayes (NB), support vector machines (SVM), random forest (RF), XGBoost (XGB), and LightGBM (LGB), to develop predictive models for diabetes. They use different datasets, such as the Diabetes Data Classification (DDC) dataset, and apply various ML classifiers to evaluate the performance of different missing value imputation (MVI) techniques, feature selection (FS) strategies, and number of selected features (NSF). Their goal is to determine the most effective ML approach for diabetes classification and prediction. They also compare the performance of different ML algorithms based on evaluation metrics such as area under the curve (AUC) and true positive rate versus false positive rate, and discuss the advantages of using ML-based approaches for early identification and diagnosis of diabetes. They acknowledge the limitations of their study and recommend exploring modern deep learning techniques as future research directions.

Arianna Dagliati [Dag+18] used a data mining pipeline, which includes clinical center profiling, predictive model targeting, predictive model construction, and model validation, to predict the onset of microvascular complications in patients with type 2 diabetes. The variables considered for the prediction models were gender, age, time from diagnosis, body mass index (BMI), glycated hemoglobin (HbA1c), hypertension, and smoking habit. Logistic regression with stepwise feature selection was used to develop the predictive models, and strategies were employed to handle missing data and class imbalance. The prediction models were built for different time scenarios, such as 3, 5, and 7 years from the first visit at the Hospital Center for Diabetes.

The methodology used by S.Saru and S.Subashree to predict diabetes involves the use of data mining techniques and algorithms. The author employs the WEKA software as a mining tool for diagnosing diabetes. The Pima Indian diabetes database, obtained from the UCI repository, is used for analysis. The dataset is studied and analyzed to build an effective model that can predict and diagnose diabetes. In the study, the author applies the bootstrapping resampling technique to enhance accuracy. They also apply different classification algorithms, such as Naive Bayes, Decision Trees, and K Nearest Neighbors (KNN), and compare their performance in predicting diabetes. The aim is to find the best classifier from these algorithms for accurate prediction of the disease. Overall, the methodology used by the author involves analyzing the dataset using data mining techniques and algorithms, and comparing the performance of different classifiers to predict diabetes.

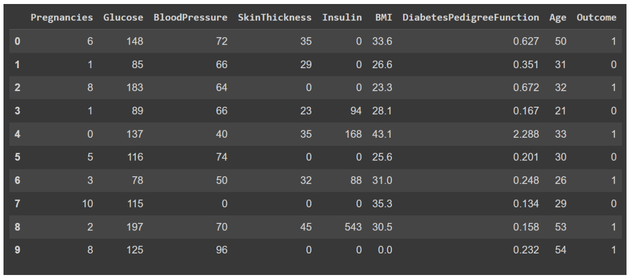
The methodology used by Muhammad Exell Febrian [Feb+23] to predict diabetes is by comparing two machine learning algorithms - k-Nearest Neighbor (KNN) and Naive Bayes. The study first collected a dataset with eight health attributes related to diabetes: pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age. Then, the dataset was pre-processed, which involved reading the dataset, viewing its contents, and filling in any missing data. Afterwards, the dataset was divided into training and testing data with a ratio of 80 to 10. The KNN and Naive Bayes algorithms were then applied to the training data to create prediction models. These models were evaluated using the Confusion Matrix, which helps assess the performance of a machine learning algorithm by comparing the actual data against the predicted outcomes. According to the results of the experiments and the evaluation using the Confusion Matrix, the Naive Bayes algorithm was found to outperform KNN in predicting diabetes based on the health attributes in the dataset.

The methodology used by lta Llaha [Olt21] to predict diabetes is data mining and machine learning. The author collected data about diabetes, including attributes such as age, body mass index, insulin levels, and glucose levels. Then, the author implemented various data mining algorithms, such as Simple Logistic, Multilayer Perceptron, Logistic, Naive Bayes, Bayes Net, SMO, and C4.5, to analyze the data and predict whether a person has diabetes or not. The author compared the performance of these algorithms using measures such as accuracy, precision, recall, and F-measure. The results showed that the decision tree algorithm had the highest accuracy rate of 79 and was able to classify diabetes data effectively. The decision tree algorithm was also chosen because it expresses the rules explicitly and is easy to interpret and understand. Overall, the author concluded that data mining and machine learning methods can be valuable tools for analyzing and predicting diabetes.

4. Dataset Description and preprocessing

4.1. Dataset Origin and Collection:

The Pima Indian dataset was originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) in collaboration with the Indian Health Service (IHS) and several Native American communities. The dataset primarily focuses on Pima Indian women aged 21 years or older, residing near Phoenix, Arizona, USA. The data collection process involved comprehensive medical examinations, including glucose tolerance tests, measurements of various physiological parameters, and demographic surveys. These assessments were performed between 1988 and 1990, ensuring a relatively recent and reliable dataset for analysis.



	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
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

Figure 2: indian pima dataset

4.2. Dataset Description:

Pregnancies: Number of times pregnant

Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test.

This is a lab test to check how your body handles the sugar. Normal person (2 hr after glucose test) should have less than 140mg/dl

Blood Pressure: Diastolic blood pressure (mm Hg).

Normal values are less than 80. Stage 1 hypertension: 80-89 Stage 2 hypertension: 90 or more Hypertensive crisis: 120 or more

Skin Thickness: Triceps skin fold thickness (mm)

For adults the normal values are 2.5 mm for men; 18 mm for women

Insulin: 2-Hour serum insulin (mu U/ml). Insulin is a hormone that helps move blood sugar.

150 mu U/ml is a critical number, in which most people with type 1 or 2 needs insulin theraphy

BMI: Body mass index (weight in kg/(height in m)^2) : $Assessifapersonisoverweightorunderweight.$

Underweight: less than 18.5 Normal weight: 18.5 - 24.9 Overweight: 25-29.9 Obese: over 30.0

Diabetes pedigree function: Provides some information on the history in relatives. This is a measure of genetic influence.

Age (years)

Target variable: Outcome 1 indicates having diabetes; 0 indicates not having diabetes.

4.3. Dataset Preprocessing:

The Pima Indian dataset provides a substantial problem in terms of class imbalance even though it is a useful resource for diabetes prediction. When the goal variable, which signals whether or not a person has diabetes, is distributed with a large bias, there is a class imbalance. The instances representing diabetes-positive patients and those representing diabetes-negative cases are noticeably imbalanced in the Pima Indian dataset. Predictive models may perform less effectively due to this imbalance.

Therefore, they can affect most classes, which will lead to poor performance in diagnosing diabetes. Dataset rebalancing techniques [Fre+21] should be used to overcome this problem, including undersampling most classes, oversampling minorities, and using more complex methods such as the Synthetic Minority Oversampling Technique (SMOTE). The prediction model examined in the Pima Native American dataset can improve the generality and accuracy of identifying individuals at risk for diabetes by reducing the problem of class impairment, which may improve outcomes of treatment and prevention measures.

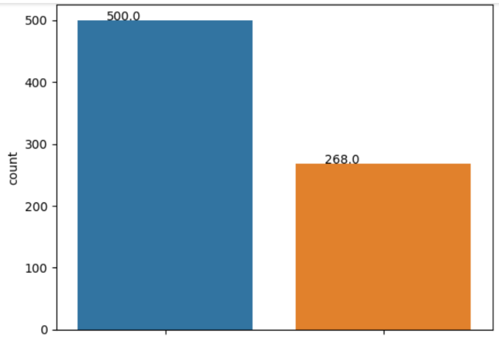
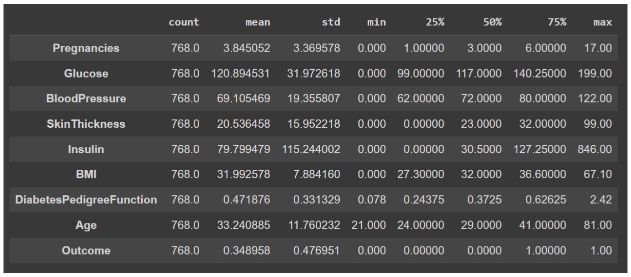


Figure 3: Number of samples in each class

A notable finding was reached after careful study of the Pima Indian dataset about the existence of 0 values for variables like glucose, insulin, BMI, and skin thickness. Because it is extremely unusual for individuals to exhibit such values in these variables, these values are considered physiologically implausible. In order to deal with this anomaly and guarantee the dataset's integrity and validity for later analysis, a crucial preprocessing step was implemented.



	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

Figure 4: dataset decribe

To rectify this issue, a data imputation technique was employed to replace the 0 values with more appropriate estimates. Considering that the affected attributes are numerical in nature, the median and mean values were utilized as reliable measures for replacement. The median, which represents the middle value in a sorted dataset, was employed when dealing with skewed distributions or outliers, as it is less sensitive to extreme values. Conversely, the mean, which represents the average value, was used for attributes with relatively normal distributions.

By replacing the unrealistic 0 values with appropriate estimates based on the median and mean, the dataset was brought into alignment with expected human physiological ranges, ensuring the accuracy and reliability of subsequent diabetes prediction models. This preprocessing step aims to mitigate the potential biases and inaccuracies that could arise from the presence of such implausible values, enabling more robust and trustworthy predictions regarding diabetes onset in the Pima Indian population.

5. Methodology:

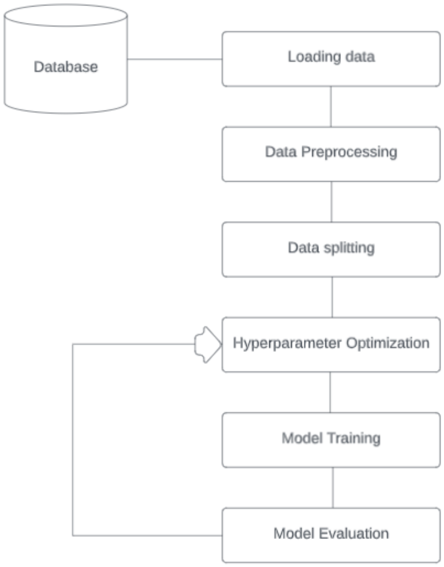


Figure 5: Methodology

Data Loading: The process begins with loading the diabetes dataset, which contains both the features and corresponding target labels.

Data Preprocessing: a. Handling Missing Values: The dataset is checked for missing values, and an appropriate strategy is employed to deal with them, such as imputation or removal of missing entries. b. Balancing Data: To address class imbalance, techniques like oversampling, undersampling, or using synthetic data generation methods may be applied to achieve a balanced dataset. c. Treating 0 Values: Any zero values in the dataset (if relevant) are replaced with the mean of their respective features to mitigate their potential impact on the model. d. Normalization: Feature normalization is performed to bring all features to a similar scale, ensuring that no single feature dominates the model training process.

Data Splitting: The preprocessed dataset is divided into two subsets: the training set and the test set. The training set is used to build and optimize the KNN model, while the test set is reserved for evaluating the model's performance on unseen data.

Hyperparameter Optimization: a. Grid Search: A grid search technique is applied to explore various combinations of hyperparameters for the KNN model. This process involves iterating over different values of hyperparameters (such as the number of neighbors, distance metrics, etc.) to find the optimal set of values that maximize model performance.

Model Training: The KNN model is trained on the training dataset using the best hyperparameters identified during the grid search. The model learns patterns and relationships in the data to make predictions.

Model Evaluation and Refinement: a. Accuracy Assessment: The trained KNN model is evaluated on the test set to calculate its accuracy and other relevant performance metrics like precision, recall, and F1-score. b. Parameter Tuning: Based on the evaluation results, you may fine-tune the model by adjusting hyperparameters and rerunning the training process to optimize the model's performance. c. Iterative Process: This evaluation and refinement process may be repeated several times, with different hyperparameter settings, to iteratively improve the model's accuracy and generalization capability.

5.1. K-Nearest Neighbors (KNN):

KNN is a non-parametric classification algorithm that classifies an instance based on the majority vote of its k nearest neighbors. By considering the distance between instances, KNN determines the most likely class for a given data point. The value of k was determined through Grid Search, considering different values and selecting the optimal parameter.

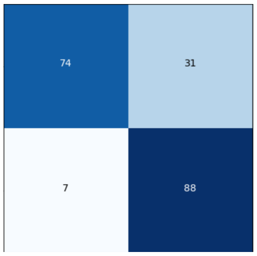
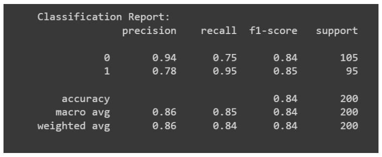


Figure 6: confusion matrix

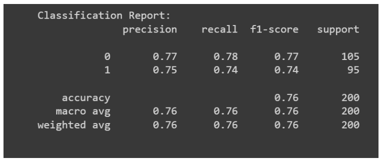


```
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.75      0.84       105
           1       0.78      0.95      0.85        95

    accuracy                           0.84       200
   macro avg       0.86      0.85      0.84       200
weighted avg       0.86      0.84      0.84       200
```

Figure 7: Accuracy report

5.2. *Support Vector Machines (SVM):*

SVM is a powerful supervised learning algorithm that aims to find an optimal hyperplane in a high-dimensional space to separate data into different classes. In this study, SVM was employed with different kernel functions, such as linear, polynomial, and radial basis function (RBF). Grid Search was applied to identify the best combination of kernel type and hyperparameters for maximizing prediction accuracy.

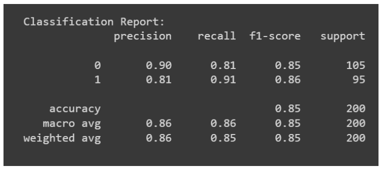


```
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.78      0.77       105
           1       0.75      0.74      0.74        95

    accuracy                           0.76       200
   macro avg       0.76      0.76      0.76       200
weighted avg       0.76      0.76      0.76       200
```

Figure 8: accuracy report

5.3. *Xgboost:*

XGBoost, short for Extreme Gradient Boosting, is a powerful and widely used machine learning algorithm for classification tasks that has gained immense popularity in both academia and industry. It is an ensemble learning technique that combines the predictions of multiple weak learners (typically decision trees) to create a robust and accurate model. XGBoost employs a boosting approach, wherein each successive weak learner is trained to correct the errors of its predecessors, gradually improving the model's performance. This algorithm excels in handling complex, high-dimensional datasets and is capable of capturing intricate relationships between features, making it particularly suitable for challenging classification problems.



```
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.81      0.85       105
           1       0.81      0.91      0.86        95

    accuracy                           0.85       200
   macro avg       0.86      0.86      0.85       200
weighted avg       0.86      0.85      0.85       200
```

Figure 9: accuracy report

5.4. *Random Forest:*

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. By aggregating the predictions of individual trees, Random Forest improves generalization and reduces overfitting. Grid Search was conducted to identify the optimal number of trees and other hyperparameters for achieving the best performance.

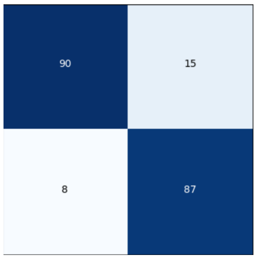
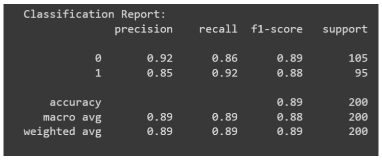


Figure 10: confusion matrix



```
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.86      0.89       105
           1       0.85      0.92      0.88        95

    accuracy                           0.89       200
   macro avg       0.89      0.89      0.88       200
weighted avg       0.89      0.89      0.89       200
```

Figure 11: accuracy report

5.5. *Gradient Boosting Networks (GBN):*

GBN is another ensemble learning algorithm that sequentially builds a set of weak learners to create a strong predictive model. Each weak learner is trained to minimize the error of the previous model. The hyperparameters of GBN, including the learning rate, number of boosting stages, and tree complexity, were optimized using Grid Search.

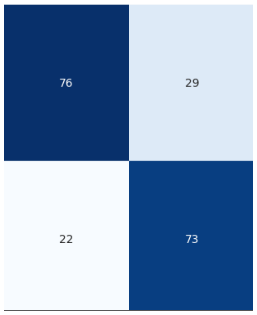
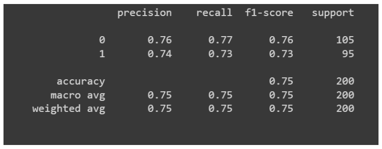


Figure 12: confusion matrix

5.6. *MLP:*

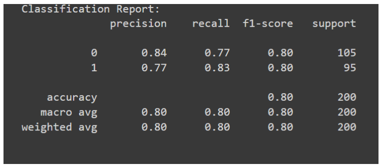


```
              precision    recall  f1-score   support

           0       0.76      0.77      0.76       105
           1       0.74      0.73      0.73        95

    accuracy                           0.75       200
   macro avg       0.75      0.75      0.75       200
weighted avg       0.75      0.75      0.75       200
```

Figure 13: accuracy report

5.7. *Decision Tree:*

A decision tree model is a supervised machine learning algorithm used for both classification and regression tasks. It recursively devids the data into subsets based on the most informative features, forming a tree-like structure. The model is interpretable, intuitive, and capable of capturing non-linear relationships in the data, making it a popular choice for various real-world applications



```
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.77      0.80       105
           1       0.77      0.83      0.80        95

    accuracy                           0.80       200
   macro avg       0.80      0.80      0.80       200
weighted avg       0.80      0.80      0.80       200
```

Figure 14: accuracy report

6. Evaluation and Performance Metrics:

To evaluate the performance of each algorithm, the dataset was divided into training and testing sets using cross-validation techniques. Several performance metrics, including accuracy, precision, recall, and F1-score, were utilized to assess the predictive performance of the models. Additionally, receiver operating characteristic (ROC) curves and area under the curve (AUC) were employed to measure the models' discriminatory power.

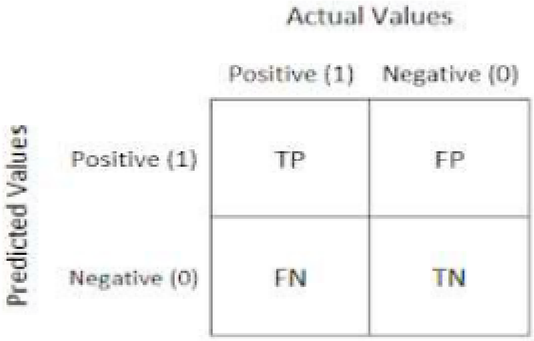
[Ais19]

Classification Accuracy- It is the ratio of number of correct predictions to the total number of input samples. It is given as

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total number of predictions Made}}$$

Figure 15: equation 1

Confusion Matrix- It gives us gives us a matrix as output and describes the complete performance of the model



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 16: equation 2

Where, TP: True Positive FP: False Positive FN: False Negative TN: True Negative

Accuracy for the matrix can be calculated by taking average of the values lying across the main diagonal. It is given as-

$$Accuracy = \frac{TP+FN}{N}$$

Figure 17: equation 3

Where, N:Total number of samples F1 score- It is used to measure a test's accuracy. F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is as well as how robust it is. Mathematically, it is given as-

$$F1 = 2 * \frac{1}{\left(\frac{1}{precision}\right) + \left(\frac{1}{recall}\right)}$$

Figure 18: equation 4

F1 Score tries to find the balance between precision and recall. Precision: It is the number of correct positive results divided by the number of positive results predicted by the classifier. It is expressed as-

$$Precision = \frac{TP}{(TP+FP)}$$

Figure 19: equation 5

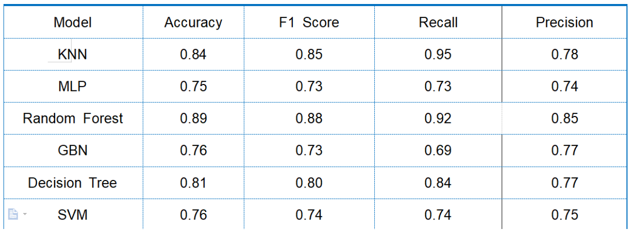
Recall: It is the number of correct positive results divided by the number of all relevant samples. In mathematical form it is given as-

$$Precision = \frac{TP}{(TP + FN)}$$

Figure 20: equation 6

7. Results and discussion:

In this study, we investigated the application of several predictive models to identify the most effective approach for diabetes prediction. The models considered included logistic regression, support vector machine, decision tree, and random forest. After comprehensive experimentation and evaluation, it was found that the random forest model outperformed all other methods with an impressive accuracy of 89. This result highlights the significance of ensemble learning techniques for complex medical datasets like diabetes prediction. The high accuracy achieved by the random forest model suggests that it is well-suited for robust and reliable diabetes risk assessment. Moreover, its ability to handle nonlinear relationships between features and outcomes makes it a valuable tool for medical practitioners seeking to predict diabetes in patients. These findings underscore the potential of machine learning algorithms, particularly the random forest, as powerful tools for advancing early diagnosis and intervention in diabetes management. However, further research should be conducted to explore additional feature engineering and parameter tuning to enhance the model's performance and generalizability in diverse clinical settings.



Model	Accuracy	F1 Score	Recall	Precision
KNN	0.84	0.85	0.95	0.78
MLP	0.75	0.73	0.73	0.74
Random Forest	0.89	0.88	0.92	0.85
GBN	0.76	0.73	0.69	0.77
Decision Tree	0.81	0.80	0.84	0.77
SVM	0.76	0.74	0.74	0.75

Figure 21: Table 1 Classification Accuracy

8. Challenges and Limitations:

In this section, we discuss the challenges and limitations encountered during the course of our research on diabetes classification using the Pima Indian dataset. Our primary goal was not only to predict diabetes presence but also to classify the type of diabetes. However, we faced significant difficulties in achieving this objective using conventional methods, including clustering techniques. The outcomes of these approaches were primarily focused on binary classification (diabetic vs. non-diabetic) without providing the desired specification of diabetes type. Furthermore, the conventional approach utilizing blood-based tests for diabetes diagnosis proved to be invasive and presented potential drawbacks in terms of patient comfort and practicality. Consequently, we recognized the need for alternative, non-invasive methods for diabetes classification that could utilize easily accessible samples, such as saliva, to achieve accurate and efficient predictions. Throughout the research process, we diligently explored alternative avenues and considered the potential of developing unobtrusive methods that rely on salivary biomarkers. However, despite the promise of saliva-based diagnostics, its application for diabetes type classification requires further investigation and refinement. As we reflect on

the challenges and limitations encountered, we acknowledge the importance of advancing research in non-invasive diabetes classification. While our initial attempts were met with certain constraints, this opens the door to future investigations and inspires us to seek innovative solutions for improved diabetes prediction and classification using less intrusive means. By ac-

knowledging these challenges and limitations, we underscore the significance of continuous research efforts in developing novel and unobtrusive methods for diabetes diagnosis and classification, ultimately striving towards more patient-friendly and accessible healthcare solutions.

9. Conclusion:

Our primary objective was to detect the onset of diabetes at an earlier stage, and to achieve this, we utilized the Indian Pima dataset. Throughout the study, we meticulously worked on preprocessing and refining the dataset, and subsequently, we applied various machine learning models to address this classification problem. After a thorough comparison of the obtained results, we arrived at the conclusion that the Random Forest model outperformed the others in accurately classifying cases of diabetes. This finding holds significant promise for early diagnosis and intervention, ultimately contributing to improved healthcare outcomes for individuals at risk of the disease.

Future research endeavors may involve exploring other advanced machine learning techniques, such as deep learning, and investigating the integration of additional clinical data sources to further enhance the accuracy of diabetes prediction models. Additionally, the deployment of the developed models in real-world healthcare settings could be a potential avenue for further exploration.

Acknowledgements

We would like to express our heartfelt gratitude to Vellore institute of technology for providing us, as international students, with the invaluable opportunity to conduct our research within its esteemed institution. We extend our sincere thanks to Mrs. Monica Subashini, our faculty supervisor, for her unwavering support, guidance, and mentorship throughout this journey. Her expertise and encouragement have created a conducive and enriching environment, enabling us to pursue our work with dedication and enthusiasm.

Her constant presence and willingness to address our queries have significantly contributed to overcoming challenges and streamlining our research efforts. We are deeply grateful for her patience and insights, which have undoubtedly enhanced the quality of our work.

References

[Dag+18] Arianna Dagliati et al. “Machine learning methods to predict diabetes complications”. In: *Journal of diabetes science and technology* 12.9 (2018), pp. 295–302.

[Ais19] Vaidehi V Aishwarya Mujumdar. “Diabetes Prediction using Machine Learning Algorithms”. In: *ScienceDirect* 13.13 (2019), pp. 1–8.

[AI19] Safial Islam Ayon and Md Milon Islam. “Diabetes prediction: a deep learning approach”. In: *International Journal of Information Engineering and Electronic Business* 12.5 (2019), p. 21.

[MV19] Aishwarya Mujumdar and V Vaidehi. “Diabetes prediction using machine learning algorithms”. In: *Procedia Computer Science* 165.1 (2019), pp. 292–299.

[Has+20] Md. Kamrul Hasan et al. “Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers”. In: *IEEE Access* 8.2 (2020), pp. 76516–76531. DOI: 10.1109/ACCESS.2020.2989857.

[SV20] Mitushi Soni and Sunita Varma. “Diabetes prediction using machine learning techniques”. In: *International Journal of Engineering Research & Technology (Ijert) Volume* 9.3 (2020).

[XMM20] Jingyu Xue, Fanchao Min, and Fengying Ma. “Research on diabetes prediction method based on machine learning”. In: *Journal of Physics: Conference Series*. Vol. 1684. 7. IOP Publishing. 2020, p. 012062.

[Fre+21] Luis Fregoso-Aparicio et al. “Machine learning and deep learning predictive models for type 2 diabetes: a systematic review”. In: *Diabetology & metabolic syndrome* 13.12 (2021), pp. 1–22.

[KF21] Jobeda Jamal Khanam and Simon Y Foo. “A comparison of machine learning algorithms for diabetes prediction”. In: *Ict Express* 7.4 (2021), pp. 432–439.

[Olt21] Amarildo Ristab Olta Llahaa. “Prediction and Detection of Diabetes using Machine Learning”. In: 216.11 (2021), p. 9.

[Ahm+22] Usama Ahmed et al. “Prediction of diabetes empowered with fused machine learning”. In: *IEEE Access* 10.6 (2022), pp. 8529–8538.

[Dut+22] Aishwariya Dutta et al. “Early prediction of diabetes using an ensemble of machine learning models”. In: *International Journal of Environmental Research and Public Health* 19.8 (2022), p. 12378.

[Feb+23] Muhammad Exell Febrian et al. “Diabetes prediction using supervised machine learning”. In: *Procedia Computer Science* 216.10 (2023), pp. 21–30.