CSE/ECE 343 : Machine Learning Mid-Project
Title: Diabetes Prediction using Machine Learning

Pragathi Vuppu vuppu22604@iiitd.ac.in
Bhanu Shandilya bhanu22134@iiitd.ac.in
Kanak Yadav kanak22611@iiitd.ac.in
Neelu neelu22318@iiitd.ac.in

Friday 29th November, 2024

1 ABSTRACT

Diabetes is a long term disease that has its prevalence in the region of individuals globally, and as a result, people should receive timely and accurate diagnosis. The old methods or models used for diagnosis depend on invasive procedures, which are not so efficient. In our project we try to resolve this problem by applying Machine Learning techniques to predict whether a person is diabetic or not. In order to do that we have worked with patients data containing clinical and demographic information, the dataset is present publically and we have generated a dataset from our side also in order to build a robust model. The models we have used by far are KNN, Decision Tree, Random Forest, Logistic Regression and able to achieve Validation accuracy of 88 % by Random Forest. To give an overview of what we have done so far is that we work on the common dataset present publicly and we tackle some problems like missing values , poor correlation between features, outliers but we handle these problems and do some manual feature engineering by introducing some new features like NewInsulin, NewGlucose,etc which actually make our models more consistent and robust. Our main motive is to make an accurate and robust model that actually can replace these invasive methods which may be costly for some people.

2 INTRODUCTION

Diabetes is a disease that affects the hormone insulin, which leads to abnormal carbohydrate metabolism and increased blood sugar levels. High blood sugar can adversely impact the blood vessels and nerves and some other organs which will eventually result in complications throughout the body. Reasons behind the diabetes remain unclear or uncertain as some scientists believe that both genetics and environment plays a key role. According to WHO the number of people with diabetes has surpassed 800 Million globally which is 4 times more than the 1990s data. Early detection of diabetes is the final solution in order to manage the disease and save lives. Our research and project focuses on the early prediction of diabetes by analyzing the risk associated with it and what type of trend it follows. For this study we gathered a diagnostic dataset having 8 attributes. But in our previous work we see that 8 attributes are actually not enough to build an accurate and robust model, So after feature engineering we got 15 attributes which actually improves model performance and in the next part of the project we use more than 20 relevant features and will show the improvement. In this project we worked upon a generalised way which included Both Male and Female while in our previous work we only worked with a dataset containing Female only.Machine Learning techniques provide well-organized results to extract knowledge by making prediction models from diagnostic medical datasets composed of diabetic patients. Though it is difficult to select the best techniques to predict based on such attributes, thus for the determination of the study different algorithms have been used for the model prediction.

3 PROBLEM STATEMENT

The major challenge in predicting diabetes cases is its detection. There are instruments available that can predict diabetes but either they are expensive or are not efficient to calculate the chance of diabetes in humans. Early detection of diabetes can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more patience, time, and expertise. Since we have a good amount of data in today's world, we can use various machine-learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

4 MOTIVATION

Machine learning techniques have been around us and have been compared and used for analysis for many kinds of data science applications. This project is carried out with the motivation to develop an appropriate computer-based system and decision support that can

aid in the early detection of diabetes, in this project we have developed a model that classifies if the patient will have diabetes based on various features (i.e. potential risk factors that can cause diabetes) using random forest classifier. Hence, the early prognosis of diabetes can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine.

5 LITERATURE SURVEY

Before diving in creating machine learning models for diabetes prediction, it is necessary reviewing what research has already been done on this matter. Researchers have been looking for different algorithms from traditional structures (Logistic Regression, SVM) to more advanced ones such as Ensemble Learning and Deep Learning. These works concentrate on get higher accuracy to predict specific disease with respect of imbalanced data, reduce a raw dimension scale and handle complex clinical datasets. The first paper, "Diabetes Prediction Using Machine Learning Classification Algorithms", focuses on predicting diabetes using the Pima Indian Diabetes Dataset containing 768 sam- ples with 9 features such as Glucose, BMI, and Insulin. The authors addressed data preprocessing challenges, such as zero or invalid values, by replacing them with mean or median values and normalized the dataset for consistency. They used multiple machine learn- ing models, including Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Decision Tree (DT). Of these, Random Forest reached the high- est accuracy of 88.31 lowed by ANN with 86 lights RF's effectiveness in handling complex relationships in data and emphasizes the importance of feature engineering for improving model performance. The sec- ond paper, "Diabetes Prediction Using Machine Learn- ing and Explainable AI Techniques", utilizing both the Pima Indian Diabetes Dataset and a private dataset (RTML) of 203 samples from Bangladesh, which did not contain the Insulin feature. The authors have used a semi-supervised XGBoost regressor to predict missing val- ues and handled class imbalance using SMOTE and ADASYN techniques. They have used various classifiers, includ- ing Logistic Regression, Random Forest, SVM, KNN, AdaBoost, and XGBoost, and evaluated the same based on accuracy, precision, recall, and F1-score. The XG- In addition, Boost model performed best achieving 81 % accuracy and AUC of 0.84. The paper also combined explainable AI tools to interpret feature importance, and Glucose and BMI were found to be key predictors.

6 DATASET

The dataset contains **253,680 samples** and **22 features**, designed for diabetes prediction. The target variable, `Diabetes_binary`, is a binary indicator for diabetes presence (0 or 1). Features include health-related attributes such as `HighBP`, `HighChol`, `BMI`, and physical/mental health days, as well as lifestyle factors like smoking, physical activity, and diet. Demographic variables such as `Age`, `Sex`, `Education`, and `Income` are also included. Missing values, represented by 0, were handled using the median, and outliers were addressed using the IQR method and Local Outlier Factor. The dataset's diversity and size make it suitable for robust machine learning modeling in diabetes prediction tasks.

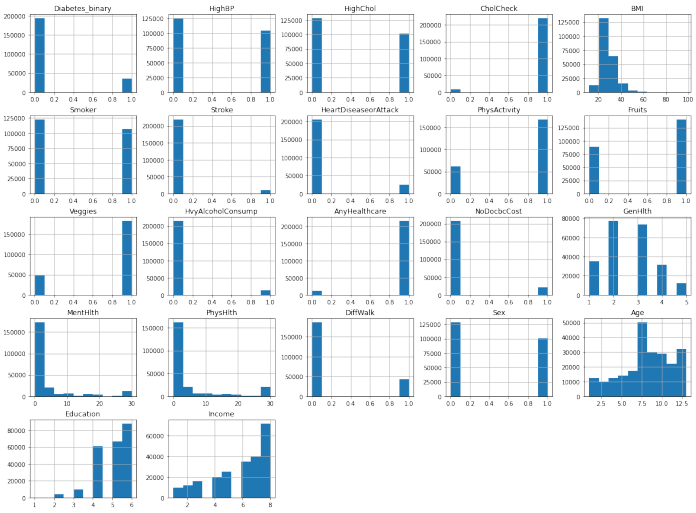


Figure 1: Histogram for different features

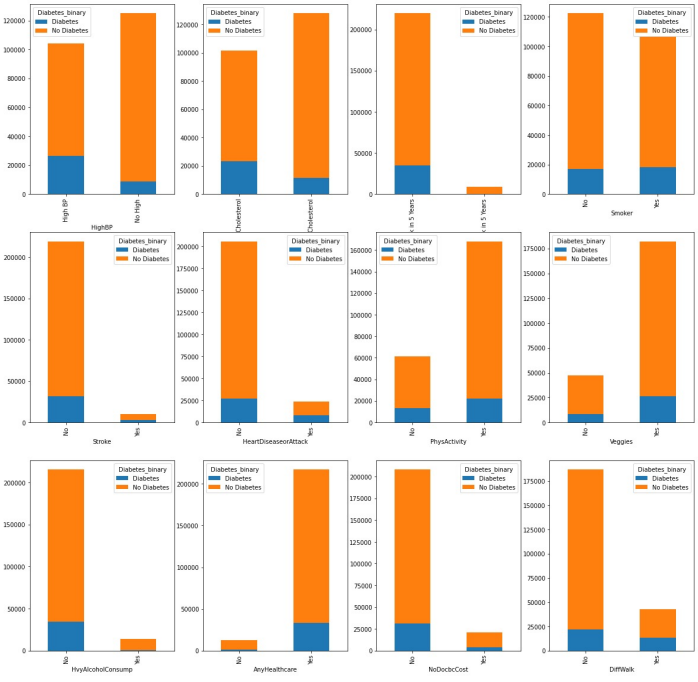
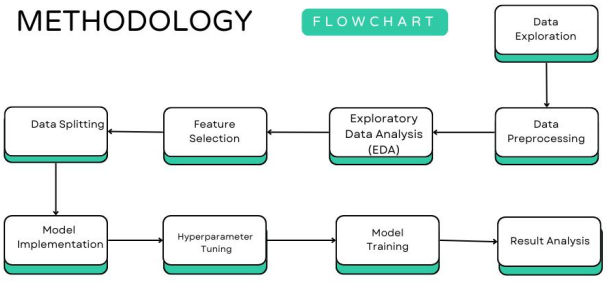


Figure 2: Individual Feature Relation with Diabetes Binary i.e label

7 METHODOLOGY

The methodology adopted for this project systematically handled data preprocessing, analysis, feature selection, and model training to ensure robust diabetes prediction. Initially, the dataset contained 22 features with 253,680 non-null values. Descriptive statistics such as mean, standard deviation, and minimum values were computed to understand the data distribution. In preprocessing, null values were removed, unique values in categorical variables were checked, and outliers were identified and addressed. Duplicate records were removed, and the cleaned dataset was referred to as dataset2. A duplicate of this dataset, data2, was created for handling string-based values. Relationships between all attributes and the target variable (Diabetes binary) were explored through visualizations, including a heatmap for correlation analysis. For feature selection, Variance Inflation Factor (VIF), ANOVA, and Chi-square tests were applied to identify the most significant predictors. The cleaned data was then split into training and testing sets. Techniques such as oversampling or undersampling were employed to handle class imbalances, followed by scaling the features for consistency. Multiple machine learning models were implemented, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machines (SVM), XGBoost, and K-Nearest Neighbors (KNN).We use MLP in because we want to see how the models which do not requires the feature extraction can perform. Each model underwent hyperparameter tuning to enhance performance before final training and evaluation. This structured approach ensured optimal use of features and robust predictions. At the end we see the losses and accuracy based on this we get SVM and XGBoost as the best model and finally we do ensemble learning and re-evaluate based on this.

7.1 FLOWCHART OF METHODOLOGY



7.2 MODEL DETAILS

Four machine learning algorithms were selected for experimentation. First one is K-Nearest Neighbors (KNN), a simple yet effective method that classifies data points based on the majority class of their nearest neighbors. The second algorithm is Random Forest, an ensemble learning method that creates multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. Decision Tree which generally prone to overfit but after tuning some parameter the tree better generalize the data. Finally, Logistic Regression, a statistical model that estimates the probability of a binary outcome, was applied to predict whether a patient is diabetic.SVM finds an optimal hyperplane to separate classes, using kernels to handle non-linear data.XGBoost is a fast, regularized ensemble method that builds trees iteratively to minimize errors.

8 RESULT AND ANALYSIS

Table 1 shows that Random Forests achieved the highest validation accuracies before and after feature engineering. Logistic Regression improved significantly with proper scaling and feature selection. While Decision tree and KNN also improved but not so significantly.

Model	Accuracy	MSE
Decision Tree	84.80%	0.1519
Random Forest	85.88%	0.1411
Logistic Regression	84.72%	0.1528
KNN	80.53%	0.1947
SVM	86.03%	0.1397
XGBoost	86.63%	0.1336
MLP	85%	0.30

Table 1: Highest Testing Accuracy Before and After Feature Engineering.

8.1 ANALYSIS

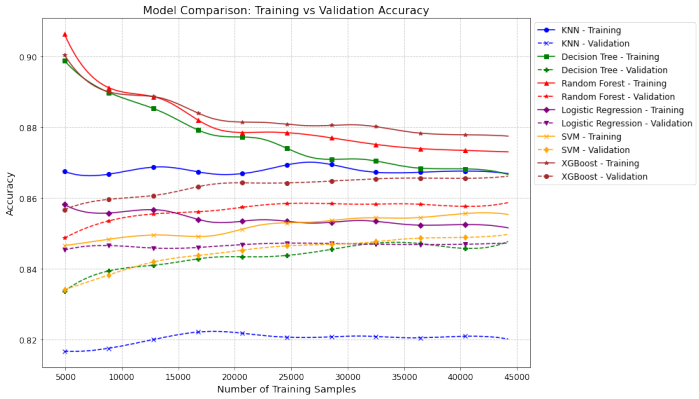


Figure 3: Impact of Training Set Size on Model Performance before Feature Extraction

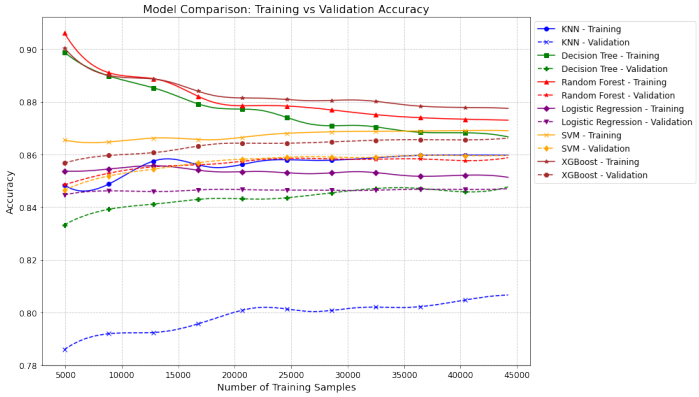


Figure 4: Impact of Training Set Size on Model Performance after Feature Extraction

8.1.1 GRAPH COMPARISON

First compare the graph Fig 2 and Fig 3 in this we can clearly see that before feature extraction and after feature extraction there is an improvement in performance and XGBoost performs better in both cases but other then KNN all models shows improvement. Now we compare the Fig 1 with Fig 1 and 2, here we clearly see that despite different dataset Random Forest ables to capture the underlying pattern from which we can conclude that if other dataset on diabetes comes, Random Forest will be a robust model.

9 CONCLUSION

9.1 LEARNING FROM THE PROJECT

The use case was diabetes prediction where practical implications of applying machine learning into healthcare were realized by incorporating data preprocessing techniques of managing missing values, removing outlying records, as well as duplicate records which help in improving the models. Techniques such as encoding ensure that the model generalizes well. Feature engineering emerged as a game-changing exercise, where domain-related feature inclusion such as NewInsulin and NewGlucose provided substantial improvements in prediction performance. Statistical tests-ANNOVA and Chi-Square, along with correlations helped in choosing more informative predictors.Model selection and tuning played a central role in balancing complexity and accuracy. Testing through various algorithms, starting from the simplest Logistic Regression to the more complex models such as Random Forest and XGBoost, showed strengths and weaknesses of each. Hyperparameter tuning further improved the robustness and generalizability of the models.Comparing models before and after feature engineering highlighted the importance of well-curated features, where XGBoost consistently performed best and Random Forest showed versatility across datasets.The need for diversification and a reasonable amount of data for good generalization was again highlighted by studying the effects of training set sizes. Oversampling techniques and ensemble approaches were novel solutions to challenges like class imbalance and scalability. Balance between interpretability and performance was critical because predictions could directly influence the outcome for patients in healthcare.Collaboration among the team members promoted diverse approaches to problem-solving and efficient workflows. Visualizations became very important in communicating findings and progress. Beyond the technical advancements, this project brought out the transformative nature of machine learning in the healthcare sector. It was in this regard that the project emphasized the need for early diagnosis and accurate predictions to improve patient care and reduce complications. These learnings have strengthened our understanding of leveraging data-driven approaches to create impactful solutions in medicine.

9.2 CONTRIBUTION OF EACH MEMBER

We ensured proper work distribution to achieve the best possible results in the project in the most efficient way. Bhanu Shandilya took care of the data collection and data visualization part as he suggested IQR to detect outliers. He also contributed to feature engineering, KNN model implementation, and created the graph based on the size of the dataset vs. accuracy, observing that overfitting decreases. Additionally, Bhanu implemented the SVM model and analyzed its perperformance. Pragathi Vuppu handled reading material (articles and research papers) and suggested the concept of feature engineering. She also implemented feature engineering with Bhanu, implemented the logistic regression model, and proved that models converge early while the ac- curacy remains the same. Pragathi also implemented the Naive Bayes model and compared its results with other models. Kanak Yadav contributed to data visualization by detecting and removing outliers. She handled data preprocessing, especially handling missing values. She implemented the Decision Tree, built a graph of increasing model complexity, and studied how the model's performance changes. Ad- ditionally, Kanak implemented the MLP model, tuning its parame- ters to optimize accuracy. Neelu took care of the data preprocessing part, such as standardizing values using RobustScaler, which is robust to outliers, and handling categorical features using one-hot encoding. She implemented the Random Forest model, checked the accuracy by change the number of estimators and graphed increasing model complexity to examine performance variations. The report was written collaboratively: Bhanu Shandilya wrote the Conclusion and Introduction, Kanak wrote the Methodology, and Model Details, Pra-gathi wrote the Literature Survey and Results and Analysis, and Neelu wrote the Abstract and Dataset Description.

10 REFERENCE

References

[1] D. Dutta, D. Paul, P. Ghosh, "Analyzing Feature Importance's for Diabetes Prediction using Machine Learning", *IEEE*, pp. 942-928, 2018.

[2] K. VijiyaKumar, B. Lavanya, I. Nirmala, S. Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes", *Proceedings of International Conference on Systems Computation Automation and Networking*, 2019.

[3] Md. F. Faruque, Asaduzzaman, I. H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus", *International Conference on Electrical, Computer and Communication Engineering (ECCE)*, February 7-9, 2019.

[4] G. Tripathi, R. Kumar, "Early Prediction of Diabetes Mellitus Using Machine Learning", *2020 8th International Conference on Reliability Info.*

[5] Author(s), "Title of the project report", *Institution*, Year. Retrieved from https://www.ssgmce.ac.in/uploads/UG_Projects/cse/Gr%20No-10-Project-Report.pdf.

[6] S. Kumar, A. S. Khurana, "Diabetes Prediction Model Using Machine Learning Algorithms", *Jaypee University of Information Technology*, Retrieved from http://www.ir.juit.ac.in:8080/jspui/bitstream/123456789/9888/1/Diabetes%20Prediction%20Model.pdf

[7] Tasin, I., Nabil, T.U., Islam, S., Khan, R.: Diabetes prediction using machine learning and explainable AI techniques. *Healthc. Technol. Lett.* 10, 1–10 (2023). They were able to achieve 88.8% accuracy using Ensemble (XGBoost).

[8] Tasin, I., Ullah, T., Islam, T., Khan, R. (Year). Diabetes prediction using machine learning and explainable AI techniques. Link https://pmc.ncbi.nlm.nih.gov/articles/PMC10107388/pdf/HTL2-10-1.pdf

[9] Nahzat, Shamriz Yaganoglu, Mete. (2021). Diabetes Prediction Using Machine Learning Classification Algorithms. European Journal of Science and Technology. 53-59. 10.31590/ejosat.899716.