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

Wednesday 20th 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. Old-fashioned diagnostic models depend on invasive procedures, and are usually inefficient. In this work we apply machine learning techniques to estimate the probability of diabetes occurrence according to the patients' data containing clinical and demographic information. As part of this study, we hope to use a combination of commonly accessed public datasets which will facilitate the development of machine learning models including Logistic Regression, Decision Trees, Random Forest, SVM so as to help healthcare providers in early diagnosis of the disease. To compare the models and identify the best approach the performance indicators inclusive of accuracy, precision, recall, and F1-score are used. The study evidences that the static and dynamic features as insights of the machine learning models can greatly improve the detection and can be a non-invasive, cost-effective addition for diabetes prediction with improved patient's outcome.

2 INTRODUCTION

Diabetes is a disease that affects the hormone insulin, follow-on in abnormal metabolism of carbohydrates, and advanced steps of sugar in the blood. This great blood sugar affects several organs of the human body which in turn complicates many sources of the body, in precise the blood strains and nerves. The details of diabetes are not nevertheless totally exposed, many researchers supposed that both the hereditary elements and environmental effects are complex therein. As exposed by the International Diabetes Federation[1], the extent of people having diabetes stretched 422 million out of 2021 that makes up 5.34 percent of the world's total adult population. Early prediction of such diseases can be controlled over the diseases and save human life. To accomplish this goal, this research work mainly discovers the early prediction of diabetes by taking into account various risk factors related to this disease. For the willpower of the study, we gathered a diagnostic dataset having 16 attributes diabetic of different patients. Later, we debate about these attributes with their conforming values. Based on these attributes, we figure a prediction model by means of various machine learning techniques to predict diabetes. 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

Prior to 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. A paper from Frontiers in Genetics (2023) discusses versus XGBoost, similarly to the paper above as well but looks at ensemble learning methods for diabetes detection. A key finding was that ensemble methods outperformed conventional models, particularly in the context of clinical and genetic data. The authors also highlighted the need for adding clinical biomarker features like BMI, Glucose and genetic predispositions to improve prediction led by biological prior knowledge. Article written at NCBI (2023): This article from the house of all PubMed articles includes machine learning models with clinical pipelines for predicting potential diabetics and subsequently making optimal decisions in treatment. The research evaluates a variety of algorithms like mean bag-of-words, SVM, logistic regression and many deep learning models. The 2 primary predictors for diabetes identified were fasting glucose level and BMI. So the two machine learning algorithms Gradient boosting and Random forests have been presented in this paper with good performance for diabetes predictions that can be used on massive medical datasets. Ensemble models, which aggregated multiple learning methods into a single model, achieved superlative accuracy in situational real-world clinical scenarios. Our results show that the model accuracy degrades over time signifying the importance of continuous monitoring and updating.

6 METHODOLOGY AND MODEL DETAILS

6.1 METHODOLOGY

When you actually see the data there is no missing values because all the missing values are 0 not Null values. So, our first task is to replace all zero's with the Null values then we plot the missing plot which actually tells that there are lots of missing value specially insulin with 374 missing values now the task is how to fill these values, for that we first plot the distribution curve and box plots and by observing them one thing is for sure that there a lots of outliers as well as the curves do not actually follows normal distribution because of which we can't replace these missing values with the mean value because mean is sensitive to outliers so instead we replace all missing values with the median values which are robust to outliers. Now we need to deal with the outliers for that we use IQR method which helps in identifying the outliers but there are some values that are still present as a outlier which then remove by using the Local Outlier finder. After that we normalize all values using the Robust Scaler which is again robust to outliers although we can use the standard scaling technique also. But when we plot graph for different data sizes the training and validation accuracy vary too much, that is model is not performing consistently because which we decide to do Feature Engineering we can also use PCA but to make our project different from others we did Feature Engineering by creating Two categorical columns i.e NewGlucose and NewInsulin which stores string values like normal, abnormal, prediabitic, etc. After that we use get dummies which is basically one hot encoding method that is used for categorical columns. After that we make correlation matrix which shows some good relationship between NewInsulin and Outcomes and NewInsulin and Outcomes. Based upon which we select manually the columns whose absolute value is 0.2 and then we do train test and split 80% data is used for training and 20% used for testing and then we again plot the same graph but this time every model performs somewhat consistently and robust to data sizes.

6.2 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.

7 RESULT AND ANALYSIS

7.1 RESULT

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	Before Feature Eng.	After Feature Eng.
Decision Tree	85.53%	87.50%
Random Forest	88.16%	88.16%
Logistic Regression	82.24%	87.50%
KNN	86.18%	87.50%

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

7.2 ANALYSIS

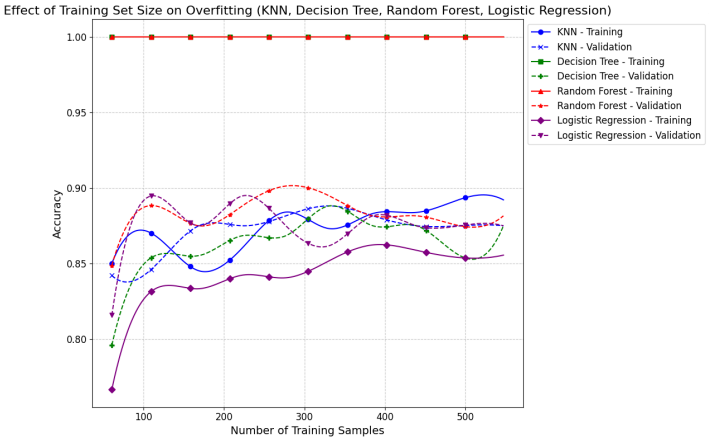

Figure 1: Impact of Training Set Size on Model Performance after Feature Engineering

The graph in Figure 1 shows how the accuracy changes for a model as the training dataset increases. The Random forest gives consistently high performance with little overfitting, as the figure suggested the stable training and validation accuracies. While Decision tree shows some overfitting with smaller dataset but as the size increases it shows strong performance. KNN shows a lot of fluctuations with increasing data size which indicates its sensitivity to data variations. But among all Logistic regression show the consistent trend with gradual improvement as training sample increase which shows its robustness to smaller data sizes.

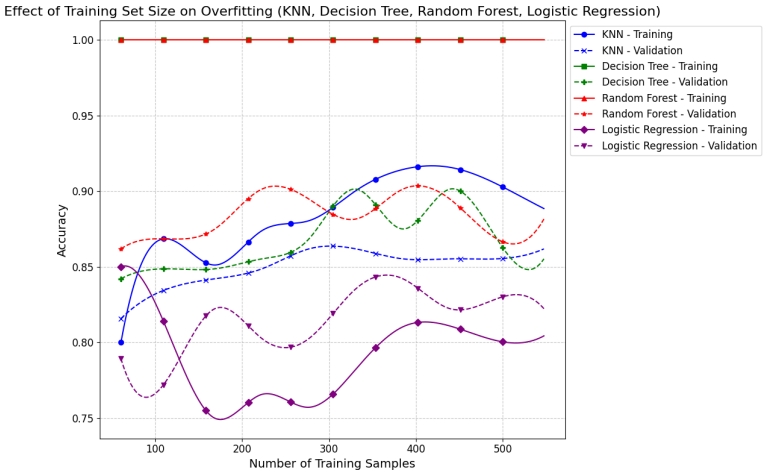

Figure 2: Impact of Training Set Size on Model Performance before Feature Engineering

The graph in Figure 2 shows how the accuracy changes for a model as the training dataset increases. KNN model shows a large fluctuations with consistently lower test accuracy then training accuracy with clearly shows significant overfitting. Decision tree model performs consistent but for smaller data set the trianing accuray is very higher then the validation accuracy which suggest that the model overfits on smaller dataset. Similar to decision tree, Random Forest model performs consistently with less fluctuations but it still overfitting on smaller dataset but also achieve the best validation accuracy among all models. Whereas the Logistic Regression models underfitting as its training accuracy is very low through out the data size and which means it is not able to capture the underlying patterns in data. Overall Random Forest gives best accuracy, Decision Tree and KNN are prone to overfitting on smaller dataset, Logistic Regression model prone to underfitting.

7.2.1 GRAPH COMPARISON

Both the graph suggest one thing for sure that Random Forest gives the best accuracy but when we talk about consistency it can be clearly seen that after feature engineering all models are somewhat consistent and less fluctuation can be seen which suggest the robustness of all models as the data size increases which is quite good if we train and validate our model on larger data set.

8 CONCLUSION

8.1 LEARNING FROM THE PROJECT

First let's talk about the dataset, just by looking at the dataset you can't find the missing values as for missing values the 0 is given but a computer machine considers it as a number (Integer) but it is a null value. So, 0 is converted to null values then replaced with median as the distribution does not allow replacement with mean. Then the outliers were detected using IQR (Interquartile range) method which is new for us but a good concept then we use Local outlier factor for outlier detection and removing them. Then performing our basic practice i.e normalizing the values and training using the models like Random Forest, Decision Tree, KNN, Logistic Regression. Where Random Forest got the highest accuracy of 88% which is quite good. While Decision Tree also performs quite similar to Random Forest but other models does not perform well so we perform the feature engineering by introducing categorical features like NewGlucose and NewInsulin which then added to dataset and best features were chosen based on correlation matrix basically features with greater then 0.2 correlation magnitude(i.e absolute value) with the 'Outcome' column. Then the model is again trained and surprisingly the accuracy of random forest is still same i.e 88% and small improvement in decision tree but after feature engineering the other two models(KNN and logistic regression) perform very well and there increases specially for logistic regression. Then two plots were made before and after feature engineering and these plot are based on increasing the test size how

the accuracy of model changes, with the help of these plots we can identify which model overfits and underfits.

8.2 WORK LEFT

Although the 88% highest accuracy is being achieved using Random Forest and Decision Tree, but as we have gone through many articles, we still think that we can increase the accuracy above 90. There are two models on which we haven't trained i.e SVM (Support Vector Machine) and Gradient Boosting. We hope that both models will perform better than the Random forest and Decision tree. But in order to choose one model we can't only reliable on the accuracy only, other factors needs to be taken in account like Precision and recall(Sensitivity) i.e confusion matrix, speed and computation efficiency, robustness and generalization, overfitting and regularization, hyperparameter tuning complexity (Gradient boosting, Random forest, etc) also helps us to choose the best model. Although we have done the feature engineering but still we will implement PCA reduction to see how the accuracy changes and how much our model is robust to outliers and data sizes.

8.3 CONTRIBUTION OF EACH MEMBER

We maintained proper work distribution for achieving 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 in feature engineering, KNN model implementation and he also made the graph based on the size of data set vs accuracy and observed overfitting decreases. Pragathi Vuppu took care of reading materials (articles and research papers) and proposed the idea of feature engineering and also implemented feature engineering with Bhanu. She implemented the logistic regression model and showed that models converge early, and the accuracy remains the same. Kanak Yadav helps in data visualization and detects the outlier and removes them, She does the data preprocessing specially handling the missing values, She implements the Decision Tree and makes a graph of increasing the model complexity and analyzes how the model performance changes. Neelu took care of the data preprocessing part like standardizing the values using RobustScaler which is robust to outliers and handling the categorical features using the one hot encoding. She implemented the Random forest and checked the accuracy by changing the estimators and made a graph of increasing the model complexity and analyzed how the model performance changes. The report part is written by everyone i.e Bhanu Shandilya write the conclusion part, Kanak write the Introduction and Methodology and model details, Pragathi write Literature survey and Result and analysis, Neelu write the Abstract and dataset related.

9 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] Aishwarya Mujumdar, Dr. Vaidehi V et al. (2019): Diabetes prediction using machine learning. They were able to achieve 96% accuracy using Logistic Regression. Further, this work can be extended to find how likely non-diabetic people are to have diabetes in the next few years.

[9] Abdulhadi N., Al-Mousa A. (2021): Diabetes detection using machine learning classification methods. In: 2021 International Conference on Information Technology (ICIT). IEEE, p. 350–354. They were able to achieve 82% accuracy with the Random Forest.