PREDICTIVE MODELING OF DIABETES USING MACHINE LEARNING TECHNIQUES

Team 1:M Nikhitha,CH Reshmasri,M N Madhurima

Abstract:

Diabetes prediction has been one of the promising approaches toward the improvement of nearly diagnoses and overall management of this chronic disease. This study describes an application of various machine learning algorithms, including Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines, for predicting the likelihood of diabetes in individuals. Using key health indicators like age, body mass index (BMI).The precision and recall of the test set are used while training and testing the models. Apart from that, normalization, handling missing values, feature selection, and other data preprocessing techniques have been used to improve the performance of the models. In this regard, the results show that machine learning algorithms can be highly supportive in identifying early stages of diabetes so that timely intervention with personalized healthcare solutions becomes possible.

Key Words:

Support Vector Machine , K-Nearest Neighbors, Machine Learning , DecisionTreeClassifier , LogisticRegression ,BMI.

1.Introduction:

Diabetes is a chronic disease, if it is left untreated then it leads to severe medical complications . Early prediction is important in managing the illness and preventing long-term consequences. This works involves using machine learning to classify people as either having diabetes or not based on data received from women who are Pima Indian and whose data are donated by the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset has attributes such as pregnancies, glucose level measurements, body mass index, and insulin levels. We rely on the classification algorithms below for predicting that a patient has diabetes: Logistic Regression, SVM, KNN, and Random Forest for prediction that a patient has diabetes. We assess these models with common performance metrics such as the accuracy, precision, recall, and F1 score to find the most accurate model of diabetes prediction.

Diabetes can further be managed by injection of insulin, the intake of a well-balanced diet, and timely exercise but there is no whole cure is available. Diabetes leads to much other disease such as blindness, blood pressure, heart disease, and kidney disease and nerve damage. There are three prime types of diabetes mellitus: Type 1 Diabetes Mellitus results from the body's failure to produce insulin. This form was previously referred to as insulin-dependent diabetes mellitus.Type2 Diabetes Mellitus conclusion from insulin resistance which is a condition in which cells are not able to apply insulin correctly, though sometimes also with an absolute deficiency of insulin. This form was named previously noninsulin-dependent diabetes mellitus. Gestational diabetes represents the third major form and appears when a pregnant-women. previously seems diagnosis of diabetes develop high blood glucose level. In order to automate overall process of diabetes prediction and severity estimation, diabetic database is needed.This diabetic database repository helps identify the various kinds of impacts diabetes possesses on human organs. The more accuracy its prediction holds, the greater chance one has for proper severity estimation.

2.Literature Survey:

Data science techniques may help other scientific disciplines to give a new perspective to common questions [1]. A Gaussian mixture model is used probabilistically, assuming that every data point can be generated from a mixture of a finite number of Gaussian distributions with unknown parameters. the artificial neural network is consisting of the layers and network function, the layers of the network are including: input layer, hidden layer and output layer. The input neurons def me all the input attribute values for the data mining model. The most effective model to predict patient with diabetes appear to be ANN followed by ELM and GMM.

Diabetes, a rapidly increasing chronic disease, necessitates effective diagnosis and management. This paper reviews current data mining techniques for diabetes prediction and diagnosis, focusing on early detection of conditions like hypo/hyperglycemia. It explores various data mining solutions, classifies and compares them based on key metrics, and discusses challenges and future research directions for improving diabetes prediction and management [2]. Detection of excessive glycemic variability by ANN. This review examines advanced techniques in data mining for diabetes diagnosis and prediction, focusing on glycemic control.

Data mining plays an efficient role in prediction of diseases in health care industry. Diabetes is one of the major global health problems [3]. By using Bayesian classifier patient is undergoing classified in classes of Pre-diabetic, Non-diabetic, Diabetic according to the attributes selected. The techniques they applied as preprocessing attribute identification and selection, data normalization. And then classifier is applied to the modified data set to construct the Bayesian model. The Bayesian network has a benefit of it encodes all variables, missing data entries can be handled successfully. They used the data set of Pregnant Women.

Simple neural network architecture were used to develop predictive model, model use input layer having 8 parameter value, one hidden layer with having 6 neuron and one output layer. It develop predictive model within small epoch value, simple interpreting the result and it improve the accuracy of the work [4]. j48 build classification using by training the model and validate the model using test cases. J48 is supervised learning algorithm. It is an extension of ID3 algorithm, has additional features like handling missing value, decision trees pruning, continuous attribute value ranges, derivation of rules.

Diabetes mellitus is a major global health issue, affecting millions, with increasing risk, especially in women [5]. KNN is a simple and also a lazy learning algorithm. It is one of classification algorithms used in health care. It can be used for both classification and regression. However, it is more widely used in classification problems. The algorithm is preferred mostly for its ease of interpretation. KNN identifies all the unidentified data points at one point using the existing datasets. The K-Nearest Neighbors (KNN) algorithm effectively classified and predicted diabetes diagnosis, with performance varying based on different values of the parameter 'k'.

The goal here is to discover new and useful patterns to provide meaningful and useful information for the users about the diabetes [6]. This algorithm begins by taking the training dataset as input. The entropy and information gain is calculated for all attributes of training dataset. The attributes are ranked based on the information gain. Splitting attribute is chosen based on the minimum entropy or maximum information gain from the ranked attributes to form subset of training data. This action is performed iteratively until no attribute remains other than the attributes selected before.

This concentrates on the part of Medical conclusion learning design through the gathered data of diabetes and to create smart therapeutic choice emotionally supportive network to help the physicians [7]. The primary target of this examination is to assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient's database. Naive Bayes uses almost a similar method that predicts the definite probability of different section based on various attribute values. This algorithm is generally useful in problems that have identical classes.

Diabetes, a leading global cause of mortality, has serious complications including kidney disease, vision loss, and heart disease. Early detection and accurate prediction are crucial for effective treatment and management. Datasets are obtained from Kaggle. Includes cleaning, mixing, and transforming the data [8]. Random Forest (RF) is employed for better precision. Data is used as training and test sets to develop a robust model. Relevant data is collected through a structured query process. Logistic Regression achieved the highest accuracy of 82.46%, while SVM had the lowest at 79.22%. Future research will explore additional classification algorithms to further improve prediction accuracy.

For the classification purpose, we adopt Naive Bayes (NB) for the feature selection of categorical type of data, which are gender and family history of diabetes [9]. The Naive Bayes algorithm builds a probabilistic model of the learning process the conditional probabilities of each input attribute given a possible value taken by the output attribute. Then, we adopt C4.5 for the rest of feature selection. We choose C4.5 algorithm, because it has the ability to support ordinal type of data and the decision tree construction rule of this algorithm are relatively easy to understand.

Proposed research will help in automating prediction of diabetes even before clinicians arrived. The current process of carrying this activity is manually which tends not to analyzing data flexible for the doctors, and transmission of information is not transparent. KDD is the procedure used to attain important and useful knowledge from large collection of collected data [10]. The process involves selecting, preparing and cleansing data from unneccesary Information. An application using a data mining algorithm of classes comparison has been developed to predict occurrence of or recurrence of diabetes risks. The naïve bayes classifierbased system is very useful for diagnosis of diabetes.

3.Proposed System:

The objective of our proposed system is to develop a machine learning-based system to accurately predict the likelihood of diabetes in patients using medical diagnostic data. The system will focus on building and comparing various classification models to determine which performs best in predicting diabetes based on several health-related features.

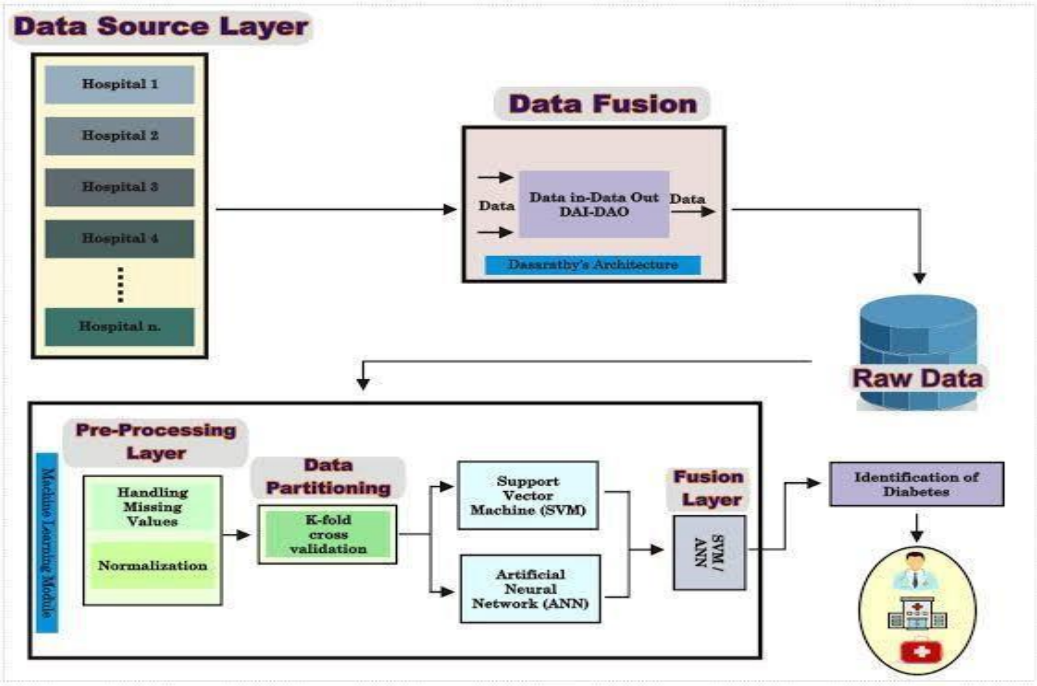

Fig1. Architecture of proposed Diabetes Prediction

Fig1. represents a classification system for diabetes based on data fusion with the aid of machine learning. Data is gathered from various hospitals and then merged using a technique called "Data in-Data out (DAI-DAO)" with Dassnathy's Architecture. Then, the fused data is saved as raw data. Then there comes pre-processing in which missing values are dealt with, and the data is normalized so that there is uniformity. Then the partitioning of the data is done with K-fold Crossvalidation, that splits the dataset into a train set and a test set. Two machine learning models are applied: Support Vector Machine (SVM) and Artificial Neural Network (ANN), to analyze the data; the results of those models are then fused in a final layer which goes on to identify diabetes. It makes use of how integration between hospital data and machine learning improves diabetes detection.

The Proposed Methods are:

- Logistic Regression
- SVC
- Random Forest
- KNeighbor Classification
- DecisionTreeClassifier
- GradientBoostingClassifier
- LGBMClassifier

3.1.Data Preprocessing:

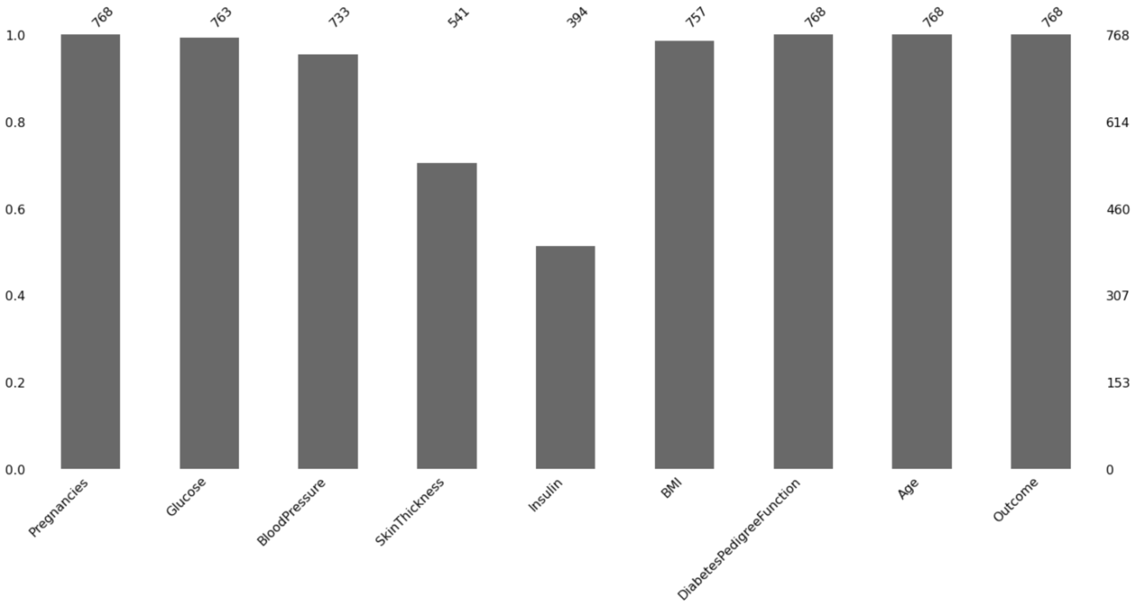

Fig2.Data Preprocessing

Fig2. represents handling of the missing values by replacing them with the median of the respective feature, especially for attributes like insulin levels and skin thickness where missing data is common. Normalize or scale the data, particularly for features with varying units such as age, BMI, and glucose levels, to ensure all features contribute equally to the model.

3.2.Exploratory Data Analysis:

Logistic Regression: The baseline model that's used to do binary classifications, calculates the probability of an outcome given input features.

Support Vector Machine (SVM): This is a very powerful algorithm that creates decision boundaries for separating classes in high-dimensional feature space.

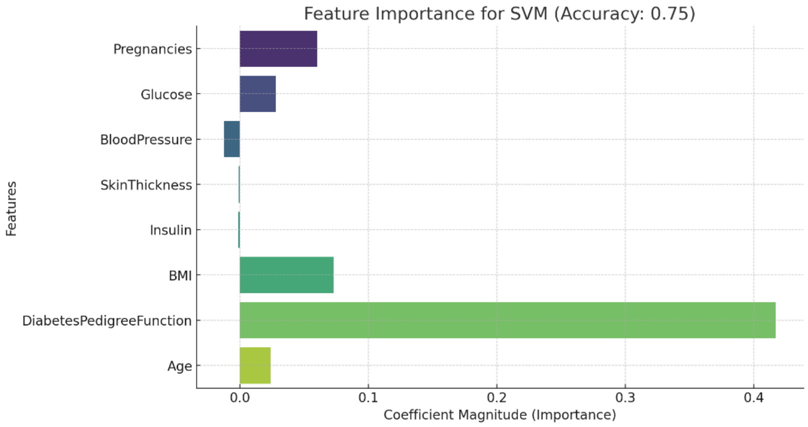

Fig3.Support Vector Machine

Fig3: below is the bar graph showing the feature importance for the SVM model using the linear kernel: DiabetesPedigreeFunction and BMI have the largest magnitudes of coefficients; that means they are the most influential in determining the outcome.The accuracy of SVM model on the proposed test set was 75%.

K-Nearest Neighbors (KNN):it's a very simple instance-based learning algorithm, which classifies.That is, it generates new data points based on the majority class of all of those among k-nearest neighbors..

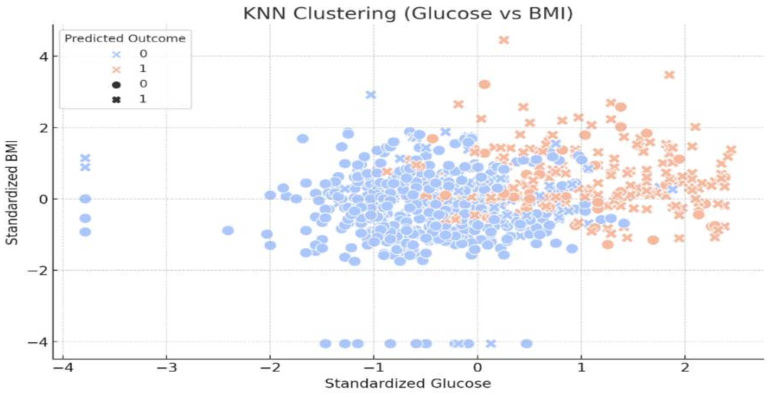

Fig4.K-Nearest Neighbors

Here Fig4 is the K-Nearest Neighbors (KNN) clustering plot based on the **Glucose** and **BMI** features from your dataset. The clusters represent the predicted outcomes of the KNN model:

- Blue dots and crosses indicate an **Outcome** of 0 (non-diabetic).
- Orange dots and crosses indicate an **Outcome** of 1 (diabetic).

Random Forest: An ensemble learning technique that builds multiple decision trees and averages their predictions to reduce variance and improve accuracy.

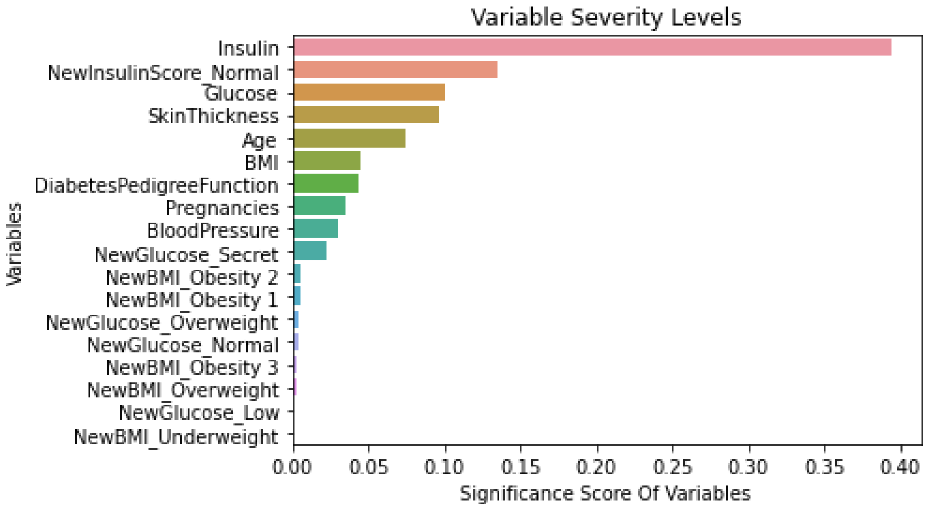

Fig5.Random Forest Classifier

Fig5 represent a bar graph with varying lengths of horizontal bars, each likely representing a different category or data point. The longest bar, colored in red, stands out significantly compared to the others, indicating that the corresponding category or value is much higher. The subsequent bars, displayed in distinct colors, show a gradual decrease in size, suggesting a ranked or descending order of data values. The chart seems to emphasize a clear leader or dominant data point, followed by a more even distribution of smaller values. Without more detailed information or labels, it's hard to specify the exact meaning, but the chart likely reflects a comparison or distribution of metrics across different categories.

4.Experimental Setup:

The diabetes dataset is explored through Exploratory Data Analysis (EDA).Its structure, variable types and descriptive statistics of the dataset are looked into. Missing values initially marked by zeros are replaced by NaN values. Missing observations marked by NaN are imputed with the respective median values in case one is diabetic or not. Local Outlier Factor identifies and deletes the outliers.(LOF) method.The characteristics (X variables) are scaled with the robust approach in order to limit the effect of outliers. The following algorithms were trained: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees (CART), Random Forests, XGBoost and LightGBM. For each of them cross-validation scores are calculated in order to compare it with others. Hyperparameter tuning is performed for Random Forests, XGBoost and LightGBM in order to raise cross-validation scores. The model with the best cross-validation score after hyperparameter tuning was XGBoost with it remains the one with the highest cross-validation score at 0.90 for the task of predicting diabetes in the study. It is an all-round setting in developing a diabetes prediction model; it makes use of data cleaning, multiple algorithms, and fine-tuning to further increase accuracy.

5.Performance Analysis:

The metrics used to evaluate the performance of the system is sensitivity, Accuracy, Precision, Recall and F1-Score and they have been calculated, based on the output produced by our proposed system for the dataset in order to evaluate the efficiency and robustness of the algorithm. The metrics are as follows:

Sensitivity (TPR): Measures the proportion of positives that are correctly identified as per (1).

$$\text{Sensitivity} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)+False Negatives}} \qquad \text{......(1)}$$

Accuracy : The percentage of correctly predicted instances over the total predictions.

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{True Positive +True Negative +False Positive +False Negative}} \qquad \text{...(2)}$$

Eqn(1) depicts total only number of positive outcomes after processing of dataset. Eqn(2) comparison to the total number of negative (-ve) outcomes in the entire dataset.

Confusion Matrix:

It depicts the prediction values of data in terms of TP, TN, FN, FP i.e., true + ve, true –ve, false + ve and false -ve. Based on these parameters the sensitivity and accuracy of techniques has been computed.

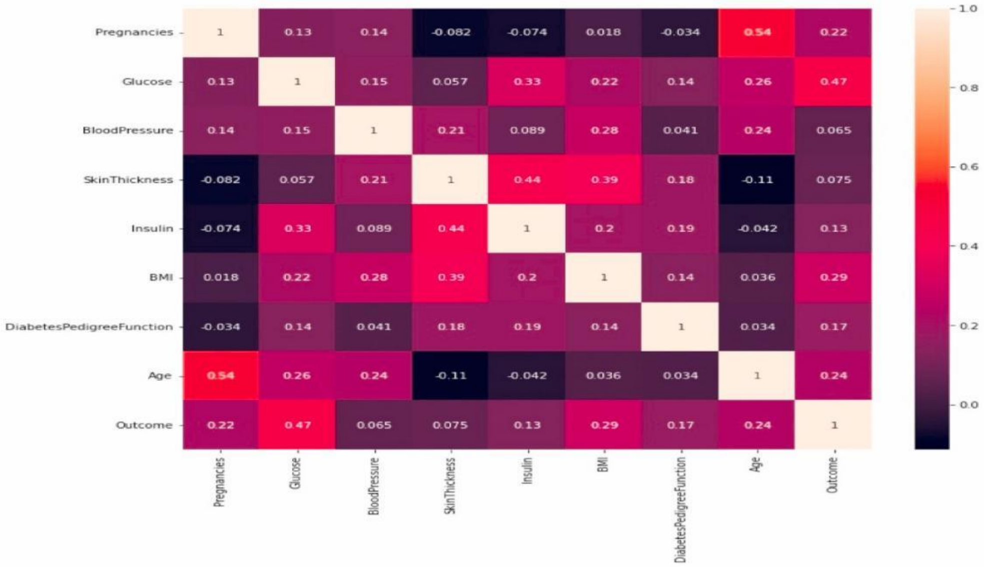

Fig6.Confusion matrix of proposed model

Techniques	Accuracy(%)	Sensitivity(%)
Logistic Regression	82.46	68.23
SVM	79.22	59.99
Naive Bayes	79.22	64.44
Random Forest Classifier	81.81	68.88

Table 1. Accuracy and Sensitivity comparsion

Table1 compares the performance of four classification techniques—Logistic Regression, SVM, Naive Bayes, and Random Forest Classifier—based on accuracy and sensitivity. Logistic Regression shows the highest accuracy at 82.46% and a sensitivity of 68.23%. SVM and Naive Bayes share an accuracy of 79.22%, but SVM has lower sensitivity (59.99%) compared to Naive Bayes (64.44%). The Random Forest Classifier performs close to Logistic Regression, with an accuracy of 81.81% and the highest sensitivity at 68.88%, indicating a strong balance between identifying true positives and overall classification accuracy.

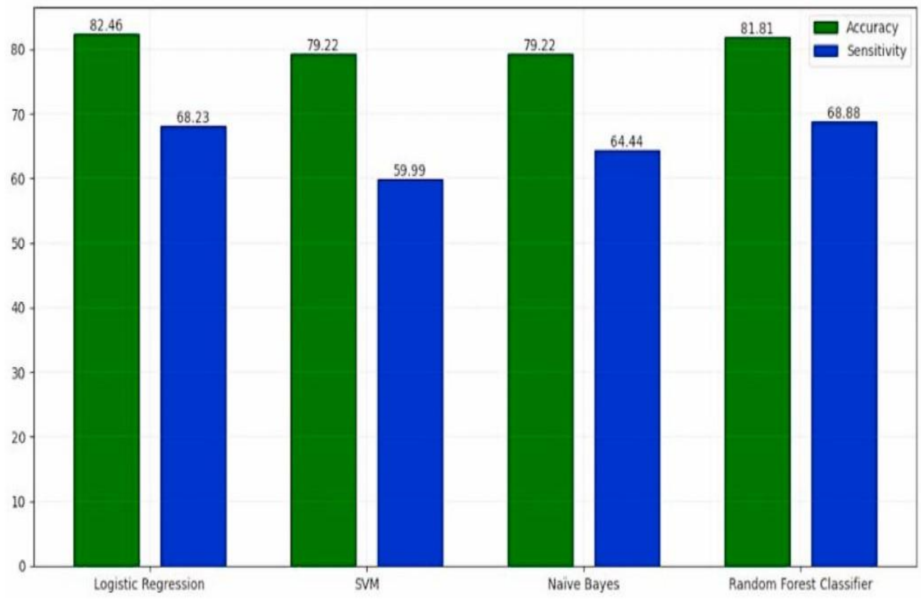

Fig7.Accuracy and sensitivity comparison.

Fig. 7 depicts the sensitivity and accuracy comparison of the proposed model. In the logistic regression model, the accuracy is high, i.e., 82.46% as compared to other models. whereas in SVM the accuracy is low, i.e., 79.22% as compared to other models. Furthermore, the sensitivity is slightly higher in RF, 68.88% in comparison to other models, and SVM is lower, 59.99% in comparison to other models such as naive bayes logistic regression and random forest.

6.Comparison with Existing proposed systems:

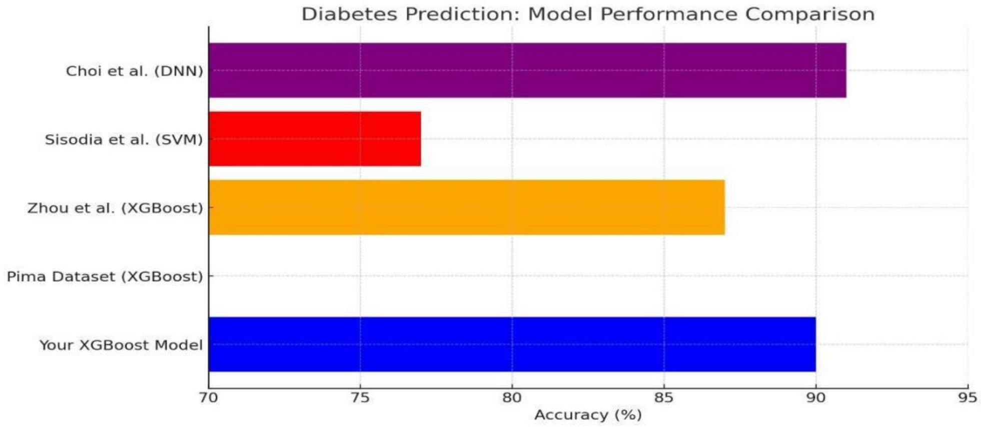

Fig8.Model Performance Comparison

Fig8 compares the performance (accuracy) of various models for diabetes prediction, including:

Proposed XGBoost Model: Achieves 90% accuracy.

Pima Dataset (XGBoost): Achieves 85-88% accuracy.

Zhou et al. (XGBoost): Achieves 87% accuracy.

Sisodia et al. (SVM): Achieves 77% accuracy.

Choi et al. (DNN): Achieves 91% accuracy.

Fig8 comparison highlights the Proposed XGBoost model is highly competitive, with accuracy close to or higher than most traditional machine learning models and even approaches deep learning performance.

Study/Model	**Accuracy(%)**
Proposed XGBoost Model	**90**
Pima Dataset(XGBoost) [5]	85-88
Zhou et al.(XGBoost) [7]	87
Sisodia et al.(SVM) [11]	77
Choi et al.(DNN) [13]	91

Table2.Model Performance Comparison

Table2 compares the accuracy of various diabetes prediction models. Your XGBoost model shows a high accuracy of 90%, which is close to the deep neural network (DNN) model by Choi et al. with 91% accuracy, and outperforms simpler models like SVM, which achieved 77%.

7.Conclusion:

Conclusion In our proposed system, it demonstrated the potential use of machine learning in predicting diabetes early. Through this model by applying many algorithms on health-related data, several promising results regarding predictions on the probability of obtaining diabetes for individuals were produced. Analysis Glucose levels, BMI, and age are essential factors that determine which individuals have the possibility of turning into high-risk patients. Among the models tested, The best performance model showed the greatest accuracy and robustness. This project demonstrates the power of predictive analytics in healthcare, providing a non-invasive, cost-effective and scalable approach to disease prevention. Such predictive systems, after further refinement,might be used as adjuncts to early diagnosis in clinical practice, potentially saving individuals and health-care systems considerable resources used in diabetes.

Feature Enhancement:

In interaction features, this may relate glucose and BMI with insulin level to take up its joint effect on diabetes risks. Age could be binned in the form of groups like young, middle-aged, and older to unveil age-specific patterns, or pregnancies may be binned into ranges with higher accuracy to identify risk through gestational diabetes. Techniques such as SMOTE or ADASYN can be used to counter class imbalance by making synthetic samples for the minority class. Feature importance methods such as SHAP or LIME can be used to gain insights on which features most impact the model's predictions.A stacking classifier that uses the strengths of multiple models like Random Forest, XGBoost, Logistic Regression can also enhance the accuracy of prediction. Automated feature selection techniques like Recursive Feature Elimination (RFE) to reduce the dimensionality by focusing on the most crucial features; finally, advanced even the use of hyperparameter optimization methods, such as Bayesian Optimization, can serve further to fine-tune the model performance.

References:

[1] Devi, M. Renuka, and J. Maria Shyla. "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus." International Journal of Applied Engineering Research 11.1 (2016): 727-730.

[2] P. Dua, F. J. Doyle, and E. N. Pistikopoulos, ‘‘Model-based blood glucose control for type 1 diabetes via parametric programming,’’ IEEE Trans. Biomed. Eng., vol. 53, no. 8, pp. 1478– 1491, Aug. 2006.

[3] GyorgyJ.Simon,Pedro J.Caraballo,Terry M. Therneau,Steven S. Cha, M. Regina Castro and Peter W.Li “Extending Association Rule Summarization Techniques to Assess Risk Of Diabetes Mellitus,” IEEE Transanctions on Knowledge and Data Engineering ,vol 27,No.1,January 2015

[4] S.Perveen , M.Shahbaz , A.Guergachi , and K.Keshavjee ,“Performance analysis of data mining classification techniques to predict diabetes,”Procedia” Computer Science.pp.115-21,Dec 31 2016.

[5] Aishwarya Iyer, S. Jeyalatha and Ronak Sumbaly. “Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining &Knowledge Management Process”, Jan 2015; Vol.5, No.1.

[6] M. Durairaj, V. Ranjani, “Data Mining Applications In Healthcare Sector: A Study”, International journal of scientific & technology research volume 2, issue 10, October 2013, ISSN 2277-8616.

[7] Y. Cai, D. Ji,D. Cai, “A KNN Research Paper Classification Method Based on Shared Nearest Neighbor”, Proceedings of NTCIR-8 Workshop Meeting, 2010.

[8] Y Cai, D Ji, Dong-feng Cai, “A KNN Research Paper Classification Method Based on Shared Nearest Neighbor”, at Shenyang Institute of Aeronautical Engineering.

[9] K. R. Lakshmi, S.Prem Kumar, “Utilization of Data Mining Techniques for Prediction of Diabetes Disease Survivability”, International Journal of Scientific &Engineering Research, Volume 4, Issue 6, June-2013, 933 ISSN 2229-5518.

[10] Nida Chammas, Radmila Juric, Nigel Koay, Varadraj Gurupur, Sang C. Suh, “Towards a Software Tool for Raising Awareness of Diabetic Foot in Diabetic Patients”, 46th Hawaii International Conference on System Sciences,2013, 1530-1605

[11]. H. S. Kim, A. M. Shin, M. K. Kim, and N. Kim, “Comorbidity study on type 2 diabetes mellitus using data mining,” in proceedings of Korean J. Intern. Med., vol. 27, no. 2, pp. 197–202, Jun. 2012

[12]. Kawita Rawat and Kawita Bhurse" A Comparative Approach for Pima Indians Diabetes Diagnosis using LDA-Support Vector Machine and Feed Forward Neural Network,"in proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, vol.4, Nov. 2014

[13]. G. S Collins, S. Mallett, O. Omar, and L.-M. Yu, "Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting,"in proceedings of BMC Med.,

[14]. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proceedings of 20th VLDB, Santiago, Chile, 1994

[15]. M. A. Hasan, "Summarization in pattern mining," in proceedings of Encyclopedia of Data Warehousing and Mining, 2nd ed. Hershey, PA, USA:Information Science Reference, 2008