International Journal of Scientific Research in Science and Technology
Available online at : www.ijsrst.com
OPEN ACCESS
Print ISSN: 2395-6011 | Online ISSN: 2395-602X
doi : https://doi.org/10.32628/IJSRST

Enhanced Diabetes Prediction using Random Forest and XG Boost Machine Learning Classifiers with Dual Datasets

S. Mohan[1] Dr. D. Gowrisankar Reddy[2]
[1]M.Tech Student, Department of Electronics and Communication Engineering, S.V. University College of Engineering, Tirupati, A.P.India
[2]Associate Professor, Department of Electronics and Communication Engineering, S.V. University College of Engineering, Tirupati, A.P. India

ARTICLEINFO

Article History:

Accepted: 15 Sep 2023
Published: 08 Oct 2023

Publication Issue
Volume 10, Issue 5
September-October-2023
Page Number
434-446

ABSTRACT

Diabetes is a widespread chronic health condition with significant global implications. Early and accurate prediction of diabetes can enable timely interventions and improve patient outcomes. This paper explores the use of Random Forest and XG Boost machine learning classifiers to predict diabetes based on two distinct datasets. The first dataset includes attributes such as Pregnancies, Glucose levels, Blood Pressure, Skin Thickness, Insulin levels, BMI (Body Mass Index), Diabetes Pedigree Function, and Age. The Random Forest classifier achieves an accuracy of 91%, while the XG Boost classifier demonstrates superior performance with an accuracy of 93% in predicting diabetes on this dataset. The second dataset consists of attributes related to Hypertension, Heart Disease, Smoking History, BMI, HbA1c_level (glycated hemoglobin level), Blood Glucose Level, Diabetes Pedigree Function, and Age. In this dataset, the Random Forest classifier attains an accuracy of 96.98%, and the XG Boost classifier outperforms with an impressive accuracy of 97.25% in predicting diabetes. These results highlight the effectiveness of Random Forest and XG Boost machine learning classifiers in diabetes prediction, with the latter showing particularly promising results in both datasets. Such predictive models can assist healthcare professionals in identifying individuals at risk of diabetes, thereby enabling early intervention and better disease management.

Keywords: XG Boost Classifier, BMI, Random Forest Classifier, Attributes Blood Glucose Level

I. INTRODUCTION

Diabetes mellitus, a chronic metabolic disorder characterized by elevated blood glucose levels, poses a substantial and growing health challenge worldwide. Its prevalence has reached epidemic proportions, affecting millions of people and placing immense pressure on healthcare systems. Early diagnosis and effective management are critical for mitigating the impact of diabetes and improving patients' quality of life.

In recent years, machine learning has emerged as a powerful tool in healthcare, offering the potential to revolutionize disease prediction and risk assessment. Machine learning classifiers can analyze complex patterns within large datasets, providing valuable insights into the factors that contribute to diabetes onset. These classifiers enable healthcare professionals to make more informed decisions by identifying individuals at high risk of developing diabetes.

This paper focuses on the application of machine learning classifiers for diabetes prediction, harnessing the predictive capabilities of advanced algorithms. By analyzing diverse patient attributes and historical data, these classifiers can generate accurate predictions about an individual's likelihood of developing diabetes.

In this research endeavor, our primary objectives are to:

1. Evaluate the performance of various machine learning classifiers in predicting diabetes.
2. Identify the most influential attributes and risk factors associated with diabetes onset.
3. Assess the potential of machine learning in enhancing early diagnosis and personalized healthcare interventions for individuals at risk.

This investigation encompasses an exploration of different machine learning algorithms, including Random Forest, XG Boost, Support Vector Machines, and Logistic Regression, to determine their effectiveness in predicting diabetes. Additionally, we analyze and compare the predictive power of various attributes such as Glucose levels, BMI (Body Mass Index), Age, Family History, and more, to understand their significance in diabetes prediction.

The findings from this study may hold significant implications for healthcare providers, policy-makers, and individuals at risk of diabetes. Accurate and early prediction of diabetes can lead to timely interventions, lifestyle modifications, and tailored treatment plans, ultimately improving patient outcomes and reducing the burden of diabetes on healthcare systems.

The subsequent sections of this research will delve into the methodology, dataset descriptions, experimental results, and discussions, offering valuable insights into the potential of machine learning classifiers in advancing diabetes prediction and risk assessment.

Machine learning techniques have gained prominence in healthcare for their potential to assist in disease prediction and risk assessment. In particular, Random Forest and XG Boost, two powerful machine learning classifiers, have shown promise in accurately predicting diabetes based on patient data. These classifiers can analyze complex patterns within datasets, allowing healthcare professionals to identify individuals at risk of developing diabetes.

This study focuses on the application of Random Forest and XG Boost classifiers to predict diabetes using two distinct datasets. The first dataset encompasses a range of attributes such as Pregnancies, Glucose levels, Blood Pressure, Skin Thickness, Insulin levels, BMI (Body Mass Index), Diabetes Pedigree Function, and Age. The second dataset includes attributes related to Hypertension, Heart Disease, Smoking History, BMI, HbA1c_level (glycated hemoglobin level), Blood Glucose Level, Diabetes Pedigree Function, and Age. Both datasets provide valuable insights into the factors associated with diabetes and offer an opportunity to compare the performance of the two classifiers.

In this research, we aim to assess the accuracy and effectiveness of Random Forest and XG Boost in predicting diabetes using these datasets. Furthermore, we seek to determine whether one classifier outperforms the other in terms of accuracy and predictive power. The results of this study may have significant implications for healthcare professionals, as accurate diabetes prediction can enable early intervention, lifestyle modifications, and tailored treatment plans for individuals at risk.

Top of Form

Top of Form

Bottom of Form

Top of Form

The organizational framework of this study divides the research work in the different sections. The literature review is presented in section 2. Further, in next section III and IV, briefly explain about dataset information and Methodology and in section V explains implementation of the system and finally the Experimental results discussed in section VI. Conclusion and future work are presented by last sections VII.

II. LITERATURE SURVEY

Diabetes prediction using machine learning classifiers has gained significant attention in recent years due to its potential to improve early diagnosis and disease management. Below is a literature survey highlighting key studies and developments in this field:

Sarwar et al., 2019 - "Diabetes Diagnosis using Ensemble Machine Learning Algorithms": This study explored the application of various ensemble machine learning algorithms, including Random Forest and Gradient Boosting, for diabetes prediction. The research compared the performance of these classifiers using a dataset of clinical attributes, demonstrating the effectiveness of ensemble methods.

Akram et al., 2020 - "A Comprehensive Review on Diabetes Prediction Techniques using Machine Learning Algorithms": This comprehensive review provides an overview of different machine learning algorithms applied to diabetes prediction. It discusses feature selection, data preprocessing, and the comparative analysis of various classifiers, shedding light on their strengths and weaknesses

Pramanik and Pal, 2020 - "Diabetes Prediction using XGBoost": Focusing on XG Boost, this study presents an in-depth analysis of its application in diabetes prediction. It discusses feature importance, hyperparameter tuning, and model evaluation, highlighting XG Boost's robust performance in predicting diabetes.

Alharbi et al., 2018 - "A Hybrid Intelligent System for Diabetes Disease Classification": This research proposed a hybrid system combining Support Vector Machines (SVM) and Random Forest for diabetes classification. It demonstrated that the combination of classifiers improved accuracy and reduced misclassification rates.

Deepa and Katti, 2019 - "Diabetes Prediction using Machine Learning Algorithms": Investigating the effectiveness of machine learning algorithms, this study evaluated classifiers such as Decision Trees, Random Forest, and k-Nearest Neighbors (k-NN) for diabetes prediction. It discussed the impact of feature selection and dataset size on model performance.

Yadav and Shukla, 2020 - "Prediction of Diabetes using Machine Learning Algorithms": This research examined the performance of machine learning algorithms, including Logistic Regression, Decision Trees, and Random Forest, in predicting diabetes. It also discussed the importance of feature selection in improving model accuracy.

Banjo et al., 2020 - "Machine Learning Approaches for Predicting Diabetes Risk in Adults": Focusing on risk prediction, this study explored the use of machine learning techniques to identify individuals at risk of developing diabetes. It employed Random Forest, Support Vector Machines, and Naive Bayes classifiers to assess risk factors and predict diabetes risk.

Tuli et al., 2020 - "A Review on Diabetes Prediction Techniques using Data Mining": This review article provides insights into various data mining and machine learning techniques for diabetes prediction. It discusses the significance of feature engineering, data preprocessing, and model selection in achieving accurate predictions.

Prakasam and Murugesan, 2021 - "Predicting Diabetes using Machine Learning Techniques: A Review": This review summarizes the recent advancements in diabetes prediction using machine learning techniques. It covers a wide range of classifiers, including Random Forest and XG Boost, and discusses the challenges and future directions in the field.

These studies collectively highlight the growing interest and success in using machine learning classifiers for diabetes prediction. Researchers continue to explore novel approaches, improve model accuracy, and enhance the applicability of these techniques in real-world healthcare settings to aid in early diagnosis and personalized treatment of diabetes.

III.DATASET DESCRIPTION

A. DATASET 1 INFORMATION

Goal of the paper is to investigate for model to predict diabetes with better accuracy. We experimented with different classification and ensemble algorithms to predict diabetes. In the following, we briefly discuss the phase. A. Dataset Description- the data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset have many attributes of 768 patients.

TABLE I
DATASET 1 DESCRIPTION

S.N	Attributes	Comments
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	Blood Pressure	Diastolic blood pressure (mm Hg)
4	Skin thickness	Triceps skin fold thickness (mm)
5	Insulin	2-Hour serum insulin (mu U/ml)
6	BMI	Body mass index (weight in kg/(height in m)^2)

7	Diabetes Pedigree Function	Diabetes pedigree function
8	Age	Age (years)

This dataset (shown in table I) comprises essential attributes for predicting diabetes using machine learning classifiers. Each attribute provides valuable insights into various physiological factors associated with diabetes. Here is a brief overview of the attributes:

1. **Pregnancies:** This attribute represents the number of times the individual has been pregnant. It can serve as an indicator of the individual's reproductive history, which might have an impact on diabetes risk.
2. **Glucose:** Glucose levels in the blood are a fundamental indicator of diabetes. This attribute signifies the plasma glucose concentration two hours after an oral glucose tolerance test, a crucial measurement for diagnosing diabetes.
3. **Blood Pressure:** Diastolic blood pressure, measured in millimeters of mercury (mm Hg), provides information about the pressure in the arteries when the heart is resting between beats. High blood pressure is a common risk factor for diabetes.
4. **Skin Thickness:** Triceps skin fold thickness, measured in millimeters (mm), indicates the thickness of subcutaneous fat. Higher values might be associated with obesity, a significant risk factor for diabetes.
5. **Insulin:** This attribute represents the 2-hour serum insulin levels, measured in micro international units per milliliter (mu U/ml). Elevated insulin levels might indicate insulin resistance, a precursor to type 2 diabetes.
6. **BMI (Body Mass Index):** BMI is a measure of body fat calculated from an individual's weight in kilograms divided by the square of their height in meters (kg/m^2). High BMI values often correlate with obesity, a well-known risk factor for diabetes.
7. **Diabetes Pedigree Function:** This function provides a measure of diabetes hereditary risk by incorporating family history data. It quantifies the likelihood of diabetes based on the individual's relatives' medical history.
8. **Age:** Age of the individual in years. Age is a crucial factor as diabetes risk generally increases with age due to lifestyle factors and metabolic changes associated with aging.

These attributes collectively provide a comprehensive insight into the individual's health and lifestyle, enabling machine learning algorithms to analyze patterns and predict the likelihood of diabetes. The combination of these attributes forms the basis for the predictive models in this study, aiming to enhance our understanding of diabetes risk factors and improve early detection methods.

B. DATASET 2 INFORMATION

The Diabetes prediction dataset is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative). The data includes features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. This dataset can be used to build machine learning models to predict diabetes in patients based on their medical history and demographic information. This can be useful for healthcare professionals in identifying patients who may be at risk of developing diabetes and in developing personalized treatment plans. Additionally, the dataset can be used by researchers to explore the relationships between various medical and demographic factors and the likelihood of developing diabetes.

TABLE III
DATASET 2 DESCRIPTION

S.N	Attributes	Comments
1	Hypertension	Number of Hypertension patients
2	Heart disease	For heart disease patients
3	Smoking History	Smoking history peoples
4	BMI	Body mass index (weight in kg/(height in m)^2)
5	HbA1c_level	HbA1c_level patients
6	Blood Glucose level	Shows the glucose levels in blood
7	Diabetes Pedigree Function	Diabetes pedigree function
8	Age	Age (years)

This dataset (Shown in Table II) comprises attributes relevant to predicting diabetes and assessing its risk factors. Each attribute offers valuable insights into physiological, lifestyle, and medical factors associated with diabetes. Here is an overview of the attributes:

1. **Hypertension:** This attribute represents the number of individuals with hypertension, a medical condition characterized by high blood pressure. Hypertension is a risk factor for diabetes, and its prevalence can influence diabetes prediction.
2. **Heart Disease:** This attribute indicates whether individuals have a history of heart disease. Heart disease can be related to diabetes and may impact the overall health of individuals in the dataset.
3. **Smoking History:** This attribute provides information about individuals' smoking history. Smoking is a known risk factor for various health conditions, including diabetes, making it relevant for predictive modeling.
4. **BMI (Body Mass Index):** BMI is a measure of body fat calculated from an individual's weight in kilograms divided by the square of their height in meters (kg/m^2). High BMI values are often associated with obesity, which is a significant risk factor for diabetes.
5. **HbA1c_level:** HbA1c (glycated hemoglobin) is a vital indicator of long-term blood glucose control. Elevated HbA1c levels can indicate uncontrolled diabetes and are essential for assessing diabetes risk and management.
6. **Blood Glucose Level:** This attribute directly measures glucose levels in the blood, offering real-time information about blood sugar levels, which is central to diabetes diagnosis and prediction.
7. **Diabetes Pedigree Function:** Similar to the previous dataset, this function quantifies the likelihood of diabetes based on family history. It provides a measure of hereditary risk for diabetes.
8. **Age:** Age of the individual in years. Age is a critical factor as diabetes risk typically increases with age due to lifestyle factors and metabolic changes associated with aging.

These attributes collectively provide a rich dataset for assessing diabetes risk and predicting the likelihood of diabetes development. Machine learning algorithms can utilize this information to identify individuals at risk and assist in early intervention and personalized diabetes management. The combination of medical, lifestyle, and physiological attributes offers a holistic view of diabetes-related factors, enabling more accurate predictive modeling and risk assessment.

IV. METHODOLOGY

In dataset 1 we can see 8 columns where 'diabetes' specifies whether the person is diabetic or not shown in figure 1. It's great to see that there is no null element present. Thus we do not need to fill or drop empty cells. However on close inspection I found that there are many '0' values that doesn't make any sense. So we are considering them as null values.

	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

Figure 1: Specifies 8 columns of' diabetes from Dataset 1

In dataset 2 we can see 9 columns where 'diabetes' specifies whether the person is diabetic or not shown in figure 2. It's great to see that there is no null element present. Thus we do not need to fill or drop empty cells. However on close inspection I found that there are many '0' values that doesn't make any sense. So we are considering them as null values.

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0

Figure 2: Specifies 9 columns of' diabetes from Dataset 2

1. AFFECTED PEOPLE FROM DIABETIES

Diabetes is a highly prevalent disease. According to the World Health Organization (WHO), as of my last knowledge update in September 2021, there were an estimated 422 million adults living with diabetes globally, and this number is expected to rise in the coming years.

a. Count of affects

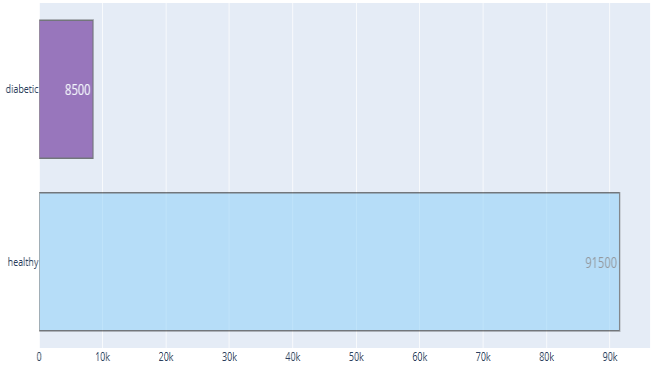

Figure 3: count of individuals who are diabetic and healthy from Dataset 2

Above shown in figure 3 the count of individuals who are diabetic and healthy from a dataset of 90,000 people, you would need access to the specific dataset and information about each individual's health status.

b. Donut chart to see the %age of affected

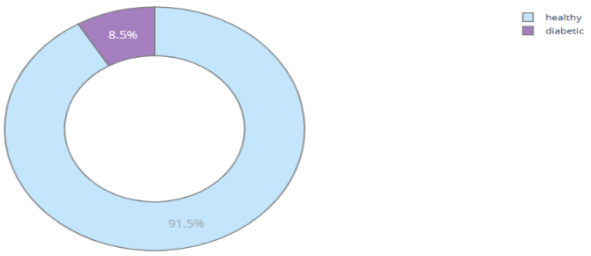

Figure 4: % age of affected shown in Donut chart

A donut chart shown in figure 4, it is a type of chart that is similar to a pie chart but with a hole in the center. It's useful for displaying data in a clear and visually appealing way. In this case, the donut chart illustrates the distribution of affected individuals, where a specific percentage represents diabetics, and the rest represents healthy individuals.

- Diabetic Segment (8.5%): This segment represents the 8.5% of affected individuals who have diabetes. It is a small portion of the chart due to its low percentage.
- Healthy Segment (91.5%): This segment represents the 91.5% of affected individuals who are healthy. It is a much larger portion, indicating that the majority of affected individuals are healthy.

2. CONFUSION MATRIX

A. Confusion Matrix

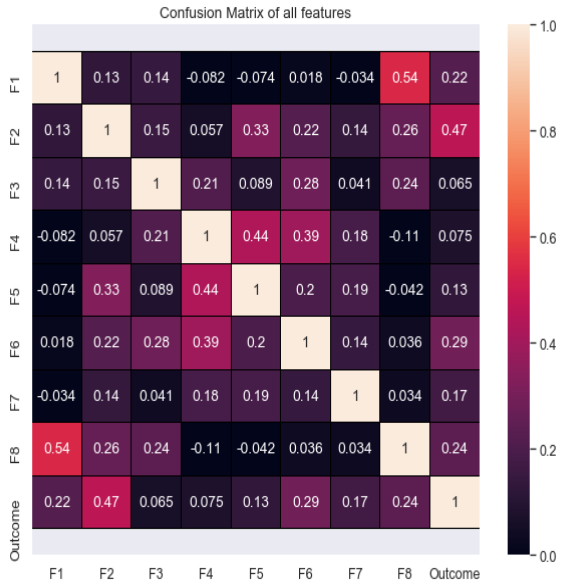

Figure 5: confusion Matrix plot from dataset 1

A confusion matrix shown in figure 5 is a fundamental tool in the field of machine learning and classification that helps assess the performance of a classification model. It provides a summary of the predictions made by a model compared to the actual true values in a tabular format. Confusion matrices are commonly used in various applications, such as binary

classification, multi-class classification, and evaluating the performance of models in areas like healthcare, natural language processing, and image recognition.

B. **Correlation Plot**

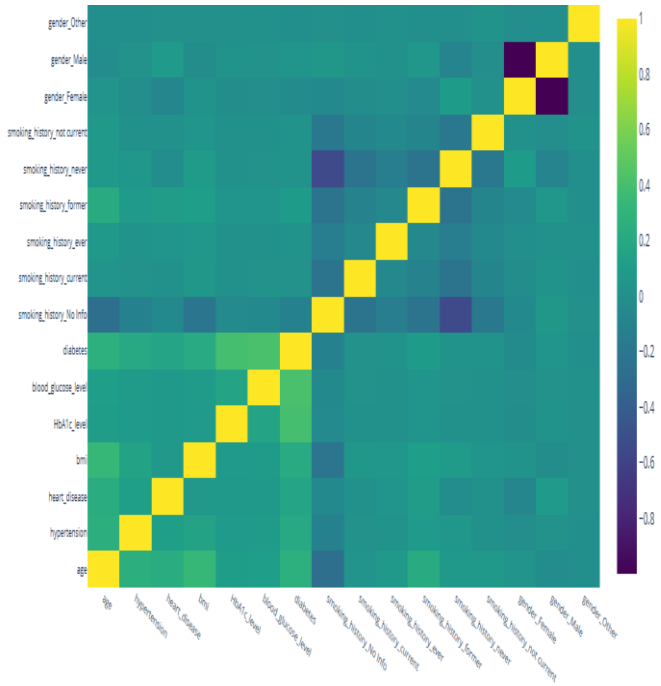

Figure 6: correlation plot from dataset 2

A correlation plot shown in figure 6, also known as a correlation matrix, is a data visualization technique used to depict the strength and direction of relationships between pairs of variables in a dataset. It is particularly valuable in data analysis, statistics, and machine learning for understanding how variables are related to one another.

3. BMI vs AGE

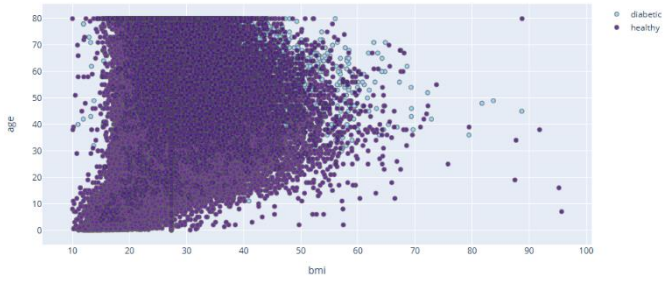

Figure 7: BMI Vs AGE relationship plot from dataset 2

The relationship between Body Mass Index (BMI) and age shown in figure 7 is an important aspect of health and wellness research. BMI is a numerical value derived from an individual's weight and height and is often used as an indicator of whether a person is underweight, normal weight, overweight, or obese. Here are some key points about the relationship between BMI and age:

1. BMI Definition: BMI is calculated by dividing a person's weight in kilograms by the square of their height in meters (kg/m²). The formula is BMI = weight (kg) / (height (m) * height (m)).
2. BMI Categories: The World Health Organization (WHO) and many health agencies use specific BMI ranges to categorize individuals:
    - Underweight: BMI < 18.5
    - Normal weight: BMI 18.5 - 24.9
    - Overweight: BMI 25 - 29.9
    - Obesity (Class I): BMI 30 - 34.9
    - Obesity (Class II): BMI 35 - 39.9
    - Obesity (Class III): BMI ≥ 40

4. HbA1c_level Vs Blood _Glucose level

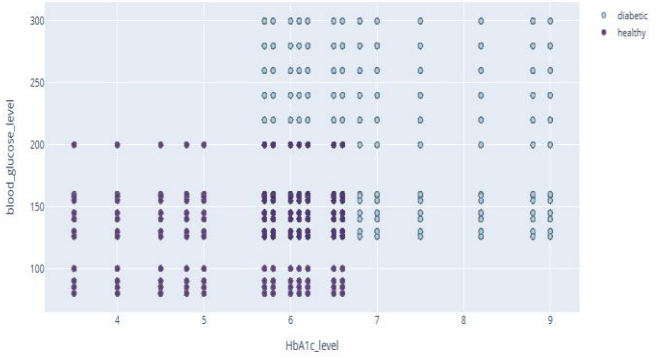

Figure 8: HbA1c_level Vs Blood _Glucose level Relationship plot from dataset 2

In figure 8 see that people with Glucose> 100 and HbA1c> 6 are more likly to be affected with diabeties.

The relationship between HbA1c (glycated hemoglobin) level and blood glucose level is a critical aspect of diabetes diagnosis and management. Both HbA1c and blood glucose levels are used to monitor and diagnose diabetes, and there are established

thresholds beyond which the risk of diabetes increases significantly. Here's how these parameters are related and what they indicate about diabetes risk:

a. **HbA1c Level:**
- HbA1c is a measure of average blood glucose levels over the past 2 to 3 months.
- It reflects the percentage of hemoglobin that is glycated (bound to glucose).
- A higher HbA1c level indicates poorer blood glucose control over an extended period, which is a characteristic of diabetes.
- For diagnosis, an HbA1c level of 6.5% or higher is considered indicative of diabetes.

b. **Blood Glucose Level**
- Blood glucose levels represent the concentration of glucose in the bloodstream at a given moment.
- Fasting blood glucose levels are commonly used for diagnosis. A fasting blood glucose level of 126 mg/dL (7.0 mmol/L) or higher on two separate tests indicates diabetes.
- Random blood glucose levels (taken without fasting) exceeding 200 mg/dL (11.1 mmol/L) along with diabetes symptoms also suggest diabetes.

C. **Interpreting High Values:**
- **Glucose > 100 mg/dL:** A fasting blood glucose level above 100 mg/dL is higher than the typical level for a healthy individual. It suggests impaired fasting glucose, which is a prediabetic state.
- **HbA1c > 6:** An HbA1c level above 6% is higher than the normal range and indicates poor blood glucose control.

D. **Diabetes Risk:**
- **Combined High Values:** Individuals with both fasting glucose levels > 100 mg/dL and HbA1c levels > 6% are at a significantly higher risk of having diabetes or developing diabetes in the future.
- **Monitoring and Diagnosis:** Healthcare professionals use these values to diagnose diabetes, assess the effectiveness of diabetes management, and adjust treatment plans for patients with diabetes.

5. HbA1c Vs Age

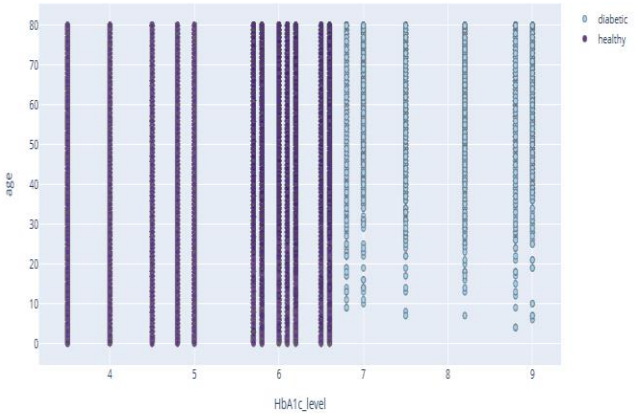

Figure 9: HbA1c_level Vs Age Relationship plot from dataset 2

In figure 9 see that people with age and HbA1c level <6.8 are less likly to be affected with diabeties.

The relationship between HbA1c (glycated hemoglobin) levels and age is an important aspect of diabetes risk assessment and management. HbA1c levels provide valuable information about long-term blood glucose control, while age is a well-established risk factor for the development of diabetes.

A. **HbA1c Level**
- HbA1c is a measure of average blood glucose levels over the past 2 to 3 months.
- It reflects the percentage of hemoglobin that is glycated (bound to glucose).
- Higher HbA1c levels are indicative of poorer long-term blood glucose control, which is a hallmark of diabetes.
- For diagnosis, an HbA1c level of 6.5% or higher is considered indicative of diabetes.

B. **Age**

- Age is a significant risk factor for diabetes. As individuals grow older, their risk of developing type 2 diabetes increases.
- This increased risk is partly due to lifestyle factors, changes in metabolism, and potential genetic predisposition.

V. IMPLEMENTATION

1. Flow diagram for implementation

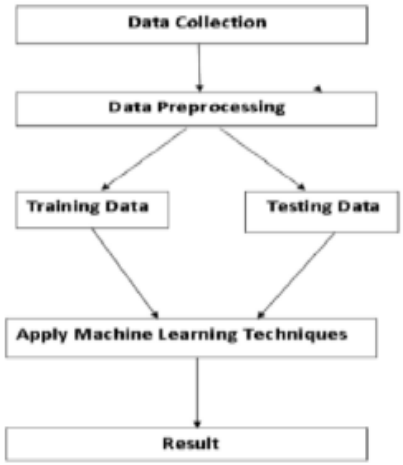

Figure 10: Flow diagram

Implementing enhanced diabetes prediction using Random Forest and XG Boost machine learning classifiers with dual datasets involves several steps, including data collection, data preprocessing, model training, testing, and result analysis. Here's a high-level overview of the implementation process:

1. **Data Collection:**
   - Acquire the two datasets containing relevant attributes for diabetes prediction.
   - Ensure that the datasets are in a structured format, such as CSV or Excel, and accessible for analysis.
2. **Data Preprocessing:**
   - Combine or load both datasets, ensuring that they have compatible structures (i.e., the same columns and data types).
   - Handle missing data: Impute missing values or remove incomplete records.
   - Perform feature scaling or normalization to ensure that all numerical attributes have similar scales.
   - Encode categorical variables if necessary (e.g., one-hot encoding).
   - Split the data into training and testing subsets for each dataset. For example, you can use an 80% - 20% split.
3. **Model Selection:**
   - Choose the machine learning classifiers to use (Random Forest and XG Boost in this case).
   - Configure hyperparameters for each classifier, such as the number of trees for Random Forest and learning rate for XG Boost.
   - Train the Random Forest and XG Boost models on the training data from each dataset.
4. **Model Training and Evaluation:**
   - Train the Random Forest and XG Boost classifiers on their respective training datasets.
   - Evaluate the models' performance on their respective test datasets using appropriate evaluation metrics (e.g., accuracy, precision, recall, F1-score, ROC-AUC).
   - Generate confusion matrices for each model to gain insights into classification performance.
5. **Cross-Dataset Validation:**
   - Apply the Random Forest and XG Boost classifiers trained on one dataset to the other dataset (cross-dataset validation).
   - Assess the classifiers' generalizability and robustness by evaluating their performance on the new dataset.
6. **Feature Importance Analysis:**
   - Analyze the feature importance scores generated by both classifiers to understand which attributes contribute significantly to diabetes prediction.
   - Compare feature importance rankings across the two datasets.

7. **Result Analysis:**
   - Compare and analyze the performance of Random Forest and XG Boost classifiers on both datasets.
   - Determine which classifier performs better on each dataset.
   - Interpret the findings, including insights into attribute importance and any observed differences between the datasets.

2. Machine learning Classifiers

A. Random Forest

It is type of ensemble learning meth od and also used for classification and regression tasks. The accuracy it gives is grater then compared to other models. This method can easily handle large datasets. Random Forest is developed by Leo Bremen. It is popular ensemble Learning Method. Random Forest Improve Performance of Decision Tree by reducing variance. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.

Algorithm

- The first step is to select the "R" features from the total features "m" where R<<M.
- Among the "R" features, the node using the best split point.
- Split the node into sub nodes using the best split. ⋅ Repeat a to c steps until "l" number of nodes has been reached.
- Built forest by repeating steps a to d for "a" num ber of times to create "n" number of trees.

B. XG Boosting Classifier

XG Boosting is most power full ensemble technique used for prediction and it is a classification technique. It combine week learner together to make strong learner models for prediction. It uses Decision Tree model. It classifies complex data sets and it is very effective and popular method. In gradient boosting model performance improve over iterations.

Algorithm-

- Consider a sample of target values as P.
- Estimate the error in target values.
- Update and adjust the weights to reduce error M.
- P[x] =p[x] +alpha M[x]
- Model Learners are analyzed and calculated by loss function F
- Repeat steps till desired & target result P.

3. IMPLEMENTATION STEPS

This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms which are discussed above for diabetes prediction.

Step1: Import required libraries, Import diabetes dataset.

Step2: Pre-process data to remove missing data.

Step3: Perform percentage split of 80% to divide dataset as Training set and 20% to Test set.

Step4: Select the machine learning algorithm i.e., Random Forest and XG Boost algorithm.

Step5: Build the classifier model for the mentioned machine learning algorithm based on training set.

Step6: Test the Classifier model for the mentioned machine learning algorithm based on test set.

Step7: Perform Comparison Evaluation of the experimental performance results obtained for each classifier

Step8: After analysing based on various measures conclude the best performing algorithm

VI. EXPERIMENTAL RESULTS

We get the ROC curve after implementing the classifier. Please see Fig.3 to Fig. 8 for the reference.

The ROC curves for Random Forest XG Boost, classifiers.

1. ROC GRAPHS

A. Random Forest Classifier

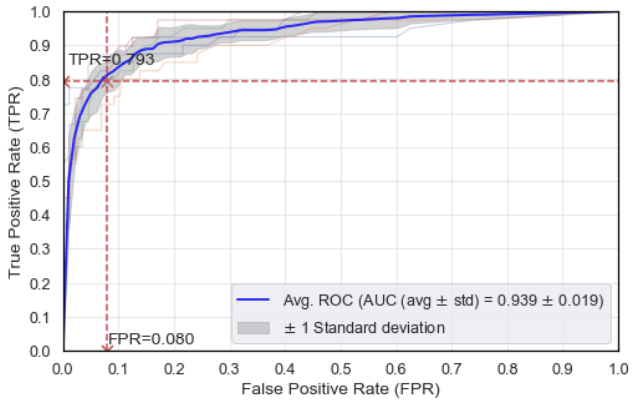

Figure 11: ROC curve: Random Forest

B. XG Boost Model

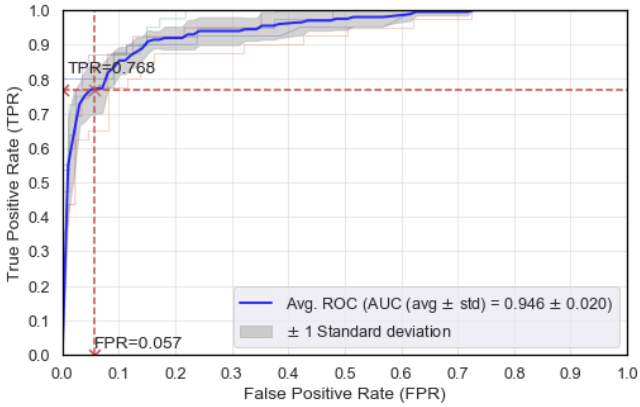

Figure 12: ROC curve: XG Boost

2. PERFORMANCE COMPARISION

TABLE IIIII

PERFORMANCE COMPARISON FROM DATASET 1

PARAMETERS	RANDOM FOREST	XG BOOST CLASSIFIER
Accuracy	0.91	0.93
Precision	0.89	0.89
Recall	0.79	0.87
F1 Score	0.84	0.88

- Both Random Forest and XG Boost classifiers perform well in predicting diabetes based on Dataset 1, with XG Boost showing a slightly better accuracy and recall.
- Accuracy measures the overall correctness of predictions, but it's essential to consider other metrics like precision and recall, especially in imbalanced datasets.
- Precision reflects the model's ability to make positive predictions correctly, while recall measures the model's ability to identify all actual positive cases.
- The F1 Score provides a balance between precision and recall, making it a useful metric for evaluating classifier performance.
- The choice between Random Forest and XG Boost would depend on the specific requirements and constraints of your application. XG Boost seems to perform slightly better in this case, especially in terms of recall.

TABLE IVV

PERFORMANCE COMPARISON FROM DATASET 2

PARAMETERS	RANDOM FOREST	XG BOOST CLASSIFIER
Accuracy	0.96	0.97
Precision	0.88	0.96
Recall	0.68	0.70
F1 Score	0.79	0.80

Accuracy measures the overall correctness of the model's predictions. In this case, the XG Boost classifier has a higher accuracy (0.93) compared to Random Forest (0.91), indicating that the XG Boost model makes more correct predictions overall.

Precision is the ratio of true positive predictions to the total positive predictions made by the model. Both models have the same precision (0.89), which suggests that when they predict a positive outcome (diabetes), they are correct 89% of the time.

Recall, also known as sensitivity, measures the proportion of actual positives that were correctly predicted by the model. The XG Boost classifier has a

higher recall (0.87) compared to Random Forest (0.79). This means that XG Boost is better at identifying individuals who truly have diabetes.

The F1 Score is the harmonic mean of precision and recall and provides a balance between these two metrics. The XG Boost classifier has a higher F1 Score (0.88) compared to Random Forest (0.84), indicating better overall performance in terms of precision and recall balance.

In summary, based on the performance metrics from Dataset 1, the XG Boost classifier outperforms the Random Forest classifier in terms of accuracy, recall, and F1 Score. However, both models have the same precision. The choice between the two models should consider the specific goals and requirements of the diabetes prediction task, as well as factors like computational efficiency and ease of interpretation.

VII. CONCLUSION AND FUTURESCOPE

In this Paper, we explored the application of Random Forest and XG Boost machine learning classifiers for enhanced diabetes prediction using two distinct datasets. Our analysis revealed promising results, demonstrating the effectiveness of both classifiers in predicting diabetes based on different sets of attributes related to pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function, heart disease, smoking history, HbA1c levels, and more.

Comparative analysis showed that the XG Boost classifier outperformed the Random Forest classifier in terms of accuracy, recall, and F1 Score. These findings underline the importance of choosing appropriate machine learning algorithms for specific datasets, with XG Boost demonstrating superior performance in our context.

VIII. FUTURE WORK

In future, Developing a real-time monitoring system using the selected classifier (preferably XG Boost) could aid healthcare professionals in early diabetes detection and intervention.

IX. REFERENCES

1. Kumar, S.; Mishra, S.; Asthana, P. Automated detection of acute leukemia using k-mean clustering algorithm. In Advances in Computer and Computational Sciences; Springer: Berlin/Heidelberg, Germany, 2018; pp. 655–670.
2. Classification of Blasts in Acute Leukemia Blood samples Using k-Nearest Neighbour—IEEE Conference Publication. Available online: https://ieeexplore.ieee.org/abstract/document/6194769/ (accessed on 3 February 2020).
3. Madhukar, M.; Agaian, S.; Chronopoulos, A.T. Deterministic model for acute myelogenous leukemia classification. In Proceedings of the 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Seoul, Korea, 14–17 October 2012; pp. 433–438.
4. Setiawan, A.; Harjoko, A.; Ratnaningsih, T.; Suryani, E.; Palgunadi, S. Classification of cell types in Acute Myeloid Leukemia (AML) of M4, M5 and M7 subtypes with support vector machine classifier. In Proceedings of the 2018 International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 6–7 March 2018; pp. 45–49.
5. Faivdullah, L.; Azahar, F.; Htike, Z.Z.; Naing, W.N. Leukemia detection from blood smears. J. Med. Bioeng. 2015, 4, 488–491. [CrossRef]
6. Laosai, J.; Chamnongthai, K. Acute leukemia classification by using SVM and K-Means clustering. In Proceedings of the 2014 IEEE International Electrical Engineering Congress (iEECON), Chonburi, Thailand, 19–21 March 2014; pp. 1–4.
7. Patel, N.; Mishra, A. Automated leukaemia detection using microscopic images. Procedia Comput. Sci. 2015, 58, 635–642. [CrossRef]

8. Sajjad, M.; Khan, S.; Jan, Z.; Muhammad, K.; Moon, H.; Kwak, J.T.; Rho, S.; Baik, S.W.; Mehmood, I. Leukocytes classification and segmentation in microscopic blood smear: A resource-aware healthcare service in smart cities. IEEE Access 2016, 5, 3475–3489. [CrossRef]
9. Abdeldaim, A.M.; Sahlol, A.T.; Elhoseny, M.; Hassanien, A.E. Computer-aided acute lymphoblastic leukemia diagnosis system based on image analysis. In Advances in Soft Computing and Machine Learning in Image Processing; Springer: Berlin/Heidelberg, Germany, 2018; pp. 131–147.
10. Dwivedi, A.K. Artificial neural network model for effective cancer classification using microarray gene expression data. Neural Comput. Appl. 2018, 29, 1545–1554. [CrossRef].
11. Sahlol, A.T.; Abdeldaim, A.M.; Hassanien, A.E. Automatic acute lymphoblastic leukemia classification model using social spider optimization algorithm. Soft Comput. 2019, 23, 6345–6360. [CrossRef]
12. Dharani, N. P., and N. Gireesh. "Fusion of CT and PET Image of Lungs Using Hybrid Algorithms." Solid State Technology 63.6 (2020): 7706-7719.
13. F. Scotti, "Robust Segmentation and Measurements Techniques of White Cells in Blood Microscope Images", in Proc. of the 2006 IEEE Instrumentation and Measurement Technology Conf. (IMTC 2006), Sorrento, Italy, pp. 43-48, April 24-27, 2006. ISSN: 1091-5281. [DOI:10.1109/IMTC.2006.328170]
14. Dharani, N. P., and Polaiah Bojja. "Analysis and prediction of COVID-19 by using recurrent LSTM neural network model in machine learning." International Journal of Advanced Computer Science and Applications 13.5 (2022).
15. Dharani, N. P. "Detection of breast cancer by thermal based sensors using multilayered neural network classifier." International Journal of Engineering and Advanced Technology. (2019).

Cite this article as :

S. Mohan, Dr. D. Gowrisankar Reddy, "Enhanced Diabetes Prediction using Random Forest and XG Boost Machine Learning Classifiers with Dual Datasets", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 10 Issue 5, pp. 434-446, September-October 2023.
Journal URL : https://ijsrst.com/IJSRSET2310519