PREDICTION OF DIABETES IN FEMALES OF PIMA INDIAN HERITAGE USING XGB CLASSIFIER AND PERFORMING COMPARATIVE STUDY WITH OTHER CLASSIFICATION ALGORITHMS

[1]**Dr. G. Sai Chaitanya Kumar** Associate Professor, Department of Computer Science and Engineering, DVR & Dr. HS MIC College of Technology, Kanchikacherla, Andhra Pradesh, India.
V. Padmaja[2], **Assistant** Professor, Department of Computer Science and Engineering
DVR & Dr. HS MIC College of Technology, Kanchikacherla, Andhra Pradesh, India.
,[3]**P. Satya Kiran**, UG Student, Department of Computer Science and Engineering
DVR & Dr. HS MIC College of Technology, Kanchikacherla, Andhra Pradesh, India.
[4]E**. Satya Sriram Dilip** ,UG Student, Department of Computer Science and Engineering
DVR & Dr. HS MIC College of Technology, Kanchikacherla, Andhra Pradesh, India
,[5]**K. Lakshmi Kameswar**i , UG Student, Department of Computer Science and Engineering
DVR & Dr. HS MIC College of Technology, Kanchikacherla, Andhra Pradesh, India.
,[6]**D. Manisha** , UG Student, Department of Computer Science and Engineering
DVR & Dr. HS MIC College of Technology, Kanchikacherla, Andhra Pradesh, India.

Abstract:
Today, diabetes is a widespread illness that affects millions of people worldwide, with women bearing the brunt of its effects. Modern medical studies have used a variety of cutting-edge technology to diagnose patients and forecast their diseases using clinical data. Machine learning (ML) is one of these technologies, which allows for more precise diagnosis and prediction. This issue is taken into account as a binary classification issue.Algorithms for supervised learning have therefore been employed. We make use of the Kaggle dataset on female Pima Indians with diabetes. Extreme Gradient Boost is referred to as XGBoost. One of the supervised machine learning modules, it supports both classification and regression issues. In our project, we make use of the XGBoost module's XGBClassifier, whichis often used for classification as our project was to predict whether the person is diabetic or not. Along with XGBClassifier we also work with some other classification algorithms like Logistic Regression, Decision Tree Algorithm, Support Vector Machine,K-Nearest Neighbour and perform the comparative study on them based on the accuracyscore, Classification Report, Confusion Matrix and finally concludes the best model for our dataset

1.Introduction:

Diabetes is a condition marked by abnormally high blood sugar (glucose) levels. Heart attack or stroke, blindness, complications during pregnancy, and renal failure are just a few of the significant health issues that diabetes can bring on. One in every nine adult women[9] in the US, or around 15 million women, have diabetes. Due to a number of reasons, women are more likely than males to develop diabetes.

Hormonal variations: Throughout their lifetime, women experience hormonal changes that may influence[10] their chance of acquiring diabetes. For instance, gestational diabetes can occur in pregnant women, increasing their risk of type 2 diabetes in the future. During menopause, women also suffer hormonal changes that may impact their insulin sensitivity and raise their risk of diabetes.

Body Structure: Women tend to have more body fat than men, especially around the hips and thighs. This type of body fat is called "subcutaneous" fat and is less metabolically active than the "visceral" fat that accumulates around the organs, which is more common in men. Visceral fat is associated with insulin resistance and an increased risk of diabetes.

Lifestyle Factors: Women may be more likely to engage in sedentary behaviours, such as sitting for long periods, and may have less physical activity than men. They may also be more likely to consume a diet high in sugar and refined carbohydrates, which can increase the risk of developing diabetes.

Socioeconomic Factors: Women may also be more likely to experience socioeconomic factors that can increase their risk of developing diabetes, such as lower income, lower educational attainment, and less access to healthcare.Conclusively, the exact reasons why women may be more prone to diabetes than men are not fully understood, it is likely a combination of hormonal, metabolic, and

 DOI10.36893.JK.2023.V13I04N16.00198-00202

lifestyle[11] factors. It is important for both men and women to maintain a healthy lifestyle, including a balanced diet and regular physical activity, to reduce their risk of developing diabetes.

2. Related Work:

To predict diabetes in females of Pima Indian descent, many algorithms have been developed. The following is a quick discussion of these works. An approach to diagnose diabetes in female Pima Indian populations has been put out by Zolfagri et al. using an ensemble of neural networks and SVM. By combining fitting and generalisation in a new data mining approach, Pham et al. have successfully predicted diabetes.Aapproach for diagnosing diabetes utilising a semi-supervised learning technique that makes use of Laplacian SVM has been put out by Wu et al. Fuzzy c-means clustering and SVM were used to diagnose diabetes by Sanakal et al. Al et al. have diagnosed type-2 diabetes using the decision tree technique. SVM was used by Kumari et al. to classify diabetes. Dey et al. as well as Zou et al.have put in place a web-based strategy to forecast diabetes using ML techniques. None of the papers, however, have comprehensively explored all the well-known supervised learning techniques. Pradhan et al. used an artificial neural network (ANN) to predict diabetes. A data mining algorithm (CP-DMA) comparison has been proposed by Karthikeyani et al. to forecast the development of diabetes. A region-based SVM method has been presented by Karatsiolis et al. to aid in the medical diagnosis of Pima Indians. Bayes network has been utilised by Guo et al. to forecast type-2 diabetes. Using ML approaches, Maniruzamman et al. conducted a comparative analysis of diabetes mellitus data. Han et al. used quick miner software to evaluate the data. Saji et al. used a multilayer perceptron to predict diabetes. Jahangir et al. have proposed an expert system to predict diabetes using an autotuned multilayer perceptron. Li et al. have proposed a weight-adjusted approach to diagnose diabetes. In have proposed an accurate diabetes risk stratification using ML techniques. Like Pradhan et al.,Sivastava et al. have predicted diabetes using ANN approach. Kala et al. have proposed an intelligent hybrid system for a diabetes diagnosis. Chen et al. have proposed a hybrid prediction model to diagnose type-2 diabetes using decision trees and k-means.Kahramanli et al. have proposed a hybrid system for diabetes and heart disease. In authors have proposed a genetic algorithm approach using NN to diagnose Pima Indians diabetes.

3. Methodology:

This section describes the methodology used to predict the diabetes in female of Pima Indian heritage. To do so we used a dataset from Kaggle data repository which consists of following columns. As previouslydiscussed, it is a binary classification problem, we use different supervised learning techniques for prediction purpose

3.1 Dataset:

The dataset actually consists of 8 columns and the target column which describes the patient is diabetic or not

The columns are

1. Pregnancies
2. Glucose
3. Blood Pressure
4. Skin thickness
5. Insulin
6. BMI
7. DiabetesPedigreeFunction
8. Age
9. Outcome

3.2 Proposed model:

We are providing a model that is built using XGBClassifier because there are other models that are built using various other classification techniques. In essence, XGBClassifier is a class in the open-source[12] machine learning toolkit XGBoost (eXtreme Gradient Boosting), which is used to create decision tree-based models.Building classification models is the sole purpose of XGBClassifier. It is a gradient boosting implementation that uses an improved gradient boosting technique to increase the

 DOI10.36893.JK.2023.V13I04N16.00198-00202

precision and efficiency of model training.Large datasets, high-dimensional feature spaces, and a range of loss functions are all things that the XGBClassifier is renowned for handling. Also, it can be tailored and offers a wide range of characteristics that may be changed to enhance the model's functionality.
XGBClassifier is a well-known tool in industry and academia for a wide range of applications, including data science, finance, and healthcare. It is effective at creating accurate and effective classification models.

3.3 How it functions:

In order to increase the model's accuracy, the XGBClassifier[13] gradient boosting approach gradually adds decision trees. This is a general description of how XGBClassifier operates:
Initialization: As the first model, XGBClassifier initialises a single decision tree. Typically, this decision tree is a very basic model with only one node.
Model fitting: XGBClassifier develops a model by fitting it to the training set of data. It computes the difference between the target variable's expected and actual values during this operation. (i.e., the loss function).
Creating the following tree: XGBClassifier constructs the following tree by identifying the features that, in the preceding phase, had the greatest impact on the loss function. Then, depending on the previously discovered features, a new decision tree is created andadjusted to enhance the performance of the model.
Adding the new tree: XGBClassifier adds the new decision tree to the existing model and re-calculates the predicted values for the training data using the updated model.
Repeat: Until a stopping requirement is satisfied, such as when the maximum number of trees has been reached, or the model's performance has ceased to improve, XGBClassifier repeatedly finds the most crucial features and adds new decision trees.
Making predictions: Once the model is trained, XGBClassifier can be used to make predictions on new data by passing the input data through the decision trees to generate a predicted output.
In summary, XGBClassifier is an iterative algorithm that adds decisiontrees sequentially to optimize the model's performance, making it a powerful and effective tool for building classification models.
Other Classification Algorithms used:

1. Logistic Regression
2. Decision Tree
3. Support Vector Classifier
4. Random Forest Classifier
5. K Nearest Neighbour

We build model for XGBClassifier along with the other 5 algorithms mentioned above and suggest the best algorithm for classification generally
Algorithm 1: A general approach for classification and prediction
INPUT: Pima Indian diabetes dataset from Kaggle
OUTPUT: Accuracy Score, Classification Report, Confusion Matrix

1. Download and import the data set into Google Collaboratory
2. Gain insights from dataset
3. Provide the data to the XGBClassifier,

LogisticRegression, Decision Tree, Support Vector, Random Forest and KNN for learning individually

4. Test the models by performance metrics like accuracyscore, confusion matrix, Classification report
5. Finalize the model with best accuracyscore
6. Visualize the classification report

 DOI10.36893.JK.2023.V13I04N16.00198-00202

ISSN: 2278-4632
Vol-13, Issue-04, No.06, April : 2023

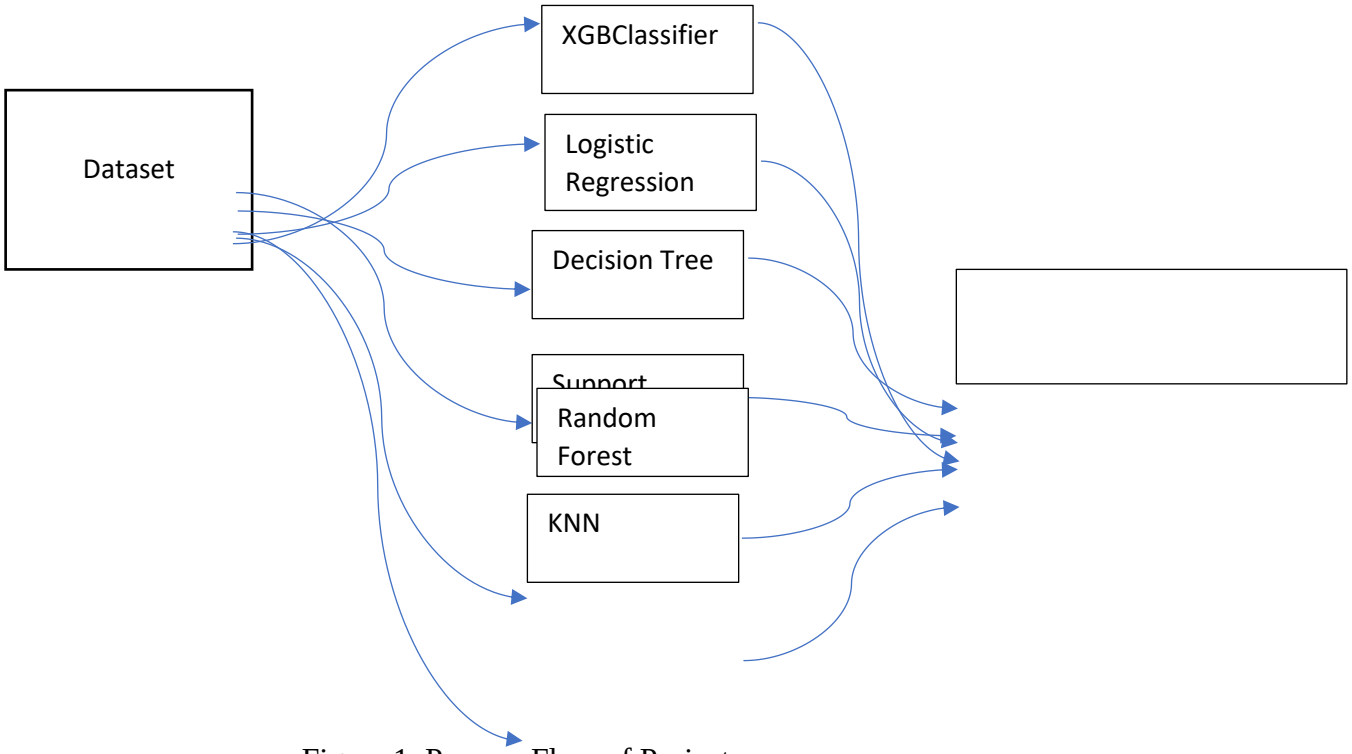

Figure 1. Process Flow of Project

In this section the results are analysed using Algorithm 1. We use Google Collaboratory an open-source platform to classify and predict diabetes in Pima Indian Heritage dataset. The results of differentmodel's build are represented in a tabular format with algorithm along with its accuracy score.

Analysis and Result:

NO	lgorithm	ccuracy ore
	GBClassifier	759740
	ogistic egression	759740
	ecision Tree	720779
	upport Vector	753246
	andom Forest	707792
6	KNN	0.662337

Table1: Accuracy scores

5. Conclusion:

We infer from the above data that XGBClassifier and Logistic Regression produced the same outcomes.As a result of its excellent accuracy and efficiency, the XGBClassifier is a potent machine learning algorithm for classification tasks that has grown in popularity in recent years. When using XGBClassifier, there are a number of critical procedures that must be taken, including as gathering

 DOI10.36893.JK.2023.V13I04N16.00198-00202

and pre-processing data, choosing pertinent features, and fine-tuning the model's hyperparameters for optimum performance

Due to its integrated regularisation approaches, XGBClassifier has the capacity to accommodate missing data and prevent overfitting, which is one of its main advantages. XGBClassifier is a suitable option for big data applications because it can also handle enormous datasets with high dimensions. However, mastery in machine learning and data analysis is also necessary for using XGBClassifier, as well as knowledge of the specific domain and the dataset being used. It is important to carefully select the right features and hyperparameters to avoid underfitting or overfitting the model.Overall, XGBClassifier is a valuable tool for building accurate and efficient classification models, and its popularity is likely to continue to grow as more applications are found for this powerful algorithm. Our work has really become easy by using google Collaboratory because it is an open-sourceand installation of libraries will be very easy and cell by cell execution was done which helps us to find errors easily in each part of our code.

6. References:

[1] Prediction of Diabetes in Females of Pima Indian Heritage: A Complete Supervised Learning Approach:SouravKumar Bhoi, Sanjaya Kumar Panda, Kalyan Kumar Jena, P. Anshuman Abhisekh, Kshira Sagar Sahoo, Najm Us Sama, Shweta Supriya Pradhan, Rashmi Ranjan SahooResearch and application of XGBoostin imbalanced data: Ping Zhang, Yiqiao Jia and Youlin Shang

[2]]Diabetes Prediction using Machine LearningTechniques: Mintushi Soni, Dr. Sunita Varma

[3] Study and Analysis of Decision Tree Based Classification Algorithms: Harsh H. Patel, Purvi Prajapati

[4] https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390

[5] []https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/

[6] https://towardsdatascience.com/log-book-xgboost-the-math-behind-the-algorithm-54ddc5008850

[7]]https://www.datacamp.com/tutorial/xgboost-in-python

[8] .Sai Chaitanya Kumar, Dr.Reddi Kiran Kumar, Dr.G.Apparao Naidu, "Noise Removal in Microarray Images using Variational Mode Decomposition Technique " Telecommunication computing Electronics and Control ISSN 1693-6930 Volume 15, Number 4 (2017), pp. 1750-1756

[9] [V. S. Rao, V. Mounika, N. R. Sai and G. S. C. Kumar, "Usage of Saliency Prior Maps for Detection of Salient Object Features," *2021* Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2021, pp. 819-825, doi: 10.1109/I-SMAC52330.2021.9640684

[10] G. S. C. Kumar, D. Prasad, V. S. Rao and N. R. Sai, "Utilization of Nominal Group Technique for Cloud Computing Risk Assessment and Evaluation in Healthcare," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp. 927-934, doi: 10.1109/ICIRCA51532.2021.9544895

[11] N. R. Sai, G. S. C. Kumar, M. A. Safali and B. S. Chandana, "Detection System for the Network Data Security with a profound Deep learning approach," 2021 6th International Conference on Communication and Electronics Systems (ICCES), 2021, pp. 1026-1031, doi: 10.1109/ICCES51350.2021.9488967

 DOI10.36893.JK.2023.V13I04N16.00198-00202