AI-Based Diabetes Prediction System

PHASE 5 DOCUMENTATION

PROJECT:AI-Based Diabetes Prediction System

OVERVIEW:

An AI-based diabetes prediction system is a technological solution that utilizes artificial intelligence and machine learning algorithms to assess an individual's risk of developing diabetes or to help manage the condition in those who already have it. Here's an overview of how such a system works:

Data Collection: The foundation of any AI-based prediction system is data. In the case of diabetes prediction, this can include personal health data such as age, gender, family history of diabetes, lifestyle factors (diet, exercise), and medical history (e.g., previous diagnoses of prediabetes or gestational diabetes).

Feature Selection: Relevant features or variables are extracted from the collected data. These features could include fasting blood sugar levels, body mass index (BMI), blood pressure, cholesterol levels, and more.

Data Preprocessing: Data preprocessing is essential to clean and transform the data into a suitable format for machine learning. This involves handling missing values, normalizing data, and dealing with outliers.

Machine Learning Algorithms: AI-based diabetes prediction systems typically use various machine learning algorithms, such as logistic regression, decision trees, random forests, support vector machines, and deep learning neural networks, to build predictive models. These models learn from historical data to identify patterns and relationships between the features and the likelihood of developing diabetes.

Training the Model: The system is trained on a labeled dataset, where the presence or absence of diabetes is known. This allows the machine learning model to learn how different variables correlate with the condition.

Model Evaluation: After training, the model's performance is evaluated using a separate set of data not seen during training (validation or test data). Common metrics for evaluating the model include accuracy, sensitivity, specificity, precision, and the receiver operating characteristic (ROC) curve.

Predictive Capabilities: Once the model is trained and evaluated, it can predict the likelihood of an individual developing diabetes in the future. Users input their relevant data, and the system provides a risk assessment.

Continuous Monitoring: Some systems are designed for continuous monitoring and can be integrated with wearable devices like smartwatches or fitness trackers. These devices can collect real-time data, providing insights into changes in health indicators and diabetes risk over time.

Alerts and Recommendations: Based on the prediction results, the system can generate alerts and recommendations for individuals. This might include advice on lifestyle changes, diet, exercise, or the need for regular check-ups with a healthcare professional.

Feedback and Improvement: Continuous learning is essential to improve the accuracy of the model. Feedback from users and updates to the model based on new data help refine predictions over time.

Privacy and Security: Diabetes prediction systems should prioritize data privacy and security to protect users' personal health information.

AI-based diabetes prediction systems have the potential to assist individuals and healthcare professionals in preventing and managing diabetes by providing early warnings and tailored recommendations. However, it's important to ensure that such systems are developed and deployed with ethical considerations and rigorous validation to ensure their effectiveness and safety in real-world applications. Additionally, they should be used as a supportive tool rather than a sole diagnostic method, with healthcare professionals involved in the decision-making process.

ABSTRACT:

Diabetes Prediction:

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis .According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes mellitus or imply diabetes is a disease caused due to the increase level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. However, early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as kidney, eye, heart, nerves, foot etc. Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is to help make predictions on medical data. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques .This project aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, KNN. This project also aims to propose an effective technique for earlier detection of the diabetes disease using Machine learning algorithms and end to end deployment using flask

CONTENT:

1. INTRODUCTION

2. RELATED WORKS

3. DATASET

4. PROPOSED METHODS

5. MODELING AND ANALYSIS

a) Logistic Regression

b) K-Nearet Nseighbors:

c) SVM

d) Naive Bayes:

e) Decision Tree:

f) Random Forest

g) adaboost Classifier

7. RESULTS AND DISCUSSION

9. CONCLUSION

INTRODUCTION:

All around there are numerous ceaseless infections that are boundless in evolved and developing nations. One of such sickness is diabetes. Diabetes is a metabolic issue that causes blood sugar by creating a significant measure of insulin in the human body or by producing a little measure of insulin. Diabetes is perhaps the deadliest sickness on the planet .It is not just a malady yet, also a maker of different sorts of sicknesses like a coronary failure, visual deficiency, kidney ailments and nerve harm, and so on. Subsequently, the identification of such chronic metabolic ailment at a beginning period could help specialists around the globe in forestalling loss of human life. Presently, with the ascent of machine learning, AI, and neural systems, and their application in various domains [1, 2] we may have the option to find an answer for this issue. ML strategies and neural systems help scientists to find new realities from existing well-being-related informational indexes, which may help in ailment supervision and detection. The current work is completed utilizing the Pima Indians Diabetes Database. The point of this frame work is to make an ML model, which can anticipate with precision the likelihood or the odds of a patient being diabetic. The ordinary distinguishing process for the location of diabetes is that the patient needs to visit a symptomatic focus. One of the key issues of bio-informatics examination is to achieve precise outcomes from the information. Human mistakes or various laboratory tests can entangle the procedure of identification of the disease. This model can foresee whether the patient has diabetes or not, aiding specialists to ensure that the patient in need of clinical consideration can get it on schedule and also help anticipate the loss of human lives .DNA makes neural networks the apparent choice. Neural networks use neurons to transmit data across various layers, with each node working on a different weighted parameter to help predict diabetes.

Presently, with the ascent of machine learning, AI, and neural systems, and their application in various domains [1, 2] we may have the option to find an answer for this issue.ML strategies and neural systems help scientists to find new realities from existing well-being-related informational indexes, which may help in ailment supervision and detection .The current work is completed utilizing the Pima Indians Diabetes Database.

CAUSES OF DIABETES:

Genetic factors are the main cause of diabetes. It is caused by at least two mutant genes in the chromosome 6, the chromosome that affects the response of the body to various antigens. Viral infection may also influence the occurrence of type 1 and type 2 diabetes. Studies have shown that infection with viruses such as rubella, Coxsackievirus, mumps ,hepatitis B virus, and cytomegalovirus increase the risk of developing diabetes.

Types of Diabetes

Type 1

Type 1 diabetes means that the immune system is compromised and the cells fail to produce insulin in sufficient amounts. There are no eloquent studies that prove the causes of type 1 diabetes and there are currently no known methods of prevention.

Type 2 Type 2 diabetes means that the cells produce a low quantity of insulin or the body can’t use the insulin correctly. This is the most common type of diabetes, thus affecting 90%of persons diagnosed with diabetes. It is caused by both genetic factors and the manner of living.

Data mining and machine learning have been developing, reliable, and supporting tools in the medical domain in recent years. The data mining method is used to pre-process and select the relevant features from the healthcare data, and the machine learning method helps automate diabetes prediction [14]. Data mining and machine learning algorithms can help identify the hidden pattern of data using the cutting-edge method; hence ,a reliable accuracy decision is possible. Data Mining is a process where several technique are involved, including machine learning, statistics, and database system to discover a pattern from the massive amount of dataset [15]. According to Nvidia: Machine learn in uses various algorithms to learn from the parsed data and make predictions.

RELATED WORKS:

Diabetes prediction is a classification technique with two mutually exclusive possible outcomes, either the person is diabetic or not diabetic. After extensive research ,we came to conclusion that although numerous classification techniques can be used for the purpose of prediction, the observed accuracy varied. On careful examination of the performance of techniques used in prevalent works, logistic regression, KNN, Naive Bayes[3], random forest, decision tree, and neural network [4], we found the mat par when applied to our dataset. KNN and logistic regression techniques were able to achieve 80% accuracy. The primary factor which influenced our algorithm selection was its adaptability and compatibility with future applications. The inevitable shift of data storage toward DNA makes neural networks the apparent choice. Neural networks use neurons to transmit data across various layers, with each node working on a different weighted parameter to help predict diabetes.

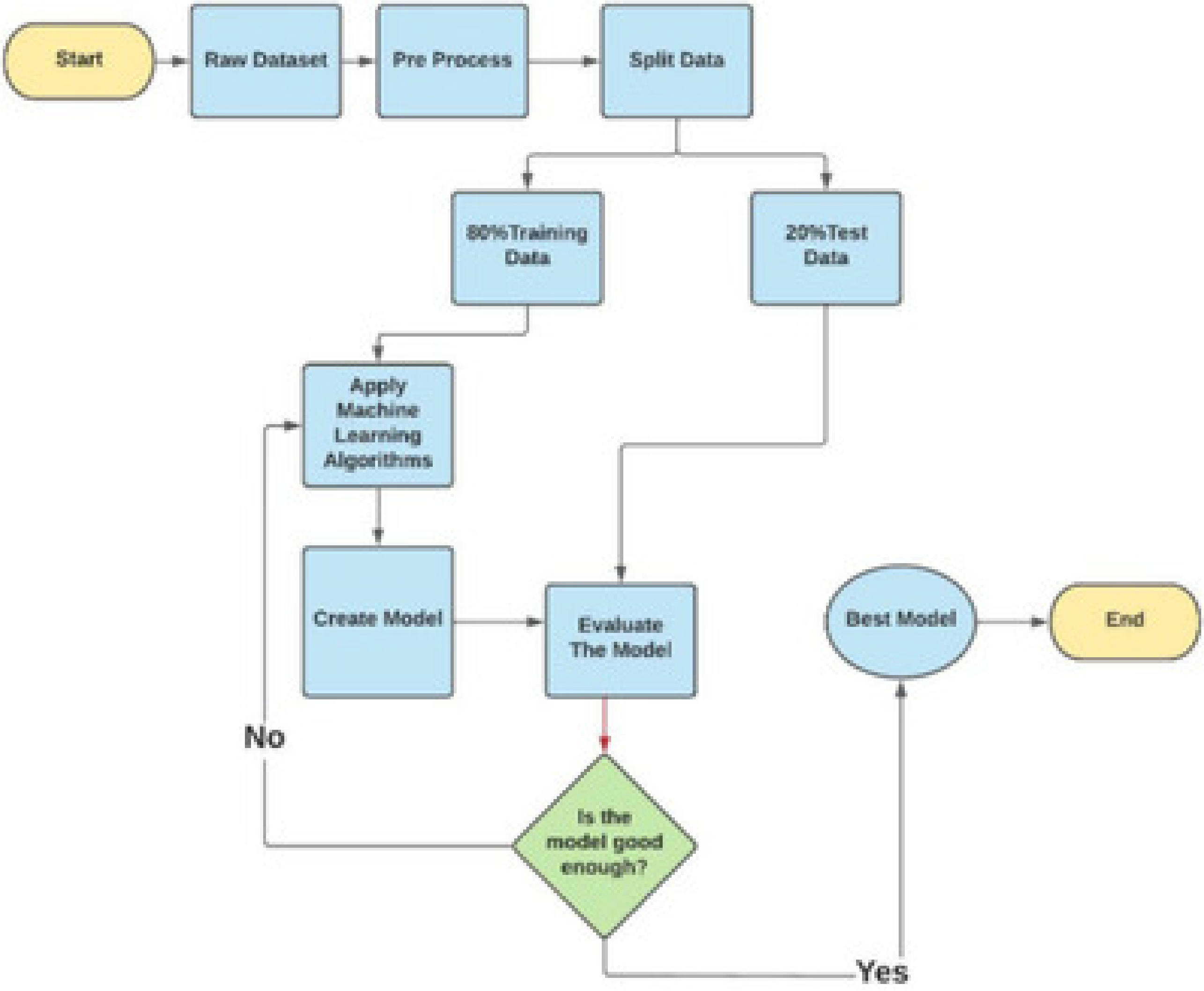
The point of this framework is to make an ML model, which can anticipate with precision the likelihood or the odds of a patient being diabetic. The ordinary distinguishing process for the location of diabetes is that the patient needs to visit asymptomatic focus. One of the key issues of bio-informatics examination is to achieve precise outcomes from the information .Human mistakes or various laboratory test scan entangle the procedure of identification of the disease. This model can for see whether the patient has diabetes or not, aiding specialists to ensure that the patient in need of clinical consideration can get it on schedule and also help anticipate the loss of human lives.

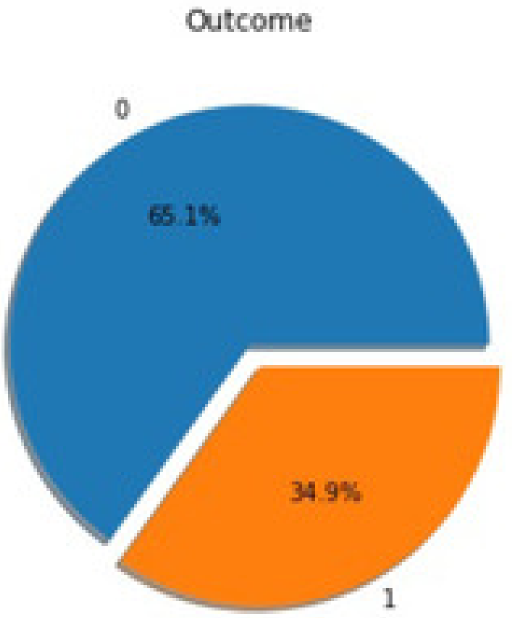
DATASET:

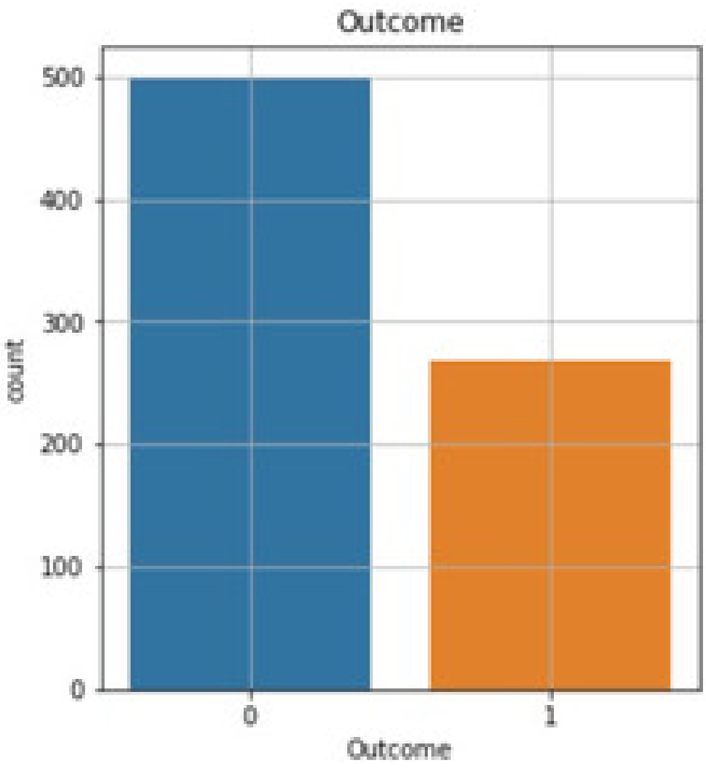
DATASET LINK:**https://www.kaggle.com/datasets/mathchi/diabetes-data-set**

PROPOSED METHOD:

This section describes the working procedures and implementation of various machine learning techniques to design the proposed automatic diabetes prediction system. Figure 1 shows the different stages of this research work. First, the dataset was collected and preprocessed to remove the necessary discrepancies from the dataset, for example, replacing null instances with mean values, dealing with imbalanced class issues etc. Then the dataset was separated into the training set and test set using the holdout validation technique. Next, different classification algorithms were applied to find the best classification algorithm for this dataset. Finally, the best-performed prediction model is deployed into the proposed website and smartphone application framework.







Data Pre-processing:

This phase of model handles inconsistent data in order to get more accurate and precise results like in this dataset Id is inconsistent so we dropped the feature .This dataset doesn’t contain missing values. So, we imputed missing values for few selected attributes like Glucose level, Blood Pressure, Skin Thickness, BMI and Age because the attributes cannot have values zero. Then data was scaled using Standard Scaler. Since there were a smaller number of features and important for prediction so no feature selection was done.

MODELING AND ANALYSIS:

Logistic Regression:

Logistic regression is a machine learning technique used when dependent variables are able to categorize. The outputs obtained by using the logistic regression is based on the available features. Here sigmoidal function is used to categorize the output.

K-Nearest Neighbors:

K-nearest neighbors (KNN) algorithm uses ‘feature similarity’ to predict the values of new data points which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. Predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variable for those k instances.

SVM:

SVM is superrbf kernel. If we select the hyper plane with low margin leads to miss classification.

Naive Bayes:

Naive Bayes classifiers are a collection of classification algorithms based on

Bayes’ Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

Decision Tree:

Decision tree is non parametric classifier in supervised learning. In this method all the details are represented in the form of tree, where leaves are corresponds to the class labels and attributes are corresponds to internal node of the tree. We have used Gini Index for splitting the nodes.

Random Forest:

Random forest is an ensemble learning method for classification. This algorithm consists of trees and the number of tree structures present in the data is used to predict the accuracy. Where leaves are corresponds to the class labels and attributes are corresponds to internal node of the tree. Here number of trees in forest used is 100 in number and Gini index is used for splitting the nodes.

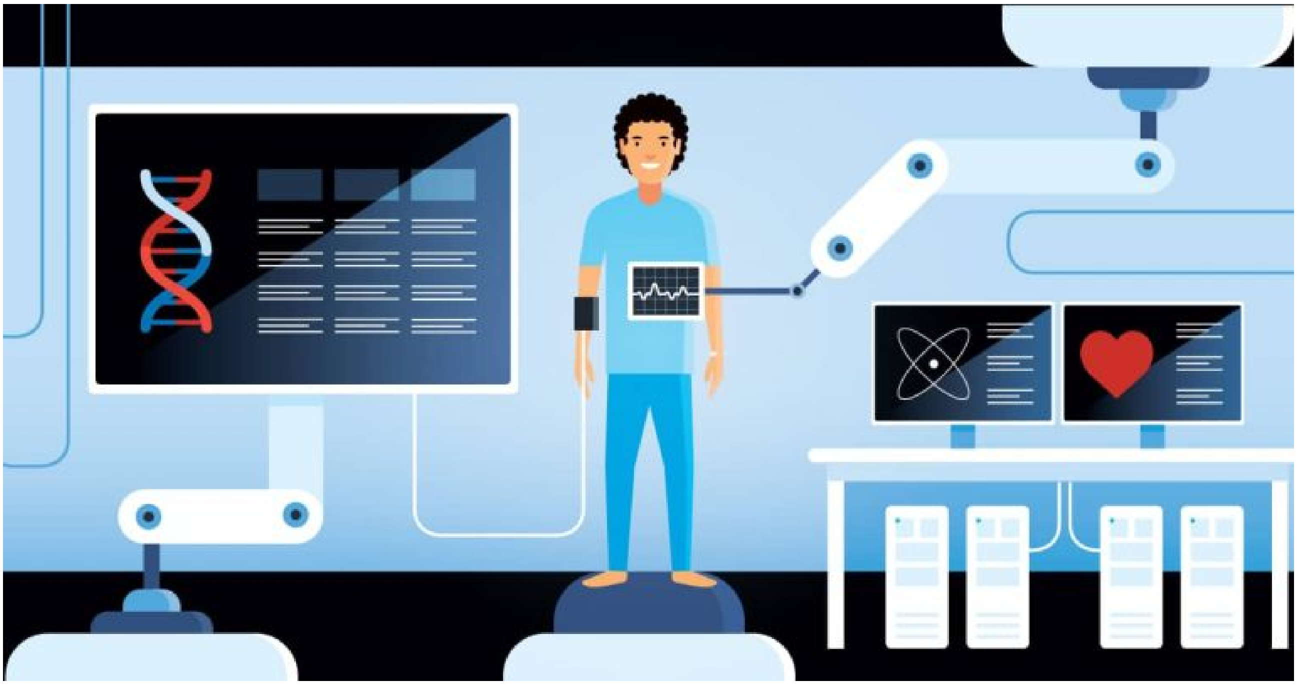
AdaBoost Classifier:

Boosting is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models are added.

AdaBoost was the first really successful boosting algorithm developed for the purpose of binary classification. AdaBoost is short for Adaptive Boosting and is a very popular boosting technique that combines multiple “weak classifiers” into a single “strong classifier”. It was formulated by Yoav Freund and Robert Schapire. They also won the 2003 Gödel Prize for their work.

RESULTS AND DISCUSSION:

Machine learning classification algorithms developed for prediction of diabetes in earlier stage. We used 70% of data for trining and 30% of data for testing. In this ratio of data splitting Here we found that Random Forest Classifier predicted with 99% of accuracy as highest accuracy for the dataset. Comparison of results of all the implemented classifiers are listed in below.



CONCLUSIONS:

Diabetes can be a reason for reducing life expectancy and quality. Predicting this chronic disorder earlier can reduce the risk and complications of many diseases in the long run. In this paper, an automatic diabetes prediction system using various machine learning approaches has been proposed. The open-source Pima Indian and a private dataset of female Bangladeshi patients have been used in this work. SMOTE and ADASYN preprocessing techniques have been applied to handle the issue of imbalanced class problems. This research paper reported different performance metrics, that is, precision, recall, accuracy, F1 score, and AUC for various machine learning and ensemble techniques. The XGBoost classifier achieved the best performance with 81% accuracy and an F1 score and AUC of 0.81 and 0.84, respectively, with the ADASYN approach. Next, the domain adaptation technique has been applied to demonstrate the versatility of the proposed prediction system. Finally, the best-performed XGBoost framework has been deployed into a website and smartphone application to predict diabetes instantly. There are some future scopes of this work, for example, we recommend getting additional private data with a larger cohort of patients to get better results. Another extension of this work is combining machine learning models with fuzzy logic techniques and applying optimization approaches.