DIABETES PREDICTION USING MACHINE LEARNING

[1]ASHWINI R, [2]S M AIESHA AFSHIN, [3]KAVYA V, [4]DEEPTHI RAJ

Department of Telecommunication Engineering
Dayananda Sagar College of Engineering Bengaluru, India

***Abstract*: Diabetes is a very common disease that affects people all over the world. Diabetes raises the risk of long-term complications such as heart disease and kidney failure. If this disease is detected early, people may live longer and healthier lives. Different supervised machine learning models trained on appropriate datasets can aid in the early detection of diabetes. The goal is to develop effective machine-learning-based classifier models for detecting diabetes in people using clinical data. K-nearest neighbour (KNN), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine are among the machine learning algorithms that will be trained using various datasets (SVM).We used efficient pre-processing techniques to improve the model's accuracy. Furthermore, we identified and prioritised a number of risk factors using various feature selection approaches. Extensive experiments have been carried out to evaluate the model's performance on various datasets. When our model is compared to some recent studies, the results show that the proposed model can provide better accuracy ranging from 2.71 percent to 13.13 percent depending on the dataset and ML algorithm used. Finally, the most accurate machine learning algorithm is chosen for further development. We use the Python Flask web development framework to integrate this model into a web application. The findings suggest that using an appropriate preprocessing pipeline on clinical data and applying ML-based classification can accurately and efficiently predict diabetes.**

Keywords*: *Decision Support Systems, Diabetes prediction, Machine learning, Support Vector Machine, Random Forest, K-Nearest Neighbor, Logistics Regression

INTRODUCTION

Diabetes mellitus is a chronic disease that is caused by a high sugar level in the circulatory system. Classification strategies are widely used in the medical field for categorizing data into different classes based on some constraints, as opposed to an individual classifier. Diabetes is a disease that impairs the body's ability to produce the hormone insulin, causing carbohydrate metabolism to become abnormal and blood glucose levels to rise. According to the WHO (World Health Organization) report on November 14, 2016 in the world diabetes day-Eye on diabetes, 422 million adults have diabetes, with 1.6 million deaths, as the report indicates, it is not difficult to guess how serious and chronic diabetes is. The four primary diseases, namely cardiovascular diseases, cancer, chronic respiratory diseases, and diabetes, kill more than 18% of the world's population and have become a major public health concern. Diabetes mellitus can be caused by obesity, age, a lack of exercise, hereditary diabetes, high blood pressure, a poor diet, and other factors. Diabetes increases the risk of developing diseases such as heart disease, stroke, kidney failure, nerve damage, eye problems, and so on.

With the right medical care, negative effects of diabetes can be prevented if it is identified in its early stages. Methods using machine learning can help in early disease detection. Since ML techniques allow computers to learn and gain knowledge from previous experience or a pre-defined dataset, they are used to develop prediction models.Since studies have shown that machine-learning algorithms perform better in diagnosing various diseases, many researchers are conducting experiments for disease diagnosis using various classification algorithms of machine learning approaches like J48, SVM, Naive Bayes, Decision Tree, Decision Table, etc.

Data ,feature and software tool

The numerous machine learning techniques for diabetes prediction that were outlined above have been used in this.

- Step1: Import the necessary libraries and the diabetic dataset as the first step.
- Step 2: Preprocess the data in step two to fill in any gaps.
- Step 3: Apply 80 percent scaling to divide the collection of data into a training set and a test set of 20 percent each.
- Step 4: Choose a machine learning algorithm from the list provided, such as K-Nearest Neighbor, Support Vector Machine, Decision Tree, Logistic Regression, and Random Forest.
- Step 5: Using the training set, develop a model classifier using the aforementioned machine learning technique.
- Step 6: Use the test set to evaluate the classifier model for the aforementioned machine learning technique.
- Step 7: Execute a comparative analysis of the test performance results for each classifier in step 7.
- Step 8: Choose the algorithm that performs the best after assessing data based on different parameters.

LITERATURE SURVEY

[1] Sneha, N. and Gangil,T., Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big Data, 6(1), p.13.(2019):

The focus of the authors' work has been on choosing the characteristics that will aid in the early detection of Diabetes Miletus utilizing predictive analysis and machine learning approaches. The UCI machine repository served as the source of the dataset. The classification process has involved the usage of 15 attributes. The classifiers employed are Support Vector Machine, Random Forest, and Naive Bayes, with accuracy rates of 77.73%, 75.39%, and 73.48% respectively.

[2] Sisodia, D. and Sisodia, DS, Prediction of diabetes using classification algorithms. Procedia computer science, 132,pp.1578-1585(2018):

A mechanism to help with disease estimation, including diabetes, is being designed by authors using the Indians Pima Diabetes Selected-database (PIDD). Three machine learning identification algorithms, Bayes Naive, SVM, and Decision Tree, are used in this study with accuracy rates of 76.3 percent, 65.1 percent, and 73.82 percent to identify diabetes at an early stage.

[3] Rahul Joshi and Minyechil Alehegn(2017):

The authors proposed ML techniques that are used to guess the data set in the initial phase in order to save a life. Using the KNN and the Nave Bayes algorithms. In this study, the proposed method provided high accuracy with an accuracy value of 90.36 percent, while decision Stump provided less accuracy than the others with an accuracy value of 83.72 percent. The most commonly used predictive algorithms in this context are Random Forest, Naive Bayes, and KNN.

The single algorithm provided less precision than the ensemble algorithm. In the majority of the tests, the decision tree performed admirably. The tools used in this hybrid study to predict diabetes data are Java and Weka. They proposed an ensemble approach to analysis and prediction of diabetes diseases using machine learning algorithms. The following algorithms are used to create this system as an ensemble hybrid model: KNN, Naive Bayes, Random Forest, and J48 are techniques used to improve performance and accuracy. J48 is one of the most widely used and accurate. All of these algorithms are used to improve accuracy and are more advanced than others. They discussed the ML techniques that are used to guess datasets at an early stage in order to save lives.

[4] Ridam Pal, Dr.Jayanta Poray and Mainak Sen (2017):

They presented Diabatic Retinopathy (DR), which is one of the leading causes of blindness in diabetic patients. They examined the performance of a set of machine learning algorithms and verified their performance on a specific data set.

[5] Dr.M.Renuka Devi and J. Maria Shyla(2016):

They discussed the analysis of various mining skills for predicting diabetes using Naive Bayes, Random Forest, Decision Tree, and J48 algorithms.

[6] Veena Vijayan V.And AnjaliC (2015):

Diabetes, according to the authors, is caused by high blood sugar levels. Various computer information systems that use classifiers to predict and diagnose diabetes have been described, including decision trees, SVM, Naive Bayes, and ANN algorithms. They proposed a technique for diabetic patient diagnosis called Prediction and Diagnosis of Diabetes Mellitus - A Machine Learning Approach. They used this technique to provide high accuracy based on the AdaBoost algorithm. We can collect the localdataset by using the relating mean value as a part of the global dataset. As well as we can train and validate the dataset collection byusing four base classifiers as a Decision tree, Support Vector Machine, Native Bayes and Decision stump.

After that, we can also calculate Body Mass Index (BMI) by using the height and weight of a person. We can easily obtain the performance accuracy, sensitivity, specificity, and error rate by analyzing all of these techniques and using the AdaBoost algorithm. They discussed diabetes, which is caused by an increase in blood sugar levels. Several computerized information systems utilizing classifiers for anticipating and diagnosing diabetes were described using decision tree, SVM, Naive Bayes, and ANN algorithms.

[7] Santhanam, T. and Padmavathi, M.S., 2015. K-means and genetic algorithms are used to reduce dimension by integrating SVM for diabetes diagnosis.

The authors proposed a k-means approach for removing noisy data and genetic algorithms for determining the optimal set of features using Support Vector Machine (SVM) as a classifier. For the reduced dataset of Pima Indians Diabetes from the UCI repository, the proposed model achieved an average accuracy of 98.79 percent.

ATTRIBUTES

The UCI Machine Learning Repository provided the Prima Indian Diabetes Dataset, which was used in this study. The data was

obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. Pregnancy record, BMI, insulin level, age, glucose concentration, diastolic blood pressure, triceps skin fold thickness, diabetes pedigree function, and so on comprise the dataset. This dataset contains data from 768 patients, all of whom are female and over the age of 21. In the dataset, the number of true cases is 268 (34.90 percent) and the number of false cases is 500 (65.10 percent).
We chose eight distinct parameters for data preparation in the following, such as,

- Pregnancies: Pregnancy records for women
- Glucose: Plasma glucose concentration
- Blood Pressure (mm Hg)
- Skin Thickness: Triceps with skin fold thickness (mm)
- Insulin: Patients, 2-Hour level of serum insulin
- BMI
- Diabetes pedigree function
- Age: Age (years)

OBJECTIVE:

• To create a system that predicts a patient's diabetic risk level with greater accuracy.

• The goal of developing a machine learning model is to accurately predict whether or not the patients in the dataset have diabetes.

• Demonstrating the effectiveness of computer-based methods.

• Helped to shorten the time between diabetes onset and clinical diagnosis.

BLOCK DIAGRAM:

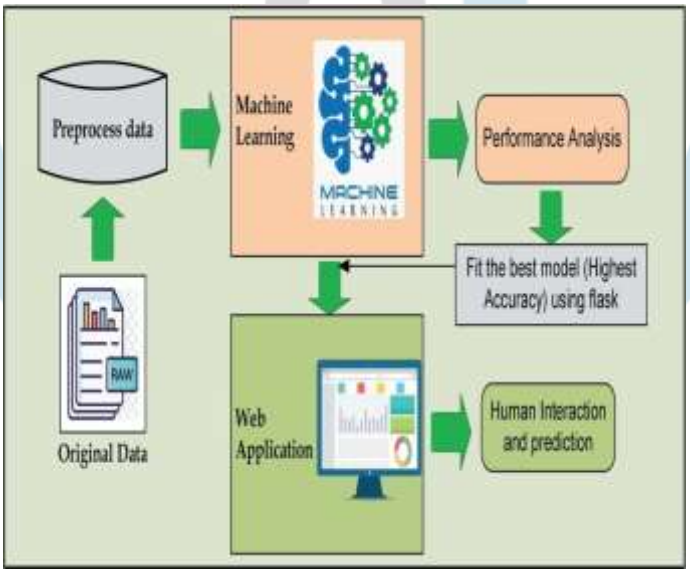

Fig 1. Block Diagram

Each phase of the proposed ML-based diabetes prediction model is depicted in Fig. Every dataset is pre-processed in the first phase. The pre-processed datasets are then fed into the various machine learning algorithms in the second stage. The output of the models is then analysed using various metrics in the third phase. Later on, the model with the highest accuracy is used to detect diabetes in any individual and is integrated with a web-based application. This web-based application was created with the Python programming language and Flask.

In a nutshell, the following are the contributions of this research:
Our first contribution is to train several machine learning algorithms for diabetes detection using four different clinical datasets. All datasets are pre-processed using various pre-processing techniques.
Second, the performance of each ML algorithm on four datasets is evaluated using several parameters such as precision, recall, f1-score, ROC curve, and accuracy. Furthermore, we identified several important features or attributes by employing various feature selection methods such as correlation, chi-square, and so on. The feature selection methods identify the features that are most closely related to diabetes disease. The performance of the ML algorithms on the reduced set of attributes was also examined.
Third, based on the performance results, a web-based application is developed to predict individuals diabetes.

METHODOLOGY

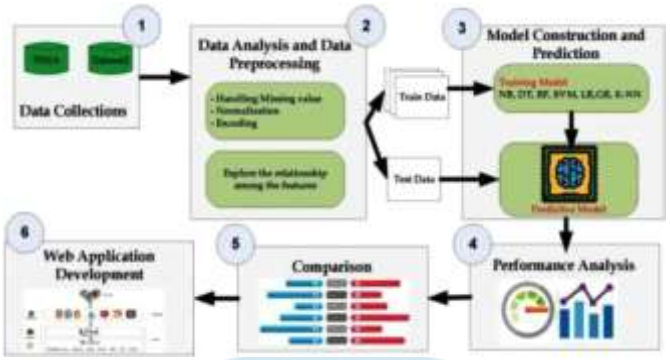

Fig 2. Work flow Diagram

1. Data Collection: To ensure the model's robustness, we collected datasets with varying numbers of factors or features. The datasets were compiled from a wide range of sources, including diabetes statistics and health characteristics obtained from people all over the world as well as various health institutes.

2. Data Analysis and Preprocessing: Several pre-processing techniques are used on the datasets before feeding them into the machine learning model to improve the model's performance. Among the pre-processing tasks are the removal of outliers and the handling of missing values, as well as data standardisation, encoding, and soon.

- Outliers Removal - The dataset may contain values that are outside of acceptable bounds and have a high variation from the rest of the attribute's value. The value of such attributes may degrade the performance of the machine learning algorithm. We used the IQR (Inter-quartileRange) method to eliminate such outliers.
- Handling missing values - To improve model performance, the mean value of each attribute was used to handle missing values.
- Label Encoding - Label encoding is the process of converting text/categorical value labels into a numerical format that machine learning algorithms can understand. The categorical values of Junkfood consumption status yes to '1′ and No to '0′, for example, have been converted.

3. Model Construction and Prediction: To build the predictive model, 80 percent of the pre-processed data was used for training, while the remaining 20% was used for testing.

4. Performance Analysis: We analysed the proposed model's results in terms of several performance metrics. The best algorithm for web application development is the one that provides the highest prediction accuracy.

5. Performance Comparison: In this step, the proposal's accuracy was compared to some recent works on diabetes prediction. The performance results indicate that the proposal has the potential to outperform previous related research.
*6.*Web Application Development: We used the Flask micro-framework and the best model to create a smart web application. To predict diabetes, a user must fill out a form with the required number of diabetes-related parameters. The application, which is hosted on a server, predicts the outcomes using the machine learning model that was chosen. In the following sections, we describe the machine learning algorithms that were used.

ADOPTED MACHINE LEARNING ALGORITHMS

- ***KNN***

The K-Nearest Neighbor algorithm is a classification algorithm that is simpler and easier to use than other data mining techniques. This technique classifies new belongings using a similarity measure where the value of k is always a positive integer number. This algorithm stores training data based on the neighbours or closest prediction of test data when it is complete.

i. Determine k, which is the number of neighbours nearby.

ii. Calculate the distance between the instance and the training samples.

In this step, the remoteness of the training samples is sorted, and the closest neighbour based on the shortest distance is determined.

iv. In this step, we obtain all of the classes from all of the training data.

v. As the query instance's prediction value, use the majority of the class of closest neighbours.

- ***Random forest***

It is supervised learning, and it is used for classification as well as regression. The random forest [1] logic is based on a bagging

technique to generate random sample features. The decision tree differs from the random forest in that the process of finding the root node and splitting the feature node is done at random. The steps are outlined below .Load the data where I tconsists of ml features representing the behavior of the dataset.

a. Load the data, which consists of ml features representing the dataset's behaviour.

b. The random forest training algorithm is known as the bootstrap algorithm or bagging technique, and it is used to randomly select n features from m features, i.e. to create random samples, this model trains the new sample to out of bag sample (1/3rd of the data) used to determine the unbiased OOB error.
c. Repeat the steps until there are n trees.
d. Determine the total number of votes received by each tree for the predicting target. The random forest's final prediction is the class with the most votes.

- ***Logistics Regression(LR)***

In the early twentieth century, logistic regression was mostly used in biological research and applications. Logistic Regression (LR) is a machine learning algorithm that is commonly used when the target variable is categorical. Recently, LR has become a popular method for solving binary classification problems. Furthermore, it displays a discrete binary product between 0 and 1. Logistic Regression calculates the relationship between feature variables by assessing probabilities (p) with the underlying logistic function.

- ***Support Vector Machine***

SVM's have proven to be extremely effective for a variety of data classification tasks. It attempts to find the best separating hyperplane between classes by locating the set of points on the class descriptors' edges. The margin is the distance between the classes. SVM algorithms find a margin with the greatest possible distance. The greater the margin, the higher the classification accuracy of the classifier. The data points on the border are referred to as support vectors. As a result, the name support vector machine was coined. The remaining training samples are thrown away. SVM's can achieve good performance even with small training samples because they use fewer training samples.

SVMs are primarily intended to deal with linearly separable binary classification data. Several modifications have been proposed in order to apply it to multi-class classification problems. Similarly, it can be used to classify nonlinear cases by employing kernel techniques. These kernels use nonlinear to linear space mapping.

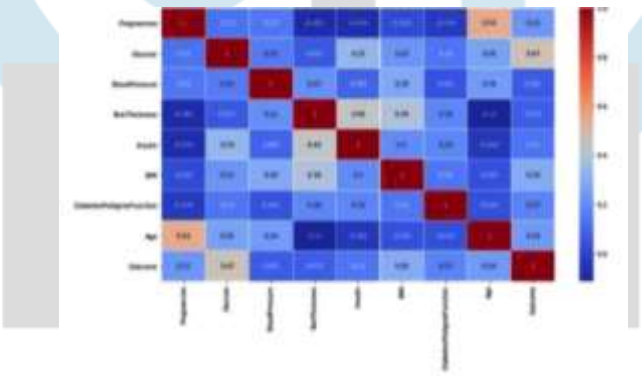

Fig 3. Correlation matrix for correlation analysis

Finally, the webpage displays the anticipated outcome (Step-5)

```
diabetes(root)
    |____templates
          |___index.html
    |____static
          |____css
          |_____js
    |____app.py
    |_____Diabetes.pkl
```

Fig 4. File structure of the web application

1. Web application development using flask

With Flask, users can add application functionality as though it were already included in the framework. Flask is a Python-based microweb platform. Demonstrates the developed application's core file formats and the four different programme modules that made up the development process:

- This file, model.pkl, includes the machine learning prediction model.
- app.py-This package contains Flask APIs that compute the projected value using our model, return it, and accept Diabetes information via GUI or API calls.
- Template-The HTML form (index.html) in this folder lets users enter details about their diabetes and displays the anticipated result.
- Static – The css file used in this folder has the styling needed for HTML form.

The proposal's application process is described in Fig

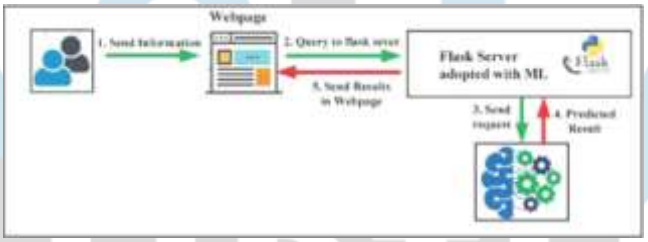

Fig 5 Working flow of the web application

2. RESULTS:

Accuracy of the models:

Algorithm	Accuracy
Logistic Regression	**83.11%**
KNN	**77.27%**
Random Forest Classifier	**79.87%**
Support Vector Machine	**81.81%**

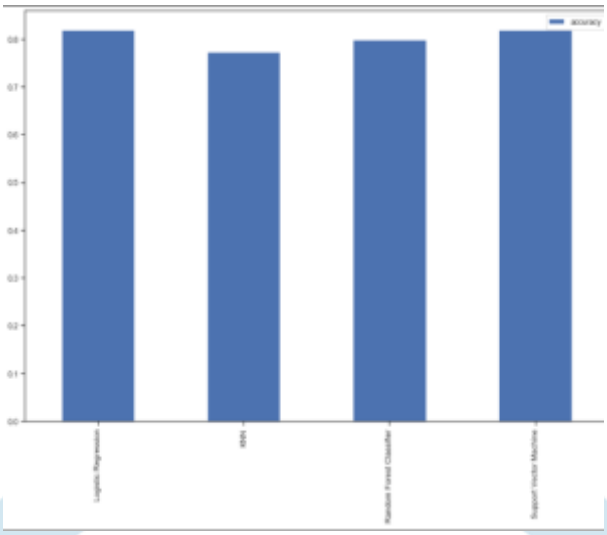

Fig 6. Accuracy graph

Website of Diabetes Prediction:

After clicking Predict probability:

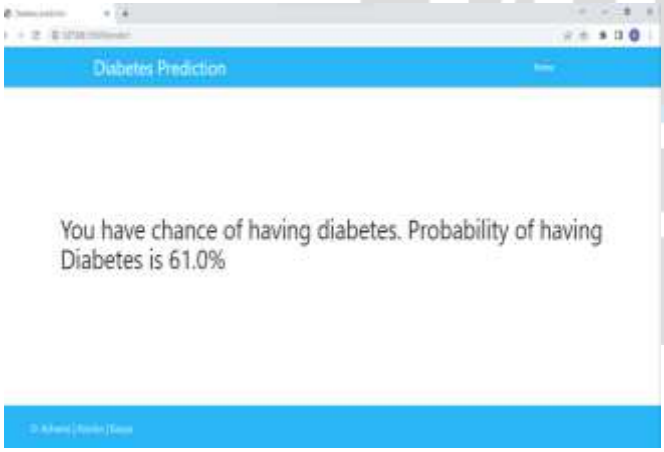

Fig 7. Website design of our project

CONCLUSION:

Doctors can diagnose and treat diabetes with the use of machine learning. We shall get to the conclusion that the performance of the machine learning models will be improved by increasing the classification accuracy.

The effectiveness of each classification technique, including logistic regression, K-nearest neighbours, SVM, and random forest, is

evaluated in terms of accuracy rate.
With the use of cutting-edge computational techniques and the availability of a sizable number of epidemiological and genetic diabetes risk datasets, machine learning has the potential to completely transform the ability to forecast the risk of developing diabetes.

The major goal is to develop and apply a machine learning system for predicting diabetes and to evaluate the method's effectiveness. SVM, KNN, logistic regression, and random forest are used in the suggested method approach. The method may also assist researchers in creating a precise and useful tool that would sit at the clinicians' table and assist them in making better decisions about the disease status.

We also discover that the existing system's accuracy is less than 70%, which is why we advise adopting a set of classifiers known as associative approaches. The combined approach makes use of the values added from two or more other techniques. Our approach gave us accuracy of more than 80 percent.

REFERENCES:

[1] Sneha, N. and Gangil, T., 2019. Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of BigData ,6(1), p.13.(2019)

[2] Sisodia,D. and Sisodia,DS,2018.Prediction of diabetes using classification algorithms.Procedia computer science,132, pp.1578-1585. (2018)

[3] Rahul Joshi and MinyechilAlehegn,- Analysis andprediction of diabetes diseases using machine learning algorithm‖: Ensemble approach, International Research Journal of Engineering and Technology Volume:04Issue:10 | Oct-2017

[4] RidamPal, Dr.Jayanta Poray,and MainakSen,
Application of Machine Learning Algorithms on Diabetic Retinopathy‖ ,2017 2nd IEEE International Conference on Recent Trends in Electronics Information & Communication Technology, May19-20,2017,India.

[5] Dr.M.Renuka Devi and J.Maria Shyla,-Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus 1, International Journal of Applied Engineering Research ISSN0973-4562 Volume11, Number 1(2016)

[6] Veena Vijayan V. And Anjali C, Prediction and Diagnosis of Diabetes Mellitus, -A Machine Learning Approach ,2015 IEEE Recent Advances in Intelligent Computational Systems(RAICS)|10-12 December 2015 |Trivandrum.

[7] Santhanam, T.and Padmavathi. M.S. ,2015. Application of K-means and genetical gorithms for dimension reduction by integrating SVM for a diabetes diagnosis. Procedia Computer Science ,47, pp.76-83. (2015).