A Review On Designing Of Diabetes Prediction

Prof. Swati Shamkuwar
Asst. Prof Dept of Information Technology G.H.Raisoni College of Engineering, Nagpur.

Bhargavi Chimote
Department of IT
GHRCE
Nagpur

Harshita Zode
Department of IT
GHRCE
Nagpur

Mitali Bhoyar
Department of IT
GHRCE
Nagpur

Abstract – Diabetes Prediction is a model that uses a concept of Machine Learning to predict a possibility of having diabetes to a person based on their medical history and personal information. The model's training and evaluation data were collected from the Diabetes Database. Using feature selection technique, the data is cleaned and made available for analysis.

In this project the main focus is on the Diabetes prediction. Here we are using the logistic regression and Naïve Bayes Classifier algorithm for prediction. Utilizing measures for accuracy, precision, remember, and F1-score, the model's success is assessed. The results show that the Naïve Bayes Classifier algorithm correctly predicts diabetes 80% of the time. By offering a dependable and effective tool for the early detection and prevention of diabetes, this initiative illustrates the potential of machine learning in healthcare.

Keywords: *Logistic Regression, Naïve Bayes, Diabetes Prediction, Machine Learning*

Introduction

Millions of individuals globally suffer from long-term diabetes, which has become more common in recent years. Effective diabetes treatment and improved patient outcomes depend on early diagnosis and prevention of the disease. By offering predictive models that can help with illness detection and control, machine learning, a branch of artificial intelligence, has the potential to revolutionize healthcare.
In this project, for developing a machine learning model, Python is used. that forecasts a person's probability of having diabetes based on their medical history and personal information. The Native American Diabetes Database, which includes information from female patients over the age of 21, served as the model's training data. The collection contains information about, among other things, age, BMI, blood pressure, and glucose levels.

To clean up the data and get it ready for analysis, we use techniques for feature selection and data preprocessing. Logistic regression is examples of machine learning methods used for prediction. Using measures for accuracy, precision, remember, and F1-score, we assess the model's success.

The outcomes of this study provide a secure and efficient instrument for the early diagnosis and prevention of diabetes, showcasing the potential of machine learning in healthcare. improved patient treatment and management choices can be made with the help of the model, which will result in improved health outcomes.

I. LITERATURE SURVEY

The research being conducted on using Python's machine learning tools to predict diabetes is significant and developing. Several machine learning models have been investigatedin various research papers for the prediction of diabetes. Electronic health records, the native
American Diabetes dataset, and other datasets dedicated to diabetes are only a few of the datasets to which these models have been applied.

Many research have looked into the use of several characteristics, including age, BMI, blood sugar levels, and family history, for the

prediction of diabetes. The most crucial elements for diabetes prediction have been identified using feature selection techniques like correlation analysis and recursive feature removal.

Moreover, the performance of the chosen models has been optimized. Accuracy, precision, recall, and F1-score have all been utilised as evaluation measures to evaluate the performance of the models.

Research has also looked into recurrent neural networks and convolutional neural networks, two examples of deep learning techniques, as possible methods for diabetes prediction. Moreover, exploratory data analysis and performance visualisation have both made use of visualisation techniques like visualisations, scatterplots, and graphical representations.

Overall, the research indicates that Python machine learning methods can be used to predict diabetes with high levels of precision and accuracy. Machine learning models for diabetes prediction are simple to implement and deploy because a variety of Python libraries are readily available, including Scikit-learn, Pandas, and Matplotlib.

II. PROPOSED METHODOLOGY

The proposed automated diabetes prediction system's appropriate procedures and implementation of several machine learning algorithms are described in this section. Figure 1 depicts the many phases of this research project.

Data collection is the initial step, during which diabetes-related data is gathered and preprocessed. Exploratory data analysis (EDA) is the next step, which involves using statistical and
graphical tools to analyze the data's features and detect anomalies and missing values.Then, using methods like correlation analysis or recursive feature reduction, feature selection is carried out to determine which characteristics are most important in predicting diabetes.

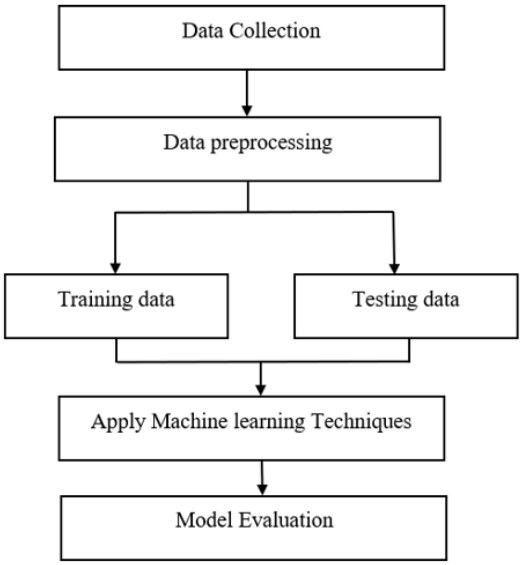

Figure 1. Proposed Methodology

To prevent overfitting, the models are trained on preprocessed data using cross-validation methods. Following that, the effectiveness of the trained models is assessed using suitable evaluation measures including accuracy, precision, recollection, and F1-score.

The model's performance is enhanced through hyperparameter tuning utilising methods like grid search and random search. In the end, the model is used to make predictions based on fresh data. The Python modules Scikit-learn, Pandas, and Matplotlib can be used to implement this suggested methodology in python.

Naïve Bayes Classifier Algorithm

A probabilistic machine learning method called Naive Bayes can be used for categorization activities like predicting diabetes. The Naive Bayes algorithm is trained on a collection of input characteristics (such as age, BMI, glucose levels, blood pressure, etc.) and associated binary labels showing whether the patient has diabetes or not in order to forecast a patient will develop diabetes.The Naive Bayes method first determines the conditional probability of each input feature given the class label (diabetes or non-diabetes). This bayesian probability is then combined using the Bayes theorem. The algorithm then predicts the class label with the greatest probability chance as its result.

The " Naïve " in Naïve Bayes originates from the concept that the input attributes are conditionally independent given the class label. The Naive Bayes algorithm remains capable of generating good results in many situations, especially when the set of input characteristics is high compared to the size of the training dataset, even though that assumption is frequently broken in reality.

We must first preprocess the data to ensure that it is in an appropriate shape for the algorithm before we can use the Naive Bayes algorithm for diabetes forecast in Python. The dataset is then divided into training and testing groups, and the model is trained on the training data using scikit-learn's Naive Bayes implementation. The model's performance is then assessed on the testing data using measures like accuracy, precision, recall, and precision.

III. SYSTEM FRAMEWORK

Data Collection: Obtaining the Native Americans Diabetes Database from the UCI Machine Learning Library or another trustworthy source.

Data preprocessing: In order to get the data ready for analysis, preprocessing involves cleansing, organizing, and transformation. Taking care of missing numbers, looking for duplicates, and eliminating errors are all included in this.

Feature Selection: Selecting the most crucial characteristics that have a big effect on diabetes outlook is known as feature selection. This involves finding and eliminating characteristics that are not important to the prediction.

Data Splitting: Splitting the data into training and evaluation groups is known as data splitting. As a result, we can use some of the data to train the model, and the remaining data

to test the model's performance.

Model Training: Model training is the process of using training data to develop machine learning algorithms. This involves selecting the suitable algorithms, such as random forest, decision trees, logistic regression, and k-nearest neighbors, and adjusting their hyperparameters.

Model Evaluation: Analyzing the accuracy, precision, recall, and F1-score measures to assess how well the trained model performed on the trial data. This aids in identifying the formula that predicts diabetes the best.

Model Deployment: Using fresh data to forecast diabetes using the top-performing machine learning algorithm.

Model Interpretation: Interpreting the results of the model to learn more about the variables that influence diabetes prediction. This can assist healthcare workers in making wise choices about the management and care of patients.

Documentation: For future reference and consistency, it is important to record every step of the process, including data gathering, preprocessing, feature selection, model training, and assessment.

IV. CONCLUSION

Techniques based on machine learning have shown tremendous potential in terms of early diabetes detection and diabetes risk prediction. Numerous machine learning techniques can be used to forecast diabetes in Python.

Starting with a collection of relevant characteristics, such as age, BMI, glucose level, blood pressure, and family history, one can create a diabetes forecast model in Python. After managing missing values, scaling, and normalizing the data, the collection can then undergo preprocessing.

In overall, creating a diabetes prediction model in Python using machine learning can assist in early detection of people at risk of getting diabetes, enabling prompt assistance and treatment.

V. REFERENCES

[1] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance's for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[2] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.

[3] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

[4] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13

[5] Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018