An International Conference-Innovation 2023
in association with International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET

Decoding Heart Health using Machine Learning

Ashlesha Katore, Mansi Kotkar, Kaustubh Jha, Prof. Yashanjali Sisodia, Prof. Prajakata Jadhav
Department of Computer Engineering, Savitribai Phule Pune University, Pune, Maharashtra, India

ABSTRACT

In this study, we are working on making a good model to predict diabetes early on. The goal is to stop the disease from getting worse and causing problems. We are using information from different datasets. Our main tool for this is something called logistic regression. We are trying two ways to pick the most important information from the data to make our model better. We are also using a few tricks to combine different predictions and make our guesses more accurate. We are doing all this using a programming tool called Python. Our findings show that logistic regression is pretty good at this job. The best accuracy we got was 78% for one dataset and 93% for the other after using our tricks to combine predictions. We also talk about how diabetes is a big problem worldwide and how important it is to find it early. Our hope is that our study helps make better tools for predicting diabetes early. This could mean doctors can help people sooner, and that is important for keeping everyone healthier.

Keywords: Python Programming, Machine Learning Algorithms, Classification Techniques

I. INTRODUCTION

Diabetes is a widespread health issue affecting millions globally. In 2019, 463 million adults had diabetes, and it is expected to reach 700 million by 2045. Diabetes leads to serious problems like blindness, kidney failure, heart attacks, strokes, and amputations. Around 84.1 million Americans have prediabetes, emphasizing the need for preventive measures. There are three main types of diabetes: Type 1, where the body cannot produce enough insulin; Type 2, where cells struggle to use insulin effectively; and gestational diabetes during pregnancy, often linked to undetected diabetes.

Although diabetes is not curable, it can be managed with proper treatment. Modern healthcare uses machine learning, like predictive modelling, to improve diagnosis and treatment. These advanced methods, using complex algorithms to identify subtle patterns, help in drug discovery and treatment planning. This focuses on creating a predictive model for diabetes to identify those at risk. Understanding factors like family history, age, diet, and high blood pressure is crucial for targeted intervention.

Our model uses machine learning algorithms like Random Forest, Decision Trees, K-Nearest Neighbours (K-NN) Algorithm, and Naïve Bayes. Random Forest performs exceptionally well in terms of accuracy and efficiency. By using this forward-looking method, we aim to enhance our understanding of diabetes, providing valuable insights for future research and intervention strategies in the ongoing fight against this health issue.

II. RELATED WORK

Mr. Santhana Krishnan J., Geetha,[1] S This study explores the predictive capabilities of two supervised data mining algorithms, namely the Naïve Bayes Classifier and Decision Tree Classification, in assessing the likelihood of heart disease in patients. The dataset is subjected to a comparative analysis of both algorithms to discern their accuracy levels. Notably, the Decision Tree model exhibits superior performance, achieving a commendable 91% accuracy, while the Naïve Bayes Classifier follows closely with an 87% accuracy rate.

P. Rama Krishna, P. Ruchita, Ch. Bharat Teja, M. Manoj Kumar, T V S Lingeswararao, [3] In this study, these algorithms were meticulously trained on a curated dataset, with Random Forest demonstrating remarkable accuracy. Beyond the present findings, this model lays the groundwork for future advancements, envisioning the integration of deep learning techniques to further refine accuracy.

Aishwarya Mujumdara, Dr. Vaidehi Vb, [4] This study explores the effectiveness of various machine learning algorithms in classifying datasets, revealing Logistic Regression as a standout performer with an impressive 96% accuracy. The introduction of a pipeline further enhances predictive capabilities, showcasing the AdaBoost classifier as the best model, achieving a remarkable accuracy of 98.8%.

III. METHODOLOGY

1. Data Collection: Describe the sources and types of datacommonly used in diabetes prediction studies. This may include electronic health records, medical surveys, laboratory measurements, and lifestyle information. The diabetes data set was originated from https://www.kaggle.com/datasets/mathchi/diabetesdata-set ,This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. Diabetes dataset containing 769 cases.

diabetes.csv

```
Pregnancies,Glucose,BloodPressure,SkinThickness,Insulin,BMI,DiabetesPedigreeFunction,Age,Outcome
6,148,72,35,0,33.6,0.627,50,1
1,85,66,29,0,26.6,0.351,31,0
8,183,64,0,0,23.3,0.672,32,1
1,89,66,23,94,28.1,0.167,21,0
0,137,40,35,168,43.1,2.288,33,1
5,116,74,0,0,25.6,0.201,30,0
3,78,50,32,88,31,0.248,26,1
10,115,0,0,0,35.3,0.134,29,0
2,197,70,45,543,30.5,0.158,53,1
8,125,96,0,0,0,0.232,54,1
4,110,92,0,0,37.6,0.191,30,0
10,168,74,0,0,38,0.537,34,1
10,139,80,0,0,27.1,1.441,57,0
1,189,60,23,846,30.1,0.398,59,1
5,166,72,19,175,25.8,0.587,51,1
7,100,0,0,0,30,0.484,32,1
0,118,84,47,230,45.8,0.551,31,1
7,107,74,0,0,29.6,0.254,31,1
1,103,30,38,83,43.3,0.183,33,0
1,115,70,30,96,34.6,0.529,32,1
3,126,88,41,235,39.3,0.704,27,0
8,99,84,0,0,35.4,0.388,50,0
7,196,90,0,0,39.8,0.451,41,1
9,119,80,35,0,29,0.263,29,1
11,143,94,33,146,36.6,0.254,51,1
10,125,70,26,115,31.1,0.205,41,1
7,147,76,0,0,39.4,0.257,43,1
1,97,66,15,140,23.2,0.487,22,0
13,145,82,19,110,22.2,0.245,57,0
```

Fig.1. Dataset

This diabetes dataset consists of 9 attributes with outcome where 0 indicates there are chances of diabetes and 1 indicates there is chances of diabetes.

2. Data Preprocessing: Explain the preprocessing steps, which involve cleaning and organizing the data to make it suitable for analysis. This includes handling missing values, outliers, and normalizing features. Emphasize the role of feature engineering in creating new attributes that could be more informative for diabetes prediction.

3. Feature Selection: Discuss the importance of selecting relevant features or attributes. Feature selection methods, such as correlation analysis or recursive feature elimination, should be introduced. Highlight the need to balance between reducing dimensionality and maintaining predictive accuracy.
4. Model Selection: Present a comprehensive overview of machine learning algorithms suitable for diabetes prediction. This may include:
   Decision Trees
   Random Forest
   Naive Bayes
   K-Nearest Neighbours (KNN)
5. Model Training: Explain how the selected machine learning algorithms are trained on a portion of the dataset. Cross-validation techniques like k-fold cross-validation should be mentioned for hyperparameter tuning and model assessment.
6. System Architecture:

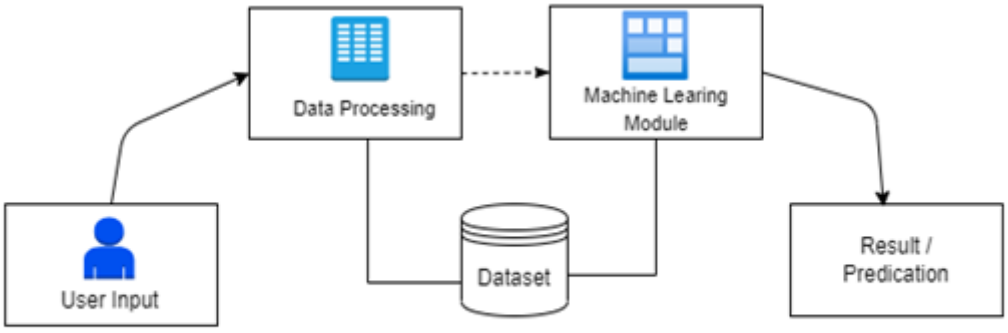

Fig.2. System Architecture

7. Data Flow Diagram:

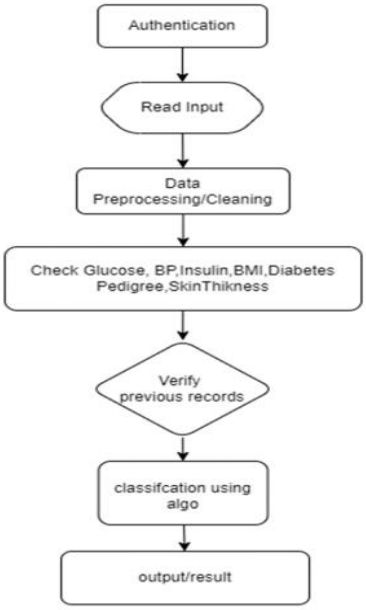

Fig.3. Data flow diagram

IV. ALGORITHMS

Decision Tree Classification Algorithm: The Decision Tree is like a smart tree that can answer questions based on certain conditions. These answers are usually categorical, like "Yes" or "No", "True" or "False" or even "1" or "0." In the context of medical datasets, the Decision Tree is often used to make predictions. The way this tree works is different from other models like K-Nearest Neighbours (K-NN) or Support Vector Machines (SVM). It creates a tree-like structure to analyse data, which is why it's called a Decision Tree. This structure consists of horizontal and vertical lines that split the data based on certain conditions related to the variables we are looking at.

The unique thing about the Decision Tree is that it considers all the attributes in the dataset. It analyses the data in a way that looks like a tree, with three important parts:

Root Node: Think of this as the main decisionmaker. Everything starts from here.

Interior Node: This node handles the conditions related to the variables we are looking at.

Leaf Node: The result, whether it is a "Yes" or "No" for our prediction, is found at a leaf node.

K-Nearest Neighbours (K-NN) Algorithm: K-NN is an intriguing machine learning algorithm that belongs to the supervised learning category. Its distinctive characteristic lies in its neighbour-based approach, making it a versatile tool for making predictions.

Neighbour-Based Predictions: At its core, K-NN aims to find a set number of training samples that are closest to a new, unknown data point in terms of distance. These closest neighbours serve as valuable references to predict the label or value of the new point.

Classification Focus: K-NN often shines in classification tasks. This means it is particularly useful when you want to categorize data into different groups. What is exciting is that it does not require a deep understanding of how the data is spread out; it simply looks at the closest neighbours to make decisions.

V. CONCLUSION

This research aimed to create a computer program using machine learning to help find heart diseases early. They used three different methods and checked how well they worked using measures like accuracy, precision, recall, and F-measure.

The Random Forest method was the best, getting a perfect 100%accuracy in predicting heart disease. This is crucial because heart issues can be very serious, and a wrong or late diagnosis can lead to dangerous outcomes, even death. The study shows that using computer programs like this can be super helpful for heart doctors to make more reliable and faster diagnoses, ultimately helping patients.

In summary, this study successfully made computer programs to predict heart diseases using fancy math. These findings can be a big deal for heart doctors. Future studies should check more things, try different methods, and make sure the data is super good to improve these heart disease prediction programs even more.

VI. REFERENCES

[1]. Mr. Santhana Krishnan J., Geetha S [2019], “Prediction of Heart Disease Using Machine Learning Algorithms”.

[2]. Shriniket Dixit, Pilla Vaishno Mohan, Shrishail Ravi Tern[2022], “Prediction of Heart Disease Using ML algorithms”.

[3]. P. Rama Krishna, P. Ruchita, Ch. Bharat Teja, M. Manoj Kumar, T V S Lingeswararao [2022]. “DIABETES PREDICTION USING MACHINE LEARNING ALGORITHMS”.

[4]. Aishwarya Mujumdara, Dr. Vaidehi Vb [2019]. “Diabetes Prediction using Machine Learning Algorithms”.