International Journal of Scientific Research and Engineering Development-– Volume 5 Issue 5, Sep- Oct 2022
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Diabetes Data Analysis and Machine Learning Based Prediction Model on Streamlit Web App

Nalla Adithya Reddy
Electronics and Communication Engineering, KL University
Email: adithyareddy639@gmail.com

----------**********----------

Abstract:

The condition known as diabetes is brought on when the immune system, which protects us from infection, targets and kills the pancreatic cells that produce insulin. This is the pathogenesis of type 1 diabetes. When cells in the muscles and liver grow resistant to insulin, type 2 diabetes develops. Undiagnosed diabetes can cause dangerous illnesses including renal damage and heart attacks. In order to discover and analyse gestational diabetes, there is a need for sound research and the improvement of learning models. This significant flaw is been resolved by advances in machine learning techniques. The goal of this study is to create a system that, using machine learning algorithms, can more accurately forecast a patient's risk of developing diabetes at an early stage. The suggested model yields the most accurate diabetic prediction results, and the findings shown that the prediction system is capable of accurately, quickly, and most significantly, instantaneously predicting the diabetes illness. Sklearn has been utilized by us to train, test, and determine the model's correctness,

Keywords **— Machine learning algorithms, diabetes detection, prediction models, Sklearn, data cleaning, type-1 diabetes, type-2 diabetes.**

----------**********----------

I. INTRODUCTION

One of the most significant medical conditions and one of the main killers in several nations across the world is diabetes mellitus (DM). Diabetes mellitus (DM) is a long-term condition brought on by high blood sugar levels. 1.6 million individuals died from DM in 2016, and 422 million people had the disease, according to study. Diabetes mellitus (DM) is mainly divided into two types: type 1 and type 2. A severe lack of the hormone insulin, which the pancreas produces, causes hyperglycaemia in type 1 diabetes. Usually, this kind affects young children. Type 2 diabetes, on the other hand, affects both adults and children often and is defined by hyperglycaemia brought on by an abnormality in insulin production, typically with a contribution from insulin resistance.

The most common hardware method for diagnosing diabetes was drawing a little sample of blood from the patient using a strip or machine and then estimating the chance that the patient had diabetes using that information. People in rural areas cannot afford to buy these devices because of their high cost. We can now roughly categorize if a person has diabetes or not using machine learning based on different input parameters including their blood pressure, BMI, and glucose level. We can make use of reliable machine learning models for class prediction.

By using machine learning, we can turn the model into a web application and host it online for free usage by anybody in the globe. A web app based on machine learning that predicts diabetes will be created using the Kaggle Diabetes Dataset.

II. METHODOLOGY

The goal of the study is to evaluate the effectiveness of two machine learning algorithms for diabetes diagnosis and prediction. Manual checks were made for any missing or empty values in the data. The data was determined to be clear enough to move forward. The Receiver Operating Characteristics Curve was created and analysed in order to diagnose the algorithms' performance (The ROC curve).

A. *Data Collection*

In this paper we have considered dataset from PIMA Indians Diabetic disease from Kaggle which has a 821 diabetic samples and is been analysed using weka tool. Weka was determined to be effective and simple to use for research and analytical purposes among the many data mining tools that were accessible. Data preparation, including data cleansing, discretization, and data transformation, is required before employing such data sets in the data mining process. The properties of the PIMA dataset are shown in the table below (Table 1). Our dataset does not take into account missing or noisy characteristics.

TABLE 1:
DATASET CHARACTERISTICS

Dataset	PIMA
Number of samples	821
Feature Attributes	8
Output Classes	2
Total Feature Attributes	9
Missing Attributes Status	None
Noisy Attributes Status	None

B. *Data Pre-processing*

a. *Feature Identification and categorization*

Both discrete and continuous categorical values may be used to define features, which might be either continuous or discrete in nature. Contrary to continuous features, which have an exceedingly high total number of values (infinite) and span a certain range, distinctive features are concerned with situations where the total number of values is relatively limited (finite) (range).

From the data collection, the following characteristics can be extracted:

i. Pregnant: Seen in Women how many times that she went under pregnancy
ii. Plasma-Glucose
iii. DiastolicBP: measured in mmHg
iv. TricepsSFT: measured in mm
v. Serum-Insulin: measured in mm U/mt in a two-hour serum insulin
vi. BMI is generally abbreviated as Body Mass Index measured in kg for w and m for height
vii. DPF
viii. Age: measured in years.
ix. Class: either in 0 or 1

When cleaning the data set, it is important to keep these qualities in mind. RapidMiner alone can handle the majority of this work. You will receive a correctly categorized characteristic table after uploading samples from the Pima Indian Data Set, changing the default characteristic names, and renaming the values of the characteristic Class from (0, 1) to (No, Yes).

b. *Data Integration and Reduction*

A technique for retrieving and combining disparate knowledge into a single integrated type and structure is known as data integration. The ability to fully integrate various information types (such as information sets, documents, and tables) by people, organizations, and applications for usage in either personal or professional activities and/or functions

Data reduction is the process of converting numerical or alphabetical digital data obtained via empirical observation or by experimentation into a rectified, organized, and simpler kind. The primary idea is to reduce indeterminate volumes of data all the way down to their useful components.

c. Data Normalization and Standardization

Using the usual scaler approach, we normalize the data in this phase. Since many variables have a tendency to differ in units, for example, age and glucose in our example dataset, we must apply the standardization approach to convert them to a common scale by dividing the dataset by the standard deviation. Although the normalizing procedure does not eliminate variation.

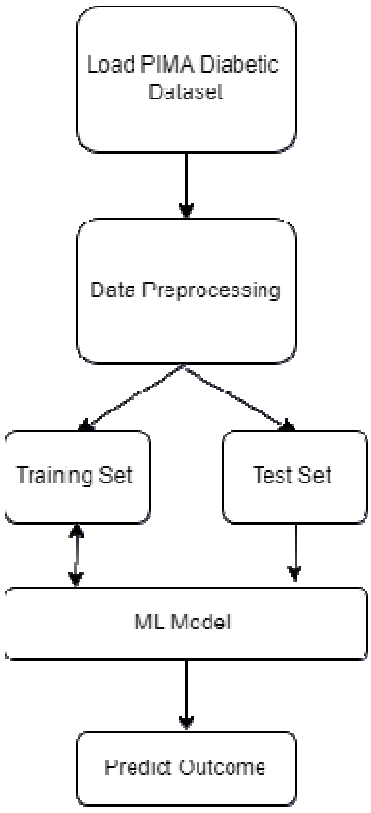


Fig. 1 Basic Architecture of Diabetic Detection

III. ABOUT STREAMLIT

A Python app framework called Streamlit is available for free. It allows us to create online apps for data science and machine learning quickly. It works with a number of the most well-liked Python libraries, including scikit-learn, Keras, PyTorch, SymPy (latex), NumPy, pandas, and Matplotlib. Since widgets are considered as variables in Streamlit, call backs are not required. Computation pipelines become more straightforward and swifter with data caching. A shared link will automatically be updated with the application when Streamlit detects changes in the associated Git repository's updates. Abbreviations and Acronyms.

When an acronym or abbreviation is used for the first time in a document, it should be defined, even if it has already been done in the abstract. It's not necessary to define acronyms like IEEE, SI, MKS, CGS, SC, dc, and rms. If at all possible, avoid using acronyms in the title or headings.

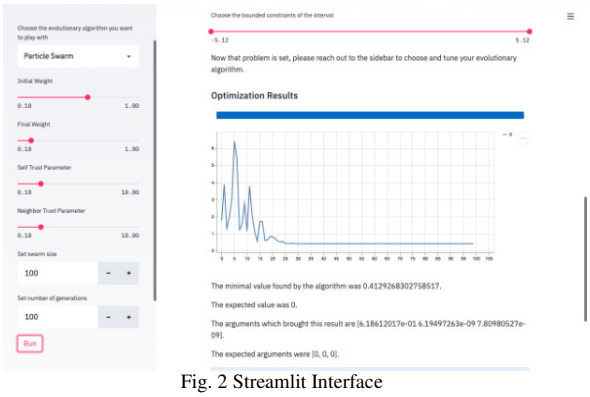


Fig. 2 Streamlit Interface

IV. RESULT AND ANALYSIS

A. Diabetes prediction with SVM and RF

The algorithm for detecting and forecasting diabetes in patients was created using the Python programming language. The subsequent procedures were carried out during the SVM implementation.

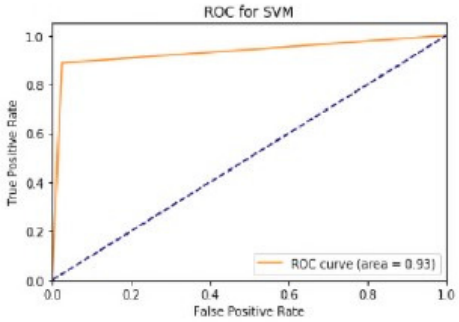
a. Importing data
b. Training and testing the data
c. Implementing SVM model
d. Prediction results

For further investigation, the findings of the SVM classification output were recorded. The following procedures were used to put the RF Classification algorithm to use in diagnosing and predicting diabetes.

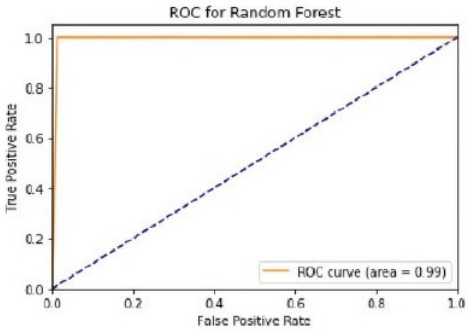
i. Importing data
ii. Training and testing data
iii. Implementing RF Model
iv. Prediction results

B. Area under the ROC curve analysis results

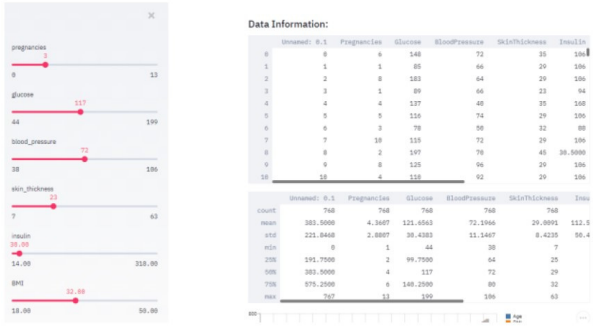
i. ROC vs SVM



ii. ROC vs RF



C. Final Output



V. CONCLUSION

This paper's goal is to develop the prediction model of the effects of diabetes. According to the, there are four issues, diabetes symptoms, such as diabetic kidney disease and diabetic eye disease, Diabetes-Related Heart Disease, and Diabetes-Related Hyperlipidaemia. According to the findings, the model has a 87.4 percent accuracy rate. This study took into account diabetes, since it was shown to be a prevalent and serious illness among Indians. The goal of this research is to create a system that can accurately anticipate diabetes in patients earlier by utilizing machine learning, which offers advance assistance for diabetes prediction accuracy rates.

REFERENCES

[1] http://article.sapub.org/10.5923.j.diabetes.20200902.01.html

[2] https://arxiv.org/abs/2005.08701

[3] https://www.toujeo.com/-/media/EMS/Conditions/Diabetes/Brands/toujeodtcghp/pdf/AboutToujeo_TRANSCRIPT.pdf?la=en

[4] https://www.kaggle.com/uciml/pima-indians-diabetes-database

[5] C. Amuthadevi, R. Umamaheswari, and M. Belinda, "Computational Intelligence Based Medical Diagnosis," J. Comput. Theor. Nanosci., vol. 15, pp. 2369-2372, Jun. 2018, doi: 10.1166/jctn.2018.7471.

[6] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," BMC Med. Res. Methodol., vol. 19, no. 1, p. 64, Mar. 2019, doi: 10.1186/s12874-019-0681-4.

[7] "The Importance of Early Diabetes Detection," ASPE, 23 -Nov-2015. [Online]. Available: https://aspe.hhs.gov/report/diabetes-national- plan-action/importance-early-diabetes-detection. [Accessed: 26-Feb- 2020].

[8] Deepti Sisodiaa, Dilip Singh Sisodiab," Prediction Diabetes and Classification Algorithm" A National Institute of Technology, G.E Road, Raipur and 492001, India, International Conference on Computational Intelligence and Data Science.

[9] A.M.Rajeswari,M.Sumaiya Sidhika,M.Kalaivani C.Deisy,"Prediction of Pre-Diabetes using Fuzzy Logic Baesd Association Classification", Thiagarajar College of Engineering, Madurai, India Proceedings of the (ICICCT 2018), cdcse@tce.edu.

