Diabetic Detection Using Machine Learning

Roosso.P[1], Sudharsan.N[2], DR. D. Rajinigirinath[3]
[1, 2] Dept of Artificial Intelligence and Data Science
[3]Head of Department, Dept of Artificial Intelligence and Data Science
[1, 2, 3] Sri Muthukumaran Institute of Technology

Abstract- *Now a days diabetes has become the major health problem among the people of all ages. The main problem in this type of diseases is its prediction. It is found that if diabetes is detected at early stage, then it can be cured. So early detection of diabetes is important. Data mining now-a-days plays an important role in prediction of diseases in health care industry. Data mining is the process of selecting, exploring, and modeling large amounts of data to discover unknown patterns or relationships useful to the data analyst.Benefit of using these systems is that accuracy of prediction rate is higher as compared to other techniques. Medical data mining has emerged impeccable with potential for exploring hidden patterns from the data sets of medical domains. These patterns can be utilized for fast and better clinical decision making for preventive and suggestive medicine. Here we have used Naive Bayes algorithm which comes under machine learning technique which help us to increase the prediction accuracy.*

Keywords- Detection, Diabetics, Machine Learning, TensorFlow, Naive Bayes algorithm.

I. INTRODUCTION

Classification strategies are broadly used in the medical field for classifying data into different classes according to some constrains comparatively an individual classifier. Diabetes is an illness which affects the ability of the body in producing the hormone insulin, which in turn makes the metabolism of carbohydrate abnormal and raise the levels of glucose in the blood. In Diabetes a person generally suffers from high blood sugar. Intensify thirst, intensify hunger and Frequent urination are some of the symptoms caused due to high blood sugar. Many complications occur if diabetes remains untreated. Some of the severe complications include diabetic ketoacidosis and nonketotic hyperosmolar coma.Diabetes is examined as a vital serious health matter during which the measure of sugar substance cannot be controlled. Diabetes is not only affected by various factors like height, weight, hereditary factor and insulin but the major reason considered is sugar concentration among all factors. The early identification is the only remedy to stay away from the complications. In this work, Naive Bayes and Logistic Regression machine learning classification algorithms are used and evaluated on the dataset to find the prediction of diabetes in a patient. Experimental performance of both algorithms is compared on various measures. The accuracy of blood glucose meter readings is dependent on the test strip material, fabrication process, operating procedures by patients, environmental conditions, and patient medication. The technical accuracy of a glucose meter is determined by comparing the glucose readings from the blood samples analyzed using a glucose meter against the blood plasma samples analyzed by laboratory methods at the same time. It is well established that the finger prick method is a reliable method for accurate glucose measurements. However, consistent penetration of the skin is painful, inconvenient and carries a risk of infection.

II. OBJECTIVE

The main objective of our model is to achieve high accuracy. Classification accuracy can be increase if we use much of the data set for training and few data sets for testing. This survey has analyzed various classification techniques for classification of diabetic and non-diabetic data. Thus, it is observed that techniques like Logistic Regression, Naive Bayes are most suitable for implementing the Diabetes prediction system. Here we are going to compare the performance of both techniques.

III. LITRETURE SURVEY

Prediction models depend on the variables or input attributes that characterize the studied phenomenon. When data collected from databases and the phenomenon to modeling is complex (e.g. patient's health prediction), in many cases there is a need to construct representative variables according to the problem and then modeling and testing. Both processes can be automatized using algorithms to assess their fit with real data. A proposed algorithm wraps this modeling and testing process and looks for the variables constructed from knowledge, helping to fit to real data, and which could be used for any modeling problem that has both quantitative and qualitative mixed information. To test this approach, real data from ten years of diabetic patients' records was used following the American Diabetes Association last recommendations to construct variables that could find high

risk patients and evaluate the variables efficacy. The method could be used to improve data warehouse's framework and for this case, to help care institutions to deploy new health politics to adjust treatments and resource management for diabetic patients.

IV. EXISTING SYSTEM

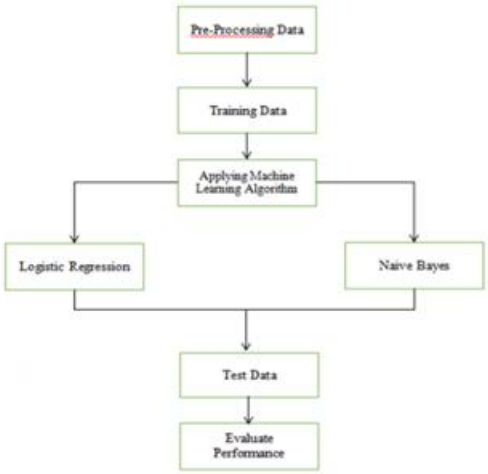
In the existing method is we used SVM Algorithm for calculate that diabetic details. SVM Algorithm is slow process for classify all the given details. In that main disadvantage is time efficiency. Patient's diabetic details start with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seek to explore complex and evolving relationships among data.

V. PROPOSED SYSTEM

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods Bayes theorem provides a way of calculating posterior probability P(c/x) from P(c), P(x) and P(x/c).It is easy and fast to predict class of test data set. It also perform well in multi class prediction. When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data. It perform well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified.

VI. ARCHITECTURE DIAGRAM

This architectural diagram represents the process of the diabetes detection. pre-processing of a data is used to clean null values and outliers in the dataset, training and testing a machine learning algorithm to get optimal solution.



Architecture Diagram

VII. ALGORITHMS USED

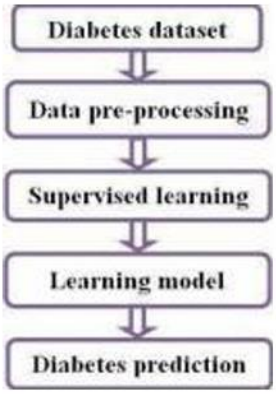
Linear regressionto detect or predict diabetes is an interesting application, although it typically would not be the first choice for a classification problem like diabetes detection. However, linear regression can still be valuable for related tasks, such as predicting quantitative measures that are indicators of diabetes, like glucose levels or insulin resistance scores. In diabetes research and prediction, linear regression might be used to estimate how various factors contribute to blood sugar levels or to explore relationships between different clinical measurements.

Naive Bayes is a powerful, simple, and commonly used classification technique in machine learning that is based on Bayes' Theorem with the assumption of independence among predictors. In the context of diabetes detection, Naive Bayes can be particularly useful due to its effectiveness in handling large datasets with many features, which is often the case in medical data.

VIII. DATA FLOW DIAGRAM

Dataflow diagram shows the flow of the data in the project .

At first we gather the diabetes data for reliable sources and then preprocess the data to give input in supervised learning model to get the output



Data flow diagram

XI. EXPERIMENTAL ANALYSIS

This project can run on commodity hardware. We ran entire projecton an Intel I5 processor with 8 GB Ram, 2 GB Nvidia Graphic Processor, It also has 2 cores which runs at 1.7 GHz, 2.1 GHz respectively. First part of the is training phase which takes 10-15 mins of time and the second part is testing part which only takes few seconds to make predictions and calculate accuracy.Based on a series of preprocessing procedures, the model is comprised of two parts, the improved K-means algorithm and the logistic regression algorithm. The Pima Indians Diabetes Dataset and the Waikato Environment for Knowledge Analysis toolkit were utilized to compare our results with the results from other researchers. The conclusion shows that the model attained a 3.04% higher accuracy of prediction than those of other researchers. Moreover, our model ensures that the dataset quality is sufficient. To further evaluate the performance of our model, we appliedit to two other diabetes datasets. Both experiments' results show goodperformance. As a result, the model is shown to be useful for the realistic health management of diabetes.). A dataset of 442 diabetic patients is used to evaluate the performance of the

X. IMPLEMENTATION

Implementing a diabetes detection model using a Naive Bayes classifier can be a practical approach to categorize patients based on their likelihood of having diabetes, based on various health indicators. Below, I provide a detailed Python implementation using a popular dataset known as the Pima Indians Diabetes Database. This dataset is frequently used for demonstrating machine learning algorithms due to its relevance and availability.

The first step in any machine learning project is to gather the data that will be used to train and test the model. For diabetes detection, the dataset typically includes various medical indicators that are known risk factors for diabetes. These might include glucose concentration, blood pressure, body mass index (BMI), age, insulin levels, skin thickness, and number of pregnancies. A commonly used dataset for such projects is the Pima Indians Diabetes Database, which includes all these features and has been used extensively for machine learning research in diabetes.

Once the data is collected, it needs to be cleaned and prepared for analysis. This involves handling missing values, which could be replaced with the mean or median of the column, or by using a more sophisticated imputation method. It’s also important to remove or correct any outliers that could skew the results. Additionally, since machine learning algorithms require numerical input, any categorical data should be converted into a numerical format through encoding methods. Lastly, features are typically normalized or standardized to ensure that the model isn't biased towards variables with larger scales.

XI. CODING

```
import pandas as pd
data = pd.read_csv('diabetes.csv')
data.head()
data.tail()
data.shape
print("Number of Rows",data.shape[0])
print("Number of Columns",data.shape[1])
data.info()
data.isnull().sum()
data.describe()
import numpy as np
data_copy = data.copy(deep=True)
data.columns
data_copy[['Glucose','BloodPressure','SkinThickness','Insulin','
BMI']] =
data_copy[['Glucose', 'BloodPressure', 'SkinThickness',
'Insulin','BMI']].replace(0,np.nan)
data_copy.isnull().sum()
data['Glucose'] =
data['Glucose'].replace(0,data['Glucose'].mean())
data['BloodPressure'] =
data['BloodPressure'].replace(0,data['BloodPressure'].mean())
data['SkinThickness'] =
data['SkinThickness'].replace(0,data['SkinThickness'].mean())
data['Insulin'] = data['Insulin'].replace(0,data['Insulin'].mean())
data['BMI'] = data['BMI'].replace(0,data['BMI'].mean())
X = data.drop('Outcome',axis=1)
```

y = data['Outcome']

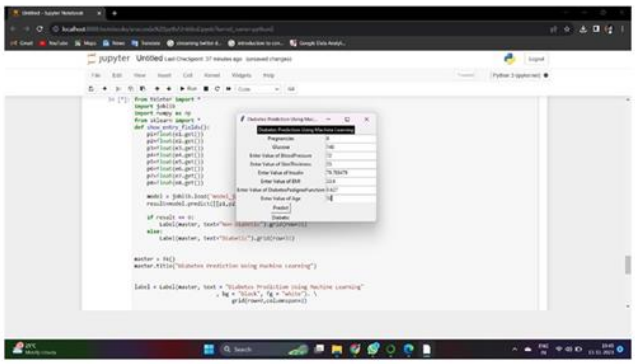
XIII. CONCLUSION

In our work, analysis is made for diabetes prediction to improve the accuracy by using machine learning classification algorithm. We compared the two-prediction model. From, that Naive Bayes classifier achieves higher accuracy than Logistic Regression classifiers. This can be used to select best classifier for predicting diabetes. For future work, it is necessary to bring in hospital's real and latest patients' data for continuous training and optimization and also the quantity of the dataset should be large enough for training and predicting. The field of diabetes prediction is constantly evolving, with new technologies and techniques emerging all the time. Here are some potential areas for future enhancement in diabetes prediction: As more and more health data become available, there are opportunities to incorporate new data sources into diabetes prediction models. For example, wearable devices that track physical activity and sleep patterns could be used to identify individuals at risk of developing diabetes based on their lifestyle habits. The future of diabetes prediction is likely to involve a combination of new data sources, advanced machine learning techniques, and personalized approaches to identifying and managing risk. With ongoing research and development, diabetes prediction models have the potential to revolutionize the way we prevent and manage this common chronic disease.

XII. RESULT

This picture shows the output of the given code above this picture first get the data from the tinker interface

To provide results for diabetes detection, I would need specific information or data, such as medical test results, symptoms described by the patient, or outcomes of diagnostic tools (e.g., glucose tolerance tests, fasting blood sugar levels, HbA1c values, etc.). Diabetes is typically diagnosed based on such medical data, which should be interpreted by healthcare professionals.



Here are some common tests used for diagnosing diabetes:

Fasting Plasma Glucose (FPG) Test: Diabetes is diagnosed if the fasting blood glucose level is 126 mg/dL (7.0 mmol/L) or higher on two separate tests.

Oral Glucose Tolerance Test (OGTT): Diabetes is diagnosed if the blood glucose level is 200 mg/dL (11.1 mmol/L) or higher 2 hours after ingesting a special sugar drink (this test is used more frequently for diagnosing type 2 diabetes and gestational diabetes).

Hemoglobin A1c Test (HbA1c): This test measures the average blood glucose control for the past two to three months. Diabetes is diagnosed if the A1c level is 6.5% or higher.

Random Plasma Glucose Test: Diabetes is diagnosed if blood glucose levels are 200 mg/dL (11.1 mmol/L) or higher and symptoms of diabetes are present.

If you have specific test results or more detailed information that you would like me to analyze, please provide them, and I can help explain what they might indicate. However, for an official diagnosis or medical advice, please consult a healthcare provider.

REFERENCES

[1] An Open Algorithm for Systematic Evaluation of Readmission Predictors on Diabetic Patients from Data Warehouses, Hector Kaschel ; Victor Rocco ; German Reinao, 2018 IEEE International Conference on Automation/XXIII

[2] [Congress of Application of data mining methods in diabetes prediction, MessanAKomi, Jun Li, YongxinZhai, Xianguo Zhang, 2017 2nd International Conference on Image Vision and he Chilean Association of Automatic

[3] Diabetic data analysis in big data with predictive method, S. Thanga Prasad ; S. Sangavi; A. Deepa ; F. Sairabanu ; R. Ragasudha, 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies.

[4] Y. Lu, A. Zhong, Q. Li, and B. Dong, "Beyond finite layer neural net works: Bridging deep architectures and numerical differential equations," in Proc. 35th Int. Conf. Mach. Learn., 2018, pp. 3276–3285.

[5] E. Haber and L. Ruthotto, "Stable architectures for deep neural net works," Inverse Problems, vol. 34, no. 1, Jan. 2018, Art. no. 014004.

[6] R. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, "Neural ordinary differential equations," in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 6571–6583.

[7] Y. Rubanova, R. T. Chen, and D. K. Duvenaud, "Latent ordinary differential equations for irregularly-sampled time series," in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 5320–5330.

[8] M.Gollapalli, A. Alansari, H. Alkhorasani, M. Alsubaii, R. Sakloua, R. Alzahrani, W. Albaker, A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: pre-diabetes, T1DM, and T2DM, Comput. Biol. Med. 147 (2022), 105757.

[9] Prediction of heart and kidney risks in diabetic prone population using fuzzy classification, S. Ananthi ; V. Bhuvaneswari, 2017.

[10] Predicting serious diabetic complications using hidden pattern detection, Saeed Farzi, Sahar Kianian, IlnazRastkhadive, 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), December 2017.