IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 2, February 2021

DOI 10.17148/IJARCCE.2021.10231

Diabetes Prediction System Using Machine Learning Algorithm

R. S. Badodekar[1], Arymann Sharma[2], Ujjwal Thakur[3], Atique Aziz Chaudhary[4]

Student, IT, Sinhgad Institute of Technology, Lonavala, India[1-4]

Abstract: Machine learning is the scientific field dealing with the ways in which machines learn from experience. The purpose of machine learning is the construction of computer systems that can adapt and learn from their experience. With the rise of Machine Learning approaches we have the ability to find a solution to many issues. Furthermore, predicting the disease early leads to treating the patients before it becomes critical. The aim of this research is to develop a system which can predict the diabetic risk level of a patient with a higher accuracy. This research has focused on developing a system based on Logistic Regression. Various machine learning techniques, its application and research papers were studied and reviewed. Logistic Regression was applied in the medical data set and higher accuracy than previous techniques was achieved. Also, Logistic Regression provided more accuracy than numerous other algorithms.

I. INTRODUCTION

Diabetes Mellitus:
Diabetes is one of the deadliest diseases in the world. It is not only a disease but also leads to different kinds of diseases.

Machine Learning:
Machine learning is the scientific field dealing with the ways in which machines learn from experience. The purpose of machine learning is the construction of computer systems that can adapt and learn from their experience. With the rise of Machine Learning approaches we have the ability to find a solution to this issue, we have developed a system using machine learning which has the ability to predict whether the patient has diabetes or not.

Supervised Learning:
In supervised learning, the system must infer inductively from a function called target function. The objective function is used to predict the value of an output variable called dependent variable, from a set of independent variables. The set of possible input values of the function are called instances. Each case is described by a set of characteristics (attributes or features). A cases subset, for which the output variable value is known, is called training data.

Unsupervised Learning:
In unsupervised learning, the system tries to discover the hidden structure of associations between variables in data. In that case, training data consists of instances without any corresponding labels. Clustering is the task of grouping a set of objects in such a way that objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters.

Reinforcement Learning:
Reinforcement Learning is given to a family of techniques, in which the system attempts to learn through direct interaction with the environment so as to maximize cumulative results. It is important to mention that the system has no prior knowledge about the behaviour of the environment and the only way to find out is through trial and error.

II. LITERATURE SURVEY

In order to have a well-versed knowledge about this topic, past research papers have been studied and all the methods have been taken into consideration when creating the proposed system which are discussed here -

1. Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine :- This study involves the implementation of FCM and SVM and testing it on a set of medical data related to diabetes diagnosis problem.
2. A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier :- This study involves data mining techniques using Random Forest Classifier to predict diabetes mellitus.
3. Towards Early Detection of Diabetic Retinopathy Using Extended Fuzzy Logic :- The aim of this research was to prepare and build an automated system for detecting Diabetic Retinopathy.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 2, February 2021

DOI 10.17148/IJARCCE.2021.10231

4. FFBAT-Optimized Rule Based Fuzzy Logic Classifier for Diabetes :- This research has attempted to make diabetes detection system using Firefly-BAT optimised Rule Based Fuzzy Logic.
5. Disease Classification Using Machine Learning Algorithms – A Comparative Study :- This paper includes a comparative study of different machine learning techniques.
6. Fuzzy expert system for diagnosing diabetic neuropathy :- This research was done to create a system for diabetic neuropathy detection.
7. A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes :- This study is mainly focused on comparison of different machine learning and data mining techniques for differentiating diabetes data effectively.
8. Analysis and Prediction of Diabetes Mellitus using Machine Learning Algorithm- :- This research has been done to create a hybrid model using SVM, Naive Bayesian Net, Decision Stump and Proposed Ensemble Method algorithms of machine learning.
9. A model for early prediction of diabetes :- This model was created using ANN, K-Means and Random Forest algorithms for early detection of diabetes.
10. Analysis of diabetes mellitus for early prediction using optimal features selection :- This research designs a predictive model using the most optimal classifier to give the closest possible result.

III. PROPOSED SYSTEM ALGORITHMS AND TECHNIQUES

The methodology consists of 5 different phases as shown in below Figure. Data Extraction, Data Pre-processing, Logistic Regression based processing, Post processing and Analysis of Results.

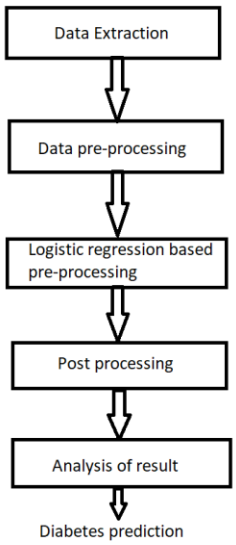


Fig. 1. Proposed System.

In statistics, the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick.This technique is also used in Engineering especially for predicting the probability of failure of a given process, system or product. It is also used in marketing applications. So here the Logistic regression algorithm is used based on the 8 input fields to predict if the person is suffering from diabetes or not.

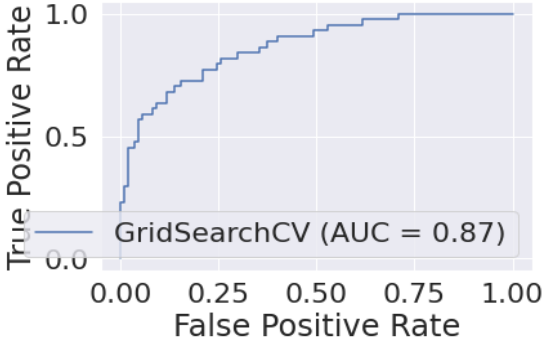


Fig. 2.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 2, February 2021

DOI 10.17148/IJARCCE.2021.10231

Here's a graph we have mention that determines a graph between true positive rate and false positive rate. Where true positive rates are determined on x-axis and false positive rates determined on y-axis

IV. ALGORITHMS AND TECHNIQUES

A. Algorithm:-

Logistic regression is a statistical model that uses a logistic function.In regression analysis logistic regression is estimating the parameters of a logistic model. A binary logistic model has a dependent variable with two possible values, such as pass or fail, where the two values are represented as "0" and "1".

```
[ ]  ## Build an model (Logistic Regression)
     from sklearn.linear_model import LogisticRegression
     log_reg = LogisticRegression(random_state=0)
     log_reg.fit(X_train,y_train);
     ## Evaluating the model
     log_reg = log_reg.score(X_test,y_test)
     ## Build an model (KNN)
     knn = KNeighborsClassifier()
     knn.fit(X_train,y_train);
     ## Evaluating the model
     knn = knn.score(X_test,y_test)
     ## Build an model (Random forest classifier)
     clf= RandomForestClassifier()
     clf.fit(X_train,y_train);
     ## Evaluating the model
     clf = clf.score(X_test,y_test)
```

B. Dataset description:-

This dataset is downloaded from kaggle and is originally from National Institute of Diabetes and Digestive and Kidney Diseases. The dataset consists of 9 columns in which one is outcome that is 1 or 0 and other 8 are the pregnancies value, glucose value,blood pressure and so on.

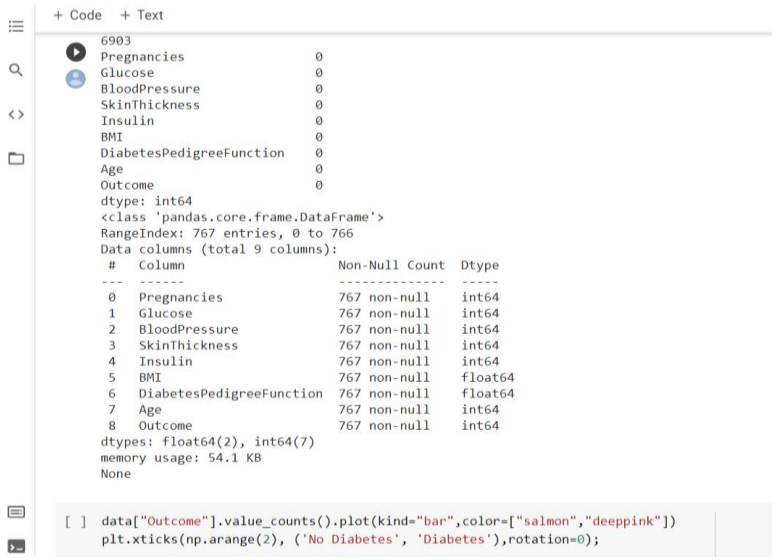


Fig. 3. Data Description.

V. IMPLEMENTATION

The implementation is done using a web application which runs a python script at the backend. The front-end .i.e. the web application is built using HTML and Materialize CSS. Flask API is used to connect the front-end and back-end script. The Machine Learning model was trained using Google Collaboratory notebooks which is an online tool for training ML models. We compared four different algorithms out of which Logistic Regression gave us the highest accuracy after Randomized CV and Grid Search CV was applied.

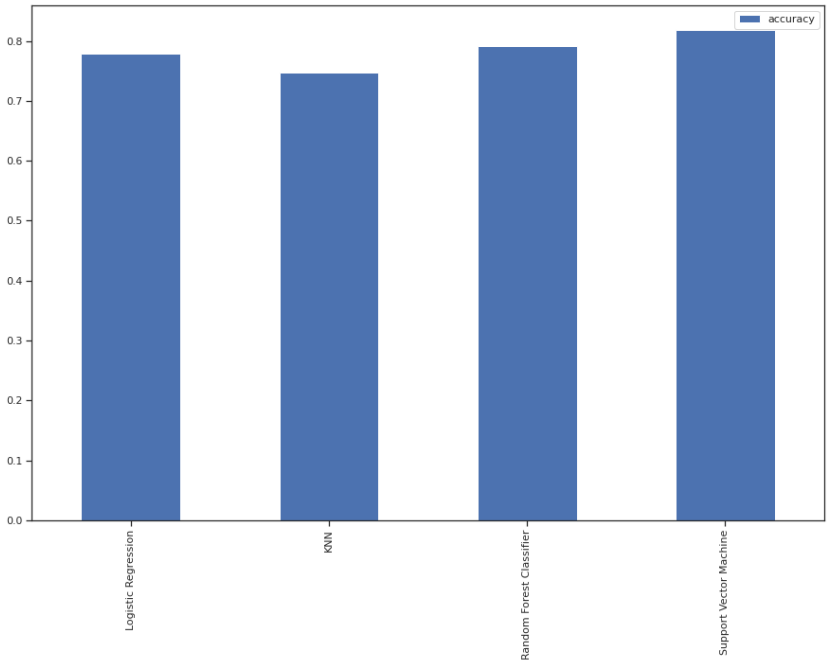
IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

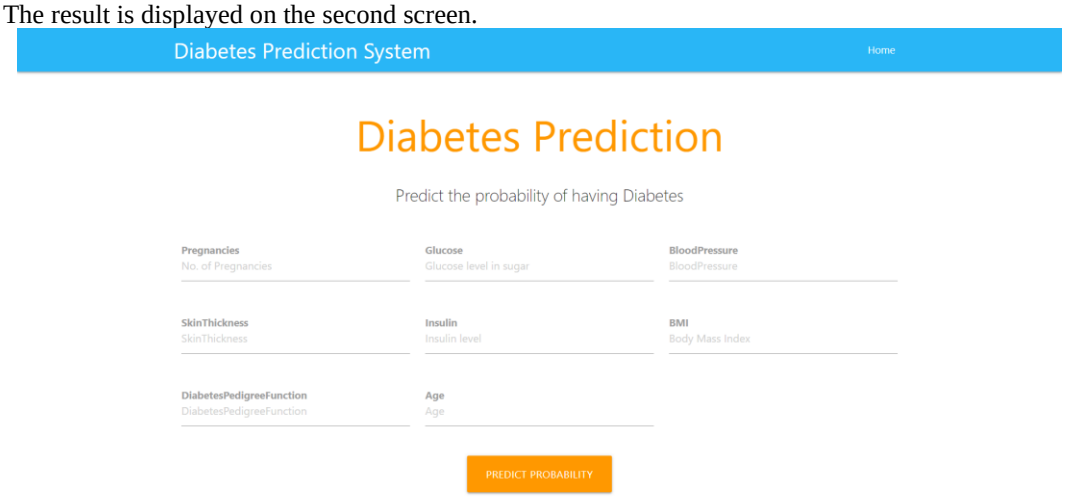
International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 2, February 2021

DOI 10.17148/IJARCCE.2021.10231



The Linear Regression Model is then pickled and used in our web application which in turn helps in prediction of the model. Pickle is the library which is mainly used for serialization and deserialization of objects in python. When the user inputs data into web application input fields the python script is run in the background and the prediction analysis is obtained. The result is displayed on the second screen.



VI. CONCLUSION

Diabetes is a heterogeneous group of diseases. It's characterized by chronic elevation of glucose in the blood. We are trying to detect and prevent the complications of diabetes at the early stage through predictive analysis by improving the classification techniques. Our proposed work also performs the analysis of the features in the dataset and selects the optimal features based on the correlation values. The Logistic Regression gives the highest specificity and the highest accuracy among all algorithms and holds best for the analysis of diabetic data. Support vector machine and NB techniques give the accuracy of 77.73% and 73.48% respectively from the existing method. Hence, it is not as efficient as Logistic Regression. Also, Logistic Regression gives an accuracy of 82.57% when further Randomized Search CV and Grid Search CV is applied.

REFERENCES

[1] Ravi Sanakal, Smt.T. Jayakumari." Prognosis of Diabetes Using Data Mining Approach- Fuzzy C Means Clustering and Support Vector Machine (2014).
[2] Mani Butwall and Shraddha Kumar." A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier." (2015).
[3] Imran, Mohammed, et al. "Towards Early Detection of Diabetic Retinopathy Using Extended Fuzzy Logic." (2016).
[4] G. Thapa Reddy a, NeeluKhare2." FFBAT-Optimized Rule Based Fuzzy Logic Classifier for Diabetes" (2016).

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 2, February 2021

DOI 10.17148/IJARCCE.2021.10231

[5] Leoni Sharmila.S, Dharuman. C, Venkatesan. P, A Novel Neuro- Fuzzy System for classification, Global Journal of Pure and Applied Mathematics, 2017, 12, 267-270. (2017).
[6] Meysam Rahmani Katigari, Haleh Ayatollahi, Mojtaba Malek, Mehran Kamkar Haghighi "Fuzzy expert system for diagnosing diabetic neuropathy " (2017).
[7] Ratana atil, sharavarita mane." A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes " (2018).
[8] Minyechil Alehegn and Rahul Joshi& Dr. Preeti Mulay." Analysis and Prediction of Diabetes Mellitus using Machine Learning Algorithm" (2018).
[9] Talha Mahboob Alama, Muhammad Atif Iqbala, Yasir Alia, Abdul Wahabb, Safdar Ijazb, Talha Imtiaz Baigb, Ayaz Hussainc." A model for early prediction of diabetes" (2019).
[10] N. Sneha and Tarun Gangil." Analysis of diabetes mellitus for early prediction using optimal features selection." (2019).