Diabetic Disease Prediction Using Machine Learning

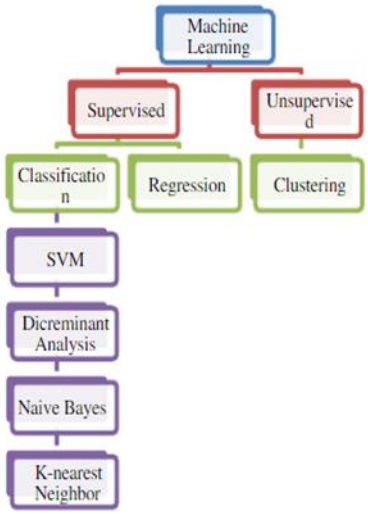
AAKRITI[1], ASHUTOSH MISHRA[2], HARSH VERMA[3], PROF. SHRIPAD DESAI[4]
[1, 2, 3, 4] *Dept. of Electrical Engineering, Bharati Vidyapeeth (Deemed to be University), College of Engineering Pune*

Abstract- ***Many of the interesting and important applications of machine learning are seen in a medical organization. The notion of machine learning has swiftly become very appealing to healthcare industries. The predictions and analysis made by the research community for medical dataset support the people by taking proper care and precautions by preventing diseases. Through a set of medical datasets, different methods are used extensively in developing the decision support systems for disease prediction. We also discuss various applications of machine learning in the field of medicine focusing on the prediction of diabetes through machine learning. Diabetes is one of the most increasing diseases in the world and it requires continuous monitoring. To check this, we explore various machine learning algorithms which will help in early prediction of this disease.***

Indexed Terms- ***Diabetes; health care; Random Forest, Xg Boost; machine learning.***

I. INTRODUCTION

Multiple opportunities for healthcare are created because machine learning models have potential for advanced predictive analytics. There are already existing models in machine learning which can predict the chronic illness like heart disorder, infections, and intestinal diseases.[1] There are also few upcoming models of machine learning to predict non-communicable diseases, which is adding more and more benefit to the field of healthcare. Researchers are working on machine learning models that will offer very early prediction of specific disease in a patient which will produce effective methods for the prevention of the diseases. This will also reduce the hospitalization of patients. This transformation will be very much beneficial to the healthcare organizations. The most explored area is the healthcare system which uses modern computing techniques is in healthcare research.



Flow of machine

Many people in the world are getting affected by diabetes and this number is increasing day by day. This disease can damage many vital organs hence the early detection will help the medical organization in treatment of it.[4] As the Many people in the world are getting affected by diabetes and this number is increasing day by day. This disease can damage many vital organs hence the early detection will help the medical organization in treatment of it.[4] As the number of diabetic patients is more there is an excessive important medical information which must be maintained. With the support of increasing technology, the researchers must build a structure that store, maintain and examine this diabetic information and further see feasible dangers.

II. MOTIVATION

During this ongoing coronavirus pandemic, when we are all bound to live a restricted life under the constant fear of infection risks, it is natural for anyone to develop anxiety. The continuous flow of negative news, the inadequacy of daily resources, everything is adding to this growing anxiety and depression.[9] Due to lockdown multiple hospitals are allocated as covid-

19 healthcare hospital due to it many doctors are busy with treatment of covid positive patients so, an individual who couldn't Consult a doctor for routine check-up cause this model for a rough assumption whether the person is diabetic or not. The common symptoms of diabetes are Urinate (pee) a lot, of tenat night, are very thirsty, lose weight without trying, are very hungry, Feel very tired etc.

According to ongoing problems, we have designed a diabetic predictor model which will help people to predict the early diabetic disease and consult the doctor as soon as possible.

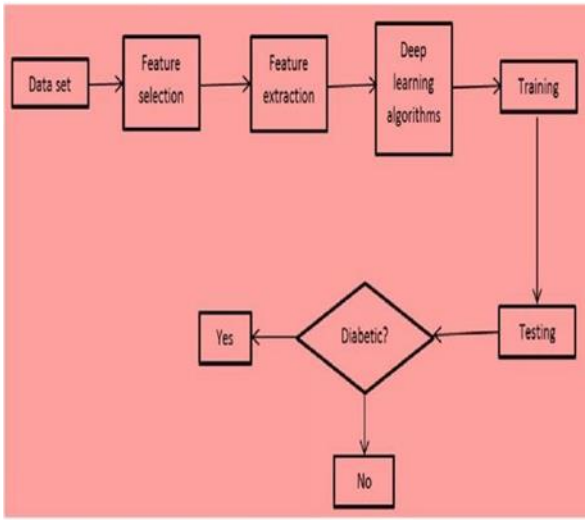
III. REVIEW OF LITERATURE

Before this model was proposed, few other people worked on similar topics to solve the persisting issue.

- One of the titles of the published paper is "An analysis of predicting diabetes using machine learning" by Mr. Ujjwal Anand and Dr. Amit Sehgal. They built two models using Random Forest and Xg boost algorithms. The accuracy by using these models is 88.7% for Random Forest and 88.5% for Xg boost model [2].
- One more paper published is titled as "Diabetes prediction using machine learning techniques" by Mitesh Sony and Dr. Sunita Varma. The proposed system assists the doctor to predict diabetes correctly, and the prediction makes patients and medical insurance providers benefitted. This model consists of various classification algorithms and gives the average accuracy of 77% [3].
- The other papers which is proposed by Aishwarya Mujumdar and Dr. Videha V. titled as "Diabetes prediction using machine learning algorithms" by using random forest and Xg boost algorithms. The accuracy is 94% for random forest and 96% for Xg boost algorithms.[4]
- Another paper which has referred is titled as "Decision tree and random forest based on classification for diabetes prediction" by Anirudh Heber Pana bur and Nitta Meenakshi. This classification model proposed an efficient method for getting the chances of diabetes. It was published in 2019. [5]

IV. DATA DESCRIPTION AND ANALYSIS

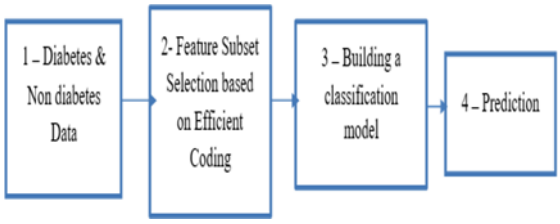
A multivariate dataset obtained from Apollo Hospitals, TN, India-containing about 400 data instances of people aged ranging from 2-90 has been used in this paper. There are about 25 features in total, majority of which are clinical with some also being physiological.



Data Description

V. PROPOSED SYSTEM

Therefore, we use machine learning algorithm to predict/detect whether the person will suffer from diabetes disease or not. A CSV file containing dataset for training and testing is attached. diabetes disease depends upon various factors like, sodium, albumin, sugar, red blood cells, blood urea, bacteria etc.



Firstly, we perform data pre-processing that includes renaming of columns, data cleaning like removing the not-a-number (NAN) values, data visualization to study the skewness, aberration, data-correlation, and distribution. Then, best features election is done using Select Best and chi2 after applying the label encoding for categorical data. Based on the feature scores, we

select the best 8 columns only. Function file is made separately which is called later in main file. Prediction of model is done using the Xg boost algorithm, and to hyper tune the parameters randomized search CV is used.

VI. IMPLEMENTATION

1. Random Forest Algorithm-

Random forest is a Supervised Learning algorithm which usesensemble learning method for classification and regression.

Random forest is a bagging technique and not a boosting technique. The trees in random forests are run in parallel. There is no interaction between these trees while building the trees.

2. Xg boost Algorithm-

Xg boost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. Xg boost is an implementation of gradient boosted decision trees designed for speed and performance.

The Xg boost library implements the gradient boosting decision tree algorithm.

uThis algorithm goes by lots of different names such as gradient boosting, multiple additive regression trees, stochastic gradient boosting or gradient boosting machines.

VII. CODE STRUCTURE

1. Random_forest. ipynb:

Data preparation files Their file contains the python code that consist of all theprocess involved in building our machine learning model.

Steps involved in this file are:

- Load Python packages.
- Pre-Process the data.
- Subset the data.
- Split the data into train and test sets.
- Build ML model.
- Predict on given data.
- Check the Accuracy of the Model.
- Check important Feature Importance.

2. Pickle_classifier.pkl:

Machine learning model which is dumped in pickle format.

These is our trained model which we have dumped in pickle format so that it can be used further and every time we don't need to trainour model.

This process / procedure of saving a ML Model is also known as object serialization - representing an object with stream of bytes, to store it on disk, send it over networks or save to a deserialization while the restoring/reloading of ML model procedure is known as deserialization.

3. app.py:

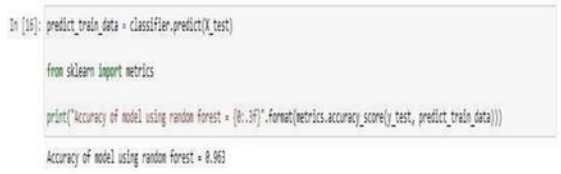
This contains Flask APIs that receives sales details through GUI or API calls, computes the predicted value backdoor model and returns it.

VIII. RESULTS

1. Random Forest Algorithm-

Random forest is a Supervised learning algorithm which uses ensemble learning method for classification and regression.

The accuracy of the diabetes prediction model using Random Forest Algorithm is shown below:



2. Xgboost Algorithm-

Xg boost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data.

The accuracy of the diabetes prediction model using Xg boost Algorithm is shown below:

IX. DATA SET

The Pima Indian Diabetes dataset that is available in the UCI repository is chosen as the sample for the experimental setup. This dataset consists of diabetic and non-diabetic records.[6] It consists of eight attributes and a class attribute. There are 2000 total instances available in the data set.

Attribute _Id	Attribute Name	Attribute Description
A1	Pregnant Times	Number of times pregnant
A2	Plasma Glucose	Plasma glucose concentration 2 hours inan oral glucose tolerance test
A3	Diastolic BP	Diastolic Blood Pressure (mm Hg)
A4	Skin Thickness	Triceps Skin Fold Thickness (mm)
A5	Serum Insulin	2-Hour Serum Insulin (U/ml)
A6	BMI	Body mass index
A7	Pedigree	Diabetes Pedigree Function
A8	Age	Age in years
A9	Class Variable (output feature)	Zero or one

CONCLUSION

The advancements and innovations in the field of science and technology pave the path to explore different possibilities in multidisciplinary domains. While building this system we have experimented with Random Forest and Xg boost algorithms, as the amount of data in the dataset is more hence those models did work well. The results were quite promising; it gave around 95% prediction accuracy. The domain still has numerous possibilities to explore.

The outputs which come after entering the inputs is quite accurate. The model which is build using Random Forest and Xg boost algorithms works well and it shows whether a person is having diabetes or not based on his/her health parameters.

ACKNOWLEDGMENT

We would like to express our special thanks of gratefulness to Professor Shripad G. Desai, Department of electrical engineering for their Guidance and support for completing the research paper. I would like to thank the faculty member of the department of electrical engineering who helped us with extended support.

REFERENCES

[1] DRAP: Decision tree and random forest- based classification model to predict diabetes, conference paper, Jan 2019.

[2] Alumax, A.A., Ahamad, M.G., Siddiqui, M.K., 2019. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University.

[3] Debary Dutta, Debroy Paul, Perthanes Ghosh, "Analyzing Feature Importance's for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[4] Parveen, S., Shahbaz, M., Guaracha, A., Keshav, K., 2016. Performance Analysis of Datamining Classification Techniques to Predict Diabetes. Procedia Computer Science 82, 115–121. Doi: 10.1016/j.procs.2014.04.016.

[5] International Diabetes Federation. Diabetes Atlas. 5th ed. Brussels, Belgium: IDF Publications. (2011) the Global Burden of Diabetes; pp. 7–13. Available from http://www.idf.org/diabetesatlas/news/fifth-edition-release. Accessed 25, May2015.

[6] V. A. Kumari and R. Chitra, "Classification of diabetes disease using support vector machine," International Journal of Engineering Research and Applications, vol. 3, no. 2, pp. 1797– 1801, 2013.

[7] G. Kare Gowda, M. Jayaram, and A. Manjunath, "Cascading k- means clustering and k-nearest neighbor classifier for categorization of diabetic patients," International Journal of Engineering and Advanced Technology, vol. 1, no. 3, pp. 147– 151, 2012.

[8] H. C. Koh and G. Tan, ―Data Mining

Application in Healthcare‖, Journal of Healthcare Information Management, vol. 19, no. June 2005.

[9] L.J. Muhammad, “Predictive Data Mining Models for Novel Coronavirus (COVID-

[10] 19) Infected Patients’ Recovery”, SN Computer Science, Vol. 1, No. 4, pp. 1-7, 2020.

[11] Sa yank Paul, "Fledgling's Guide to Feature Selection in Python", Available at https://www.datacamp.com/local rea/instructional exercises/featureless on-python, Accessed at 2021.