International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 05 | May 2021 www.irjet.net p-ISSN: 2395-0072

ML Based Web App for Diabetics Prediction

Dr.Vrajesh Maheta[1], Aniket Yadav[2], Nasreen Bagde[3], Prasad Solase[4]

[1]Sr. Asst. Professor, Department of Information Technology, Terna Engineering College, Nerul, New Mumbai, India
[2,3,4] Department of Information Technology, Terna Engineering College, Nerul, New Mumbai, India

---

Abstract - *Diabetes is a disease caused when the immune system that helps us to fight infection attacks and destroys the insulin producing cells of the pancreas. This is how type 1 diabetes is caused. Type 2 diabetes is caused when cell in muscles , the liver become resistant to insulin. Because these cells don't take in enough sugar. Nowadays diabetes is becoming the severe threat to life, if not diagnosed early can result fatal outcomes. It is also one of the causes that might sometimes cause heart attack to some people, especially if they have some heart problem. India is one of the largest home for most number of diabetic patients in the world. Its becoming severe year by year, if not taken care properly many lives will be lost. Hence we are trying to use machine learning which can help us predict whether the person is having diabetes or not. With every day advancing field of machine learning, we are trying to build such a model that will help us determine the prediction of the patient. It will be mostly helpful to the people of rural areas where there is not much of testing available and is also costly.*

***Index Terms*— Diabetes, prediction, machine learning, SVC, Random Forest, Data Cleaning, ML.**

1. INTRODUCTION

The most popular diabetes detection gadget made in the form of hardware, where the strip or machine was used to take a pinch of blood of the patient and then based on that it used to show the probability of whether the patient is having diabetes or not. Such gadgets are also costly which cannot be bought by people in rural areas. Whereas with the help of machine learning we can now basically classify whether the person is diabetic or not based on some input features like glucose level, BP, BMI etc. We can use powerful machine learning models like Random Forest, SVM, Logistic Regression etc, to predict the class. And the beauty of machine learning is that we can make a web app of the model and can host it on the web, so that anyone in the world can use it for free. We will be using Pima Indian Diabetes Dataset from kaggle to make a ML based web app for predicting diabetes.

2. LITERATURE SURVEY

2.1 Diabetes Prediction Using Machine Learning

Machine Learning is the emerging field in the world, various research are done every day on how to improve the accuracy of the models. New and more powerful models are developed to tackle the ever-increasing data in the world. With so much data available new models are made to improve the predictions. Usually in the past ML was mostly used in Forecasting, Sales, Finance etc or was mostly a research field. But now health care sector has seen a big jump in use of ML as a early diagnostic tool. With the help of ML and Hospitals data about patient we can create a very good model that can help us predicting the outcome. In our diabetes probelm also it can be used as the early detector tool to diagnose diabetes.

2.2 Diabetes Prediction Using Classifications Algorithms

Aiswarya Iyer (2015) used classification technique to study hidden patterns in diabetes dataset. Naïve Bayes and Decision Trees were used in this model. Comparison was made for performance of both algorithms and effectiveness of both algorithms was shown as a result.[5]. Classification algorithm are used highly as they powerful tools to classify the outcome. Since , in health care sector we usually need to predict whether the person is having certain disease or not, whether he/she is gonna live or not which is a classification problem unlike regression where we have to predict the continuous number. Hence classification are most suitable in health sector.

2.3 Data Cleaning

Data cleaning is the process in which the unwanted data is removed, replaced or imputed to make the data reliable for predicting results with more accuracy. There is a saying in data science that the ML model is as good as the data, "Garbage in Garbage Out". If the data is good the model will perform well on the data and will give results with good accuracy and precision, whereas if the data is bad the model will not fit perfectly to the data and the result will not be reliable. Some common data cleaning techniques include - searching for NULL values, identifying outliers, replacing null values , Deleting NULL values or if the data is too small replacing outlier with some common imputation method like mean imputation, median imputation, mode imputation, we can also use regression algorithm to impute the values, but it can be time consuming. Sometimes its not necessary to delete the outliers or impute it with mean/median/mode because it can sometime help us with identifying the problem why something has caused that. To identify outlier one of the most common technique used is called inter-quartile range, the values outside this range is considered to be outlier. Data cleaning process can vary from data to data and with respect to business/research problem we are trying to solve.

3. PROPOSED SYSTEM

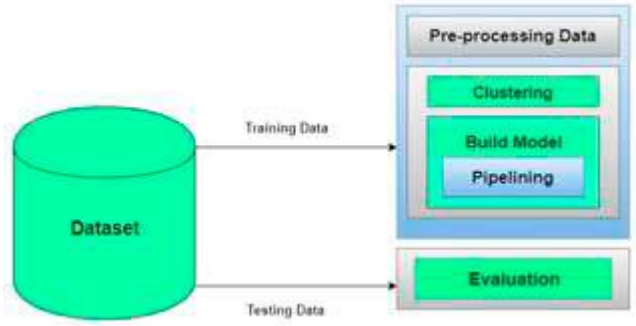
A.The Supervised Learning/Predictive Models

Supervised learning algorithms are used to construct predictive models .A predictive model predicts missing value using other values present in the dataset. Supervised learning is basically that we are given a set of inputs and labels as a output. So based on the given input and labels we train the model and try to predict on the test set. For example in our diabetes data set we have 8 input features and 1 outcome feature that is labeled, since we are predicting based on that label our problem is supervised learning problem.

B. Unsupervised Learning / Descriptive Models Descriptive Unsupervised Learning is basically that the given data set has no labels given to it unlike supervised where the label is given. Based on the data only the clustering has to be formed and classes are named to those clusters. Credit card fraud detection is the best example for it.

C. Semi-supervised Learning

Semi Supervised learning method uses both labeled and unlabeled data on training dataset. Classification, Regression techniques come under Semi Supervised Learning. Logistic Regression, Linear Regression are examples of regression techniques.



Img-1: 1)Dataset Collection, 2)Data-Preprocessing, 3)Build model, 4)evaluation

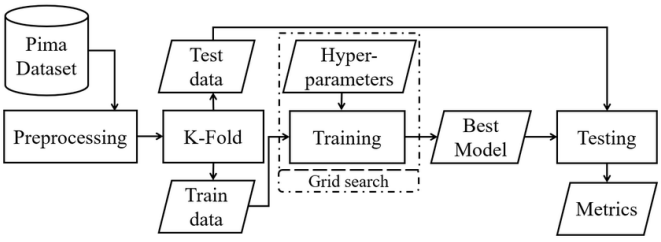


Fig -1: Architecture diagram

3.1 Data Collection:

All the data are collected from one source i.e kaggle, a huge open dataset platform where all data science enthusiast participate in the competions all around the globe. The name of the dataset is "PIMA Indian Diabetes dataset". The dataset was collected by National Institute of Diabetes and Digestive and Kidney Diseases. The data has 768 rows and 8 columns. The column names inlcude: Pregnancies, BP, Glucose, Insulin, Age, DPF, SkinThickness, BMI.

3.2 Data Pre-processing:

This phase of model belongs to data cleaning, imputing methods where the raw data is cleaned and processed to be used for the model. Our diabetes data set contains missing values. Our diabetes data set consists of null values in the form of numerical value zero, and also insulin have high values of outliers. We have used IQR to detect and impute outliers based on their median value.

3.3 Data Normalization/Standardization:

In this phase we standardize the data using standard scaler method. Since many values tend to differ in units like in our example dataset - we have age and glucose which totally differ in units , so to make them on equal scale we use standardization technique which basically divides its dataset by standard deviation to give us common scale. Even though the variance is not lost in standardizing method.

3.4 K-fold CrossValidation and Test-Train Split:

In this phase we cross validate our data into 4 folds. This helps us to ensure that we don't overfit or underfit our data when applying the machine learning model. Then we divide 70 % of our data to be as the training data and 30 % to be as the testing data.

3.5 Model Building:

This is the phase which includes model building for prediction of diabetes. We have implemented various machine learning algorithms for diabetes prediction. These algorithms include Support Vector Classifier, Random Forest Classifier, Decision Tree Classifier, Logistic Regression, K-Nearest Neighbour. While building our model we used Grid Search algorithm for hyper-parameter tuning of the models. Then based on accuracy score of each models we chose the final model for deployment.

3.6 Front-end and deployment:

In this phase we used pythons wonderful streamlit library that is used to make front-end or the dashboards of the machine learning problems. And for deployment we used heroku platform to deploy our main prediction model and EDA dashboard.

4. RESULTS

Fig2: Pregnacy graph

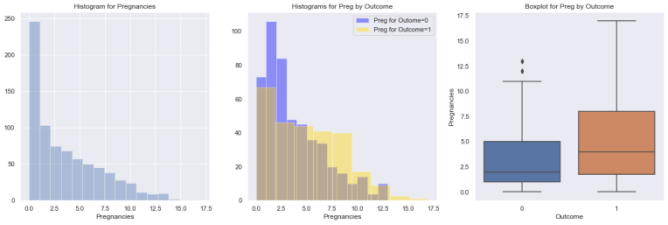


Fig3: Glucose graph

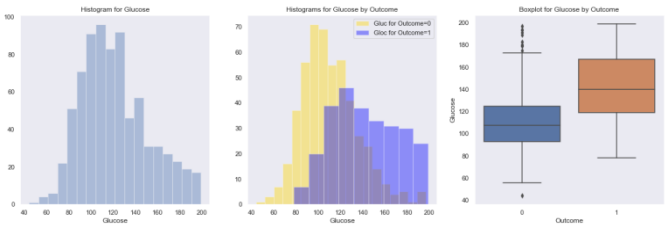


Fig4: Blood Pressure graph

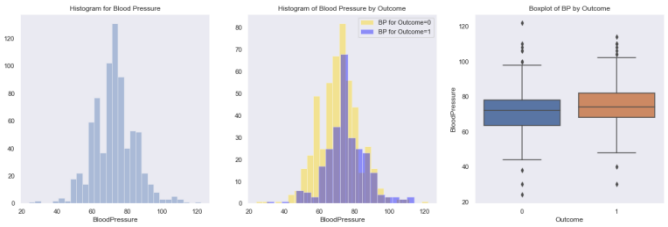


Fig5: Skin-Thickness graph

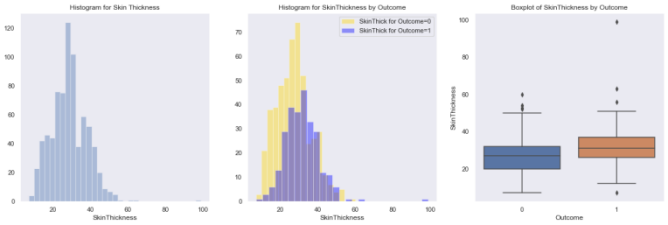


Fig6: Performance Metrics:

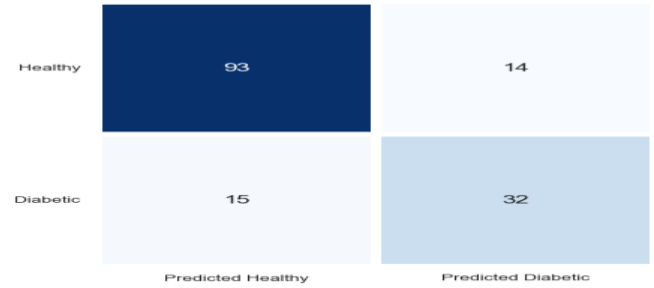
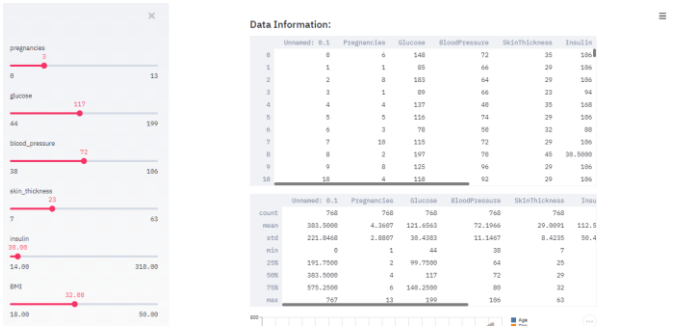


Fig7: Final Output:



5. CONCLUSION

In this project we have performed successful machine learning implementation of Diabetics prediction and have deployed it to heroku platform. Our final ML model used is Random Forest since it had the best accuracy among other ML models.

ACKNOWLEDGEMENT

Our sincere appreciation goes to the Engineering Product Innovation Center (E.P.I.C) and staff members of the Department of Information Technology, terna engineering college, nerul.

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://devinincerti.com/2015/10/13/diabetes-highcost.html

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

.