Prediction of diabetes classification in the population of Texas

Yeamin Faria Chowdhury

12/02/2021

Introduction

- Diabetes is a common chronic disease which is caused by insufficiency in secreting glucose hormone in the body.
- While type 1 diabetes occurs in younger group of people, type 2 diabetes is mostly prevalent in middle-aged or elderly people. The risk factors for type 2 diabetes are obesity, hypertension, physical inactivity and other environmental and lifestyle factors (Zou et al., 2018).
- The prevalence of diabetes in America has been recently increased due to increased incidence of obesity among people (Doughty and Jones, 2011).
- Decreased quality of lifestyle, increased medical bills, increase of other health complications like stroke, kidney disease, nervous system disease and blindness increases among people suffering from diabetes.

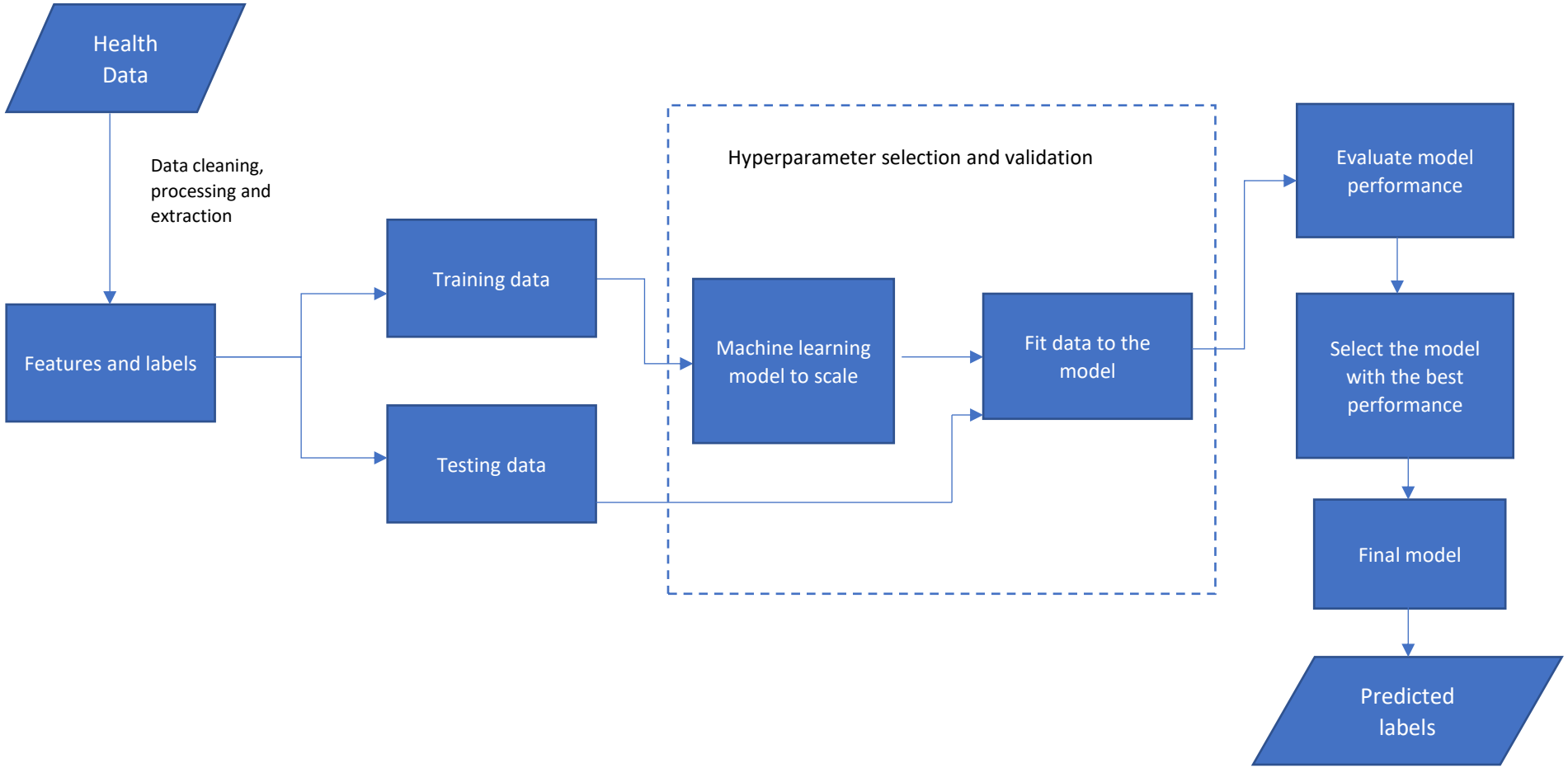
Objectives

- To observe the crucial factor among obesity, physical inactivity and high blood pressure for diabetes in the population of Texas .
- To determine the percent of population in Texas having high diabetes based on the risk factors.
- To determine the optimal machine learning model for predicting diabetes in the population of Texas.

Data Collection

- The dataset used in this study has the measure of the prevalence of diabetes among US adults (18+). It is obtained from the project by Robert Wood Johnson Foundation (RWJF) in conjunction with the Centers for Disease Control and Prevention (CDC) Foundation. It contains estimates of data for 27 health measures from the entire United States at four geographic levels.
- The measure of diabetes is divided into two classes based on the quantile scheme- 0 for values less than 12.5 and 1 for greater values.
- The three common influencing factors for diabetes are observed in this study. These are – High Blood pressure, Obesity and physical inactivity, collected from the same dataset.
- After splitting the samples into 75% training and 25 % testing, I got 2999 for training and 750 for testings.

The workflow



Logistic regression

- The logistic regression uses the sigmoid function, an S-shaped curve which to classify the y variable. The result obtained from logistic regression splits into binary values, 0 and 1. The formula for this curve is:

$$y = \frac{1}{1 + e^{-x}}$$

- A threshold value is used, above which all the x variables will be 1, and the values below the threshold will be classified as 0. The comparison of the predicted and observed value provides the loss function, which has a concave shape.

Support vector machine

- The support vector machine will try to find a hyperplane in N-dimensional space based on the training data to separate the variables into two distinct classes. The ideal hyperplane is the one that has the maximum distance from the two classes.

Random forest

- The random forest is an ensemble method that takes into account many decision trees to get the prediction. The training data is split into the decision trees and each tree returns an output based on some attributes. To determine a better grouping and to control over-fitting, the best mean value is used.

Neural network

- Neural networks are made of neurons or nodes and have three layers – the input layer, hidden layer, and the output layer. The hidden layer performs most of the computation. The nodes are connected to the next layer as channels with weights assigned to them. We can decide the number of hidden layers and their nodes based on cross-validation. A non-linear transformation is applied to the nodes when they are passed through the hidden layers.
- A 5-fold validation divides the samples into 5 parts, 4 parts for training and 1 part for validation. The best hyperparameter for the Neural network for which the mean test score is the highest are:

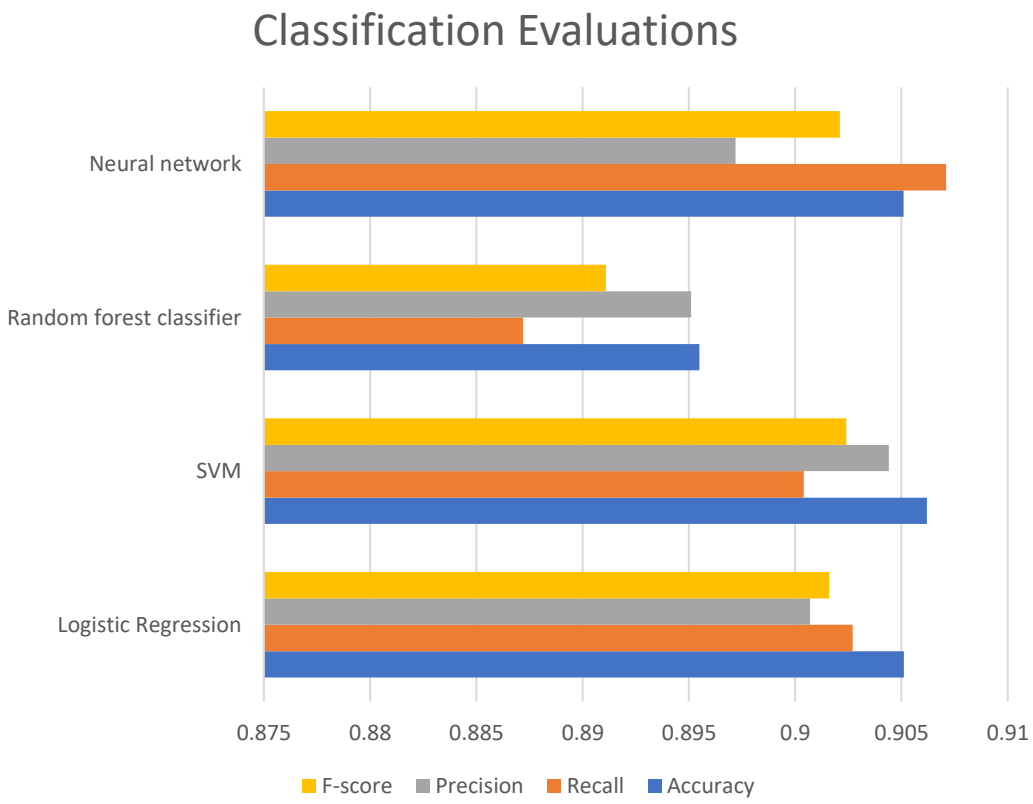
C = 1, solver = ‘sag’ (Stochastic Average Gradient descent)

Results and Discussion

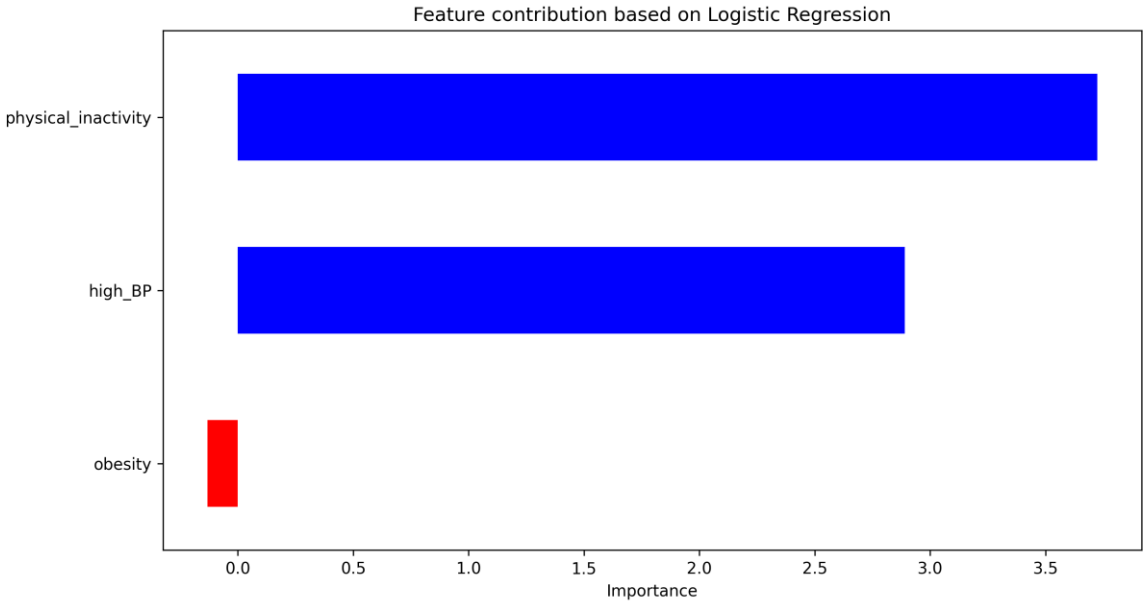
- The classification results in four possible prediction outcomes. These are:
- **True positives (TP):** True positives are the number of samples in class 1 are predicted as class 1.
- **True negative (TN):** The number of samples in class 0 are predicted as 0.
- **False negative (FN):** The number of samples of class 1 predicted as class 0.
- **False positive (FP):** The number of samples of class 0 that are predicted as class 1.
- The four metrics are used for classification evaluation. These are:
- **Accuracy** = (TP + TN)/(TP + TN + FP + FN)
- **Recall** = TP/(TP + FN)
- **Precision** = TP/(TP + FP)
- **F-score** = Average of precision and recall

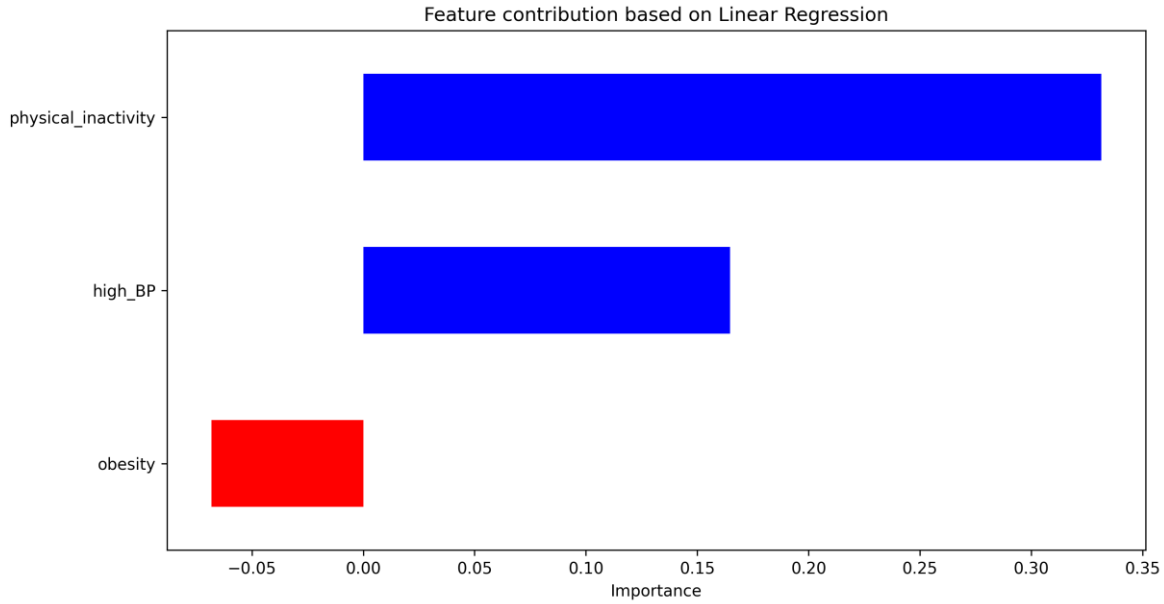
Results and Discussion

	Logistic Regression	SVM	Random forest classifier	Neural network
Accuracy	0.90512	0.9062	0.8955	0.9051
Recall	0.9027	0.9004	0.8872	0.9071
Precision	0.9007	0.9044	0.8951	0.8972
F-score	0.9016	0.9024	0.8911	0.9021



Results and Discussion





Physical inactivity is the most important factor for predicting diabetes in the population of Texas based on Logistic and Linear regression. Obesity goes in the opposite direction.

Conclusion

- In conclusion, it can be said that all the models performs well for this dataset, prediction from support vector machine has the highest F-score, 0.9024**.** The Logistic regression predicts that 49% of the population of Texas should have high Diabetes and physical inactivity is the most contributing risk.

Reference

- Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y. and Tang, H., 2018. Predicting diabetes mellitus with machine learning techniques. *Frontiers in genetics*, *9*, p.515.