|  |  |  |
| --- | --- | --- |
| ***GRADED ASSIGNMENT*** | | |
| *Course* | *:* | *TIPP – AAI* |
| *Module Name* | *:* | *Machine Learning Fundamentals* |
| *Submitted By:* | *:* | *Koay Seng Tian (*[*sengtian@yahoo.com*](mailto:sengtian@yahoo.com)*)*  *19 February 2020* |
|  |  |  |

## Introduction

## The Pima are a group of Native Americans living in Arizona. A genetic predisposition allowed this group to survive normally to a diet poor of carbohydrates for years. In recent years, due to a shift of dietary disposition to processed foods, it had made them develop the highest prevalence of type 2 diabetes and for this reason they have been the subject of many studies.

This dataset came originally from the 'National Institute of Diabetes and Digestive and Kidney Diseases'. The objective is to predict, based on diagnostic measurements whether a patient has diabetes. Patients were females at least 21 years old of Pima Indian heritage. This is a classification task for machine learning.

Data Source: <https://www.kaggle.com/kumargh/pimaindiansdiabetescsv>

1. Machine Learning Process

The machine learning process are as shown.

2.1 Load the dataset.

2.2 Check the dataset for abnormality.

2.2.1 Are the values in reasonable range? The data are in reasonable range.

2.2.2 Any NULL values in the dataset? No null value was detected

2.2.3 Any missing values in the dataset? No missing value in the dataset.

2.2.4 Are the zero values in the dataset reasonable?

2.2.4.1 It is reasonable to have zero pregnancy, but it is unreasonable for age to be zero.

2.2.4.2 Zero values in the dataset were replaced by media values.

2.2.5 Check class imbalance.

2.2.5.1 The ratio of No Diabetes vs Diabetes patient outcomes is 65:35.

2.3 Normalize dataset.

2.3.1 Standard scaler (scikit-learn) was used to normalize the dataset.

2.4 Feature selection.

2.4.1 Correlation matrix generated. The first four features were selected to train the model.

Outcome 1.000000

Glucose 0.492782

BMI 0.312038

Age 0.238356

Pregnancies 0.221898

SkinThickness 0.214873

Insulin 0.203790

DiabetesPedigreeFunction 0.173844

BloodPressure 0.165723

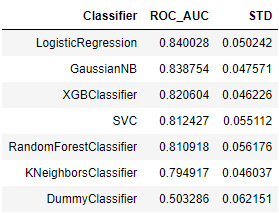
2.5 Select, train and evaluate model.

2.5.1 Dataset is splitted by a ratio of 70:30. 70% is splitted to be the training dataset and the remaining (30%) is allocated for the test dataset.

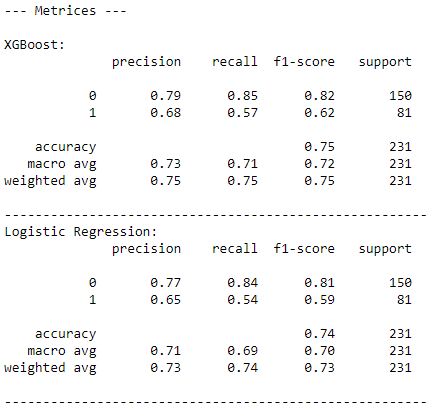
2.5.2 Several models from Scikit-learn were tested to determine the top three models. Also, another model (XGBoost) was installed and preliminary testing was performed. XGBoost was one of the popular algorithms used in Kaggle competitions in 2017.

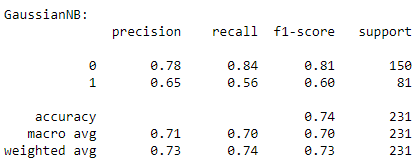
XGBoost: <https://xgboost.readthedocs.io/en/latest/>

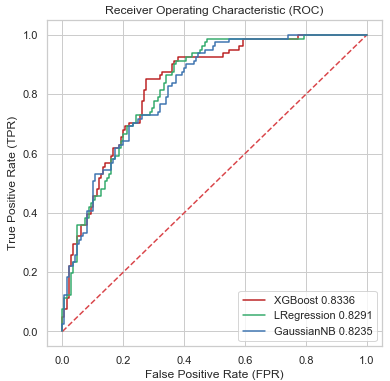
2.5.3 Top three models based on ROC\_AUC metrics were: Logistic Regression, GaussianNB and XGBClassifier tested on training dataset.



2.5.4 The model selected based on test dataset is XGBoost. The ROC\_AUC metric was the highest amongst the three models. The result of the three models are as shown.





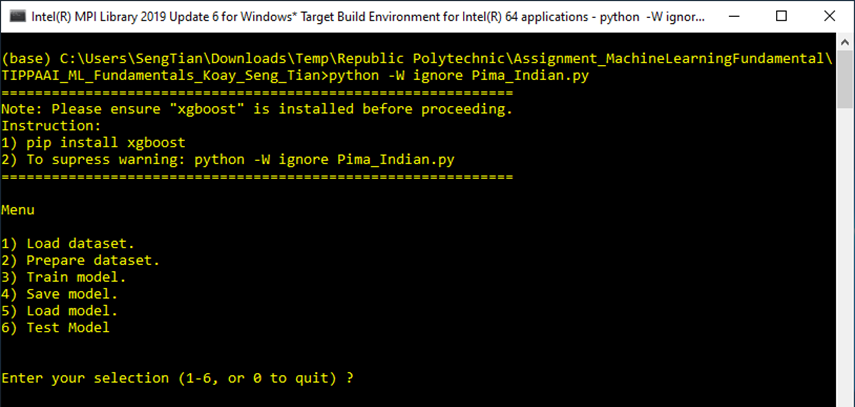


2.6 Save and load the trained XGBoost model.

2.7 Test the model.

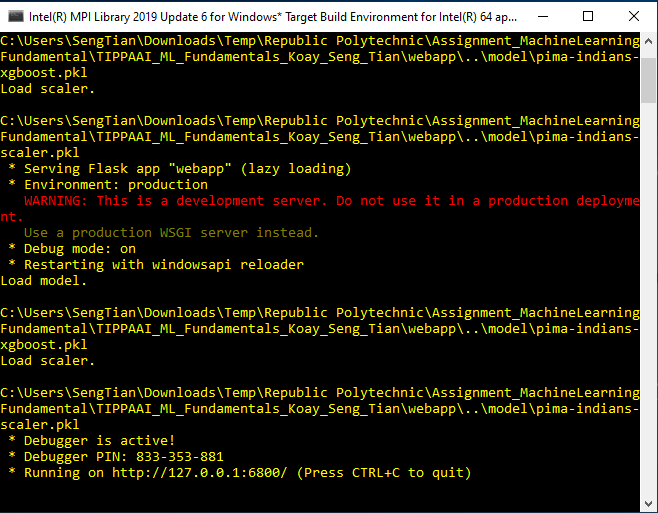
2.7.1 The model can be tested either using the command line app (script) or running it as a flask app.

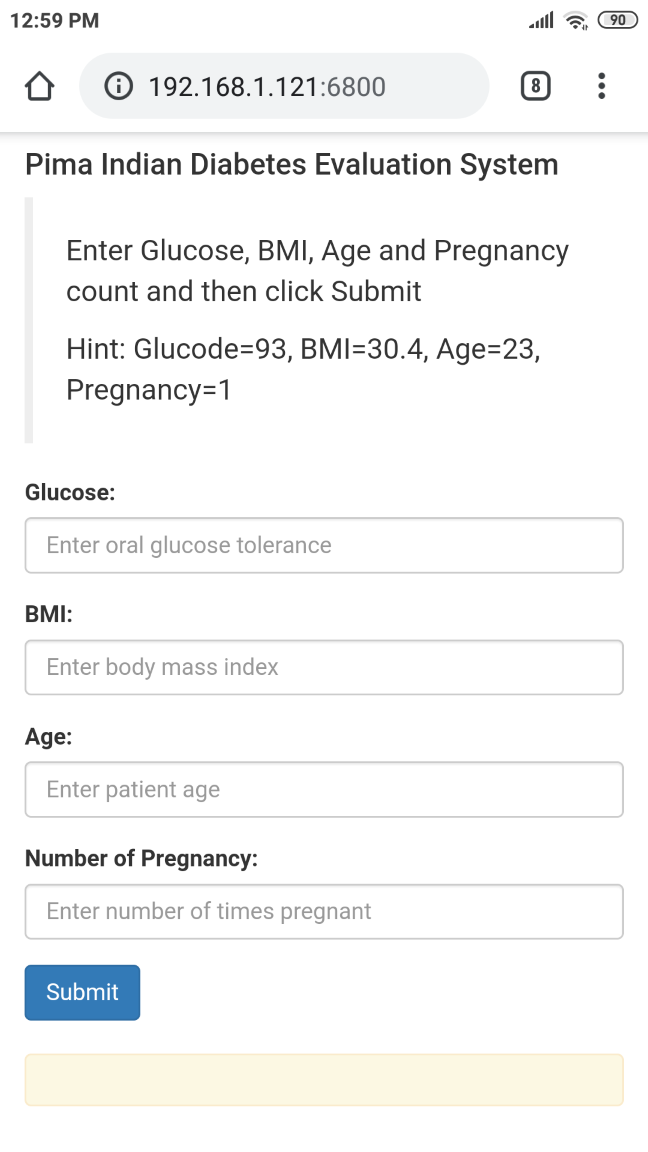
2.7.2 Run the python script with ‘-W ignore’ to suppress the warning messages: **python -W ignore Pima\_Indian.py**.



2.7.3 Run the Flask app via mobile phone or desktop’s web browser. This is a browser-based app. To run the webapp, go to the webapp directory and issue the python command i.e. **python webapp.py**. Open a web browser and enter the URL as [**http://127.0.0.1:6800**](http://127.0.0.1:6800)**.**

A screen captured video is also aviable (Screenrecorder-2020-02-15-11-48-05-218.mp4) and stored in the webapp directory. You can watch the video by drag the file to a chrome browser.





3.0 Summary

XGBoost algorithms was selected based on its higher ROC\_AUC metric. XGBoost is also a fast algorithm. However, the three algorithms were quite close when metric like True Positives (TP) and True Negatives (TN) were compared. Much can still be done, for example, to further fine tune the hyperparameters of each of the top three models. Like any medical related dataset, the dataset was also skewed. As mentioned, the dataset is skewed by a ratio of 65:35 i.e. 65% were diagnosed to have no diabetes. When time permits, I will like to up-sample the minority (have diabetes) or down-sampling the majority (no diabetes) to see whether the prediction outcome can be further improved.