CSCI933: Assignment. #2

Due on Apr. 21, 2020 at 10:00pm

Professor Chao Sun

Group

1. 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 the recent years, because of a sudden shift from traditional agricultural crops to processed foods, together with a decline in physical activity, made them develop the highest prevalence of type 2 diabetes and for this reason they have been subject of many studies.

The type of dataset and problem is a classic supervised binary classification. Given a number of elements all with certain characteristics (features), we want to build a machine learning model to identify people affected by type 2 diabetes.

To solve the problem we will have to analyse the data, do any required transformation and normalisation, apply six machine learning algorithm GaussianNB, RandomForest, K Nearest Neighbour,Decision Tree,Logistic Regression and SVM. Here are some introduction:

2. Data preparation

Through figure 1 simple statistics, it can be found that the ratio between sick and not sick is 67:33.

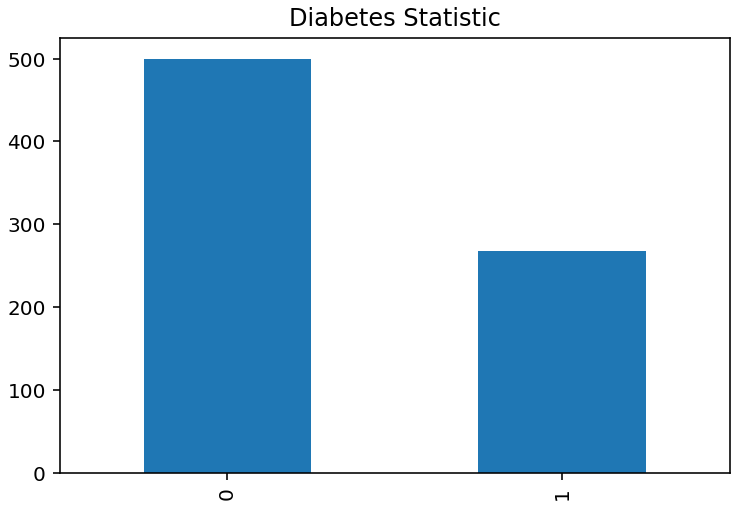

Figure 1: Diabetes Statistic

Then drawing histograms to visualize the data distribution of the eight features.

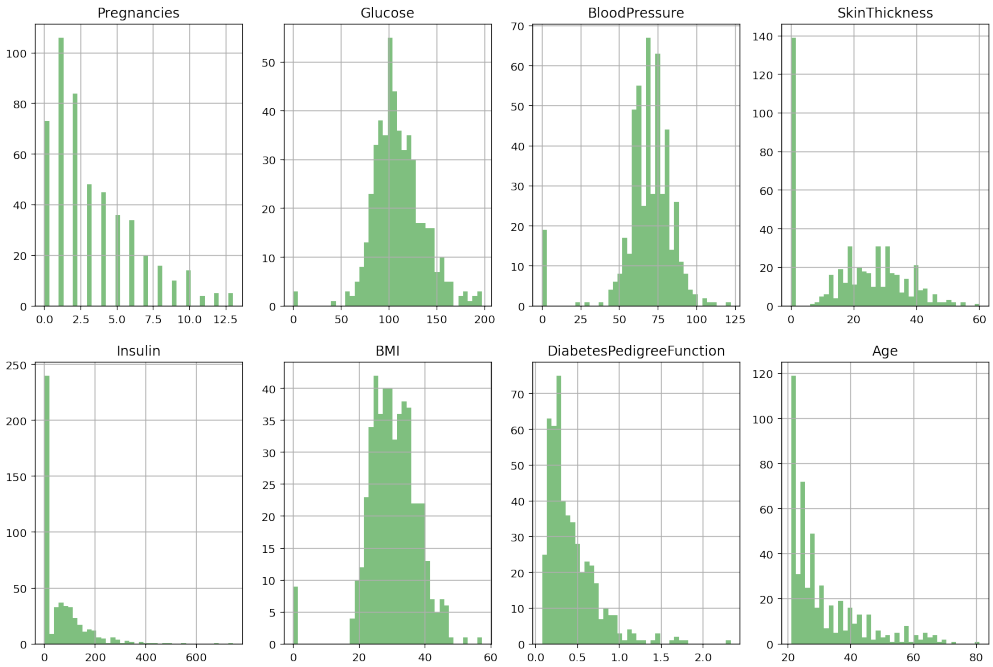

Figure 2: Features Distribution

It can be seen that there are some zero data, which may affect the accuracy of the training and affect the experimental results. Therefore, instead of directly removing the zero data, we chose to use the median to replace them, tere are the processed distribution:

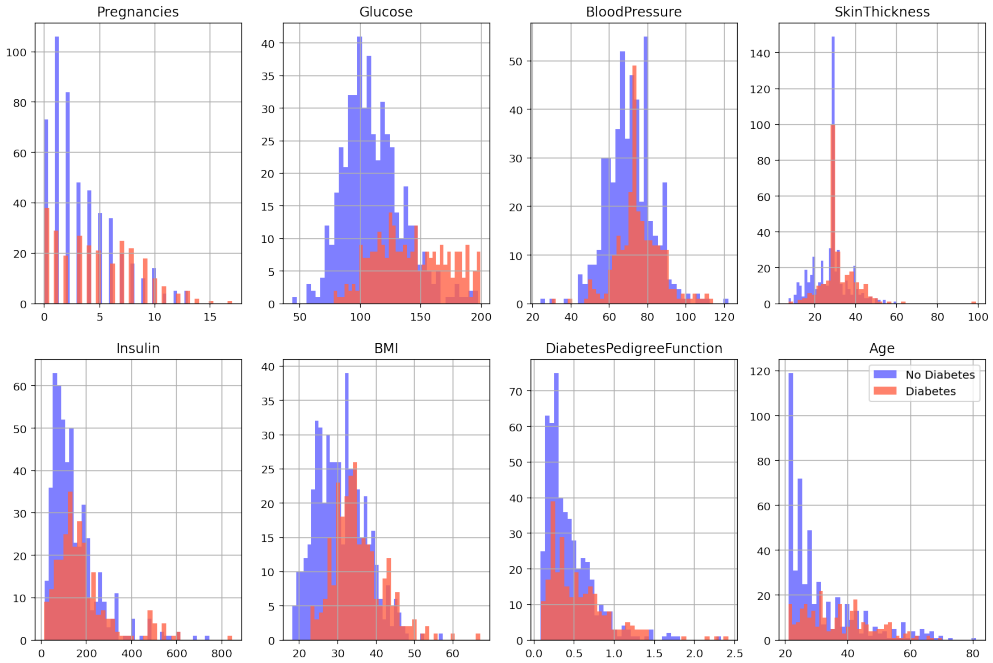

Figure 3: Features Processed Distribution

After that we split data to 80:20 for training and testing. In particular, because this is an imbalance dataset, adding stratify = y to ensure that the output of the split still maintains the same proportion of class 1 and 0. This is the end of Data preparation.

3. Classifier

3.1 Logistic Regression

A logistic regression model will be developed in an attempt to predict whether a patient is likely (represented by 1) or not likely (represented by 0) to develop diabetes in the next 5 years. First, load the file to fetch the data, and split the data to the training set and test set by calling the function train_test_split. Besides, all the 8 characteristics are applied to trained the model. Once we have our training set and test set we can define a LogisticRegression model and fit it to our training data. Once trained, the model can then be used to make predictions against the test set.

3.2 SVM

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems.
Suppose we are given a training dataset of n points of the form,

$$(x_1, y_1)...(x_n, y_n)$$

where y_i represents the classes, it is -1 or 1.
Any hyperplane can be written as the set of points x_i satisfying:

$$\omega * x_i - b = 0$$

By maximizing the distance between the planes, we could get a classifier to separate the data set. For the non-linear data set, it is suggested a way to create classifiers by applying the kernel trick. The idea is to classify non-linear data by cleverly mapping our space to a higher dimension.
Some common kernels include: Polynomial(we used in this project), Gaussian radial basis function and Hyperbolic tangent.
Polynomial kernel function could be described by the equations:

$$k * (x_i, x_j) = (x_i * x_i)^d$$

3.3 KNN

The implementation of KNN is relatively simple: (1 Calculate the distance from the unknown instance to all known instances; (2 Select parameter K; (3 According to the majority-voting rule, classify unknown instances as the largest number of categories in the sample. n general, a larger K value during classification can reduce the impact of noise, but will blur the boundaries between categories. Therefore, the value of K is generally small (K $<$ 20). The measure of distance:

$$d(x, y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2}$$

3.4 GaussianNB

GaussianNB implements the Gaussian Naive Bayes classification algorithm, and this model assumes that the probability density function of all features conforms to a Gaussian distribution, whose the mean and variance of the distribution are estimated using maximum likelihood.

$$P(x_i|j) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2})$$

3.5 Decision Tree

Decision tree learning uses a top-down recursive method, and its basic idea is to construct a tree with the fastest decrease in entropy value using information entropy as a measure.
Feature selection means that one feature is selected from the many features as the criterion of the current node split. There are different quantitative evaluation methods for the selected feature, so as to derive different decision trees. According to the selected feature evaluation criteria, the child nodes are generated recursively from top to bottom, and the decision tree stops growing until the data set is inseparable. Decision trees are easy to overfit and generally require pruning to reduce the size of the tree structure and alleviate overfitting

3.5 RandomForest

Standard Random Forest (RF) is based on decision tree as a basic learner, an extended algorithm based on Bagging, which adds random attribute selection in each round of decision tree training. when training in RF, for each node of the decision tree, the entire attribute set from the current node Randomly select a subset containing k attributes, and then select an optimal partition attribute

from this subset. Usually choosing k as:

$$k = \log_2^d \ or \ k = \log_2^d + 1$$

This is the result based on the above classifier(See the code attachment for the specific implementation process).

```
(machinelearning) guozebiao@guozebiao-UX430UNR:~/PycharmProjects
LR
              precision    recall  f1-score   support

         0.0       0.81      0.86      0.84       125
         1.0       0.71      0.63      0.67        67

    accuracy                           0.78       192
   macro avg       0.76      0.75      0.75       192
weighted avg       0.78      0.78      0.78       192

AC 0.78125
```

Figure 4: Logistic Regression

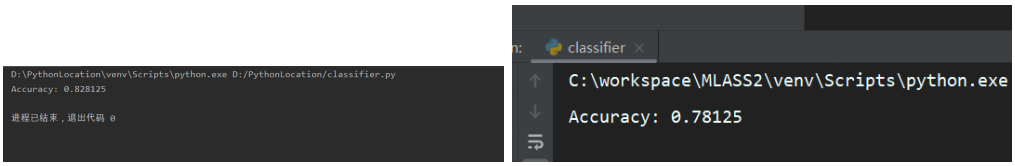

Figure 5: DT vs SVM

	model	acc	recall	f1	logloss
0	GaussianNB	0.754112	0.607641	0.632677	10.092624
1	KNN	0.737758	0.565670	0.600649	10.092603
2	RandomForest	0.757324	0.542193	0.628289	9.195482

Figure 6: GNB vs KNN vs RFC

4.Evaluation

Through the adjustment and optimization of superparameters and thresholds, three classifiers are left behind.

At the same time, from Figure 1 considering that this is a imbalance dataset set, the accuracy rate is not a good evaluation model, so we use the Recall rate to evaluate, while considering loss and time consumed.

	model	acc	recall	f1	logloss
0	GaussianNB	0.754112	0.607641	0.632677	10.092624
1	KNN	0.737758	0.565670	0.600649	10.092603
2	RandomForest	0.757324	0.542193	0.628289	9.195482

Figure 7: Baeline Models

	model	acc	recall	f1	logloss	timetaken
0	GaussianNB	0.754112	0.907407	0.657718	0.759611	0
1	KNN	0.763854	0.907407	0.620253	0.771097	1
2	RandomForest	0.771998	0.888889	0.680851	0.497248	39

Figure 8: Optimal Models

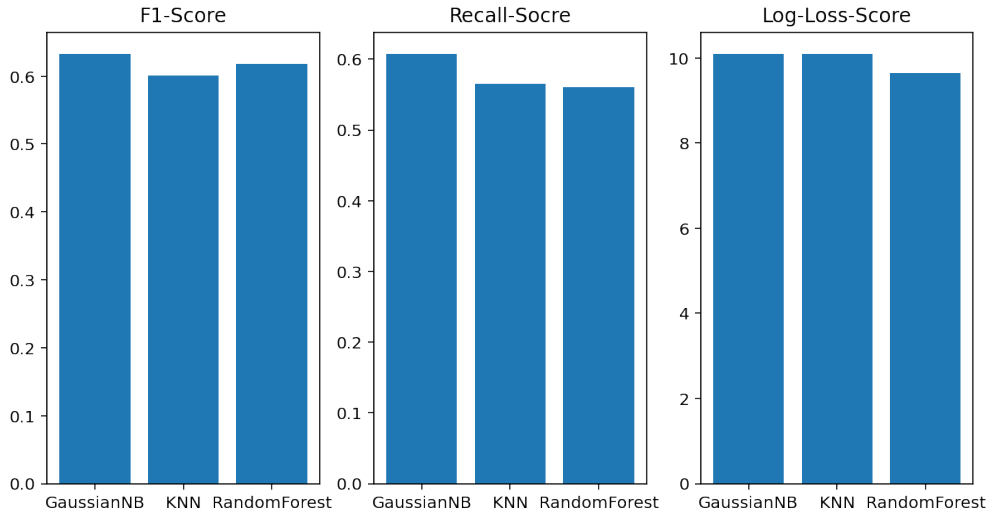

Figure 9: Baseline Record

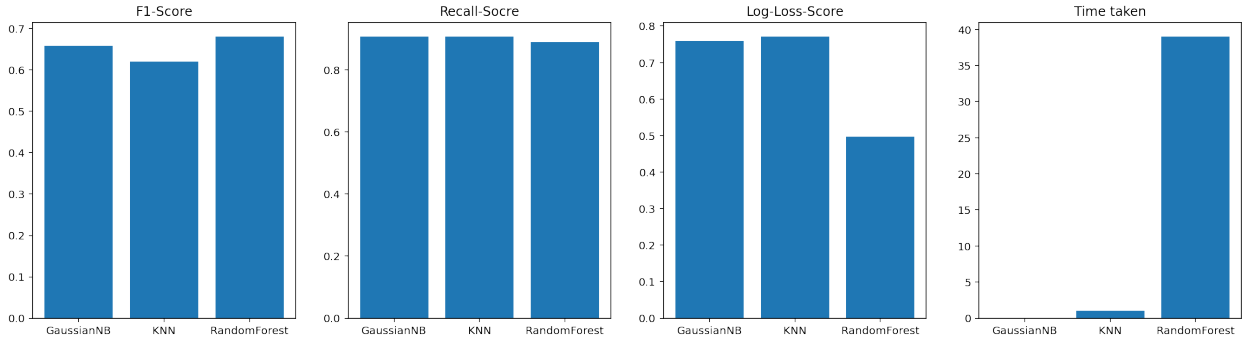

Figure 10: Optimal Record

5.Conclusions

Although RandomForest can achieve good F1-Score and low loss, it consumes too much time. So from the comprehensive evaluation results, GaussianNB can achieve great results.

By GaussianNB classification, it also has better support for top feature.

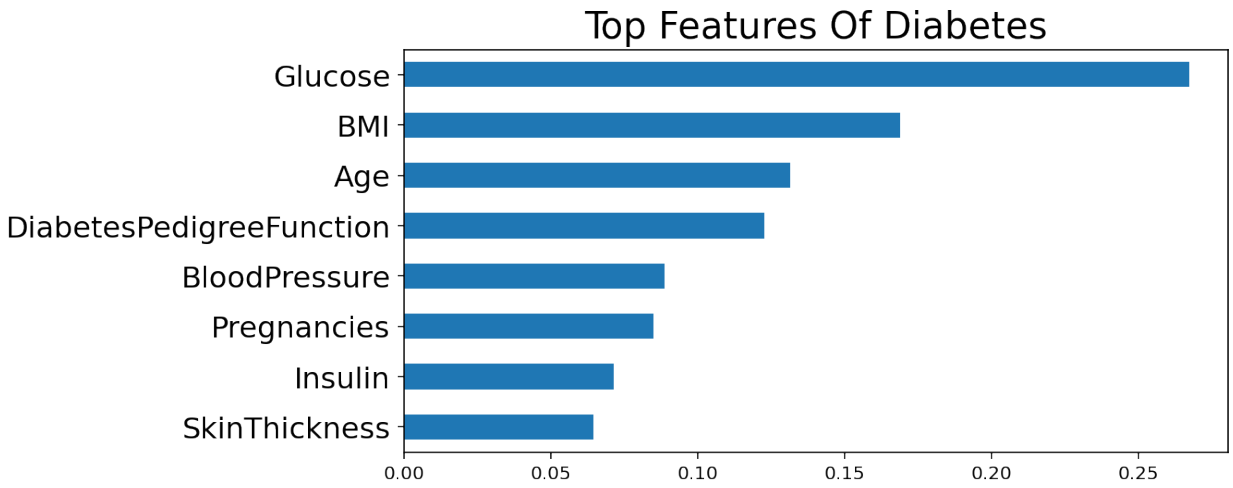

Figure 11: Top Features

So, we can judge subgroup of Glucose and BMI that are more likely to have diabetes.

6. Individual Effort

Yao Xiao-2019180015: 30%

Ruochen Liu-2019180018: 25%

Zebiao Guo-2019180020: 25%

Jun Wu-2019180006: 20%