Comparison of Models about Cardiovascular Disease Prediction

1 Abstract

The dataset selected was titled "Cardiovascular Disease dataset", It was downloaded from the Kaggle website (https://www.kaggle.com/sulianova/cardiovascular-disease-dataset), which consists of 70 000 records of patients and 11 features including personal information and results of medical examination, such as age, gender, systolic blood pressure, diastolic blood pressure, etc. The target 'cardio' is binary data, if cardio" equals 1 when patients suffer from cardiovascular disease, and if patients are healthy, it is 0. We can use machine learning to predict the presence or absence of cardiovascular disease (CVD), and explore what factors are deciding and important to determine whether people have cardiovascular disease or not.

2 Background and Research problem

To better predict someone contracts a cardiovascular disease or not. I used two different algorithms to generate two different models and get their scores comprising of precision, accuracy, recall, and f1-score. Then using the MacNemar test to verify whether there is any different performance between the two models. The null hypothesis for this study is that we will see no significant difference in performance between the two models. The alternative is that they perform differently.

3 Approach Description

There are three important stages before generating models, including data cleaning, exploratory data analysis, and standardize data. Because I changed the data set, this part I should summarize a little. This data set has 6 categorical variables and 5 numerical variables, so it is quite essential to standard them in order to train the best models. This part used the class of StandardScaler of preprocessing of sklearn library to normalize them.

The next part is about building models by using different classification algorithms. Firstly, this data is split up into training (70%) and testing (30%) data. For training data, there are 48983 records. For testing data, there are 20993 records.

Secondly, I have no idea what parameters are best to build models and predict correctly. In this case, I used the Grid Search and cross-validation algorithm to help me find the best parameters which will produce the most accurate estimates. For the Logistic Regression Classifier, the parameters I will tune are the type of penalty and the inverse regularization strength. For the Random Forest Classifier, I will tune the number of trees, the maximum depth of the trees, the maximum features, the minimum samples’ leaf, and the minimum samples’ split. These parameters are to reduce overfitting occurring, and get the best generalization performance.

Thirdly, I will use the models to get the predicted value by X_test data in order to see whether the model is useful or not and get the classification report to compare the models. Then I will use the McNemar test to see whether there are any significant different performances between the two models. if the p-value is less than 0.05, I will reject the null hypothesis, indicating they have different performance for prediction and vice versa.

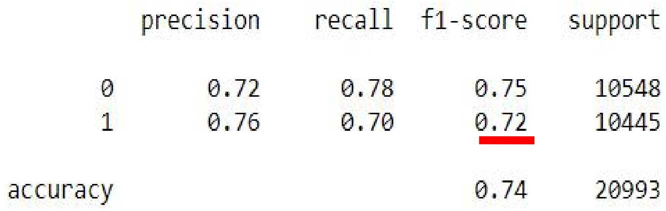
4 Evaluation Setup

If the result of the McNemar test show the two models perform differently, I will use the f1 score as our metric, as it is the combination of precision and recall. Also, I will use the feature importance coefficients to determine what factors have the greatest influence on the target, which will help us to analyze what features are highly related to the disease diagnosis.

5 Results and analysis

For the two models, as shown in appendix 1, for Random Forest Classification, the best score is 0.74 and the best parameter: n_estimator: 8, max_depth: 8, max_features: 0.8, min_samples_leaf: 2, min_samples_split: 9. For Logistic Regression, the best score is 0.72 and the best parameter: C: 0.1, penalty: l2, solver: liblinear.

As shown in figure 2, the result of the McNemar test told us that we should reject the null hypothesis, which means Random Forest and Logistic Regression have different classification performances. So I used f1-score by generating the classification report to compare which one is better to predict whether one person has the disease or not. The result can be seen in Figure 1 and Figure 2.

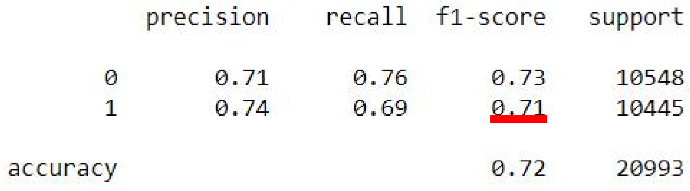


```
              precision    recall  f1-score   support

           0       0.72      0.78      0.75     10548
           1       0.76      0.70      0.72     10445

    accuracy                           0.74     20993
```

Figure 1: The Classification report of Random Forest Classifier



```
              precision    recall  f1-score   support

           0       0.71      0.76      0.73     10548
           1       0.74      0.69      0.71     10445

    accuracy                           0.72     20993
```

Figure 2: The Classification report of Logistic Regression Classifier

Figure 1 and Figure 2 showed the results of different metrics for the two models. We can see the f1-score that is underscored. The Random Forest classification is 0.72, which is a little bit higher than that of Logistic Regression 0.71. Also, other metrics such as precision, recall, and accuracy, Random Forest classification is higher than Logistic Regression. Therefore, I think the former is better than the latter in terms of prediction.

On the other hand, It is necessary to analyze The Confusion Matrix.

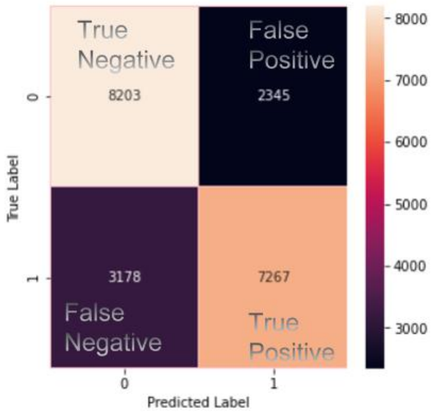


Figure3: Random Forest Confusion matrix

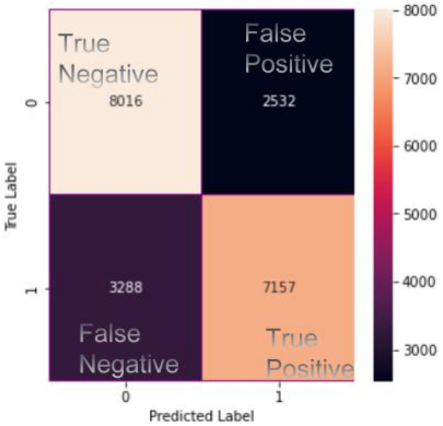
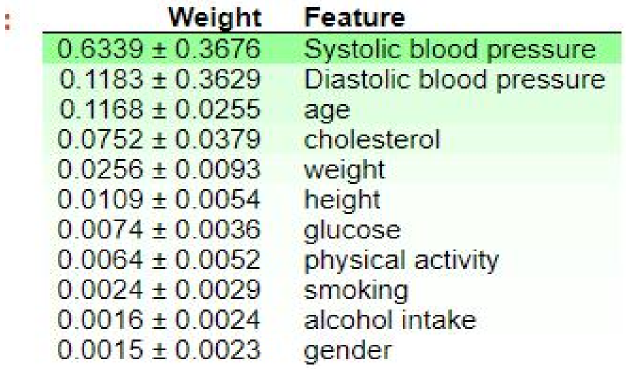


Figure4: Logistic Regression Confusion Matrix

As shown in Figures 3 and 4, True Positive is 7267 and 7157 respectively, False Negative is 3178 and 3288 respectively. FP of Random Forest is less than that of Logistic Regression, which means the model predicted the cardiovascular disease as a non-cardiovascular disease, it is a big threat to the patient's life. But it doesn't matter if the model predicts a non-disease case as a disease (False Positive) because they can do a physical examination later. In this case, the fewer False Negatives we have, the better. So from this metric, I also think Random Forest Classification is better.

Next, I imported the eli5 library which is very useful to help us explain and analyze the model. As shown in Figures 5 and 6, the weight of systolic blood pressure is approximately 0.6 - 0.7, which is extremely high compared with other features. This also indicated that if people who have high systolic blood pressure should pay more attention to their physical health, as higher relation, higher probability. The following are diastolic blood pressure, age, and cholesterol, but these factors' weight is around 0.1, having a little influence on the disease diagnosis compared with weight, height, glucose, and so on. At the same time, we can say that there is no relationship between gender and cardiovascular disease.

In conclusion, systolic blood pressure is highly related to cardiovascular disease, for what is the meaning of systolic and diastolic blood pressure, appendix 4 can be seen. However, there are still some drawbacks to my project. Firstly, the f1-score and other metrics are not high for the two models, it is quite risky to make the reckless decision about the disease diagnosis according to this model. So I think we need other more useful and significant features to train these models and get the best prediction. Secondly, I only compared two models, maybe other models like KNN, SVM can be helpful to our prediction.



Weight	Feature
0.6339 ± 0.3676	Systolic blood pressure
0.1183 ± 0.3629	Diastolic blood pressure
0.1168 ± 0.0255	age
0.0752 ± 0.0379	cholesterol
0.0256 ± 0.0093	weight
0.0109 ± 0.0054	height
0.0074 ± 0.0036	glucose
0.0064 ± 0.0052	physical activity
0.0024 ± 0.0029	smoking
0.0016 ± 0.0024	alcohol intake
0.0015 ± 0.0023	gender

Figure 5: Feature Importance Weight

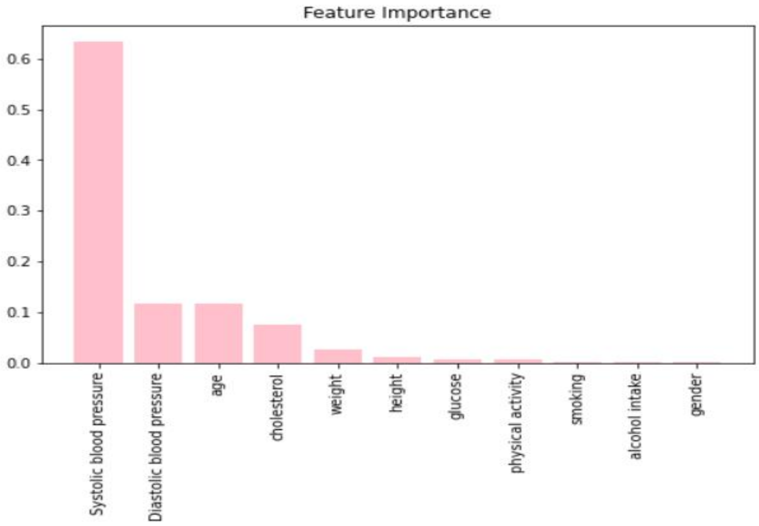


Figure 6: Feature Importance Plot

6 conclusion

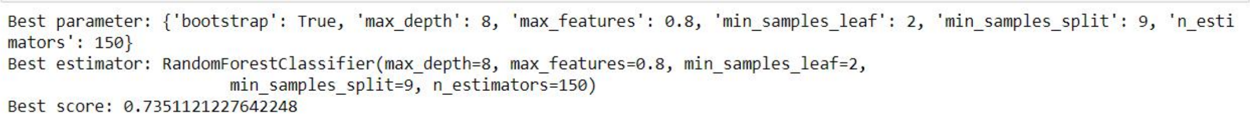
It is my first time doing data analysis and mining using machine learning. I think it is fun and interesting to explore data and gain insights from the data. So I will recommend this solution to solve the problem because it is difficult to discover underlying stuff and causes by just observing the data, and sometimes, maybe we will make incorrect decisions due to our wrong intuition or experience. In the future, as data analysts, we not only need to master the professional skills of data analysis, such as python or R language, but also the professional knowledge of the industry that is analyzed, such as medical care, finance, and e-commerce and so on. Only in this case can we become all-around data analysts and enhance our irreplaceability in our professional field. Finally, it is a nice semester, and I learned much knowledge about data science. Thank all of you!

References

[1] https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

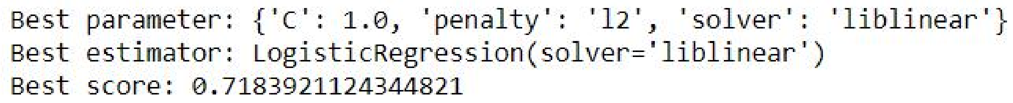
[2] https://www.diffen.com/difference/Diastolic_vs_Systolic

Appendices



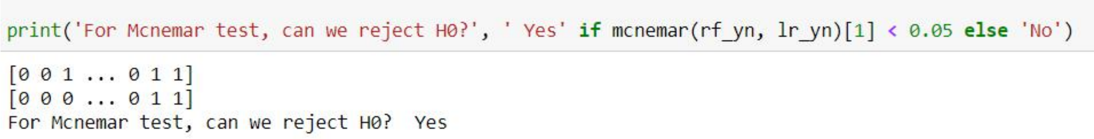
```
Best parameter: {'bootstrap': True, 'max_depth': 8, 'max_features': 0.8, 'min_samples_leaf': 2, 'min_samples_split': 9, 'n_esti
mators': 150}
Best estimator: RandomForestClassifier(max_depth=8, max_features=0.8, min_samples_leaf=2,
                       min_samples_split=9, n_estimators=150)
Best score: 0.7351121227642248
```

Figure 1 - best parameters and score of Random Forest Classification



```
Best parameter: {'C': 1.0, 'penalty': 'l2', 'solver': 'liblinear'}
Best estimator: LogisticRegression(solver='liblinear')
Best score: 0.7183921124344821
```

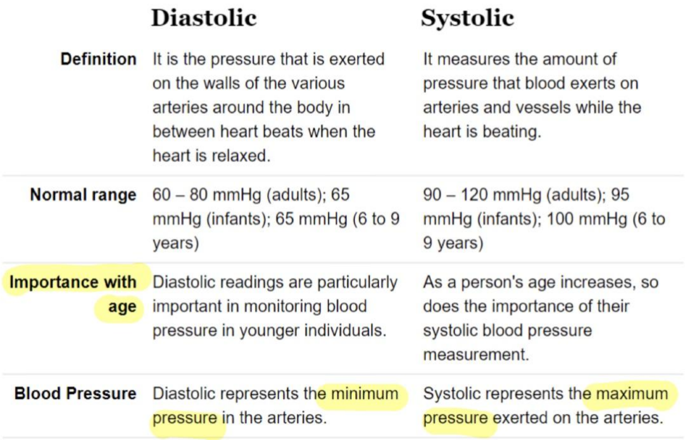
Figure 2 - best parameters and score of Regression Logistic Classification



```
print('For Mcnemar test, can we reject H0?', ' Yes' if mcnemar(rf_yn, lr_yn)[1] < 0.05 else 'No')
```

```
[0 0 1 ... 0 1 1]
[0 0 0 ... 0 1 1]
For Mcnemar test, can we reject H0?  Yes
```

Figure 3 - Mcnemar test result



	Diastolic	Systolic
Definition	It is the pressure that is exerted on the walls of the various arteries around the body in between heart beats when the heart is relaxed.	It measures the amount of pressure that blood exerts on arteries and vessels while the heart is beating.
Normal range	60 – 80 mmHg (adults); 65 mmHg (infants); 65 mmHg (6 to 9 years)	90 – 120 mmHg (adults); 95 mmHg (infants); 100 mmHg (6 to 9 years)
Importance with age	Diastolic readings are particularly important in monitoring blood pressure in younger individuals.	As a person's age increases, so does the importance of their systolic blood pressure measurement.
Blood Pressure	Diastolic represents the minimum pressure in the arteries.	Systolic represents the maximum pressure exerted on the arteries.

Figure 4 - The difference between diastolic and systolic blood pressure