Heart Disease Prediction with Various Machine Learning Models

Shihao Hu
Department of Computer Science
Western University
London, Canada
shu223@uwo.ca

Xiangli Zhang
Department of Computer Science
Western University
London, Canada
xzha453@uwo.ca

***Abstract*—Predicting and discovering cardiac disease has always been a difficult and time-consuming undertaking for doctors. To treat cardiac disorders, hospitals and other clinics are giving costly therapies and operations. As a result, detecting cardiac disease in its early stages will be beneficial to people all around the world, allowing them to take the required treatment before it becomes serious. Heart disease has been a major issue in recent years, with the primary causes being excessive alcohol use, tobacco use, and a lack of physical activity. Machine learning has shown to be useful in making decisions and predictions from given volumes of data produced by the healthcare industry over time. Logistic regression (LR), Random forest (RF), support vector machine (SVM), neural network (NN), and nave Bayes (NB) are some of the supervised machine learning algorithms employed in this prediction of heart disease. This paper is dedicated to evaluating the effectiveness of these models for heart disease detection and some possibilities for optimization.**

I. INTRODUCTION

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, killing an estimated 17.9 million people each year, accounting for 31% of all deaths. Heart attacks and strokes account for four out of every five CVD deaths, with one-third of these deaths occurring before the age of 70. CVDs are a common cause of heart failure. People with heart disease or who are at high risk for heart disease require early detection and management, which a machine learning model can provide. In this paper, we first analyze the feature distribution and correlation of the data. Then we train different machine learning models for the data. We then compare the accuracy of different machine learning models. Finally, we select the best model to find feature importance for the data. Also, we obtain a classification report and a confusion matrix of each model for analyzing their performance. The next step we are done is we use five cross-validation methods to evaluate each model and get an average score for validating the performance of each model. In the end, we infer the importance of each public feature to the target feature and draw some conclusions based on the analyzed results. The results we obtained will contribute some fundamental research findings and knowledge to the academia of machine learning and disease diagnosis category, and also will help doctors in identifying the risk of whether a patient will have heart disease. This paper is composed in the following structure: Section 2 introduces the background and related work, section 3 declares our research objectives, significance and the methodology that we use, and section 4 demonstrates all the resulting details. Finally, section 5 concludes this paper along with the view of future work and some lessons we learnt from this research.

II. BACKGROUND AND RELATED WORK

The heart is one of the main parts of the human body after the brain. The primary function of the heart is to pump blood to the whole body parts. Heart disease refers to any condition that can impair the heart's ability to operate. The most frequent heart disorders are coronary artery disease and heart failure. Heart disease refers to any condition that can impair the heart's ability to operate. There are several types of heart illness in the world; the most frequent heart disorders are coronary artery disease and heart failure.

The main cause of coronary heart disease is a narrowing or blockage of the coronary arteries. Heart failure can be caused by a variety of conditions. Medical scientists have divided these elements into two groups: risk factors that cannot be adjusted and risk factors that can be changed. Risk factors that cannot be changed include family history, sex, and age. High blood pressure, high cholesterol, smoking, physical inactivity, and high blood pleasure are all risk factors [1]. Due to the prevalence of bad practices, heart disease has risen substantially in recent years to become the top cause of mortality in adults in the United States. These include a decrease in physical activity as technology increasingly replaces human physical activity and bad eating habits, both of which are connected to an increased risk of heart disease. The National Cardiac, Lung, and Blood Institute defines heart disease as a disruption in the heart's normal electrical system and pumping processes. Where the disease impairs the heart's ability to properly pump blood. Furthermore, the World Health Organization (WHO) estimates that 17.9 million people worldwide die each year from cardiovascular illnesses, accounting for 31 percent of all deaths.

This necessitates the development of a cost-effective system capable of providing a preliminary diagnosis of a patient based on inexpensive medical testing[2]. Numerous current studies are being conducted by researchers on the prediction and analysis of cardiac disease. The author uses the random forest in [3] with the Cleveland dataset to investigate cardiac disease.

For the investigation, the author employed the Chi-Square feature selection model and a feature selection model based on a genetic algorithm (GA). They demonstrated that their proposed model with Genetic algorithm feature selection provided more accuracy than current models in experimental findings. The outcomes, on the other hand, are assessed using established machine learning models. The author of [4] discussed the use of data mining techniques to forecast cardiac disease. They used approaches including the KNN algorithm, decision tree algorithm, neural network classifications, and Bayesian classification algorithms to research and assess. The use of the genetic algorithm in feature selection for heart disease important traits was also investigated by the author. Researchers have shown that machine learning algorithms perform exceptionally well when assessing medical data sets in recent years. These data sets will be fed straight into machine learning algorithms, which will perform in accordance with their nature and produce results.

III. METHODS

A. RESEARCH OBJECTIVES

1) Analyze the feature distribution and correlation of the data.
2) Train different machine learning models for the data.
3) Compare the accuracy of different machine learning models.
4) Select the best model to find feature importance for the data.

Since the prediction of heart disease has always been a issue, the study and explore of different machine learning model will contribute fundamental research findings and knowledge into the academia of machine learning and disease diagnosis category, opening the way for future research on advanced topics incorporating unique machine learning algorithms and optimizing approach by achieving the objectives. In practice, it will also assist hospitals in identifying the risk of whether a patient will have heart disease, and help medical companies develop drugs to prevent heart disease. Thus, the objectives we set could help us achieve the final goal more consistently and effectively.

B. RESEARCH METHODOLOGY

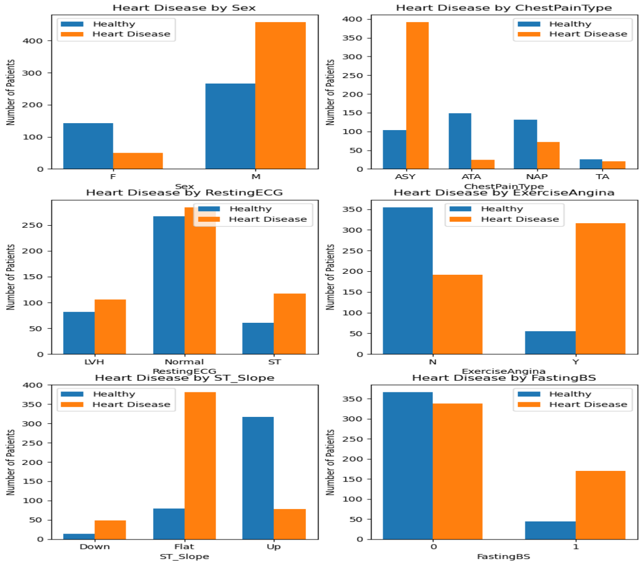
The problem to be solved by this project is how to identify which patients have heart disease and which patients don't have heart disease. This dataset was developed by integrating other datasets that were previously available but had not been combined. This dataset combines 5 heart datasets with 11 shared features to provide the largest heart disease dataset available for research purposes to date[5]. The final dataset contains 918 observations totally. The response variable is called 'HeartDisease', which is a binary response where 1 indicates a patient has heart disease and 0 indicates a patient has no heart disease. The variable dictionary is shown in Table 1.

TABLE I
DATA DICTIONARY

Variable	**Variable Explanation**
Age	age of the patient [years]
Sex	sex of the patient [M: Male, F: Female]
ChestPainType	chest pain type
RestingBP	resting blood pressure [mm Hg]
Cholesterol	serum cholesterol [mm/dl]
FastingBS	fasting blood sugar [1: if FastingBS ¿ 120 mg/dl, 0: otherwise]
RestingECG	resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality, LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
MaxHR	maximum heart rate achieved [Numeric value between 60 and 202]
ExerciseAngina	exercise-induced angina [Y: Yes, N: No]
Oldpeak	oldpeak = ST [Numeric value measured in depression]
ST_Slope	the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
HeartDisease	output class [1: heart disease, 0: Normal]

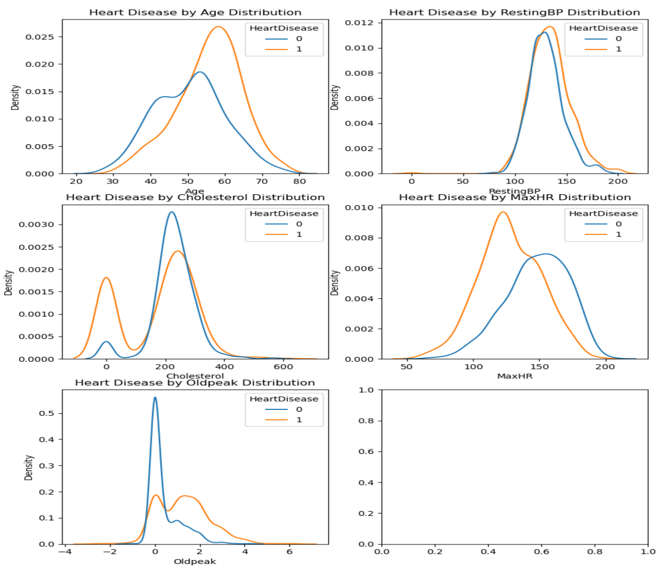
In the dataset, the 11 variables will be used as input values, 6 of which are categorical variables.. Their distribution is shown in figure 1.

Fig. 1. Distribution of categorical variables



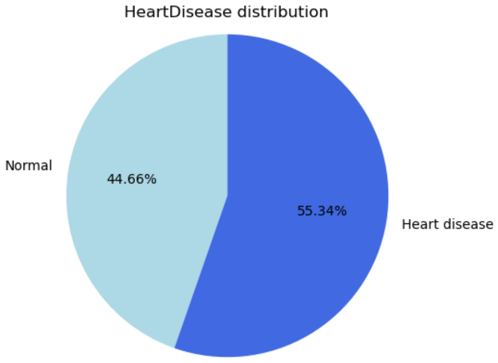
The remaining 5 variables are numeric variables. Their distribution is shown in figure 2.

Fig. 2. Distribution of numeric variables



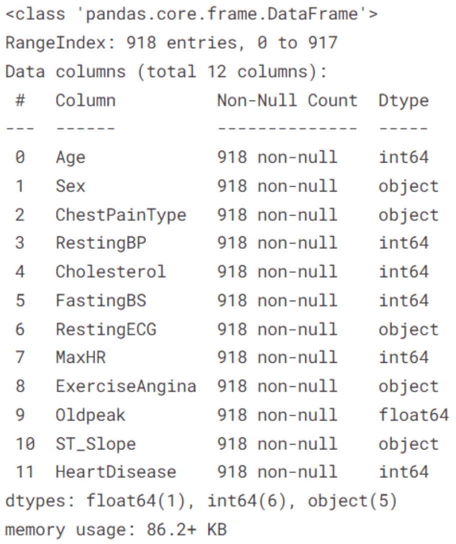
In the classification task, unbalanced data refers to an issue in which the classes are not evenly distributed. During the training process, it may cause the model to be more prone to categorise the data into the majority category, reducing the model's true generalisation capabilities. Figure 3 shows that 'Heart disease' is labelled on 55 percent of the training data, whereas 'normal' is labelled on 45 percent. Because the classes are nearly evenly divided, no extra sampling procedures are required to balance the data.

Fig. 3. Distribution of numeric variables



12 variables will be used to predict heart disease, and Figure 4 shows that none of the 12 variables in the dataset have null values, indicating that there are no missing values.

Fig. 4. Data information



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             918 non-null    int64
 1   Sex             918 non-null    object
 2   ChestPainType   918 non-null    object
 3   RestingBP       918 non-null    int64
 4   Cholesterol     918 non-null    int64
 5   FastingBS       918 non-null    int64
 6   RestingECG      918 non-null    object
 7   MaxHR           918 non-null    int64
 8   ExerciseAngina  918 non-null    object
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    object
 11  HeartDisease    918 non-null    int64
dtypes: float64(1), int64(6), object(5)
memory usage: 86.2+ KB
```

The goal of the research is to create a classification model that uses Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Multi-layer Perceptron (MLP), and Gaussian Naive Bayes (GaussianNB) to distinguish between patients who have heart disease and those who do not.

Logistic regression is a statistical analytic approach for predicting a binary result, such as 'heart disease' or 'normal' for this research. By evaluating the relationship between 11 independent factors, a logistic regression model can predict the outcome of the 'heartdisease' variable.

Random forest is a supervised machine learning algorithm that is commonly used to solve classification and regression problems. For this research, it is a classification problem. RF creates decision trees from several samples and uses the majority vote to classify them.

Support Vector Machine is a supervised machine learning technique that can be used for classification and regression. We shall utilize SVMs because they are more often used in classification situations. SVMs are based on the concept of determining the optimum hyperplane for dividing a dataset into two classes.

Multilayer perceptron is a neural network connecting multiple layers in a directed graph. For this model, we have 1 hidden layer with 100 hidden units. Aside from the input nodes, each node has a nonlinear activation function. The activation function we used is the Relu function. MLP uses backpropagation as a supervised learning technique. Our model optimizes the log-loss function using adam, with a learning rate of 0.001. The maximum number of iterations is 2000.

Gaussian Naive Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data. Naive Bayes is a group of supervised machine learning classification algorithms based on the Bayes theorem. It is a simple classification technique but has high functionality. They find use when the dimensionality of the inputs is high.

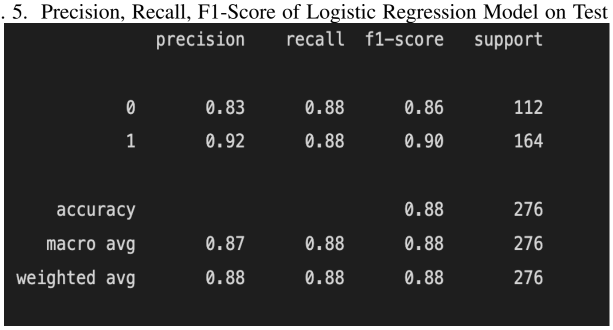
IV. RESULTS

We trained each of the models on a training set of 642 observations. The test set of 276 observations is used for prediction to assess the generalisation capabilities of trained models. The test set is preprocessed in the same way as training data is before being fed to models.

A. Logistic Regression

Figure 5 shows the values of precision, recall, and F1 score according to the prediction result on test set from Logistic Regression, while Figure 6 is the corresponding confusion matrix. The F1 score of the Logistic Regression is 0.88, which is regarded as the baseline of the model performance.

Fig. 5. Precision, Recall, F1-Score of Logistic Regression Model on Test Set

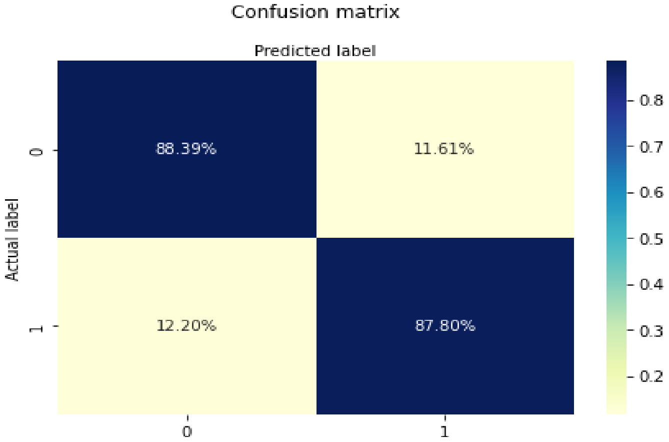


```
              precision    recall  f1-score   support

           0       0.83      0.88      0.86       112
           1       0.92      0.88      0.90       164

    accuracy                           0.88       276
   macro avg       0.87      0.88      0.88       276
weighted avg       0.88      0.88      0.88       276
```

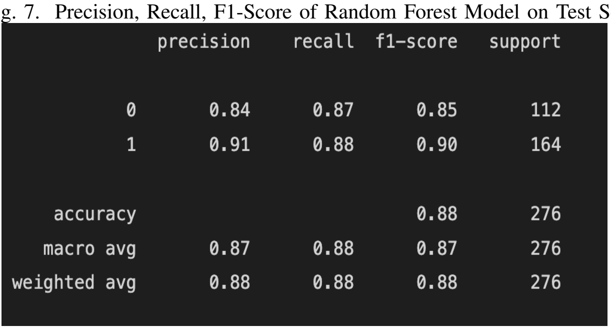
Fig. 6. Confusion Matrix of Logistic Regression Model on Test Set



B. Random Forest Classification

Figure 7 shows the values of precision, recall, and F1 score according to the prediction result on test set from Random Forest Model, while Figure 8 is the corresponding confusion matrix.. The F1 score of the Random Forest Model is 0.88

Fig. 7. Precision, Recall, F1-Score of Random Forest Model on Test Set

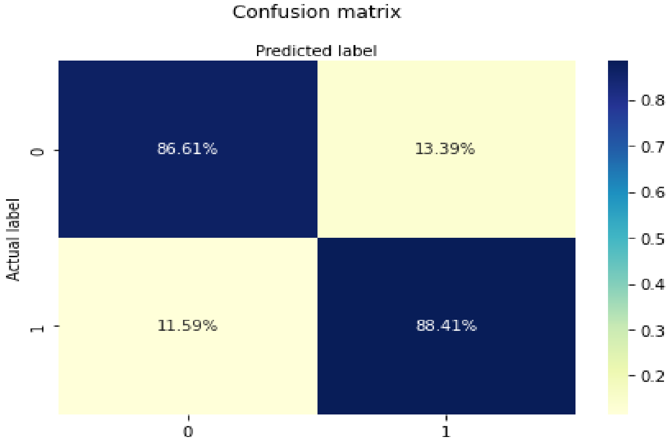


```
              precision    recall  f1-score   support

           0       0.84      0.87      0.85       112
           1       0.91      0.88      0.90       164

    accuracy                           0.88       276
   macro avg       0.87      0.88      0.87       276
weighted avg       0.88      0.88      0.88       276
```

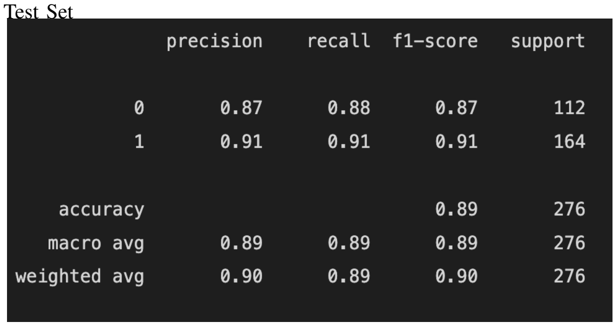
Fig. 8. Confusion Matrix of Random Forest Model on Test Set



C. Support Vector Classification

Figure 9 shows the values of precision, recall, and F1 score according to the prediction result on test set from Support Vector Classification (SVC) Model, while Figure 10 is the corresponding confusion matrix. The F1 score of the Support Vector Classification Model is 0.89, which is slightly higher than the result from LR and Random Forest Model.

Fig. 9. Precision, Recall, F1-Score of Support Vector Classification Model on Test Set

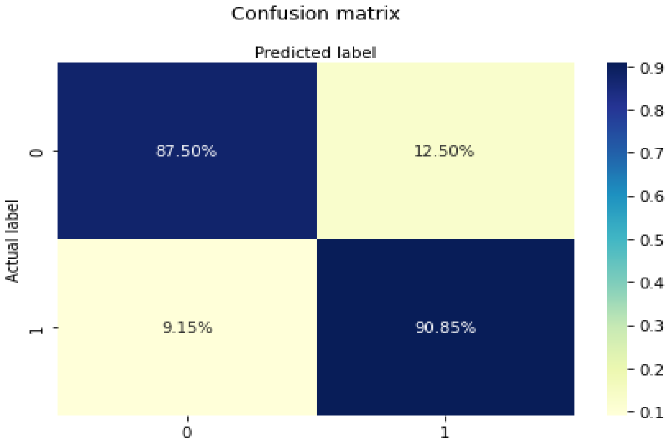


```
              precision    recall  f1-score   support

           0       0.87      0.88      0.87       112
           1       0.91      0.91      0.91       164

    accuracy                           0.89       276
   macro avg       0.89      0.89      0.89       276
weighted avg       0.90      0.89      0.90       276
```

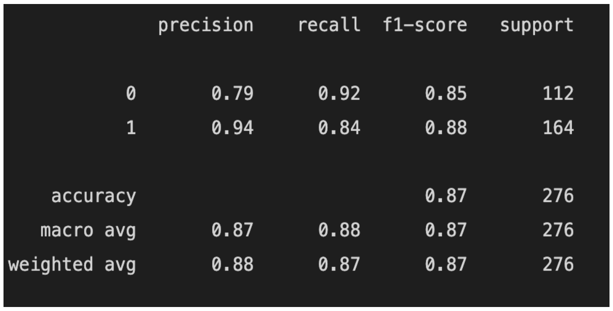
Fig. 10. Confusion Matrix of Support Vector Machine Model on Test Set



D. Multi-layer Perception classifier

Figure 11 shows the values of precision, recall, and F1 score according to the prediction result on test set from Multi-layer Perception Model, while Figure 12 is the corresponding confusion matrix.. The F1 score of the Multi-layer Perception Model is 0.87, which is slightly lower than the result from LR and Random Forest Model.

Fig. 11. Precision, Recall, F1-Score of Multi-layer Perception Model on Test Set

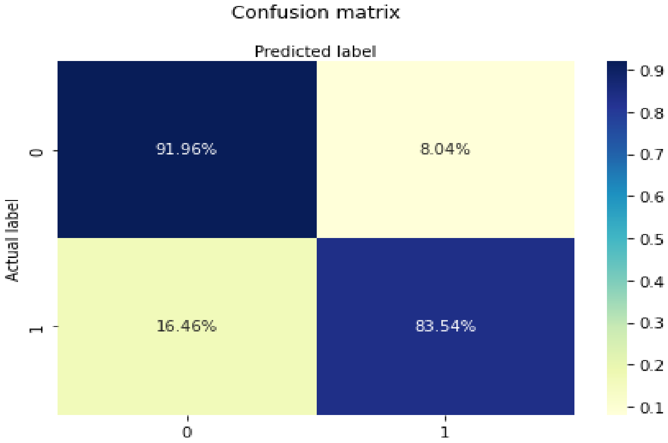


```
              precision    recall  f1-score   support

           0       0.79      0.92      0.85       112
           1       0.94      0.84      0.88       164

    accuracy                           0.87       276
   macro avg       0.87      0.88      0.87       276
weighted avg       0.88      0.87      0.87       276
```

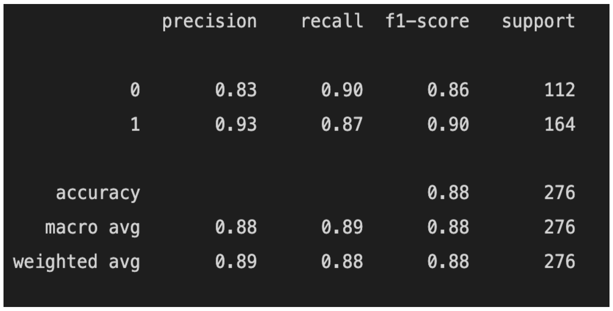
Fig. 12. Confusion Matrix of Multi-layer Perception Model on Test Set



E. Gaussian Naive Bayes

Figure 13 shows the values of precision, recall, and F1 score according to the prediction result on test set from Gaussian Naive Bayes Model, while Figure 14 is the corresponding confusion matrix.. The F1 score of the Gaussian Naive Bayes Model is 0.88, which is the same as the result from LR and Random Forest Model.

Fig. 13. Precision, Recall, F1-Score of Gaussian Naive Bayes Model on Test Set

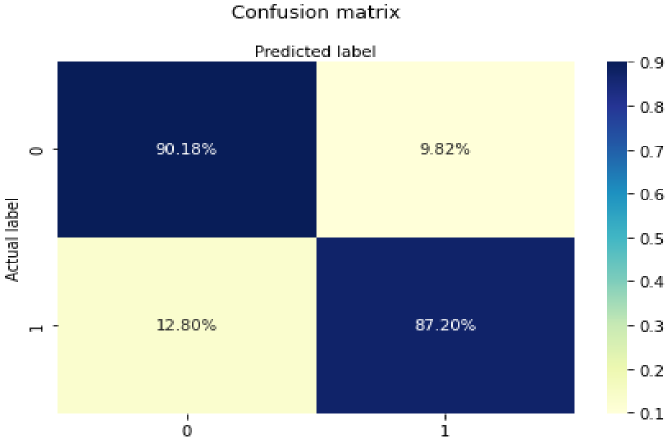


```
              precision    recall  f1-score   support

           0       0.83      0.90      0.86       112
           1       0.93      0.87      0.90       164

    accuracy                           0.88       276
   macro avg       0.88      0.89      0.88       276
weighted avg       0.89      0.88      0.88       276
```

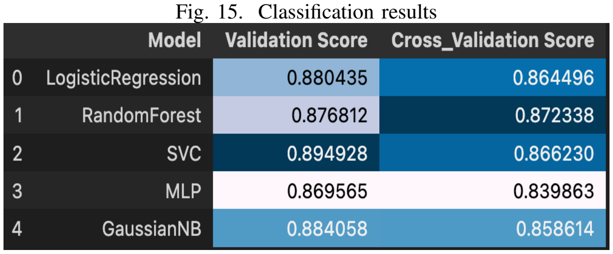
Fig. 14. Confusion Matrix of Gaussian Naive Bayes Model on Test Set



V. CONCLUSIONS & FUTURE WORK

Since the dataset is small, we also apply cross-validation to train the model entire dataset and obtain an average accuracy score for each model. Figure 15 shows the accuracy scores under the two training methods.

Fig. 15. Classification results

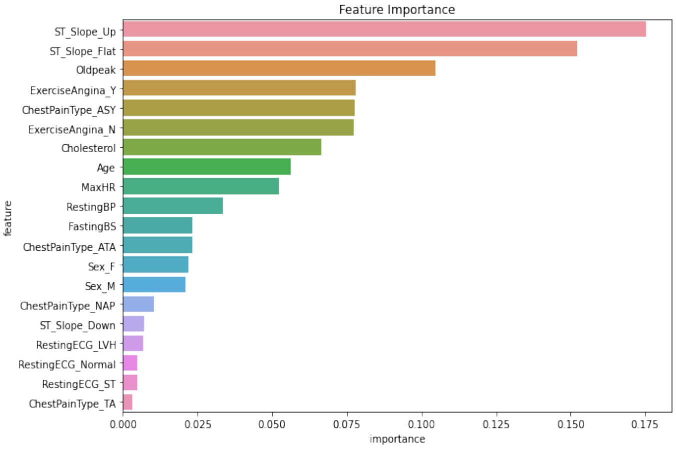


	Model	Validation Score	Cross_Validation Score
0	LogisticRegression	0.880435	0.864496
1	RandomForest	0.876812	0.872338
2	SVC	0.894928	0.866230
3	MLP	0.869565	0.839863
4	GaussianNB	0.884058	0.858614

Comparing the accuracy scores of different models, it can be seen that SVC predicts best when there is no cross-validation.

When there is cross-validation, RF has the best prediction. Thus RF generalizes better, has better prediction performance on new data, and with less overfitting. Based on the RF model, we obtained the feature importance of the data, as shown in Figure 16.

Fig. 16. Feature importance distribution



From Figure 16, we could indicate the top three features that have an impact on heart failure, which are ST_Slope_Up, ST_Slope_Flat and Oldpeak. Also, we could infer that the slope of the peak exercise ST-segment influence the target feature (heart failure) the most, especially when the ST slope is going up and flat. Moreover, we could see that the type of chest pain has minimal effect on heart failure.

In the future, we will try to use a larger dataset and will try to apply the decision tree model and K-nearest neighbour model to train the data and compare the test accuracy and cross-validation score with those models we already investigated. And we will also compare, pick up the best model and utilize it if possible. From this project, we learned that the score of test accuracy is not the only metric for evaluating the model, it is necessary to use cross-validation for making the result more persuasive.

References

[1] R. Katarya and P. Srinivas, “Predicting Heart Disease at Early Stages using Machine Learning: A Survey,” 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 302-305, doi: 10.1109/ICESC48915.2020.9155586.

[2] R. Atallah and A. Al-Mousa, “Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method,” 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS), 2019, pp. 1-6, doi: 10.1109/ICTCS.2019.8923053.

[3] Jabbar, M. A., B. L. Deekshatulu, and Priti Chandra. “Intelligent heart disease prediction system using random forest and evolutionary approach.” Journal of Network and Innovative Computing 4.2016 (2016): 175-184.

[4] Soni, Jyoti, et al. “Predictive data mining for medical diagnosis: An overviewof heart disease prediction.” International Journal of Computer Applications 17.8 (2011): 43-48.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.