Portfolio Report for CSE 575: Statistical Machine Learning

Shreyansh Khandelwal
Ira A. Fulton Schools Of Engineering — School of Computing and Augmented Intelligence
Arizona State University
Email: skhand17@asu.edu

***Abstract*—The number of cardiovascular ailments has reached an all-time high in recent years. According to estimations, around 18 million deaths (32% fatalities) were caused by CVDs in 2019, with heart attack and stroke deaths accounting for 85% of them. Early detection and prevention are critical for the development of successful health care. Machine learning has proven to be quite beneficial in producing predictions and detecting anomalies. We tested a range of cutting-edge models to predict the possibility of heart illness, and we obtained results on which models performed better and could be used to save many lives in the future.**

***Index Terms*—Cardiovascular disease, machine learning, dataset, confusion matrix, ROC and AUC curve, accuracy, precision, recall, f-1 scores, Logistic regression, GridSearchCV, RandomSearchCV.**

I. Introduction to Project-I

Cardiovascular diseases (CVD) encompass a range of conditions affecting the circulatory system, primarily involving the heart and blood vessels. These conditions can be categorized into various types, including coronary CVD, cerebrovascular CVD, peripheral arterial CVD, rheumatic CVD, congenital heart CVD, and conditions like deep vein thrombosis and pulmonary embolism. Strokes and heart attacks, which result from blockages in blood vessels, are among the most significant acute events associated with CVD.

The spread of electronic patient records, with their computer-readable entries e.g. Magnetic Resonance Imaging (MRI), signals like ECG (Electrocardiography), clinical information like blood sugar, blood pressure, cholesterol levels etc. as well as the physician's interpretation is opening new possibilities for medical data mining and a world of virtual research [1]. Heart disease is recognized as a leading cause of death in both high-income and low-income countries, according to the World Health Organization. Machine learning (ML), a subset of artificial intelligence (AI), is gaining prominence in the field of cardiovascular care. ML employs models that take input data, such as images or text, to predict outcomes through mathematical optimization and statistical analysis. ML techniques are increasingly used for predicting and diagnosing CVD. Researchers have harnessed data mining techniques to diagnose heart disease and provide progressive measures for early detection and insights into disease patterns. Data classification, a common task in machine learning, plays a vital role in this process. While ECG machines are widely used in cardiac centers and hospitals for detecting CVD, they have limitations. Therefore, experts and practitioners are exploring classical machine learning techniques to enhance heart disease diagnosis, which serves as the motivation behind the project. This summary outlines the significance of CVD, the role of machine learning in its diagnosis, and the motivation behind using classical machine learning techniques in this context. Medical experts may overlook important details while assessing a patient's heart alignment; as a result, a heart alignment prediction approach based on machine learning algorithms might help in such circumstances [2].

II. DESCRIPTION OF THE SOLUTION

Our core objective was to detect heart disease using machine learning models trained on two distinct datasets and compare their performance. In a collaborative team of six, we established the initial environment, set benchmarks, and outlined our project framework. Primarily, I focused on dataset 1, obtained from the UCI database and widely used by machine learning researchers. The project began with comprehensive data exploration through exploratory data analysis, a pivotal step in understanding the dataset and revealing patterns. We meticulously compared feature variables, some outside our expertise, against the target variable, indicating heart disease presence (1) or absence (0). A correlation matrix was constructed to highlight variable impacts, setting the stage for subsequent analyses and model development. The subsequent phase centered on model training, where we selected four models namely Logistic Regression, K-Nearest Neighbors, RandomForest Classifier, and Gradient Boost Classifier to predict the target variable. Decision trees are powerful and popular for both classification and prediction. They are also useful for exploring data to gain insight into the relationships of a large number of candidate input variables to a target variable [3].

A. Dataset Exploration

This project aims to predict the presence or absence of heart disease using various features as predictors. These features include age, sex (coded as 1 for males and 0 for females), chest pain (CP), resting blood pressure (trestbps in mm Hg), serum cholesterol levels (in mg/dL), fasting blood sugar levels (above or below 120 mg/dL), resting electrocardiographic (ECG) data, maximal heart rate during physical activity (thalach), exercise-induced angina (exang), ST depression during exercise (Oldpeak), the slope of the ST segment during exercise, the number of main vessels colored using fluoroscopy (ca), and the result of the thallium stress test. The study explores the relationships between these feature variables and the target variable, which indicates the presence (1) or absence (0) of heart disease. Data exploration and preliminary analysis were performed to identify patterns and correlations. A correlation matrix was created to assess the relationships between the independent features and the target variable, helping to determine whether they have a positive or negative influence on the presence of heart disease. Understanding these relationships is crucial for developing a predictive model for heart disease and highlights the importance of domain knowledge and data exploration in the field of machine learning and healthcare analysis.

1) Dataset I Exploration: I started off creating a scatter plot to see age relation with maximum heart rate to understand about the dataset-I that I worked upon. The scatter plot can be viewed in 1.

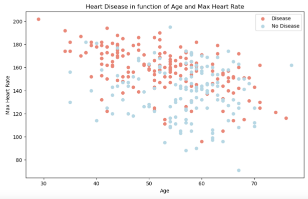


Fig. 1. Heart Disease as Age and Max Heart Rate

Moving forward, I created a correlation matrix where I could figure out the positive and negative correlations of independent variables with target variable. The correlation matrix can be viewed in 2.

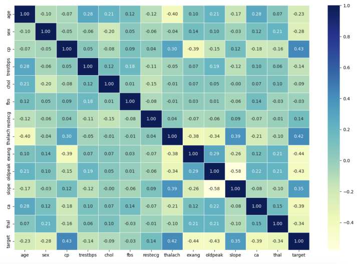


Fig. 2. Correlation Matrix

Few other data exploratory that I performed as part of this project work includes Heart disease frequency per chest pain type as can be viewed in 3 and normal distribution of age with target. Please refer 4.

2) Dataset I Evaluation and Correlation: Data analysis was conducted in order to identify trends within the datasets. The analysis commenced with a comparison of several feature variables among themselves, as well as their comparison with the target variable. The target variable 1 indicates the presence of cardiac disease, while the value 0 signifies its absence. Furthermore, a correlation matrix was constructed to ascertain the impact of independent factors on our target variable, discerning whether this influence is positive or negative. In order to enhance our understanding and analysis of the data, we generated additional bar charts.

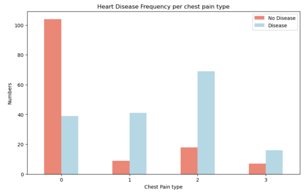


Fig. 3. Heart disease frequency per chest type

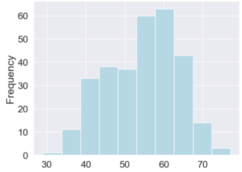


Fig. 4. Normal Distribution of age with target

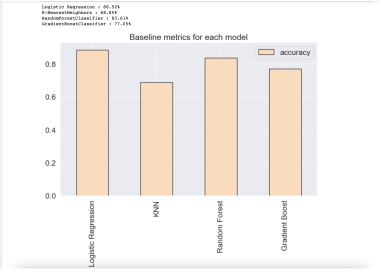
B. Models Description

Our next step was to perform Model training. We have selected four models: Logistic Regression, K-Nearest Neighbors', Random Forest Classifier and Gradient Boost classifier to predict the target variable. The Scikit-Learn machine learning map aided us in understanding which category we fall into and what classification models we can pursue. We trained and split our Model into 80:20 ratio, a rule of thumb is to use 80% for training and 20% for testing. When tested on test data, these models revealed our baseline metrics; we could see that Logistic Regression surpassed the rest, whereas KNN accuracy was the least. Gradient Boost and Random Forest provided almost the same accuracy. Following this, we evaluated our models using evaluation metrics.

I have conducted an analysis and implemented models on the dataset. Each of the remaining team members individually conducted data analysis on dataset 2 and identified two models from dataset 2 as their respective objectives for the project. I conducted experiments using various combinations of the aforementioned models. To optimize their performance, we employed the GridSearchCV and RandomSearchCV techniques. The baseline metrics were revealed by these models during evaluation on the test data. The baseline metrics could be viewed in the 5.

III. RESULTS AND CONCLUSION

The hyper-parameters of all models were tweaked in order to optimize their respective parameters and enhance the accuracy of each model. The hyper-parameter tuning process primarily involved the utilization of RandomSearchCV and GridSearchCV techniques. Each time we changed the hyper-parameter settings, we got different results. The validation set was not utilized due to the limited size of the initial dataset, which consisted of just 303 patient records that were used for both testing and training purposes. However, we employ



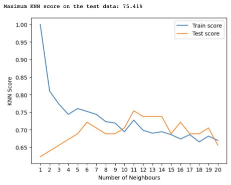
```
Logistic Regression : 88.52%
K-NearestNeighbors : 68.85%
RandomForestClassifier : 83.61%
GradientBoostClassifier : 77.05%
```

Fig. 5. Baseline metrics for each model

a technique known as k-fold cross-validation.Cross-validation is a statistical method used to estimate the skill of machine learning models [4]. The data is partitioned into k-folds and the Model is evaluated on each fold in order to mitigate reliance on a single train and test split. A five-fold train and test split was employed, as it is a commonly utilized approach among machine learning researchers.

Ultimately, our strategy entailed optimizing the hyper-parameters of the model and thereafter evaluating their performance through comparison. Additionally, we conducted cross-validation, plotted receiver operating characteristic (ROC) and area under the curve (AUC) curves, constructed a confusion matrix, calculated precision, recall, and F-1 score metrics, and determined the significance of the features.

After comparing KNN metrics with different hyper-parameter values (n_neighbors), I noticed that it performed best when n_neighbors == 11. The maximum score retrieved after tuning this model was around 75% as seen in fig 6.



```
Maximum KNN score on the test data: 75.41%
```

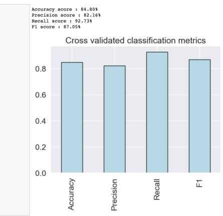
Fig. 6. Tuning with KNN

The best hyperparameters I got after tuning each of my machine learning models using RandomSearchCV and GridSearchCV are listed below:

- Logistic Regression:
  - Input: C is np.logspace(-4, 4, 20).
  - Input: solver is liblinear.
  - Output: C is 0.2335721.
  - Output: solver is liblinear.
- Random Forest Classifier
  - Input: n_estimators is np.arange(10, 1000, 50).
  - Input: max_depth is [None, 3, 5, 10].
  - Input: min_samples_split is np.arange(2, 20, 2).
  - Input: min_samples_leaf is np.arange(1, 20, 2).
  - Output: n_estimators is 210.
  - Output: min_samples_split is 4.
  - Output: min_samples_leaf is 19.
  - Output: max_depth is 3.

On applying the best combination of parameters using RandomSearchCV, I could see a slight increase in each model's accuracy.

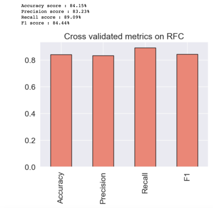
- Logistic Regression Evaluation Metric on Accuracy, Precision, Recall and F-1 score can be viewed in fig 7.



```
Accuracy score : 84.80%
Precision score : 82.16%
Recall score : 92.73%
F1 score : 87.05%
```

Fig. 7. Metric on Logistic Regression

- Random Forest Classifier Evaluation metrics on accuracy, precision, recall and f-1 score can be viewed in fig 8.



```
Accuracy score : 84.15%
Precision score : 83.23%
Recall score : 89.09%
F1 score : 84.44%
```

Fig. 8. Metric on Random Forest Classifier

After looking at these results, we found that Logistic Regression surpassed other machine learning models and thus I decided to construct a ROC and AUC curve along with the confusion matrix.

- ROC and AUC curve can be viewed in fig 9.

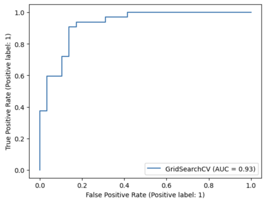


Fig. 9. ROC And AUC Curve

- Confusion matrix on Logistic Regression can be viewed in fig 10.

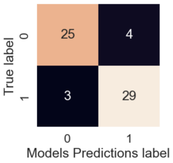


Fig. 10. Confusion Matrix

IV. CONTRIBUTIONS

I have primarily focused on Dataset I and conducted extensive hyper-parameter tuning to enhance the performance of each model. To optimize hyperparameters, RandomSearchCV and GridSearchCV were employed to identify the optimal parameters for each model. The cross-validation technique of 5-fold was employed to generate distinct training and testing sets for data set I. After considering all hyperparameters, it was determined that Logistic Regression had superior performance compared to the other models. The dataset I was subjected to a sequential method. The data was initially examined to ascertain the associations between various independent and dependent variables. Subsequently, the process involves the identification of the most appropriate ensemble of machine learning models that align with my specific requirements. According to the material provided by scikit-learn, it has been determined that the aforementioned models would be very suitable for application on my dataset. The dataset was divided into training and testing sets using an 80-20 split, with 80% of the data allocated to the training set and the remaining 20% allocated to the testing set. A baseline metric was established for all the models, followed by the utilization of hyper-parameter tweaking to identify the optimal hyperparameters for each model. Cross-validation train and test split was employed in order to mitigate reliance on a single split.

A. Additional Work

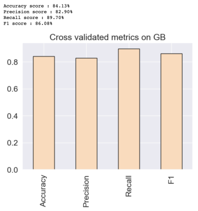
I have initially used only models such as KNN, Logistics regression and Random Forest Classifier to train and test my data set I. For this portfolio report I have made use of Gradient Boost classifier to check its best parameters using Random and GridSearchCV to see if its performance got increased or not. To accurately predict CVDs in the present study, Shapley values were used to create a Gradient Boosting model with an Area Under the Curve of 0.927% for predicting the risk of a heart disease diagnosis [5]. The following parameters combination values were best when I trained and tested my dataset using Gradient Boost Classifier. The best possible parameters for Gradient Boost are listed below:

- Gradient Boost Classifier
  - Input: n_estimators is np.arange(10, 50, 10).
  - Input: max_depth is range(1, 16, 2).
  - Input: learning_rate is np.arange(0.1, 1, 0.1).
  - Input: random_state is np.arange(0, 10, 1).
  - Output: random_state is 0.
  - Output: n_estimators is 20.
  - Output: max_depth is 1.
  - Output: learning_rate is 0.2.

The evaluation metric on the dataset-I can be viewed in fig 11.

V. LESSONS LEARNED

During the course, I acquired knowledge on the process of creating and configuring the environment, including Anaconda and Jupyter Notebook. Additionally, I gained proficiency in utilizing data manipulation libraries such as Pandas and NumPy. I have acquired the knowledge and skills necessary to do training and testing of datasets utilizing the scikit-learn library.



```
Accuracy score : 84.13%
Precision score : 82.90%
Recall score : 89.70%
F1 score : 86.08%
```

Fig. 11. Evaluation metric on Gradient boost

The utilization of evaluation criteria, including Accuracy, Precision, Recall, and F-1 score, has been incorporated in my analysis. A comprehensive comprehension of the confusion matrix and its application in evaluating the performance of a model was pivotal in my comprehension.

VI. TEAM MEMBERS

The team members list is mentioned below:

- Satish Raj — Email: ssakhine@asu.edu
- Jay Ganesh - Email: jganesh2@asu.edu
- Swati Mahapatra - Email: ssmahapa@asu.edu
- Aakash Mood - Email: acmood@asu.edu
- Bornabh Bhuyan - Email: bbhuyan1@asu.edu

VII. ACKNOWLEDGMENT

I would like to extend my appreciation to my team members, namely Swati Mahapatra, Jay Ganesh, Bornabh Bhuyan, Satish Raj, and Aakash Mood, for their unwavering support during the duration of the project. I would also like to express my gratitude to Dr. Yingzhen Yang, who served as the instructor for the course, as well as Utkarsh Nath, the teaching assistant, and Saumya Gangwar, the grader, both affiliated with Arizona State University. Their consistent support and encouragement during the duration of the course are greatly appreciated.

REFERENCES

[1] Joshi, S., Shenoy, P.D., Venugopal, K.R. and Patnaik, L.M. (2010) Data Analysis and Classification of Various Stages of Dementia Adopting Rough Set Theory. International Journal on Information Processing, 4, 86-89

[2] V. Kakulapati, K. Ankith, K. Vaibhav, P.R. Charan Predictive analysis of heart disease using stochas-tic gradient boosting along with recursive feature elimination

[3] Michael, J.A.B. and Gordon, S.L. (2004) Data Mining Techniques for Marketing, Sales, and Customer Relationship Management. 2nd Edition, Wiley Publishing, Inc., Indianapolis.

[4] Brownlee, J. (2023) A gentle introduction to k-fold cross-validation, MachineLearningMastery.com. Available at: https://machinelearningmastery.com/k-fold-cross-validation/ (Accessed: 20 October 2023).

[5] Baghdadi, N.A., Farghaly Abdelaliem, S.M., Malki, A. et al. Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. J Big Data 10, 144 (2023). https://doi.org/10.1186/s40537-023-00817-1