F21DL – DATA MINING AND MACHINE LEARNING

COURSEWORK

TITLE: Heart Disease Prediction & ECG Image Classification of Cardiac Patients

GROUP Dubai_PG 12 MEMBERS:

- **Abhay Krishnan - H00481165**
- **Aghil Subramanian Kizhukkulathil - H00468078**
- **Akul Vinod Adichikkatt - H00481041**
- **Edwin Binu - H00482595**
- **Muhammad Hilal Aslam - H00484782**

TABLE OF CONTENTS

Heart Disease Prediction & ECG Image Classification of Cardiac Patients

Overview

Heart disease is a leading cause of death globally. Understanding predictive factors through data analysis can help in early diagnosis, improving preventive measures and treatment strategies, ultimately saving lives.

The coursework is divided into two parts

1. Predicting heart disease using machine learning techniques. (Tabular Dataset)
2. ECG image classification using Neural Network. (Image Dataset)

Introduction

1. The Aim is to build predictive models using structured health data using various classifiers like Logistic Regression, Decision Trees, Random Forest… to analyze and predict heart disease using attributes such as cholesterol levels, chest pain types, and blood pressure
2. The is Aim is to applying deep learning techniques, such as convolutional neural networks (CNNs), these visual patterns can be automatically analyzed to classify different types of heart abnormalities, such as arrhythmias, ischemia, or myocardial infarction.
    - ECG (Electrocardiogram) images contain distinct patterns that reflect various heart conditions.

Data Collection

Source Link and License Link

Dataset 1 HeartDiseaseClassification.ipynb - Predicting Heart Disease

Link of the Dataset *https://www.kaggle.com/datasets/mexwell/heart-disease-dataset*

Link of License *https://creativecommons.org/licenses/by/4.0/*

Dataset 2 ECGImageClassification.ipynb - ECG Image Classification

Link of the dataset *https://www.kaggle.com/datasets/evilspirit05/ecg-analysis*

Link of License *https://www.mit.edu/~amini/LICENSE.md*

Files and Folders

*   The repository contains 2 folder ('Dataset 1 HeartDiseaseClassification' and 'Dataset 2 ECGImageClassification') and a readme file

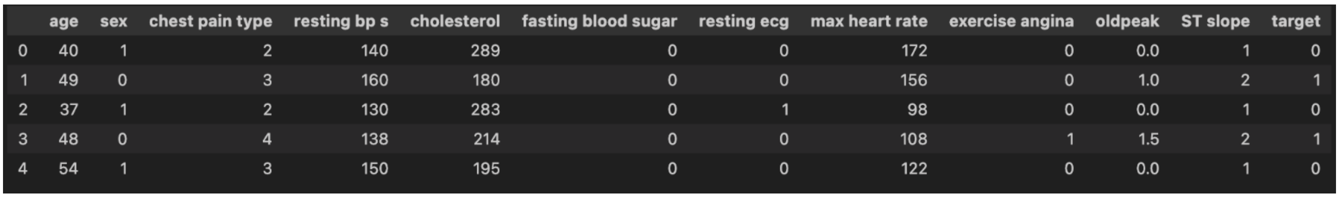
*   'Dataset 1 HeartDiseaseClassification' is the Heart Disease Prediction and contains the 'HeartDiseaseClassification.ipynb' which contains the code for the analysis, 'heart_statlog_cleveland_hungary_final.csv' contains the data for the analysis

*   'Dataset 2 ECGImageClassification' is the ECG Image classificaiton and contains the 'ECGImageClassification.ipynb' which contains the code for the analysis and ECG_DATA folder containing the dataset used for the analysis

Dataset Description and Analysis

Dataset 1 HeartDiseaseClassification.ipynb

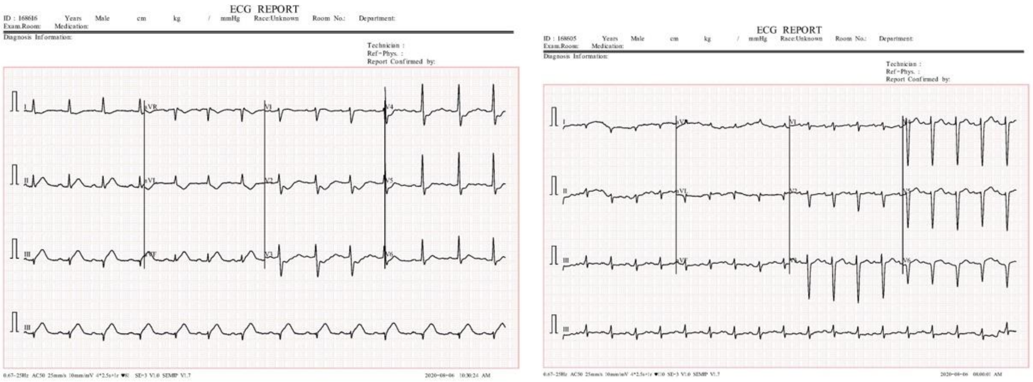
This image shows the overview of the tabular data, which showcases the attributes that we are using for the model: Age, Sex, Chest Pain Type, Resting Blood Pressure, Serum Cholesterol, Fasting Blood Sugar, Resting Electrocardiogram Results, Maximum Heart Rate Achieved, Exercise Induced Angina, Oldpeak (ST Depression), The Slope of Peak Exercise ST Segment, Class (Target).



	age	sex	chest pain type	resting bp s	cholesterol	fasting blood sugar	resting ecg	max heart rate	exercise angina	oldpeak	ST slope	target
0	40	1	2	140	289	0	0	172	0	0.0	1	0
1	49	0	3	160	180	0	0	156	0	1.0	2	1
2	37	1	2	130	283	0	1	98	0	0.0	1	0
3	48	0	4	138	214	0	0	108	1	1.5	2	1
4	54	1	3	150	195	0	0	122	0	0.0	1	0

Dataset 2 ECGImageClassification.ipynb

This image represents the ECG of a MI patient and patient that has abnormal heartbeat.



Analysis

Dataset 1 HeartDiseaseClassification.ipynb :

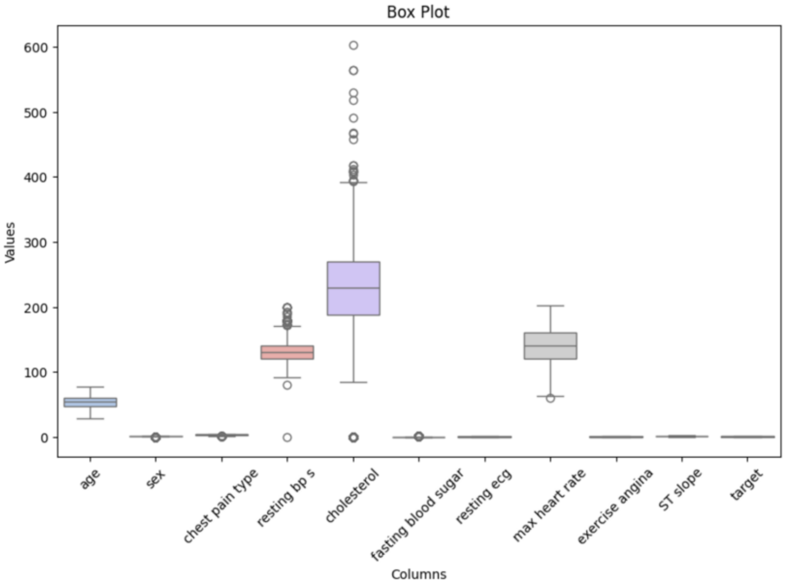
➢ Data Loading and Inspection

The data is loaded using pandas library, and it is proceeded with initial inspection by describing the dataset and visualization of correlation (using seaborn)

Then the data is preprocessed to check the missing values, were none were found

➢ **Data Preprocessing**:

The data is check for outliers using box plot



The data is feature scaled using Standard scalers as there are outliers found.

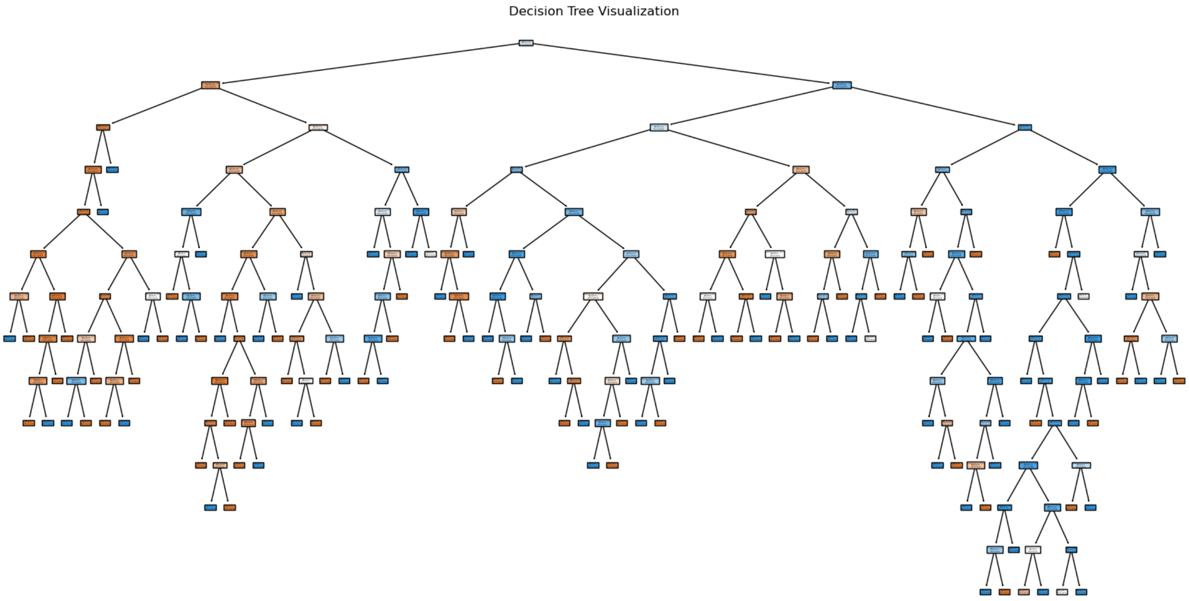
➢ Data Splitting:

The data is then split into a training set and testing set.

➢ Model Development:

Decision tree parameter estimation is used to try out Different splits and Depths to find the best parameter using GridSearchCV Method Grid Search Cross-Validation is method that uses cross validation and Grid Search method to find the best hyper-parameter

- The best Parameter for Max Depth is 13 and Min sample split is 3.



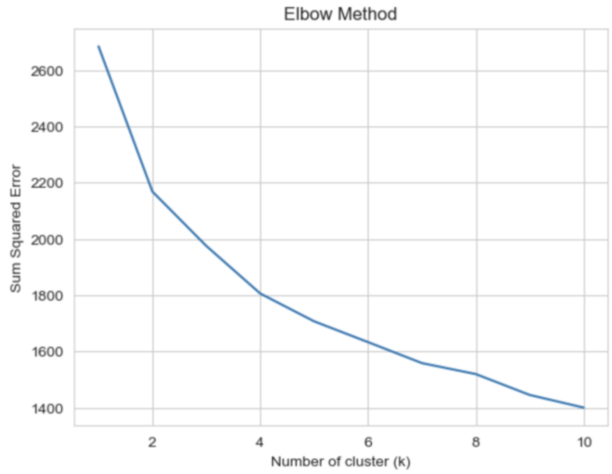
Accuracy Improvement:

The model is then fed a loop to predict heart disease of a patients with relevant attributes using different machine learning algorithm Such as: DecisionTree, RandomForest, Logistic Regression, KNN, GradientBoosting, NaiveBayes, SVM and compare the accuracy between the model for better prediction.

To improve the accuracy added a K-fold algorithm with split 5 and ran the loop with it

Clustering:

KMeans clustering is ran through the data. To find the optimal K, has used Elbow Method. The Elbow method uses within-cluster-sum-of-square (WCSS) vs K value graph. The optimal K value is at the point where the graph forms an elbow



Dataset 2 ECGImageClassification.ipynb:

➢ Data Loading and Inspection

The required libraries including pandas, os, TensorFlow are imported

➢ Data Splitting

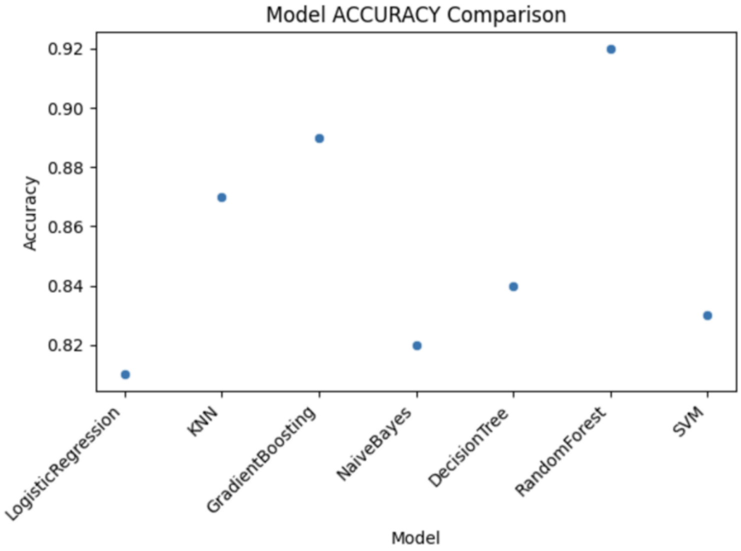
The dataset which is already split in the folder is imported and classified accordingly using os library

➢ Model Development

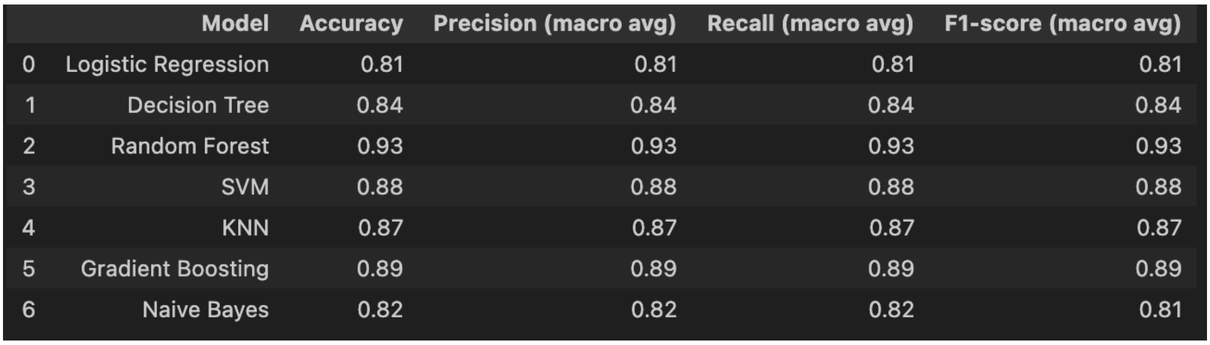
The data then is run through a traintest generation model, where the data is augmented and converted into a greyscale image. The data is trained through a MLP model with a epochs of 25. A CNN model is created using the TensorFlow library to analyze the data. The model is training MLP- 25 epochs ensuring it learns effectively from the data.

Results

Dataset 1 HeartDiseaseClassification.ipynb : From the correlation analysis we found that Chest pain type, exercise angina, ST slope, max heart rate, oldpeak are Strong Predictors, Age, fasting blood sugar, and sex. Are moderate predictor and Cholesterol and resting blood pressure are weak predictor.



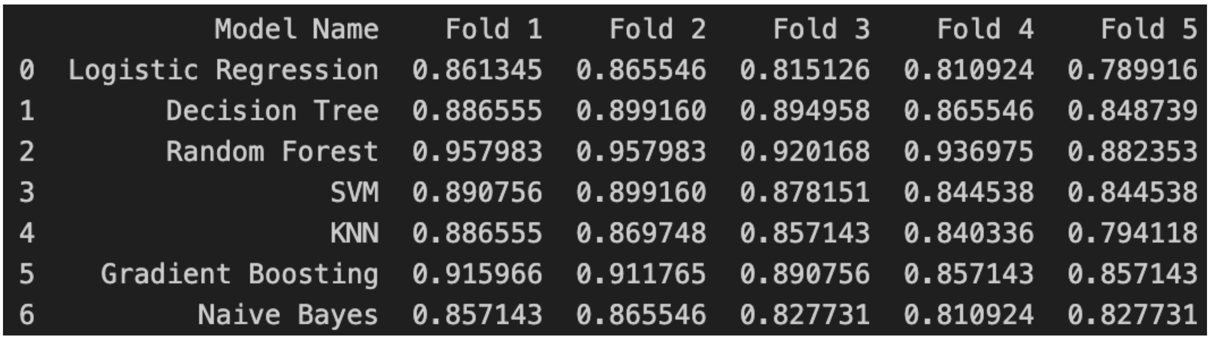
The accuracy of the model is listed in the table below



	Model	Accuracy	Precision (macro avg)	Recall (macro avg)	F1-score (macro avg)
0	Logistic Regression	0.81	0.81	0.81	0.81
1	Decision Tree	0.84	0.84	0.84	0.84
2	Random Forest	0.93	0.93	0.93	0.93
3	SVM	0.88	0.88	0.88	0.88
4	KNN	0.87	0.87	0.87	0.87
5	Gradient Boosting	0.89	0.89	0.89	0.89
6	Naive Bayes	0.82	0.82	0.82	0.81

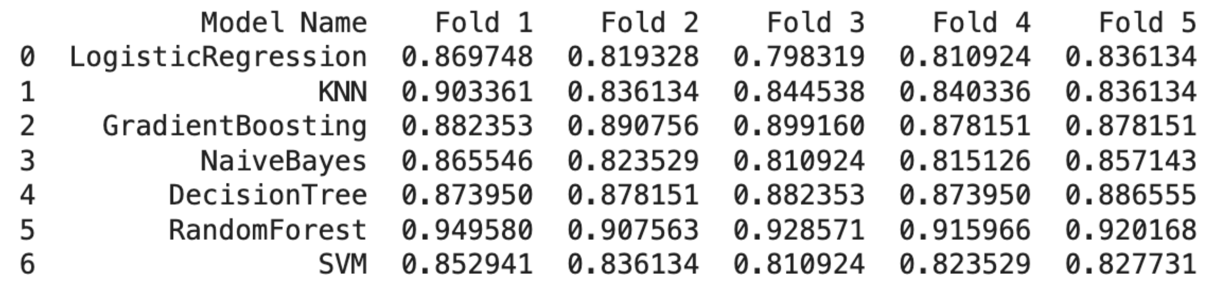
The best model with the best accuracy is Random Forest with an Accuracy of 0.93

After performing K - Fold algorithm on the model there was a significant improvement in the accuracy, further validating the model



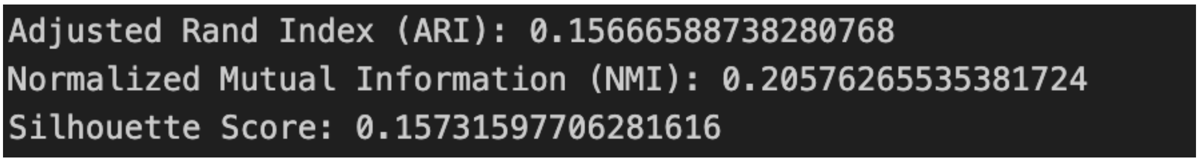
```
            Model Name    Fold 1    Fold 2    Fold 3    Fold 4    Fold 5
0  Logistic Regression  0.861345  0.865546  0.815126  0.810924  0.789916
1        Decision Tree  0.886555  0.899160  0.894958  0.865546  0.848739
2        Random Forest  0.957983  0.957983  0.920168  0.936975  0.882353
3                  SVM  0.890756  0.899160  0.878151  0.844538  0.844538
4                  KNN  0.886555  0.869748  0.857143  0.840336  0.794118
5    Gradient Boosting  0.915966  0.911765  0.890756  0.857143  0.857143
6          Naive Bayes  0.857143  0.865546  0.827731  0.810924  0.827731
```

After applying the K fold algorithm, the best model with the highest accuracy is Random Forest with an Accuracy of 0.95



```
           Model Name    Fold 1    Fold 2    Fold 3    Fold 4    Fold 5
0  LogisticRegression  0.869748  0.819328  0.798319  0.810924  0.836134
1                 KNN  0.903361  0.836134  0.844538  0.840336  0.836134
2    GradientBoosting  0.882353  0.890756  0.899160  0.878151  0.878151
3          NaiveBayes  0.865546  0.823529  0.810924  0.815126  0.857143
4        DecisionTree  0.873950  0.878151  0.882353  0.873950  0.886555
5        RandomForest  0.949580  0.907563  0.928571  0.915966  0.920168
6                 SVM  0.852941  0.836134  0.810924  0.823529  0.827731
```

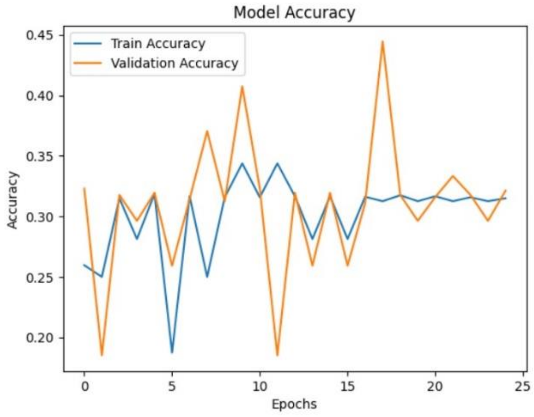
K Means Clustering gave an output of

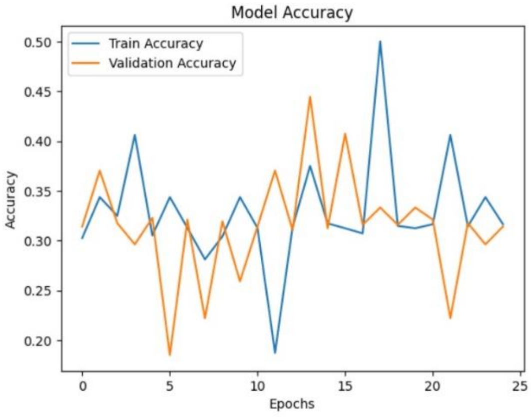


```
Adjusted Rand Index (ARI): 0.15666588738280768
Normalized Mutual Information (NMI): 0.20576265535381724
Silhouette Score: 0.15731597706281616
```

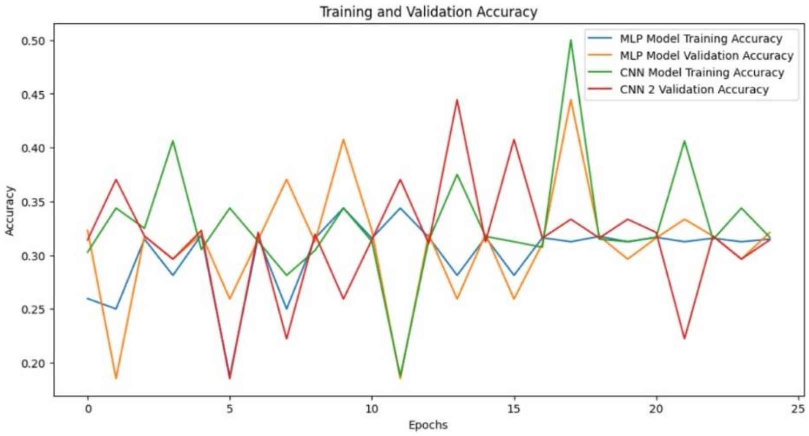
Dataset2: The model MLP with an accuracy of 0.45

The model MLP with an accuracy of 0.50





When comparing the MLP and CNN model, we can see a improvement in the CNN model by 0.5 percentage



Conclusion

Dataset 1 HeartDiseaseClassification.ipynb

Among all models tested, Random Forest consistently had shown the best performance across different metrics (Accuracy, Precision, Recall, and F1-score)

- Without K-Fold validation, the Random Forest achieved an accuracy of **93%**.
- K-Fold cross-validation improved the accuracy for most models, especially Random Forest and Gradient Boosting.
- After applying K-Fold cross-validation, the accuracy of Random Forest increased to **95%**
- The scores (ARI: 0.1567, NMI: 0.2058, and Silhouette: 0.1573) indicate that the clustering structure is weak, suggesting the data may not be naturally separable into distinct clusters or the features may need more preprocessing or feature engineering.

The Random forest is the more reliable and accurate algorithm for this dataset, with a accuracy of 95 %

Dataset 2 ECGImageClassification.ipynb

Both models have high variability in accuracy across epochs, which is due to Insufficient Training data or Noisy Data

Reference

- K Means Algorith - (https://www.geeksforgeeks.org/k-means-clustering-introduction/)
- Elbow Method (https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/)
- GridSearchCV (https://www.geeksforgeeks.org/hyperparameter-tuning-using-gridsearchcv-and-kerasclassifier/)
- Matplotlib (https://matplotlib.org/stable/index.html) seaborn (https://seaborn.pydata.org/)
- Sowmiya, C. and Sumitra, P. (2017). Analytical study of heart disease diagnosis using classification techniques. *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*. doi:https://doi.org/10.1109/itcosp.2017.8303115.

Group Declaration

- Edwin Binu - – Data Preparation and GridSearchCV
- Abhay Krishnan – Algorithm and K Fold
- Aghil Subramanian Kizhukkulathil – Elbow Method and KMeans
- Akul Vinod Adichikkatt – Data Preparation and MLP Algorithm
- Muhammad Hilal Aslam - Graphs and CNN Algorithm