2024-2025

# Predicting Cardiovascular Disease Risk and Identifying High-Risk Patients Using Machine Learning

**Coursework for Data Mining and Machine Learning (F21DL)**

BY :- HARIKA MAHALAKSHMI SRIDHAR (H00464781)
SYED ARIF ALI (H00484788)
BHAGYA PERERA (H00481008)
PRIYANSHU DWIVEDI (H00330801)
RAHUL KUMAR (H00474374)

# Student Declaration of Authorship

HERIOT WATT UNIVERSITY
UK | DUBAI | MALAYSIA

| | |
|---|---|
| **Course code and name:** | **F21DL - Data Mining and Machine Learning** |
| **Type of assessment:** | **Group** |
| **Coursework Title:** | **Machine Learning Portfolio (PG-level 11)** |
| **Student Name:** | **BHAGYA PERERA** |
| **Student ID Number:** | **H00481008** |

**Declaration of authorship. By signing this form:**

- **I declare that the work I have submitted for individual assessment OR the work I have contributed to a group assessment, is entirely my own. I have NOT taken the ideas, writings or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings or inventions of others, or of any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgements section.**

- **I confirm that I have read, understood and followed the University's Regulations on plagiarism as published on the University's website, and that I am aware of the penalties that I will face should I not adhere to the University Regulations.**

- **I confirm that I have read, understood and avoided the different types of plagiarism explained in the University guidance on Academic Integrity and Plagiarism**

**Student Signature *(type your name):*** **BHAGYA PERERA**

**Date:** ***22/11/2024***

# Student Declaration of Authorship

HERIOT WATT UNIVERSITY
UK | DUBAI | MALAYSIA

| | |
|---|---|
| **Course code and name:** | **F21DL - Data Mining and Machine Learning** |
| **Type of assessment:** | **Group** |
| **Coursework Title:** | **Machine Learning Portfolio (PG-level 11)** |
| **Student Name:** | **HARIKA MAHALAKSHMI SRIDHAR** |
| **Student ID Number:** | **H00464781** |

**Declaration of authorship. By signing this form:**

- **I declare that the work I have submitted for individual assessment OR the work I have contributed to a group assessment, is entirely my own. I have NOT taken the ideas, writings or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings or inventions of others, or of any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgements section.**

- **I confirm that I have read, understood and followed the University's Regulations on plagiarism as published on the University's website, and that I am aware of the penalties that I will face should I not adhere to the University Regulations.**

- **I confirm that I have read, understood and avoided the different types of plagiarism explained in the University guidance on Academic Integrity and Plagiarism**

**Student Signature *(type your name):*** **HARIKA MAHALAKSHMI SRIDHAR**

**Date:** ***22/11/2024***

# Student Declaration of Authorship

HERIOT WATT UNIVERSITY
UK | DUBAI | MALAYSIA

| | |
|---|---|
| **Course code and name:** | **F21DL - Data Mining and Machine Learning** |
| **Type of assessment:** | **Group** |
| **Coursework Title:** | **Machine Learning Portfolio (PG-level 11)** |
| **Student Name:** | **PRIYANSHU DWIVEDI** |
| **Student ID Number:** | **H00330801** |

**Declaration of authorship. By signing this form:**

- **I declare that the work I have submitted for individual assessment OR the work I have contributed to a group assessment, is entirely my own. I have NOT taken the ideas, writings or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings or inventions of others, or of any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgements section.**

- **I confirm that I have read, understood and followed the University's Regulations on plagiarism as published on the University's website, and that I am aware of the penalties that I will face should I not adhere to the University Regulations.**

- **I confirm that I have read, understood and avoided the different types of plagiarism explained in the University guidance on Academic Integrity and Plagiarism**

**Student Signature *(type your name):*** **PRIYANSHU DWIVEDI**

**Date:** ***22/11/2024***

# Student Declaration of Authorship

HERIOT WATT UNIVERSITY
UK | DUBAI | MALAYSIA

| | |
|---|---|
| **Course code and name:** | **F21DL - Data Mining and Machine Learning** |
| **Type of assessment:** | **Group** |
| **Coursework Title:** | **Machine Learning Portfolio (PG-level 11)** |
| **Student Name:** | **RAHUL KUMAR** |
| **Student ID Number:** | **H00474374** |

**Declaration of authorship. By signing this form:**

- **I declare that the work I have submitted for individual assessment OR the work I have contributed to a group assessment, is entirely my own. I have NOT taken the ideas, writings or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings or inventions of others, or of any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgements section.**

- **I confirm that I have read, understood and followed the University's Regulations on plagiarism as published on the University's website, and that I am aware of the penalties that I will face should I not adhere to the University Regulations.**

- **I confirm that I have read, understood and avoided the different types of plagiarism explained in the University guidance on Academic Integrity and Plagiarism**

**Student Signature *(type your name):*** **RAHUL KUMAR DWIVEDI**

**Date:** ***22/11/2024***

# Student Declaration of Authorship

HERIOT WATT UNIVERSITY
UK | DUBAI | MALAYSIA

| | |
|---|---|
| **Course code and name:** | **F21DL - Data Mining and Machine Learning** |
| **Type of assessment:** | **Group** |
| **Coursework Title:** | **Machine Learning Portfolio (PG-level 11)** |
| **Student Name:** | **SYED ARIF ALI** |
| **Student ID Number:** | **H00484788** |

**Declaration of authorship. By signing this form:**

- **I declare that the work I have submitted for individual assessment OR the work I have contributed to a group assessment, is entirely my own. I have NOT taken the ideas, writings or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings or inventions of others, or of any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgements section.**

- **I confirm that I have read, understood and followed the University's Regulations on plagiarism as published on the University's website, and that I am aware of the penalties that I will face should I not adhere to the University Regulations.**

- **I confirm that I have read, understood and avoided the different types of plagiarism explained in the University guidance on Academic Integrity and Plagiarism**

**Student Signature *(type your name):* *SYED ARIF ALI***

**Date: *22/11/2024***

# Table of Contents

**Introduction** ..... 2

**Dataset Description, Analysis, and Experimental Setup** ..... 2

**Framingham Heart Disease Dataset:** ..... 2

**UCI Heart Disease Dataset:** ..... 4

**Cardiomegaly Image Dataset:** ..... 6

**Results of Framingham Heart Disease Dataset:** ..... 7

**Results of UCI Heart Disease Dataset:** ..... 7

**Results of Cardiomegaly Image Dataset:** ..... 8

# Introduction

The research aims to assist cardiovascular health screening by machine learning algorithms to assess an individual's likelihood of developing heart disease and enlarged heart conditions. By integrating patient medical records and diagnostic imaging data, the study seeks to create predictive models that can identify potential cardiac risks at an early stage.

The three datasets we selected for analysis are:

1) Framingham Heart Disease Dataset

2) UCI Heart Disease Dataset

3) Cardiomegaly Image Dataset

## Dataset Description, Analysis, and Experimental Setup

### Framingham Heart Disease Dataset:

The Framingham Heart Disease dataset is frequently used to study heart disease.

1. **Details:**
    a. It consists of about 4240 patient records with about 16 different features.
    b. The dataset target is to predict the chance of a patient suffering from heart disease in the next 10 years.
    c. c. There is an imbalance in the number of healthy vs. sick patients. Our analysis corrects this with the help of up-sampling, SMOTE, and testing results on both using XGB.
2. **Exploratory Data Analysis (EDA):**
    a. The dataset includes several features such as age, weight, glucose, and the number of cigarettes smoked per day.
    b. The presence of missing values (the blood sugar data had almost 400 missing values) and the class imbalance in the target variable (TenYearCHD) required us to use a technique called SMOTE to balance the dataset, as most records indicate that patients are not at risk.
    c. The following features also had missing values but in smaller amounts: education, cigsPerDay, BPMeds, and totChol which has less than 150 missing values each.
3. **Experimental Setup**
    a. **Data Cleaning:** Missing values in important features like glucose, education, blood pressure medication and total cholesterol were filled in. we used median value for numerical features like glucose and the most common value for categorical features such as educational level.
    b. **Feature Scaling:** We scaled the numerical features like age, weight, and systolic blood pressure by applying MinMaxScaler. Features are scaled to make sure that features with very different ranges do not affect the model too much.
    c. **Class Imbalance Handling:** Our target variable (TenYearCHD) was highly imbalanced as there were many more healthy patients than sick ones, we used a technique called SMOTE to create more examples of sick patients. This has shown to help model learn to better detect minority class instances.
    d. **Feature Selection:** To simplify the model and focus on the most relevant features, we employed a feature selection technique. This process helped us identify the key features that are most relevant to heart disease risk. The two methods that were used are:
        i. **SelectKBest:** In this method, the features are ranked based on their statistical tests, such as ANOVA F-values.
        ii. **Random Forest:** Random Forest was used to compute feature importance, highlighting the most influential features.
    e. **Modelling:** Machine learning models used to predict the 10-year risk of heart disease
        i. **Logistic Regression:** A simple model to start with, a baseline model for binary classification.

ii. **Support Vector Machine (SVM):** This model is particularly effective in handling complex datasets with many features.

iii. **Random Forest:** This model helps combine multiple decision trees for robust predictions.

iv. **K-Nearest Neighbours (KNN): This** model classifies data based on similarity used for understanding baseline performance.

v. **Multilayer Perceptron (MLP):** A neural network which can learn complex patterns in the data.

vi. **Naïve Bayes:** A fast and efficient model for classification.

f. **Clustering:**

The dataset was analysed using K-means clustering to explore natural groupings within the data. Goal of clustering was to identify meanigful clusters among patiernts based on their health-related features, providing insights into patterns that may not be apparent from individual variables.

i. **Optimal Number of Clusters:** We used Elbow Method and Silhouette Analysis to determine the optimal number of clusters (k), both methods gave us an optimal value of 3, balancing within-cluster vairance and separation between clusters.

ii. **Visualization:** We used Principal Component Analysis (PCA) reduced the data to two dimensions for visual representation of clusters and a heatmap of standardized cluster centers revealed key characteristics differentiating the three clusters.

**Results of clustering:** The PCA plot shows clear separation between clusters, while the heatmap highlights differences in features like age, glucose levels, and BMI, suggesting varying risk profiles among the clusters.

g. **Evaluation Metrics:** To assess the performance of the models, the following metrics were used:

i. **Accuracy:** Measures the overall proportion of correct predictions.

ii. **Precision, Recall, and F1-Score:** These metrics are useful for imbalance datasets, providing insights into model's ability to correctly identify positive cases and minimize false positives and negatives.

iii. **AUC-ROC:** This metric evaluated the model's ability to distinguish between classes.

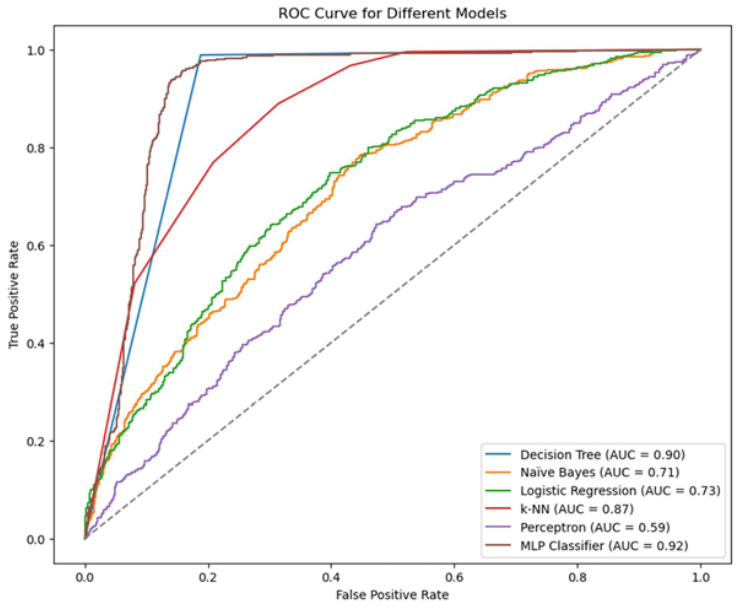


*Figure 1: MLP and Decision Tree achieve the highest AUC scores, which shows their strong discriminatory power. Simpler models performed less effectively as indicated by their lower AUC scores (Naïve Bayes and Logistic Regression)*

# UCI Heart Disease Dataset:

The UCI Heart Disease dataset is a widely used resource for studying heart disease classification and prediction. It enables researchers to analyze medical and demographic attributes to determine the likelihood of heart disease.

1. **Details**: The chosen dataset is the UCI Heart Disease Data obtained from Kaggle.
    a. This dataset contains 16 columns and 920 rows.
    b. The dataset is used for a classification task to predict the presence or absence of heart disease.
    c. Target Classes: 0 – Absence of Heart disease, 1: Presence of heart disease.
    d. The features can be categorized as demographic (age, sex, etc.) and clinical features(Chest Pain, chol, thal, etc.).
    e. **Target (num):** The target variable indicating heart disease presence:
    0: No heart disease
    >0: Presence of a heart disease with (1, 2, 3, or 4 depending on severity).

**Objective:**

The primary task is to predict whether a person has heart disease based on the given attributes. This involves analyzing both demographic and medical measurements to develop an accurate classification model.

2. **Exploratory Data Analysis (EDA):**
    a. **DataFrame Construction:** A pandas DataFrame is used to hold the dataset's rows and columns, facilitating experimentation and classification using various machine learning models.
    b. **Dropping Irrelevant Columns:** The id and dataset columns, which hold no relevance to classification tasks, are removed.
    c. **Encoding Categorical Features:** Several machine learning algorithms require numerical inputs. Therefore, Label Encoding is applied to categorical columns (cp, restecg, slope, thal, ca) to convert string-based categories into numeric values.
    d. **Class Imbalance:** The target column (num) shows a critical class imbalance, with Class 4 being underrepresented compared to Class 0. To mitigate this, Classes 1 through 4 are merged into a single category, transforming the target column into a binary classification feature (0 = No heart disease, 1 = Presence of heart disease).
    e. **Handling Erroneous and Missing Values**:
    **Null Values:** Columns like trestbps (resting blood pressure) and chol (cholesterol) contain values of 0, which are medically unfeasible. These are marked as null for further treatment.
    **Missing Data:** Features such as ca, thal, oldpeak, and chol have significant missing values. Imputation methods are applied to handle these to avoid severe data loss.
    f. **Data Imbalance in Features**: The sex column shows a disparity in the number of male and female samples, which could affect model performance. This imbalance is noted for consideration in further analysis.
    g. **Correlation Matrix Analysis:**

| Feature | Correlation with Heart Disease (num) | Key Insights |
|---|---|---|
| ca (major vessels) | 0.52 | Strongest positive correlation with heart disease |
| oldpeak (ST depression) | 0.44 | Positively correlated with heart disease |
| thalach (max heart rate) | -0.37 | Negatively correlated; lower values indicate risk |
| age | 0.34 | Positive correlation with disease prevalence |
| exang (exercise angina) | 0.34 | Positively associated with disease presence |

- Maximum heart rate (thalach) decreases with age, a general physiological trend.
- ST depression and ca stand out as the most critical features for predicting heart disease.

3. **Experimental Setup:**
    a. **Missing Value Analysis:**
        i. **Identification**: Missing values were identified using visualizations, with features like ca, showing significant missing data.
        ii. **Imputation Strategy**: Implemented mean imputation strategy through:

- **SimpleImputer:** Used for features with less than 10% missing values, employing mean imputation.
- **KNNImputer**: Applied for features with 10% to 20%. Missing values, leveraging nearest neighbors for imputation.

iii. This strategy maintained the original data distribution and minimized data loss.

**b. Feature Scaling:**

i. **StandardScaler:** Applied to normalize feature values, ensuring all variables contribute equally to machine learning algorithms.

**c. Clustering:**

**K-Means Algorithm:**

i. **Cluster Optimization:** The optimal number of clusters was determined using the Elbow Method (k=4) and Silhoutte Scores (peak at k=4).

ii. **Cluster Analysis:** Clusters revealed the following key patterns:

- Features like age and thalach exhibited the strongest variations across clusters,
- Moderate variations in features such as cp and trestbps.
- Clear visual separation of clusters in PCA space.
- Cluster validity was supported by Silhouette Scores.

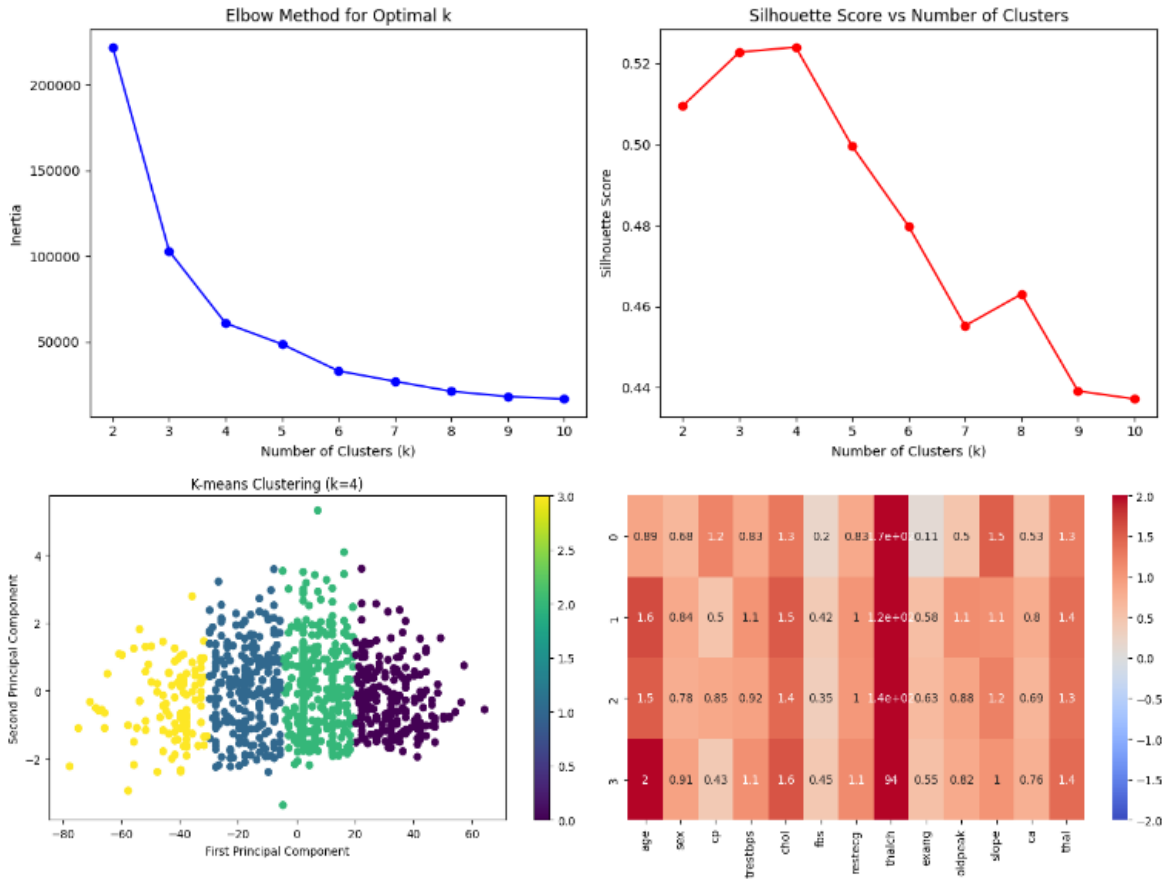


*Figure 2: This figure shows the elbow method, silhouette scores, K-means clustering (k=4) in PCA-reduced space, and a heatmap of feature correlations.*

## Cardiomegaly Image Dataset:

Cardiomegaly refers to an enlargement of the heart, which often leads to heart failures and problems such as high blood pressure, heart valve disease, and cardiomyopathy. This dataset contains chest x-ray images and is frequently used for studying cardiomegaly disease.

1. **Details:**
    a. **Dataset size**: It consists of 5552 patient records. We have split data set in to ratio of 80:20 for train and test respectively. Train set consists of 2219 cardiomegaly images and 2219 non-cardiomegaly images.
    b. **Objective**: Our aim is to build a CNN model which can predict cardiomegaly and non-cardiomegaly
    c. **Class Balance:** The dataset has equal number of cardiomegaly and non-cardiomegaly images. Hence, dataset is balanced.
2. **Experimental Setup:**

This data set consists of Gray-scaled images and image size is 128*128. Therefore, we have only one channel.

i. **Preprocessing:**
    a. **Gray-Scaled Images**: Since the images are Gray scaled, directly used this image data set with size 128*128*1.
    b. **Normalization**: Normalized pixel values by dividing by 255 to get the values between 0-1 since RGB range is 0-255.
    c. **Label Encoding**: Label values are encoded to 0,1 with cardiomegaly as 1 and non-cardiomegaly as 0

ii. **Create CNN model:**
    a. **Convolutional layers**: CNN model consists of 4 convolutional layers. 3*3 kernel used.
    b. Feature map sizes: Feature maps of sizes 32,64,128,128 respectively for each of the 4 CNN layers and the activation function is Relu
    c. Pooling Layers -2*2 pooling layers with stride 1
    d. Fully connected Layer with 128 feature maps.
    e. For the Output layer the activation function is Sigmoid since range is (0,1) good for binary classification
    f. Dropout layers are used to avoid overfitting

iii. **Evaluation Metrics:** To assess the performance of the models, the following metrics were used:
    a. **Accuracy:** Measures the overall proportion of correct predictions.
    b. **Precision, Recall, and F1-Score, Confusion metrics:** Providing insights into model's ability to correctly identify positive cases and minimize false positives and negatives.
    c. **Adam optimizer** used with different learning rates (0.1,0.01,0.001,0.0001)
    d. **Different batch sizes** used to train the model. (32,63,128,256,512)
    e. 100 epochs are used however **early stopping** is used to avoid the overfitting. It checks for the minimum value loss. If there is no optimal value loss, during the next 10 epochs the model stops and reassign the best weights.
    f. The best weights are used to evaluate the model using test data
    g. **Cross fold validation** is used to check the vary of accuracies and average measures have been taken as the final measures.

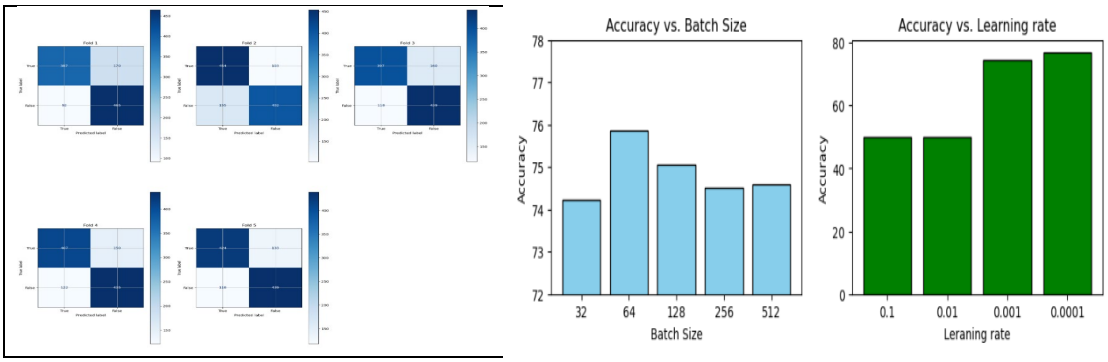


*Figure 3: Confusion Matrices and Model Accuracy Analysis*

# Overall Results, Discussion, and Conclusion

## Results of Framingham Heart Disease Dataset:

i. **Model Performance:** The following table summarizes the performance of all tested machine learning models on the Framingham Dataset:

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Decision Tree | 89.85% | 83.48% | 98.87% | 90.52% |
| MLP Classifier | 89.24% | 83.31% | 97.67% | 89.90% |
| K-NN | 78.54% | 73.12% | 88.94% | 80.26% |
| Logistic Regression | 66.49% | 65.38% | 67.30% | 66.33% |
| Naïve Bayes | 63.16% | 67.06% | 48.87% | 56.53% |
| Perceptron | 56.34% | 57.25% | 43.29% | 49.30% |

ii. **Conclusion:** From the results we can conclude that:

a. **Decision Tree:** Demonstrated the highest performance in terms of F1-Score and recall, making it particularly adept at identifying true positive cases.

b. **MLP Classifier:** Model exhibited comparable performance with a slightly lower F1-score but maintained robust accuracy and recall.

c. **Logistic Regression & Naïve Bayes:** Even though the dataset's imbalance was mitigated using SMOTE, simpler models like Logistic Regression and Naïve Bayes still showed lower recall and F1-scores. This indicates that their linear assumptions and probabilistic nature might not effectively capture the non-linear and complex patterns in the data.

d. **Perceptron:** Our least performing model due to its linear approach, which struggled with the dataset's complexity.

## Results of UCI Heart Disease Dataset:

i. **Model Performance:**

a. Baseline Training and Evaluation:

i. K-Nearest Neighbors (KNN):

1. Achieved the highest accuracy (0.819) and balanced performance
2. Works well because similar patient profiles tend to have similar diagnoses
3. Used 9 neighbors with uniform weights

ii. Logistic Regression:

1. Achieved accuracy of 0.779
2. Particularly effective due to the linear relationship between some medical indicators and heart disease

iii. Decision Tree:

1. Accuracy of 0.765
2. Max depth of 5 prevented overfitting
3. Useful for capturing non-linear relationships in medical data

iv. Naive Bayes:

1. Accuracy of 0.779
2. Performed well on the dataset

v. Perceptron:

1. Accuracy of 0.745
2. Simplest neural network implementation
3. Used balanced class weights to handle imbalance
4. Early stopping helped prevent overfitting

vi. Neural Network (MLP):

1. Accuracy of 0.765
2. Two-layer architecture (100, 50 neurons), Good at capturing complex patterns

## Results of Cardiomegaly Image Dataset:

i. **Results:** Train set is split into validation set with the ratio of 80:20 (in %) . Test set is strictly used only for testing. Training is performed with different batch sizes, learning rates and cross fold validation.

1. **Batch sizes**
   a. According to the evaluation the best batch size was 64.
   b. Test Accuracy of this model is approximately 75.82% .
   c. Precision, recall, f1_score are 0.7285,0.8240 and 0.7733 respectively.
   d. Therefore, out of the cardiomegaly images, 82.40% are predicted correctly. Out of the images which are predicted as cardiomegaly 72.85% were cardiomegaly.
2. **Learning rates**
   1. According to the evaluation the best learning rate is 0.0001. Accuracy is 77.02% which is high when compared to the other learning rates
   2.
3. **Cross fold validation**
   a. This technique is used to evaluate the performance of a model by splitting the dataset into 5 equal size folds.
   b. The model is then trained and evaluated 5 times, each time using a different fold for testing and the remaining folds for training.
   c. For each fold:
      i. Train the model using data from all folds except the current fold. (other 4 folds).
      ii. The current fold is used as the validation set.
      iii. For each fold using the test dataset the model is evaluated
   d. Calculate the average performance: The result is the average of the performance metrics (like accuracy, precision, recall, etc.) from each fold.

*Model with different parameters*

| Optimizer | Learning Rate | batch Size | Cross Fold | Accuracy | Loss | Precision | Recall | F1-Score |
|---|---|---|---|---|---|---|---|---|
| Adam | 0.001 | 32 | NO | 0.742369 | 0.54336 | 0.7027 | 0.8402 | 0.7653 |
| Adam | 0.001 | 64 | NO | 0.75852 | 0.54209 | 0.72857 | 0.82405 | 0.77337 |
| Adam | 0.001 | 128 | NO | 0.75044 | 0.54172 | 0.7066 | 0.85637 | 0.77435 |
| Adam | 0.001 | 256 | NO | 0.745062 | 0.539195 | 0.709035 | 0.83123 | 0.765289 |
| Adam | 0.001 | 512 | NO | 0.7459604 | 0.5540325 | 0.72682 | 0.78815 | 0.7562446 |
| Adam | 0.1 | 32 | NO | 0.5 | 0.7189 | 0.5 | 1.0 | 0.6667 |
| Adam | 0.01 | 32 | NO | 0.5 | 0.693173 | 0.5 | 1.0 | 0.6667 |
| Adam | 0.0001 | 32 | NO | 0.77019 | 0.54761 | 0.74876 | 0.8132 | 0.7796 |
| Adam | 0.0001 | 128 | YES | 0.7628 | 0.543142 | 0.754456 | 0.78276 | 0.767255 |

*Figure 4: Model performances with different parameters.*

ii. **Conclusion**

Data preprocessing techniques have followed to normalize the data and encode the labels to 0 and 1. Our CNN model consists of 4 convolutional layers followed by pooling layers, fully connected layer, dropout layer and output layer. Sigmoid function used as the activation function since this is a binary classification. Model evaluation done under the different batch sizes, learning rates and using cross fold validation. Learning rate 0.0001 gives the best accuracy with Adam optimizer of 77.02%