



**Faculty of Engineering & Technology**  
**Electrical & Computer Engineering Department**

**Machine Learning and Data Science - ENCS5341**

**Assignment #3**

---

**Prepared by:**

Ahmed Zubaidia                      1200105

Mohammad Makhamreh              1200227

**Instructor:** Dr. Yazan Abu Farha

**Section:** 2

**Date:** 24/1/2024

## Abstract

This project aims to apply and evaluate machine learning models for predicting the presence of heart disease based on clinical features. The chosen models for this task include k-Nearest Neighbors (kNN), Logistic Regression, and Random Forest. The evaluation metrics used for assessing model performance are accuracy, precision, recall, F1-score, and the confusion matrix.

## Contents

<b>Dataset .....</b>	<b>1</b>
<b>Experiments and Results.....</b>	<b>3</b>
Baseline Model: k-Nearest Neighbors .....	3
The Proposed ML Models .....	4
Logistic Regression:.....	4
Random Forest: .....	5
<b>Performance analysis.....</b>	<b>6</b>
Error Analysis and Patterns: .....	6
<b>Conclusion .....</b>	<b>7</b>

## Table of Figures:

Figure 1: Descriptive Statistics value used .....	1
Figure 2: histograms for each feature. ....	2
Figure 3: KNN model, k=1, k=2 accuracy.....	3
Figure 4: validation test accuracy for logistic regression. ....	4
Figure 5: confusion matrix for logistic regression and precision and recall.....	4
Figure 6: validation test accuracy for random forest .....	5

## Dataset

The dataset is based on data from 303 patients and is drawn from the “processed.cleveland.data” file at the UCI Machine Learning Repository. It contains fourteen clinical features like age, sex, kinds of chest pain, resting blood pressure, cholesterol, fasting blood glucose, resting electrocardiographic outcomes, etc. The goal is to use these features for binary classification of heart disease [1].

Exploratory Data Analysis (EDA)

### Descriptive Statistics:

- The dataset comprises 303 patient records. The mean age is approximately 54.44 years, with a standard deviation of 9.04 years, indicating a wide age range from 29 to 77 years.
- The target variable, indicating heart disease presence, shows that 45.87% of the patients have heart disease.

### Visualizations:

- Histograms of distribution for each feature were plotted. These plots showed range and frequency of values for each clinical feature.
- the visualizations help identify patterns like The distribution of cholesterol and The range of patients 'maximum heart rates.

	Age (years)	...	Heart Disease Presence (1: Present, 0: Absent)
count	303.000000	...	303.000000
mean	54.438944	...	0.458746
std	9.038662	...	0.499120
min	29.000000	...	0.000000
25%	48.000000	...	0.000000
50%	56.000000	...	0.000000
75%	61.000000	...	1.000000
max	77.000000	...	1.000000

Figure 1: Descriptive Statistics value used

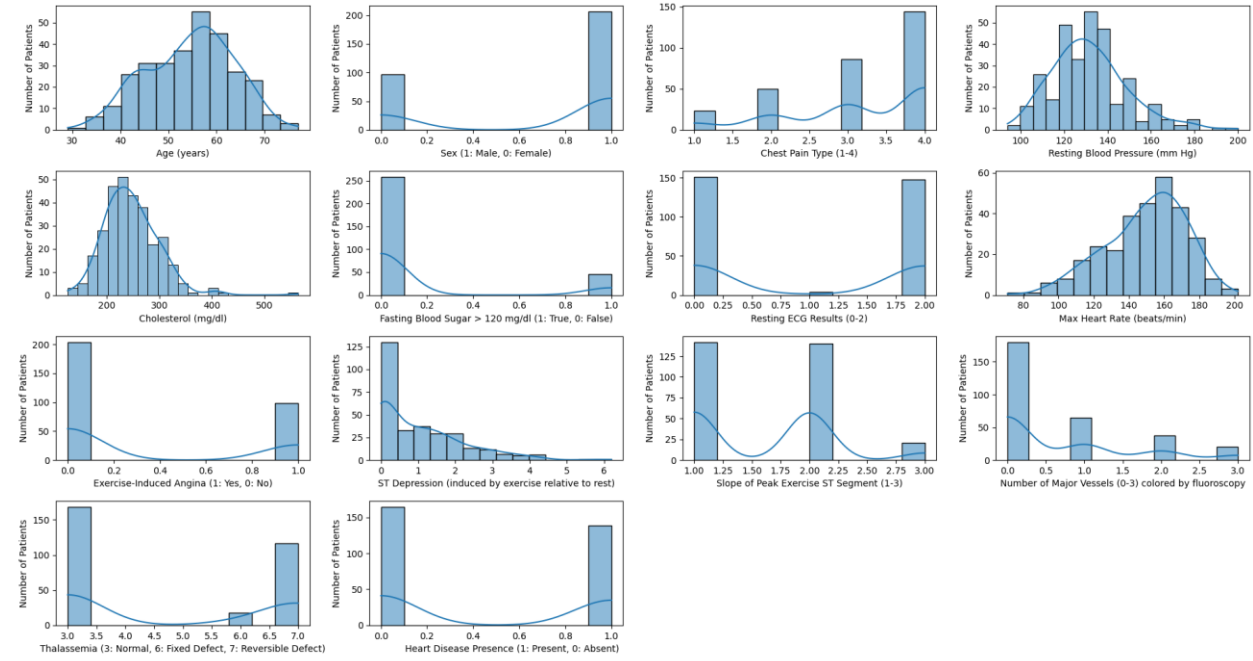


Figure 2: histograms for each feature.

Histograms revealed that most patients were older, and that heart disease risk was greater with age. More men (1) than women (0) were shown, indicating differences in heart disease rates by gender. High blood pressure and cholesterol were regular risk factors for heart issues. Strangely enough, high blood sugar was less frequent, suggesting other things might be a little more important in this group. Heart rate variations during stress and patterns of chest pain demonstrated the complexity of heart disease symptoms. A few patients had severe artery blockage or genetic conditions such as thalassemia. These insights highlight some of the important features which are usually common in heart disease.

## Experiments and Results

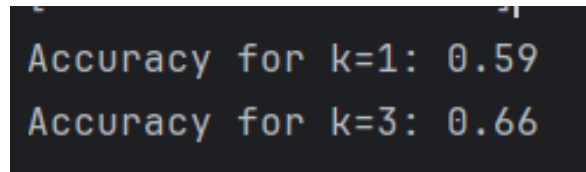
### Baseline Model: k-Nearest Neighbors

#### Implementation and Results

The k-Nearest Neighbors (kNN) algorithm was applied as the baseline model with two variations:  $k = 1$ ,  $k = 3$ . The accuracy of the model was evaluated.

$k = 1$  Accuracy: 59%. This shows how the model performed when classifying the heart disease status according to the closest data point in the training set.

$k = 3$  Accuracy: 66%. This increase in accuracy indicates that a larger neighborhood where the model makes its decision-making has better predictions.



```
Accuracy for k=1: 0.59
Accuracy for k=3: 0.66
```

*Figure 3: KNN model,  $k=1$ ,  $k=2$  accuracy*

## The Proposed ML Models

### Model Selection and Hyperparameter Tuning.

Two additional models were selected for evaluation: Logistic Regression & Random Forest. These models were selected because they were strong and effective in handling classification tasks. Hyperparameter tuning was performed for each model to optimize their performance.

#### Logistic Regression:

- Hyperparameter Tuning: The regularization strength (**C**) was tuned with values [0.01, 0.1, 1, 10, 100].
- Outcome: best performance was obtained with **C = 0.01**, for an validation accuracy score of 83.46% and 83.60 for testing. This indicates that higher regularization (lower C value) prevents overfitting and improves the generalization capability of the model.

#### Confusion Matrix for Logistic Regression:

[[27, 2], [8, 24]], indicating 27 true negatives, 24 true positives, 2 false positives, and 8 false negatives.

#### Classification Report:

Precision for Class 0 (No Heart Disease): 77%

Precision for Class 1 (heart disease) 92%

Recall for Class 0: 93%

Recall for Class 1: 75%

```
Logistic Regression Best Parameters: {'logisticregression__C': 0.01}
Logistic Regression - Validation Score: 0.8346938775510203 Test Accuracy: 0.8360655737704918
```

Figure 4: validation test accuracy for logistic regression.

Confusion Matrix for Logistic Regression:				
[[27 2]				
[ 8 24]]				
Classification Report for Logistic Regression:				
	precision	recall	f1-score	support
0	0.77	0.93	0.84	29
1	0.92	0.75	0.83	32

Figure 5: confusion matrix for logistic regression and precision and recall.



### Random Forest:

- Hyperparameter Tuning: The number of trees (**n\_estimators**) varied among [10, 50, 100, 200].
- Outcome: The optimal number of trees was 50 with (81% validation accuracy) and 88.52 for testing accuracy. This performance improvement is attributed to the ensemble approach of Random Forest in which several decision trees make a more stable and accurate prediction.
- **Confusion Matrix for Random Forest:**
  - [[27, 2], [5, 27]], indicating 27 true negatives, 27 true positives, 2 false positives, and 5 false negatives.
- **Classification Report:**
  - Precision for Class 0: 84%
  - Precision for Class 1: 93%
  - Recall for Class 0: 93%
  - Recall for Class 1: 84%

```
Confusion Matrix for Random Forest:
[[27  2]
 [ 5 27]]
Classification Report for Random Forest:
              precision    recall  f1-score   support

     0       0.84         0.93         0.89         29
     1       0.93         0.84         0.89         32
```

```
Random Forest Best Parameters: {'n_estimators': 50}
Random Forest - Validation Score: 0.8223639455782312 Test Accuracy: 0.8852459016393442
```

Figure 6: validation test accuracy for random forest

## Performance analysis

In this project, three distinct machine learning models were used for the prediction of heart disease presence:

K-Nearest Neighbors (kNN), Logistic Regression and Random Forest. The kNN model as a baseline was evaluated in 2 variations ( $k = 1$  and  $k = 3$ ). Accuracy 59% and 66% respectively were noticed. This accuracy improvement from  $k = 1$  to  $k = 3$  indicates that a bigger neighborhood offers a contextual basis for prediction, which is essential in medical datasets where individual features can be misleading in isolation.

Logistic Regression model was set with regularization strength (C) of 0.01, which achieving optimal performance. In validation set with, 83.46% accuracy was recorded and 83.60% in testing set. A strength of this model was the 92% accuracy for correctly identifying patients with heart disease. However, a recall rate of 75% was also lower, hinting that the model is more conservative in predicting heart disease, potentially resulting in more false negatives involving cases of heart disease not being identified.

The Random Forest model with 50 trees predicted heart disease with an accuracy of 88.52%. This indicates its strong capability in correctly identifying both the precision and recall of the condition. The model performed balancedly with an accuracy of 84% for no heart disease and 93% for heart disease prediction. It had similar recall rates for Class 0 and Class 1, respectively. This demonstrates that the model can identify the right patients with heart disease and avoid false alarms. In general, only 7 of 61 predictions were incorrect and the Random Forest model proved to be a reliable tool for this medical diagnostic task.

## Error Analysis and Patterns:

In the k-Nearest Neighbors (kNN) model errors was higher when neighboring data points had mixed classifications, especially in ambiguous cases. When the model considered more neighbors ( $k = 1$  to  $k = 3$ ), accuracy improved. However, this improvement was not without limits as the kNN model has important constraints in complex classification scenarios.

Logistic Regression model was typically conservative and more heart disease detections were missed frequently when patients' symptoms were less pronounced or atypical. This pattern indicates the need for more detailed feature interaction analysis and symptom representation in the model.

Errors in the Random Forest model were less frequent but mostly occurred when the ensemble of decision trees had conflicting information from the different features. This highlights the difficulty of diagnosing heart disease with such disparate and contradictory clinical indicators.

## Conclusion

This project examining k-Nearest Neighbors, Logistic Regression and Random Forest models of heart disease prediction revealed some important features of machine learning in medical diagnostics. The kNN model highlighted the need for contextual analysis, Logistic Regression highlighted the difficulty in detecting subtle symptoms and Random Forest demonstrated high general accuracy and the difficulty in interpreting diverse clinical data. In summary, this project highlights the model selection and tuning in healthcare applications and gives an eye for the future advancements in machine learning for medical diagnostics.

## References

- [1] "1," [Online]. Available: <https://archive.ics.uci.edu/dataset/45/heart+disease>.