Department of

Electrical & Electronics Engineering

Abdullah Gül University

Lab 4 – Random Forest

BIOMEDICAL SYSTEM DESIGN CAPSULE – MACHINE LEARNING

Submitted on: 27.03.2025

Mehmet Fatih GÖĞÜŞ

Elif Nur BAYSAR

Instructor: Oğuzhan AYYILDIZ

OBJECTIVE

The objective of this lab is to analyze the impact of preprocessing techniques on the performance of a classification model using the Heart Disease Dataset from Kaggle. Normalization (Min-Max) and binning (equal-width) will be applied to examine their effects on model accuracy and performance metrics. The Random Forest algorithm will be used for classification, and K-Fold cross-validation will be performed to ensure a reliable evaluation. Performance will be measured through ROC-AUC scores and a confusion matrix. Finally, results will be compared to determine whether preprocessing significantly improves classification performance.

BACKGROUND

Heart disease remains a major health concern worldwide, which makes early detection an essential task in healthcare. Machine learning models are increasingly used to analyze patient data and predict the presence of heart disease. However, the accuracy and reliability of these models depend on the quality of the dataset and the preprocessing techniques applied.

The dataset used in this study consists of 1,000 instances and 14 attributes, including features such as age, cholesterol levels, resting blood pressure, maximum heart rate achieved, and presence of exercise-induced angina. The target variable, disease, indicates whether a patient has heart disease (1) or not (0). Since the dataset contains features with varying scales and distributions, preprocessing techniques such as normalization and binning will be evaluated to determine their impact on model performance.

In this lab, exploratory data analysis (EDA) will be conducted to understand the dataset's structure. Then, preprocessing techniques will be applied before training the Random Forest model. Performance will be assessed using ROC-AUC scores and a confusion matrix, and a comparison will be made between models trained with and without preprocessing to determine the effectiveness of these techniques.

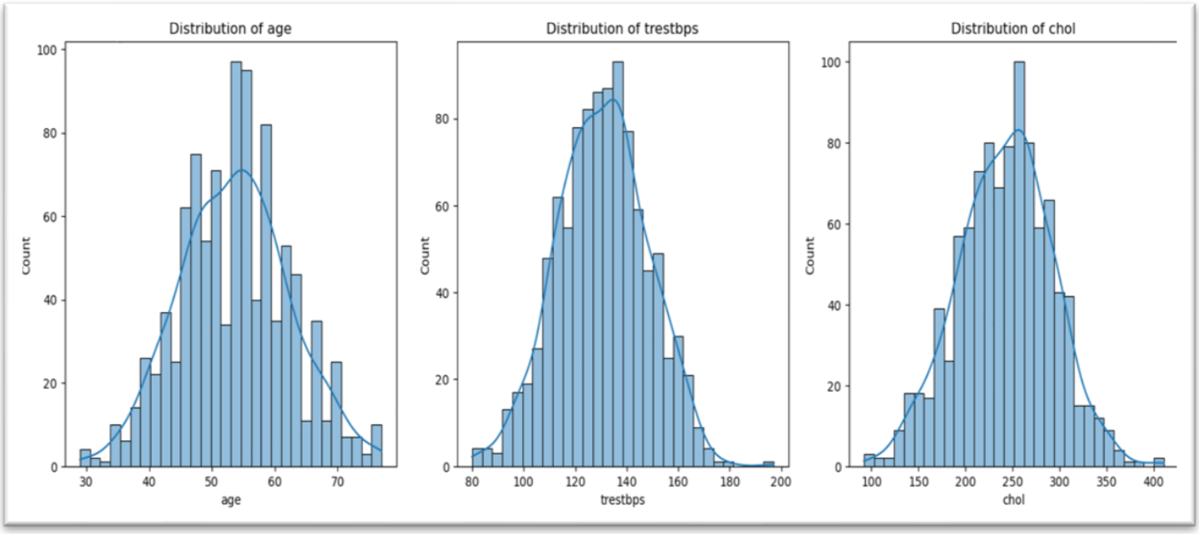


Figure 1 - Data Distribution

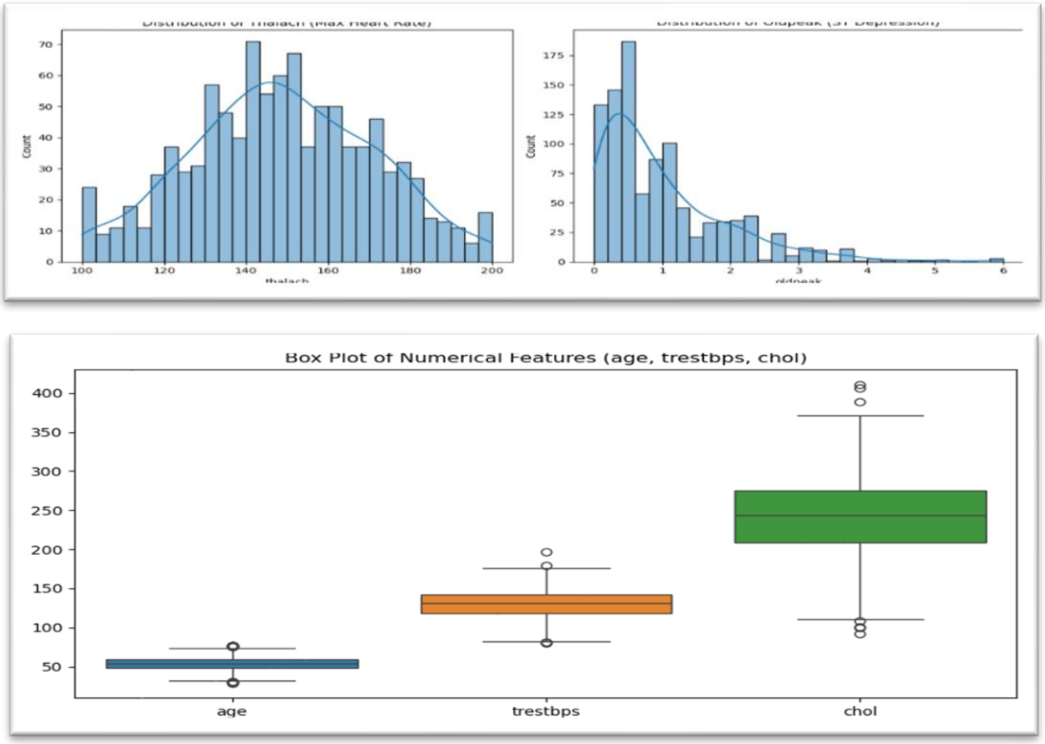


Figure 2 – Data Distribution

The dataset contains both numerical and categorical features, each with different distributions. As seen in Figures 1 and 2, the numerical features like age, resting blood pressure (trestbps), cholesterol (chol), maximum heart rate (thalach), and ST depression (oldpeak) have varying patterns. Most of them appear to follow a normal distribution, but some, like oldpeak, are right-skewed, meaning most values are low while a few are much higher. The box plot shows that cholesterol has the widest range and most outliers, while age is the most compact. Trestbps has a few outliers, but its spread is moderate. Cholesterol's high variability could affect the model's predictions.

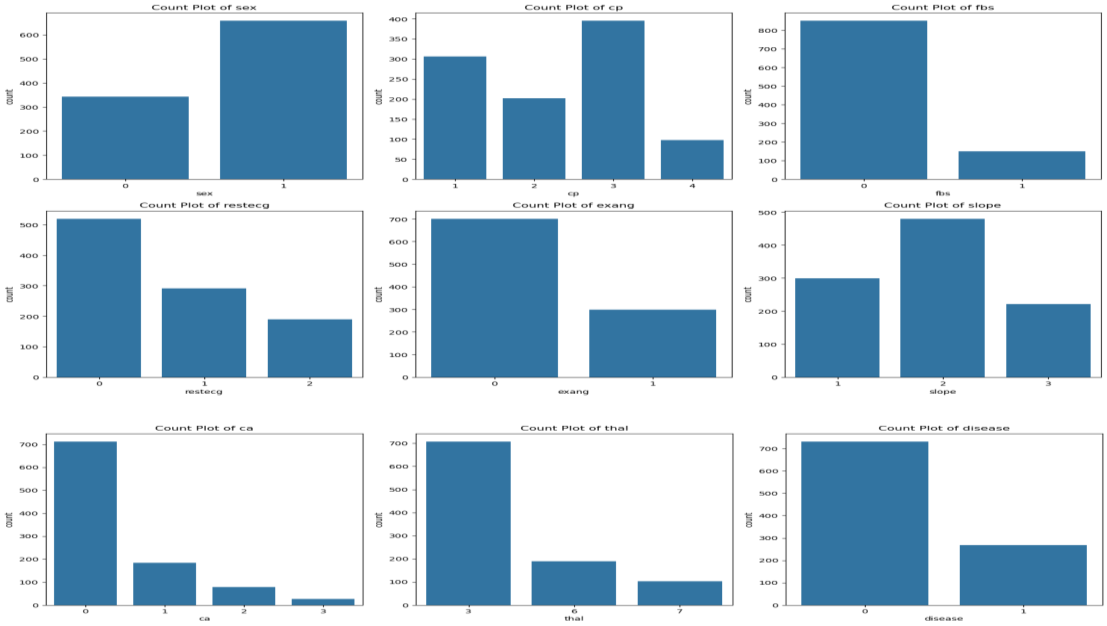
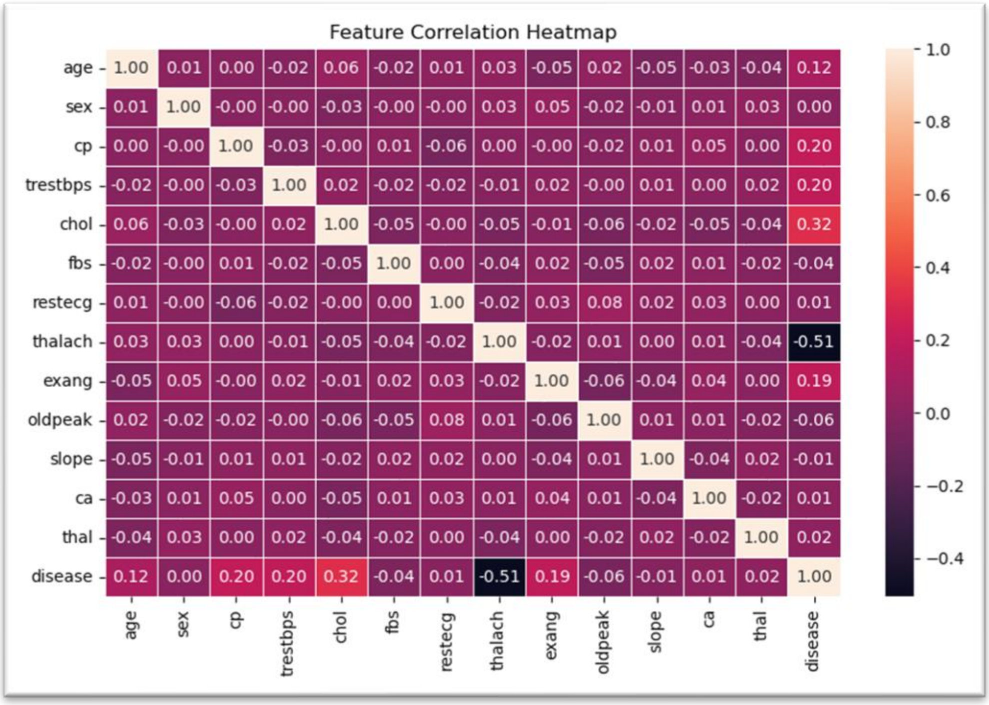


Figure3 - Data Distribution of Categorical Features

As shown in Figure 3, the categorical features, such as sex, chest pain type (cp), fasting blood sugar (fbs), ECG results (restecg), exercise-induced angina (exang), slope (ST Segment Slope), major vessels (ca), and thalassemia type (thal), show different class distributions. Some categories, like male patients in the sex feature, are more common, and there is an imbalance in the disease feature, where more patients do not have heart disease than those who do. Certain categories, like specific values in cp, ca, and thal, have fewer samples, which might impact classification performance.



Feature Correlation Heatmap

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	disease
age	1.00	0.01	0.00	-0.02	0.06	-0.02	0.01	0.03	-0.05	0.02	-0.05	-0.03	-0.04	0.12
sex	0.01	1.00	-0.00	-0.00	-0.03	-0.00	-0.00	0.03	0.05	-0.02	-0.01	0.01	0.03	0.00
cp	0.00	-0.00	1.00	-0.03	-0.00	0.01	-0.06	0.00	-0.00	-0.02	0.01	0.05	0.00	0.20
trestbps	-0.02	-0.00	-0.03	1.00	0.02	-0.02	-0.02	-0.01	0.02	-0.00	0.01	0.00	0.02	0.20
chol	0.06	-0.03	-0.00	0.02	1.00	-0.05	-0.00	-0.05	-0.01	-0.06	-0.02	-0.05	-0.04	0.32
fbs	-0.02	-0.00	0.01	-0.02	-0.05	1.00	0.00	-0.04	0.02	-0.05	0.02	0.01	-0.02	-0.04
restecg	0.01	-0.00	-0.06	-0.02	-0.00	0.00	1.00	-0.02	0.03	0.08	0.02	0.03	0.00	0.01
thalach	0.03	0.03	0.00	-0.01	-0.05	-0.04	-0.02	1.00	-0.02	0.01	0.00	0.01	-0.04	-0.51
exang	-0.05	0.05	-0.00	0.02	-0.01	0.02	0.03	-0.02	1.00	-0.06	-0.04	0.04	0.00	0.19
oldpeak	0.02	-0.02	-0.02	-0.00	-0.06	-0.05	0.08	0.01	-0.06	1.00	0.01	0.01	-0.02	-0.06
slope	-0.05	-0.01	0.01	0.01	-0.02	0.02	0.02	0.00	-0.04	0.01	1.00	-0.04	0.02	-0.01
ca	-0.03	0.01	0.05	0.00	-0.05	0.01	0.03	0.01	0.04	0.01	-0.04	1.00	-0.02	0.01
thal	-0.04	0.03	0.00	0.02	-0.04	-0.02	0.00	-0.04	0.00	-0.02	0.02	-0.02	1.00	0.02
disease	0.12	0.00	0.20	0.20	0.32	-0.04	0.01	-0.51	0.19	-0.06	-0.01	0.01	0.02	1.00

Figure 1 - Feature Correlation Heat Map

The heatmap, Figure 4, shows how different features in the dataset relate to each other and to the target variable (disease). Most of the values are close to zero, meaning that there aren't very strong correlations between features. However, some interesting patterns stand out.

- Thalach (Maximum Heart Rate) has the strongest negative correlation with disease (-0.51), meaning that lower max heart rate is more common in people with heart disease.
- Cholesterol (chol), chest pain type (cp), and resting blood pressure (trestbps) have the highest positive correlations with disease, but they're still not very strong (around 0.2-0.32). This suggests that higher cholesterol and certain types of chest pain might be linked to heart disease, but they aren't the only factors.
- Some features, like sex, age, and fasting blood sugar (fbs), don't seem to have a strong connection with disease, meaning they might not be that useful for making predictions.

Overall, there's no single feature that directly determines whether someone has heart disease. Instead, the model will likely need to use multiple features together to make accurate predictions.

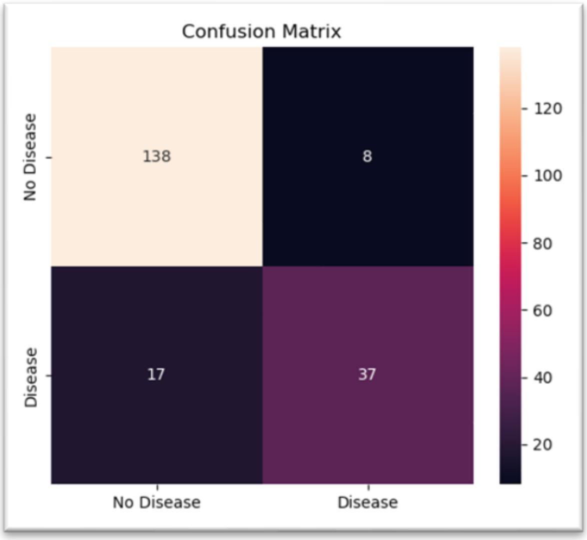


Figure 5 - Confusion Matrix

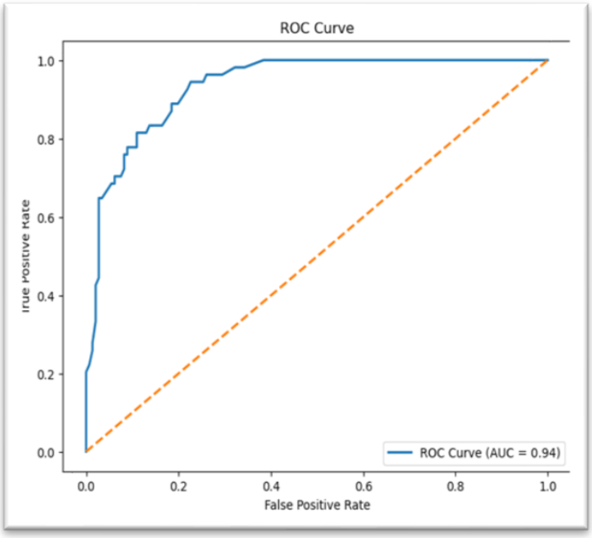
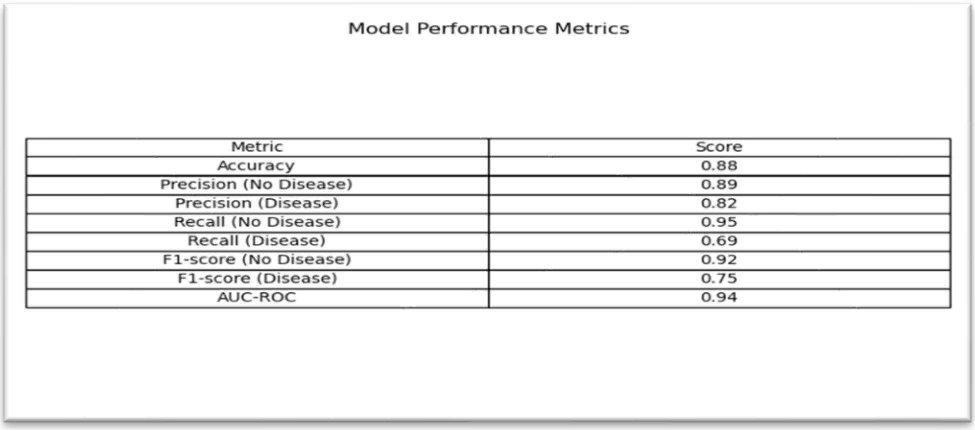


Figure 6 - ROC Curve

The confusion matrix (Figure 5) shows that the Random Forest model correctly classified most cases, with 138 true negatives and 37 true positives. However, 8 false positives and 17 false negatives indicate some misclassifications, with false negatives being critical in medical settings. The model maintains a good balance between sensitivity and specificity.

The ROC curve (Figure 6) confirms strong performance with an AUC score of 0.94, showing high discriminatory power. The curve stays well above the diagonal, indicating the model effectively distinguishes between patients with and without heart disease. While performance is strong, reducing false negatives remains an area for improvement to enhance reliability in medical predictions.



Model Performance Metrics

Metric	Score
Accuracy	0.88
Precision (No Disease)	0.89
Precision (Disease)	0.82
Recall (No Disease)	0.95
Recall (Disease)	0.69
F1-score (No Disease)	0.92
F1-score (Disease)	0.75
AUC-ROC	0.94

Figure 7 - Model Performance Metrics

The model performance metrics shown in Figure 7 indicate that the Random Forest classifier performed well on the heart disease dataset, achieving an overall accuracy of 88%. The precision and recall values for both classes show a slight imbalance with the model performing better in identifying "No Disease" cases compared to "Disease" cases. Specifically, the recall for "No Disease" is 0.95, meaning most healthy individuals were correctly classified, whereas the recall for "Disease" is 0.69, indicating that some patients with heart disease were misclassified as healthy.

The F1-scores, which balance precision and recall, further highlight this imbalance. The F1-score for "No Disease" (0.92) is higher than that for "Disease" (0.75), suggesting that the model is more confident in correctly classifying healthy patients. All in all, the AUC-ROC score of 0.94 confirms that the classifier has a strong ability to differentiate between the two classes overall.

Application of Min-Max Normalization

In this part, we applied Min-Max normalization to scale all features between 0 and 1, ensuring that no feature dominates the others due to larger numerical values. We then evaluated its impact on model performance.

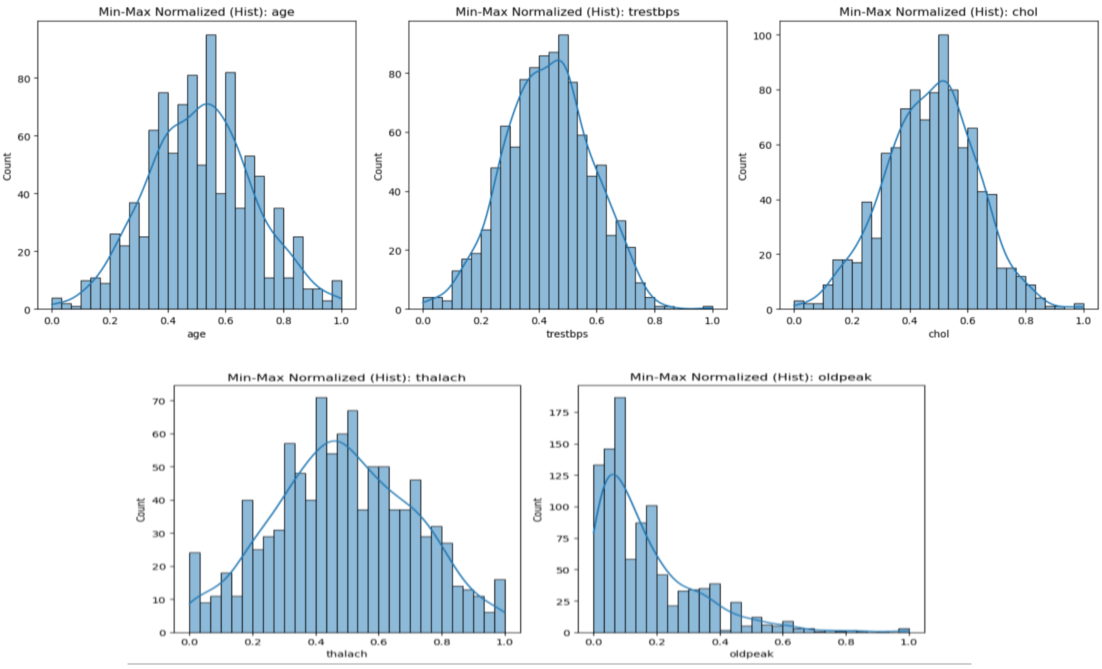
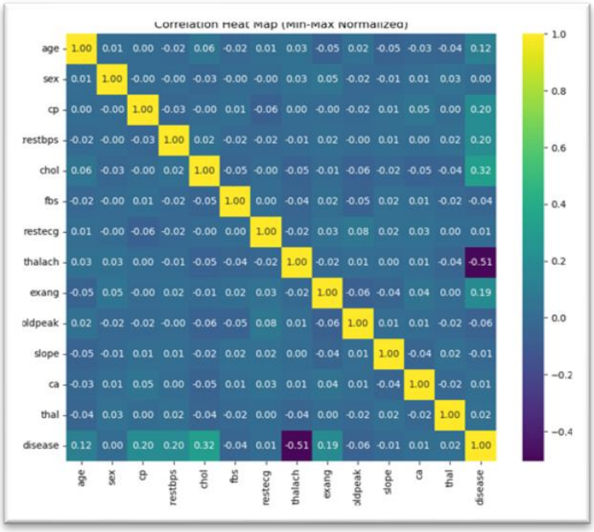


Figure 8 – Min-Max Normalization

The histograms, Figure 8, show the distribution of several key features in the dataset after applying Min-Max normalization. Overall, most of the features exhibit a roughly normal distribution, such as age, cholesterol (chol), and resting blood pressure (trestbps), which have a bell-shaped curve centered around the middle of the normalized range. This suggests that these features are well-distributed and do not have extreme skewness.

However, some features like oldpeak (ST depression) are highly skewed with most values concentrated towards the lower end of the range. This could indicate that a majority of patients have low ST depression values, but a few have significantly higher ones, creating a long tail in the distribution.



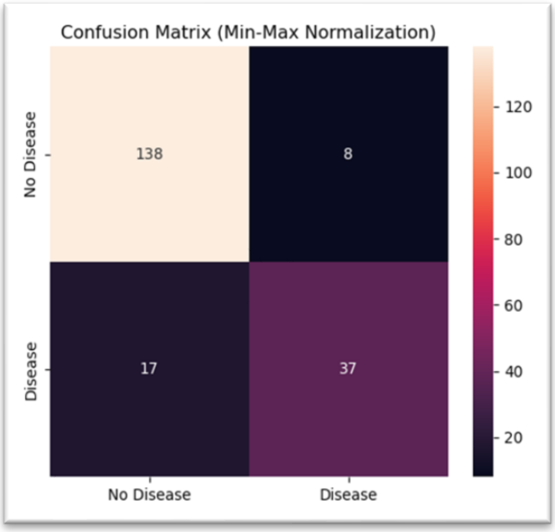


Figure 9 - Heat Map with Min-Max Normalization

Figure 10 - Confusion Matrix with Min-Max Normalization

The correlation heatmap (Figure 9) shows that thalach (max heart rate) has the strongest correlation with disease (-0.51), meaning higher heart rates are linked to a lower risk of heart disease. Cholesterol (chol) and trestbps (resting BP) have weak positive correlations, which suggests they may contribute but aren't strong predictors. Overall, the model likely relies on multiple features rather than a single dominant one.

The confusion matrix after Min-Max normalization (Figure 10) shows the same performance to the original data, with 138 true negatives and 37 true positives, but 17 false negatives remain an issue. Normalization didn't impact results, and the reason could be the model was already performing well.

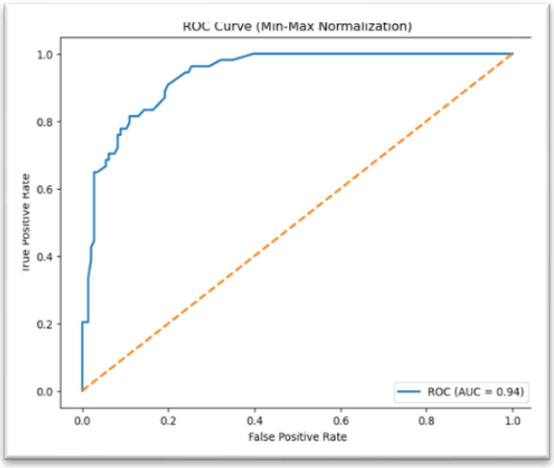
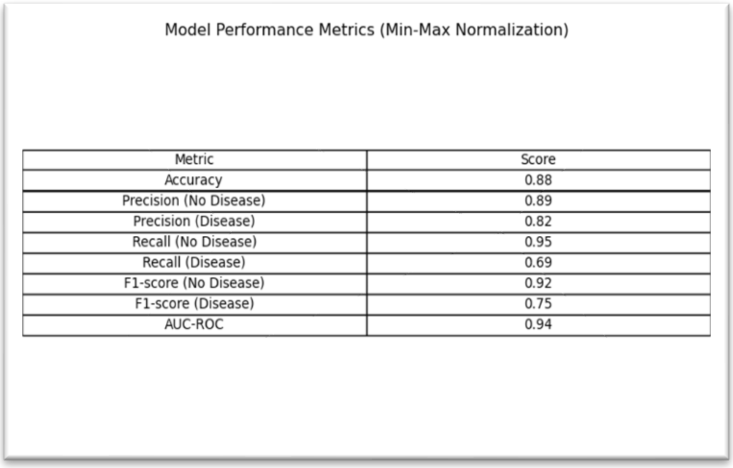


Figure 11 - ROC Curve with Min-Max Normalization



Model Performance Metrics (Min-Max Normalization)

Metric	Score
Accuracy	0.88
Precision (No Disease)	0.89
Precision (Disease)	0.82
Recall (No Disease)	0.95
Recall (Disease)	0.69
F1-score (No Disease)	0.92
F1-score (Disease)	0.75
AUC-ROC	0.94

Figure 12 - Model Performance Metrics with Min-Max Normalization

The ROC curve, Figure 11, after Min-Max normalization remains almost identical to the previous one, with an AUC score of 0.94. This suggests that the model's ability to distinguish between patients with and without heart disease was not affected by normalization. The curve remains well above the diagonal baseline, indicating strong classification performance. The high AUC value confirms that the model is making accurate predictions overall, but since recall for disease cases is still relatively low, some misclassifications (especially false negatives) persist.

The performance metrics table, Figure 12, also shows that Min-Max normalization did not bring improvements. The accuracy remains at 88%, and key metrics like precision and recall for both classes are unchanged. The recall for disease cases is still at 0.69. While normalization helped scale features, it did not impact classification results, suggesting that the Random Forest model is robust to different feature scales.

Application of Log-Based Normalization to the Skewed Feature (oldpeak)

Here, we transformed skewed features using log normalization to reduce the effect of extreme values and improve data distribution. The classification results were analyzed to assess whether this transformation helped in better prediction.

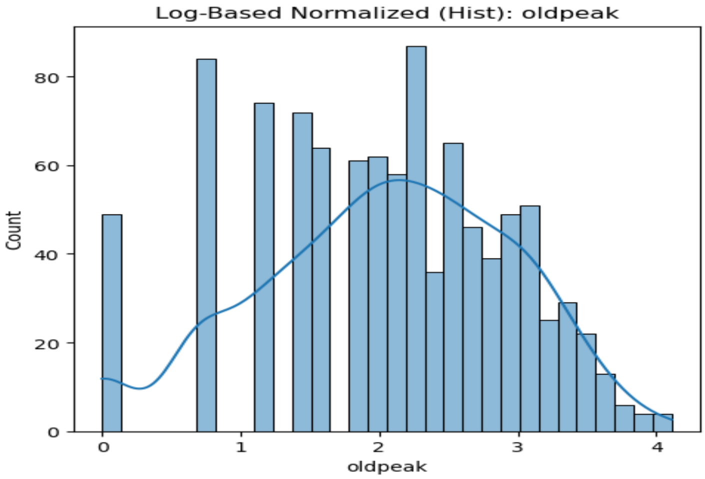
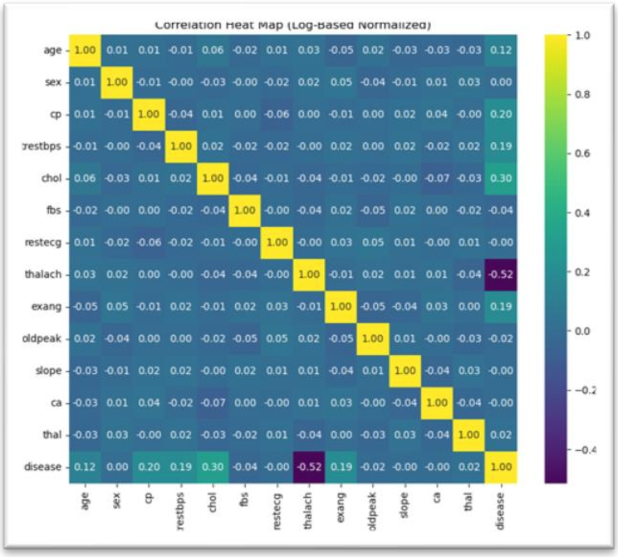


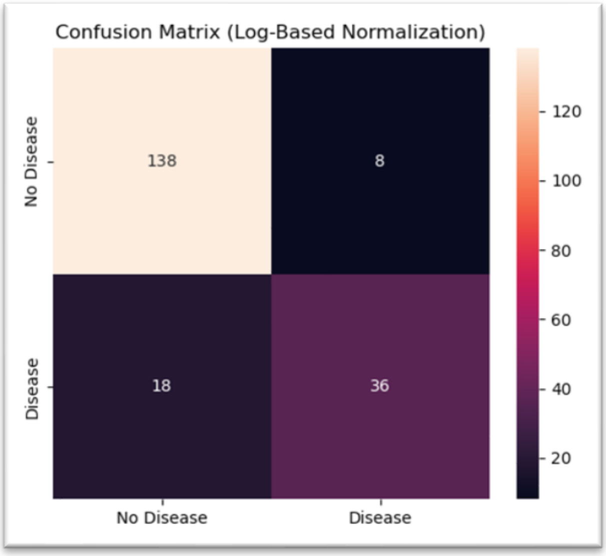
Figure 13 – Oldpeak Distribution with Log Normalization

The histogram of the "oldpeak (ST Depression)" feature after log-based normalization, Figure 13, shows a more balanced distribution compared to its original skewed form. Previously, the oldpeak (ST Depression) values were heavily right-skewed, with most data points concentrated near zero. After applying log transformation, the distribution appears more spread out, reducing the extreme skewness and making it more suitable for machine learning models. This transformation helps in cases where features contain large variations in scale, as it compresses higher values while keeping lower values more distinguishable.



	age	sex	cp	restbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	disease
age	1.00	0.01	0.01	-0.01	0.06	-0.02	0.01	0.03	-0.05	0.02	-0.03	-0.03	-0.03	0.12
sex	0.01	1.00	-0.01	-0.00	-0.03	-0.00	-0.02	0.02	0.05	-0.04	-0.01	0.01	0.03	0.00
cp	0.01	-0.01	1.00	-0.04	0.01	0.00	-0.06	0.00	-0.01	0.00	0.02	0.04	-0.00	0.20
restbps	-0.01	-0.00	-0.04	1.00	0.02	-0.02	-0.02	-0.00	0.02	0.00	0.02	-0.02	0.02	0.19
chol	0.06	-0.03	0.01	0.02	1.00	-0.04	-0.01	-0.04	-0.01	-0.02	-0.00	-0.07	-0.03	0.30
fbs	-0.02	-0.00	0.00	-0.02	-0.04	1.00	-0.00	-0.04	0.02	-0.05	0.02	0.00	-0.02	-0.04
restecg	0.01	-0.02	-0.06	-0.02	-0.01	-0.00	1.00	-0.00	0.03	0.05	0.01	-0.00	0.01	-0.00
thalach	0.03	0.02	0.00	-0.00	-0.04	-0.04	-0.00	1.00	-0.01	0.02	0.01	0.01	-0.04	-0.52
exang	-0.05	0.05	-0.01	0.02	-0.01	0.02	0.03	-0.01	1.00	-0.05	-0.04	0.03	0.00	0.19
oldpeak	0.02	-0.04	0.00	0.00	-0.02	-0.05	0.05	0.02	-0.05	1.00	0.01	-0.00	-0.03	-0.02
slope	-0.03	-0.01	0.02	0.02	-0.00	0.02	0.01	0.01	-0.04	0.01	1.00	-0.04	0.03	-0.00
ca	-0.03	0.01	0.04	-0.02	-0.07	0.00	-0.00	0.01	0.03	-0.00	-0.04	1.00	-0.04	-0.00
thal	-0.03	0.03	-0.00	0.02	-0.03	-0.02	0.01	-0.04	0.00	-0.03	0.03	-0.04	1.00	0.02
disease	0.12	0.00	0.20	0.19	0.30	-0.04	-0.00	-0.52	0.19	-0.02	-0.00	-0.00	0.02	1.00

Figure 14 – Correlation Heat Map with Log Normalization



	No Disease	Disease
No Disease	138	8
Disease	18	36

Figure 15 – Confusion Matrix with Log Normalization

The correlation heatmap after log normalization, Figure 14, is nearly the same as before, with thalach (-0.52) still showing the strongest correlation with disease (slightly more negative than -0.51 in Min-Max). Oldpeak remains unchanged, meaning while log normalization smoothed its distribution, it didn’t improve its predictive power.

The confusion matrix, Figure 15, shows similar performance with 138 true negatives and 36 true positives, but one more false negative (18 instead of 17 from Min-Max). This suggests log normalization did not improve recall for disease cases. Overall, the choice between Min-Max and log normalization does not drastically impact results but further tuning could help boost performance.

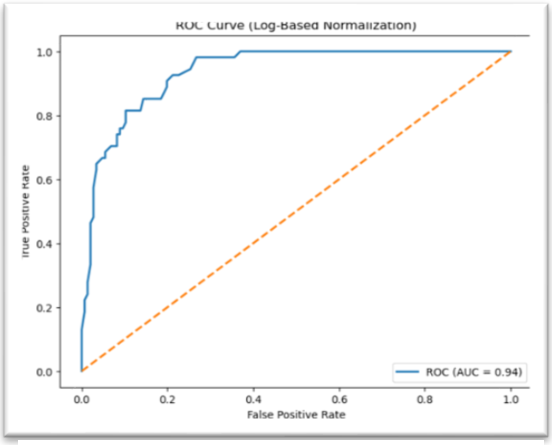
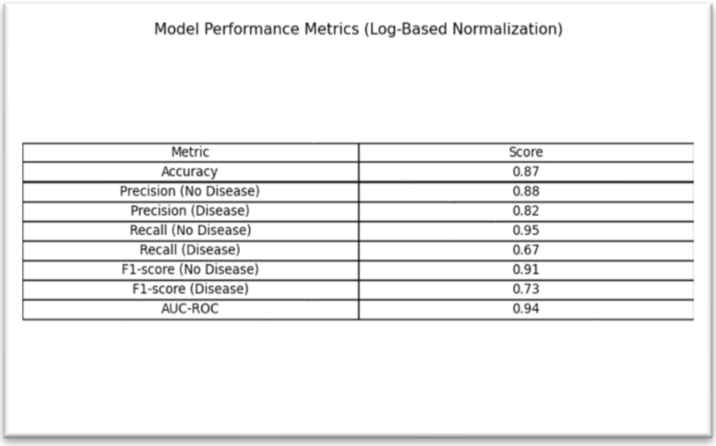


Figure 16 – ROC Curve with Log Normalization



Model Performance Metrics (Log-Based Normalization)

Metric	Score
Accuracy	0.87
Precision (No Disease)	0.88
Precision (Disease)	0.82
Recall (No Disease)	0.95
Recall (Disease)	0.67
F1-score (No Disease)	0.91
F1-score (Disease)	0.73
AUC-ROC	0.94

Figure 17 – Model Performance Metrices with Log Normalization

The ROC curve after log normalization, Figure 16, remains nearly identical, with an AUC of 0.94, meaning classification performance didn’t change much. However, recall for disease cases slightly decreased, making log normalization less effective for reducing false negatives.

The performance metrics (Figure 17) show a slight accuracy drop (0.88 → 0.87) and lower recall for disease (0.69 → 0.67), leading to more false negatives. The F1-score for disease also dropped (0.75 → 0.73), confirming weaker disease detection. However, AUC-ROC stayed at 0.94, meaning the model’s overall ability to separate classes wasn’t affected.

Bonamutial and Prasetyo (2023) report that although normalization and standardization enhance some regression models, the random forest regressor performs best without scaling, yielding lower error and higher accuracy *[1]*. Overall, all this makes sense because Random Forest doesn’t rely on feature scaling, so normalizing the data had no impact on the model’s decision-making process. The results confirm that for tree-based models like Random Forest, normalization isn’t that necessary and doesn’t improve performance.

Application of Equal-Width Binning (Bin Numbers = 4)

We converted continuous numerical features into 4 discrete bins of equal width to observe how simplifying data impacts the classifier. The model's accuracy, recall, and overall performance were compared to the original data to see if binning affected its ability to distinguish between classes.

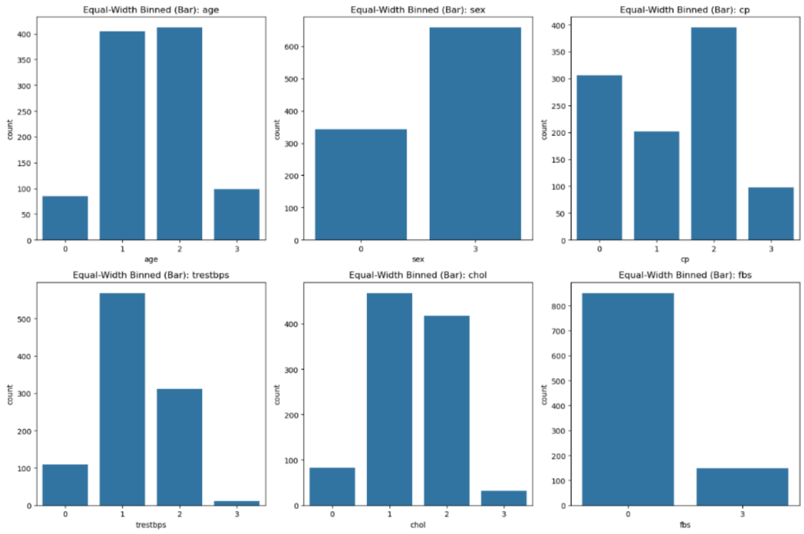
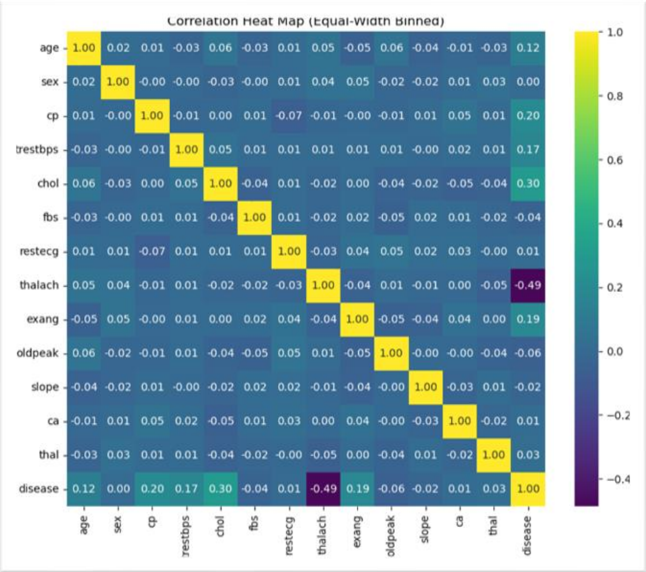


Figure 18 – Data Distribution with Equal-Width Binning

Figure 18 shows how equal-width binning grouped features into categories. Some features, like age, trestbps (Resting Blood Pressure), and cholesterol, are mostly concentrated in the middle bins, while others, like cp and fbs (Fasting Blood Sugar), have more balanced distributions. However, binning might have caused information loss, especially for features with uneven distributions, like cholesterol and trestbps.



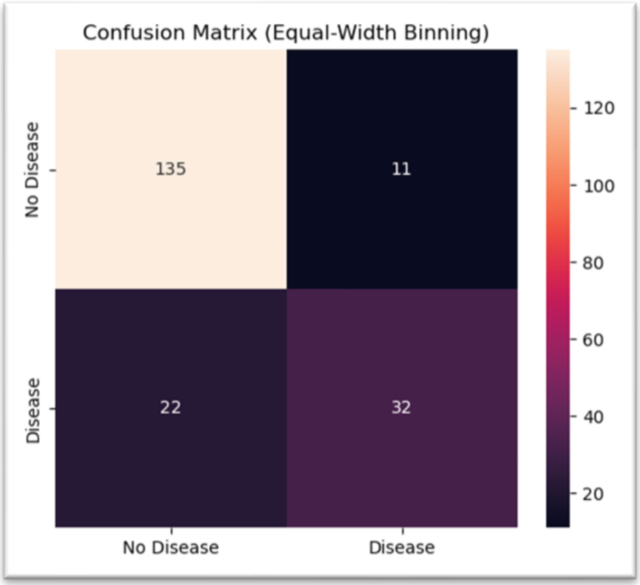


Figure 19 – Correlation Heat Map with Equal-Width Binning

Figure 20 – Confusion Matrix with Equal-Width Binning

The correlation heatmap, Figure 19, after equal-width binning shows that feature relationships have remained mostly the same, but the correlation values have slightly changed. The correlation between thalach (max heart rate) and disease dropped slightly to -0.49 (compared to -0.51 in the original data). Other features, such as cholesterol and trestbps (Resting Blood Pressure), still have weak positive correlations with disease. This suggests that binning might have slightly impacted on the relationships between variables, but the overall trends remain similar.

The confusion matrix after equal-width binning, Figure 20, shows a decline in model performance compared to the original data. The model now has more false negatives (22 vs. 17 originally) and more false positives (11 vs. 8 originally). This means the model is struggling more to correctly classify disease cases, reducing recall. The loss of information due to binning might make it harder for the model to capture patterns effectively.

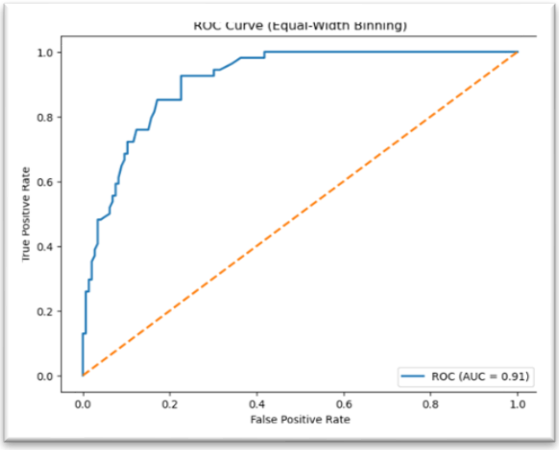
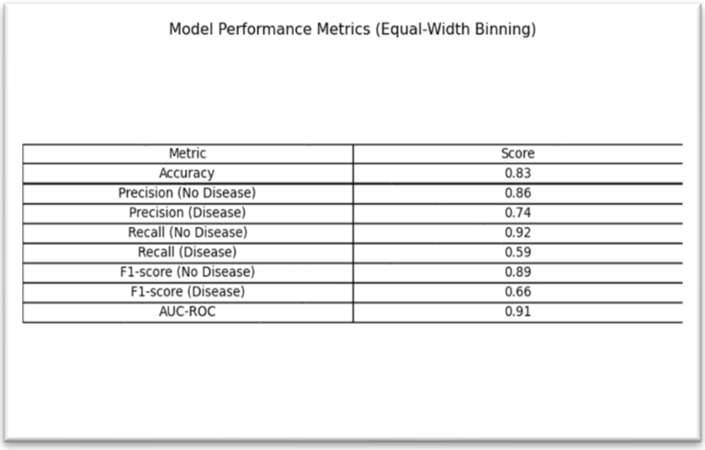


Figure 21 – ROC Curve with Equal-Width Binning



Model Performance Metrics (Equal-Width Binning)

Metric	Score
Accuracy	0.83
Precision (No Disease)	0.86
Precision (Disease)	0.74
Recall (No Disease)	0.92
Recall (Disease)	0.59
F1-score (No Disease)	0.89
F1-score (Disease)	0.66
AUC-ROC	0.91

Figure 22 – Model Performance Metrices with Equal-Width Binning

The ROC curve after equal-width binning, Figure 21, shows a noticeable drop in performance compared to the original data and other preprocessing techniques. The AUC score decreased from 0.94 (original and normalized versions) to 0.91, which indicates that the model is now slightly less effective at distinguishing between the two classes. This drop suggests that binning may have removed some valuable information from the data, leading to reduced classification power.

The performance metrics table, Figure 22, further confirms this decline. The accuracy dropped from 0.88 (original) to 0.83, and the recall for disease cases fell significantly from 0.69 (original) to 0.59. This means the model is missing more actual disease cases, which is a major concern in medical applications. The F1-score for disease cases also dropped from 0.75 to 0.66, which shows that binning has negatively impacted the model’s ability to correctly classify patients with heart disease.

Application of Equal-Width Binning (Bin Numbers = 12)

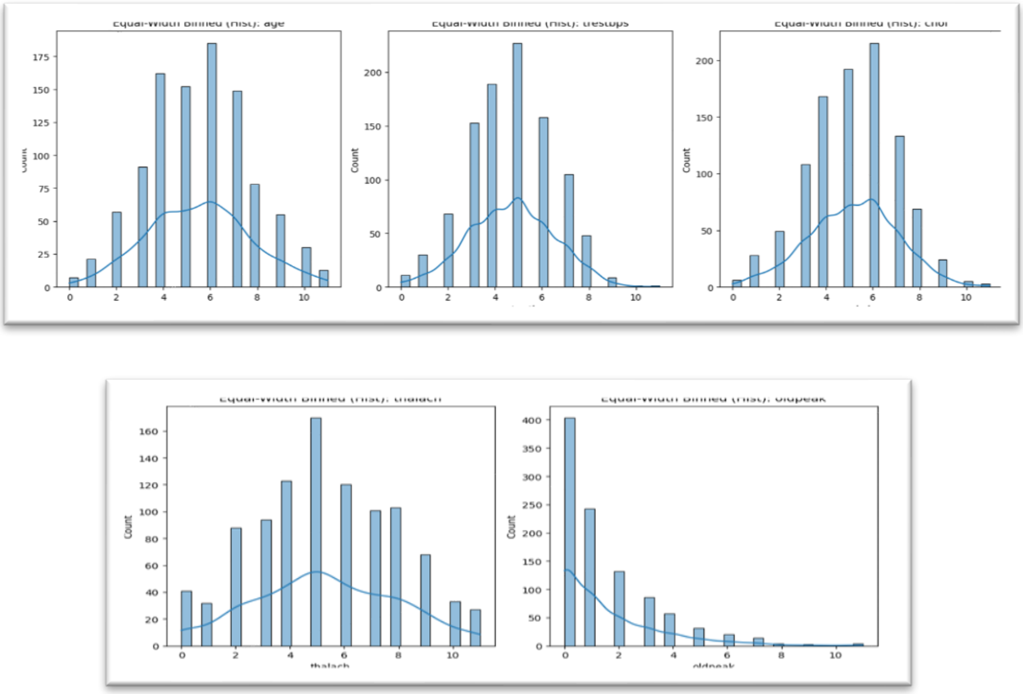


Figure 23 -Data Distribution with Equal-Width Binning

The 12-bin equal-width binning provides a more detailed view than the 4-bin version, preserving more data patterns. Features like age, trestbps, chol, and thalach still show a normal-like distribution, while oldpeak remains skewed as seen in Figure 23. This finer binning reduces information loss but still introduces artificial cutoffs.

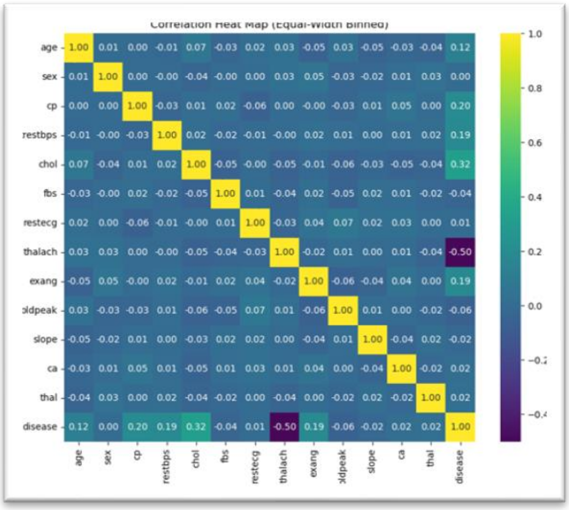


Figure 24 – Correlation Heat Map with Equal-Width Binning

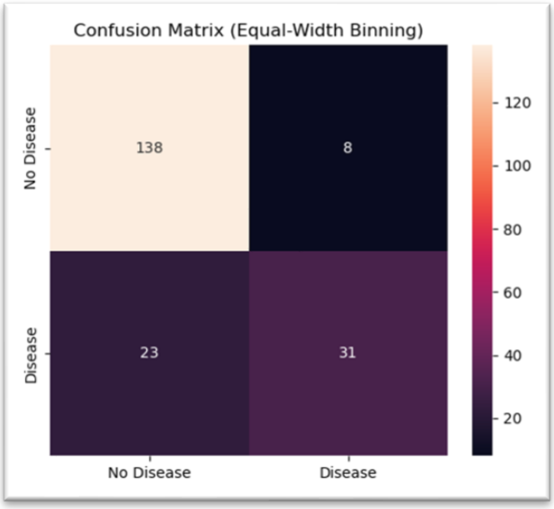


Figure 25 – Confusion Matrix with Equal-Width Binning

The 12-bin correlation heatmap, Figure 24, shows similar trends to previous versions. Thalach (-0.50) still has the strongest correlation with disease, while other features remain weakly correlated.

As shown in Figure 25, with 12-bin equal-width binning, the model misclassified 23 disease cases as "No Disease," which is worse than the original (17) and slightly worse than the 4-bin version (22). False positives (8) remained the same. The drop in recall shows that binning still negatively impacts disease detection.

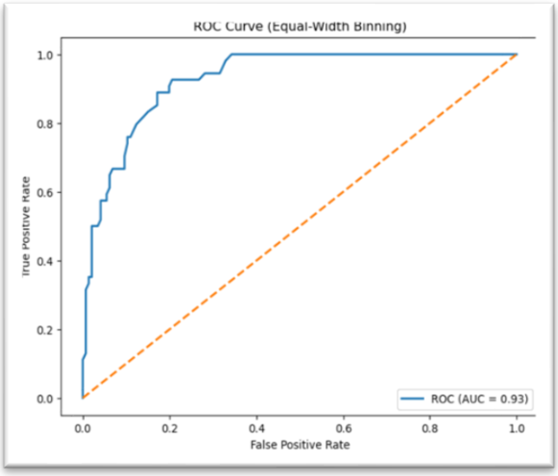
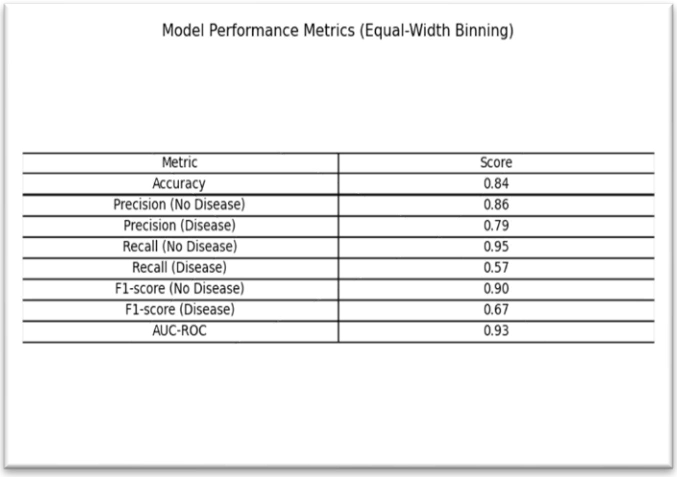


Figure 26 – ROC Curve with Equal-Width Binning



Model Performance Metrics (Equal-Width Binning)

Metric	Score
Accuracy	0.84
Precision (No Disease)	0.86
Precision (Disease)	0.79
Recall (No Disease)	0.95
Recall (Disease)	0.57
F1-score (No Disease)	0.90
F1-score (Disease)	0.67
AUC-ROC	0.93

Figure 27 –Model Performance Metrices with Equal-Width Binning

As seen in Figure 26, the AUC-ROC score (0.93) is slightly better than the 4-bin version (0.91) but still lower than the original (0.94). This means the model's ability to distinguish between classes slightly improved compared to fewer bins but is still weaker than using raw data.

As demonstrated in Figure 27, accuracy (0.84) improved slightly compared to 4-bin binning (0.83) but is still below the original (0.88). Recall for disease (0.57) dropped even further, increasing false negatives. While binning with more intervals retains more details, it still harms classification performance compared to normalization.

Overall, both 12-bin and 4-bin equal-width binning reduced model performance compared to the original data, especially in detecting disease cases. The 12-bin version performed slightly better than the 4-bin, with a higher AUC-ROC (0.93 vs. 0.91). Both methods increased false negatives, showing that binning is not the best preprocessing choice for this dataset.

Application of K - Fold Cross Validation

To ensure a more reliable performance estimate, we used 5-Fold cross-validation, splitting the data into five different training and test sets. The average results were compared to previous methods to check for improvements in stability and generalization.

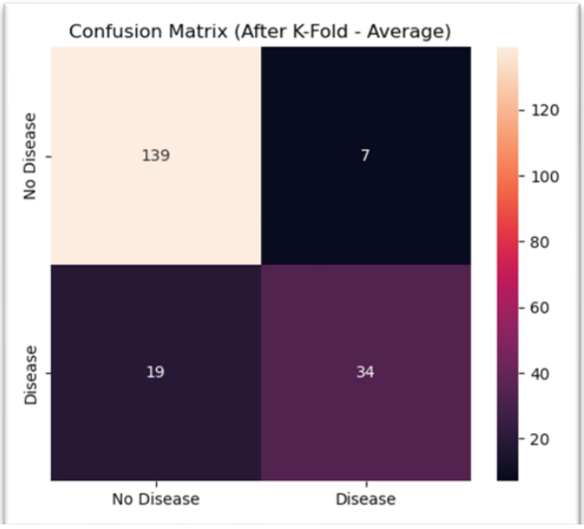


Figure 28 – Confusion Matrix with K - Fold Cross Validation

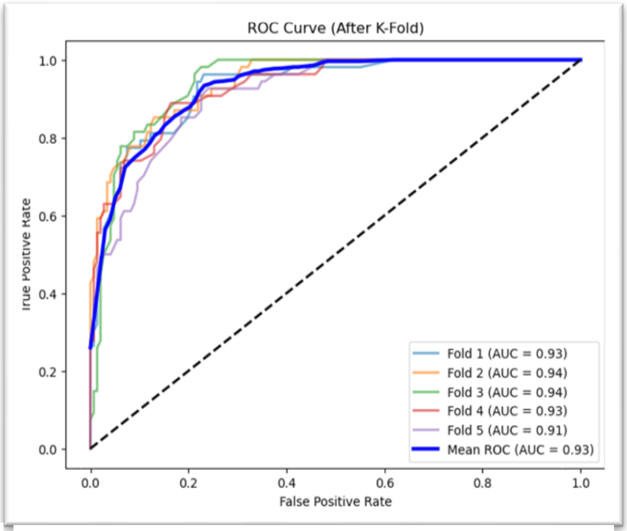
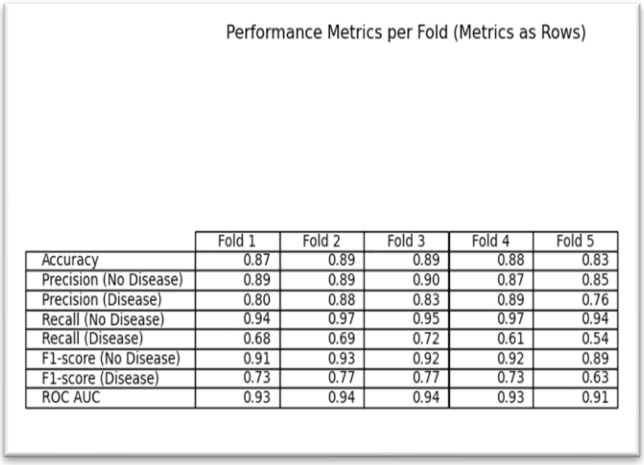


Figure 29 – ROC Curve with K - Fold Cross Validation

The confusion matrix after K-Fold cross-validation, Figure 28, shows slight improvements in the model's performance. Compared to the original confusion matrix, false positives decreased from 8 to 7, meaning fewer healthy patients were misclassified as having the disease. However, false negatives slightly increased from 17 to 19, which shows that a few more actual disease cases were missed. Despite this, the model remains consistent in its ability to correctly classify both "No Disease" and "Disease" cases.

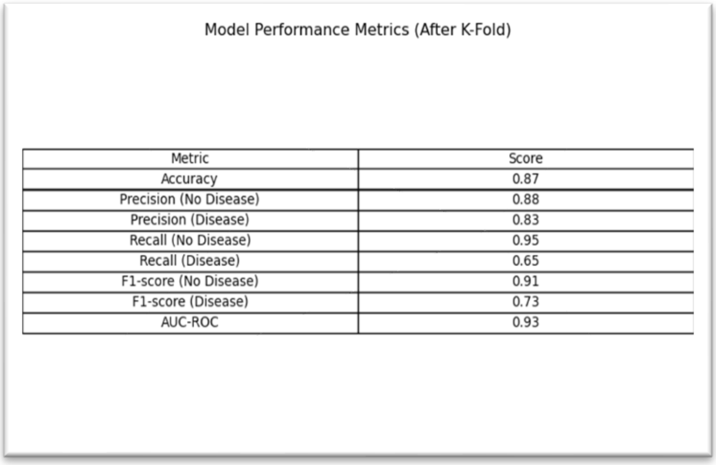
The ROC curve after K-Fold validation, Figure 29, further supports the model’s stability. The individual AUC scores for each fold range between 0.91 and 0.94, with an average AUC of 0.93, which is slightly lower than the original 0.94. This slight variation across folds is expected due to differences in training and test splits, but overall, the model performs well across different subsets of data.



Performance Metrics per Fold (Metrics as Rows)

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Accuracy	0.87	0.89	0.89	0.88	0.83
Precision (No Disease)	0.89	0.89	0.90	0.87	0.85
Precision (Disease)	0.80	0.88	0.83	0.89	0.76
Recall (No Disease)	0.94	0.97	0.95	0.97	0.94
Recall (Disease)	0.68	0.69	0.72	0.61	0.54
F1-score (No Disease)	0.91	0.93	0.92	0.92	0.89
F1-score (Disease)	0.73	0.77	0.77	0.73	0.63
ROC AUC	0.93	0.94	0.94	0.93	0.91

Figure 30 – Model Performance Metrices per Fold



Model Performance Metrics (After K-Fold)

Metric	Score
Accuracy	0.87
Precision (No Disease)	0.88
Precision (Disease)	0.83
Recall (No Disease)	0.95
Recall (Disease)	0.65
F1-score (No Disease)	0.91
F1-score (Disease)	0.73
AUC-ROC	0.93

Figure 31 - Model Performance Metrices with K - Fold Cross Validation

The performance metrics per fold, Figure 30, show that the model's performance remains fairly stable across different splits, but there are some variations. Accuracy ranges from 0.83 to 0.89, and recall for disease cases fluctuates between 0.54 and 0.72, which indicates that some folds perform better than others in detecting heart disease. The AUC-ROC values are consistently high (between 0.91 and 0.94), which means that the classifier maintains strong predictive power across all folds.

The average performance metrics after K-Fold validation, Figure 31, confirm these observations. The accuracy (0.87) is slightly lower than the original (0.88), but still strong. The recall for disease cases dropped to 0.65 from the original 0.69, which means the model is still struggling with false negatives. The precision and F1-score values remain close to the original results, which suggests that K-Fold validation did not drastically change the model’s ability to classify correctly, but it provided a more reliable estimate of its real-world performance.

Overall, K-Fold validation helped reduce overfitting and gave a better generalization of the model's performance. The results confirm that the model maintains strong predictive power across different data splits, with only minor variations in performance.

DISCUSSION

In this lab, we tested how different preprocessing techniques, Min-Max normalization, log-based normalization, and equal-width binning, affect the performance of a Random Forest classifier on the Heart Disease Dataset. We also applied K-Fold cross-validation to check the model's generalization. After comparing results with the original dataset, we analyzed whether these preprocessing methods improved classification performance or not.

Min-Max normalization didn't change much in the model's accuracy, recall, or AUC-ROC score, which makes sense because Random Forest models aren't sensitive to feature scaling. Log-based normalization also had minimal impact but slightly decreased recall for disease cases, which makes it less effective for improving classification performance.

Equal-width binning, however, had a more noticeable negative impact. Both 4-bin and 12-bin versions led to a drop in accuracy and recall, increasing false negatives. The 12-bin version performed slightly better than the 4-bin since it retained more details, but overall, binning made it harder for the model to distinguish between patients with and without heart disease. AUC-ROC scores also dropped, which shows that binning reduced classification power compared to normalization.

K-Fold cross-validation gave a more reliable performance estimate by reducing bias from a single train-test split. The results showed minor variations between folds with an average AUC-ROC of 0.93, slightly lower than the original 0.94. There was a small increase in false negatives, but overall, the model's classification ability remained stable, confirming its consistency across different subsets of data.

CONCLUSION

Overall, different preprocessing techniques had varying effects on model performance. Min-Max and log-based normalization didn't significantly change results, proving that Random Forest models don't require feature scaling. On the other hand, equal-width binning negatively affected classification, increasing false negatives and reducing recall, making it less effective for this dataset. K-Fold cross-validation showed that the model remained reliable across different data splits with only slight variations in performance.

These results suggest that normalization techniques are unnecessary for Random Forest models, while binning reduces classification effectiveness. The best results came from using the original dataset, which reinforces the idea that preprocessing methods should be chosen based on the specific needs of the model. Future work could explore other techniques like feature selection or class balancing to improve recall and reduce misclassification in heart disease detection.

REFERENCES

[1] M. Bonamutial and S. Y. Prasetyo, "Exploring the Impact of Feature Data Normalization and Standardization on Regression Models for Smartphone Price Prediction," *2023 International Conference on Information Management and Technology (ICIMTech)*, Malang, Indonesia, 2023, pp. 294-298, doi: 10.1109/ICIMTech59029.2023.10277860