Exploratory Data Analysis of a Heart Disease Dataset

R. Hernanz Hernández[1],

[1] Superior Polytechnic School, CEU San Pablo, Madrid, Spain, raquel.hernanzhernandez@usp.ceu.es

Abstract

Exploratory data analysis is essential for identifying structure, variability, and inconsistencies across clinical datasets. This study applies this technique to a public heart disease dataset to assess data quality and explore feature behaviour. Using Python-based statistical and visualization tools, distributions, correlations, and outliers were analysed. The results highlight skewed variables and irregular value ranges, underscoring the importance of early data evaluation to ensure reliable preprocessing and modelling in cardiovascular disease research.

1. Introduction

Cardiovascular diseases (CVDs) remain among main causes of mortality worldwide, with 17.9 million deaths registered and representing 32% of global deaths according to the American Heart Association (AHA) [1]. Advances in data science have encouraged the use of computational and data-driven workflows to support diagnosis, risk stratification and early detection of CVDs [2], [3]. Before implementing predictive models, exploratory data analysis (EDA) is essential to evaluate data structure, variability, and possible inconsistencies. This paper performs an EDA of a public heart disease dataset to assess data quality, visualize feature distributions, and explore relationships between clinical variables [4]. The analysis aims to provide a clear overview that supports future preprocessing and modelling stages in cardiovascular research.

2. Materials and methods

The heart disease dataset is from Kaggle platform [4]. It originates from four databases of 1988 (Cleveland, Hungary, Switzerland and Long Beach). The EDA was performed in Python using Google Colab environment (version 2025-13-11), and the notebooks were uploaded into a GitHub repository [5]. Statistical and visualization libraries such as *Pandas*, *NumPy*, *Scikit-learn* or *Tensorflow* were employed to examine data structure, detect outliers, and analyse variable distributions [6], [7], [8], [9].

3. Dataset features

In the original dataset there were 76 feature columns. However, the current version of 2025 from Kaggle has 14 features (*age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, target*) with 1025 instances in total [4]. In **Appendix A**, **Table 4,** descriptions and value of each feature are addressed in the exhibit. Initially, there are various key observations:

- There are no missing values in the rows, therefore, there are no issues related to Missing At Random values (MAR) and Missing Not At Random values (MNAR).
- According to the methods employed for finding duplicates, the dataset has 723 duplicates, however, it doesn't contain an identifier column that differentiate patients. Therefore, removing duplicates is discarded.
- Shares attributes with research datasets from other studies (e.g. the glycaemia, cholesterol, sex, age, heartbeat rate) [2], [3], [10], [11].
- Some discrepancies and patterns between the expected clinical behaviour of certain risk factors observed arise from limitations of the original dataset and the heterogeneity across the four cohorts.

4. Statistical summary

The *pandas.DataFrame.describe()* method was used to obtain statistical values (count, mean, SD, min, max, and quantiles) [7]. **Table 1,** focuses on the most relevant for inference and on those showing abnormalities (outliers and non-normal distribution).

Feature	Mean ± SD	Range	Brief observations
age	54.4 ± 9.1 years	29-77 years	Most patients are between 45–63 years old.
sex	0.70 ± 0.46	0-1	~70% male, 30% female
trestbps	131.6 ± 17.5 mmHg	94-200 mmHg	Normal to slightly elevated blood pressure.
chol	246 ± 51.6 mg/dL	126-564 mg/dL	Some extreme outliers, e.g., one with 526 mg/dL.
thalach	149 ± 23.0 bps	71-202 bps	Heart rate normally distributed, with some extreme outliers.
oldpeak	1.07 ± 1.18 mm	0-6.2 mm	ST segment depression shows right-skewed distribution.
ca	0.75 ± 1.03	0-3 vessels	At least 1-2 vessels are affected
thal	2.32 ± 0.62	0-3	Most values = 2 ("normal").

Table 1. *Statistical summary of the features*

These findings align with prior work from other cohorts [1], [3], [10], [12]:

- CVDs affect older populations more frequently, as prevalence and mortality increase significantly with age and the accumulation of comorbidities.
- The gender bias suggests the dataset may be more representative in the male population. Some research datasets do differentiate between male and females.
- Higher blood pressure and cholesterol level are common in CVDs patients, representing major modifiable risk factors for coronary artery disease and stroke.
- Elevated ST-segment depression values indicate myocardial ischemia under stress, aligning with findings of AI-assisted ECG detection of ischemic patterns.
- Multiple affected coronary vessels are associated with advanced atherosclerosis and poorer prognosis.

Additionally, the box plots are located in the **Appendix B**. **Figure 1**, corresponds to the box plots of *age*, *trestbps*, *chol* and *thalach*, while **Figure 2** is only for *oldpeak.*

5. Plot analysis

To facilitate the interpretation of histograms and scatter plots, the features were grouped according to plot type.

5.1. Histograms

- **Interval and ratio features (Appendix B, Figure 3, Figure 4):** *age* and *thalach* show normal or slightly skewed distributions. In contrast, *chol* and *trestbps* are right-skewed due to the presence of extreme outliers (e.g., cholesterol ≥500 mg/dL and BP ≥180-200 mmHg); *oldpeak* is left-skewed, indicating that most patients have values clustered close to 0.
- **Nominal and ordinal features (Appendix B, Figure 4):** the distribution of *sex* confirms a significant imbalance, with approximately 70% of the cohort being male. *Thal*, *slope*, and *cp* have dominant categories (e.g. 50% of *thal* values are 2). For *fbs*, *restecg*, and *exang*, the data concentrate heavily in a few categories, showing clear modes.

5.2. Scatter Plots (pairplots)

- **Continuous physiological parameters (*thalach*, *oldpeak*, *chol*, *trestbps*, *age*; Appendix C, Figure 5):** *thalach* decrease with *age* and diseased patients (red) tend to cluster at lower *thalach* values at higher ages. *Oldpeak* separates classes clearly, with higher values in diseased individuals. Both *chol* and *trestbps* show strong overlap and weak correlation with disease. Age is slightly higher among patients with the disease.
- **Symptoms and effort (*cp*, *exang*, *slope*, *oldpeak*, *thalach*; Appendix C, Figure 6):** *cp* concentrates healthy individuals (blue) in the lower pain categories. *Exang* distinguishes between groups, particularly enhanced when combined with *oldpeak*. *Slope* is also an effective differentiator, alongside *oldpeak*. *Oldpeak* itself shows a marked increase in patients with the disease. *Thalach* again shows its inverse relationship with cardiovascular risk.
- **Anatomical and functional makers (*ca*, *thal*, *slope*, *oldpeak*; Appendix C, Figure 7):** *ca* is a highly predictive variable (higher vessel counts appear almost exclusively in diseased patients). *Thal* indicates that reversible or fixed defects are more common in diseased patients. *Slope* demonstrates a strong association with disease when flat or descending slopes. *Oldpeak* consistently shows higher values in diseased patients.
- **Classic risk factors (*sex*, *age*, *trestbps*, *chol*, *fbs*; Appendix C, Figure 8):** *sex* is strongly associated with the disease, with a clear male predominance. *Age* trends upward in disease presence. In constrast, *trestbps* and *chol* exhibit substantial overlap and low discriminative power. Finally, *fbs* offers minimal separation and limited predictive value.

5.3. Bivariate analysis (Feature vs Target)

Bivariate analysis (**Appendix C**, **Figure 9**) indicates that *age* is slightly higher and *thalach* notably lower in diseased patients, whereas *oldpeak* shows a marked elevation, reflecting exercise-induced ischemia. *Trestbps* and *chol* overlap strongly between classes, reducing discriminative utility. *Sex* displays a male predominance among CVDs cases, while *cp* and *exang* separate groups effectively, with typical angina and exercise-induced angina more common in affected individuals.

Anatomical and functional markers (*ca*, *thal*, and *slope*) offer the strongest predictive value, as diseased patients cluster in higher vessel counts, abnormal *thal* categories, and flat/downsloping ST-segment slopes. *Fbs* and *restecg* provide minimal class distinction.

Overall, *cp*, *exang*, *thalach*, and *oldpeak* emerge as the most informative features, aligning with established clinical evidence linking them to cardiovascular disease.

5.4. Conclusion summary

The graphical analysis confirms that ischemic and function related variables (*oldpeak, thalach, cp, exang, slope,* and *ca*) achieve the clearest class separation, consistent with prior evidence on key CVDs predictors [2], [3]. Traditional risk factors such as *chol, trestbps, fbs,* and *restecg* display substantial overlap, reflecting heterogeneity and measurement limitations of public CVDs datasets [12]. These findings align with broader epidemiological trends described by the American Heart Association and with performance patterns observed in contemporary multivariable risk models [1].

6. Correlation

The following **Table 2**, contains the most correlated attributes in the correlation matrix (**Figure 10, Appendix D**). Additional parallel coordinates plots (**Figure 11**, **Figure 12**, **Figure 13**, **Figure 14**, **Figure 15**; **Appendix D**) provide a complementary visualization of these relationships. The *r* denotes the Pearson Correlation Coefficient (PCC).

Feature Pair	r	Interpretation
exang/target	+0.44	Exercise-induced angina is strongly associated with disease presence
cp/target	+0.43	Type of chest pain is a strong positive predictor of CVD
thalach/target	-0.42	Lower maximum heart rate indicates higher disease likelihood
oldpeak/target	+0.40	ST segment depression correlates with ischemic cardiac response
slope/target	+0.35	Flat or descending ST slope suggests higher cardiovascular risk
exang/oldpeak	+0.31	Angina is associated with greater ST depression
cp/thalach	+0.31	Patients with typical angina tend to have lower heart rates

Table 2. *Correlation summary of feature pairs*

The strongest correlations involve clinically meaningful features, such as chest pain type (*cp*), exercise-induced angina (*exang*), ST-segment depression (oldpeak), maximal heart rate (*thalach*), and the slope of the ST segment (*slope*). These variables capture core physiological responses to myocardial ischemia and reduced exercise tolerance, which explains their consistent association with disease presence [1].

Their behaviour aligns with classical diagnostic patterns observed in exercise-ECG assessment, where anginal symptoms and ST-segment deviations reflect ischemic burden, while lower *thalach* denotes impaired chronotropic competence [1], [3]. Consistently, these same attributes emerge as high-value predictors in AI-based cardiovascular models, where ischemic ECG signatures and functional performance indicators constitute central inputs for robust risk discrimination [1], [2], [12].

7. Dataset modifications and feature transformation

No missing values were found in the dataset, and duplicate removal was not performed due to the absence of patient identifier. Outlier were screened by K-Nearest Neighbours and Density neighbours count (**Appendix E**, **Figure 16**, **Figure 17**, **Figure 18, Figure 19**), resulting in 104 flagged observations. Several strategies were applied to mitigate it, including:

1- Normalization (**Appendix E**, **Figure 20**, **Figure 23**): this transformation standardised value ranges without altering the overall distribution of interval and ratio attributes, such as *age*, *chol*, and *trestbps*.
2- Logarithmic transformation (**Appendix E**, **Figure 24**, **Figure 25**, **Figure 26**, **Figure 27**, **Figure 28**, **Figure 29**): applied to right-skewed variables, particularly *chol* and *oldpeak*, effectively reducing skewness and compressing extreme tails.
3- Removal of the outliers **(Appendix E, Figure 20, Figure 21**): used to detect implausible measurements, including cholesterol (≥500 mg/dL) and blood pressure (≥180-200 mm Hg).
4- Outlier value capping (**Appendix E**, **Figure 20**, **Figure 22**): extreme values were capped at selected percentiles to limit their influence while preserving dataset size and structure.

Additional, discretization and encoding were applied to selected variables to support feature engineering and categorical analysis. These preprocessing procedures are consistent with the transformations commonly reported in cardiovascular machine learning studies, where normalization, log-scaling, outlier handling and categorical encoding help manage measurement variability [3], [12]. Prior work also confirms that these transformations often improve model behaviour, while large epidemiological sources, such as those referenced by the AHA, emphasise the need strict standardization across heterogeneous cohorts [1], [10].

8. PCA and feature selection

8.1. Principal Component Analysis

In the cardiovascular modelling literature, Principal Component Analysis (PCA) is acknowledged as a dimensionality-reduction technique but is generally confined to exploratory analysis rather than core predictive modelling. [2]. A recent systematic review on machine-learning applications for acute coronary syndrome (ACS) reports PCA as one of the methods used to evaluate comorbidity patterns and their relationship with myocardial infarction risk, as described in the RISTI technical review on CVDs identification [12].

Evidence indicates that while PCA (**Appendix F**, **Figure 30, Figure 31**, **Figure 32**) does not provide substantial dimensionality reduction in the dataset. However, the first components capture clinically meaningful axes reflecting ischemic burden and traditional risk factors, reinforcing its exploratory utility in line with previous ACS analyses.

8.2. Feature Selection

Filter-based methods consistently identified *oldpeak*, *thalach*, *cp*, *ca*, and *thal* as the strongest predictors, while *fbs* and *restecg* showed minimal relevance. This matches cardiovascular machine learning workflows, where removing low-information variables is essential. Shu et al. highlight feature selection and projection as key dimensionality-reduction steps in CVDs classification [2], and the RISTI review documents PCA and related analytical tools for refining ACS predictors [12]. Sun et al. similarly emphasizes reducing high-dimensional clinical data to extract meaningful patterns [3].

Wrapper methods produced model-dependent subsets: linear RFE favoured strong linear predictors (*sex*, *cp*, *thalach*, *oldpeak*, *ca*, and *thal*), whereas Random Forest SFS prioritized features improving non-linear performance (*age*, *cp*, *trestbps*, *chol*, *thalach*, and

oldpeak). This reflects machine learning practice in cardiology, where feature refinement adapts to model structure. The RISTI review notes PCA-based dimensionality reduction in ACS modelling [12], while Shu *et al.* describe feature selection as necessary for handling heterogeneous CVDs populations [2]. Sun *et al.* also stress focusing on the most informative features to improve predictive accuracy [3].

9. Modelling application

Table 3 presents the models performance obtained by applying the unmodified dataset to a Decision Tree (DT), using GridSearch (GS) optimization and cost-complexity pruning, and to a Deep Learning (DL) model [8], [9]. Both are widely adopted in cardiovascular machine-learning studies, and they were evaluated using standard classification metrics, including accuracy, the Area Under the Curve (AUC), the precision, the recall or sensitivity and the F1-score [2], [10].

Model	GS - DT	Prunned - DT	DL
Accuracy	88.3%	88.3%	94.6%
AUC	96.8%	96.8%	98.9%
Precision	91.0%	91.0%	95.0%
Recall	86.0%	86.0%	94.0%
F1-score	88.0%	88.0%	95.0%

Table 3. *Standard classification metrics based on Decision Tree models and Deep Learning performances*

The optimized Decision Tree achieved 88.3% accuracy and the highest AUC (96.8%), with splits aligned with established cardiovascular risk factors highlighted in recent AHA surveillance reports [1]. These results align with prior evidence showing that tree-based methods capture clinically relevant threshold interactions in structured datasets [10]. The Receiver Operating Characteristic (ROC) curve and the structure of the DT is shown in **Appendix G**, **Figure 33** and **Figure 34**.

The Pruned Decision Tree achieved equivalent performance (88.3% accuracy and 96.8% AUC), while offering a more compact and interpretable structure. Pruning reduced noise-driven branches while preserving the most informative clinical features, in line with findings from contemporary cardiovascular AI reviews emphasizing the importance of model simplicity for clinical interpretability [3]. The pruned model is depicted in **Appendix G**, **Figure 35** and **Figure 36**.

Following feature standardization, a fully connected neural network was trained incorporating LeakyReLU activation functions and L2 regularization, optimized with Adam/AdamW and early stopping. This Deep Learning model showed substantial improvement compared to earlier configurations, whose initial accuracies were 72-75%. The model achieved 94.6% accuracy and an AUC of 98.9%, outperforming the decision tree approaches in terms of discrimination. Despite the strong performance, results should be interpreted with caution due to dataset size and require further validation. Its behaviour is consistent with cardiovascular AI reviews, which report that appropriate preprocessing and regularization can substantially enhance neural network performance in structured clinical datasets [3]. Training behaviour and classification performance are shown in **Appendix G**, **Figure 37** and **Figure 38**; the first figure one is the individual ROC Curve of DL, and the second one is a comparison of DT and DL ROC Curves.

Taken together, the Deep Learning approach provides the strongest discriminative performance on this dataset, while decision tree-based models remain highly competitive and offer superior interpretability, which is particularly relevant for transparent clinical decision-support applications.

10. Conclusions

EDA shows that ischemic and functional variables (*oldpeak*, *thalach*, *cp*, *exang*, *slope*, and *ca*) provide the clearest class separation, whereas traditional risk factors display limited discriminative power due to dataset heterogeneity. Identified outliers were managed (removed or capped) but excluded from the modelling process. PCA confirmed its value as an exploratory tool but offered limited dimensionality reduction, while feature selection methods consistently reinforced the relevance of the same key predictors. In the modelling stage, deep learning achieved the best performance after appropriate feature standardization and regularization, while decision tree approaches provide interpretability. Overall, these results indicate that this dataset constitutes a robust and reusable foundation for further cardiovascular analytical and modelling studies.

11. Bibliography

[1] S. S. Martin *et al.*, "2025 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association," *Circulation*, vol. 151, no. 8, Feb. 2025, doi: 10.1161/CIR.0000000000001303.

[2] S. Shu, J. Ren, and J. Song, "Clinical Application of Machine Learning-Based Artificial Intelligence in the Diagnosis, Prediction, and Classification of Cardiovascular Diseases," *Circulation Journal*, vol. 85, no. 9, p. CJ-20-1121, Aug. 2021, doi: 10.1253/circj.CJ-20-1121.

[3] X. Sun, Y. Yin, Q. Yang, and T. Huo, "Artificial intelligence in cardiovascular diseases: diagnostic and therapeutic perspectives," *Eur J Med Res*, vol. 28, no. 1, p. 242, Jul. 2023, doi: 10.1186/s40001-023-01065-y.

[4] D. Lapp, "Heart Disease Dataset," 2025, *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

[5] R. Hernanz, "Projects_II_HeartDataset," 2025. [Online]. Available: https://github.com/RaquelHernanz/Projects_II_HeartDataset

[6] N. Developers, “NumPy (Version 1.26),” 2024. [Online]. Available: https://numpy.org/

[7] pandas D. Team, “pandas: Python Data Analysis Library (Version 2.2),” 2024. [Online]. Available: https://pandas.pydata.org/

[8] T. Developers, “TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems (Version 2.15),” 2024. [Online]. Available: https://www.tensorflow.org/

[9] scikit-learn Developers, “scikit-learn: Machine Learning in Python (Version 1.4),” 2024. [Online]. Available: https://scikit-learn.org/

[10] R. Nadarajah *et al.*, “Prediction models for heart failure in the community: A systematic review and meta-analysis,” *Eur J Heart Fail*, vol. 25, no. 10, pp. 1724–1738, Oct. 2023, doi: 10.1002/ejhf.2970.

[11] D. Levano-Rodriguez, F. Cerdan-Leon, R. Gonza-Morales, G. Rojas-Peña, R. Maldonado-Fernandez, and D. Davila-Valencia, “Predicción de Enfermedad Cardiovascular Utilizando el Algoritmo de Machine Learning ExtraTreesClassifier: Predicting Cardiovascular Disease Using the ExtraTreesClassifier Machine Learning Algorithm,” *Scientia*, vol. 28, no. 1, 2025.

[12] J. Mardini - Bovea *et al.*, “Modelos de identificación de enfermedades cardiovasculares implementando técnicas de aprendizaje máquina: una revisión sistemática de la literatura,” *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, no. 53, pp. 87–105, Mar. 2024, doi: 10.17013/risti.53.87-105.

Appendix

The appendix section serves as a repository of supplementary visual and tabular materials referenced throughout the EDA.

Appendix A - Dataset feature specifications

Feature	Description	Values
age	Age of the patients (years)	Numeric Interval (29-77)
sex	Gender of the patient	Nominal (0 = Female, 1 = Male)
cp	Chest pain type	Ordinal (0 = Typical angina 1 = Atypical angina 2 = Non-anginal pain 3 = Asymptomatic)
trestbps	Resting blood pressure (mmHg)	Numeric Interval
chol	Serum cholesterol (mg/dl)	Numeric Interval
fbs	Fasting blood sugar > 120 mg/dl	Nominal (0 = False, 1 = True)
restecg	Resting electrocardiographic results	Ordinal (0 = Normal 1 = ST-T wave abnormality 2 = Left ventricular hypertrophy)
thalach	Maximum heart rate achieved (bpm)	Numeric Interval
exang	Exercise-induced angina	Nominal (0 = No, 1 = Yes)
oldpeak	ST segment depression induced by exercise relative to rest (mm)	Numeric Ratio
slope	Slope of the peak exercise ST segment	Ordinal (0 = Upsloping 1 = Flat 2 = Downsloping)
ca	Number of major vessels coloured by fluoroscopy	Numeric Interval (0 – 4)
thal	Thalassemia type	Ordinal (0 = Unknown 1 = Fixed defect 2 = Normal 3 = Reversible defect)
target	Heart disease presence **(Analysis Target)**	Nominal (0 = No heart disease 1 = heart disease)

Table 4. *Feature characteristics from the dataset*

Appendix B - Descriptive statistics and distribution analysis

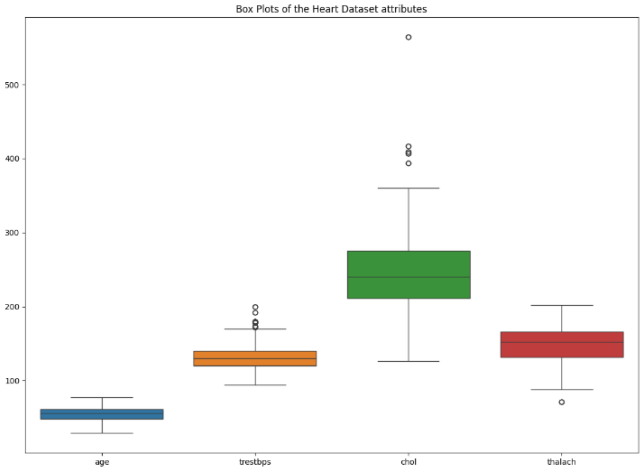

Figure 1. *Box Plots of age (years), trestbps (mmHg), chol (mg/dl) and thalach (bpm)*

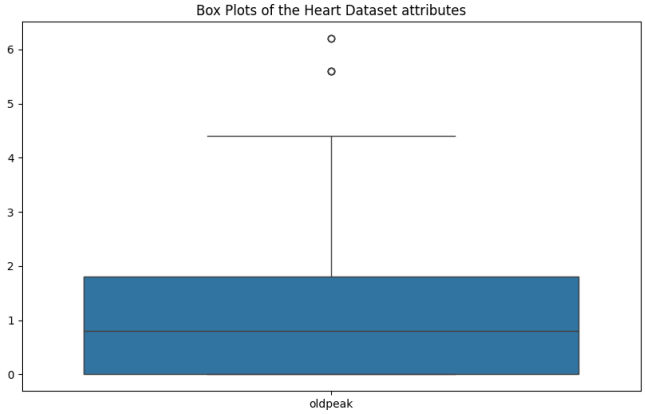

Figure 2. *Box Plot of oldpeak (mm) feature*

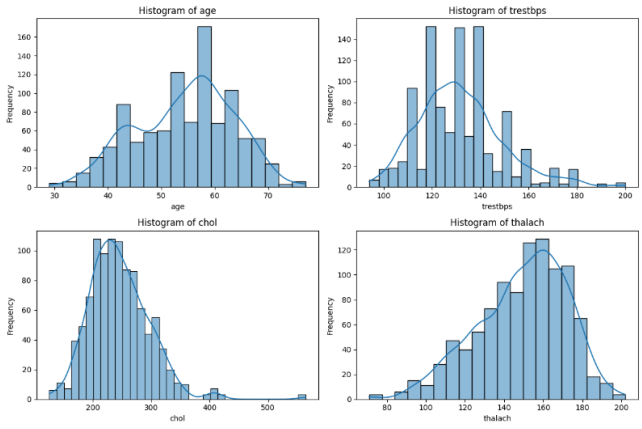

Figure 3. *Histograms of age (years), trestbps (mmHg), chol (mg/dl) and thalach (bpm)*

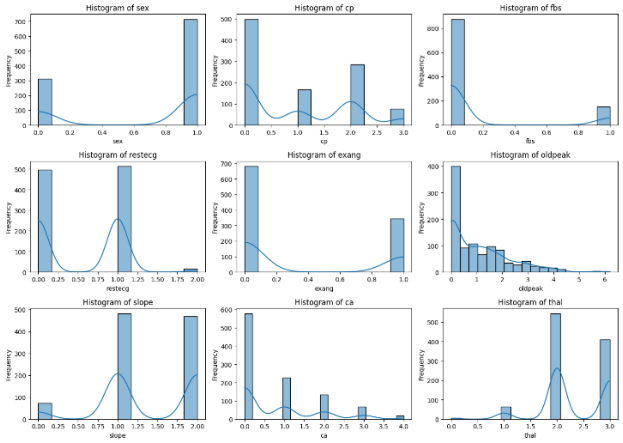

Figure 4. *Histograms of sex, cp, fbs, restecg, exang, oldpeak (mm), slope, ca and thal*

Appendix C - Bivariate and Multivariate Scatter Analysis

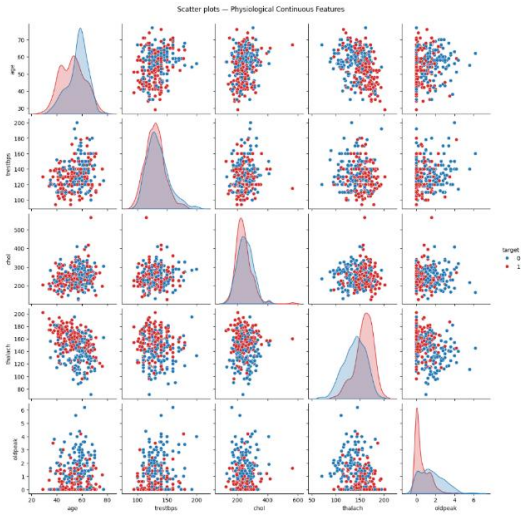

Figure 5. *Scatter plots of Continuous Physiological parameters (thalach, oldpeak, chol, trestbps)*

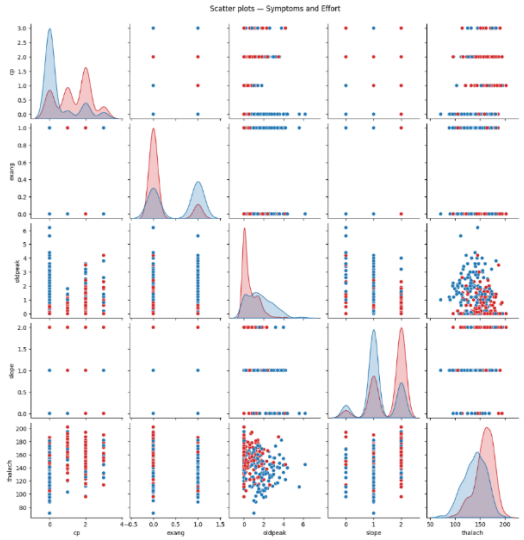

Figure 6. *Scatter plot of Symptoms and effort (cp, exnag, slope, oldpeak, thalach)*

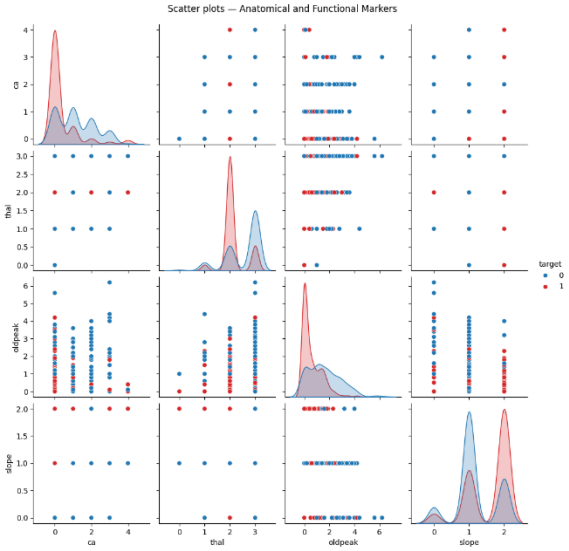

Figure 7. *Scatter plot of Anatomical and functions makers (ca, thal, slope, oldpeak)*

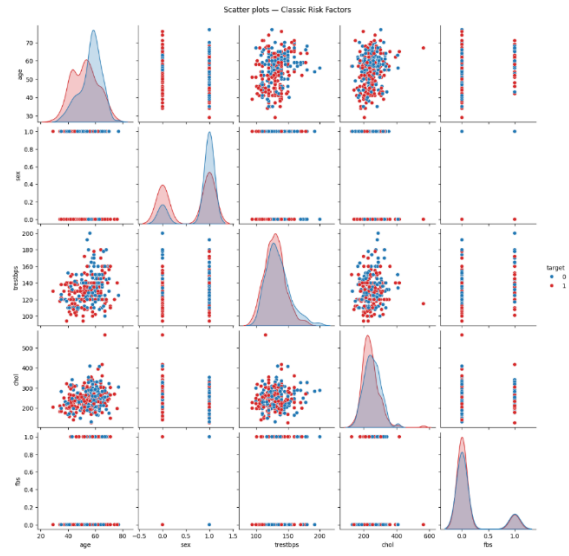

Figure 8. *Scatter plot of Classic risk factors (sex, age, trestbps, chol, fbs)*

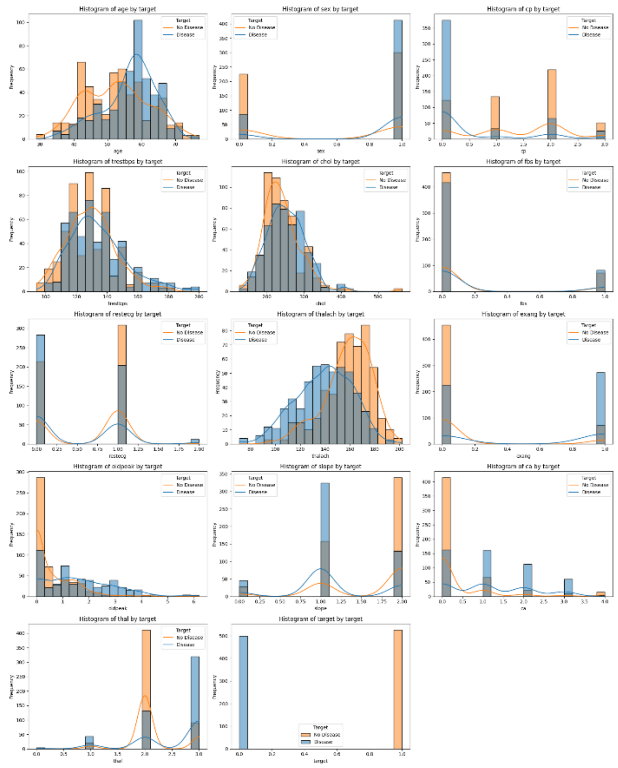

Figure 9. *Target vs Feature plots (orange corresponds to "No disease" and blue "Disease")*

Appendix D - Correlation and Parallel Coordinates (PC) analysis

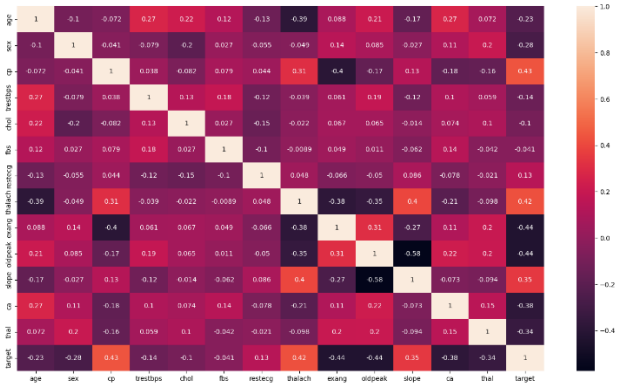

Figure 10. *Correlation matrix*

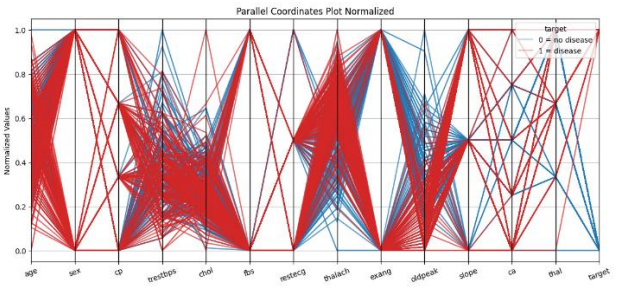

Figure 11. *Normalized Parallel coordinates (blue as" No disease" and red "Disease")*

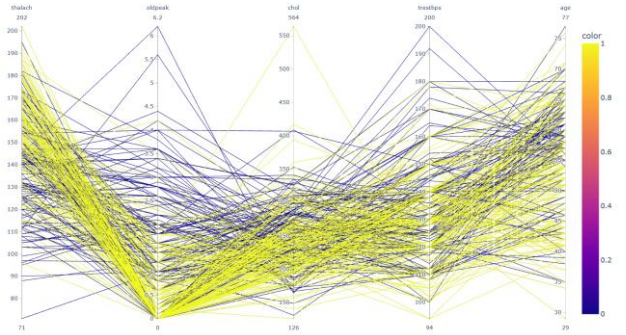

Figure 12. *PC of Continuous Physiological parameters (thalach, oldpeak, chol, trestbps)*

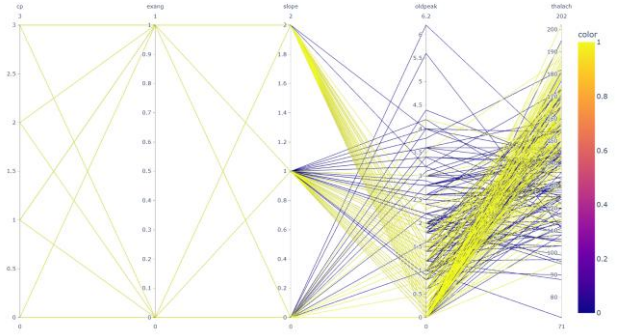

Figure 13. *PC of Symptoms and effort (cp, exnag, slope, oldpeak, thalach)*

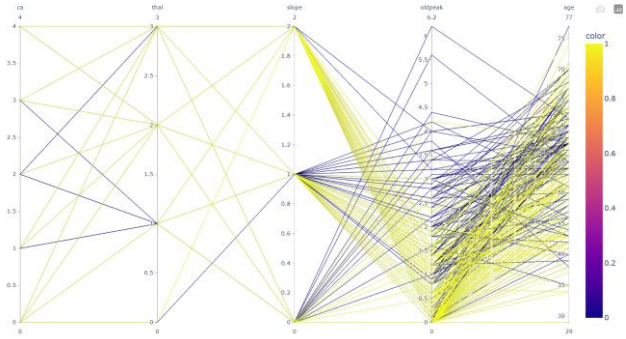

Figure 14. *PC of Anatomical and functions makers (ca, thal, slope, oldpeak)*

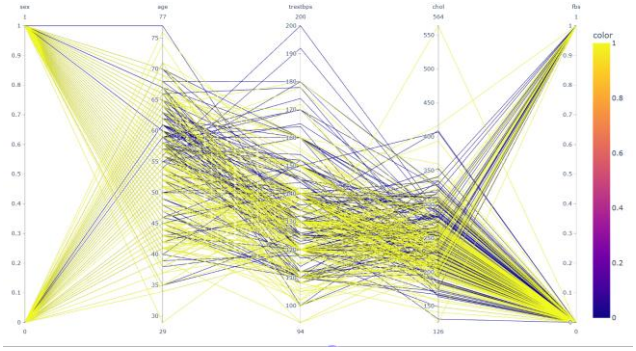

Figure 15. *PC of Classic risk factors (sex, age, trestbps, chol, fbs)*

Appendix E - Data transformation visualization

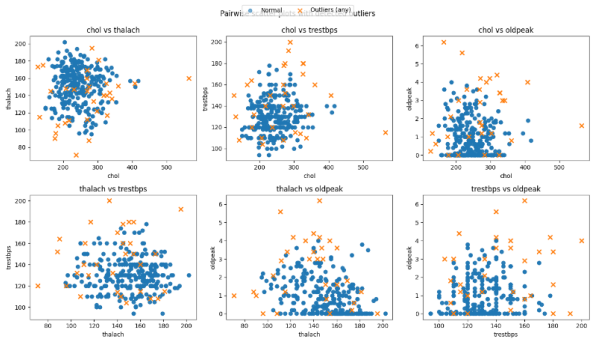

Figure 16. *Pairwise scatter plots with outliers highlighted*

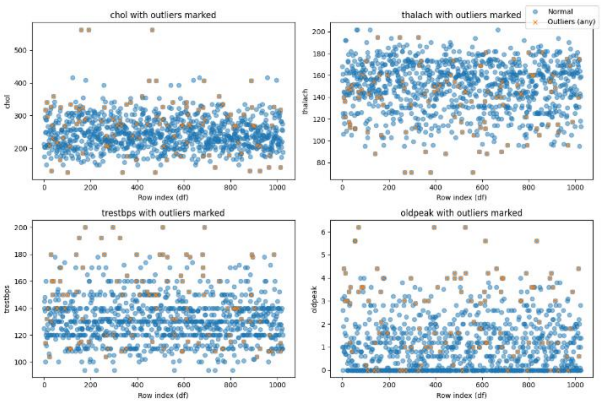

Figure 17. *Feature-wise index plots highlighting outliers*

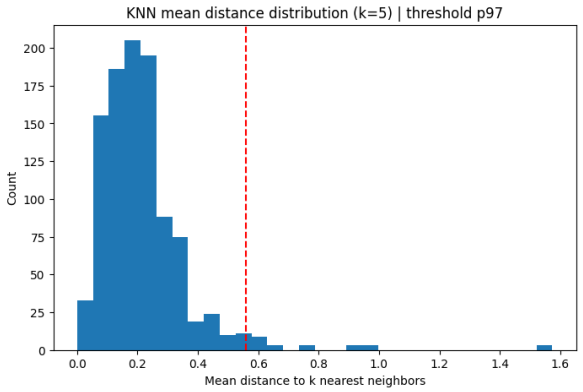

Figure 18. *K-Nearest Neighbours plot with K= 5 and threshold p97*

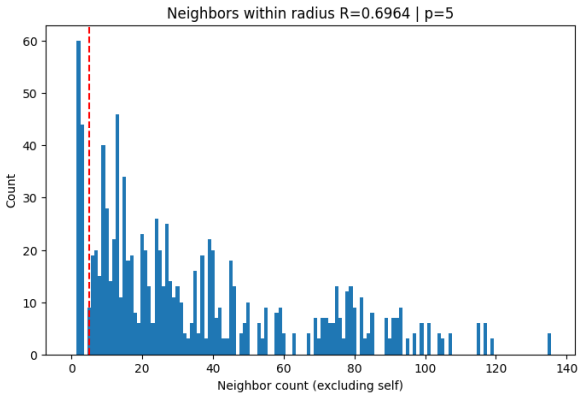

Figure 19. *Density neighbours count plot with R=0.6964 and p = 5*

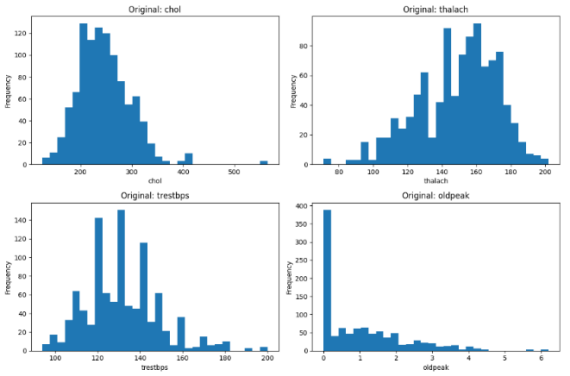

Figure 20. *Original chol (mg/dl), thalach (bpm), trestbps (mmHg), oldpeak (mm) histogram plot*

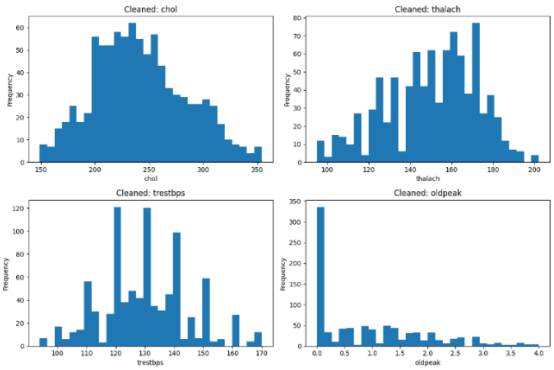

Figure 21. *Chol (mg/dl), thalach (bpm), trestbps (mmHg), oldpeak (mm) histograms after outlier removal*

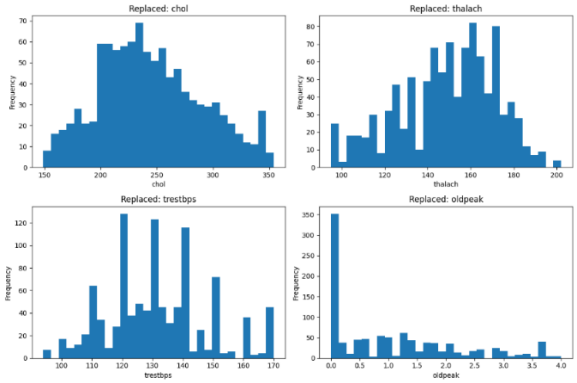

Figure 22. *Chol (mg/dl), thalach (bpm), trestbps (mm Hg), oldpeak (mm) after outlier replacement*

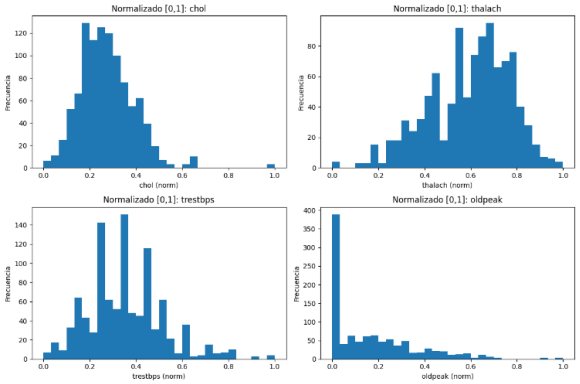

Figure 23. *Chol (mg/dl), thalach (bpm), trestbps (mmHg), oldpeak (mm) feature histograms after normalization*

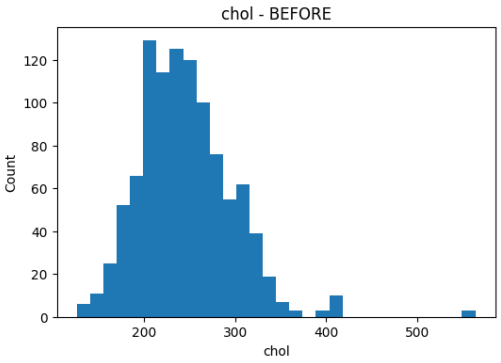

Figure 24. *Chol (mg/dl) feature histogram*

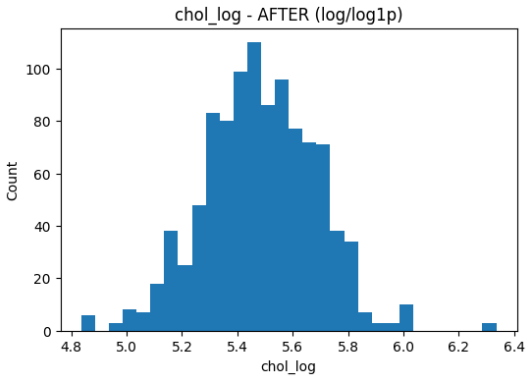

Figure 25. *Chol (mg/dl) after log-transformation*

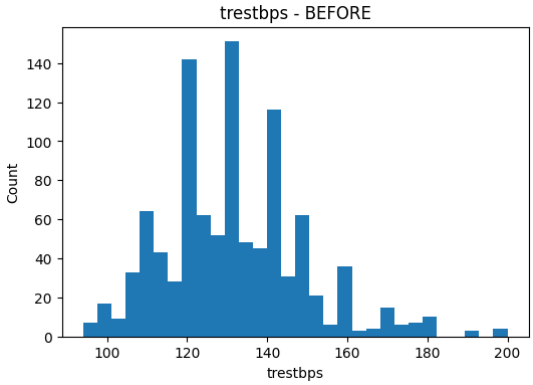

Figure 26. *Trestbps (mmHg) feature histogram*

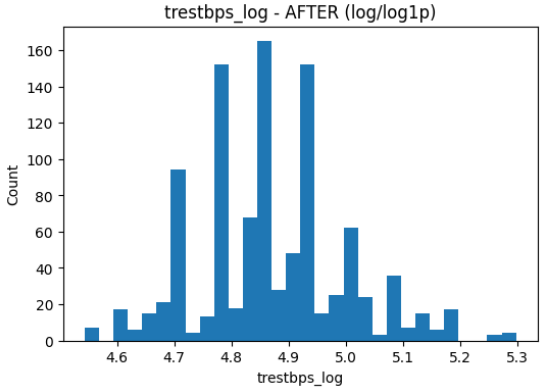

Figure 27. *Trestbps (mmHg) after log-transformation*

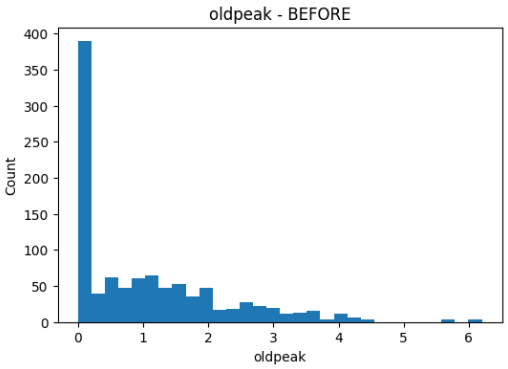

Figure 28. *Oldpeak (mm) feature histogram*

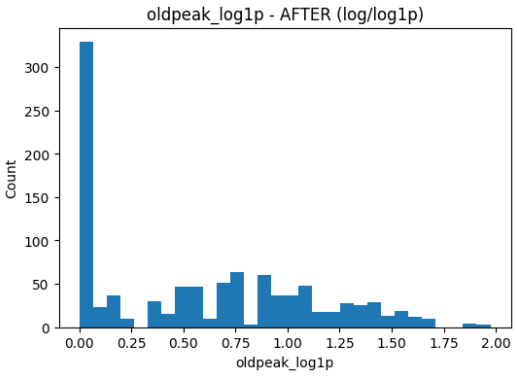

Figure 29. *Oldpeak (mm) after log-transformation*

Appendix F - Principal Component Transformation

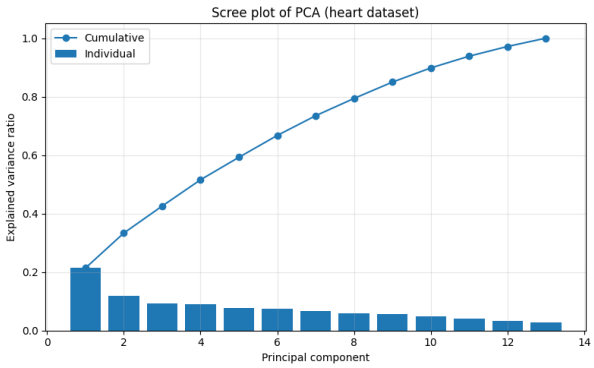

Figure 30. *Scree plot of PCA*

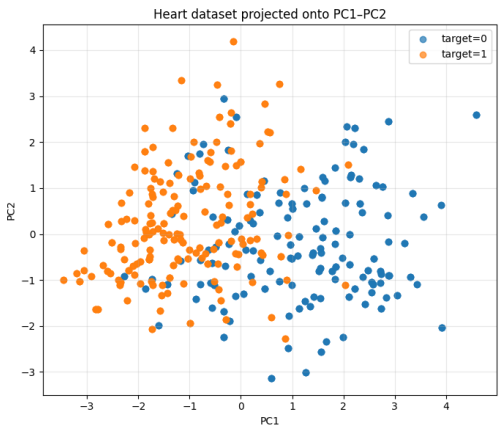

Figure 31. *Heart dataset projection on PC1-PC2*

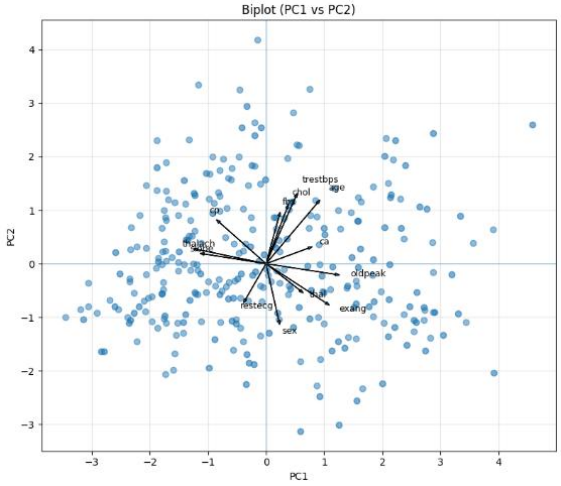

Figure 32. *Biplot (PC1 vs PC2)*

Appendix G - Classification Model Visualizations

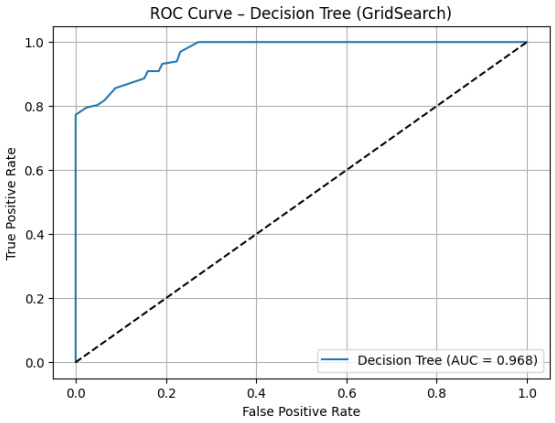

Figure 33. *ROC Curve of Decision Tree with GridSearch*

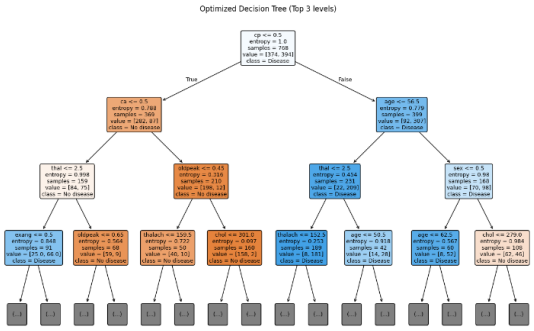

Figure 34. *Decision Tree with GridSearch*

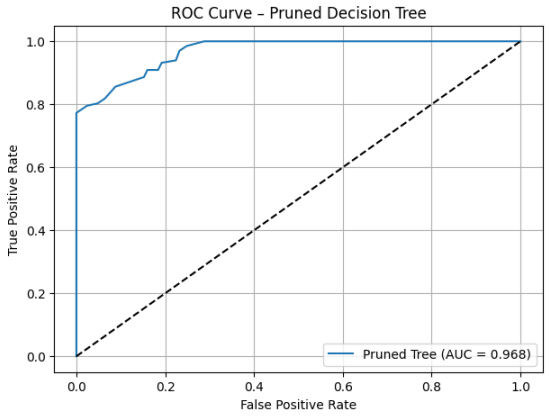

Figure 35. *ROC Curve of Pruned Decision Tree*

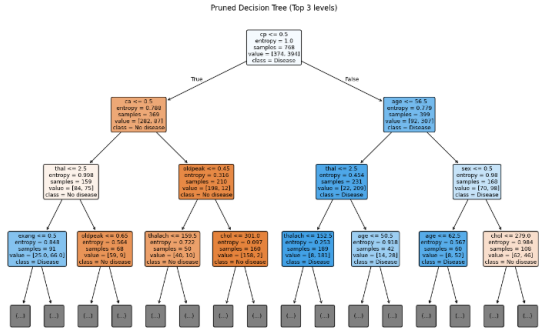

Figure 36. *Pruned Decision Tree*

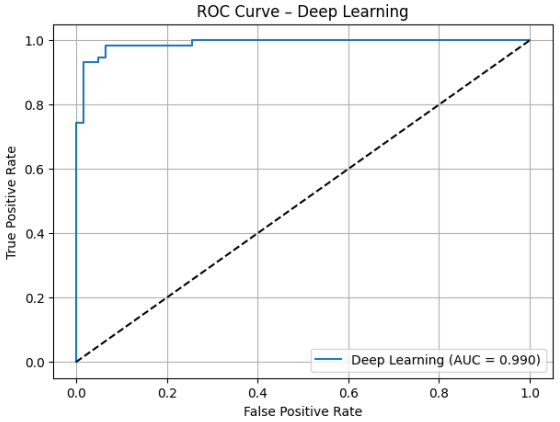

Figure 37. *ROC Curve of Deep Learning*

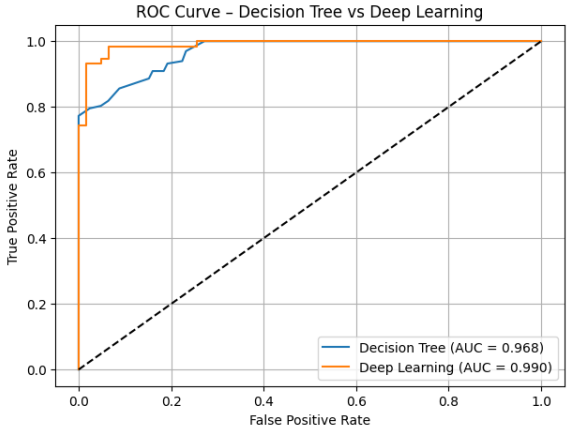

Figure 38. *ROC Curve of Decision Tree vs Deep Learning*