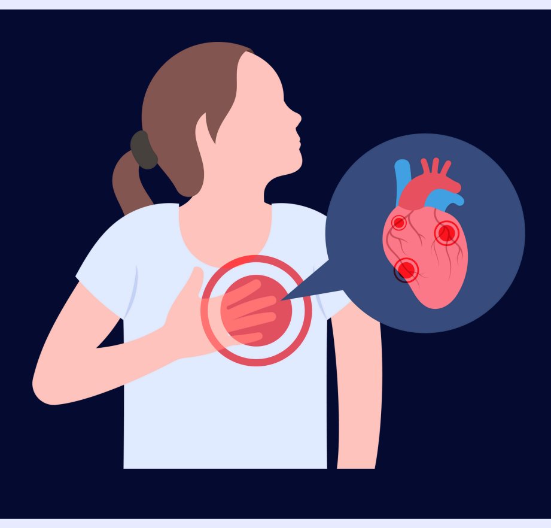
HEART DISEASE PREDICTION USING END-TO-END MACHINE LEARNING PROCESS

Authors:
Joycelin (2702213713), Kelvin Fardiman (27023364225), Kyan Pratama Lolobua (2702241925), Nusantara Kusuma (2702328636)
Computer Science and Statistics - School of Computer Science - BINUS University



INTRODUCTION

Cardiovascular diseases (CVDs) remain the leading cause of death globally, claiming approximately 17.9 million lives each year, accounting for 31% of all deaths worldwide. Notably, four out of five CVD-related deaths result from heart attacks and strokes, with nearly one-third occurring prematurely in individuals under the age of 70. Heart disease, a key consequence of CVDs, can often lead to heart failure if not detected early. The dataset used in this study comprises 11 clinical features relevant to predicting the likelihood of heart disease.

This project explores an end-to-end machine learning pipeline for predicting heart disease using structured health data. The study includes data preprocessing, feature engineering, model training, and performance evaluation across multiple algorithms to identify the most effective approach.

DATASET

This project utilizes the Heart Failure Prediction Dataset, curated by Fedesoriano on Kaggle (September 2021). The dataset was constructed by merging five reputable heart disease datasets (Cleveland, Hungarian, Switzerland, Long Beach VA, and Stalog) from the UCI Machine Learning Repository into a single, comprehensive source. All records are standardized across 11 common clinical features, resulting in 918 unique patient observations. This makes it the largest publicly available heart disease dataset for research purposes to date.

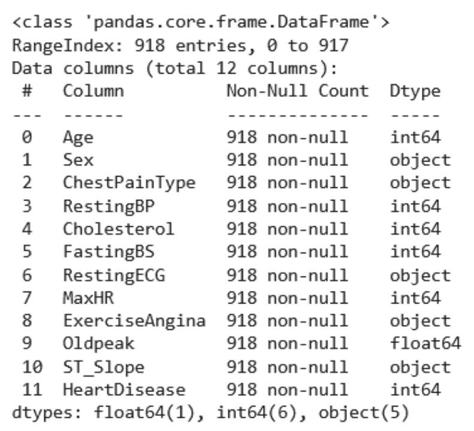
Attributes Description:

- Age: Patient’s age (years)
- Sex: Biological sex [M = Male, F = Female]
- ChestPainType: Chest pain type [TA = Typical Angina, ATA = Atypical Angina, NAP = Non-Anginal Pain, ASY = Asymptomatic]
- RestingBP: Resting blood pressure (mm Hg)
- Cholesterol: Serum cholesterol (mg/dl)
- FastingBS: Fasting blood sugar > 120 mg/dl [1 = True, 0 = False]
- RestingECG: Resting electrocardiogram results [Normal, ST, LVH]
- MaxHR: Maximum heart rate achieved (60–202)
- ExerciseAngina: Exercise-induced angina [Y = Yes, N = No]
- Oldpeak: ST depression induced by exercise
- ST_Slope: Slope of the peak exercise ST segment [Up, Flat, Down]
- HeartDisease: Output class [1 = Heart Disease, 0 = Normal]

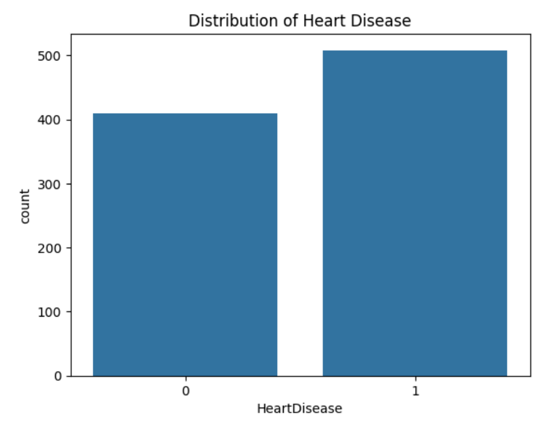
END-TO-END MACHINE LEARNING PROCESS

Data Cleaning & Exploratory Data Analysis

- Imported the dataset (heart.csv) with 918 records and 12 columns.
- Checked for missing values: all columns have zero nulls.
- Verified data type: all categorical and numerical features are properly formatted.
- Outliers were identified in numerical features, but no modifications were made in order to preserve the integrity of the medical data.
- Confirmed class distribution:
  508 patients with heart disease (label = 1)
  410 patients without (label = 0)

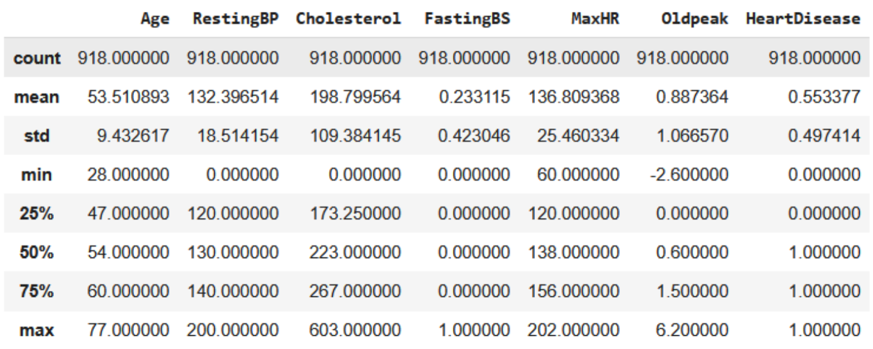


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             918 non-null    int64
 1   Sex             918 non-null    object
 2   ChestPainType   918 non-null    object
 3   RestingBP       918 non-null    int64
 4   Cholesterol     918 non-null    int64
 5   FastingBS       918 non-null    int64
 6   RestingECG      918 non-null    object
 7   MaxHR           918 non-null    int64
 8   ExerciseAngina  918 non-null    object
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    object
 11  HeartDisease    918 non-null    int64
dtypes: float64(1), int64(6), object(5)
```



Dataset Information

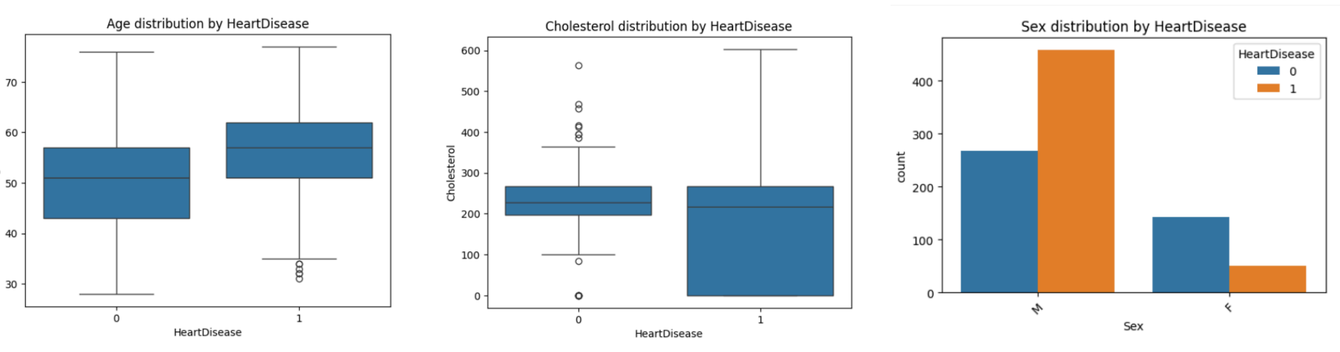
- Generated descriptive statistics for numerical features using .describe().



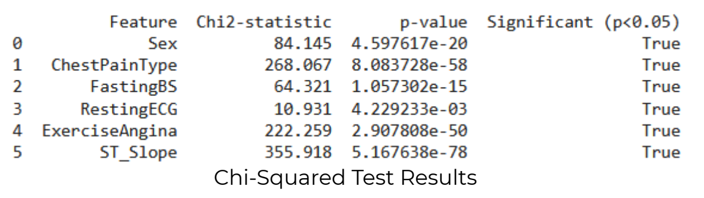
	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
count	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000
mean	53.510893	132.396514	198.799564	0.233115	136.809368	0.887364	0.553377
std	9.432617	18.514154	109.384145	0.423046	25.460334	1.066570	0.497414
min	28.000000	0.000000	0.000000	0.000000	60.000000	-2.600000	0.000000
25%	47.000000	120.000000	173.250000	0.000000	120.000000	0.000000	0.000000
50%	54.000000	130.000000	223.000000	0.000000	138.000000	0.600000	1.000000
75%	60.000000	140.000000	267.000000	0.000000	156.000000	1.500000	1.000000
max	77.000000	200.000000	603.000000	1.000000	202.000000	6.200000	1.000000

Descriptive Statistics

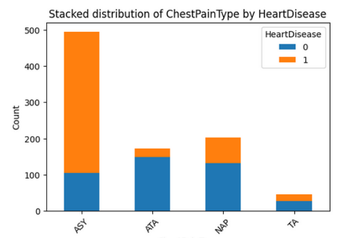
- Created distribution plots to examine how features like Age, Cholesterol, and Sex vary by heart disease status.



- Chi-Squared tests showed that all categorical features, including ChestPainType, had significant associations with HeartDisease (e.g., ASY was highly prevalent in positive cases).

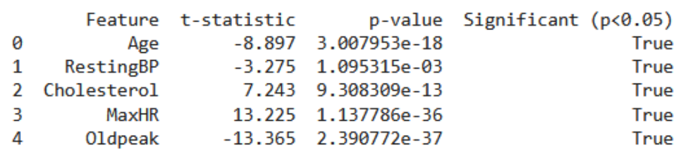


```
          Feature  Chi2-statistic       p-value  Significant (p<0.05)
0             Sex          84.145  4.597617e-20                  True
1   ChestPainType         268.067  8.083728e-58                  True
2       FastingBS          64.321  1.057302e-15                  True
3      RestingECG          10.931  4.229233e-03                  True
4  ExerciseAngina         222.259  2.907808e-50                  True
5        ST_Slope         355.918  5.167638e-78                  True
```



Chi-Squared Test Results

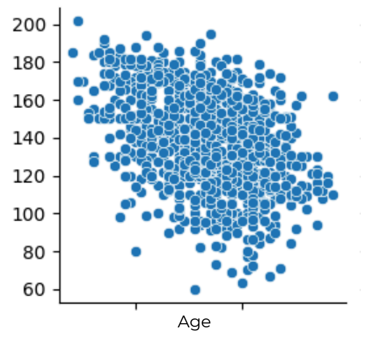
- Conducted t-tests on numerical features, showing statistically significant differences between classes (with vs. without heart disease).



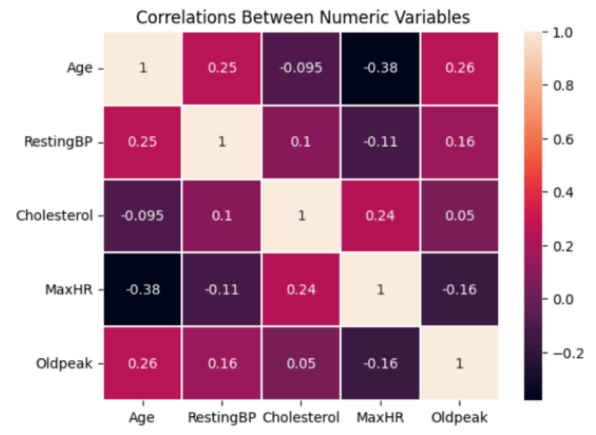
```
       Feature  t-statistic       p-value  Significant (p<0.05)
0          Age       -8.897  3.007953e-18                  True
1    RestingBP       -3.275  1.095315e-03                  True
2  Cholesterol        7.243  9.308309e-13                  True
3        MaxHR       13.225  1.137786e-36                  True
4      Oldpeak      -13.365  2.390772e-37                  True
```

T-Test Results

- A scatterplot matrix and correlation heatmap were used to visualize relationships and detect potential multicollinearity between numerical features.



Scatter Plot of MaxHR VS Age

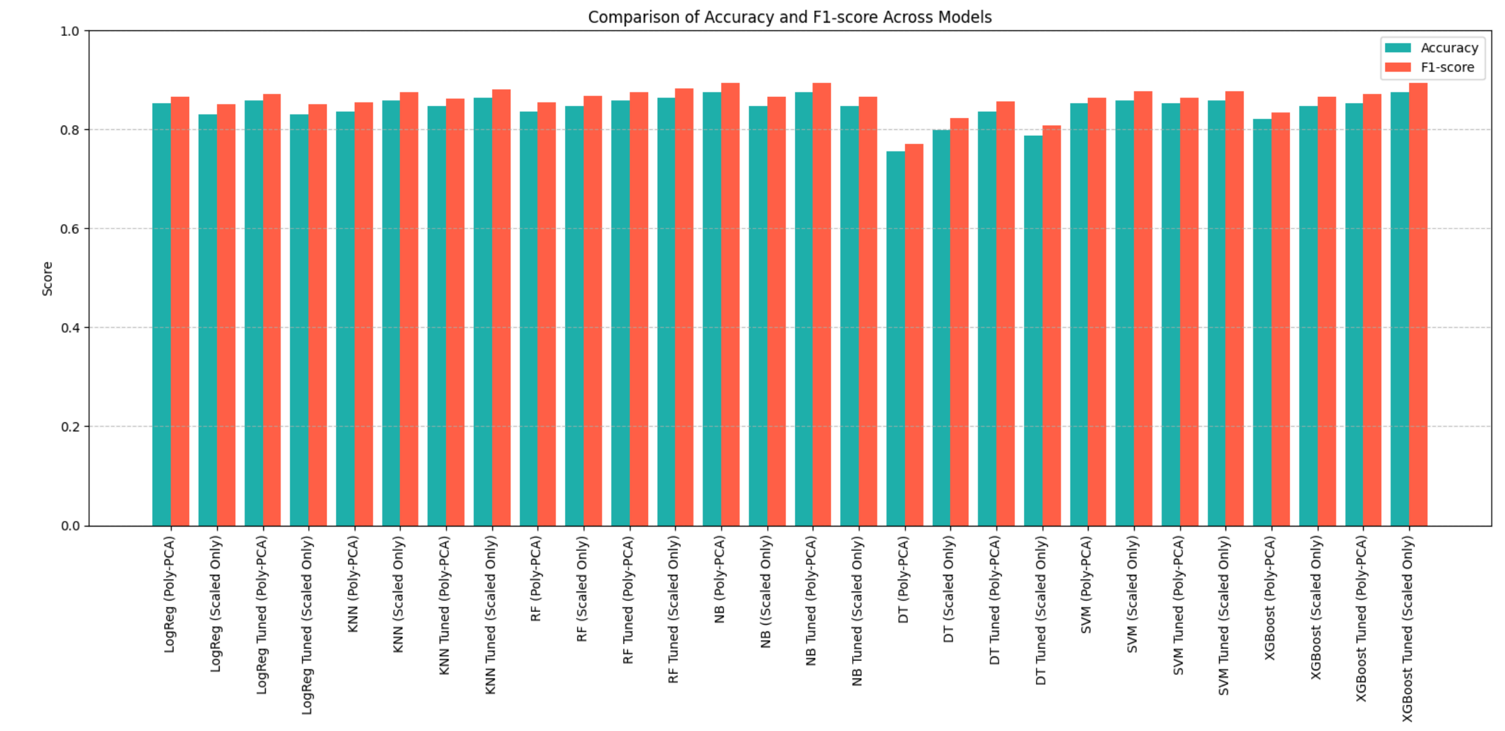


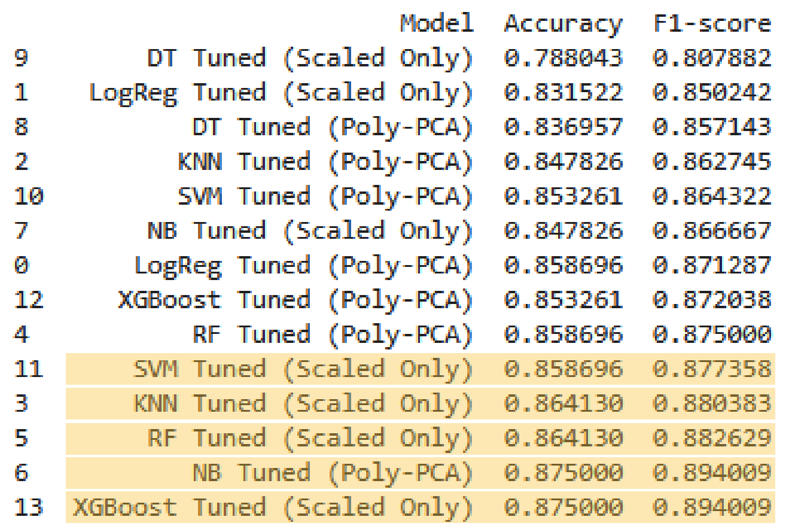
Feature Engineering

- One-hot encoding applied to categorical features (Sex, ChestPainType, RestingECG, and ST_Slope) and label encoding for ExerciseAngina
- Feature scaling using StandardScaler for numerical features
- Polynomial Features (interaction only): Created 2nd-degree interaction terms to enhance feature space
- Feature Selection: Used ANOVA F-test (SelectKBest) to select top 10 predictors
- Dimensionality Reduction: PCA applied on polynomial features to reduce dimensionality while retaining 95% variance.

Model Building and Evaluation

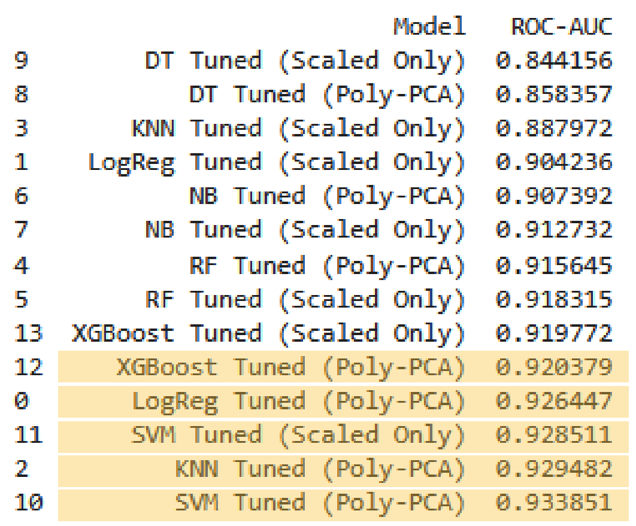
- Dataset was split into 80% training and 20% testing sets to evaluate model generalization performance.
- Models were trained on scaled dataset (11 features) and Poly-PCA Dataset (reduced interactions)
- Trained 7 supervised models, includings Logistic Regression, KNN, Decision Tree, Random Forest, Naïve Bayes, SVM, XGBoost
- Used GridSearchCV for tuning hyperparameters
- Evaluated with Accuracy, F1-score, and ROC-AUC





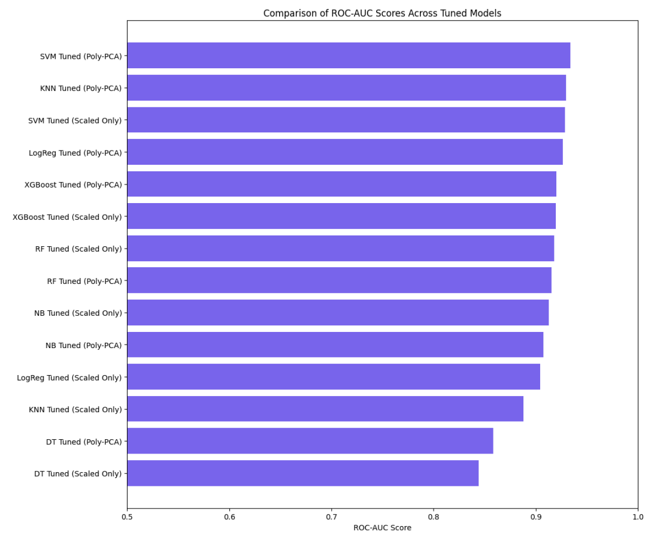
```
                        Model  Accuracy  F1-score
9      DT Tuned (Scaled Only)  0.788043  0.807882
1  LogReg Tuned (Scaled Only)  0.831522  0.850242
8         DT Tuned (Poly-PCA)  0.836957  0.857143
2        KNN Tuned (Poly-PCA)  0.847826  0.862745
10       SVM Tuned (Poly-PCA)  0.853261  0.864322
7      NB Tuned (Scaled Only)  0.847826  0.866667
0     LogReg Tuned (Poly-PCA)  0.858696  0.871287
12   XGBoost Tuned (Poly-PCA)  0.853261  0.872038
4         RF Tuned (Poly-PCA)  0.858696  0.875000
11    SVM Tuned (Scaled Only)  0.858696  0.877358
3     KNN Tuned (Scaled Only)  0.864130  0.880383
5      RF Tuned (Scaled Only)  0.864130  0.882629
6         NB Tuned (Poly-PCA)  0.875000  0.894009
13 XGBoost Tuned (Scaled Only)  0.875000  0.894009
```

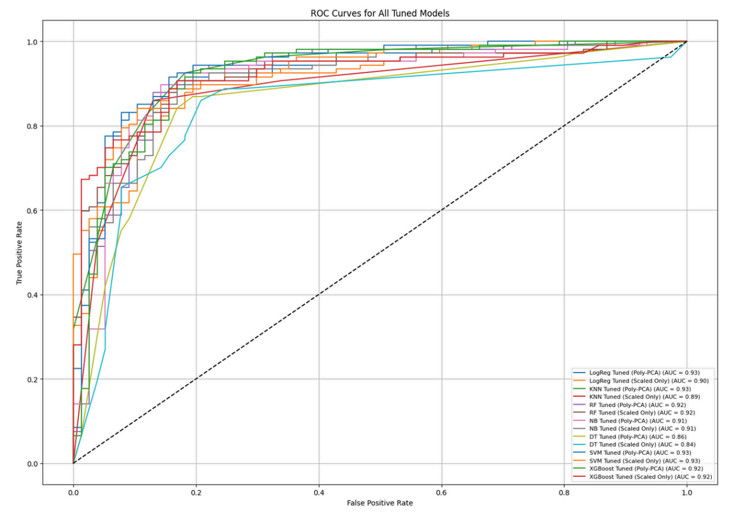
Models’ Accuracy and F1-Score



```
                          Model   ROC-AUC
9        DT Tuned (Scaled Only)  0.844156
8           DT Tuned (Poly-PCA)  0.858357
3       KNN Tuned (Scaled Only)  0.887972
1    LogReg Tuned (Scaled Only)  0.904236
6           NB Tuned (Poly-PCA)  0.907392
7        NB Tuned (Scaled Only)  0.912732
4           RF Tuned (Poly-PCA)  0.915645
5        RF Tuned (Scaled Only)  0.918315
13  XGBoost Tuned (Scaled Only)  0.919772
12     XGBoost Tuned (Poly-PCA)  0.920379
0       LogReg Tuned (Poly-PCA)  0.926447
11      SVM Tuned (Scaled Only)  0.928511
2          KNN Tuned (Poly-PCA)  0.929482
10         SVM Tuned (Poly-PCA)  0.933851
```

Models’ ROC-AUC Score





CONCLUSION

This project demonstrates the potential of an end-to-end machine learning pipeline in predicting heart disease using structured clinical data. Various preprocessing steps were applied, including encoding, scaling, interaction feature generation, and dimensionality reduction using PCA.

Among all evaluated models, SVM Tuned (Poly-PCA) achieved the best ROC-AUC (93.39%) and strong F1-score (86.4%). XGBoost Tuned (Scaled) and Naïve Bayes Tuned (Poly-PCA) also showed outstanding results, both achieving F1-scores of 89.4% with AUCs above 90%.

Both categorical and numerical variables were statistically associated with the target variable. Additionally, feature interaction and PCA boosted model performance in specific cases, especially for kernel-based models such as SVM and KNN, as well as probabilistic models like Naïve Bayes.

While model performance was promising, overall accuracy and F1-score remained below 90%, indicating that further refinements are necessary to reach clinical grade reliability. Future improvements may include:

- Integrating richer clinical data
- Applying advanced models (e.g., deep learning)
- Validating the models on external datasets to assess generalizability and reduce overfitting

These enhancements could help further unlock the potential of machine learning in supporting early detection and decision making in cardiovascular healthcare.

REFERENCES

Fedesoriano. (2021, September). Heart Failure Prediction Dataset. Retrieved June 6, 2025 from https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data

World Health Organization. (2021, June 11). Cardiovascular diseases (CVDs). Retrieved June 13, 2025, from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.