Heart Disease Risk Prediction using Machine Learning Models

Student ID	Name	Section
206001007	Zehra KOLAT	1
220911757	Kutay ŞAHİNLER	2
2309111082	Göktuğ ŞAHİN	3

Table 1: Team Members

COE305 - Machine Learning

FEMILDA JOSEPHIN JOSEPH SHOBANA BAI

Istinye University
Faculty of Engineering and Natural Sciences

Heart Disease Risk Prediction using Machine Learning Models

206001007 Zehra KOLAT Section 1
220911757 Kutay ŞAHİNLER Section 2
2309111082 Göktuğ ŞAHİN Section 3

11/12/2025

Stage 3

Contents

1 Dataset

Dataset:	`https://www.kaggle.com/datasets/pratyushpuri/heart-disease-dataset-3k-rows-python-code-2025`
Source:	Kaggle
Samples (rows):	3070 (Before SMOTE)
Features (columns):	17 (including given results)
Target variable:	`heart_disease`
Train–test split ratio:	80–20
Regression / Classification task:	Classification

2 Chosen Models

Algorithm Chosen	**Justification of choosing the algorithm**
Naive Bayes	It is simple and very fast, making it a strong baseline model.
Logistic Regression	Interpretable, efficient, and the standard choice for binary classification.
Support Vector Machine (SVM)	Heart-disease features form non-linear patterns, and SVM captures those relationships better than linear models.
Random Forest	Robust, handles feature interactions, and achieves strong accuracy on tabular medical data.

Table 2: Models

3 Model Evaluation

Model	Accuracy (Mean ± SD)	Precision	Recall	F1-Score	AUC
Naive Bayes	0.59	not computed	not computed	0.6514	not computed
Logistic Regression	0.63	not computed	not computed	0.6610	not computed
SVM	0.6398	0.64	0.64	0.64	0.6786
Random Forest	0,6791	not computed	not computed	0,9240	not computed

Table 3: Model Performance Metrics of Training Results

Model	Accuracy (Mean ± SD)	Precision	Recall	F1-Score	AUC
Naive Bayes	0.58	0.59	0.58	0.58	0.6001
Logistic Regression	0.61	0.62	0.61	0.62	0.6380
SVM	0.6221	0.64	0.62	0.63	0.6473
Random Forest	0,63	0,52	0,55	0,53	0,6372

Table 4: Model Performance Metrics of Test Results

4 Analysis & Observations

1. Which model performed best and why?

Across all metrics, Random Forest shows the strongest performance on the training set with very high F1-score (0.9240) and strong k-fold scores (0.68–0.73 range). This indicates that the ensemble method is effective at capturing non-linear relationships and interactions in the data, outperforming linear models like Logistic Regression and Naive Bayes.

However, the test results tell a different story: although Random Forest still achieves the highest test accuracy (0.63), its precision/recall/F1 scores drop significantly (0.52–0.55). This suggests Random Forest has the most raw predictive power, but it is also the most sensitive to overfitting.

For stable and balanced performance, SVM is consistent, with only a small drop. SVM generalizes best in the dataset.

Conclusion:

Random Forest has the best training capacity and potential, SVM / Logistic Regression provide the most stable generalization, Naive Bayes consistently performs the lowest due to its strong independence assumptions.

Model	Reason
1. SVM	Best generalization, stable train-test performance
2. Logistic Regression	Reliable, no overfitting, decent metrics
3. Random Forest	High train scores but clear overfitting
4. Naive Bayes	Simple model; too weak for this dataset

Table 5: Ranking Across Models

2. Were there any overfitting or underfitting signs?

Random Forest - Clear Overfitting

Test accuracy drops to (0.63) from (0.68)

F1 drop is even more noticable (0.92–0.53)

K-fold variance is high (0.59 to 0.73)

Overfitting is visible even from the moon. The model learns training data too well but fails to generalize.

SVM - Stable

Training accuracy 0.6398 → test 0.6221

Slight performance drop, but consistent across metrics

K-fold scores (0.5566–0.6110) stable

SVM shows good generalization with minimal overfitting.

Logistic Regression - Stable

Train 0.62 → test 0.61 (almost identical)

K-fold accuracy mean 0.6138 with low variance (±0.0307)

This indicates the model is well-regularized and does not overfit.

Naive Bayes - Slight Underfitting

Train = test performance (0.58)

Low scores across the board

Small k-fold variance (±0.0199)

NB's simplicity causes underfitting—it cannot capture the complexity of the data.

3. How consistent were the results across folds?

Logistic Regression

Most consistent model across folding.

Mean accuracy 0.6138 with only ± 0.0307 variation. This shows reliability and stable performance across different subsets of data.

Naive Bayes

Also very consistent (± 0.0199), but consistently low.

Stable, but not expressive enough.

SVM

Fold scores range from 0.5566 to 0.6110.

Moderate variation but acceptable. SVM is stable but slightly sensitive to fold composition.

Random Forest

Highest variability ($0.59 \rightarrow 0.73$).

Unstable depending on which data split it sees. This supports the overfitting observation—RF learns specific patterns from some folds but not others.

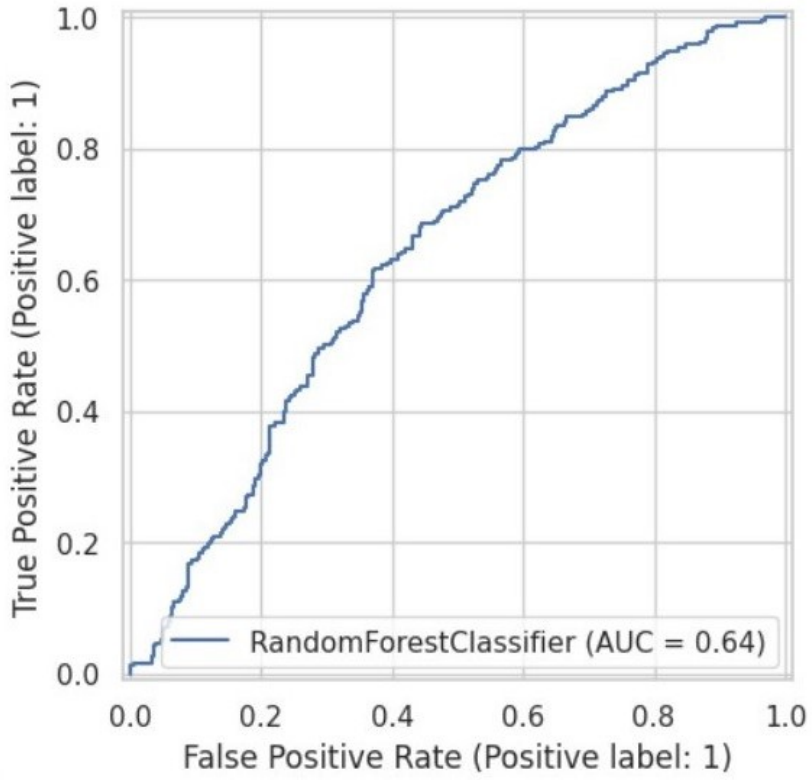

Figure 1: Aoc visualization for Random Forest Model

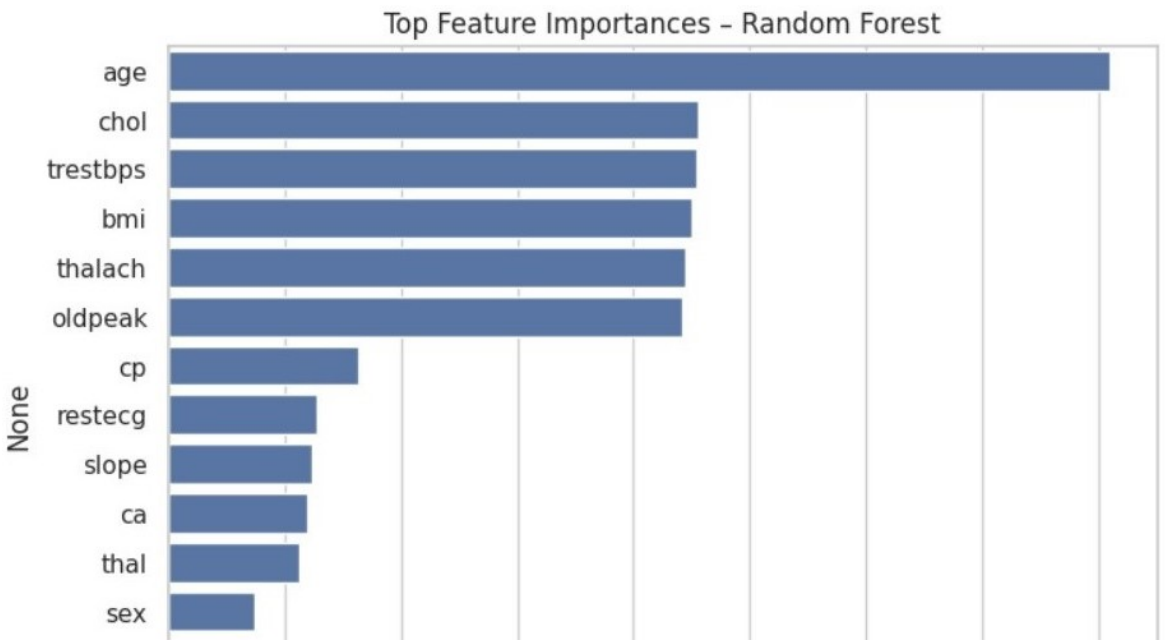

Figure 2: Feature importances for Random Forest Model