# **Report: Heart Disease Prediction Using Machine Learning**

## **1. Introduction**

Heart disease remains one of the leading causes of mortality worldwide. Early detection using machine learning (ML) techniques can support timely diagnosis and treatment. This study applies ML models to the **Heart Disease Dataset** (heart.csv) to evaluate predictive performance and identify the most important risk factors.

The workflow included:

* Data exploration and preprocessing
* Visualization of feature patterns
* Training and hyperparameter tuning of three models: **K-Nearest Neighbors (KNN)**, **Decision Tree (DT)**, and **Random Forest (RF)**
* Model evaluation, feature importance analysis, and key conclusions

## **2. Dataset Insights**

The dataset contained **303 records** and **14 attributes** describing patient health indicators such as age, chest pain type (cp), resting blood pressure, cholesterol level, maximum heart rate (thalach), exercise-induced angina, and others. The target variable indicates the presence (1) or absence (0) of heart disease.

Data quality checks showed:

* No missing values were present.
* Duplicates were dropped to avoid bias.
* The target classes were fairly balanced, reducing concerns about model bias.

This clean dataset provided a strong basis for machine learning analysis.

**3. Visualization Findings**

Exploratory Data Analysis (EDA) was performed to highlight key patterns:

* **Age Distribution**: Most patients were aged 40–60, a high-risk demographic.
* **Correlation Heatmap**: Strong positive correlation between *thalach* (max heart rate) and the presence of heart disease; negative correlation with *oldpeak* (ST depression).
* **Age vs. Thalach Scatter Plot**: Younger patients tended to have higher maximum heart rates, often associated with positive heart disease cases.
* **Target Distribution**: Nearly balanced classes ensured unbiased model evaluation.

The findings confirm that these features are closely linked to heart disease prediction.

## **4. Model Performance and Comparison**

Three models were trained and tuned using **RandomizedSearchCV** with cross-validation, optimizing for the F1-score.

| **Model** | **Accuracy** | **Best Hyperparamters** |
| --- | --- | --- |
| KNN | 78.7 % | n\_neighbors=5, weights=‘distance’ |
| Decision Tree | 77.0 % | max\_depth=5, min\_samples\_split=2 |
| Random Forest | 85.2 % | n\_estimators=200, max\_depth=10 |

**Random Forest** outperformed both KNN and Decision Tree.

**5. Feature Importance (Random Forest)**

The Random Forest model provided insights into which features most strongly influenced predictions. The top three features were:

1. **Chest Pain Type (cp)** – strongest predictor of heart disease
2. **Thallium Stress Test Result (thal)** – second most predictive feature
3. **Number of Major Vessels (ca)** – key indicator of vascular health

## **6. Impact of Hyperparameter Tuning**

Hyperparameter tuning played a pivotal role in enhancing the Random Forest model. Using **Randomized Search CV**, we explored multiple parameter combinations to identify the most effective configuration. The optimal settings were:

* Maximum Depth: 10
* Minimum Samples per Leaf: 2
* Number of Estimators: 200

With these adjustments, the Random Forest’s accuracy improved from 82.0% (default) to 85.2%(tuned). This improvement demonstrates how fine-tuning model parameters can significantly reduce overfitting, improve generalization, and ultimately deliver more reliable predictions.

## **7. Key Conclusions**

* The dataset was clean, balanced, and suitable for classification tasks.
* Exploratory analysis highlighted **chest pain type, thalach, oldpeak, and ca** as important risk indicators.
* Among the tested models, **Random Forest** performed best, achieving **85.2%** accuracy after tuning.
* Hyperparameter tuning was crucial, improving model accuracy and preventing overfitting.
* Ensemble models like Random Forest are more reliable than single learners (Decision Tree) for medical prediction tasks.