****

**Assessment Report**

on

**“Heart Disease Prediction”**

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AI)**

By

Names : Anant Singhal(202401100300037)

Abishek Singh(202401100300009)

Anirudh Kumar(202401100300045)

Aryan Khokhar(202401100300068)

Khushi Kathak(2024011003

Section: A

**Under the supervision of**

“Bikki Gupta”

**KIET Group of Institutions, Ghaziabad**

**May, 2025**

**1. Introduction**

Cardiovascular diseases are one of the leading causes of death globally. Early prediction of heart disease can assist in timely intervention and treatment. This project aims to develop a machine learning model to predict the presence of heart disease in patients using various medical parameters, thereby aiding clinical decision-making and improving patient outcomes.

**2. Problem Statement**

To build a classification model that predicts whether an individual has heart disease based on their medical history and health metrics. The model uses supervised machine learning to analyze features such as age, cholesterol level, blood pressure, ECG results, and more.

**3. Objectives**

 Load and explore the Heart Disease Dataset.

 Visualize relationships between features and the target variable.

 Preprocess data using encoding, scaling, and train-test splitting.

 Build and evaluate multiple machine learning classification models.

 Visualize the results with confusion matrices, ROC curves, and feature importance.

 Determine the best-performing model and discuss future improvements.

**4. Methodology**

* **Data Collection**: The user uploads a CSV file containing the dataset.
* **Data Preprocessing**:

 Categorical features encoded using Label Encoding and One-Hot Encoding.

 Numeric features scaled using StandardScaler.

 Dataset split into 70% training and 30% testing.

 Checked class distribution; if imbalanced, handled using class\_weight or SMOTE (if required).

* **Model Building**:

Trained and compared:

* Logistic Regression
* Decision Tree
* Random Forest
* K-Nearest Neighbors (KNN)
* Support Vector Machine (SVM)

* **Model Evaluation**:

 Metrics used: Accuracy, Precision, Recall, F1 Score, ROC-AUC.

 Visualized:

* Confusion matrix (heatmap)
* ROC curve for model comparison
* Feature importance (for tree-based models)

**5. Data Preprocessing**

* Missing values were checked; none found.
* Categorical columns like ChestPainType, RestingECG, and ST\_Slope were encoded using one-hot encoding.
* Features scaled for models sensitive to feature magnitude.
* Target column (HeartDisease) encoded as binary: 0 = No, 1 = Yes.

**6. Model Implementation**

All models were trained on the same processed dataset:

* **Logistic Regression**: Baseline model for linear decision boundary.
* **Decision Tree & Random Forest**: To capture non-linear relationships and rank feature importance.
* **KNN**: Based on similarity (distance metrics).
* **SVM**: For optimal hyperplane separation with kernel trick.

**7. Evaluation Metrics**

The following metrics are used to evaluate the model:

* **Accuracy**: Proportion of correct predictions
* **Precision**: TP / (TP + FP) – correctness of positive predictions
* **Recall**: TP / (TP + FN) – sensitivity to actual positives
* **F1 Score**: Harmonic mean of precision and recall.
* **ROC-AUC Score**: Area under the ROC curve; higher is better

**8. Results and Analysis**

 **Random Forest** gave the best performance among all models, achieving the highest ROC-AUC and F1 scores.

 Confusion matrices revealed balance in classification accuracy.

 Feature importance plots showed ChestPainType, Oldpeak, and MaxHR as the most influential predictors.

 ROC curves highlighted model separation power across thresholds.

**9. Conclusion**

This project successfully built and evaluated a heart disease prediction model using multiple machine learning algorithms. Among them, Random Forest performed the best in terms of accuracy and interpretability. This model could serve as a decision support tool in clinical environments. Future work can focus on:

* Hyperparameter tuning using GridSearchCV.
* Incorporating deep learning models (e.g., neural networks).
* Deploying the model as a web or mobile app for practical use.

**10. References**

* [Scikit-learn documentation](https://scikit-learn.org)
* Pandas documentation
* [Matplotlib documentation](https://matplotlib.org)
* Seaborn documentation
* Research papers and articles on heart disease prediction using ML





