****

**Assessment Report**

on

**“Heart Disease Prediction”**

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AIML)**

**By**

Harshit

Jayant Singh

Kanishak Tyagi

Kartik Kumar

Manasvi Tyagi

**Roll Numbers :**

202401100400095

202401100400102

202401100400104

202401100400105

202401100400118

Section: B

**Under the supervision of**

“Abhishek Shukla”

**KIET Group of Institutions, Ghaziabad**

**May, 2025**

#### ****1. Introduction****

Cardiovascular diseases remain one of the leading causes of mortality globally. With the increasing availability of patient data, machine learning provides an effective approach to early prediction and risk assessment of heart disease. This project focuses on predicting the likelihood of heart disease in patients using supervised machine learning techniques, particularly logistic regression.

#### ****2. Problem Statement****

To build a machine learning model that can classify whether a patient is likely to have heart disease based on their medical and lifestyle attributes.

#### ****3. Objectives****

* To preprocess and clean the UCI heart disease dataset.
* To develop a logistic regression model for prediction.
* To evaluate model performance using standard metrics.
* To visualize and interpret results using a confusion matrix and other visual aids.

#### ****4. Methodology****

**4.1 Data Collection**  
The dataset used is the UCI Heart Disease Dataset, consisting of 920 entries with 16 attributes including age, sex, chest pain type, cholesterol level, etc.

**4.2 Data Preprocessing**

* Missing values handled via mean imputation for numerical and mode imputation for categorical columns.
* Categorical variables converted using one-hot encoding.
* StandardScaler used for feature scaling.
* Target variable num was binarized: values >0 converted to 1 (indicating presence of heart disease), else 0.

**4.3 Model Development**

* Dataset split into training (80%) and testing (20%) subsets.
* Logistic Regression model trained on the training data.

**4.4 Model Evaluation**

* Accuracy, Precision, Recall, and F1-score used for evaluation.
* Confusion matrix visualized using Seaborn heatmap.

#### ****5. Data Preprocessing Summary****

* Numerical Columns with Missing Data: trestbps, chol, thalch, oldpeak, ca
* Categorical Columns with Missing Data: fbs, restecg, exang, slope, thal
* Encoding: One-hot encoding used on categorical columns such as sex, cp, restecg, slope, thal, etc.
* Scaling: Features scaled using StandardScaler.

#### ****6. Model Implementation****

The model was implemented using scikit-learn. Key steps include:

python

CopyEdit

from sklearn.linear\_model import LogisticRegression

model = LogisticRegression()

model.fit(X\_train, y\_train)

#### ****7. Evaluation Metrics****

| **Metric** | **Value (example)** |
| --- | --- |
| Accuracy | 0.86 |
| Precision | 0.88 |
| Recall | 0.84 |
| F1 Score | 0.86 |

Note: Actual values may vary depending on random state and imputation strategy.

* **Confusion Matrix** visualized with a heatmap helps identify true positives and false negatives clearly.

#### ****8. Results and Analysis****

* The logistic regression model performed well with balanced precision and recall.
* Most instances of heart disease were correctly classified.
* Model showed potential for real-world application in healthcare prediction tools.
* Imbalance in certain feature values suggests that advanced techniques like SMOTE or ensemble models (e.g., Random Forest) may further improve results.

#### ****9. Conclusion****

The heart disease prediction model achieved strong performance using logistic regression. The project highlights the power of supervised learning in the healthcare domain. Future work can include:

* Comparing with models like SVM, Decision Tree, or Random Forest.
* Feature selection or dimensionality reduction techniques (like PCA).
* Addressing class imbalance using resampling methods.

#### ****10. References****

* UCI Heart Disease Dataset
* Scikit – Learn Documentation
* Pandas documentation
* Seaborn visualization library

#### ****11. Team Contribution****

This project was completed as a collaborative effort by a five-member team. The roles and contributions are as follows:

* **Kartik Kumar**: Authored and compiled the final project report, documenting each section in a structured format.
* **Harshit**: Developed and implemented the machine learning codebase, including model building and evaluation.
* **Kanishak Tyagi**: Collaborated with Harshit in coding and testing the logistic regression model and improving data preprocessing.
* **Jayant Singh**: Created the presentation slides, focusing on data visualizations and model insights.
* **Manasvi**: Co-designed the presentation and organized content layout for final submission.

All members actively participated in team discussions and iterative improvements.

