Assessment Report

on

“ Predict Heart Disease”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

Computer Science Engineering - Artificial Intelligence

(CSE-AI)

By

SUHAWANI SHUKLA (20240110300254, CSE-AI D)

Under the supervision of

“MR.ABHISHEK SHUKLA”

KIET Group of Institutions, Ghaziabad

May, 2025

Introduction

Heart disease is a leading cause of death worldwide, and early detection is key to improving patient outcomes. Traditional diagnostic methods can be time-consuming and costly, but machine learning provides an effective alternative. This project leverages machine learning to predict the likelihood of heart disease based on various health factors, such as age, cholesterol levels, and maximum heart rate. By training a logistic regression model on a dataset containing these features, the goal is to build a system that can predict whether a patient is at risk of heart disease.

METHODOLOGY

1.Data Collection:

The **Heart Disease UCI dataset** is used, containing medical features and a target variable indicating heart disease.

2. Data Preprocessing:

- ○ **Label Encoding** for categorical variables (e.g., gender).
- ○ **Feature Scaling** using **StandardScaler** to normalize numerical data.

2. **Model Selection**:
Logistic Regression is used for binary classification.

3. **Model Training and Evaluation**:
The model is trained on the training data, and performance is evaluated using metrics like

accuracy, **precision**, and **recall**.

Prediction:
 The model predicts heart disease for new patients, providing a classification and probability.

CODE

```
# Step 1: Upload CSV File (Google Colab / Jupyter compatible)
from IPython.display import display
import pandas as pd
import io
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix,
classification_report
import seaborn as sns
import matplotlib.pyplot as plt

# For uploading file
from google.colab import files
uploaded = files.upload("Predict Heart Disease (2).csv")

# Get the uploaded file path
file_path = next(iter(uploaded))

# Step 2: Load and Validate Data
df = pd.read_csv(io.BytesIO(uploaded[file_path]))

# Display first few rows
print("Sample data:")
print(df.head())
```

```
# Optional: Check missing values
if df.isnull().sum().any():
    print("\nWarning: Missing values detected!")
else:
    print("\nNo missing values found.")

# Step 3: Preprocess the Data
def preprocess_data(df):
    # Encode categorical if needed
    if df['sex'].dtype != 'int':
        df['sex'] = df['sex'].map({'male': 0, 'female': 1})

    # Split features and target
    X = df.drop('target', axis=1)
    y = df['target']

    # Scale features
    scaler = StandardScaler()
    X_scaled = scaler.fit_transform(X)

    return X_scaled, y

X_scaled, y = preprocess_data(df)

# Step 4: Train the Model
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y,
test_size=0.2, random_state=42)
model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)

# Step 5: Evaluate the Model
y_pred = model.predict(X_test)

# Accuracy
accuracy = accuracy_score(y_test, y_pred)
print(f"\n✅ Model Accuracy: {accuracy * 100:.2f}%")

# Confusion Matrix
cm = confusion_matrix(y_test, y_pred)
```

```
print("\n🧮 Confusion Matrix:")
print(cm)

# Plot Confusion Matrix
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues")
plt.title("Confusion Matrix")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.show()

# Classification Report
print("\n📊 Classification Report:")
print(classification_report(y_test, y_pred))
```

Output/Result

```
No missing values found.



✅ Model Accuracy: 88.52%



🧮 Confusion Matrix:

[[25  4]

 [ 3 29]]
```

References/Credits

Dataset:

- **Heart Disease UCI dataset**: UCI Repository

Libraries & Algorithms:

- **Logistic Regression**: Scikit-learn
- **Pandas**: Pandas Documentation
- **Scikit-learn**: Scikit-learn Docs
- **Tkinter**: Tkinter Documentation