Assessment Report

on

“Problem Statement”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

CSE(AIML)

By

Name: Sujal Kumar

Roll Number: 202401100400193

Section: C

Under the supervision of

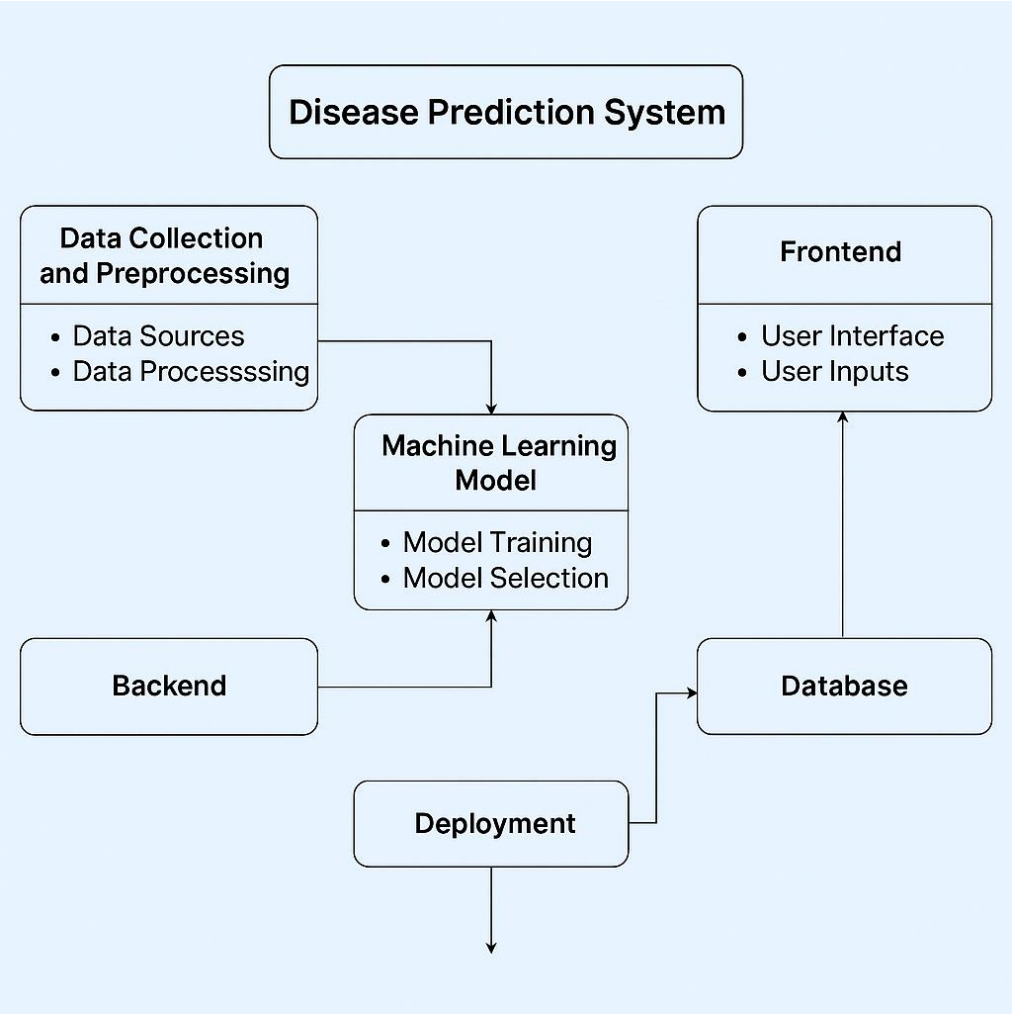
“ABHISHEK SHUKLA”

KIET Group of Institutions, Ghaziabad

INTRODUCTION

The project aims to predict the risk of a specific disease in patients based on their genetic markers, clinical symptoms, and lifestyle data. The prediction is carried out using supervised machine learning models, such as Logistic Regression, Support Vector Machines (SVM), and Random Forest. The system classifies patients into different risk categories, which helps healthcare professionals make more informed decisions about patient care. This is a crucial step in personalized medicine and early detection of diseases.

In this project, various data sources such as genetic information, clinical data (symptoms, history), and lifestyle factors (smoking, diet) are combined to train a machine learning model that predicts whether a patient is at high risk for a particular disease. The performance of the model is evaluated using metrics like accuracy, precision, recall, and F1-score.



METHODOLOGY

1. **Data Collection and Preprocessing:**
    - The data consists of three primary sources: genetic data (genetic markers), clinical data (patient symptoms and history), and lifestyle data (e.g., smoking, exercise habits).
    - These datasets are cleaned and merged to create a single dataset for model training.
    - Preprocessing steps include:
        - Handling missing data
        - Encoding categorical features
        - Normalizing numerical features
        - Feature engineering
2. **Model Selection and Training:**
    - A supervised machine learning approach is used for classification. The algorithms considered include Logistic Regression, SVM, and Random Forest.
    - The data is split into training and testing datasets (usually a 70-30 split).
    - Hyperparameter tuning and cross-validation are performed to optimize the models.
3. **Model Evaluation:**
    - The model's performance is evaluated using various metrics, including accuracy, precision, recall, and F1-score, to ensure the model can correctly classify patients at risk.
4. **Deployment:**
    - Once trained, the model can be used for inference on new patient data to predict their disease risk.

CODE

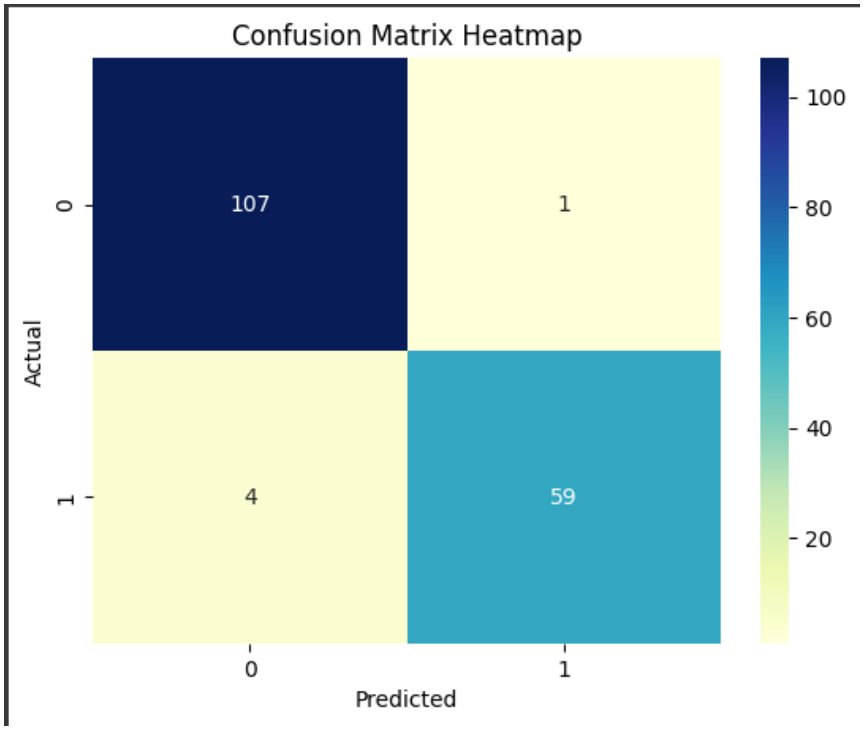
Below is the main code used in the project for training and evaluating the model. This includes data preprocessing, training the machine learning models, and evaluating their performance.

```
# -------------------------
#  Step 0: Upload Dataset
# -------------------------
from google.colab import files
uploaded = files.upload()

# Replace with the name of the uploaded file
import pandas as pd
df = pd.read_csv("dataset18.csv")  #  Your actual file

# -------------------------
#  🖌️ Step 1: Preprocessing
# -------------------------
# Drop unwanted columns
df = df.drop(columns=["id", "Unnamed: 32"], errors="ignore")

# Convert target: M = 1 (Malignant), B = 0 (Benign)
df["diagnosis"] = df["diagnosis"].map({"M": 1, "B": 0})

# Separate features and target
X = df.drop("diagnosis", axis=1)
y = df["diagnosis"]

# Standardize the features
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Train-test split
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.3, random_state=42)
```

```
# -------------------------
#  Step 2: Model Training
from sklearn.ensemble import RandomForestClassifier
model = RandomForestClassifier(random_state=42)
model.fit(X_train, y_train)

# 🍼 Step 3: Prediction & Evaluation
# -------------------------
from sklearn.metrics import confusion_matrix, classification_report, accuracy_score, precision_score, recall_score
import seaborn as sns
import matplotlib.pyplot as plt

# Predict
y_pred = model.predict(X_test)

# Confusion Matrix Heatmap
cm = confusion_matrix(y_test, y_pred)
sns.heatmap(cm, annot=True, fmt="d", cmap="YlGnBu")
plt.title("Confusion Matrix Heatmap")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.show()

# Metrics
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)

print(" Evaluation Metrics:")
print(f" Accuracy :  {accuracy:.2f}")
print(f" Precision:  {precision:.2f}")
print(f" Recall   :  {recall:.2f}")

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

Output/Result:

The model predicts whether a patient is at risk of a particular disease based on their genetic, clinical, and lifestyle data. Below is a sample output showing the model's performance evaluation metrics on a test dataset.




```
Evaluation Metrics:
Accuracy :  0.97
Precision:  0.98
Recall   :  0.94

Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.99      0.98       108
           1       0.98      0.94      0.96        63

    accuracy                           0.97       171
   macro avg       0.97      0.96      0.97       171
weighted avg       0.97      0.97      0.97       171
```

References/Credits:

- **Datasets**: The datasets used in this project were sourced from Kaggle.com.
- **Machine Learning Library**: This project used Scikit-learn for implementing the machine learning algorithms.