Assesment Report

on

“Problem Statement”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

CSE AIML

By

Shavyam Chitranshi(202401100400173)

Under the supervision of

MR. Abhishek Shukla

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May, 2025

Introduction-

Cancer is one of the most common cancers globally, and early detection can significantly improve patient outcomes. Physicians often rely on clinical and genetic data to classify tumors as **Malignant (M)** or **Benign (B)**. Malignant tumors are cancerous, while benign tumors are non-cancerous.

This project utilizes machine learning to automate tumor classification based on features like size, shape, and texture derived from clinical observations. The dataset includes measurements such as radius_mean, texture_mean, and perimeter_mean, which describe various physical attributes of cell nuclei.

By employing the Random Forest Classifier algorithm, this project aims to build a predictive model to assist medical professionals. Machine learning not only increases accuracy but also saves time in identifying and analyzing tumor types, which can positively impact patient care.

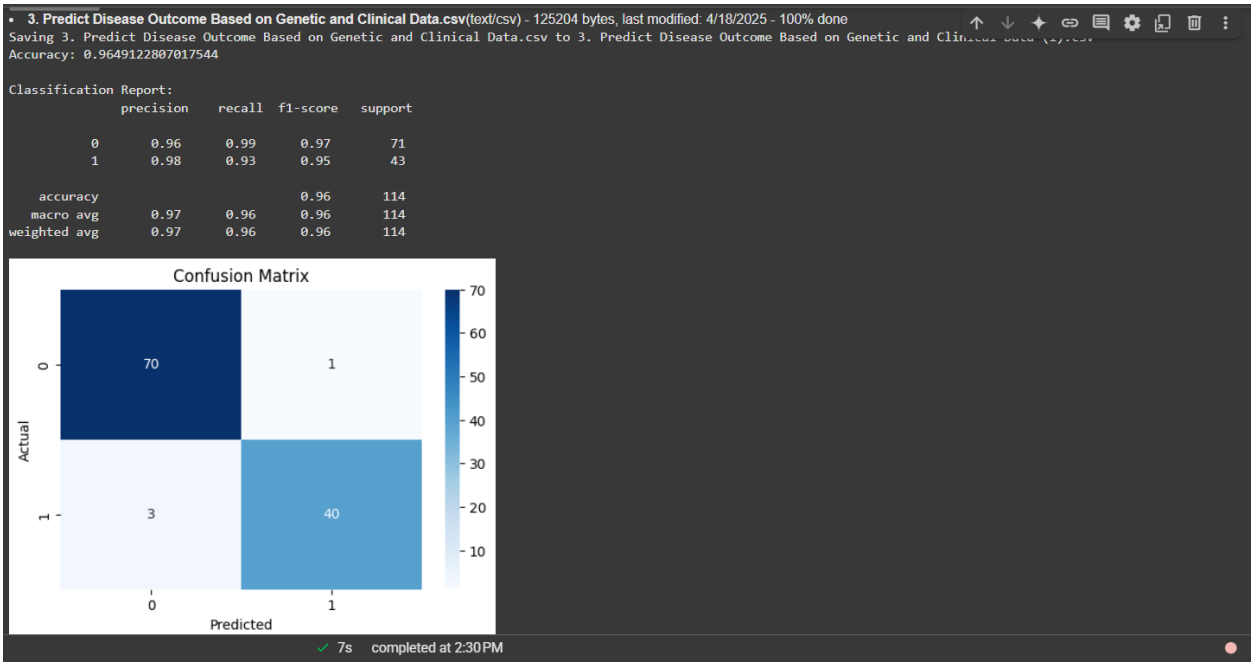
Methodology-

This project focuses on predicting the outcome of disease diagnosis using genetic and clinical data. By leveraging machine learning techniques, particularly the Random Forest Classifier algorithm, the goal is to classify tumors as **Malignant (M)** or **Benign (B)** based on feature data extracted from clinical and genetic datasets. The dataset contains features such as radius, texture, perimeter, and area metrics, which represent tumor characteristics critical to classification. It presents the confusion table which tells us about the accuracy of the model with the accurate dataset.

```
import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import classification_report, confusion_matrix,
accuracy_score

import matplotlib.pyplot as plt

import seaborn as sns


from google.colab import files

uploaded = files.upload()


df = pd.read_csv("3. Predict Disease Outcome Based on Genetic and Clinical
Data.csv")

df.drop(['id', 'Unnamed: 32'], axis=1, inplace=True)

df['diagnosis'] = df['diagnosis'].map({'M': 1, 'B': 0})


X = df.drop('diagnosis', axis=1)

y = df['diagnosis']


X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
```

```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

clf = RandomForestClassifier(random_state=42)
clf.fit(X_train_scaled, y_train)

y_pred = clf.predict(X_test_scaled)

print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred))

cm = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(6,4))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix")
plt.show()
```

OUTPUT-



REFERENCES-

- **Machine Learning Library**: scikit-learn
- **Visualization Libraries**: matplotlib and seaborn

- **Python for Data Analysis**: pandas documentation
- **KAGGLE**