Assesment Report

on

“Predict Disease Outcome Based on Genetic and Clinical Data”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

Introduction to AI

By

Sakshi Kumari (202401100400163)

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

Problem Statement:

Predict Disease Outcome Based on Genetic and Clinical Data

Use supervised machine learning to classify patients based on genetic markers, clinical symptoms, and lifestyle factors, predicting whether they are at risk for a particular disease.

Name: Sakshi Kumari
Roll Number: 202401100400163
Subject: Introduction to Artificial Intelligence
Assessment: MSE 2
Tool Used: Google Colab, Python, scikit-learn

Introduction

Early detection of disease, particularly life-threatening ones such as breast cancer, significantly improves treatment success rates. Machine learning enables automatic identification of patterns in genetic and clinical data, which can assist healthcare professionals in diagnosing patients.

This project focuses on using supervised learning to predict the likelihood of a patient having breast cancer based on features extracted from digitized images of fine needle aspirates (FNAs) of breast masses. The dataset contains features like radius, texture, smoothness, and other related attributes.

🛠 Methodology

1. **Dataset Loading and Exploration**
    - The dataset was imported from a .csv file containing 569 records with 30 features and a diagnosis label.
    - Removed non-informative columns (id, Unnamed: 32).
2. **Data Preprocessing**
    - Categorical label diagnosis was encoded (Malignant = 1, Benign = 0).
    - Features were standardized using StandardScaler to bring all values to the same scale.
3. **Train-Test Split**
    - The dataset was split into 80% training and 20% testing subsets using train_test_split.
4. **Model Selection**
    - A **Random Forest Classifier** was used due to its performance on classification tasks and ability to highlight feature importance.
5. **Model Training & Evaluation**
    - The model was trained using the training data.
    - Evaluation metrics like accuracy, precision, recall, and F1-score were calculated using the test set predictions.

Code

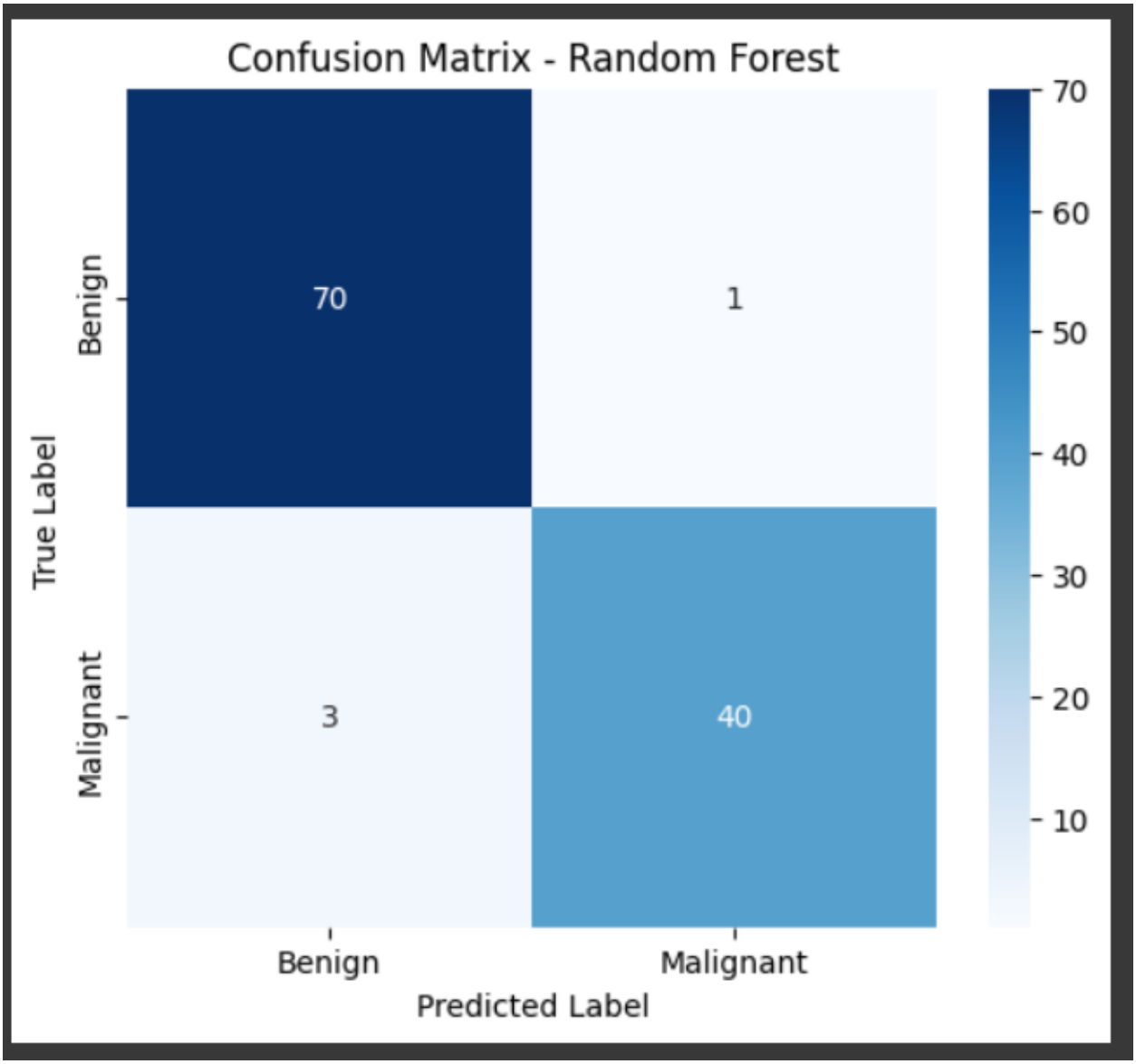
```
# STEP 1: Upload file
from google.colab import files
uploaded = files.upload()

# STEP 2: Import libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

# STEP 3: Load the uploaded CSV
df = pd.read_csv("3. Predict Disease Outcome Based on Genetic and Clinical Data (1).csv")

# STEP 4: Data Cleaning
df.drop(columns=["id", "Unnamed: 32"], inplace=True)
df["diagnosis"] = df["diagnosis"].map({"M": 1, "B": 0})

# STEP 5: Feature selection and splitting
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)


# STEP 6: Normalize features

scaler = StandardScaler()

X_train_scaled = scaler.fit_transform(X_train)

X_test_scaled = scaler.transform(X_test)


# STEP 7: Train model

model = RandomForestClassifier(random_state=42)

model.fit(X_train_scaled, y_train)


# STEP 8: Evaluate

y_pred = model.predict(X_test_scaled)


# Text-based output

print("Accuracy:", accuracy_score(y_test, y_pred))

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


# STEP 9: Confusion matrix heatmap

cm = confusion_matrix(y_test, y_pred)

plt.figure(figsize=(6, 5))

sns.heatmap(cm, annot=True, fmt="d", cmap="Blues", xticklabels=["Benign", "Malignant"],
yticklabels=["Benign", "Malignant"])

plt.xlabel("Predicted Label")

plt.ylabel("True Label")
```

```
plt.title("Confusion Matrix - Random Forest")

plt.show()
```

Output/Result

```
Choose Files  3. Predict D...cal Data.csv
• 3. Predict Disease Outcome Based on Genetic and Clinical Data.csv(text/csv) - 125204 bytes, last modified: 18/4/2025 - 100% done
Saving 3. Predict Disease Outcome Based on Genetic and Clinical Data.csv to 3. Predict Disease Outcome Based on Genetic and Clinical Data (2).csv
Accuracy: 0.9649122807017544
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.99      0.97        71
           1       0.98      0.93      0.95        43

    accuracy                           0.96       114
   macro avg       0.97      0.96      0.96       114
weighted avg       0.97      0.96      0.96       114
```



📚 References/Credits

- Dataset Source: UCI Breast Cancer Wisconsin Dataset
- Libraries: pandas, scikit-learn, matplotlib, seaborn
- Environment: Google Colab
- Guide/Documentation: scikit-learn official docs