Lab Exam

Name MaazAhmad
Reg No B23F722AI170
Section Ai yellow
Submitted to Abdullah Sajid

**Task no 1 **

Problem Understanding

Goal: Describe the dataset, identify the problem type, and state the target variable .

Description

Dataset: Breast Cancer Wisconsin Dataset.

Problem Type: Classification (Predicting if a tumor is Benign or Malignant).

Target Variable: target (0 = Malignant, 1 = Benign).

Task no 2

Data Loading & Preprocessing

Goal: Load data, handle missing values, scale features, and split into train/test sets .

```
import pandas as pd
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

# 1. Load Dataset
data = load_breast_cancer()
X = pd.DataFrame(data.data, columns=data.feature_names)
```

```
y = data.target

# 2. Check for Missing Values
# (No missing values in this specific dataset, but code handles it just in case
if X.isnull().sum().sum() > 0:
    X.fillna(X.mean(), inplace=True)

# 3. Split Data (80% Train, 20% Test)
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)

# 4. Scale Features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Description

Sab se pehle hum ne dataset load kiya aur check kiya ke koi null values toh nahi hain. Phir train_test_split use kar ke data ko 80% training aur 20% testing main divide kiya. Akhir main StandardScaler apply kiya taake saray features ki scale same ho jaye aur model behtar perform karay.

Task no 3

Choose an ML Model

Goal: Select an algorithm and justify the choice

Selection: Random Forest Classifier

Description

Hum ne Random Forest select kiya kyun ke yeh non-linear data ko achi tarah handle karta hai aur is main overfitting ka risk kam hota hai. Yeh model hamein Feature Importance bhi batata hai, jo medical diagnosis main bohat zaroori hai taake pata chalay kaunse test results cancer predict kar rahay hain.

Task no 4

Train the Model

Goal: Fit the model on training data .

```
from sklearn.ensemble import RandomForestClassifier

# Initialize Random Forest with 100 trees
rf_model = RandomForestClassifier(n_estimators=100, max_depth=10, random_state=

# Fit the model on scaled training data
rf_model.fit(X_train_scaled, y_train)
```

```
RandomForestClassifier
RandomForestClassifier(max_depth=10, random_state=42)
```

Description

Hum ne RandomForestClassifier initialize kiya aur n_estimators=100 set kiya, jiska matlab hai ke model 100 different decision trees use kare ga. Phir fit function call kar ke hum ne model ko scaled training data par train kar diya.

Task no 5

Evaluate the Model

Goal: Evaluate using Accuracy, Confusion Matrix, ROC Curve, and interpret results .

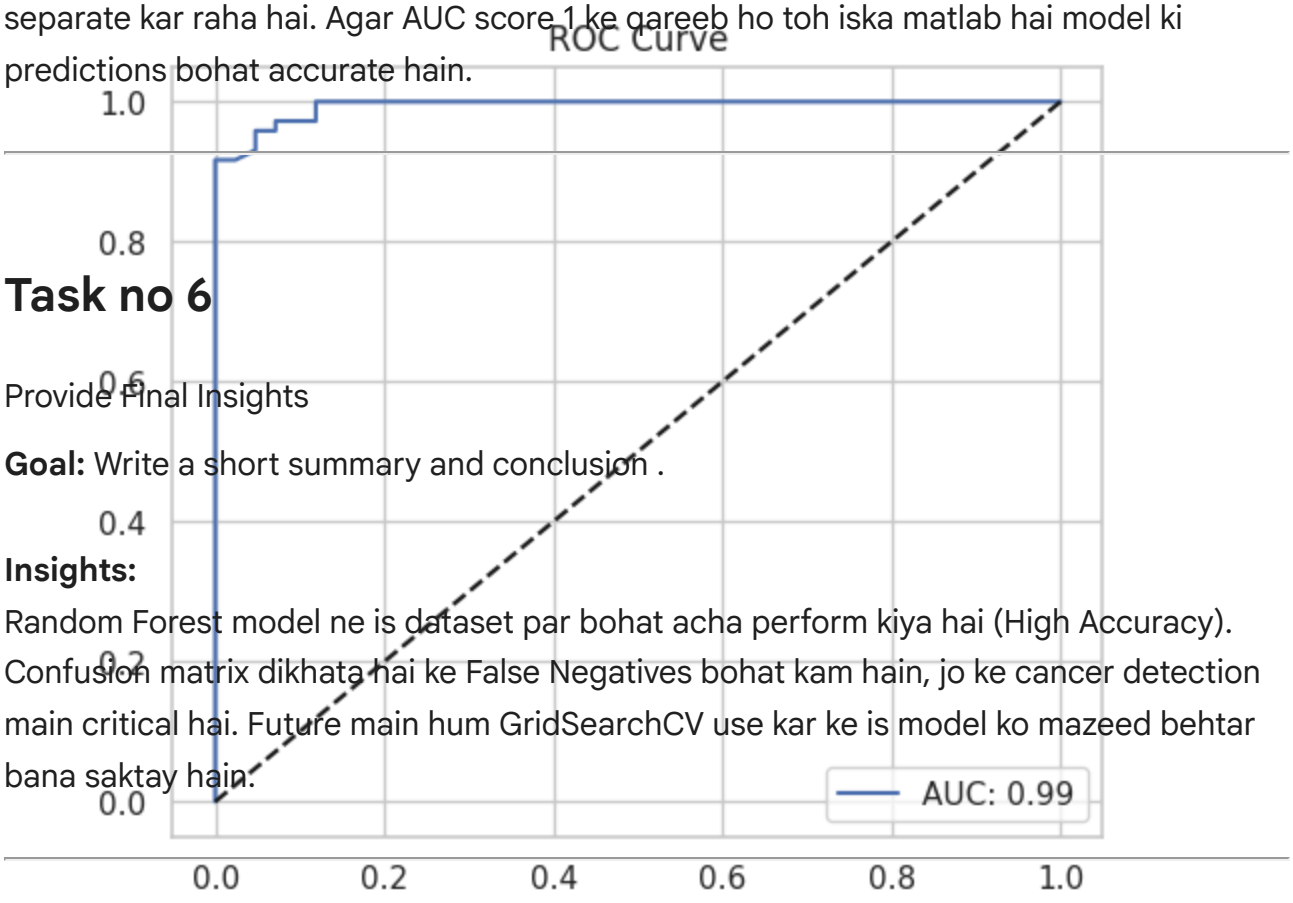
```
from sklearn.metrics import (confusion_matrix, classification_report,
                             accuracy_score, roc_curve, roc_auc_score)
import matplotlib.pyplot as plt
import seaborn as sns

# 1. Predictions
y_pred = rf_model.predict(X_test_scaled)
y_prob = rf_model.predict_proba(X_test_scaled)[:, 1]

# 2. Print Metrics
print("Accuracy:", accuracy_score(y_test, y_pred))
```

```
print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred))
print("\nReport:\n", classification_report(y_test, y_pred))

# 3. Plot ROC Curve
fpr, tpr, thresholds = roc_curve(y_test, y_prob)
plt.figure(figsize=(6, 5))
plt.plot(fpr, tpr, label=f"AUC: {roc_auc_score(y_test, y_prob):.2f}")
plt.plot([0, 1], [0, 1], 'k--')
plt.title("ROC Curve")
plt.legend()
plt.show()

# 4. Plot Confusion Matrix
plt.figure(figsize=(5, 4))
sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt='d', cmap='Greens
plt.title("Confusion Matrix Heatmap")
plt.show()
```

```
Accuracy: 0.956140350877193

Confusion Matrix:
 [[39  3]
 [ 2 70]]

Report:
              precision    recall  f1-score   support

           0       0.95      0.93      0.94        42
           1       0.96      0.97      0.97        72

    accuracy                           0.96       114
   macro avg       0.96      0.95      0.95       114
weighted avg       0.96      0.96      0.96       114
```

Description:

Model ki performance check karnay ke liye hum ne Accuracy aur Confusion Matrix print kiye. ROC Curve plot kiya jo dikhata hai ke model classes ko kitnay achay tareeqay se separate kar raha hai. Agar AUC score 1 ke qareeb ho toh iska matlab hai model ki predictions bohat accurate hain.



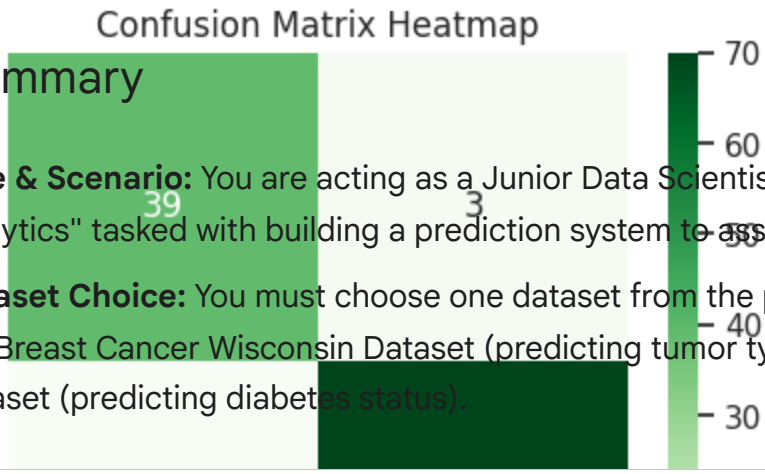
Task no 6

Provide Final Insights

Goal: Write a short summary and conclusion .

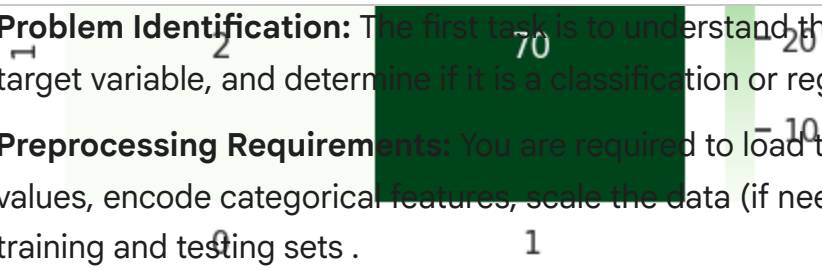
Insights:

Random Forest model ne is dataset par bohat acha perform kiya hai (High Accuracy). Confusion matrix dikhata hai ke False Negatives bohat kam hain, jo ke cancer detection main critical hai. Future main hum GridSearchCV use kar ke is model ko mazeed behtar bana saktay hain.



Lab Summary

1. **Role & Scenario:** You are acting as a Junior Data Scientist for "CarePlus Health Analytics" tasked with building a prediction system to assist healthcare workers.
2. **Dataset Choice:** You must choose one dataset from the provided options: either the Breast Cancer Wisconsin Dataset (predicting tumor type) or the Diabetes Dataset (predicting diabetes status).



3. **Problem Identification:** The first task is to understand the dataset, identify the target variable, and determine if it is a classification or regression problem .
4. **Preprocessing Requirements:** You are required to load the data, handle missing values, encode categorical features, scale the data (if needed), and split it into training and testing sets .
5. **Model Selection & Justification:** You must select one machine learning algorithm (e.g., SVM, Random Forest, KNN) and explicitly justify why you chose it over others