Experiment No : 6 **Roll no :**

Aim : To study and implement random forest algorithm in machine learning.

Theory : Random Forest is an ensemble learning method primarily used for classification and regression tasks. It was introduced by Leo Breiman in 2001 as an improvement over decision trees and bagging methods. The fundamental idea behind Random Forest is to build a large number of decision trees during training and output the mode of the classes (for classification) or mean prediction (for regression) of the individual trees.

Random Forest reduces overfitting and improves accuracy by combining the predictions of multiple decision trees, each built on different subsets of the data and features. It is widely used due to its simplicity, robustness, and ability to handle high-dimensional data.

2. How Random Forest Works

The Random Forest algorithm follows these main steps:

Step 1: Bootstrapping (Data Sampling)

- From the original dataset of size N, multiple bootstrap samples (also of size N) are drawn **with replacement**. This means some data points may appear multiple times, while others may be left out.

Step 2: Building Decision Trees

- For each bootstrap sample, a **decision tree** is grown.
- At each node of the tree, a **random subset of features (m << total features)** is selected, and the best split is found only from this subset. This is a key difference from standard decision trees, which consider all features.
- Trees are grown **to maximum depth** without pruning (unlike traditional decision trees).

Step 3: Aggregation (Ensemble Prediction)

- **For classification**: Each tree votes for a class, and the majority vote determines the final prediction.
- **For regression**: The average of predictions from all trees is taken as the final output.

This process introduces two types of randomness:

1. Randomness in data (bootstrapping).
2. Randomness in feature selection at each split.

Together, these reduce correlation between individual trees and enhance the generalization ability of the model.

3. Mathematical Formulation

Let’s denote the training dataset as D={(xi,yi)}i=1ND = \{(x_i, y_i)\}_{i=1}^{N}. A Random Forest constructs TT decision trees, each built from a bootstrap sample Dt⊂DD_t \subset D. The prediction y^\hat{y} for a new input xx is:

Classification:

y^=mode{ht(x)}t=1T\hat{y} = \text{mode}\{h_t(x)\}_{t=1}^{T}

Regression:

y^=1T∑t=1Tht(x)\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x)

Where ht(x)h_t(x) is the prediction of the ttht^{th} tree.

4. Advantages of Random Forest

- **High Accuracy**: Due to ensemble nature and reduced overfitting.
- **Handles High-Dimensional Data**: Effective even when the number of features is very large.
- **Robust to Outliers and Noise**: Since multiple trees average out errors.
- **Feature Importance**: Provides insight into the relative importance of each feature.
- **No Need for Scaling:** Works well without normalization of input features.

6. Applications of Random Forest

- **Medical Diagnosis**: Disease prediction using patient features.
- **Finance**: Credit scoring, fraud detection.
- **Marketing**: Customer segmentation, churn prediction.
- **Ecology**: Species classification, environmental modeling.
- **Natural Language Processing**: Text classification, sentiment analysis.

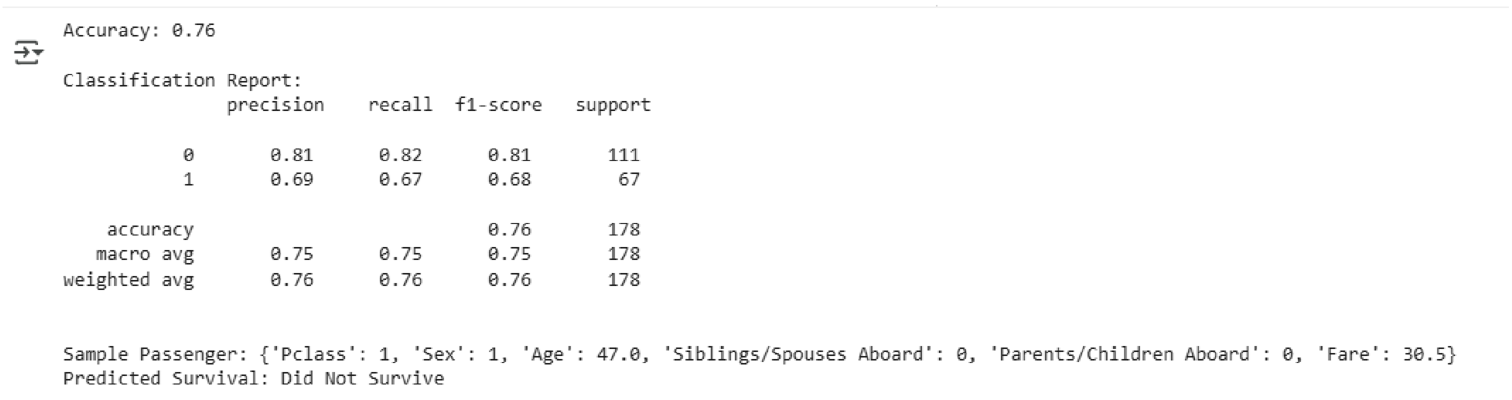
Source Code :

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report
import warnings
warnings.filterwarnings('ignore')

titanic_data = pd.read_csv('titanic.csv')

titanic_data = titanic_data.dropna(subset=['Survived'])
```

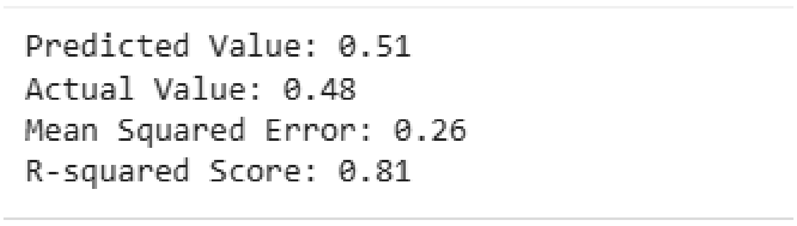
```python
X = titanic_data[['Pclass', 'Sex', 'Age', 'Siblings/Spouses Aboard',
'Parents/Children Aboard', 'Fare']]
y = titanic_data['Survived']

X.loc[:, 'Sex'] = X['Sex'].map({'female': 0, 'male': 1})

X.loc[:, 'Age'].fillna(X['Age'].median(), inplace=True)

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

rf_classifier = RandomForestClassifier(n_estimators=100,
random_state=42)

rf_classifier.fit(X_train, y_train)

y_pred = rf_classifier.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
classification_rep = classification_report(y_test, y_pred)

print(f"Accuracy: {accuracy:.2f}")
print("\nClassification Report:\n", classification_rep)

sample = X_test.iloc[0:1]
prediction = rf_classifier.predict(sample)

sample_dict = sample.iloc[0].to_dict()
print(f"\nSample Passenger: {sample_dict}")
print(f"Predicted Survival: {'Survived' if prediction[0] == 1 else 'Did
Not Survive'}")
import pandas as pd
from sklearn.datasets import fetch_california_housing
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, r2_score

california_housing = fetch_california_housing()
california_data = pd.DataFrame(california_housing.data,
columns=california_housing.feature_names)
california_data['MEDV'] = california_housing.target

X = california_data.drop('MEDV', axis=1)
y = california_data['MEDV']

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

rf_regressor = RandomForestRegressor(n_estimators=100, random_state=42)

rf_regressor.fit(X_train, y_train)

y_pred = rf_regressor.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

single_data = X_test.iloc[0].values.reshape(1, -1)
```

```
predicted_value = rf_regressor.predict(single_data)
print(f"Predicted Value: {predicted_value[0]:.2f}")
print(f"Actual Value: {y_test.iloc[0]:.2f}")

print(f"Mean Squared Error: {mse:.2f}")
print(f"R-squared Score: {r2:.2f}")
import requests

url =
"https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/stuff/titan
ic.csv"
response = requests.get(url)

with open('titanic.csv', 'wb') as f:
    f.write(response.content)

print("titanic.csv downloaded successfully.")
```

Output :



```
Accuracy: 0.76

Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.82      0.81       111
           1       0.69      0.67      0.68        67

    accuracy                           0.76       178
   macro avg       0.75      0.75      0.75       178
weighted avg       0.76      0.76      0.76       178


Sample Passenger: {'Pclass': 1, 'Sex': 1, 'Age': 47.0, 'Siblings/Spouses Aboard': 0, 'Parents/Children Aboard': 0, 'Fare': 30.5}
Predicted Survival: Did Not Survive
```



```
Predicted Value: 0.51
Actual Value: 0.48
Mean Squared Error: 0.26
R-squared Score: 0.81
```

7. Conclusion

Random Forest is a powerful, flexible machine learning model that leverages the strength of multiple decision trees to make robust predictions. Its use of randomization and ensemble learning makes it resistant to overfitting and highly accurate across a range of tasks. Despite its interpretability trade-offs, it remains a go-to method for many real-world data science applications.