Assignment – 9

B.Rithwik

2303A52330

Batch – 35

Question - 1

```
import pandas as pd
from google.colab import drive

import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score

file_path = '/content/drive/MyDrive/SML Dataset/breast_cancer_survival.csv'
data = pd.read_csv(file_path)

label_encoder = LabelEncoder()

data['Patient_Status'] = label_encoder.fit_transform(data['Patient_Status'])

categorical_columns = ['Gender', 'Tumour_Stage', 'Histology', 'ER status', 'PR status', 'HER2
status', 'Surgery_type']
for col in categorical_columns:
    data[col] = label_encoder.fit_transform(data[col])

X = data.drop(['Patient_Status', 'Date_of_Surgery', 'Date_of_Last_Visit'], axis=1)
y = data['Patient_Status']

model = RandomForestClassifier()

test_sizes = [0.2, 0.3, 0.4]
results = {}

for test_size in test_sizes:
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=test_size,
random_state=42)
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
```

```
    accuracy = accuracy_score(y_test, y_pred)
    results[f'Test Size {test_size}'] = accuracy

print("Accuracy for different test sizes:")
for test_size, accuracy in results.items():
    print(f"{test_size}: {accuracy * 100:.2f}%")
```

OUTPUT –

Accuracy for different test sizes:
Test Size 0.2: 76.12%
Test Size 0.3: 78.22%
Test Size 0.4: 76.87%

Question – 2

```
import pandas as pd
from google.colab import drive


import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

file_path = '/content/drive/MyDrive/SML Dataset/breast_cancer_survival.csv'
data = pd.read_csv(file_path)

label_encoder = LabelEncoder()
data['Patient_Status'] = label_encoder.fit_transform(data['Patient_Status'])

categorical_columns = ['Gender', 'Tumour_Stage', 'Histology', 'ER status', 'PR status', 'HER2
status', 'Surgery_type']
for col in categorical_columns:
    data[col] = label_encoder.fit_transform(data[col])

X = data.drop(['Patient_Status', 'Date_of_Surgery', 'Date_of_Last_Visit'], axis=1)
y = data['Patient_Status']

model = LogisticRegression(max_iter=1000)

test_sizes = [0.2, 0.3, 0.4]
results = {}

for test_size in test_sizes:
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=test_size,
random_state=42)
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    results[f'Test Size {test_size}'] = accuracy

print("Accuracy for different test sizes using Logistic Regression:")
for test_size, accuracy in results.items():
    print(f"{test_size}: {accuracy * 100:.2f}%")
```

OUTPUT -

```
Accuracy for different test sizes using Logistic Regression:
Test Size 0.2: 77.61%
Test Size 0.3: 78.22%
Test Size 0.4: 77.61%
```