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%