MCA-419(B) Data Science Practical – 8 Important Programs with Viva Questions & Answers

Question 1: Calculate Mean, Median and Variance using NumPy

Aim: To calculate mean, median and variance of given data using NumPy library.
Algorithm / Steps:
• Import NumPy library
• Create a list of numbers
• Use mean(), median() and var()
• Display the results
Program:

```
import numpy as np

data = [10, 20, 30, 40, 50]

print("Mean =", np.mean(data))
print("Median =", np.median(data))
print("Variance =", np.var(data))
```

Probable Viva Questions & Answers:

Q: What is NumPy?
A: NumPy is a Python library used for numerical and scientific computing.
Q: What is mean?
A: Mean is the average of all values.
Q: What is variance?
A: Variance measures how far values are spread from the mean.
Q: Which function is used for median?
A: np.median() function is used.

Question 2: Load CSV file and perform DataFrame operations using Pandas

Aim: To load a CSV file and perform basic DataFrame operations using Pandas.
Algorithm / Steps:
• Import pandas library
• Load CSV using read_csv()
• Display records
• Apply filtering
Program:

```
import pandas as pd

df = pd.read_csv("data.csv")

print(df.head())

print(df[df['Age'] > 20])
```

Probable Viva Questions & Answers:

Q: What is Pandas?
A: Pandas is a Python library used for data manipulation and analysis.
Q: What is a DataFrame?
A: A DataFrame is a 2D table-like data structure.
Q: Which function loads CSV file?
A: read_csv() function loads CSV file.
Q: What is filtering?
A: Selecting data based on condition.

Question 3: Normalize and Standardize data using Scikit-learn

Aim: To normalize and standardize numerical data using Scikit-learn.
Algorithm / Steps:
• Import preprocessing module
• Create dataset
• Apply StandardScaler
• Display output
Program:

```
from sklearn.preprocessing import StandardScaler
import numpy as np

data = np.array([[10],[20],[30],[40]])
scaler = StandardScaler()
print(scaler.fit_transform(data))
```

Probable Viva Questions & Answers:

Q: What is standardization?
A: Scaling data to have mean 0 and variance 1.
Q: Which class is used for standardization?
A: StandardScaler is used.
Q: Why scaling is needed?
A: To improve model performance.
Q: What is normalization?
A: Scaling data between 0 and 1.

Question 4: Encode categorical data using Label Encoding

Aim: To convert categorical data into numerical form.
Algorithm / Steps:
• Import LabelEncoder
• Create categorical list
• Apply fit_transform()
• Display output
Program:

```
from sklearn.preprocessing import LabelEncoder

colors = ['Red','Blue','Green','Blue']

le = LabelEncoder()

print(le.fit_transform(colors))
```

Probable Viva Questions & Answers:

Q: What is Label Encoding?
A: Converts categories into numbers.
Q: Why encoding is needed?
A: ML models work with numbers only.
Q: Which method converts data?
A: fit_transform()
Q: Name another encoding method.
A: One-Hot Encoding.

Question 5: Implement Simple Linear Regression

Aim: To implement simple linear regression model.
Algorithm / Steps:
• Import libraries
• Prepare dataset
• Train model
• Predict output
Program:

```
from sklearn.linear_model import LinearRegression
import numpy as np

x = np.array([1,2,3,4,5]).reshape(-1,1)
y = [2,4,6,8,10]

model = LinearRegression()
model.fit(x, y)
print(model.predict([[6]]))
```

Probable Viva Questions & Answers:

Q: What is regression?
A: Finding relationship between variables.
Q: What is dependent variable?
A: Output variable.
Q: Which library is used?
A: Scikit-learn.
Q: What does predict() do?
A: Predicts output.

Question 6: Implement Multiple Linear Regression

Aim: To implement multiple linear regression model.
Algorithm / Steps:
• Load dataset
• Split variables
• Train model
• Predict output
Program:

```
from sklearn.linear_model import LinearRegression
import pandas as pd

data = {'x1':[1,2,3], 'x2':[2,3,4], 'y':[3,5,7]}
df = pd.DataFrame(data)

X = df[['x1','x2']]
y = df['y']

model = LinearRegression()
model.fit(X, y)
print(model.predict([[4,5]]))
```

Probable Viva Questions & Answers:

Q: What is multiple regression?
A: Regression with more than one input.
Q: What is independent variable?
A: Input variable.
Q: Which function trains model?
A: fit()
Q: Which function predicts output?
A: predict()

Question 7: Implement Logistic Regression

Aim: To implement logistic regression for classification.
Algorithm / Steps:
• Import model
• Prepare dataset
• Train model
• Predict class
Program:

```
from sklearn.linear_model import LogisticRegression

X = [[1],[2],[3],[4]]
y = [0,0,1,1]

model = LogisticRegression()
model.fit(X, y)
print(model.predict([[5]]))
```

Probable Viva Questions & Answers:

Q: What is logistic regression?
A: Used for classification problems.
Q: Output of logistic regression?
A: Binary class (0 or 1).
Q: Which library is used?
A: Scikit-learn.
Q: What does fit() do?
A: Trains the model.

Question 8: Plot Histogram and calculate Skewness & Kurtosis

Aim: To visualize data and calculate skewness and kurtosis.
Algorithm / Steps:
• Import libraries
• Create dataset
• Plot histogram
• Calculate skewness & kurtosis
Program:

```
import matplotlib.pyplot as plt
import pandas as pd

data = [10,20,20,30,40,50]

plt.hist(data)
plt.show()

print("Skewness:", pd.Series(data).skew())
print("Kurtosis:", pd.Series(data).kurt())
```

Probable Viva Questions & Answers:

Q: What is histogram?
A: Graph showing data distribution.
Q: What is skewness?
A: Measure of asymmetry.
Q: What is kurtosis?
A: Measure of peakedness.
Q: Which library plots histogram?
A: Matplotlib.