SYMBIOSIS INSTITUTE OF TECHNOLOGY
DEPARTMENT OF AI & ML

CA- STAIML (2023-2027)

Deadline: 11/04/2024, 7:00 p.m.

Question 1	**R-based programming**	
(a)	Consider a time-series dataset from link and create a R function that returns the following (a) Mean (b) Median (c) Std Deviation (d) Variance (e) Skewness and Kurtosis (f) Histogram (g) Scatter plot (h) Q1, Q2, Q3, and Q4 (i) Pearson's Correlation Coefficient (j) Box plot	CO 1
(b)	Convert the above calculated results in form a dataframe for all possible numeric attributes of the dataset thereof.	CO 2
(c)	Demonstrate and explain the logical reasoning behind the working of for-loop and while loop in R with a suitable example.	
(d)	Consider a time-series dataset from link and do as directed (i) Calculate the correlation coefficient between two continuous variables. (ii) Determine whether the correlation is statistically significant. (iii) Visualize the relationship between variables using a scatter plot.	
Question 2	**Python-based programming**	
(a)	Write a Python class using OOPs concept that returns the moving average of a time-series sequence. The class instance should accept moving average window as the only input argument.	
(b)	(i) Calculate the mean, median, mode, standard deviation, and range of a given dataset using Python libraries such as NumPy and Pandas. (ii) Generate summary statistics for a dataset, including count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum values.	
(c)	(i) Fit a linear regression model to predict a dependent variable based on one or more independent variables using libraries like scikit-learn. (ii) Interpret the coefficients of the linear regression model. (iii) Evaluate the goodness of fit of the model using metrics such as R-squared, adjusted R-squared, and root mean squared error (RMSE). Use the model to make predictions on new data points.	
(d)	(i) Create a histogram to visualize the distribution of a numerical variable. (ii) Plot a scatter plot to explore the relationship between two continuous variables. (iii) Generate a box plot to compare the distribution of a numerical variable across different categories. (iv) Create a bar chart to visualize the frequency of categorical variables.	

Question 3	PowerBI-based programming	
(a)	Given a dataset (https://www.kaggle.com/datasets/milanzdravkovic/pharma-sales-data) The dataset is built from the initial dataset consisted of 600000 transactional data collected in 6 years (period 2014-2019), indicating date and time of sale, pharmaceutical drug brand name and sold quantity, exported from Point-of-Sale system in the individual pharmacy. Selected group of drugs from the dataset (57 drugs) is classified to the following Anatomical Therapeutic Chemical (ATC) Classification System categories: M01AB - Anti-inflammatory and antirheumatic products, non-steroids, Acetic acid derivatives and related substances M01AE - Anti-inflammatory and antirheumatic products, non-steroids, Propionic acid derivatives N02BA - Other analgesics and antipyretics, Salicylic acid and derivatives N02BE/B - Other analgesics and antipyretics, Pyrazolones and Anilides N05B - Psycholeptics drugs, Anxiolytic drugs N05C - Psycholeptics drugs, Hypnotics and sedatives drugs R03 - Drugs for obstructive airway diseases R06 - Antihistamines for systemic use Sales data are resampled to the hourly, daily, weekly and monthly periods. Data is already pre-processed, where processing included outlier detection and treatment and missing data imputation. Construct a PowerBI dashboard for an intelligent business analytics task.	
(b)	Consider a financial dataset https://www.kaggle.com/datasets/qks1lver/financial-data-of-4400-public-companies . These are data scraped from Yahoo Finance. Annually and quarterly balance sheets, cash flow statements, income statements over the past 4 years (for annual data) or the past 4 quarters (for quarterly data). **Construct a PowerBI dashboard for an intelligent business analytics task.**	