

**MACHINE LEARNING ALGORITHM USING PYTHON**

**COMPREHENSIVE EXAMINATION AY-2021**

**Date: 19 October 2021 Time: 04:15 PM to 06:15 PM**

**Term: IV Total Marks: 40**

**For Part A – Login in to SOCRATIVE. Time 15 Minutes. 20 Questions = 10 Marks.**

**For Part B: Answers to be Coded in Python (colab). Time: 1 Hour & 45 Minutes**

**Part B - THREE Questions: 3Q \* 10 Marks = 30 Marks.**

**FOR DATASET: Refer the folder “MLAP ET DATASETS October 2021”**

NOTE:

* Each and every line of command should have your comment. Eg: import pandas as pd. # loading files as Data frames, data manipulation and analysis.
* The First line of Code should have
  1. “Question Number”
  2. “YOUR FULL NAME” and
  3. ID NO” as Header.
* All answers are to be written in the Jupiter Notebook itself as Comments with Appropriate Question number and heading.
* Assume necessary parameters if it is not given explicitly in Question paper.

**ANSWER THE FOLLOWING QUESTIONS & SAVE YOUR FILES IN JUPYTER NOTEBOOK (Colab). < Q.NO-YOURNAME- ID NO > Eg: <1-Mahesh-18HPA1001>**

* You can either attend Question I (Logistic Regression) [OR] Question II. (MLR).
* You can either attend Question III (K- Means Clustering) [OR] Question IV (PCA)
* You can either attend Question V (Random Forest) [OR] Question VI (Recommendation system)
* Each Question should be stored as separate Files (. ipynb files)

**PART-B (30 Marks)**

1. Use there is **hiring.csv**. This file contains hiring statics for a firm such as experience of candidate, his written test score and personal interview score. Based on these 3 factors, HR will decide the salary. Given this data, you need to build a machine learning model for HR department that can help them decide salaries for future candidates.
2. Create descriptive statistics of all features using summary statistics and graphs
3. Build a new multiple linear regression model (80:20 - Train-Test Split) using only the significant features.
4. Explain the model validity using the results (Adjusted R2, RMSEA, model accuracy, etc) and explain the diagnostics analysis
5. Using this predict salaries for following candidates: (i) 5 yr experience, 8 test score, 7 interview score (ii) 10 yr experience, 9 test score, 8 interview score

**(OR)**

1. Use **HR\_data set** to build logistics regression model.
2. Now do some exploratory data analysis to figure out which variables have direct and clear impact on employee retention (i.e., whether they leave the company or continue to work)
3. Build a new Logistics linear regression model (80:20 - Train-Test Split) using only the significant features.
4. Explain the model validity using the results (Adjusted R2, RMSEA, model accuracy, etc) and explain the diagnostics analysis
5. Measure the accuracy of the model and explain the confusion matrix
6. DATASET: **customerspends.csv** [Cluster Analysis]
   1. Create a scatter plot to depict the customer spends on apparel and beauty and healthcare products and identify the number of clusters existing in the dataset.
   2. Normalize the features using StandardScaler and plot them in the scatter plot again.
   3. Use dendrogram and elbow method to verify if the number of clusters suggested is same as clusters visible in the scatter plot in Question A & B.
   4. Create the number of clusters as suggested by the elbow method using K-means algorithm and print the records of customers in each cluster and the cluster center of each cluster separately. Explain the clusters intuitively.

**(OR)**

1. Execute PCA on Heart disease dataset **heart.csv.**
2. Remove outliers using Z score. Usual guideline is to remove anything that has Z score > 3 formula or Z score < -3
3. Normalize the features using StandardScaler and plot them in the scatter plot again.
4. Create multiple components using PCA and create regression analysis using the created components in step C.
5. Use **Position\_Salaries.csv** to predict position salaries using RandomForestRegressor.
6. Measure prediction score using default n\_estimators (10)
7. Measure the accuracy of the model with explanations

**(OR)**

1. Use **movies.csv** and **ratings.csv** datasets to build movies recommendations system
2. Combine data from movies and ratings datasets.
3. Perform EDA (Exploratory Data Analysis)
4. Filter out the ratings records for the movies that belong to only “Action” genre
5. Calculation correlation matrix
6. Then recommend top 5 similar movies to the following movies: a. Heat and b. Eraser