ML3 Internal Graded Assessment

We combined two datasets in single CSV file. Use the below script to separate the data.

Use data_clust for dimensionality reduction and clustering

Use data_recom for recommendation system.

```
import pandas as pd
data = pd.read_csv('dataset.csv')
data_clust = data.iloc[0:897,0:15]
data_recom = data.iloc[:,15:]
```

The column 'M3' talks about the relay position, which is the target column. Don't use this while doing clustering and dimensionality reduction.

SECTION A

1. Data Preprocessing (5 marks)

a. Read the dataset and perform required cleaning and preprocessing prior to model building. Justify the pre-processing approaches used. (2 Marks)

b. Perform at least one univariate and bivariate analysis for the data_clust (2 Marks)

c. Keep the input features of data_clust (independent variables) in the variable "inp_data_dime: and output feature (M3) in the variable out. (1 Mark)

SECTION B

Answer the following questions (10 marks)

2. Apply K means clustering and identify the ideal value of K using elbow and silhouette method.

SECTION C

Answer the following questions (10 marks)

3. Apply PCA on the data. How many PCs are required to reproduce the 95% characteristics of original data. What is the top 5 features contributing in PC1? (5 marks)

4. Build the following ML model and compare its performance: (5 Marks)

a. ML model with original inp_data_dime and out

b. ML model with inp_data_dime_pca and out

(Note: the PCA components (inp_data_dime_pca) must capture the 95 percent variance in the data)

SECTION D

5. Recommendation Systems (15 marks)

a. Build the popularity-based recommendation system and suggest top 5 items. (5 Marks)

b. Build collaborative recommendation engine to recommend a top product/item to the specific user. Measure the model quality in terms of RMSE. (10 Marks)

We combined two datasets in single CSV file. Use the below script to separate the data.

Use data_dime for dimensionality reduction and data_recom for recommendation system.

```
import pandas as pd

data = pd.read_csv('dataset3.csv')

data_dime = data.iloc[0:5891,0:30]

data_recom = data.iloc[:,31:34]
```

SECTION A

1. Data Preprocessing (5 marks)

a. Read the dataset and perform required cleaning and preprocessing prior to model building. (1 MARK)

b. Calculate five-point summary for numerical variables. Summarize observations for categorical variables – no. of categories, % observations in each category. (1 MARK)

c. Perform univariate and bivariate analysis (2 MARKS)

d. Scale / Transform/ clean the data so that it is suitable for model building. Drop "SalePrice" before using clustering methods, as this is the target attribute. (1 MARK)

SECTION B

Answer the following questions (20 marks)

2. a. Use inp_data_dime. Apply PCA and compute all the possible principle components (PCs). How many PCs are required to reproduce the 95% characteristics of original data. Plot it with appropriate diagram. Also print the top 5 eigen vectors (5 marks)

2.b. Create a random matrix (M) of size 20 x 8 and compute singular values, left singular matrix and right singular matrix using Singular Value Decomposition. Try to reproduce the M back using singular values and vectors. (5 Marks)

2.c. Apply SVD on inp_data_dime and compare the SVD transform data with PCA transformed data. Also compare the top 5 singular vectors with eigen vector. How many Singular vectors are required to reproduce the 95% charecteristics of original data. (5 Marks)

2. d. Clustering: Use PCA dimensions to cluster the data. Apply K-means and Agglomerative clustering. (5 Marks)

SECTION C

3. Recommendation Systems (15 marks)

a. Build the popularity-based recommendation system and suggest top 5 items. (5 Marks)

b. Build collaborative recommendation engine to recommend a top product/item to the specific user. Measure the model quality in terms of RMSE. (10 Marks)