**Open Ended Lab – Data Warehousing and Mining**

**Objective:** To test the data mining analytical skills of students to solve the problem by using the knowledge they have gained in their previous labs

**Time Required**: 3 hrs

**Programming Language**: Python

**Software Required**: Anaconda

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Task:**

The automobile manufacturer is seeking to identify the closest competitors to their newly developed vehicle prototypes before launching the new model. To achieve this, they need to group existing vehicles on the market based on similarities, determine which group is the most similar to the prototypes, and use this information to identify the primary competitors for their new model.

The objective is to utilize clustering techniques to identify clusters of vehicles that possess unique characteristics. This analysis will provide an overview of the current market of vehicles and aid manufacturers in deciding on the development of new models based on the identified distinct clusters.

You can download the dataset from the link given below:

<https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data/CognitiveClass/ML0101ENv3/labs/cars_clus.csv>

Build your own pipeline and justify it. Also show the implementation and results of your solution through code.

**Rubrics for Evaluation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameter** | **Poor**  **(0)** | **Weak**  **(1)** | **Good**  **(2)** | **Excellent**  **(3)** |
| **Data Preparation**  **CLO-2**  **(C-3)** | The student did not perform any data cleaning, or the cleaning process is completely incorrect or inadequate. | The student attempted to clean the data, but the result is mostly incorrect or inadequate. | The student successfully cleaned the data but with some errors or omissions, or some parts are incomplete or unclear. | The student successfully cleaned the data, demonstrating a good understanding of data cleaning techniques and best practices. |
| **Data Analysis**  **CLO-2**  **(C-3)** | No attempt made to analyze the data or inappropriate statistical techniques were used. | Inappropriate statistical techniques are used, or the understanding of the statistical analyses performed is incomplete | The understanding of the statistical analyses performed is mostly clear and accurate. | Appropriate statistical techniques are used to analyze the data. The understanding of the statistical analyses performed is clear and accurate. |
| **Data Modeling**  **CLO-2**  **(C-3)** | The student did not demonstrate any understanding of data modeling concepts. | The student demonstrated a poor understanding of data modeling concepts. | The student demonstrated a good understanding of data modeling concepts. | The student demonstrated an excellent understanding of data modeling concepts. |
| **Model Selection and Building**  **CLO-2 (C-3)** | The student does not build any data mining models, or the building process is completely incorrect or inadequate. | The student attempts to build data mining models but the result is mostly incorrect or inadequate. | The student successfully built data mining models but with some errors or omissions, or some parts are incomplete or unclear. | The student successfully built accurate, robust, and interpretable data mining models, demonstrating a good understanding of model building techniques and best practices. |
| **Model Evaluation**  **CLO-2**  **(C-3)** | The student does not use any evaluation metrics, or the selected metrics are completely incorrect or inadequate. | The student attempted to use evaluation metrics such as Accuracy, F1 etc. | The student successfully used evaluation metrics but with some errors or omissions, or some parts are incomplete. | The student successfully used appropriate evaluation metrics, demonstrating a good understanding of evaluation metrics and their interpretation. |