

Data Visualization Project

Deliverables:

➤ Each team has to use the assigned dataset to their project to deliver the following:

1. Load the dataset and show the first ten rows.
2. Describe the dataset and show:
 - a. The type of distribution of the data.
 - b. If there are any outliers in the data, show them.
 - c. Show the top three columns containing high variety in the number of categories in case of categorical features in the data; or the top three columns with the highest variance in case of all features were numeric.
 - d. If there are any correlations between variables.
3. Create a meaningful interactive dashboard for the project:
 - a. It should contain at least two charts.
 - b. The dashboard should include a drop-down list or radio button that allows the user to select different values. Based on the selected value, the shape of the charts should change.
4. According to your dataset, apply at least 4 preprocessing steps that should be performed.
5. After applying preprocessing steps, build a machine learning model (Classification or regression) according to your dataset and the type of the problem.
6. Evaluate the model with the appropriate measure and save it as a .pkl file
7. Write a function to be called for predicting new coming rows.
8. Import the dataset into sql server as a table and write some analytics queries (window functions) that extract important insights from the data.

Bonus (2 Marks)

1. Develop web application with two main functionalities (train, predict)
2. In python develop flask api. One for train and one for test
3. If you choose train from web application, train api will be called to apply the ML pipeline and save the model as pkl
4. In case you select predict from web application, predict api will be called to predict the new given rows.

Datasets Description:

Dataset	Description
Dataset1	Dataset1 is a classification problem where the target column named "class".
Dataset2	Dataset2 is a classification problem where the target column named "Churn".
Dataset3	Dataset3 is a classification problem where the target column named "country_destination".
Dataset4	Dataset4 is a classification problem where the target column named "Credit_Score".
Dataset5	Dataset5 is a classification problem where the target column named "sars-cov-2_exam_result".
Dataset6	Dataset6 is a classification problem where the target column named "Class".
Dataset7	Dataset7 is a classification problem where the target column named "y".
Dataset8	Dataset8 is a classification problem where the target column named "diabetes".

General Instructions:

1. Students should work in groups of 3 to 4.
2. Each group should only use the assigned dataset.
3. All team members should work and understand all parts of the project.
4. Discussions will be held on 29th of December, on campus.