CTSE AI/ML Workflow Lab Activity

Due date : 11:55PM on 24th of March

This will be worth 1.5 marks of the total lab mark allocation.

This Activity Assumes that you have Followed the Self Guided Tutorial related to the Azure Machine Learning Studio shared last week.

Objective: Explore different modules and features of Azure Machine Learning Studio.

Main Modules in Azure Machine Learning Studio:

1. Data Import/Export Modules (e.g., Import Data, Export Data, Web Service Input/Output, etc.)
2. Data Transformation Modules (e.g., Clean Missing Data, Encode Categorical Data, etc.)
3. Feature Engineering Modules (e.g., Feature Hashing, Feature Scaling, etc.)
4. Machine Learning Modules (e.g., Train Model, Cross Validate Model, Score Model, etc.)
5. Evaluation Modules (e.g., Evaluate Model, Confusion Matrix, ROC Curve, etc.)

Data Import Export Module:

1. Supervised Learning:
   a. Adult Census Income Binary Classification Dataset
   c. Bank Marketing Binary Classification Dataset
   d. Breast Cancer Binary Classification Dataset
   e. Diabetic Retinopathy Binary Classification Dataset

2. Unsupervised Learning:
   a. Iris Dataset (Clustering)
   b. MNIST Dataset (Dimensionality Reduction)
   c. Anomaly Detection Dataset (Anomaly Detection)
   d. Recommendation Algorithm Dataset (Recommendation System)

Exploring Data

There are multiple ways for exploring data in Azure ML Studio the best and the easiest way is to follow the following steps.

1. Load your dataset into Azure Machine Learning Studio. You can do this by uploading your dataset into your workspace and creating a new dataset, or by using a pre-existing dataset from the Azure Machine Learning Studio's sample datasets.
2. Once your dataset is loaded, select it and click on the "Visualize" button by right clicking on the dataset.

3. You can use select a column to visualize the column data, and view the statistical information related to the data.
4. Make a note of things like Missing Values, unique values, and Feature Type.
    a. When we have too many missing values in a column it is difficult for the model to learn. Such column should be ideally removed from our dataset.
    b. If we have categorical data it is better to encode them.
    c. If we have numerical data it is better to scale them or standardize them to a common scale to speed up the training

Data Transformations Module:

1. Missing Values Treatment:

    - Replace with Mean/Median/Mode: This method replaces missing values with the mean/median/mode value of the column. For example, if you have missing values in a column containing age values, you can replace them with the mean age value.
    - Drop Rows/Columns: This method drops the rows or columns containing missing values. For example, if you have missing values in a row representing a customer, you can drop that row.

2. *Encoding Categorical Variables: (Not available in the Classic Version)*

    - *One-Hot Encoding: This method creates dummy variables for each category in a categorical variable. For example, if you have a categorical variable representing the type of car (e.g., sedan, SUV, truck), you can create three dummy variables representing each category.*
    - *Label Encoding: This method assigns a unique integer to each category in a categorical variable. For example, if you have a categorical variable representing the type of car, you can assign the integer 1 to sedan, 2 to SUV, and 3 to truck.*

3. Feature Scaling and Normalization:

    - *Min-Max Scaling: This method scales the values in a column to be between 0 and 1. For example, if you have a column representing income, you can scale the values to be between 0 and 1.*
    - Normalize: This method scales the values in a column to have zero mean and unit variance. For example, if you have a column representing height, you can scale the values to have zero mean and unit variance.

Activity:

1. **Choose a data set from the given supervised learning datasets.**
2. **Explore the dataset using the exploration method shown above.**

a. For each feature in the dataset visualize and check the statistical data.

b. Based on the Analysis and the statistical data,

i. What features do you think will be the most appropriate for the Model Training?

ii. What features should be cleaned?

1. Apply a suitable technique(Drop row, drop column, replace missing values)

iii. What features should be encoded?

iv. What features should be scaled?

1. Apply Normalization to scale the data

3. Compile findings into a pdf report and share your analysis. Include any screenshots as needed.

a. Rename the pdf with your IT Number and submit.