COS40007 Artificial Intelligence for Engineering

Portfolio Assessment-3: "Let's develop AI model by your own decision'"

Due: by Sunday of Week 5 (25/08/2024 23:59 PM) in Canvas

Aim

The aim of this task is for you to demonstrate your understanding of developing ML model by exploring ana analysing data by your own and comparing different ML models with different feature set and hyperparameter tuning. Then convert your ML to AI deployment.

About the dataset

Download the provided dataset (vegemite.csv)

This dataset contains machine process and machine settings data of vegemite production. To get a desired consistency of vegemite the process and settings value must meet some desired value. There are 3 values in the class 0 to 2 refer to different consistency level of solid in vegemite production.

Disclaimer: The dataset used in this task was originally collected for a funded research project by Bega Cheese. The dataset here is used solely for educational purposes and can be only used for completing activities of this studio. By any mean this dataset is not shareable to others or any public domain.

Step 1: Data Preparation

This dataset has more than 15000 data points. Let's take out 1000 data points out from this dataset that we can use to test real-time similar to what we did in Studio 4. To do this.

1) First you need to shuffle the dataset
2) Randomly take out 1000 data points (rows) such as way that each class in those 1000 samples has near equal distribution (e.g. at least 300 samples from each class)

Use the remaining 14000+ data points to train your ML model.

For constructing features answer the following question and fix if you find such problem in the dataset

1) Does the dataset have any constant value column. If yes, then remove them
2) Does the dataset have any column with few integer values? If yes, then convert them to categorial feature.
3) Does the class have a balanced distribution? If not then perform necessary undersampling and oversampling or adjust class weights.

4) Do you find any composite feature through exploration? If so, then add some composite feature in the dataset.
5) Finally, how many features you have in your final dataset?

Step 2: Feature selection, Model Training and Evaluation

6) Does the training process need all features? If not, can you apply some feature selection technique to remove some features? Justify your reason of feature selection
7) Train multiple ML models (at least 5 including DecisionTreeClassifier) with your selected features.
8) Evaluate each model with classification report and confusion matrix
9) Compare all the models across different evaluation measures and generate a comparison table.
10) Now select your best performing model to use that as AI. Justify the reason of your selection
11) Now save your selected model

Step 3: ML to AI

12) Now take the 1000 rows that you have not used (we put aside at the beginning)
13) Load the model
14) Iteratively convert columns in each row in the format of your training feature set
15) Find class prediction using the loaded model and compare with the original label
16) Measure the performance of your best model for 1000 unseen data points.
17) Now measure the performance of other model using these 1000 data points. Have you observed same result of model selection that you identified through evaluation ?

Step 4: Develop rules from ML model

In your feature set take only the columns the ends with 'SP' and remove others. SP are the set points that human can control. Others are process variable (PV) that is generated by machines and human can not control. Now generate some rules of recommended set points ranges for a class value.

- Using only SP features generate a decisiontree model
- Print the tree using export_text
- Can you now define some rules of SP values for each class ?
  For example,
  For class 1 (FFTE Production solids SP >= 39.5 and FFTE Production solids SP < 42)
- If you finalised some rules write them in the final submission document.

Submission

Create a folder and place all of your data file (including intermediate data file) and code in that folder. Then create a sharable link of that folder

The portfolios assessment submission should be a document (word or pdf) with the following

- Your name and Student number
- The studio class you attend (for example you attend Studio 1-1 then write Studio 1-1)
- Answer the questions for Step 1: Data Preparation **[2.5 marks]**
  (also provide link of your source code and data)
- Answer the questions for Step 2: Feature selection, Model Training and Evaluation **[3.5 marks]**
- Answer the questions for Step 3: ML to AI **[2 marks]**
- Step 4: Develop rules from ML model - provide the outcome of your decision tree and the decision rules you create after observing your ML model **[2 marks]**

Total **10 marks**