COS40007 Artificial Intelligence for Engineering

Portfolio Assessment 3: "Develop an AI model by your own decision'"

Due: by Sunday of Week 5 (06/04/2025 23:59 PM) in Canvas

Aim

This task aims to demonstrate your understanding of developing an ML model by exploring and analysing data on your own and comparing different ML models with different feature sets 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 the desired consistency of vegemite, the process and settings must meet some desired values. There are three values in the class 0 to 2 that refer to different consistency levels 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 only be used to complete activities for this studio. By any means, this dataset is not shareable with others or in any public domain.

Step 1: Data Preparation

This dataset has more than 15000 data points. Let us take out 1000 data points 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) in such a 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 it if you find such a 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 a 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 features through exploration? If so, then add some composite features to the dataset.
5) Finally, how many features do 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 techniques to remove some features? Justify your reason for feature selection.
7) Train multiple ML models (at least five, including DecisionTreeClassifier) with your selected features.
8) Evaluate each model with a 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 as AI. Justify the reason for 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 it with the original label.
16) Measure the performance of your best model for 1000 unseen data points.
17) Now, measure the performance of other models using these 1000 data points. Have you observed the same result of model selection that you identified through evaluation?

Step 4: Develop rules from the ML model

In your feature set, remove the columns that end with 'SP' and remove others. SP are the set points that humans can control. Others are process variables (PV) that machines generate, which humans cannot control. Now, generate some rules for recommended set point ranges for a class value.

- Using only SP features generates a decision tree 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 have finalised some rules, write them in the final submission document.

Submission

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

The portfolio 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 a link to 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 the 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**