COS40007- Artificial Intelligence for Engineering

Portfolio Assessment – 1

Name – Rupayan Banerjee
Student Number – 103538229

Dataset selected – Combined Cycle Power Plant Dataset

I am currently pursuing an undergraduate degree in Computer Science and majoring in Software Development and Artificial Intelligence. I selected the Electrical Engineering-based dataset since it correlates the most closely to my learning discipline. I have done a few units related to physics and they mostly involved the electrical sections, so this was the only one that made sense for me to select. Also, I wanted to explore more about the possibilities of introducing the power of AI in fields like this.

As for my EDA conducted in Studio 1, I will provide the details below. A separate document for it can also be found on my GitHub repository, which I will link to later in the document.

Analysing the Combined Cycle Power Plant database in details these are the main takeaways from its EDA:

- Exhaust Vacuum (V) and Ambient Temperature (AT) show strong positive correlations with PE (energy output), indicating they are key predictors for energy generation in the Steam Turbine.
- Ambient Pressure (AP) shows moderate correlations with PE, implying a fair bit of impact on the energy output
- Relative Humidity (RH) shows weak correlations with PE, implying limited direct impact on energy output.
- Ambient Pressure (AT) and Relative Humidity (RH) have a somewhat moderate correlation, which could be used later.
- Ambient Temperature (AT) and Exhaust Vacuum (V) have high correlation, indicating potential interactions between Gas and Steam Turbine performance.

In Studio 2, I went through class labelling for the target variable, which, in my case, was Electrical Energy Output (PE). I made sure I divided the range of values into 5 categorical sections ranging from below 430 being categorised as 1 and every subsequent section being 20 more than the last one. This was carefully done in accordance with the range and mean of the data at hand to make sure we don't end up with a clustered graph.

In the feature engineering section, I have normalised all the columns except PE since it is our target variable. Then I created 2 new composite features, AT_V and AT_RH, in accordance with the EDA conducted. Then in the feature selection section, I picked the features AT, V, AP and the composite features AT_V and AT_RH and of course the PE_Label for the categorical data format of PE, and put all of them in a single CSV file.

As for the training and decision tree model creation I followed through the example structure provided and created 5 different models for 5 different CSV file that I have already created. It showed me clearly which models are effective and which aren't in this scenario.

The final comparison table among all the 5 models that I had built is provided below. A separate document is also available on my GitHub repository, which I will provide the link to later in the document.

After performing the required actions to the data files and building 5 different models for a decisionTree classifier, we have the following results.

Iteration	Model 1	Model 2	Model 3	Model 4	Model 5
1	85.68%	85.13%	85.86%	85.27%	85.09%
2	85.20%	85.34%	85.72%	86.45%	85.30%
3	85.48%	85.51%	86.21%	85.61%	85.51%
4	85.30%	85.34%	85.75%	85.65%	85.27%
5	84.85%	85.41%	85.75%	86.24%	85.48%

From what we can observe in this table Model 3 comes out as a clear winner in terms of accuracy, since it has consistently been at the highest accuracy for 4 out of my 5 test iterations. When it comes to the lowest accuracy model, it is a bit more difficult since both models 1 and 5 came out to be the ones with the lowest accuracy 2 times each within the 5 test iterations.

Appendix

Link to all the code files - https://github.com/rupayan-banerjee/CodeBase.git