DM2024 ISA5810 Lab2 Homework Report

Data Processing and Feature Engineering

1. Text Representation

- TF-IDF Vectorizer:
  - Used TF-IDF (Term Frequency-Inverse Document Frequency) to convert text data into numerical representations.
  - Set the maximum features to 30,000, as experiments showed that using 5,000 features was insufficient for this competition.
- CountVectorizer:
  - Attempted CountVectorizer, but it included many unimportant words, making it less effective compared to TF-IDF.

2. Label Encoding

- Applied label encoding on the categorical data to transform it into a numerical format.
- Transformed data into a shape of “(,8)”.

Modeling

1. Initial Attempts

- DecisionTreeClassifier:
  - Tried a DecisionTreeClassifier for initial experiments.
  - The results were poor, indicating that the model lacked the complexity needed to capture the patterns in the data.

2. Neural Network (NN)

- Built a Neural Network to address the classification task.
  - Initial Configuration:
    - Started with 4 hidden layers.
    - Public leaderboard score: 0.4.
- Improvements:
  - Increased the number of hidden layers for better learning capacity.

- Implemented early stopping to prevent overfitting during training.

Future Directions

- Using Large Language Models (LLMs):
- Plan to explore LLMs (e.g., BERT, GPT) to extract more meaningful features from the text data, potentially enhancing classification performance.