1. GitHub link of your code

   https://github.com/danzel-crazy/2023-Machine-Learning.git

2. Reference if you used any code from other resources

   https://www.kaggle.com/code/samuelcortinhas/tps-aug-22-failure-prediction/notebook#5.-Preprocessing

   https://www.kaggle.com/code/azminetoushikwasi/classification-comparing-different-algorithms#5.9.-Voting-Classifier

3. Brief introduction

   I use logistic regression as my main model and also do some data preprocessing in order to reach the baseline score for this final project

4. Methodology (Data pre-process, Model architecture, Hyperparameters, ...)

   - For data pre-process, I first searched other people's notebooks which described this dataset. However, it turns out that the parameters are quite discret and have little obvious relation between each other. But I do find out some interesting stuff. First, there are many missing values in different product codes , so I use the method in sklearn to fill in the values. Second , some features are characters, so I need to give them labels.
   - For Model architecture, I first tried for a neural network , but the best score for me is 0.516 , so I will try for another one. Then I read the paragraph on the second resource. I discovered a cool tool - vote classifier, so I decided to use it. But the hard part is to choose the right algorithm to vote and the voting weights , my highest score is 0.592.Finally, I

    turn to discussion for help. I then use the most suitable model for this dataset - logistic regression and do some feature engineering so that I can reach the baseline in the end.

   - For features, I do find that some groups of the features are similar in the distribution, so I use their previous notebook for the features. And calculate the correlation for feature selection.

5. Summary

   I think this final project is quite challenging. I spend a lot of time browsing the data and trying to find a good pattern, and the choice of model also bothers me a lot. I have an idea at the end of this project. The idea is about the product code. Since the same product code have same value for the first four(?) features , if I can perfectly separate each product code and use it as prediction of the test data, maybe I will have a better score. I haven’t figured out how to use this characteristic, but I do think this idea has its potential. Finally, I find that I lack the ability to dive into the data and perform a good processing. I should try to gain data mining experience in the near future.