**Classification I: Discussion**

**Code:**

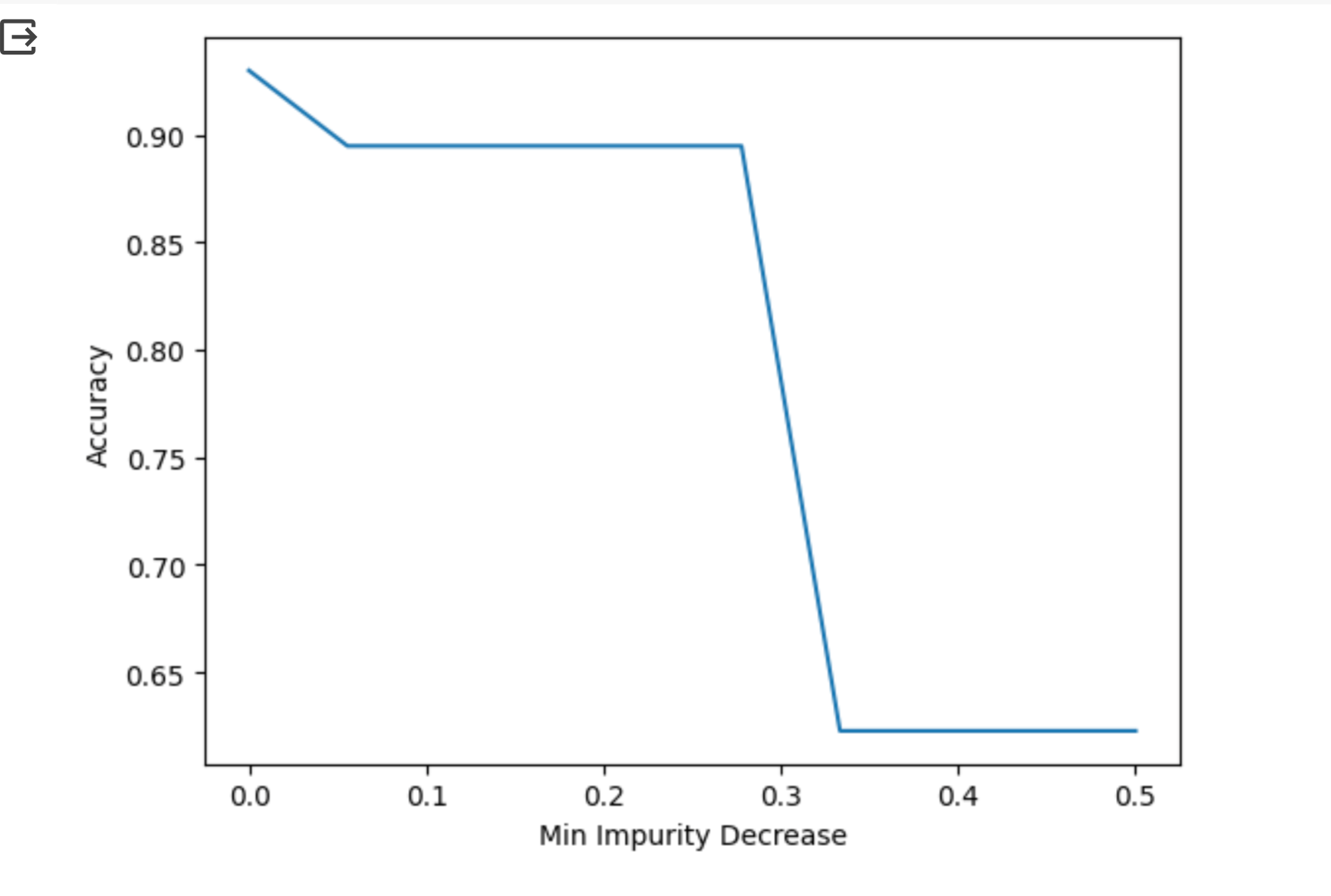
In Part A, we first perform data preprocessing by scaling the features between 0 and 1.

We then explored node impurity measures : Gini index and entropy) and used those values to determine which feature to split. The definitions of functions are put at the top.

In Part B, we learned about a built in package in Python: Scikit-learn, which we can directly build a decision tree. The training size is set to be 0.8 to ensure the model is well-trained with sufficient data. Metrics such as accuracy, sensitivity, specificity, and F1 score are calculated in the custom function and provide insights on the decision tree’s performance. And by evaluating the relationship between accuracy and min-impurity, we can get an idea of trande-off between model complexity and performance.

And in this assignment, when calculating the metrics, either the Gini index or the entropy of one feature, we use the for loop, which iterates through each feature and row. The time complexity is O(n^2). We may consider using vector operations, which apply calculations to a sequence of data, to speed up the process.

**Result:**



The graph indicates the relationship between decision tree accuracy and the minimum impurity decrease parameter. This parameter sets a threshold for early stopping in tree growth. The larger the parameter, the less the split and the simpler the tree construction.

Generally, the graph shows a decreasing pattern. As the parameter increases, the accuracy decreases. Specifically, when min\_impurity\_decreases from 0.0-0.2, and the required decrease in impurity is relatively low, the model has high accuracy. At point 0.3, the accuracy has a sharp drop from 0.9 to 0.6 and then comes to a flat.

We then consider choosing the parameter to be between 0.1 - 0.3.

**Discussion:**

It is crucial to learn that parameters like minimum impurity decrease affect tree performance. And when choosing a suitable model, we need to consider both bias, variance, and interpretability. And, although attributes in fitting the model are not equally important, by finding the most important attributes, we can better fit the model with great efficiency.

**Self-assessment:**

A or A-? I got a great understanding of the node impurity measurements, the confusion matrix, and a general picture of how the decision tree was implemented. I also clearly illustrated how model accuracy varies based on different decision tree parameters. Overall, I think I have mastered the knowledge well:)