Prepare Features :
Collects information about dataset website attempts called entries which are converted into numbers, decides if an entry is censored or not based on a scoring rule

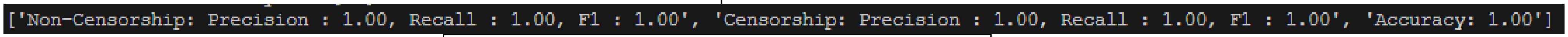
Feature Example :
Does the response from the website say 'BLOCKED"

Scoring Rule:
Adds points for signs of censorship i.e - 'BLOCKED' in response
Subtracts points for signs of non-censorship i.e - website matches a normal pattern
Labeling:
Based on the scoring rule the data is labeled as censorship or not

Split/Test
Data is split such as the manager class uses 70% of data for learning the decision tree, and 30% of the data for testing

Decision Tree
Looks at labels and builds a tree for example if response has 'BLOCKED' the tree goes one way - makes predictions like a flowchart

Performance
We can see how the decision tree model performed when tested or how model predicts whether it is a censorship or non-censorship with a few metrics.
Precision : How trustworthy the model predicts when it is either
Recall : How good the model is at finding all cases of either non-censorship/censorship
F1 : Combination of both Precision and Recall into one value
Accuracy : Overall percentage of correct predictions



```
['Non-Censorship: Precision : 1.00, Recall : 1.00, F1 : 1.00', 'Censorship: Precision : 1.00, Recall : 1.00, F1 : 1.00', 'Accuracy: 1.00']
```

Some Problems

- A perfect score can look good but there might be some problems such as :
- **overfitting** - Memorizing test data instead of learning
- **"cheating"** - Creates the labels using the same info the decision tree uses to predict
-
- **Example**: If "BLOCKED" in the website response adds a point to the score (making it more likely to be a scenario of censorship), the model learns to just check for "BLOCKED" to predict censorship. In a real scenario it might miss censorship that doesn’t follow the exact same clues

Fixes For Future

- Using a much larger datasets with variety of censorship patterns to challenge the model.
- Create labels using human judgment not the same clues the model uses, in other words assign labels based on human judgement not just the features that the model will rely on.
- Test the model on new and unseen data to ensure the model learns patterns and not just memorizes the patterns