ML LAB 3

Comparative Analysis Report

Name: Meenakshi Pathiyil
SRN: PES2UG23CS335

1. **Algorithm Performance**

- Mushroom dataset achieved the highest accuracy (100%). This is because the mushroom dataset is a perfectly separable classification task. Here, the features are highly correlated with the target.
- Usually the dataset size does play an important role - larger datasets provide training examples, which usually improves the algorithm's ability to generalize and thus reduces overfitting. However, dataset size alone does not guarantee better performance, there are other factors as well that contribute to the overall performance of an algorithm
- More features provide more potential splits, however, they also increase the tree complexity

2. **Data Characteristics Impact**

- When the data is imbalanced, the model tends to favor the majority class, since predicting the majority class leads to a higher accuracy. Due to this, the overall performance metrics appear high, but the metrics that treat all classes equally show weaker performance. This difference occurs because the model does not learn minority classes as effectively as it does the majority class
- Binary valued features usually work better as they simplify the splits and reduce the tree depth. Multi-valued features often lead to larger, more complex trees

3. **Practical Applications**

- The mushroom dataset type is relevant to cases like medical diagnoses, quality control, etc. The Tic-Tac-Toe dataset type is relevant to cases like board game engines. The nursery dataset type is relevant to cases like recommendation systems, patient triage, etc.
- The interpretability advantages for each domain are as follows:
    - Mushroom - The rules are simple and the splits are easy to follow

    - Tic-Tac-Toe - The tree is deeper and more complex, but still interpretable as the features are straightforward
    - Nursery - The tree is huge due to multiple attributes thus making the interpretability a lot lesser than the other dataset types
- The performance for the mushroom dataset is already really good and does not need further improvement. The tic-tac-toe dataset could you use an algorithm that can capture more complex interactions and generalize better. The nursery dataset could benefit from using feature selection to reduce the tree complexity