
School of Computer Science,
College of Science, University of
Tehran

Dr. Sajedi

Decision Trees

Deadline: 14 Farvardin 1402

1 Objective

The objective of this assignment is to familiarize you with decision tree classification using the Mushroom Data Set. You will practice data loading, preprocessing, basic model parameter tuning, basic feature selection, evaluation metrics computation, and result analysis.

2 Dataset Description

The dataset consists of descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family. It comprises 8124 samples with 22 features. The first column represents the label (2 different classes; edible or poisonous), and the subsequent 22 columns correspond to input features.

3 Instructions

1. Load the dataset and separate labels from features. Assign 'y' to labels and 'X' to features.
2. Shuffle the data and split it into training, validation, and test sets in a 70:15:15 ratio. Name them X_train, y_train, X_val, y_val, X_test, and y_test respectively.
3. Perform preprocessing on the data if required.
4. Search through different criteria and maximum depths to find the best parameters for your decision tree model; any extra parameters are considered a bonus. Experiment with different parameter sets and select the one with the best performance on the validation set. Avoid using the test set for parameter tuning.
5. Train the Decision Tree model using the best parameters obtained.

6. Search through at least 5 pairs of features and plot them based on those features, and select the 2 most separable features and describe your reasons in the report.

7. Include evaluation metrics (accuracy, precision, recall, and F1-score) for each parameter set in the report.

8. Include confusion matrices for both validation and test sets for all parameter sets.

9. Analyze at least three confusion matrices in your report, highlighting insights and observations.

4 Additional Guidance

- Ensure your code is in .ipynb format and thoroughly documented.
- Utilize `pandas.read_csv()` to read the data.
- Alongside your code, submit a report file containing a comprehensive analysis of your results.
- Utilize appropriate visualizations and statistical techniques to reinforce your analysis and conclusions.