

Applied machine learning: Decision Trees

Problem sheet 3

The Iris dataset is a well-known dataset in machine learning. It consists of 150 samples of iris flowers from three species: *Setosa*, *Versicolor*, and *Virginica*. Each sample contains four features:

- Sepal length (cm)
- Sepal width (cm)
- Petal length (cm)
- Petal width (cm)

The goal is to predict the species of an iris flower (*Setosa*, *Versicolor*, or *Virginica*) based on its features.

Tasks

1. **Dataset Splitting:** Load the Iris dataset and split it into training, validation, and test sets. Shuffle the dataset before splitting.
2. **Entropy Calculation:** Implement a function to calculate the entropy of a dataset given its class labels.
3. **Information Gain:** Define a function to compute the information gain of a potential split based on a feature and threshold.
4. **Data Splitting:** Write a function that splits the data into two subsets based on a feature index and a threshold value.
5. **Decision Tree Construction:** Build a decision tree using the following recursive procedure:
 - At each node, determine the best feature and threshold to split the data.
 - If all samples in a node belong to the same class or the maximum depth is reached, make the node a leaf and assign it the majority class label.
6. **Prediction:** Implement a function to make predictions for new data points using the constructed decision tree.
7. **Evaluation:** Evaluate the performance of the decision tree on the test set.