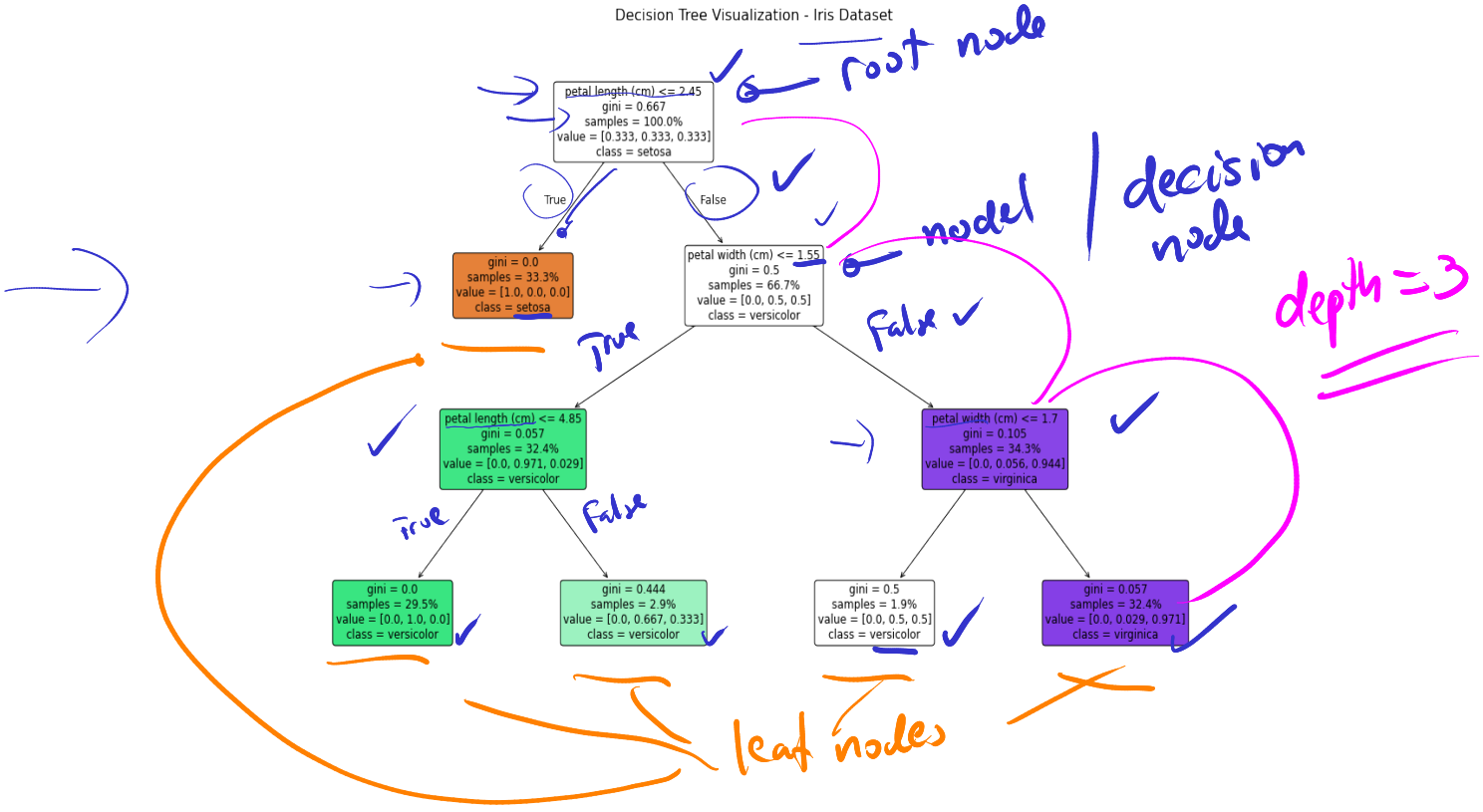
Introduction to Decision Trees

1. Introduction

Decision Trees are one of the most intuitive and interpretable machine learning algorithms. They can be used for both classification and regression tasks. The main idea is to split the data recursively based on feature values that lead to the most "pure" subsets.

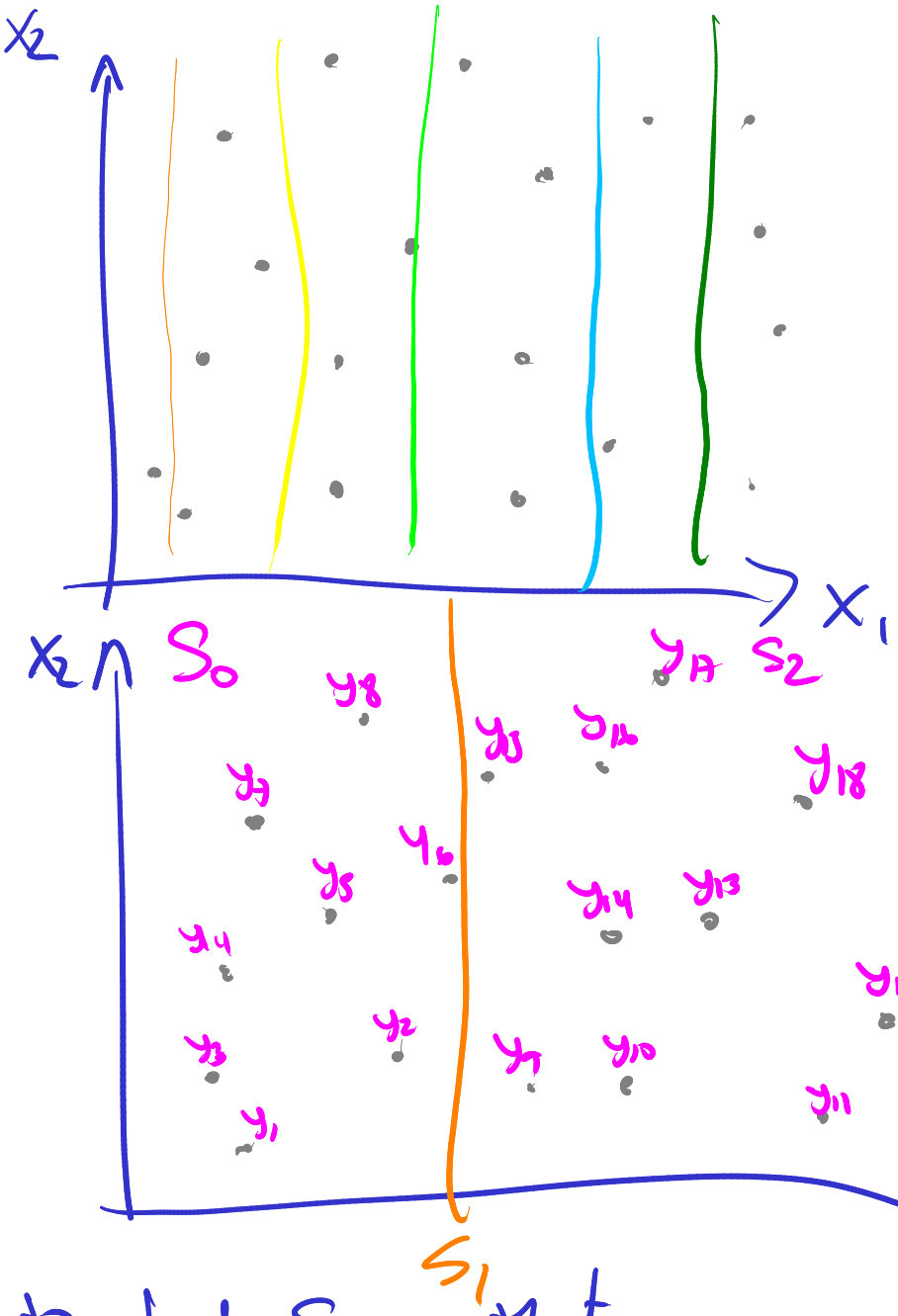
2. How They Work

A Decision Tree recursively divides the dataset into smaller subsets based on conditions on input features. Each node in the tree represents a decision rule, and each leaf node represents a final output (class label or regression value).



The most common criteria used for splits include:

- **Gini Impurity** – for classification
- **Entropy** (Information Gain) – for classification
- **Mean Squared Error (MSE)** – for regression



Regression

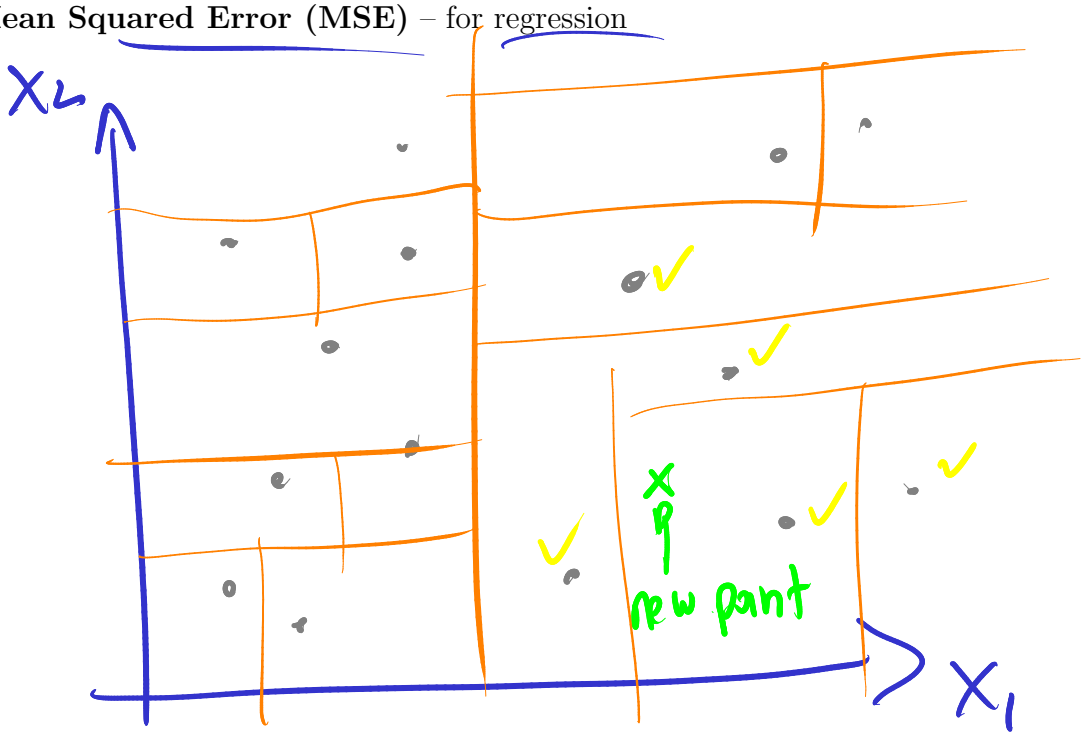
The goal is to find S_1 so that

The MSE of S_0 plus the MSE of S_1

is Minimized

The most common criteria used for splits include:

- **Gini Impurity** – for classification
- **Entropy** (Information Gain) – for classification
- **Mean Squared Error (MSE)** – for regression



3. Applications in Data Science

- **Customer Segmentation**: Classify customers into distinct groups based on behavior.
- **Credit Risk Modeling**: Assess the likelihood of loan repayment.
- **Medical Diagnosis**: Predict disease outcome based on symptoms.
- **Churn Prediction**: Identify customers likely to leave a service.

4. Advantages

- Easy to interpret and visualize.
- Handles both numerical and categorical data.
- No need for feature scaling.
- Works well even with non-linear relationships.

5. Disadvantages

- Prone to overfitting (especially deep trees).
- Small changes in data can lead to a completely different tree.
- Greedy nature may not yield optimal global tree.
- Less accurate than ensemble methods like Random Forests or Gradient Boosting.

6. Conclusion

Decision Trees are a powerful and interpretable tool for supervised learning. While they can overfit, especially on noisy data, they provide a strong foundation for more advanced ensemble models like Random Forests and Gradient Boosting.