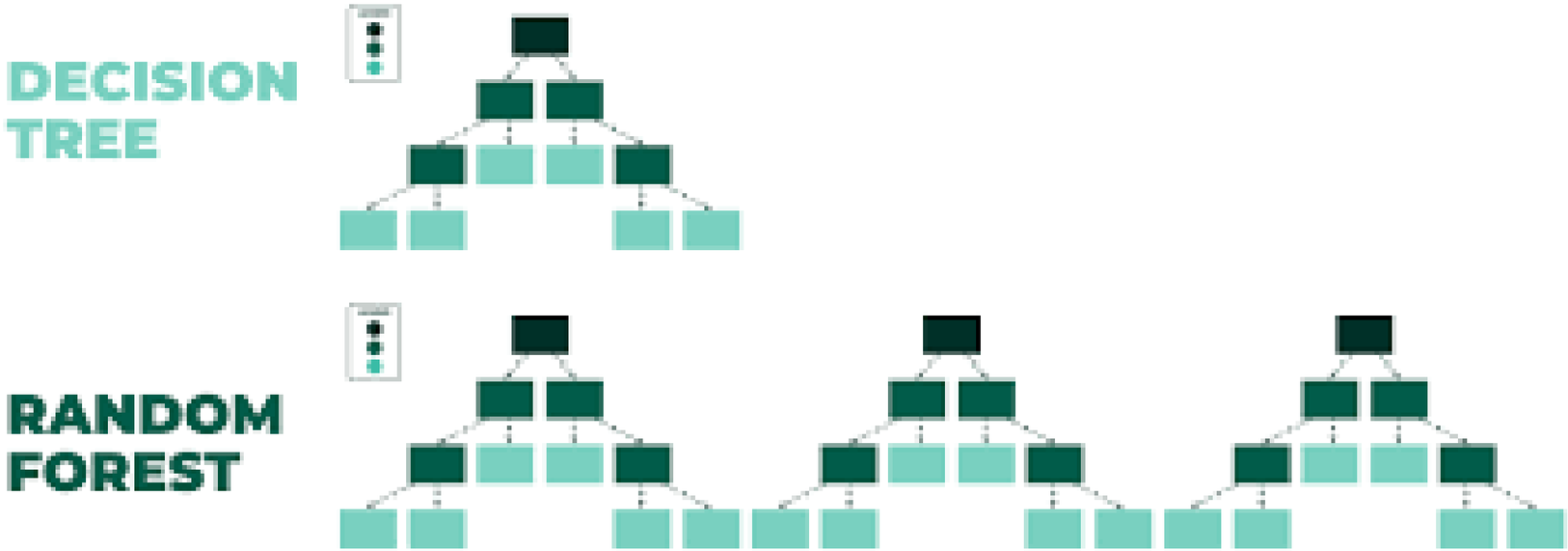
DSEM 6150 ASSIGNMENT 3

Decision Tree and Random Forest

Presented by: Ruchi Kapadiwala



Problem Statement

- A premier financial institution has embarked on a digital transformation initiative to streamline its loan approval process.
- With the objective of providing real-time loan eligibility to potential customers, the organization seeks to leverage customer information garnered from an online application form.

Data Description

This information encompasses various demographic details such as

- **Gender**
- **Age**
- **Residential Status**
- **Occupation**
- **Employment Tenure**
- **Financial Standing**
- **Etc.**

Solution

The technical approach for building decision trees and random forests involves two key steps:

01

Choosing a Splitting Criterion

Two common splitting criteria are used in decision tree algorithms:
(i) Entropy-based Information Gain
(ii) Gini Index.

02

Building the actual Decision Tree or Random Forest model

Entropy-based Information Gain

- Measures reduction in uncertainty of target variable after split
- Split with highest information gain is chosen as next split in tree
- High entropy = evenly distributed classes
- Low entropy = one class dominates

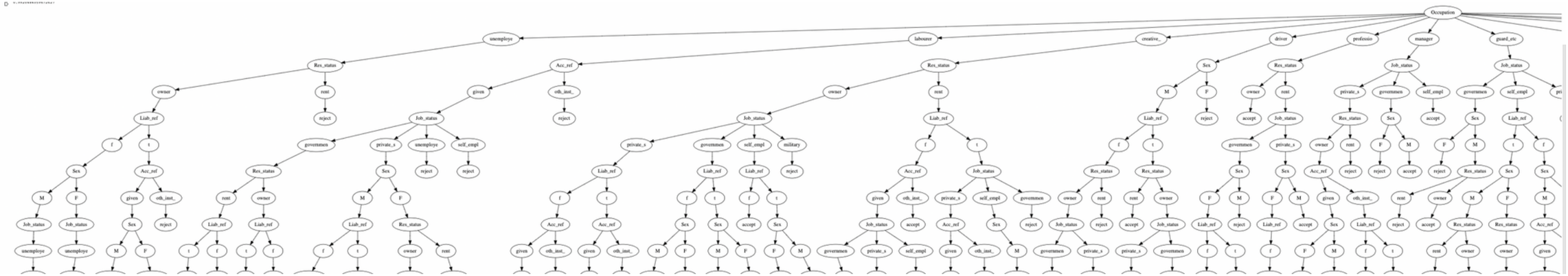
Gini Index

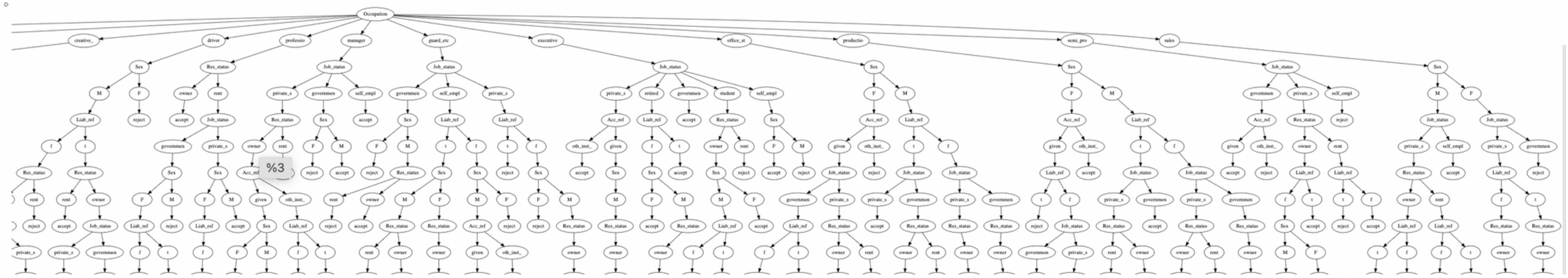
- Measures probability of misclassifying a randomly chosen instance after split
- Split with lowest Gini index chosen as next split in tree
- Ranges from 0 (perfect split) to 1 (entirely incorrect split)

Decision Tree

What is a Decision Tree?	How is it represented graphically?	Why is this algorithm so popular?	How is the tree constructed?
• An algorithm used in machine learning and data mining for prediction and classification tasks.	• As a tree-like structure, where : • Each internal node represents a test of an attribute • Each branch represents the outcome of the test • Each leaf node represents the final prediction	• Simplicity • Interpretability • Ability to handle both numerical and categorical data, missing values, and handle complex relationships between features.	• By recursively splitting the data into subsets based on the feature with the highest information gain to reduce uncertainty about the target variable.

Decision Tree Output





Random Forest

What is a Random Forest?	How does the Random Forest Algo work?	How is the final prediction made?	Why is this algorithm so popular?
• A machine learning algorithm that combines multiple decision trees to make a prediction.	• It works by constructing many decision trees (hence "forest") on different samples of the training data and combining their outputs to improve the overall accuracy and stability of the model.	• The final prediction is made by having each decision tree in the forest vote and choosing the prediction with the most votes	• Using multiple decision trees helps reduce overfitting, a problem in single decision tree models, by averaging out biases and capturing diverse relationships

Random Forest Output