Institute of Information Technology
Jahangirnagar University
Savar, Dhaka.

Assignment-2
PMIT6113: Data Mining and Knowledge Discovery

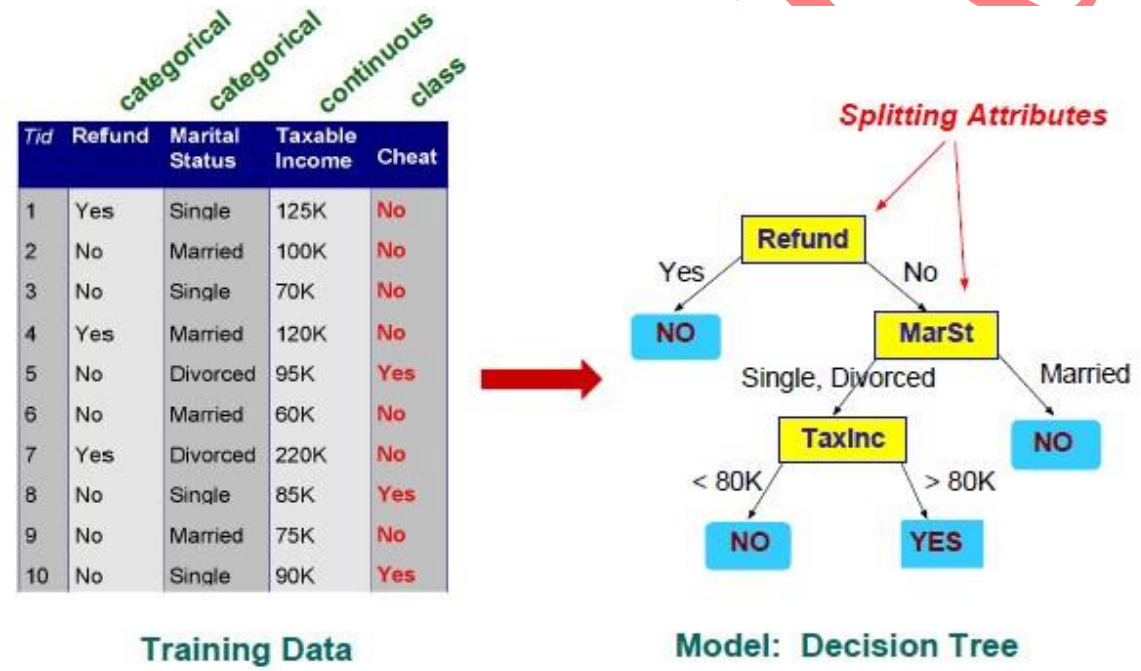
Submitted To
Md. Fazlul Karim Patwary

Submitted By
Name: **JM Mubasshir Rahman**
Section: A
ID: **242104**

Q: 1. Define tree based and rule-based classification.

Ans: In data mining, tree-based and rule-based classification are widely used methods for extracting insights and making predictions from large datasets. Below, we discuss one of these approaches:

Decision Tree Classification: Decision tree classification produces results in the form of a binary tree-like structure, making it highly interpretable, especially for marketing professionals who may need to identify key variables relevant to tasks like managing customer churn. This model works by establishing rules that guide the prediction of the target variable. The decision tree algorithm offers a clear and straightforward description of the data's distribution, making it easier to understand the patterns and relationships within the data.
Example of Decision Tree:



Rule-Based Classification: Rule-based classification in data mining is a method where class decisions are determined by applying a series of "if...then...else" rules. This approach can be defined as a type of classification governed by a set of *IF-THEN* rules, which are written as:

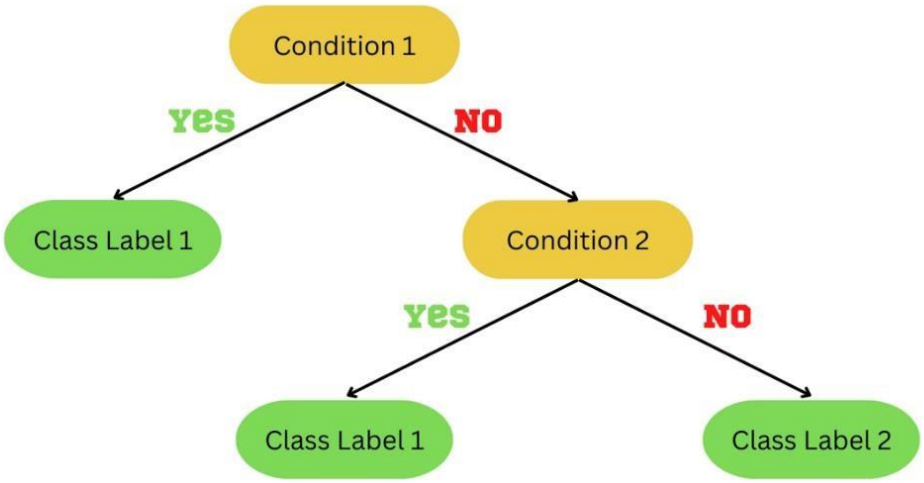
"IF condition THEN conclusion."

Structure of the IF-THEN Rule

An IF-THEN rule consists of two parts:

- **Rule Antecedent**: This is the "if condition" part of the rule, found on the left-hand side (LHS). The antecedent includes one or more conditions or attributes, typically connected using the logical AND operator.

- **Rule Consequent**: This is the "then conclusion" part, located on the right-hand side (RHS). The rule consequent provides the class prediction.



For example, consider a rule, *R1*, which could be written as:

R1: IF age = youth AND student = yes THEN buy_computer = yes

Alternatively, we could express rule *R1* as:

R1: (age = youth) ∧ (student = yes) ⇒ (buy_computer = yes)

If these conditions hold true for a given data entry, then the rule's antecedent is considered satisfied, allowing us to apply the consequent or the predicted class outcome.

Q: 2. What are the advantages of tree-based classification?

Ans: One of the primary benefits of decision trees is their ability to simplify the interpretation and visualization of complex, nonlinear data patterns. Decision trees are widely appreciated for their clarity and practical application. Below are several key advantages of tree-based classification:

- **Low Cost to Construct**: Decision trees are relatively inexpensive to build.
- **High Speed in Classification**: They are extremely efficient at classifying new, unseen data points.
- **Simplicity in Interpretation**: Small decision trees are easy to interpret and analyze.
- **Comparable Accuracy**: For many straightforward datasets, decision trees offer accuracy similar to other classification methods.
- **Minimal Data Preparation**: Decision trees require less effort in data preprocessing compared to many other algorithms.

- **No Need for Normalization or Scaling**: They do not require data normalization or scaling, simplifying the preparation process.
- **Flexible Data Handling**: Decision trees can work with various data types, including numerical, categorical, and Boolean data.
- **Intuitive Model Structure**: Decision tree models are straightforward, making them easy to explain to both technical and non-technical stakeholders.
- **Automatic Feature Selection**: During training, decision trees automatically focus on the most important features, making them suitable for high-dimensional datasets.
- **Ease of Interpretation**: The structure of decision trees makes them highly interpretable and easy to understand.

These characteristics make decision trees an effective and versatile tool for classification tasks in various fields.

Q: 3. Write the steps of tree-based classification?

Ans: The sequential steps involved in performing tree-based classification on a dataset are as follows:

- **Data Collection, Understanding, and Preparation**: Gather the necessary data, gain insights into its characteristics, and prepare it for analysis, including handling any missing or inconsistent values.
- **Feature Selection or Extraction (Optional)**: Identify or derive key features that may improve model performance.
- **Data Splitting**: Divide the dataset into training and testing subsets to evaluate model performance later.
- **Model Selection and Training**: Choose a decision tree model and train it on the training dataset to capture data patterns.
- **Model Evaluation**: Assess the model's performance on the testing data, often using metrics such as accuracy or precision.
- **Model Fine-Tuning (Optional)**: Adjust model parameters to enhance its performance, if needed.
- **Final Model Training**: Retrain the model on the entire dataset to maximize learning from all available data.
- **Model Deployment and Prediction**: Deploy the trained model to make predictions on new data.

These steps help ensure a systematic approach to building an effective tree-based classification model.

Q: 4. Suppose you have two variables (A and B) to select one in constructing a tree. Under variable A distribution of class variable are 5 Cheat and 15 No Cheat in one hand and in other hand it is 3, 10 where as under variable B it is 3, 7 in one hand and 6, 8 in other hand respectively. Which variable you should select? Use Gini coefficient or entropy to give your answer.

Ans: To determine which variable (A or B) to select for constructing a decision tree, we can use the Gini coefficient or entropy as a measure of impurity. Lower values of Gini coefficient or entropy indicate better splits.

Gini Coefficient: For each variable (A and B), calculate the Gini index for each split. The Gini index is given by the formula:

$$GINI(t) = 1 - \sum_j [p(\frac{j}{t})]^2$$

where p(j | t) is the relative frequency of class j at node t.

Entropy: For each variable (A and B), calculate the entropy for each split. Entropy is given by the formula:

$$Entropy = -\sum_j p(\frac{j}{t}) \log_2 p(\frac{j}{t})$$

Where (j | t) is the relative frequency of class j at node.

Let's calculate the Gini coefficient for both variables **A** and **B:**

For variable A:

Gini(A) = $1-(5^2/20^2 + 15^2/20^2)$ for one split

= 0.375

Gini(A) = $1-(3^2/13^2 + 10^2/13^2)$ for the other split

= 0.355

For variable B:

Gini(B) = $1-(3^2/10^2 + 7^2/10^2)$ for one split

= 0.42

Gini(B) = $1-(6^2/14^2 + 8^2/14^2)$ for the other split

= 0.490

Compare the Gini impurity values for each variable in both hands, and choose the variable with the lower Gini impurity. The variable **A** with lower Gini coefficient is generally preferred for splitting in a decision tree.

Q: 5. What do you mean by model reliability?

Ans: In data mining, model reliability means how dependable and accurate a model is when it comes to extracting patterns, trends, and valuable insights from large datasets. For making solid decisions and predictions, reliable models are very important. Some key points for understanding model reliability are as follows:

- **Accuracy and Precision**: A reliable model should be both accurate and precise, meaning its predictions should match closely with the real data.
- **Generalization**: A good model should work well not only on the data it was trained on but also on new, unseen data.
- **Robustness**: Reliable models are robust, meaning they can handle data variations and noise without being too affected by minor fluctuations or outliers.
- **Consistency**: Consistency is essential; a reliable model should produce similar results for the same input every time.
- **Interpretability**: For data mining to be useful, the model should be understandable, allowing us to draw meaningful insights from the data patterns.
- **Validation Techniques**: Reliable models are typically tested through methods like cross-validation to check their accuracy and stability.
- **Feature Importance**: Knowing which features impact predictions helps in understanding the reliability and relevance of the model.
- **Handling Imbalanced Data**: In real-world data, sometimes one class is more frequent than another. Reliable models take this imbalance into account.
- **Monitoring and Maintenance**: It's essential to keep monitoring and updating the model regularly, especially in dynamic environments, to ensure it remains reliable over time

Q: 6. Define the terms: True positive and False Negative and F-measure.

Ans: In data mining and machine learning, we use several metrics to assess how well a classification model performs. Three important terms are:

- **True Positive (TP)**: A true positive occurs when the model correctly identifies a positive instance. In simpler terms, it is the number of cases that actually belong to the positive class and are accurately predicted as positive by the model.
- **False Negative (FN)**: A false negative happens when the model incorrectly predicts a negative class for an instance that truly belongs to the positive class. In other words, these are cases that are actually positive but are wrongly classified as negative by the model.
- **F-measure (F1 Score)**: The F-measure, also called the F1 score, is a metric that balances both precision and recall to provide a single evaluation score. It is particularly helpful when the class distribution is uneven. The F1 score is calculated with the formula:

$$F1 = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

Precision refers to the ratio of true positives to the total of true positives and false positives, while recall (or sensitivity) is the ratio of true positives to the total of true positives and false negatives. The F1 score ranges from 0 to 1, with higher values indicating better model performance.

These metrics help provide a comprehensive view of a classification model's effectiveness, especially in identifying and balancing true positives and managing misclassifications.

Q: 7. What is the main principal of Gini index? – explain. When we have to use Gini index in splitting?

Ans: The Gini Index is a measure of impurity used in decision tree algorithms for tasks like data mining and classification. It quantifies how mixed or "impure" a set of data is based on its class distribution.

The main idea behind the Gini Index in data mining is to evaluate the quality of a possible split while building a decision tree. It helps in selecting the attribute and split point that will produce the most uniform or "pure" subsets, resulting in a clearer separation between classes.

Gini Index Formula: The formula for the Gini Index for a node with two classes (positive and negative) is:

$$\text{GINI}(t) = 1 - \sum_j [p(j|t)]^2$$

[Note: p (j | t) is the relative frequency of class j at node t.

Maximum (1 - 1/nc) when records are equally distributed among all classes, implying least interesting information.

Minimum (0.0) when all records belong to one class, implying most interesting information.

C1	0
C2	0
Gini = 0.000	

C1	1
C2	5
Gini = 0.278	

C1	2
C2	4
Gini = 0.444	

C1	3
C2	3
Gini = 0.555	

Examples for Computing GINI:

$$\text{GINI}(t) = 1 - \sum_j [p(j|t)]^2$$

C1	0
C2	6

$P(C1) = 0/6 = 0 \quad P(C2) = 6/6 = 1$

$Gini = 1 - P(C1)^2 - P(C2)^2 = 1 - 0 - 1 = 0$

C1	1
C2	5

$P(C1) = 1/6 \quad P(C2) = 5/6$

$Gini = 1 - (1/6)^2 - (5/6)^2 = 0.278$

C1	2
C2	4

$P(C1) = 2/6 \quad P(C2) = 4/6$

$Gini = 1 - (2/6)^2 - (4/6)^2 = 0.444$

When to use the Gini index in splitting:

Decision Trees: The Gini index is commonly used in decision tree algorithms, such as CART (Classification and Regression Trees). It is particularly useful for binary classification problems.

Categorical Variables: The Gini index is effective when dealing with categorical variables. It can handle multi-class problems but is often used in binary classification.

Impurity Measure: Use the Gini index when the goal is to minimize impurity or inequality in the resulting nodes of the decision tree. The algorithm aims to find splits that lead to nodes with predominantly one class.

Q: 8. What do you mean by the term precision and recall? When do we use these?

Ans: In data mining and machine learning, ***precision*** and ***recall*** are important metrics used to evaluate the performance of classification models, especially when dealing with imbalanced data or situations where certain types of errors have different impacts.

Recall: *Recall*, also known as sensitivity or the true positive rate, measures a model's ability to correctly identify all relevant instances of the positive class. It is defined as the ratio of true positive predictions to the total number of actual positive instances (true positives plus false negatives).

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- **High Recall**: A high recall score means the model is good at capturing positive instances and minimizes false negatives. In other words, it rarely misses actual positive cases.

Precision: *Precision*, also known as positive predictive value, measures how accurate the model's positive predictions are. It is defined as the ratio of true positive predictions to the total number of positive predictions made by the model (true positives plus false positives).

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

- **High Precision**: A high precision score indicates that the model has a low false positive rate, meaning that when the model predicts a positive class, it is usually correct.

When to Use Precision and Recall

- **Imbalanced Datasets**: Precision and recall are particularly useful when dealing with imbalanced datasets, where one class is much more frequent than the other. In such cases, accuracy alone may not be a reliable indicator of model performance.
- **Trade-off Analysis**: Precision and recall help analyze the trade-off between false positives and false negatives. Depending on the application, one metric might be more important than the other. For instance, in medical diagnosis, high recall might be prioritized to ensure that as many true positive cases are identified as possible, even if it means accepting more false positives.
- **Real-world Impact**: In some applications, the cost of false positives and false negatives can be very different. Precision and recall allow us to assess these costs and choose a model that best aligns with the specific requirements of the problem.
- **Medical Diagnosis and Anomaly Detection**: In scenarios like medical diagnosis or anomaly detection, where missing a positive case (false negative) could have serious consequences, high recall is often prioritized to reduce the risk of false negatives.
- **Information Retrieval**: Precision and recall are commonly used in information retrieval tasks, such as evaluating search engine performance. Precision measures the relevance of retrieved documents, while recall assesses the completeness of the retrieval.

These metrics are essential for understanding a model’s strengths and weaknesses in specific contexts and allow for a more targeted approach to model evaluation, especially in critical applications.