**MCSE 666: Assignment 09**

**Roll: 079 05536**  **Name: Shied Shahriar Housaini**

**Why Decision tree ID3 is called ID3 and C4.5 is called C4.5?**

**ID3 (Iterative Dichotomiser 3):** ID3 was developed by Ross Quinlan in 1986. The "3" in its name signifies that it was the third version or iteration of the algorithm. The "Iterative" part of the name reflects the fact that the algorithm iteratively builds the decision tree, dividing the dataset into subsets and making decisions based on information gain at each step. Decision tree ID3 is called ID3 because it stands for Iterative Dichotomiser 31. It is an algorithm to generate a decision tree from a dataset2. The name reflects the fact that the algorithm iteratively splits the features into two or more groups at each step1.

**C4.5:** C4.5, also created by Ross Quinlan, is the successor to ID3. The "4.5" in its name indicates that it was developed after ID3, marking the fourth iteration or version of the algorithm. C4.5 is called C4.5 because it is an extension of the ID3 algorithm that uses the **information gain ratio** instead of information gain to select the root node3. C4.5 is the precursor to the C5.0 algorithm, which is also known as See5 on Windows4.

Reference:

(1) What is the meaning of the decision tree algorithm name "c4.5"?. https://stackoverflow.com/questions/39491172/what-is-the-meaning-of-the-decision-tree-algorithm-name-c4-5.

(2) ID3 algorithm - Wikipedia. https://en.wikipedia.org/wiki/ID3\_algorithm.

(3) What is the difference between ID3 and C4 5 decision trees. https://short-informer.com/what-is-the-difference-between-id3-and-c4-5-decision-trees/.

(4) Decision Tree Algorithms: ID3, C4.5, And CART Explained And Compared. https://www.lokidom.com/2023/02/decision-tree-algorithms-id3-c45-and.html.

(5) What is the meaning of the decision tree algorithm name "c4.5". https://stackoverflow.com/questions/39491172/what-is-the-meaning-of-the-decision-tree-algorithm-name-c4-5.

(6) ID3 algorithm - Wikipedia. https://en.wikipedia.org/wiki/ID3\_algorithm.

(7) Decision Tree Algorithms: ID3, C4.5, And CART Explained And Compared. https://www.lokidom.com/2023/02/decision-tree-algorithms-id3-c45-and.html.

**What is the significance of C4.5 over ID3?**

* Decision tree algorithms are methods to generate a tree-like structure from a dataset that can be used for classification and prediction tasks.
* ID3 and C4.5 are two popular decision tree algorithms that Ross Quinlan invented.
* C4.5 is an extension and improvement of ID3 in several ways.

## **Differences and Advantages of C4.5 over ID3**

* **Continuous attributes**: C4.5 can handle continuous attributes, while ID3 only works with categorical attributes.
  + Continuous attributes are numerical values that can take any range, such as height, weight, temperature, etc.
  + Categorical attributes are discrete values that belong to a finite set, such as color, gender, type, etc.
  + C4.5 can split continuous attributes into two subsets by creating a threshold value.
    - For example, if the attribute is height, C4.5 can split the data into two groups: those who are taller than 170 cm and those who are shorter than 170 cm.
* **Missing values**: C4.5 can handle missing values, while ID3 cannot.
  + Missing values are data points that have no value for some attributes.
  + For example, if the attribute is weight, some data points may not have a weight value due to measurement errors or incomplete records.
  + C4.5 can deal with missing values by assigning them a probability based on the distribution of the values of that attribute in the data.
    - For example, if the attribute is weight and 60% of the data points have a weight value between 50 kg and 70 kg, then a missing value will be assigned a 60% probability of belonging to that range.
* **Information gain ratio**: C4.5 uses information gain ratio instead of information gain as the criterion to select the best attribute to split the data at each node.
  + Information gain measures how much the entropy (uncertainty) of the target variable decreases after splitting on an attribute.
  + The information gain ratio is the ratio of information gain to the intrinsic value of an attribute, which measures how evenly the attribute splits the data.
  + This helps to avoid bias towards attributes with many values.
    - For example, suppose the attribute is ID number and it has a unique value for each data point. In that case, it will have a high information gain but a low information gain ratio, because it will split the data into many small subsets that are not useful for classification.
* **Pruning**: C4.5 performs pruning after building the tree, while ID3 does not.
  + Pruning is a technique to reduce the size and complexity of the tree by removing branches that are not relevant or useful for classification.
  + This helps to avoid overfitting, which is when the tree fits the training data too well but performs poorly on new data.
  + C4.5 uses a validation set to evaluate the accuracy of the tree and prunes branches that do not improve or worsen the accuracy.
* **Handling of Skewed Datasets:**
  + ID3 can create biased trees when dealing with imbalanced or skewed datasets where one class is significantly larger than the others.
  + C4.5 is more robust when dealing with skewed datasets. Its use of gain ratio and pruning helps mitigate the effects of class imbalance, leading to more balanced and accurate decision trees.
* **Rule Generation:**
  + ID3 focused on tree structures as the primary output.
  + C4.5 not only provided the tree but also allowed the generation of classification rules based on the tree. This made it more interpretable and user-friendly.
    - In a credit risk assessment system, a bank might prefer having explicit rules for accepting or rejecting loans rather than interpreting a complex decision tree. C4.5's rule generation is beneficial in such cases.

## **Conclusion**

* C4.5 is more advanced and accurate than ID3, but it may also be more complex and computationally expensive.
* Depending on the data and the problem, either algorithm may perform better in different scenarios.
* C4.5's versatility, robustness, and improved performance in various aspects made it more widely applicable and popular in both academic research and practical applications. It served as a foundation for many subsequent decision tree algorithms and data mining techniques.
* C4.5's success led to the development of various software implementations and adaptations, making it widely accessible to researchers and practitioners. These tools facilitated the application of C4.5 to a broad range of fields and industries.

Reference:

(1) Study of The ID3 and C4.5 Learning Algorithms - Open Access Pub. https://openaccesspub.org/medical-informatics-and-decision-making/article/1334.

(2) C4.5 and ID3 algorithms with emphasis on practical details. https://stackoverflow.com/questions/13615096/c4-5-and-id3-algorithms-with-emphasis-on-practical-details.

(3) C4.5 Algorithm in Data Mining | T4Tutorials.com. https://t4tutorials.com/c4-5-algorithm-in-data-mining/.

(4) C4.5 and ID3 algorithms with emphasis on practical details. https://stackoverflow.com/questions/13615096/c4-5-and-id3-algorithms-with-emphasis-on-practical-details.

(5) Study of The ID3 and C4.5 Learning Algorithms - Open Access Pub. https://openaccesspub.org/medical-informatics-and-decision-making/article/1334.

(6) Comparative Analysis of Decision Tree Algorithms: ID3, C4.5 ... - Springer. https://link.springer.com/chapter/10.1007/978-81-322-2205-7\_51.

(8) Open Access Pub https://doi.org/10.14302/issn.2641-5526.jmid-20-3302.