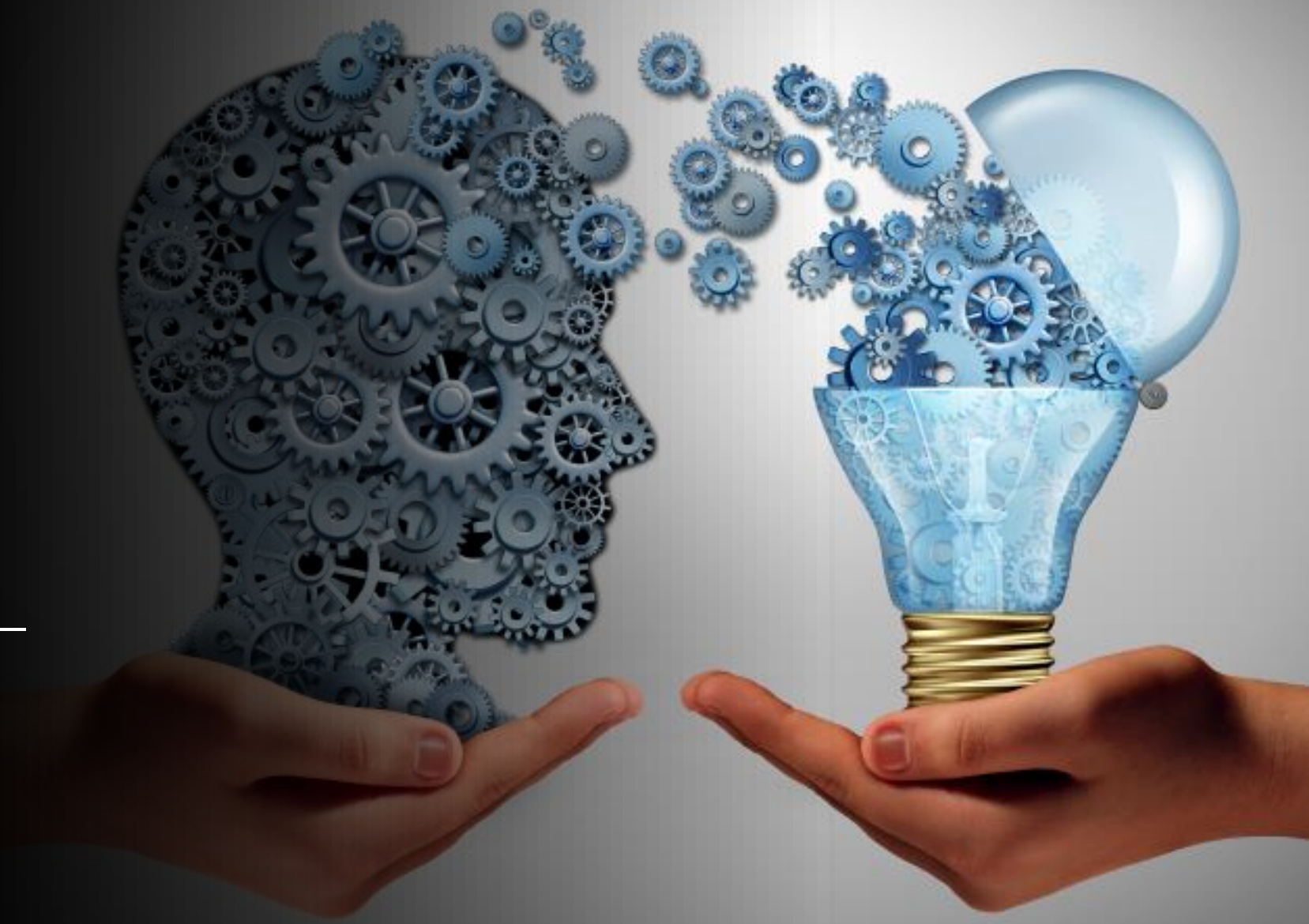
Decision Trees

Arsal Abbasi



Contents

- Motivation
- Overview of Decision Trees
- Construction
- Types
- Implementation

Motivation

Problems

- Making the right decision is critical for success in all areas of work.
- People make decisions in various ways based on their experiences, new research, and personal judgment.
- individuals and groups making decisions face increased complexity because they can no longer handle massive amounts
  of data.

Solution

- As a result, reliable decision support strategies are required to assist people in making informed decisions.

Motivation

- In the area of machine learning, decision trees are a strong and popular technique. They are a well-liked option for both novices and specialists due to their adaptability, intuitiveness, and simplicity of understanding. Decision trees are particularly helpful for managing huge datasets [1] with numerous characteristics and may be utilized for both classification and regression tasks.

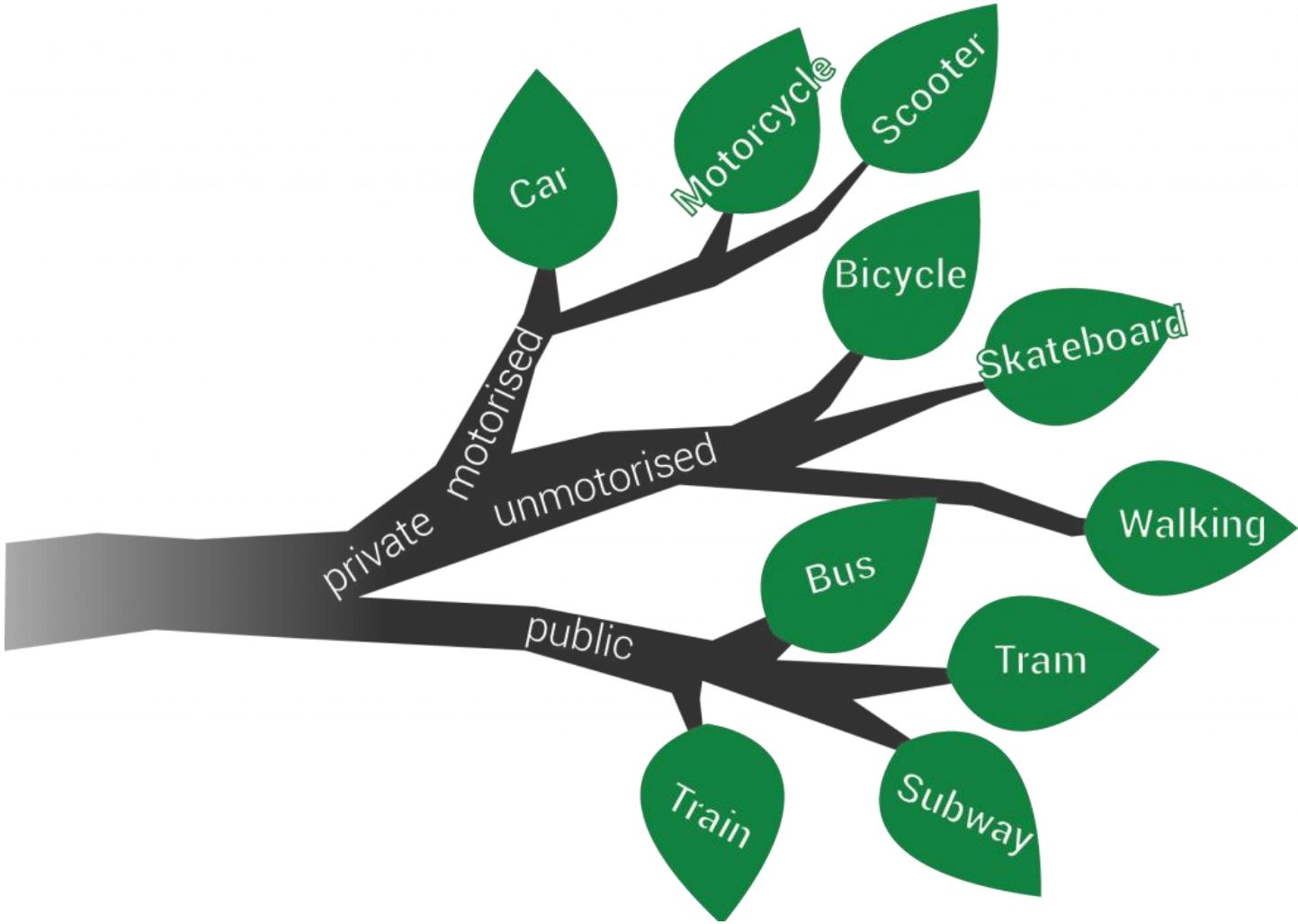


Fig.1 Decision Tree representation

Overview of Decision Trees

A decision tree is a method of categorizing data based on certain characteristics.

Decision tree mimics a tree with a root at the top and several branches leading to various outcomes [2].

The initial point has no connections to the other points, whereas the other points have one link to the starting point[2].

points having connections are referred to as internal or test nodes, whilst locations without connections are referred to as leaves.

Each internal node divides the data into several groups based on a unique feature [2].

Each leaf is allocated to a class that represents the best goal value [2].

Decision Tree

- Response to direct mailing

It can be easily deduced that:

- Customers aged more than 30 years are not likely to respond to direct mailing.
- Customers who are less than 30 years and are male are likely to respond to direct mailing.

Important observation

- As long as there is mixture of data, there is room to split further. This is known as the purity of the node, an impure node can be branched further to improve purity [3].

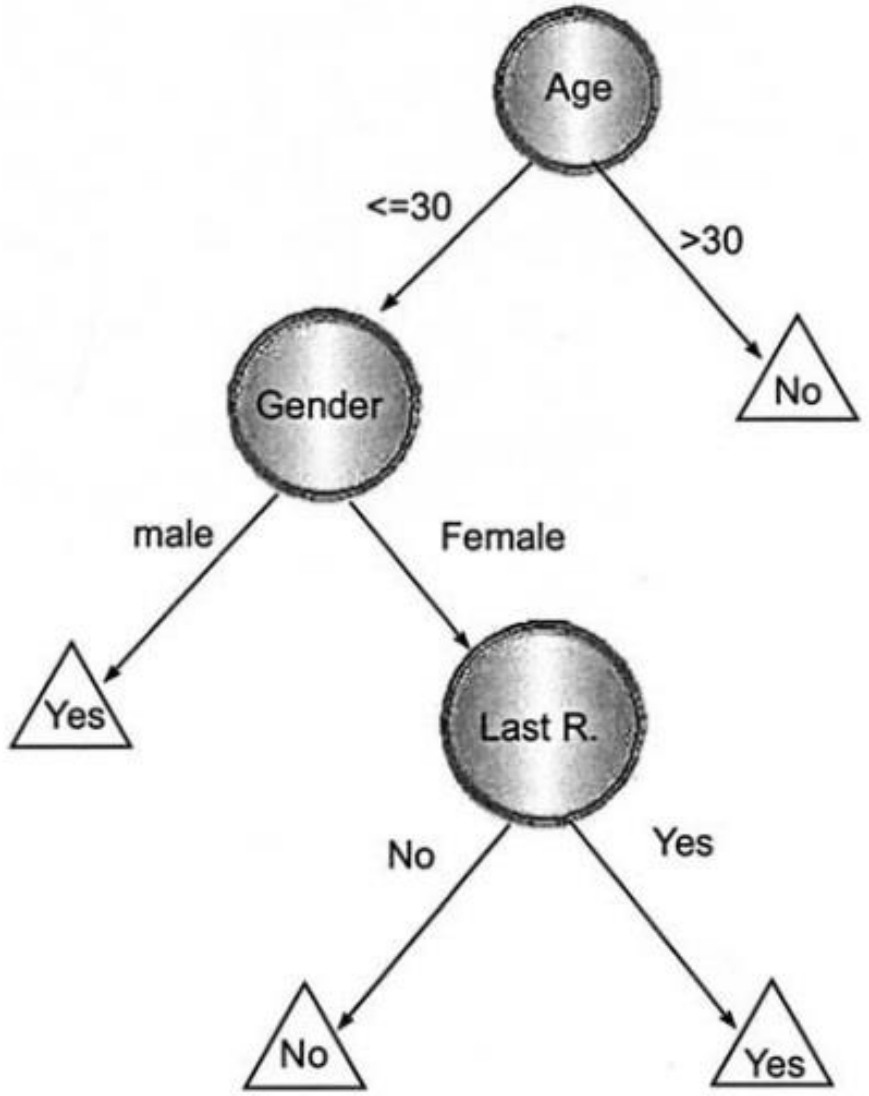


Fig.2 Decision tree presenting a response to direct mailing [RM05]

Splitting of decision trees

- Entropy Method
- Information Gain Method
- Gini Index Method

- The Entropy and Information Gain method focuses on a node's purity and impurity [3].
- The Gini Index calculates the likelihood of a random instance being misclassified when it is chosen at random. The lower Gini Index, the smaller the chance of misclassification [3].

Gini Index can be calculated using the formula [3] :

$$Gini = 1 - \sum_{i=1}^{j} P(i)^2$$

- Where j represents the number of classes and P(i) represents the number of observations in the node

Construction

- A decision tree is created by reviewing a set of training objects using the "divide and conquer" method.
- The method begins by determining if all of the objects in the set belong to the same decision class [4].
- If they do, a tree consists of a single node, which is a leaf with the necessary decision[4].
- If not, an attribute with at least two different decision classes is chosen, and the collection of objects is separated according to the category of the chosen attribute [4].
- The chosen attribute becomes an attribute (test) node in a growing decision tree, and the inciting operation is repeated on the remaining objects with respect to the division for each branch from that node until a leaf (a decision) is reached.
- This procedure is repeated recursively, with each attribute chosen to maximize the separability of cases with various outcomes, until a tree is built that can aid in decision making depending on the qualities of new examples [4].

Types

- Classification Trees
  - Used for classification problems.
  - The goal is to find a model that can predict Y given new X values.

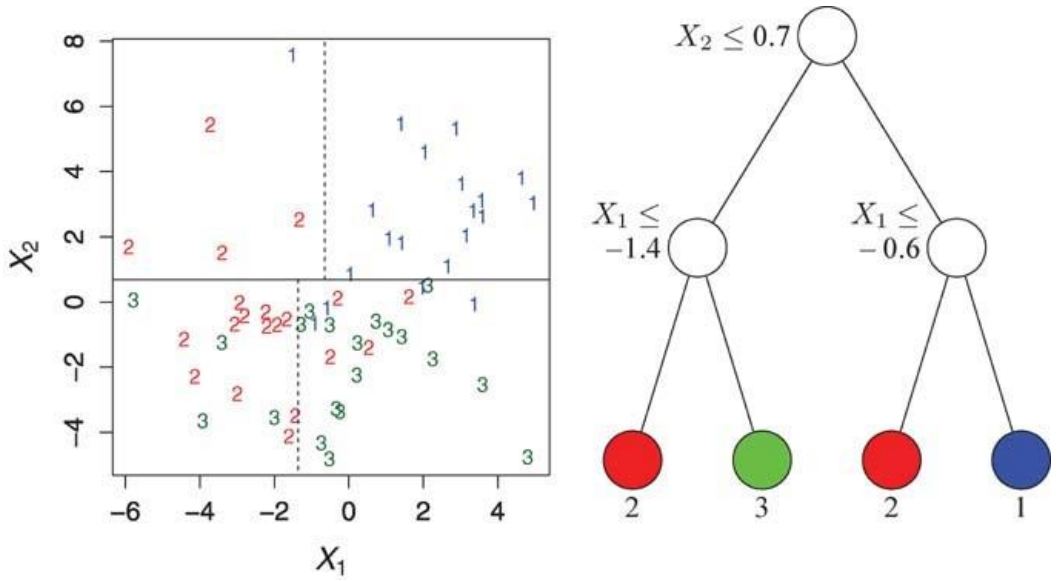


Fig. 2: Partitions (left) and decision tree structure (right) for a classification tree model [5]

Types

- Regression Trees
  - They do not have classes [6].
  - In regression trees, the Y variable takes ordered values, and a regression model is fitted to each node to provide predicted Y values [5].
  - Regression trees help to predict a value that can be continuous or ordered (like a number), and predict the way we measure how well our prediction works is by looking at the difference between our predicted value and the actual value.

Implementation

- Iris dataset was used
- Contains six features; SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalwidthCm, Species, Class labels
- Iris data in scatter was visualised using Matplotlib and seaborn libraries.

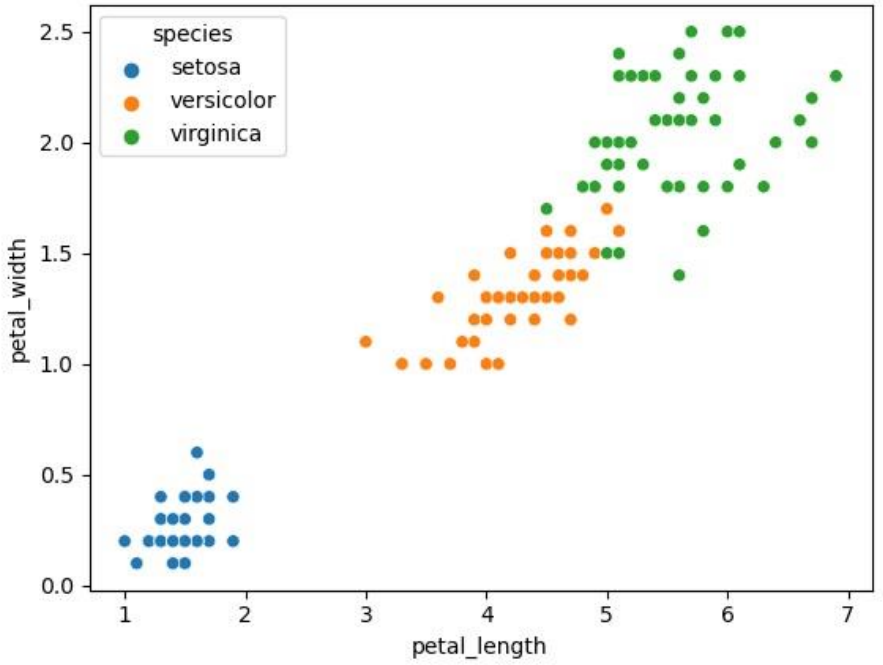


Fig. 4 Visualization of iris data

Implementation code

```python
import pandas as pd
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.tree import export_graphviz
from IPython.display import Image

iris = pd.read_csv('https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data', header=None)
iris.columns=['sepal_length', 'sepal_width', 'petal_length', 'petal_width', 'species']

X = iris.drop('species', axis=1)
y = iris['species']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Create and train the decision tree classifier
clf = DecisionTreeClassifier()
clf.fit(X_train, y_train)

DecisionTreeClassifier
DecisionTreeClassifier()

# Make predictions on the testing data
y_pred = clf.predict(X_test)
accuracy = metrics.accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

# Export the decision tree to a dot file
export_graphviz(clf, out_file='tree.dot', feature_names=X.columns)

# Convert the dot file to a PNG image
!dot -Tpng tree.dot -o tree.png

# Display the decision tree image
Image(filename='tree.png')
```

t[7]:

Results

- The data was split into two parts, 70% data was used fo training and 30% was used for testing. The accuracy of the model was calculated to be 100%.
- The top node of the tree represents the complete data set.
- If the petal length is less than or equal to 2.45 cm, the species predicted is Setosa.
- If the petal width is less than or equal to 1.75 cm, the species is Versicolor with the gini index of 0.214
- otherwise, the species is Virginica with the gini index of 0.059.

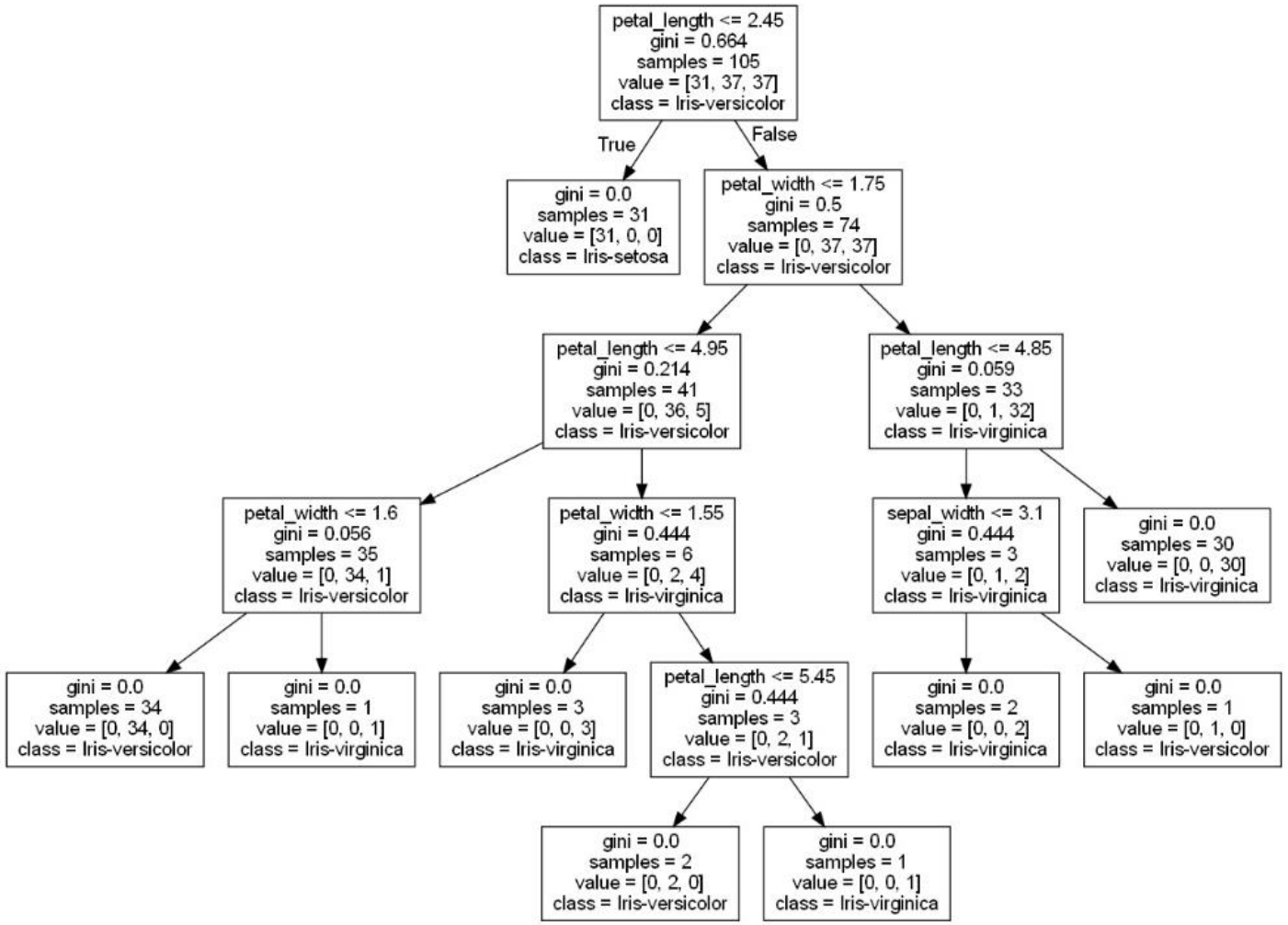


Fig. 5 Implemented Decision Tree

Summary

In conclusion, decision trees are an effective tool for classification and regression applications. They produce a tree-like structure by recursively partitioning the data based on the predictor variables. Each tree leaf node is a prediction for the target variable. Decision trees are particularly useful for dealing with enormous data sets since they can handle both categorical and continuous data. They are also easily interpret able, making the decision making process understandable

References

1. Kingsford, Carl; Salzberg, Steven L: What are decision trees? Nature biotechnology, 26(9):1011–1013, 2008.
2. okach, Lior; Maimon, Oded: Decision trees. Data mining and knowledge discovery handbook, pp. 165–192, 2005.
3. Dash, Shailey: , Decision Trees Explained—Entropy, Information Gain, Gini Index, CCP Pruning.. — towardsdatascience.com. https://towardsdatascience.com/ decision-trees-explained-entropy-information-gain-gini-index-ccp-pruning- 4d78070db36c. [Accessed 27-May-2023]
4. Podgorelec, Vili; Kokol, Peter; Stiglic, Bruno; Rozman, Ivan: Decision trees: an overview and their use in medicine. Journal of medical systems, 26:445–463, 2002.
5. Loh, Wei-Yin: Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery, 1(1):14–23, 2011.
6. Timofeev, Roman: Classification and regression trees (CART) theory and applications. Humboldt University, Berlin, 54, 2004.