

**Kurdistan Regional Government- Iraq**

**Ministry of Higher Education & Scientific Research**

**University of Halabja**

**College of Science**

**Computer Science Department**

**Salary Prediction using Decision Trees**

**Prepared by :**

Brwa Nahman – Mohammed Tahsin

Mohammed Ahmad

**Supervised by :**

Dilman Abdalla Salih

**2022-2023**

# Introduction to Decision Trees

A supervised machine-learning approach called a decision tree is used to resolve classification and regression issues. Each internal node is a test on an attribute, each branch indicates the test's result, and each leaf node represents a class label or a predicted value. The structure is similar to a tree.

Decision trees are built by recursively partitioning the training data set. The algorithm starts with the entire training data set at the root node. At each step, the algorithm chooses the attribute that best splits the data set. The data set is then partitioned into two or more sub-datasets based on the value of the chosen attribute. The algorithm continues recursively partitioning the sub-datasets until all of the data points in each sub-dataset belong to the same class.

Decision trees are a popular machine learning algorithm because they are easy to understand and interpret. They are also relatively efficient to train and predict. However, decision trees can be susceptible to overfitting, which means that they can learn the training data too well and perform poorly on new data. To prevent overfitting, decision trees can be pruned, which means that some of the branches of the tree are removed.

Decision trees are versatile machine-learning algorithms that can be used to solve a variety of problems. They are commonly used for classification problems, such as spam filtering and fraud detection. They can also be used for regression problems, such as predicting house prices and customer churn.

Here are some of the advantages of using decision trees:

* Easy to understand and interpret
* Relatively efficient to train and predict
* Versatile and can be used to solve a variety of problems

Here are some of the disadvantages of using decision trees:

* Can be susceptible to overfitting
* Can be computationally expensive to train large trees
* Not always the best choice for problems with a large number of features

Overall, decision trees are a powerful and versatile machine learning algorithm that can be used to solve a variety of problems. They are easy to understand and interpret and relatively efficient to train and predict. However, decision trees can be susceptible to overfitting, so it is important to use them carefully.

# How used Decision Tree in this code?

The decision tree model is used in this code for its ability to handle both categorical and numerical data, making it a suitable choice for the salary data, which includes both categorical (gender, education level, job title) and numerical (salary) features.

The decision tree algorithm works by recursively splitting the data into smaller and smaller subsets based on the features, until a stopping criterion is met, such as reaching a maximum depth or a minimum number of samples required to continue splitting. The algorithm creates a tree-like structure that represents the decisions made at each split, and the leaves of the tree represent the predicted target values.

To use the decision tree model in scikit-learn, we first import the `DecisionTreeRegressor` class from the `sklearn.tree` module. We then create an instance of the `DecisionTreeRegressor` class, specifying any desired hyperparameters such as the maximum depth of the tree. We can then fit the model to the training data using the `fit()` method, and use the `predict()` method to make predictions on the test data. We can also evaluate the performance of the model using metrics such as the R-squared score or mean squared error.

One important consideration when using the decision tree model is the potential for overfitting, especially when the maximum depth of the tree is not carefully chosen. A tree with too much depth may fit the training data too closely, leading to poor performance on new, unseen data. Regularization techniques such as pruning can help to address this issue.

Overall, the decision tree model is a powerful tool for both classification and regression problems, and its ability to handle mixed data types makes it a useful choice for many real-world datasets. When using the decision tree model, it is important to carefully choose hyperparameters and use regularization techniques to avoid overfitting.