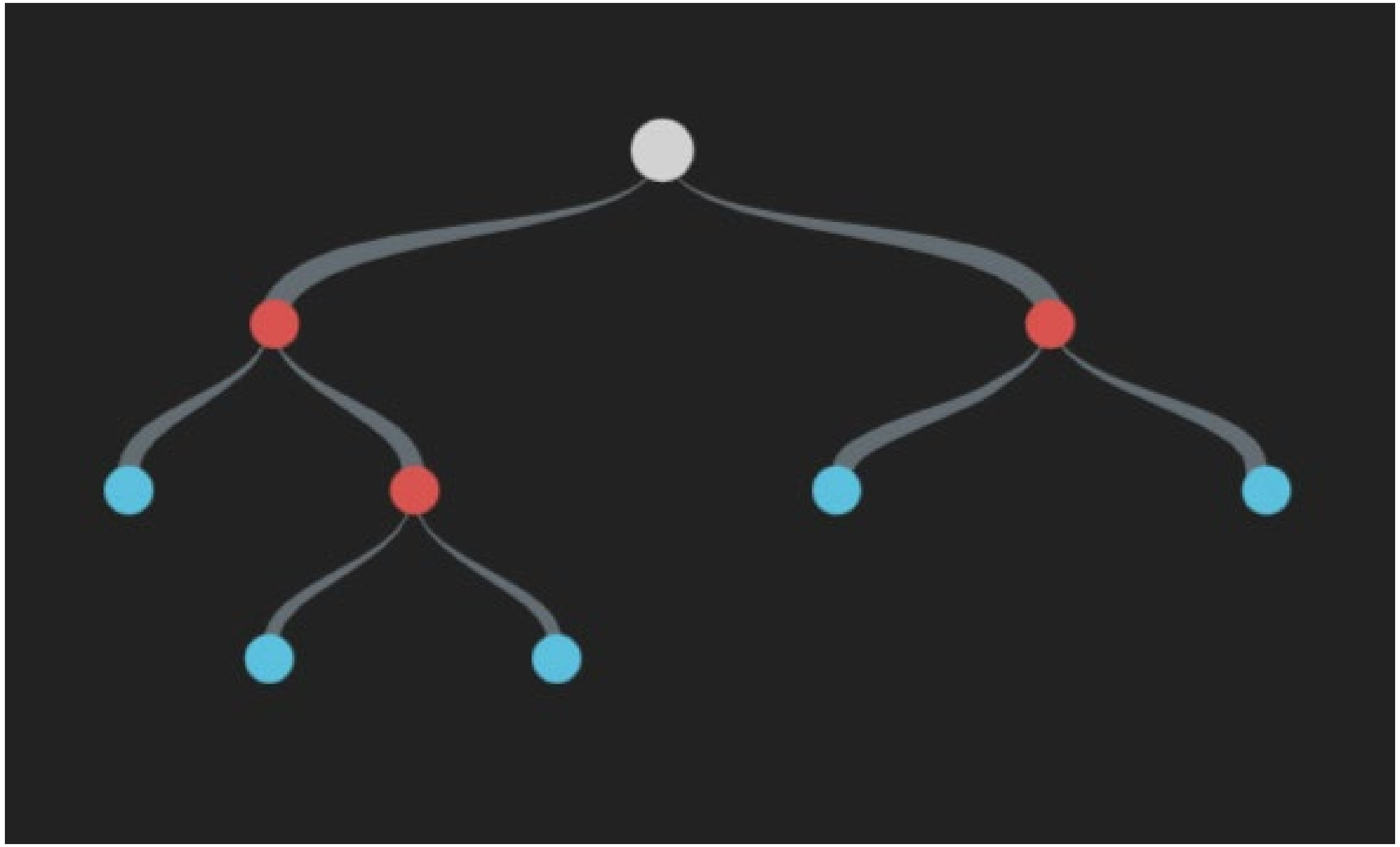
Topic: Decision Tree Regression
Instructor: Prof. Yasser Alginahi
Date: 6th October 2023

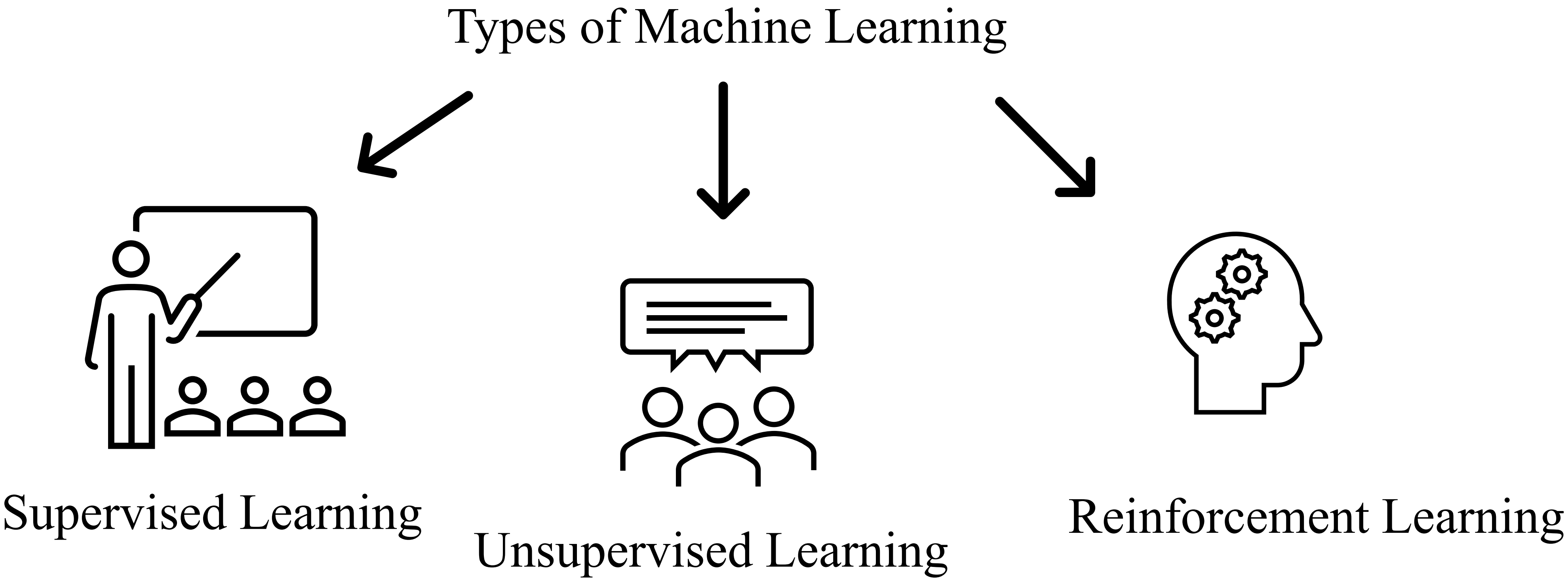
Name
Harshidh Panicker
Jayesh Thakor
Krupit Patel

Decision Tree Regression:

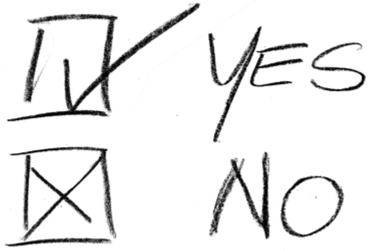


What is Machine Learning?

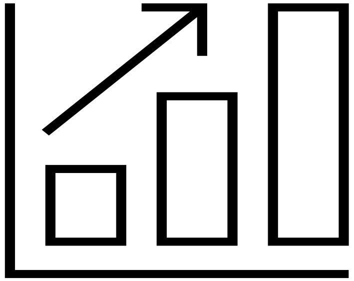
"Machine learning is a type of Artificial Intelligence where the systems gets the ability to automatically learn and improve based on experience"



Types of Supervised Learning



Classification is about assigning labels or categories to input data.



Regression involves predicting a continuous value based on input features.

A decision tree is used for both classification and regression problems.

Decision Tree

A decision tree is a flow chart created by a computer algorithm to make decisions or numeric predictions based on information in a given data set.

It is a versatile supervised machine-learning algorithm, and also non – parametric in nature.

Types of Decision Tree Algorithm

- CART (Classification and Regression Trees): It can handle both categorical and continuous data. Cross-entropy (also called information gain) is a criterion for splitting data in classification tasks and the mean squared error in regression tasks.

- ID3 (Iterative Dichotomiser 3): It uses entropy and information gain as metrics to decide how to split the data at each node.

- C4.5: It is an extension of ID3 and is more versatile.

- C5.0: It is an enhanced version of C4.5, known for its better performance and handling of missing values.

CART Decision Tree Algorithm

- CART is a decision tree algorithm that stands for Classification and Regression Trees and is commonly used in machine learning for both classification and regression tasks.

Key Points:

- 1) The Gini impurity is used by CART as the splitting criterion for classification , where Gini impurity quantifies the likelihood that an element would be incorrectly classified if it were randomly categorized in accordance with the distribution of classes in the node. The intention is to reduce the Gini impurity. Whereas, Mean Squared Error (MSE) is used by CART as the splitting condition for regression. The objective is to reduce the MSE, which represents the target variable's volatility in the node.

CART Decision Tree Algorithm

2) CART starts with a single node containing all the data points. In order to split the data into child nodes, the optimal feature and split point are chosen iteratively. In both classification and regression, it chooses the split that minimizes the Gini impurity. Until a stopping condition is satisfied, the algorithm splits nodes continuously. This requirement may be a minimal Gini impurity/MSE threshold, a certain number of samples per leaf, or a predetermined tree depth.

3) CART may apply pruning to avoid overfitting. Pruning removes branches that do not significantly improve the tree's predictive power. A validation dataset is pruned by evaluating the quality of each branch and assessing whether eliminating a subtree (branch) improves the accuracy of the dataset as a whole.

8

ID3 Decision Tree Algorithm

- Iterative Dichotomizer 3, or ID3, is one of the first and most important decision tree methods used for classification tasks.

Key points:

1) ID3 uses information gain as the splitting criterion. The goal of ID3 is to maximize information gain. Features with higher information gain are chosen for splitting.

2) ID3 applies the idea of entropy for the calculation of information gain. Entropy measures the impurity or disorder in a dataset.

ID3 Decision Tree Algorithm

3) It can handle problems involving binary and multiple classes in classification. For each categorical attribute, ID3 generates a branch for each possible attribute value.

4) ID3 does not include a pruning step. This means it might create deep, complex trees that are prone to overfitting.

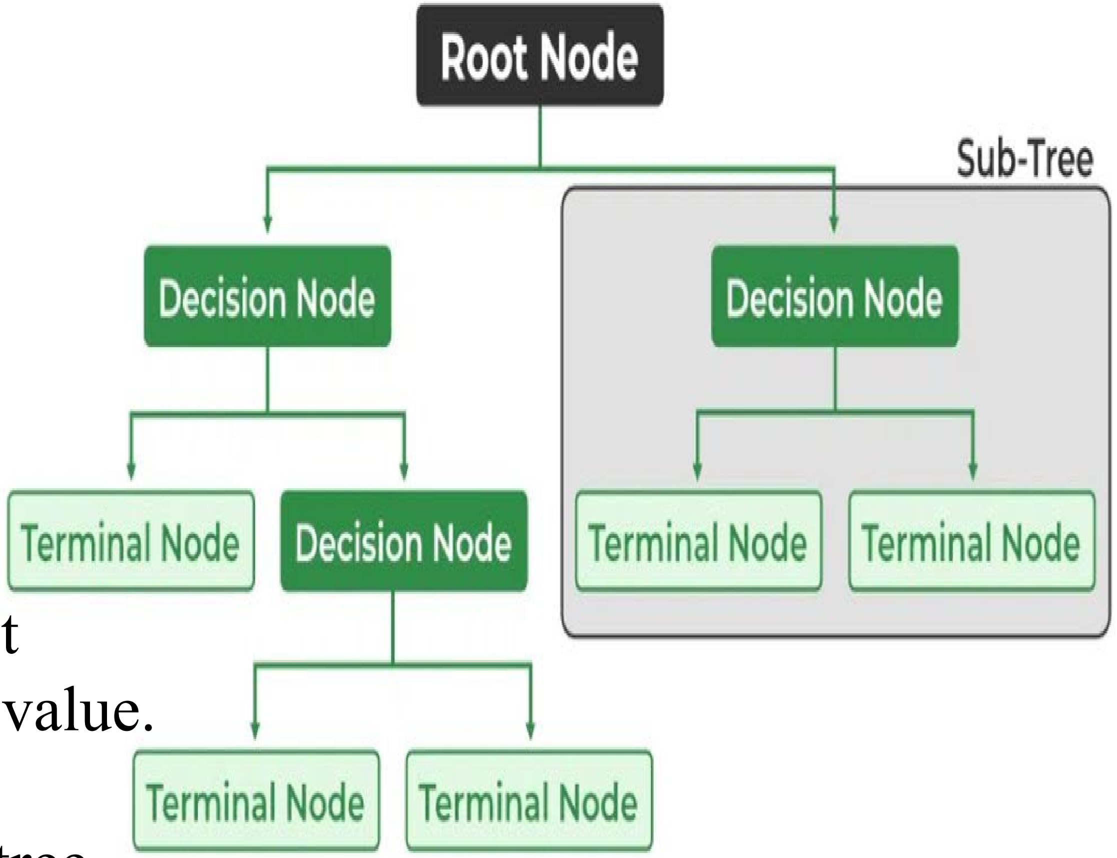
10

C4.5 and C5.0 Decision Tree Algorithm

The decision tree algorithms C4.5 and C5.0 are used for both classification and regression tasks. C5.0 is a commercial version and an enhancement of C4.5. It retains the core principles of C4.5. Here's a brief overview of C4.5 and C5.0:

1) Both techniques are very comprehensible, and the resulting decision trees are easy to understand.
2) They are capable of handling both classification and regression problems.
3) Pruning prevents overfitting.
4) They are capable of handling both categorical and numerical characteristics.
5) They perform effectively with noisy and erroneous data.
6) The choice of splitting criterion can be sensitive to data variations.

Terminologies

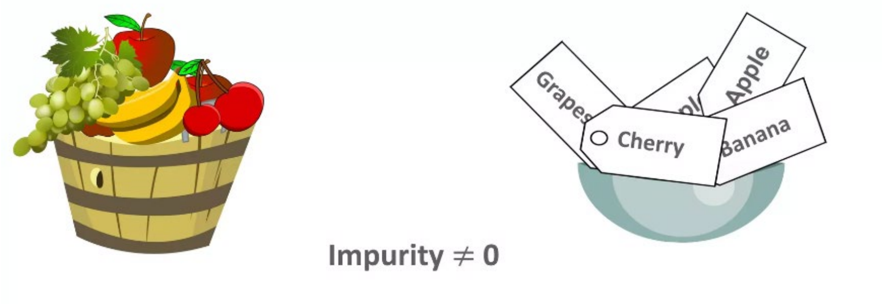


[Source: https://www.geeksforgeeks.org/decision-tree/]

- **Root Node:** It is the topmost node in the tree.
- **Decision Node:** A node that symbolizes a choice regarding an input feature.
- **Leaf/Terminal Node:** A node that cannot be split further and indicates a class label or a numerical value.
- **Branch/Sub-Tree:** A subsection of the decision tree starts at an internal node and ends at the leaf nodes.
- **Pruning:** The process of removing branches from the tree.

Terminologies

What is **Impurity** in the Decision Tree?



Impurity: It refers to the degree of randomness or uncertainty in a set of examples.

Methodology

A machine learning approach called decision tree regression can be applied to both classification and regression applications. The Decision Tree Regression approach is as follows:

Step 1: Import Libraries

- Import necessary libraries.

Step 2: Data Collection

- Collect the dataset containing the features and the corresponding target variable. Make that the data is accurate, free of missing values, and in a format that will work for training a machine learning model.

Step 3: Data Preprocessing

- Execute the needed preprocessing operations, such as managing missing values, encoding category variables, and scaling numerical features as necessary. Although feature scaling has little effect on decision trees, preprocessing is still crucial.

Methodology

Step 4: Split the Data

- Split training and testing sets from the dataset. The model is trained on the training set, and its performance on test data is assessed on the testing set.
- Here, the decision-making process at each node when selecting how to split the data depending on the values of features is greatly aided by Attribute Selection Measures.
- The Decision Tree algorithm uses ASM to determine the attribute to employ for making judgments and forming branches in the tree during the decision node splitting stage. The algorithm is guided by various attribute selection metrics to choose the attributes that would provide the most relevant splits in the data.

Methodology

- **Attribute Selection Measures (ASM)** are pivotal in guiding the split data step of a decision tree algorithm. These measures, such as Information Gain for classification or Variance Reduction for regression, quantitatively evaluate the relevance of attributes in making effective splits. By selecting the attribute with the highest ASM at each node, the algorithm optimizes the reduction in impurity or variance, ensuring the resulting subsets are more homogeneous.
- This recursive process of ASM-guided splitting creates a decision tree that captures the most relevant features of the data, preventing overfitting and contributing to model interpretability.
- ASM not only influences the construction of the tree but also helps identify the importance of different features in predicting the target variable, ultimately optimizing the performance of the decision tree model on both training and testing data.

Attribute Selection Measures (ASM)

Calculating Attribute Selection Measures (ASM):

- ASM techniques, such as **Information Gain, Entropy or Gini Index (for classification tasks)**, **Variance, or Variance Reduction (for regression tasks),** are used to quantify the effectiveness of each attribute for making splits.

Entropy

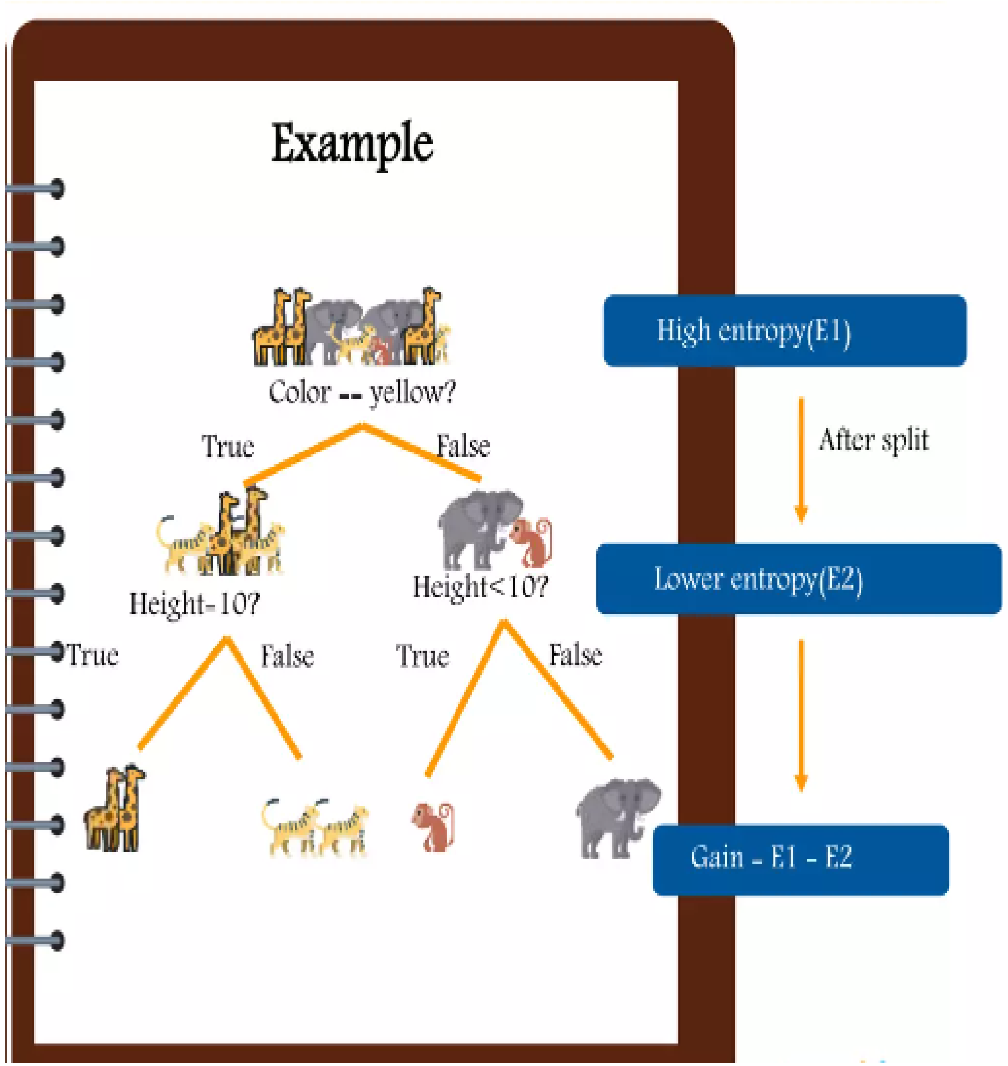
It is a metric to measure the impurity in a given attribute.

It is a degree of Randomness in Data.

Entropy can be Calculated as :

Entropy(s)= - P(true)logP(true) -P(false)logP(false)

Where,
S=Total number of Samples
P(true)=Probability of true
P(false)=Probability of false



[Source: https://tinyurl.com/44ue8ckz]

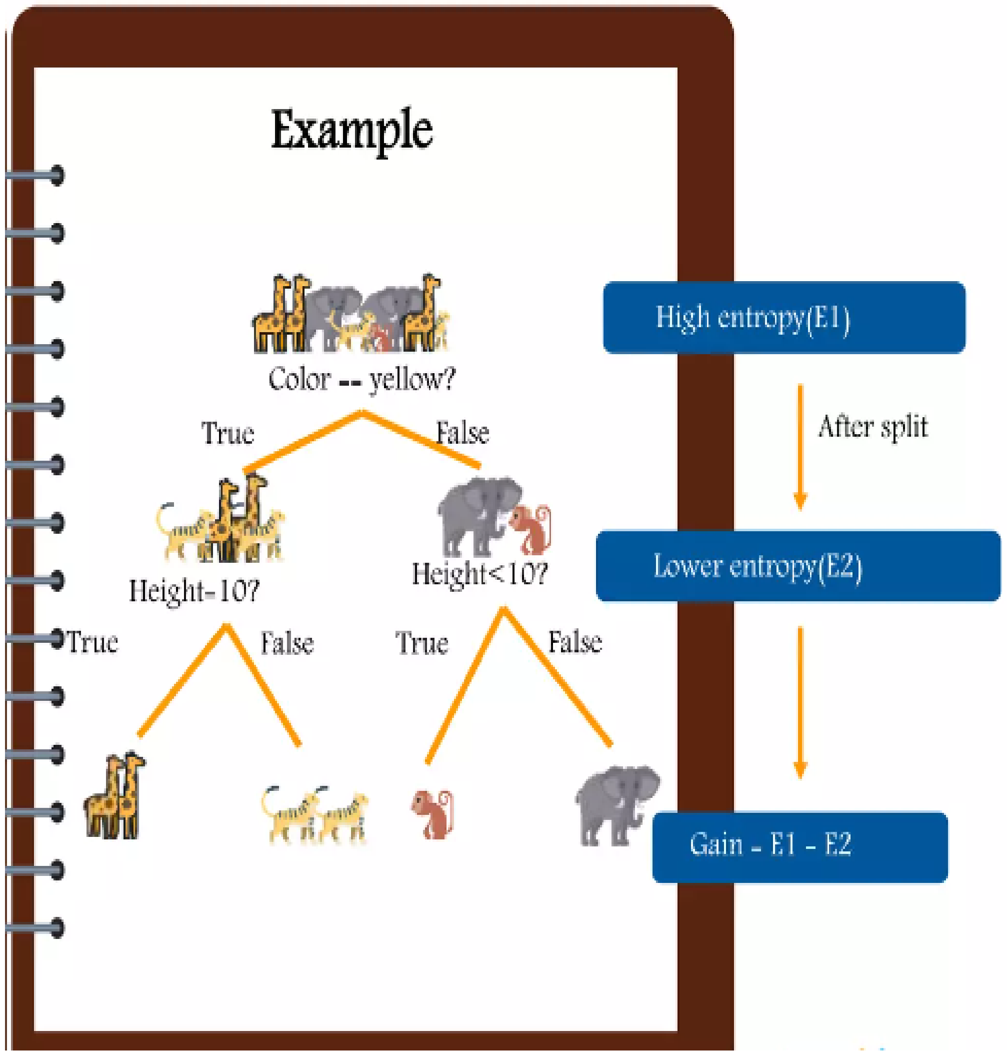
Information Gain

Information Gain: It is the measure of the decrease in entropy after the dataset is split.

It Calculates how much information a feature Provides us about a class.

A Decision tree algorithm always tries to maximize the value of Information gain and a node/ attribute having the highest information gains split first

Information Gain = Entropy(S)-[(Weighted Avg)*Entropy(each Feature)]



[Source: https://tinyurl.com/44ue8ckz]

Variance

Variance measures how much the predicted and the target variables vary in different samples of a dataset.
It is used for regression problems in decision trees.
The formula for variance is calculated as follows:

Variance : $\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$

- N is the number of data points in the node.
- y_i is the prediction for the i-th data point in
the node.
- $\bar{y}$ is the mean prediction in the node.

Variance Reduction

Variance reduction in decision trees refers to the reduction in variance achieved when a node is split into child nodes .

This reduction is often used as a criterion to select the best attribute for splitting.

The formula for variance reduction is calculated as follows:

Variance Reduction = $Variance_{Parent} - \frac{N_{left}}{N_{parent}} \times Variance_{left} - \frac{N_{right}}{N_{parent}} \times Variance_{right}$

Where:

- $Variance_{Parent}$ is the variance in the parent node before the split.
- N_{left} and N_{right} are the number of data points in the left and right child nodes, respectively.
- $Variance_{left}$ and $Variance_{right}$ are the variances in the left and right child nodes after the split.

Methodology

Step 5: Build the Decision Tree Model

- Import the DecisionTreeRegressor from scikit-learn and create an instance of it. You can customize parameters like the maximum depth of the tree, minimum samples required to split a node, etc.

- A Decision Tree Regressor is a machine learning model used for regression tasks. Unlike classification tasks where the goal is to predict a categorical outcome, regression tasks involve predicting a continuous numerical value. Decision Tree Regressors work by recursively splitting the dataset into subsets based on the values of input features, ultimately making predictions at the leaf nodes.

22

Methodology

Step 6: Make Predictions

- Use the trained model to make predictions on the testing set.

Step 7: Evaluate the Model

- Assess the model's performance using regression metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared.
- Evaluate how well the model generalizes to new, unseen data.

Step 8: Visualize the Decision Tree (Optional)

- If needed, visualize the structure of the Decision Tree to understand how it makes decisions.
- If the model performs satisfactorily and meets your requirements, deploy it for making predictions on new, unseen data.

Methodology

Step 9: Hyperparameter Tuning (Optional)

- Fine-tune hyperparameters, such as the maximum depth of the tree or minimum samples required to split a node, to optimize model performance.

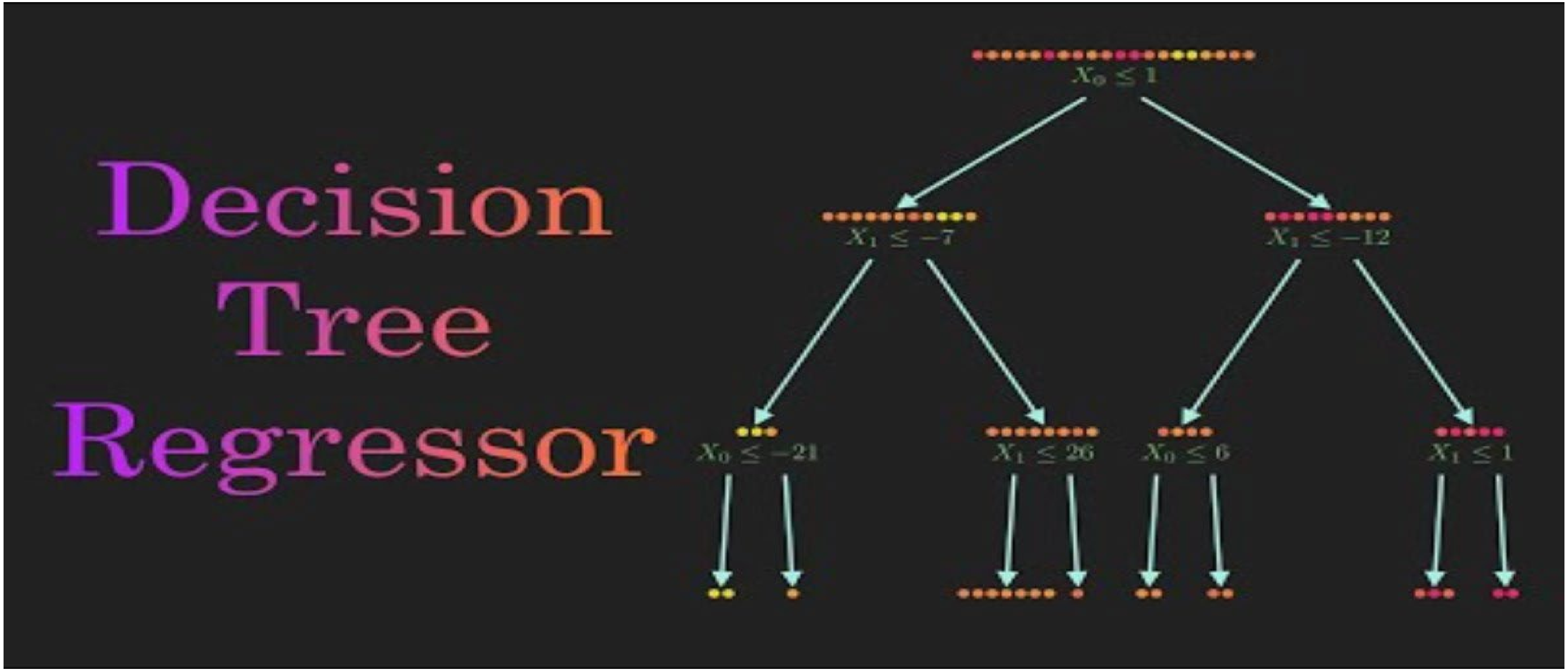
Sample Python code of CART Decision Tree Regression

```
# importing libraries
import numpy as np
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeRegressor

# data collection and data processing
X = np.array([...])  # Input features
y = np.array([...])  # Output values

#Splitting Data:
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Creating the decision tree regressor
regressor = DecisionTreeRegressor()
```

Sample Python code of CART Decision Tree Regression

```
# Training the model
regressor.fit(X_train, y_train)

#Making Predictions
y_pred = regressor.predict(X_test)

#Evaluating the Model:
from sklearn.metrics import mean_squared_error
mse = mean_squared_error(y_test, y_pred)

#Visualizing the Results:
plt.scatter(X_test, y_test, color='black', label='Actual data')
plt.plot(X_test, y_pred, color='blue', linewidth=3, label='Decision
Tree Regression’)

#Displaying the Plot:
plt.show()
```

Decision Tree Regressor



[Source: https://tinyurl.com/4vepj9u3]

Example of CART Decision Tree Regression

Adobe Acrobat
Document

Advantages of Decision Tree Regression

- The decision tree model is simple to analyze, understand, and visualize, and it can be utilized for both classification and regression issues.
- Decision trees can handle both numerical and categorical variables.
- A decision tree requires less work to prepare the data for pre-processing than other algorithms do.
- Decision trees can reveal information about the significance of a feature. By looking at the splits and node impurity values, we can figure out which characteristics have the biggest effects on the target variable.

Disadvantages of Decision Tree Regression

Overfitting

When a model completely matches the training data but fails to generalize to the testing unseen data, it is said to be overfit. When the model memorizes the noise of the training data and misses key patterns, it is said to be overfit. A decision tree that is completely fitted works admirably on training data but poorly on unobserved test data.

High Variance

Overfit models tend to have high variance. A small variance in the data tends to cause a big difference in the tree structure, which causes instability.

Adequate Data

In order to identify the appropriate patterns and make wise conclusions, decision trees require an adequate amount of data, thus they perform better when there is a lot of it. It's challenging for them to learn and perform successfully when they have very little data.

Decision Tree Applications: Classification

- Decision trees are widely used for classification tasks where the goal is to assign a label or category to a given input.

Email Spam Detection

Medical Diagnosis

Credit Card Score

Decision Tree Applications: Regression

- Decision trees can also be used for regression tasks, where the goal is to predict a continuous numerical value.

House Price Prediction

Product Demand Forecasting

Stock Price Prediction

Conclusion

- Decision Tree Regression is a machine learning technique for predicting continuous numerical values.
- It creates a tree structure by repeatedly segmenting the dataset according to important features.
- Decision trees can be interpreted and reveal correlations between features and targets.
- They are susceptible to overfitting when the tree becomes too deep or complicated.
- Overfitting can be reduced using methods like pruning and establishing a limit tree depth.
- Decision Tree Regression is useful in many different fields, such as finance for predicting stock prices, healthcare for predicting patient outcomes, and retail for projecting sales.

References

[1] "The Best Guide On How To Implement Decision Tree In Python," Simplilearn.com. https://www.simplilearn.com/tutorials/machine-learning-tutorial/decision-tree-in-python

[2] GeeksForGeeks, "Decision Tree - GeeksforGeeks," GeeksforGeeks, Oct. 16, 2017. https://www.geeksforgeeks.org/decision-tree/

[3] P. Vadapalli, "Pros and Cons of Decision Tree Regression in Machine Learning," upGrad blog, Dec. 24, 2020. https://www.upgrad.com/blog/pros-and-cons-of-decision-tree-regression-in-machine-learning/

[4] scikit-learn, "1.10. Decision Trees — scikit-learn 0.22 documentation," Scikit-learn.org, 2009. https://scikit-learn.org/stable/modules/tree.html

[5] "Decision Tree In Machine Learning | Decision Tree Algorithm In Python |Machine Learning |Simplilearn," www.youtube.com. https://www.youtube.com/watch?v=RmajweUFKvM&t=949s&ab_channel=Simplilearn (accessed Oct. 06, 2023).

References

[6] “Decision Tree Intuition: From Concept to Application,” Velocity Business Solutions Limited, Jan. 30, 2020. https://www.vebuso.com/2020/01/decision-tree-intuition-from-concept-to-application/#:~:text=There%20are%204%20popular%20types

[7] “Decision Tree Regression Clearly Explained!,” www.youtube.com. https://www.youtube.com/watch?v=UhY5vPfQIrA (accessed Apr. 05, 2022).