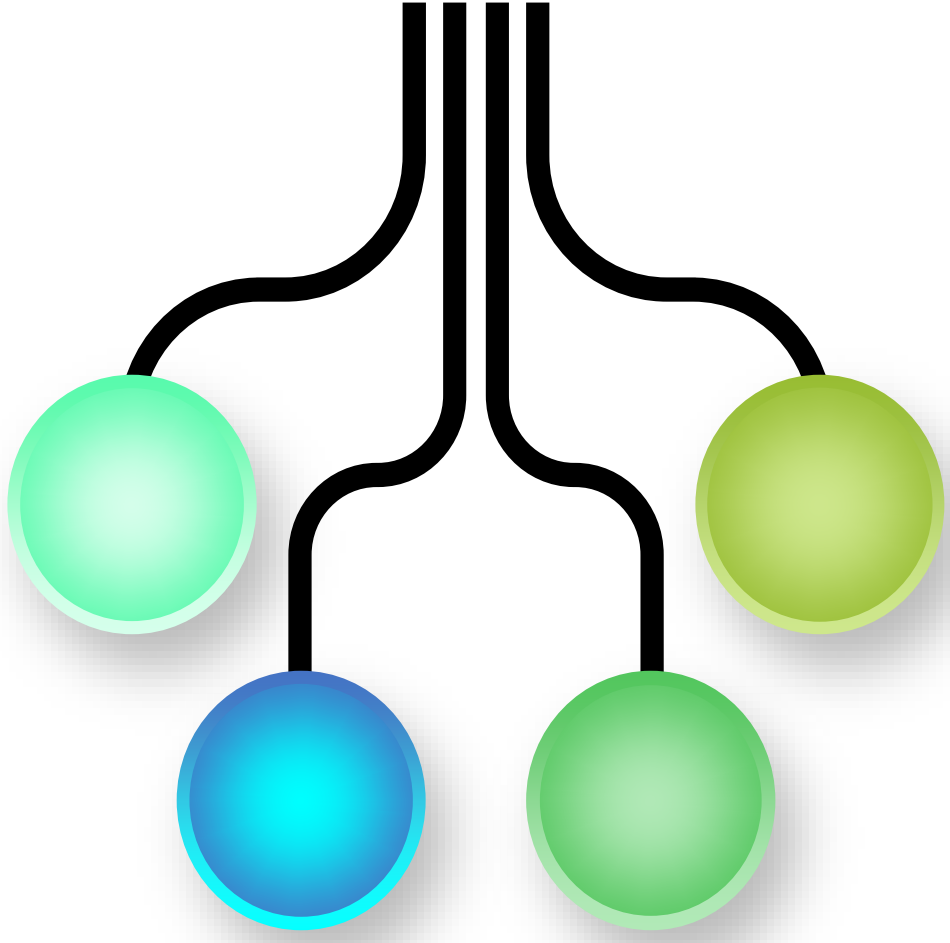


Decision Tree Regression

Presented By:

Harigovind Mohandas
Amjad Ahammed
Melvin Babu Peter

Instructor: Dr. Yasser Alginahi

Contents

- Introduction
- Decision Tree Algorithm
- Terminologies
- Examples
- Splitting of a node
- Where is ML?
- Decision tree Regression
- Variance reduction method
- Example with code
- Applications
- Recap

Introduction

Regression: a statistical method used to model and analyze the relationships between a dependent variable and one or more independent variables.

- In machine learning, regression algorithms predict continuous values based on input features.

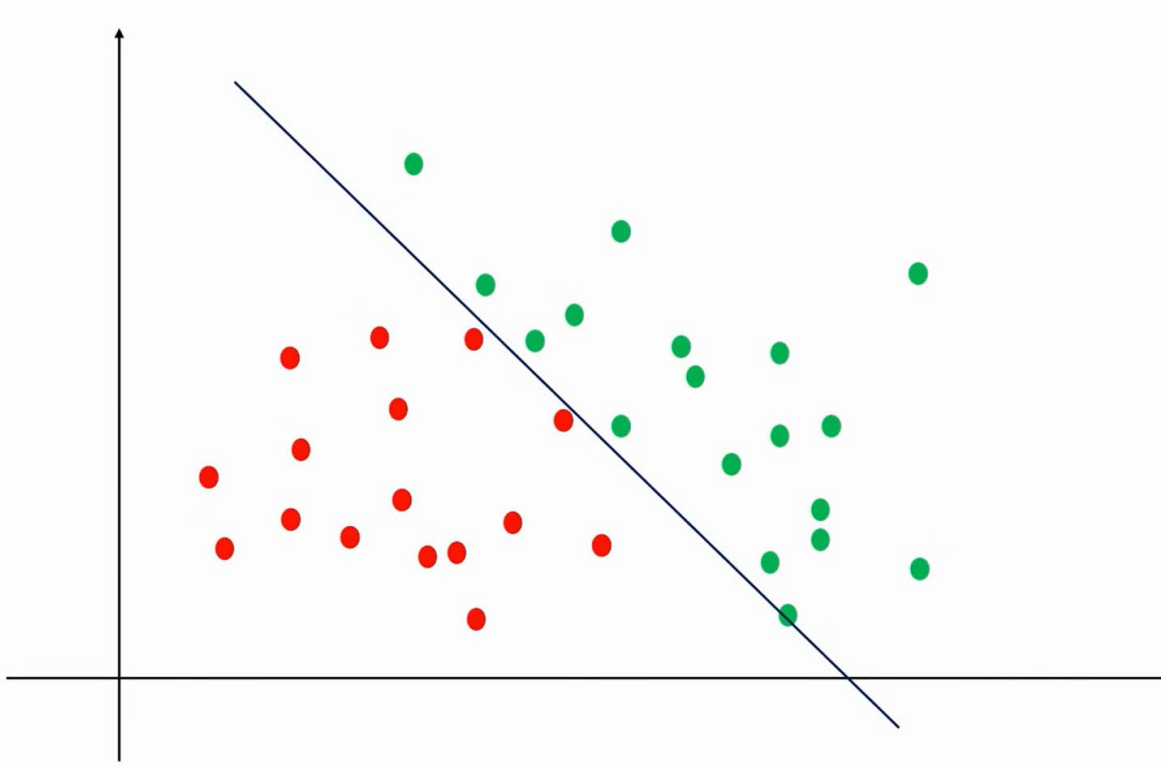
Linear Regression: Fits a linear equation to the observed data.

Polynomial Regression: Uses a polynomial equation, allows for more complex, curved relationships between variables.

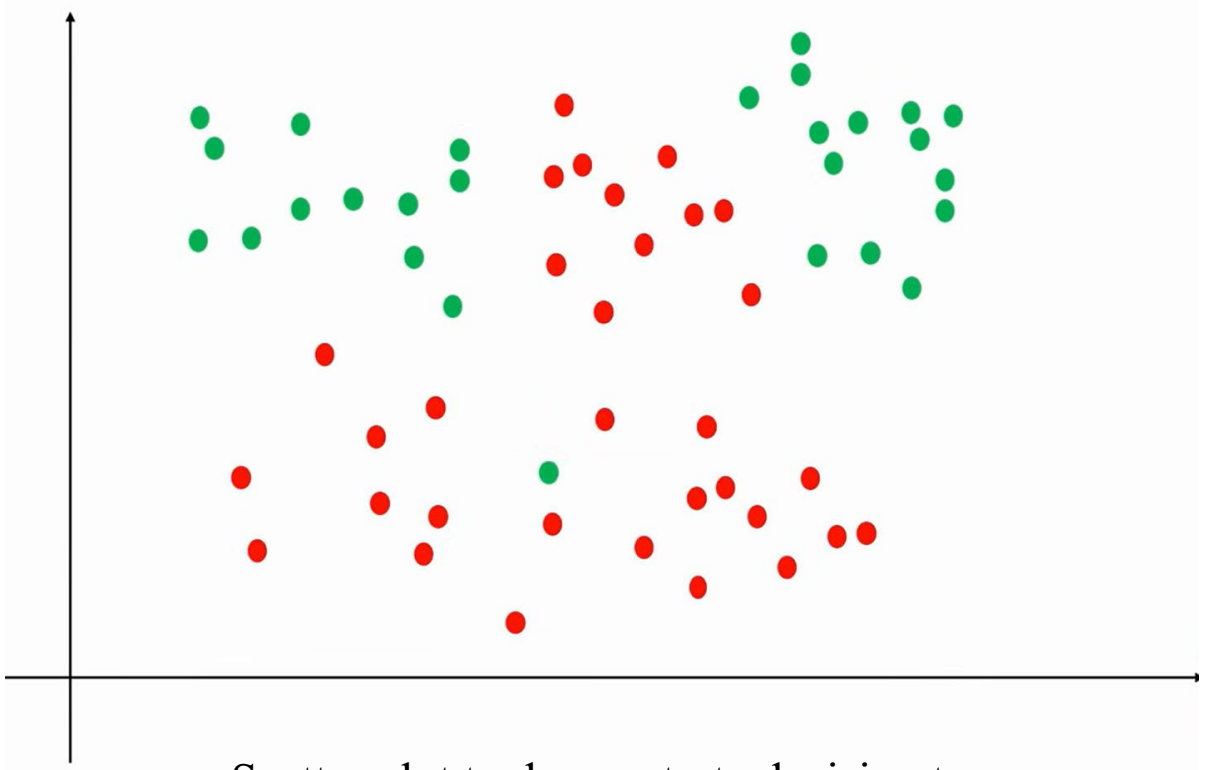
Multiple Regression: Estimates the coefficients of a linear equation that best fits the observed data.

Classification and Regression Trees (CART)

CART are a set of supervised learning models used for problems involving classification and regression



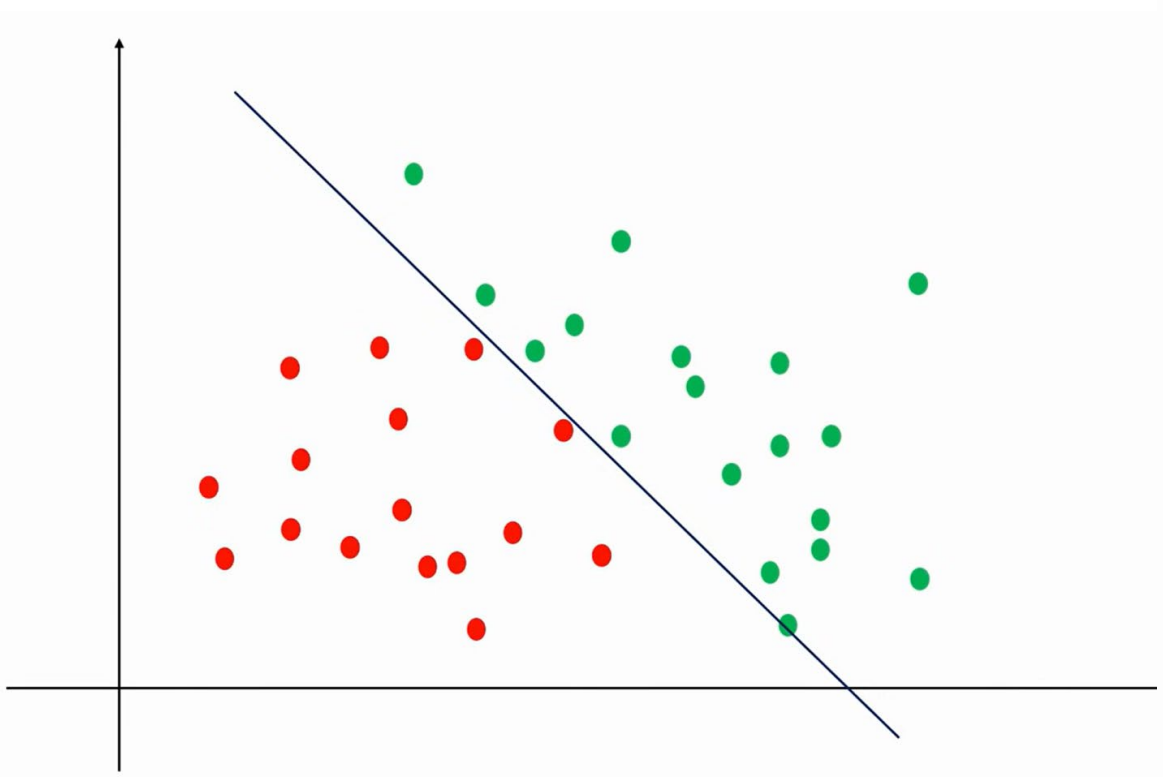
Scatter plot showing linear regression



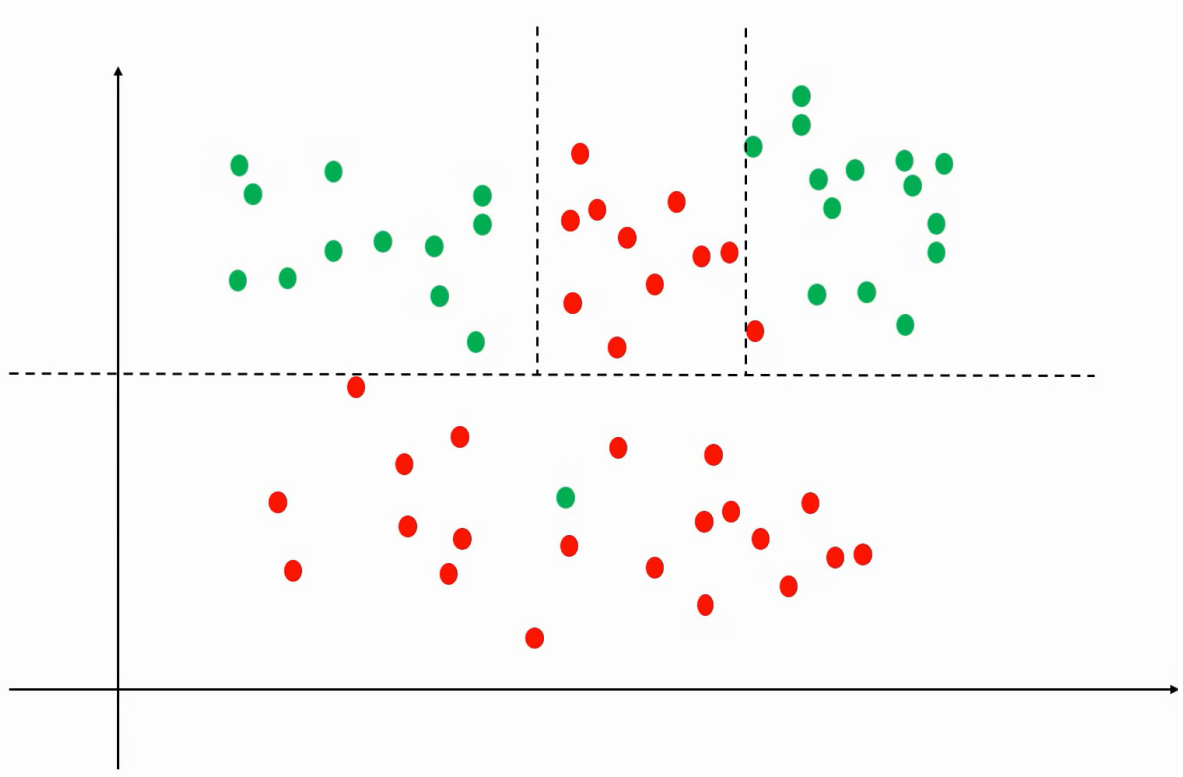
Scatter plot to demonstrate decision trees

[Source: https://www.youtube.com/watch?v=PHxYNGo8NcI]

Classification and Regression Trees (CART)



Scatter plot showing linear regression



Scatter plot to demonstrate decision trees

[Source: https://www.youtube.com/watch?v=PHxYNGo8NcI]

Decision Tree Algorithms

A supervised machine learning algorithm

It is one of the popular machine learning algorithms.

Tree-like structure constructed on the basis of attributes/features .

Non-linear regression technique.

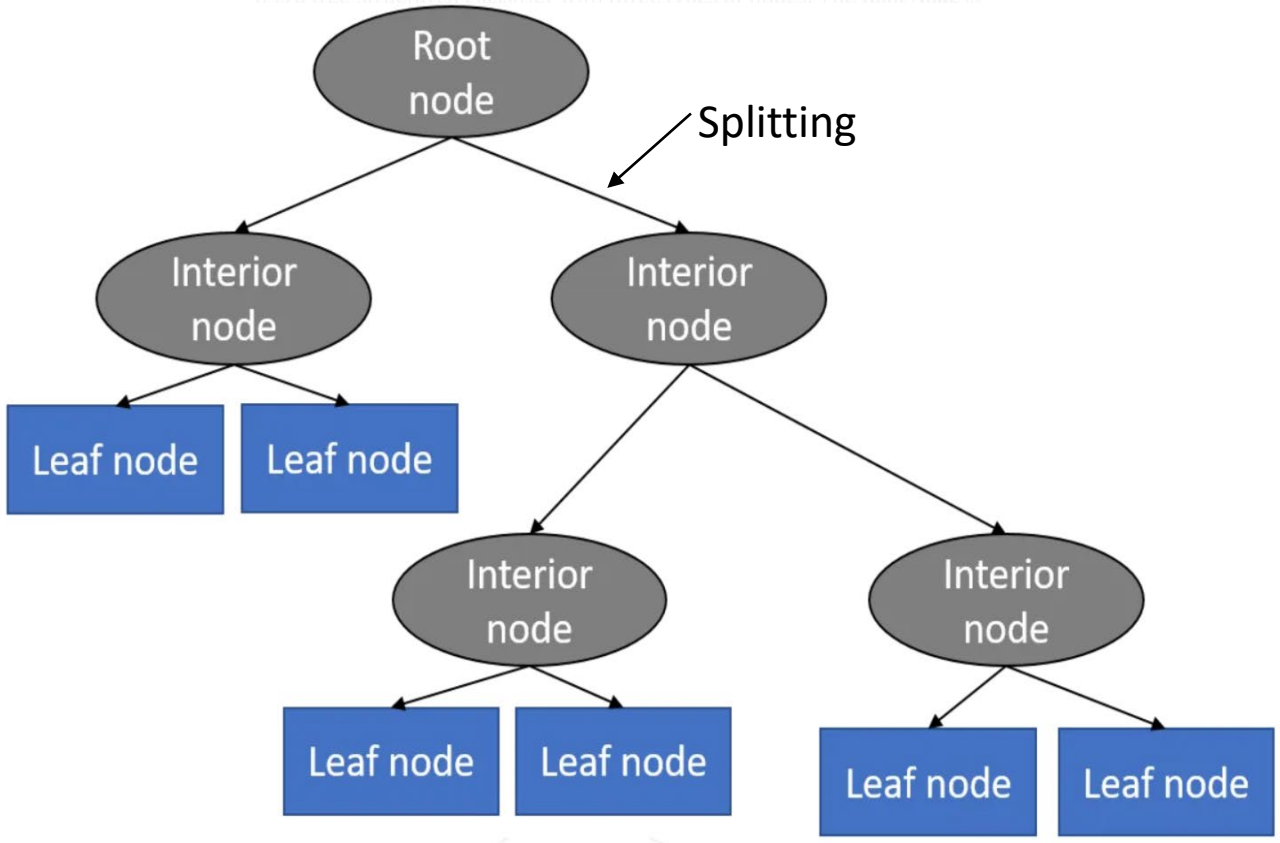
In Regression Trees, each leaf node represents a numeric value

An algorithm that predicts continuous values by making decisions based on features.

It works by recursively splitting the data set into subsets until certain criteria are met.

Applications: Predicting house prices, stock prices, temperature forecasts, etc.

Decision Tree Terminologies

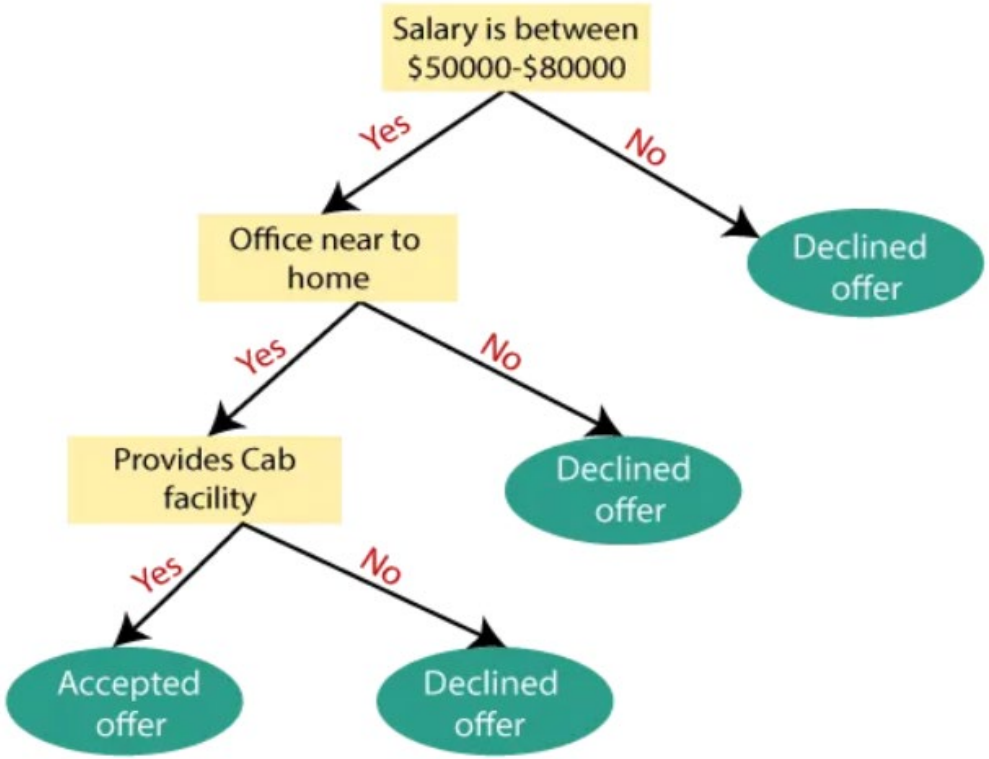


1. **Root Node:** It represents the entire population or sample, and this further gets divided into two or more homogeneous sets.
2. **Leaf/ Terminal Node:** Nodes do not split is called Leaf or Terminal node.
3. **Decision Node:** When a sub-node splits into further sub-nodes, then it is called a decision node.
4. **Branch / Sub-Tree:** A subsection of the entire tree is called a branch or sub-tree.
5. **Parent and Child Node:** A node, which is divided into sub-nodes, is called the parent node of sub-nodes, whereas sub-nodes are the child of the parent node.
6. **Splitting:** It is a process of dividing a node into two or more sub-nodes.

[Source: https://python-course.eu/machine-learning/decision-trees-in-python.php]

Decision Tree Example 1

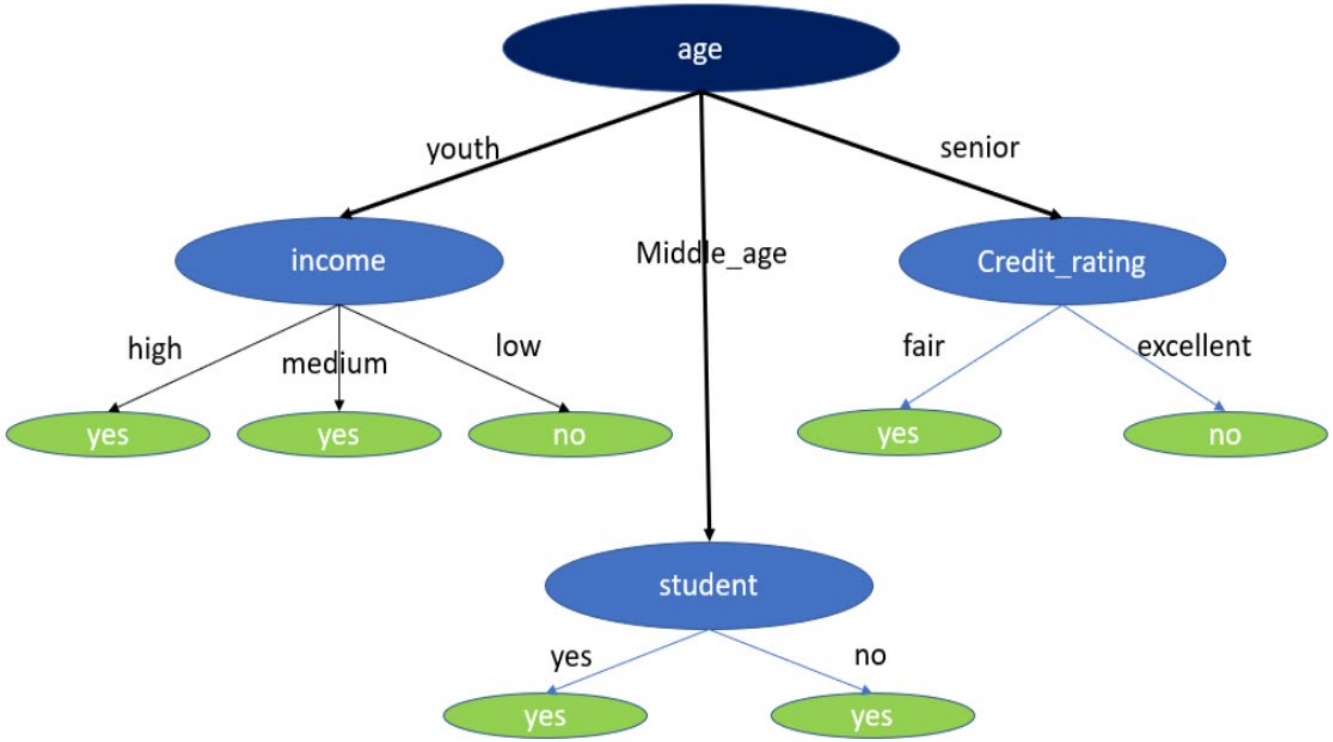
Decision Tree of a Job Offer for a Candidate



[Source: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm]

Decision Tree Example 2

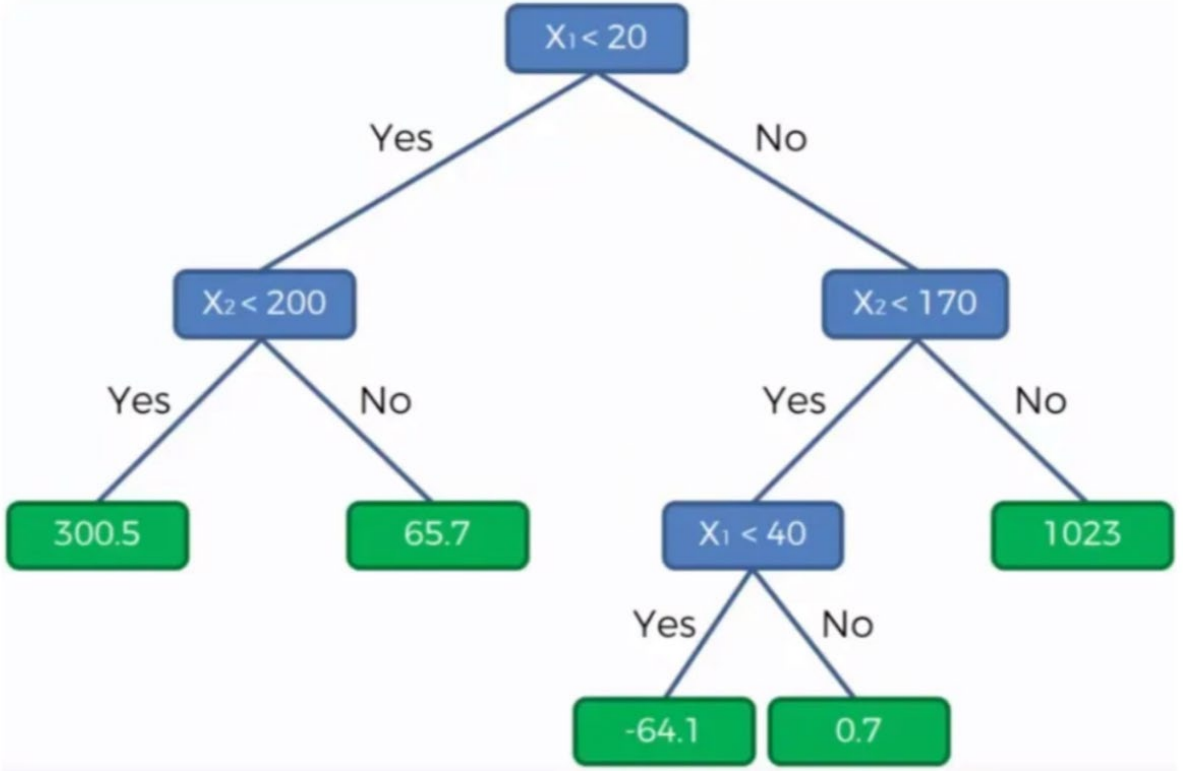
Example buying a computer based on age, income and credit rating



[Source: https://towardsmachinelearning.org/decision-tree-algorithm/]

Decision Tree Example 3

Example of Decision tree regression with numerical (continuous) variables



[Source: https://maniksonituts.medium.com/what-is-decision-tree-regression-dcd0ea40a323]

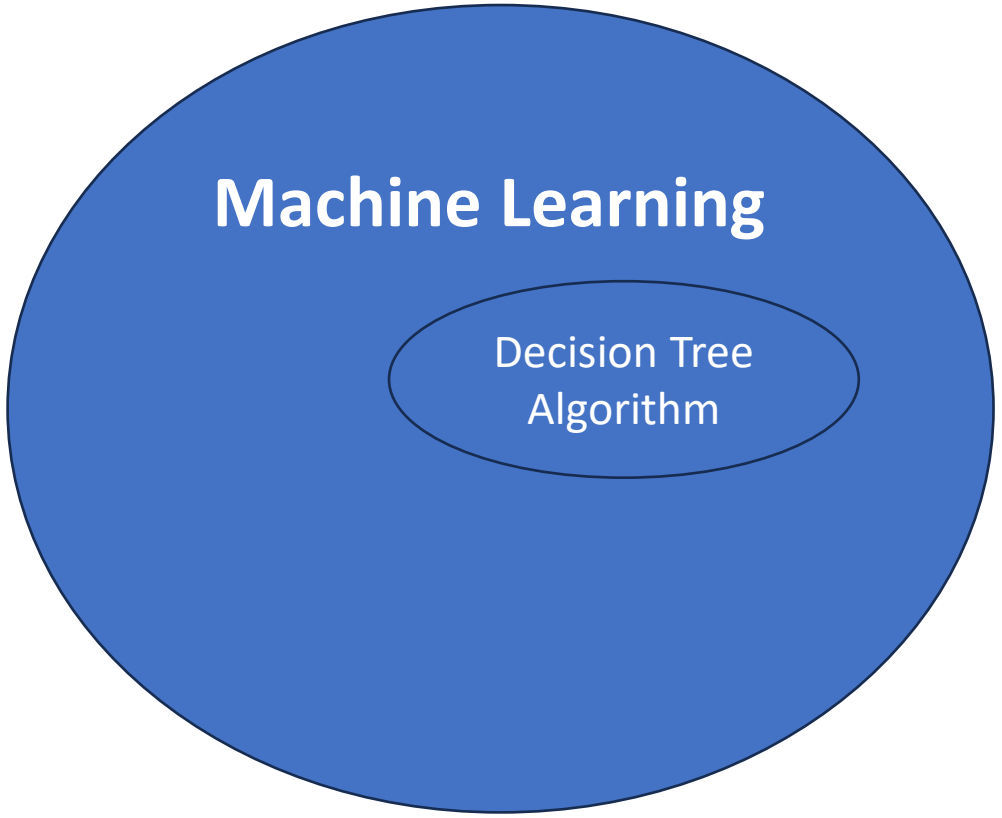
Splitting of Tree

- **Node Splitting:** Divides a node into multiple sub-nodes to achieve purity.
- Uses metrics like Variance or Mean Squared Error (MSE) to evaluate splits.
- If a node is fully homogeneous, its variance is zero.

  Goal: Choose a split that reduces variance the most.
- When a node has <5% of total points or adds no new information, it becomes a leaf.
- Splitting continues until further segmentation is no longer possible

Where is ML?

- **Position**: A specific algorithm under the supervised learning category of ML.
- **Purpose**: Predicts continuous values based on decisions made from features.
- Handles both numerical and categorical data.
- Can capture non-linear patterns without requiring transformation.



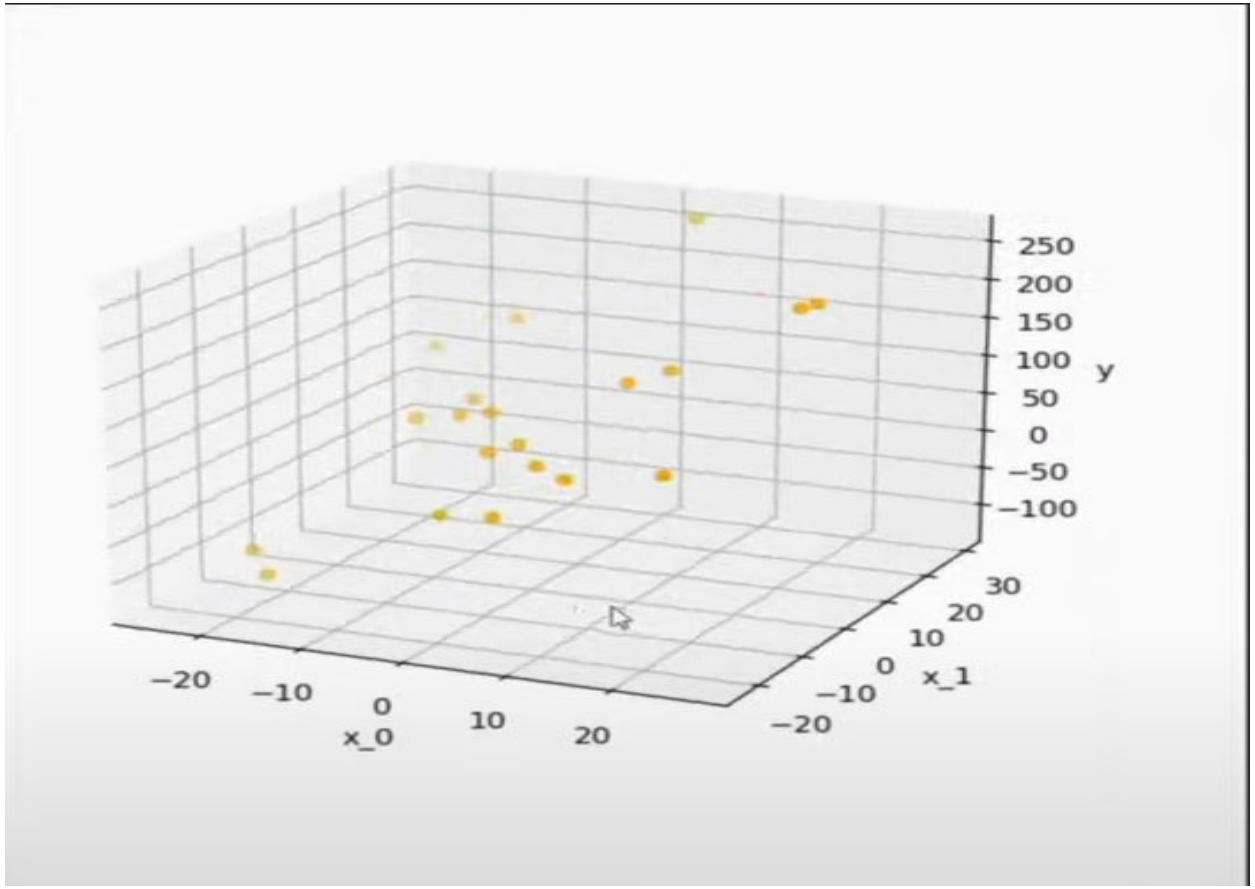
Decision Tree Regression

- A regression model is used to find the dependent (Target) variable using an independent variable.
- Regression is used when the output or target variable is continuous.
- A regression tree is a decision tree used for the task of regression to predict continuous valued outputs.
- A regression model is fit to the target variable using each of the independent variables.

Decision Tree Regression

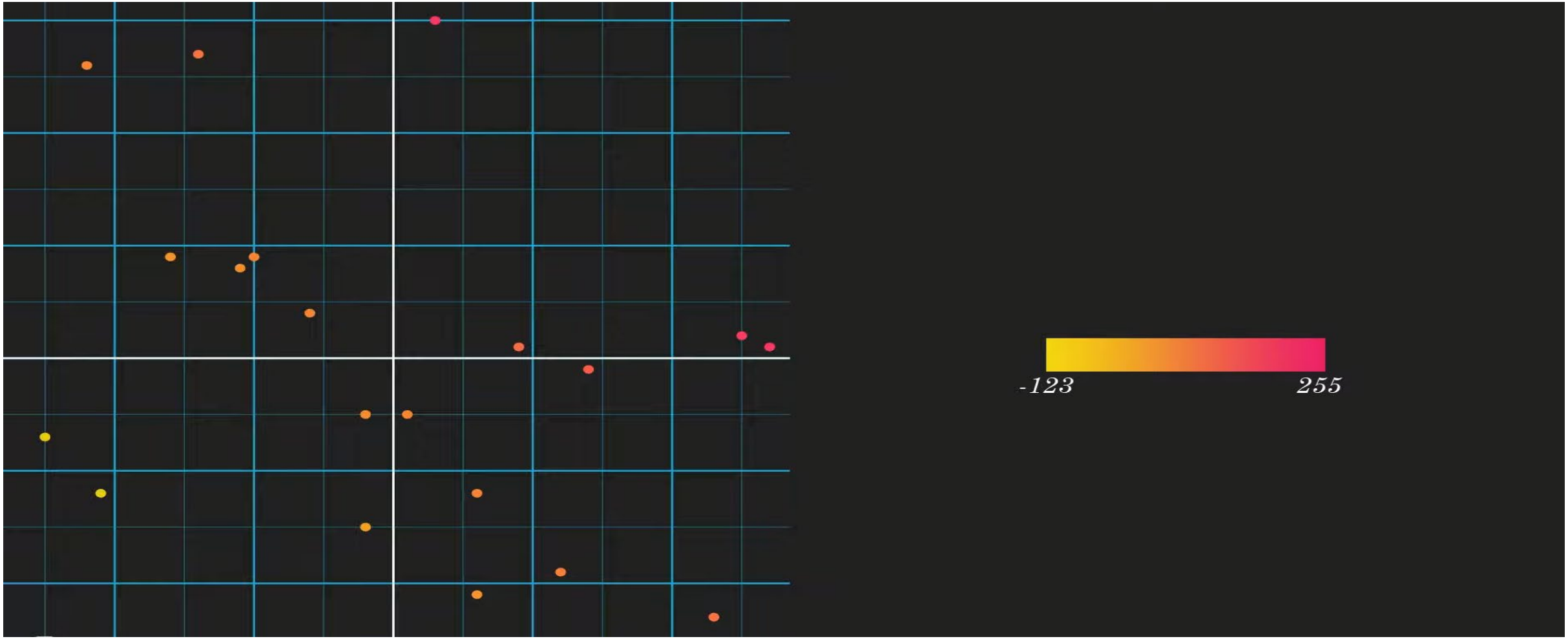
Dataset

S.NO	X0	X1	Y
1	-25	-7	-120
2	-21	-12	-118
3	-2	-15	10
4	1	-5	80
5	-2	-5	70
6	-6	4	60
7	-10	9	50
8	-11	8	40
9	-16	9	30
10	-14	27	140
11	-22	26	20
12	6	-12	110
13	6	-21	100
14	12	-19	90
15	23	-23	120
16	9	1	130
17	1	27	254
18	14	-1	150
19	25	2	252
20	3	30	251



Dataset represented in 3-D [Source: https://tinyurl.com/54454bxx]

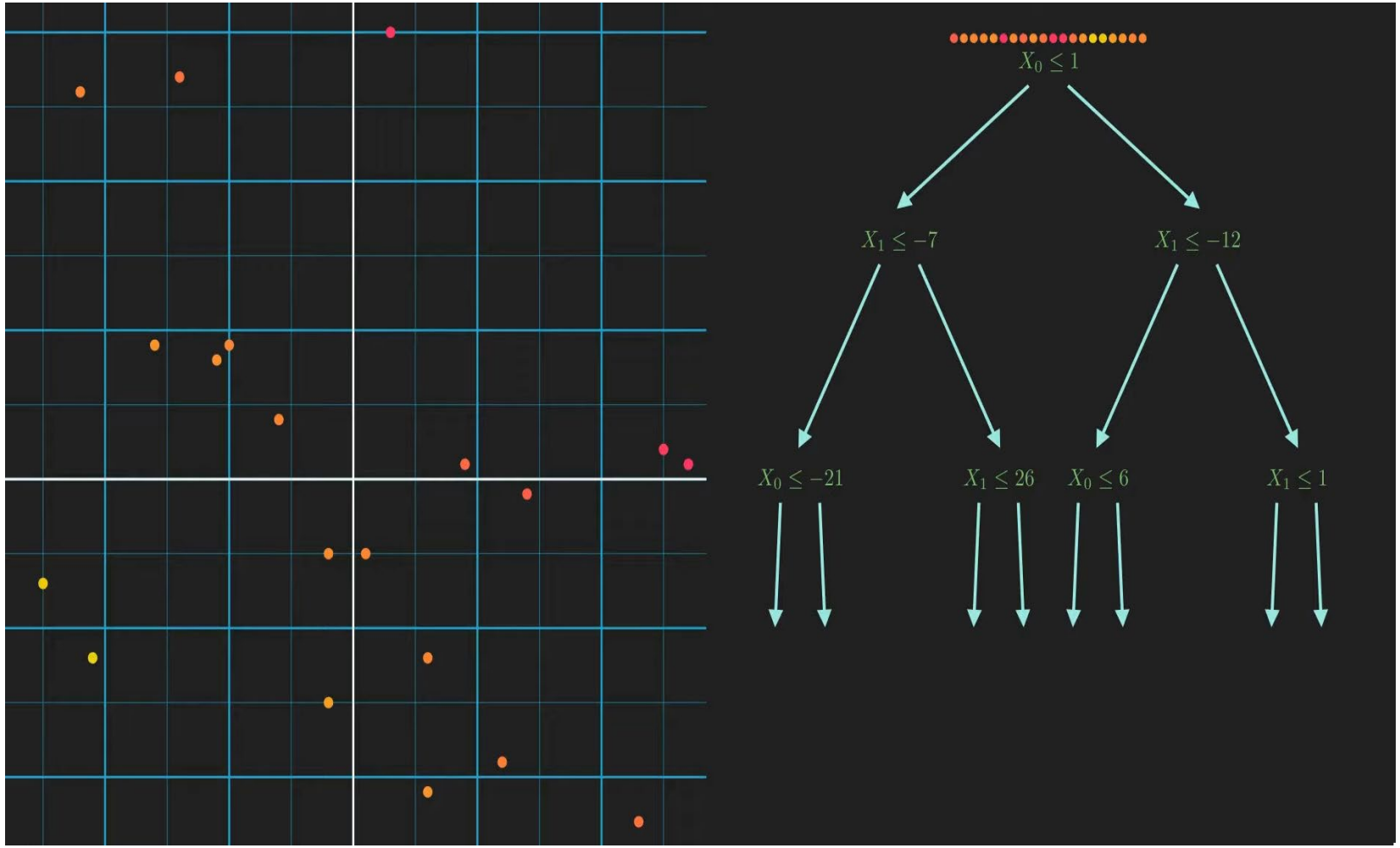
Decision Tree Regression



Data points represented in 2-D [Source: https://tinyurl.com/54454bxx]

Decision Tree Regression

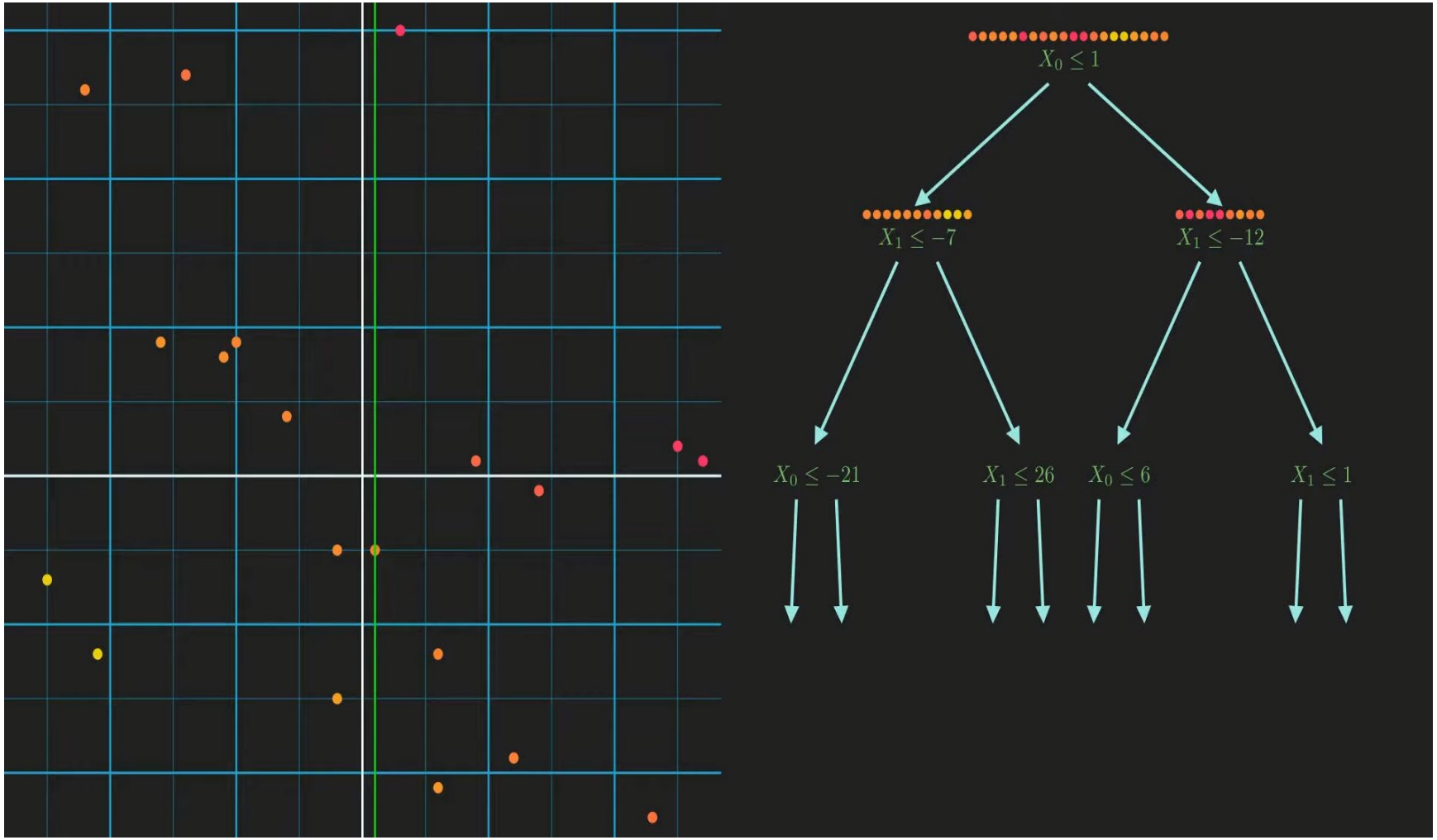
- The node X_0<=1 is called the root node.
- If the point satisfies the condition, then it will move to the left child node
- If the point does not satisfy the condition, it will move to the right child node.
- Child nodes are also called leaf nodes or decision nodes.



Decision tree [Source:https://tinyurl.com/54454bx]

Decision Tree Regression

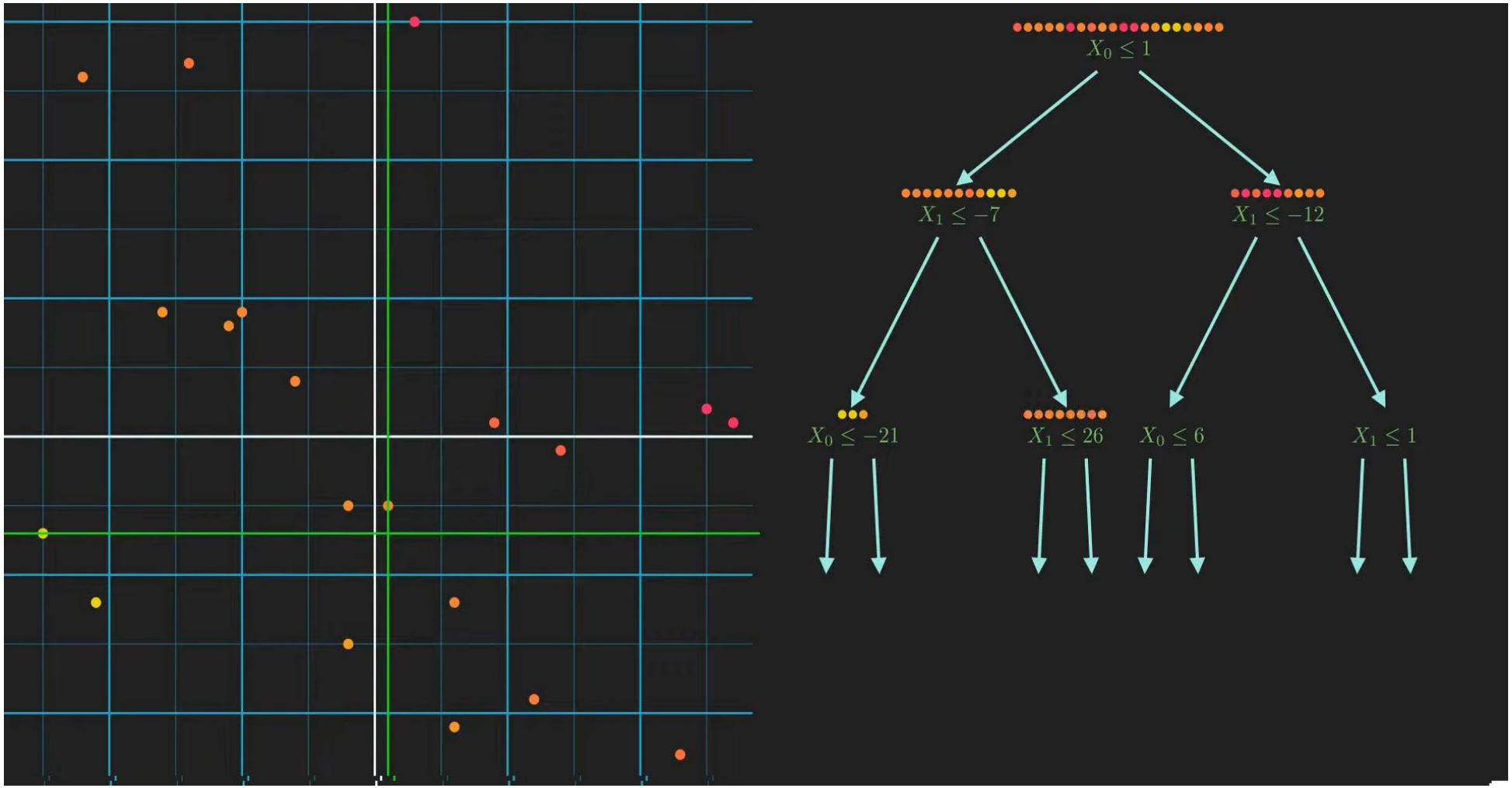
- All the points on the left side and on the line satisfy the condition X_0<=1, So they are placed on the left child and the rest on the right child node
- Yellow points are on the left and red points are on the right indicating the reduction in impurity after the spilt.



Decision tree with root node splitting [Source: https://tinyurl.com/54454bxx]

Decision Tree Regression

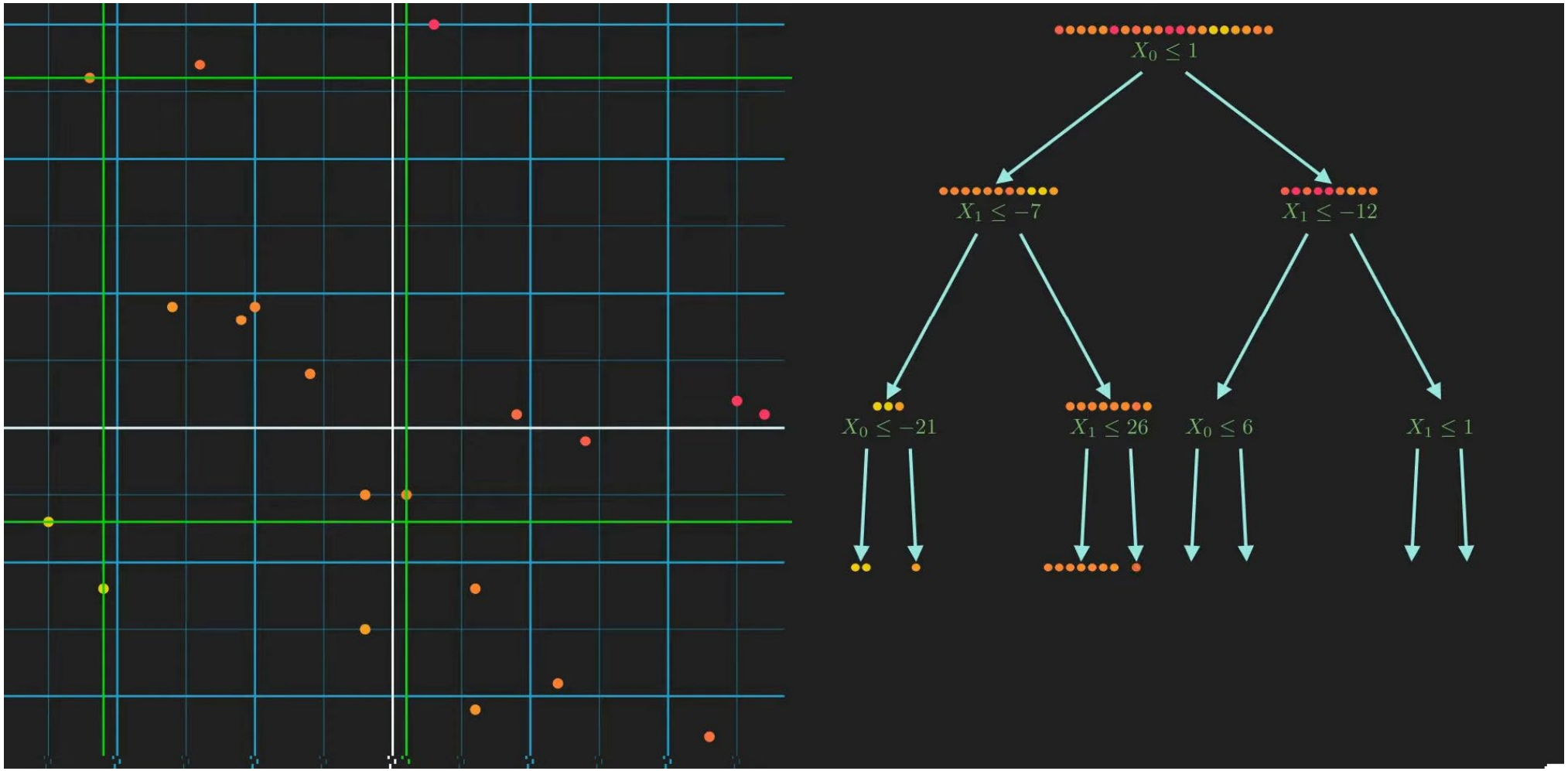
- The points on the left child node are split again using condition X_1<=-7.



Decision tree with further splitting [Source: https://tinyurl.com/54454bxx]

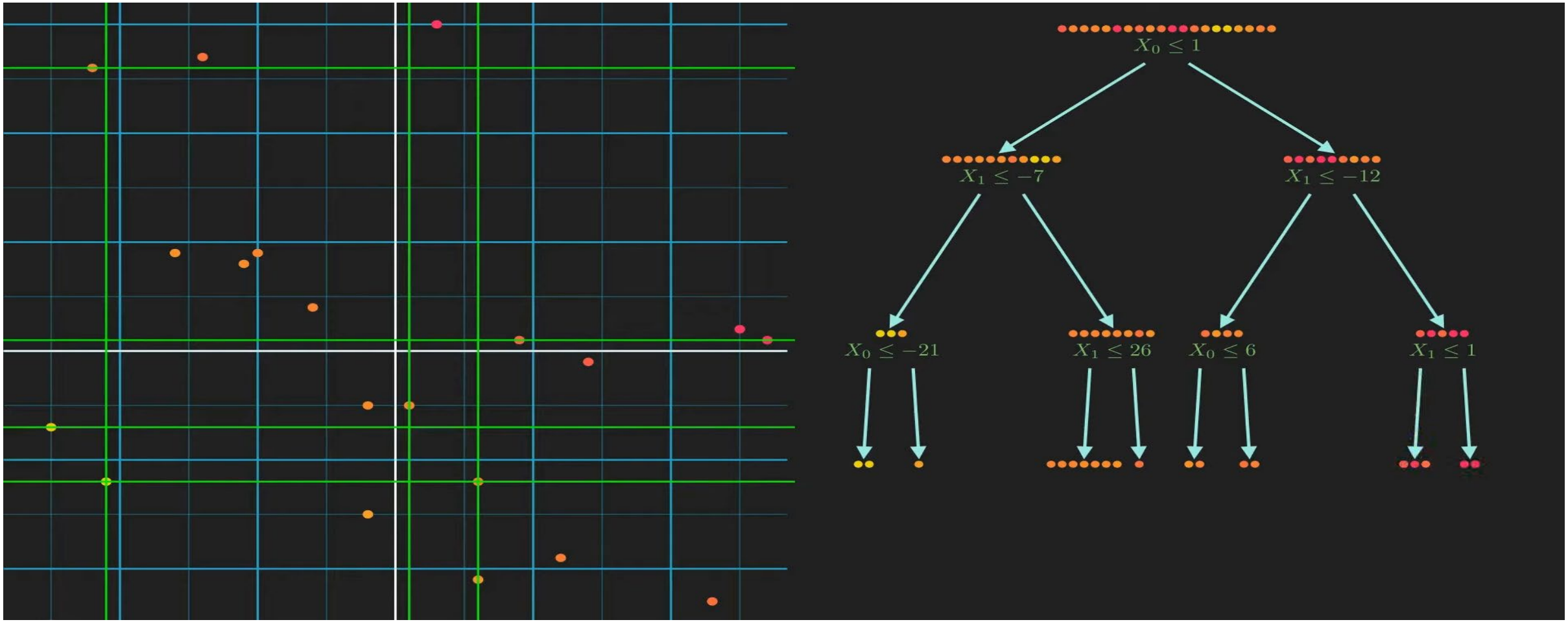
Decision Tree Regression

- The final nodes we get after all the splitting are called leaf nodes.
- All 4 leaf nodes are almost pure.



Decision tree with complete left-side split [Source: https://tinyurl.com/54454bxx]

Decision Tree Regression

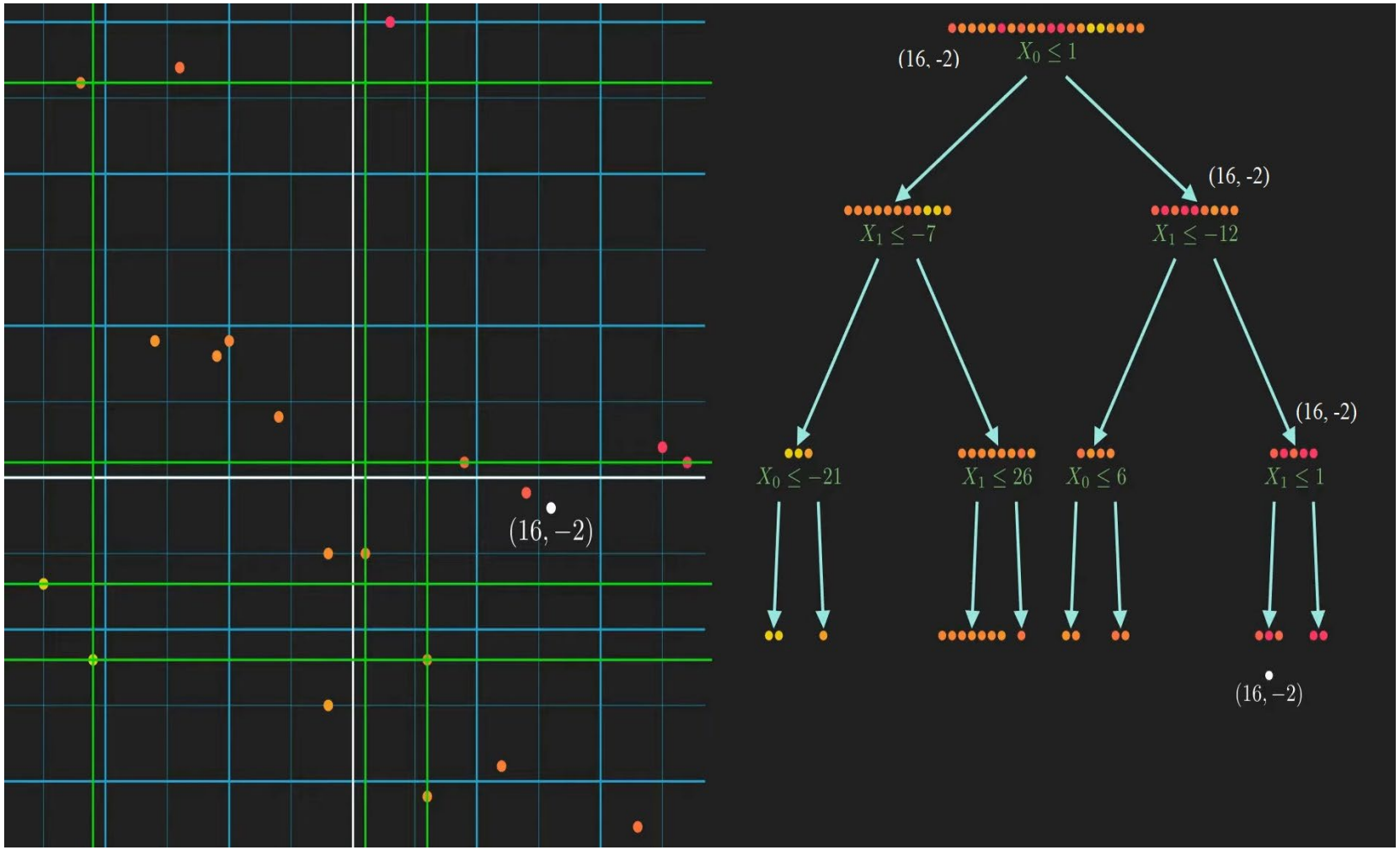


Final Decision tree [Source: https://tinyurl.com/54454bxx]

Prediction of target variable (Y) for new data point

- The predicted value of the target variable is the mean of the Y values of data points present in the leaf node where the new point belongs.
- The Y value of point (16,-2) is

Y_pred =(130+254+158)/3
= 180.67



Prediction of target variable for new data point [Source: https://tinyurl.com/54454bxx]

Variance Reduction Method

- To select the best splitting condition for nodes Variance Reduction Method is used.
- Variance is the measure of the impurity of a dataset, higher value of variance means higher the impurity.

$$Var = \frac{1}{n}\sum(y_i - \bar{y})^2$$

n = Number of data points
y_i = y value of ith data point
$\bar{y}$ = mean value of y

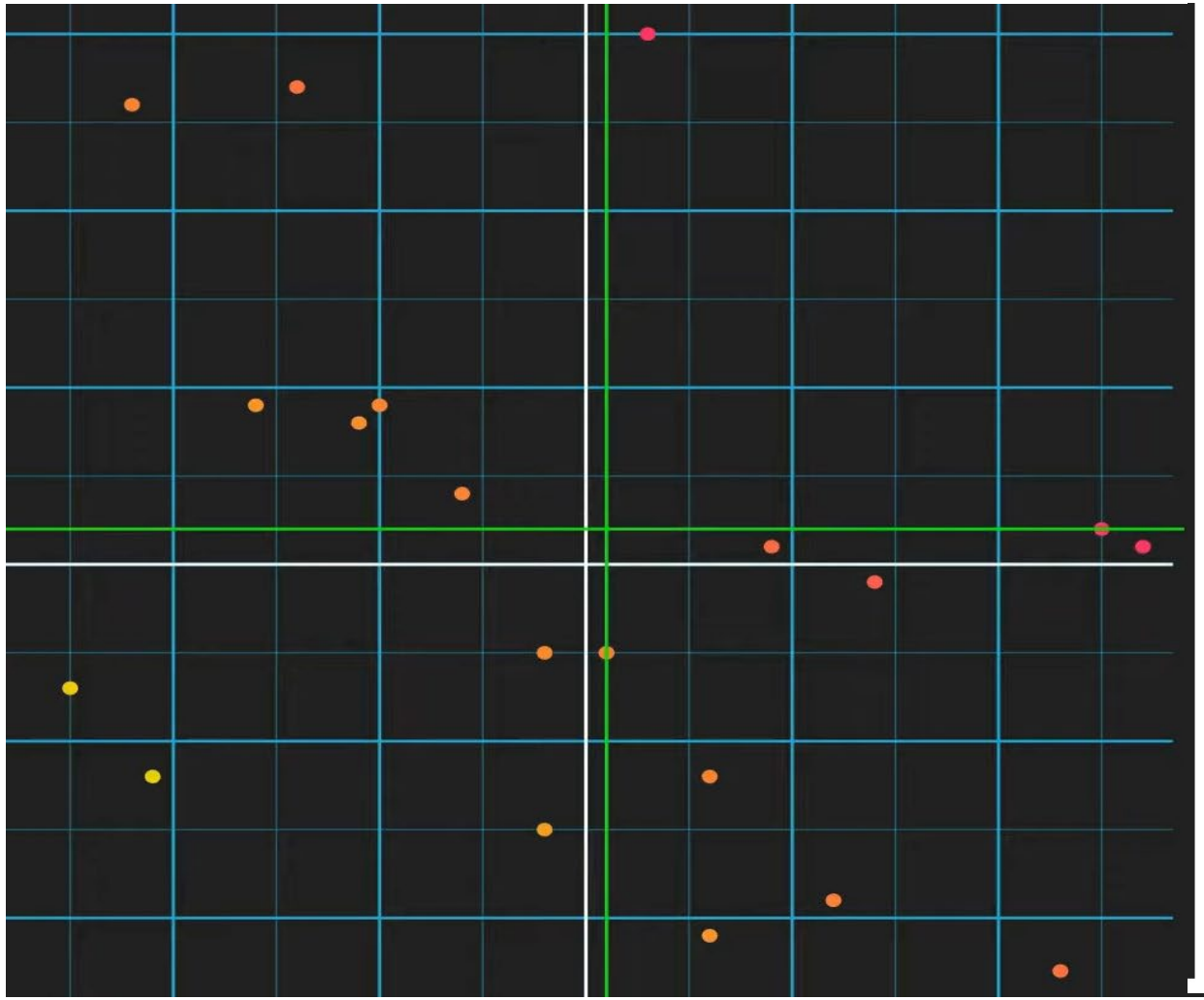
Variance Reduction Method

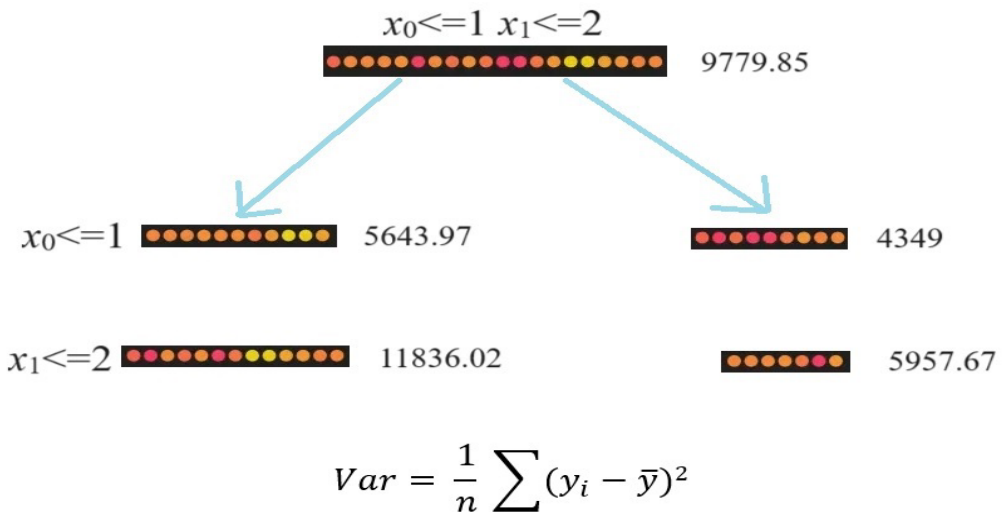
- Variance reduction is calculated using the following equation

$$\text{Variance Reduction (Var Red)} = \text{Var(parent)} - \sum w_i \text{ Var (child } i\text{)}$$

w_i = Weights (Relative size of child node with respect to the parent node)

Variance Reduction Method

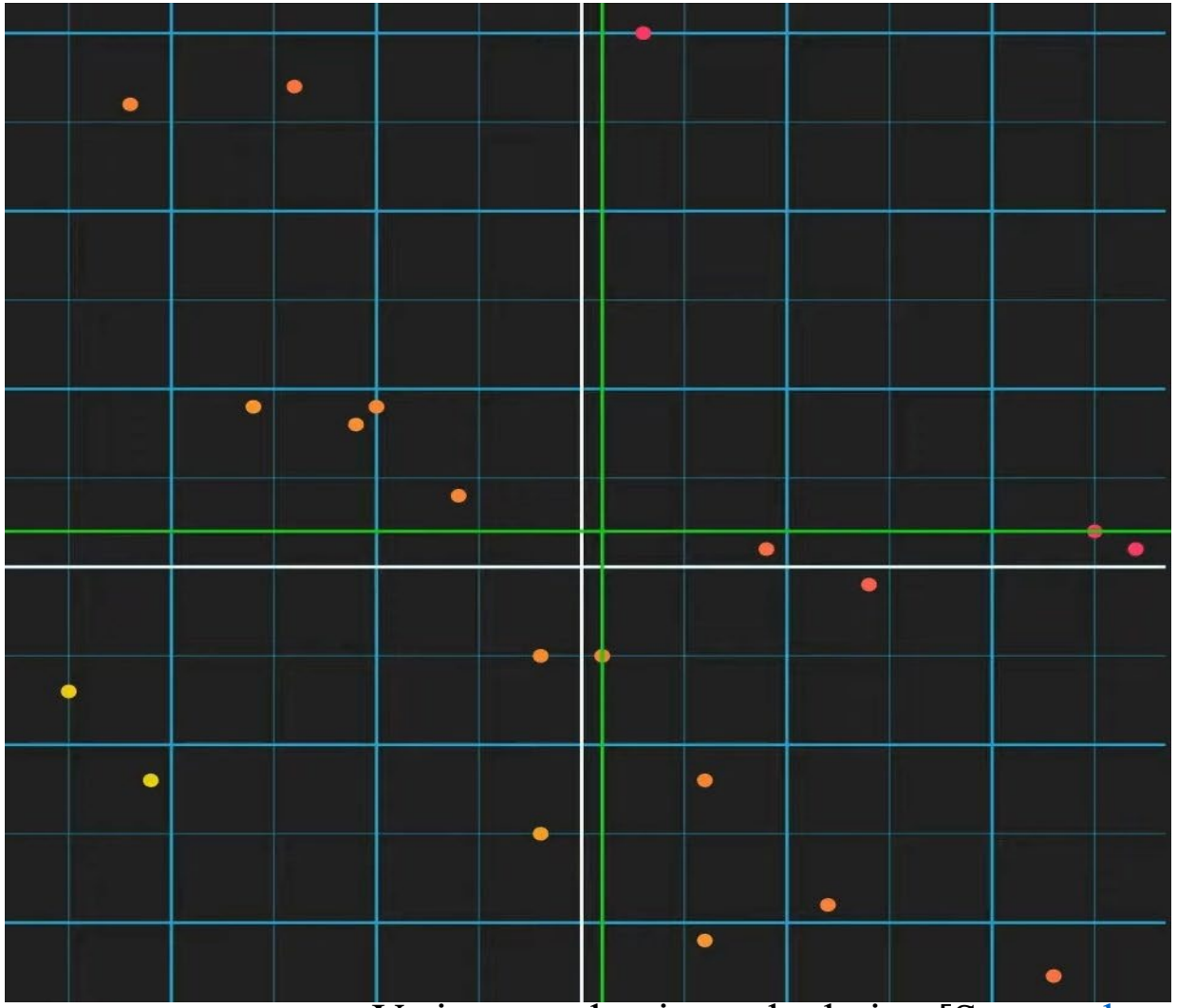


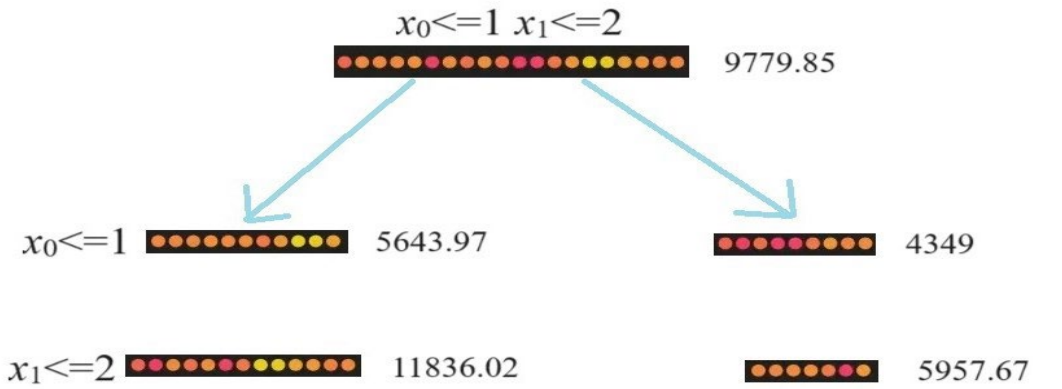


$$Var = \frac{1}{n}\sum(y_i - \bar{y})^2$$

Variance calculation [Source: https://tinyurl.com/54454bxx]

Variance Reduction Method





Variance Reduction (Var Red) = Var(parent) - Σ w_i Var (child i)

Var Red1= 9779.85 - (11/20) * 5643.97 - (9/20) * 4349 = 4718.62

Var Red2 = 9779.85 - (13/20) * 11836.02 - (7/20) * 5957.67 = 1.25

Var Red1 > Var Red2

Variance reduction calculation [Source: https://tinyurl.com/54454bxx]

For Further Reference

SuperDataScience - Decision Tree Regression Clearly Explained
[source: https://youtu.be/_wZ1Lo7bhGg?feature=shared]

Normalized Nerd - Decision Tree Regression Clearly Explained
[source: https://youtu.be/UhY5vPfQIrA?si=d-r1HVzmpszId551]

Maximum Depth of Decision Tree

- While coding decision tree regression, we use a max_depth parameter to control the maximum depth of the tree.
- If the max_depth parameter is set to high, the tree learns fine details of the training data and learns from the noise, which results in overfitting.

Example

The data shows Production Cost and its respective profit for different products.

Production Cost	Profit
100	1000
500	3000
1500	5000
3500	8000
5000	6500
6000	7000
8000	15000
9500	20000
12000	21000
14000	25000
15500	27000
16500	30000
25000	52000
30000	80000

Code for Example

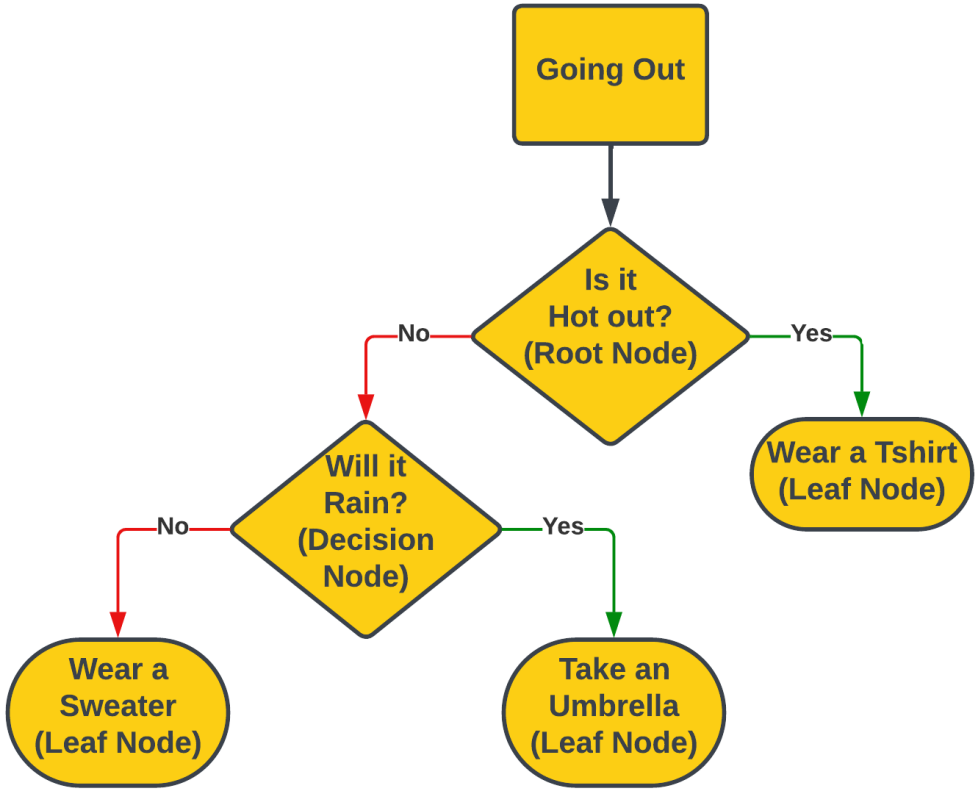
Decision Tree Regression – Colab
[Source: https://tinyurl.com/3atk5tn4]

Decision Tree Regression – GeeksforGeeks Example
[Source: https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/]

Further Reference on sklearn.tree.DecisionTreeRegressor
[Source: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html]

Practical Applications

Decision Making in our daily life



Practical Applications

Health Care

In healthcare, decision trees can help medical professionals with diagnoses. For example, based on symptoms (decision nodes), a doctor can narrow down the possible conditions (leaf nodes).

Financial Analysis

In the financial sector, decision trees are used in options pricing and strategy development. They can model possible future price movements based on different market conditions to help investors make informed decisions.

Recap

- Regression is a statistical method for modeling relationships between dependent and independent variables. Regression predicts continuous values based on input features.
- CART is a set of supervised learning models used for classification and regression.
- CART uses decision trees to make predictions based on data.
- Predicted value of the target variable is the mean of the Y values of data points present in the leaf node.
- Variance reduction helps select the best splitting conditions for nodes.
- Control overfitting by setting a maximum depth.
- It has practical applications in health care and financial analysis.

References

[1] N. Singh, “Decision Tree Regression,” *The Startup*, Nov. 26, 2020. https://medium.com/swlh/decision-tree-regression-c977b732eb51

[2] “Decision Tree Algorithm,” *TowardsMachineLearning*, Jul. 09, 2021. https://towardsmachinelearning.org/decision-tree-algorithm/

[3] “Decision Tree Regression Clearly Explained!,” *www.youtube.com*. https://youtu.be/UhY5vPfQIrA?feature=shared (accessed Oct. 06, 2023).

[4] “Python | Decision Tree Regression using sklearn,” *GeeksforGeeks*, Oct. 04, 2018. https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/

[5] “sklearn.tree.DecisionTreeRegressor — scikit-learn 0.23.2 documentation,” *scikit-learn.org*. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

[6] A. Singh, “Why: Practical Applications of Decision Trees (Part 2),” *Medium*, Jun. 08, 2023. https://medium.com/@diehardankush/why-practical-applications-of-decision-trees-ae09e04b2b16 (accessed Oct. 06, 2023).