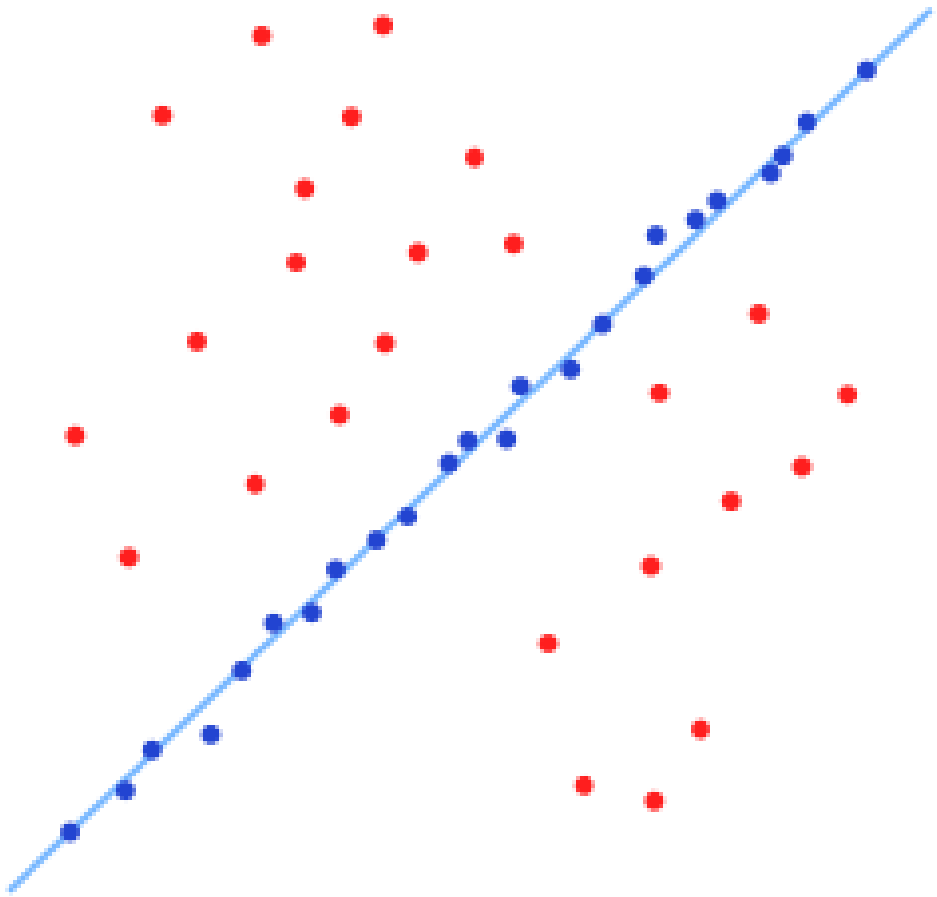
Machine Learning for Artists Week four

Alternatives to linear regression

Problems with linear models

1. Linear models are highly sensitive to outliers (atypical points) in the training dataset.
2. Most phenomena in real life are not accurately represented by linear models. (We can say, roughly speaking, that most phenomena are not linear).

We can solve the first problem using RANSAC.
This method uses randomized sampling to eliminate outliers.
But we are still left with the second problem.



Source: wikipedia

To tackle the second problem:

Instead of a linear model, we can try nonlinear regression methods.

Alternatives to linear models

There are many such alternatives. We are going to consider only two in this lesson.

1. KNN Regression
2. Decision Tree Regression

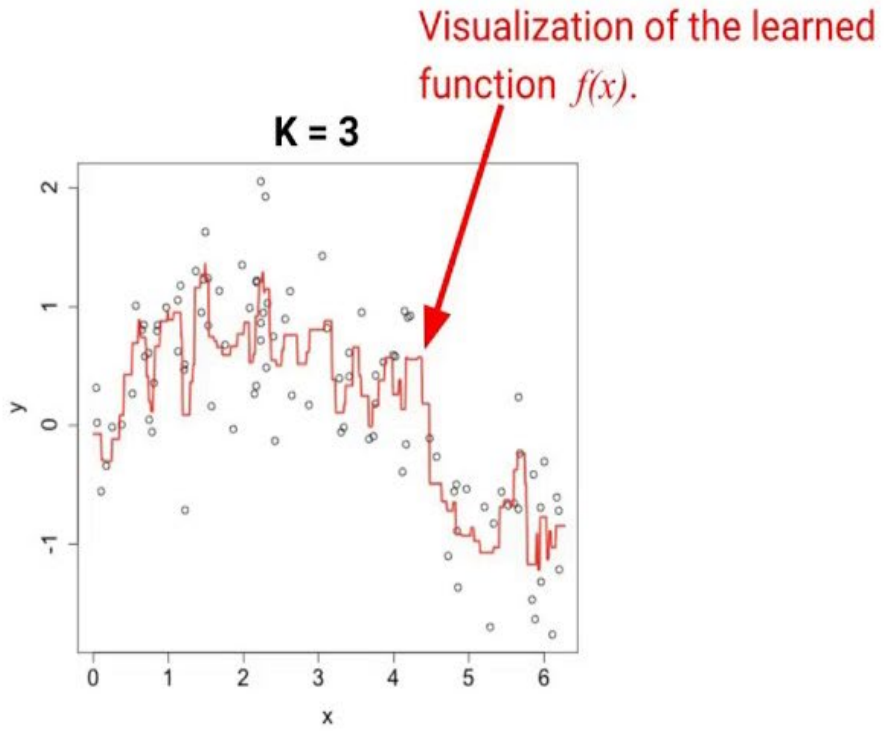
Alternatives to linear models: KNN Regression

- *This approach assigns to any input a value that is theaverage of the values of its nearest neighbours.*

- How do we select nearest neighbors?
- In KNN regression, we choose the K datapoints closest to the input we wish to regress, using some measure of distance.
- In Radius NN regression, we choose those datapoints whose distance to the input is within a specified radius.

- The value of K or of the radius are chosen by the programmer, or using other methods.

Alternatives to linear models: KNN Regression

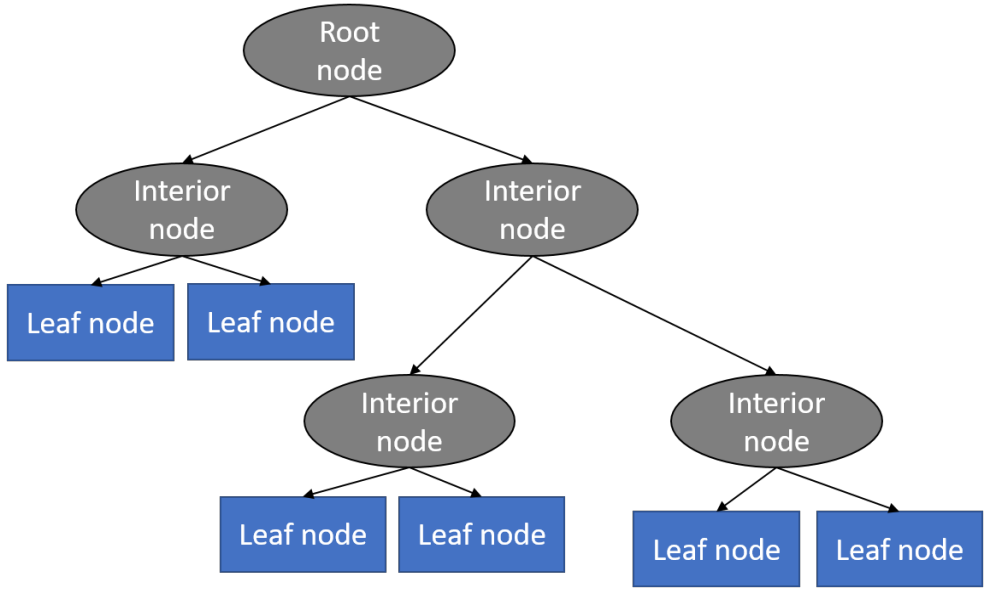


Source: https://www.youtube.com/watch?app=desktop&v=BvQVovMefsM

Alternatives to linear models:
Decision Tree Regression

A hierarchical tree is a collection of nodes and edges branching out from a root.

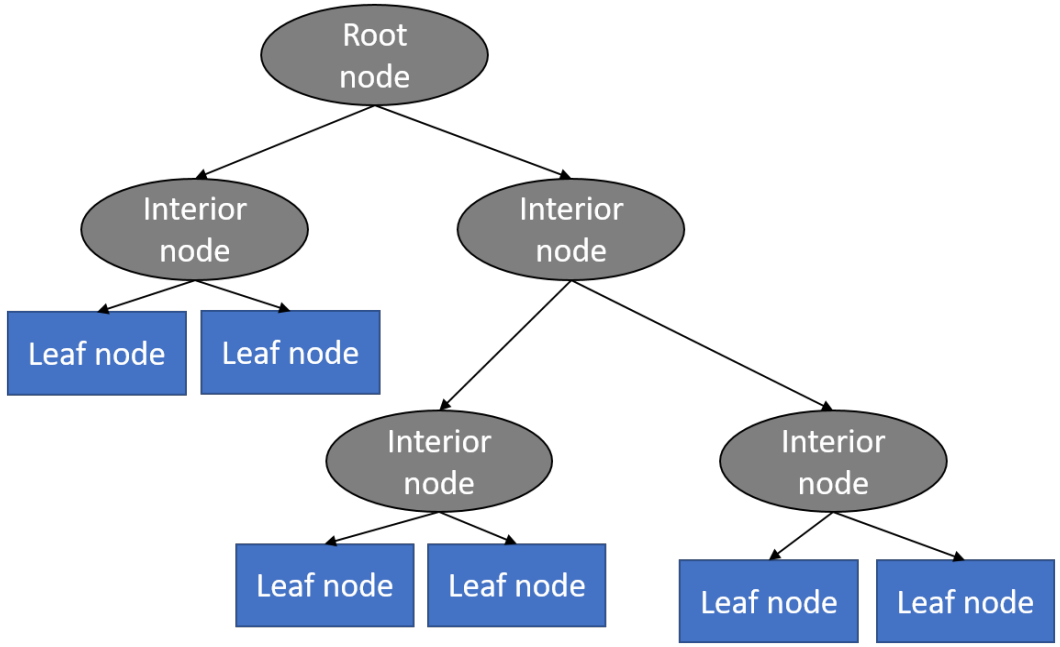
Every node (except the root) has exactly one incoming edge.



Source: https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda

Decision Tree Regression

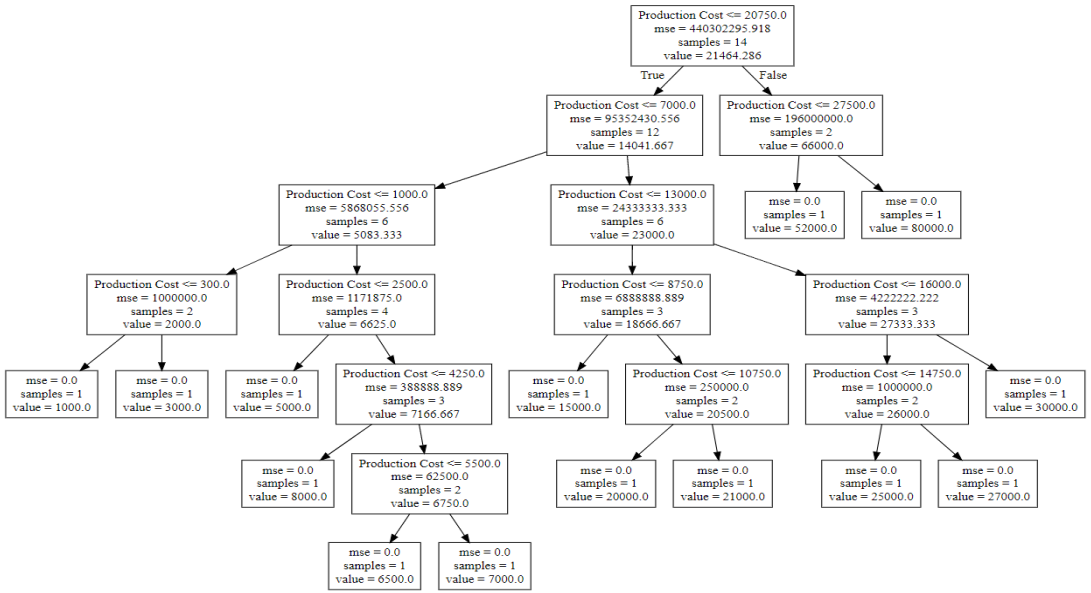
- Nodes are of two kinds, terminal and non-terminal.
- Non-terminal nodes are divided: they have exactly two outgoing edges (left and right). Non-terminal nodes are also called “interior nodes”.
- Terminal nodes have no outgoing edges. Terminal nodes are also called “leaf nodes”.



Source: https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda

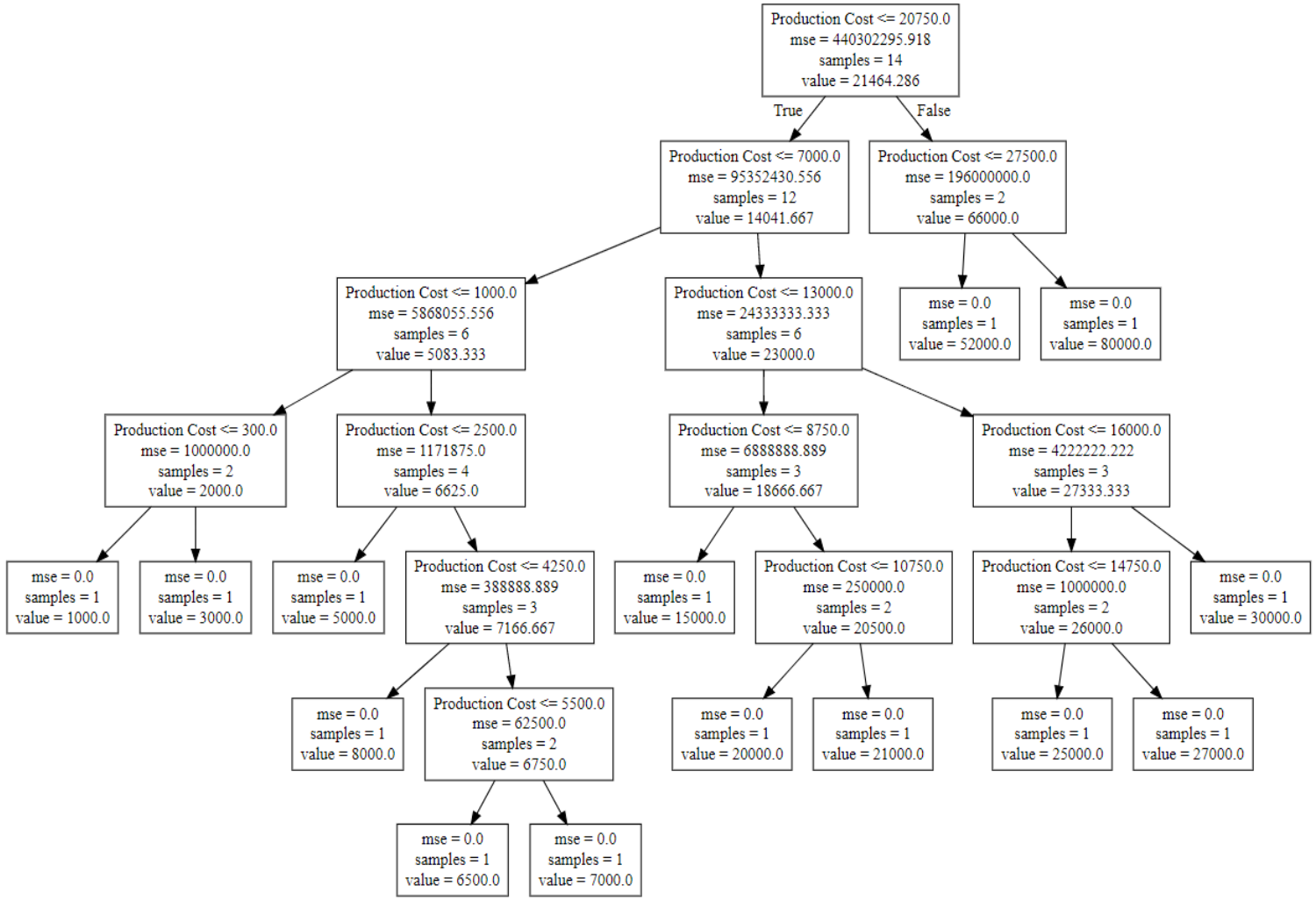
Decision Tree Regression

- A decision tree is the representation of a sequence of tests or questions used to arrive at a decision regarding the output value of some input.
- Each test or question is used to eliminate some possible outputs.
- Every non-terminal node represents such a test.



Source: https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/

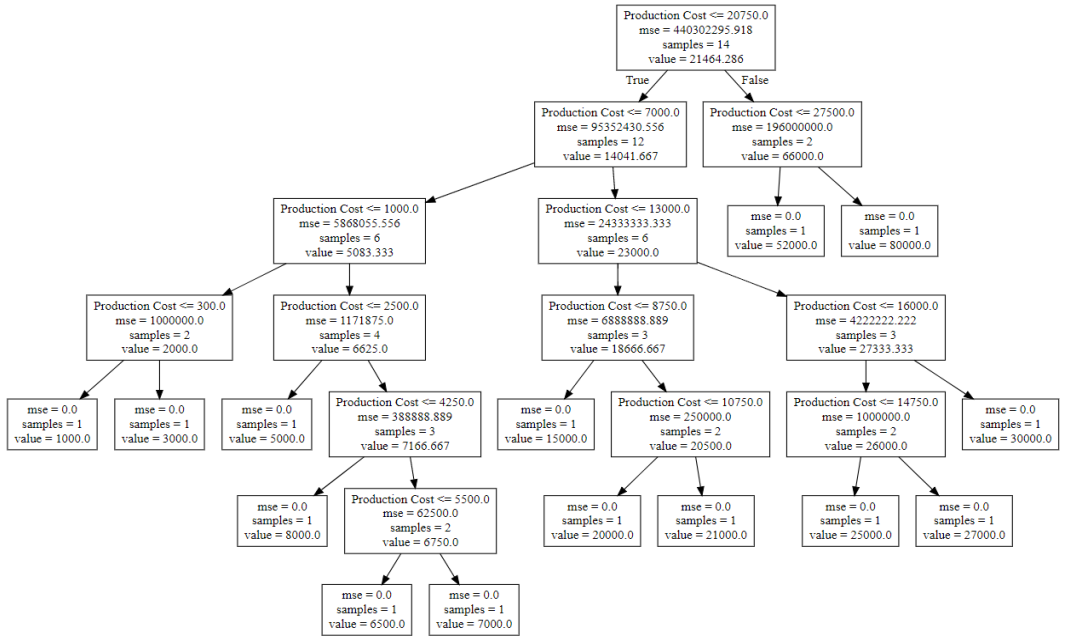
Decision Tree Regression



Source: https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/

Decision Tree Regression

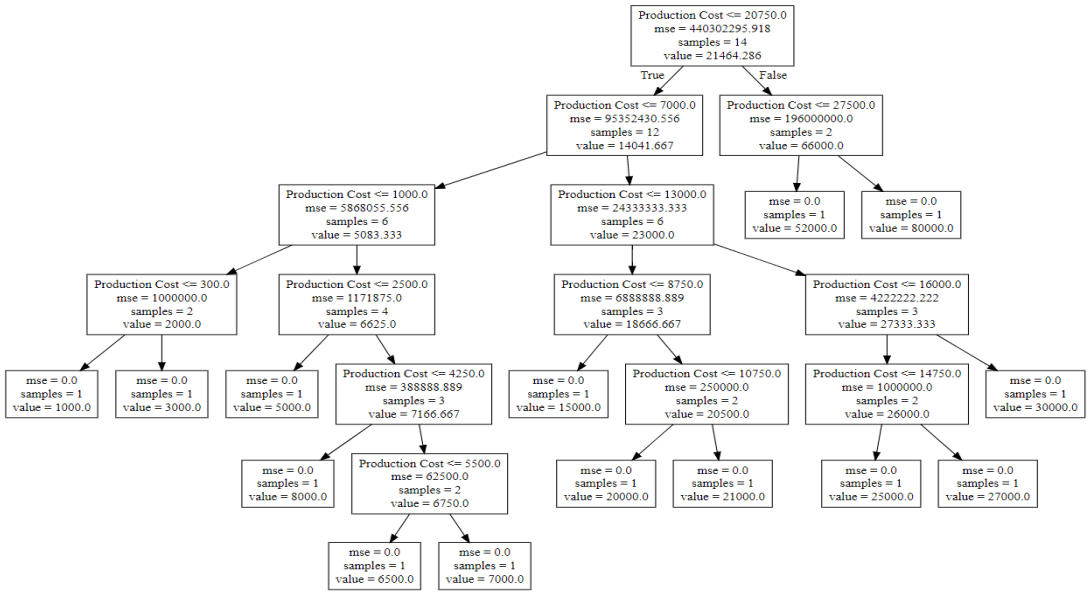
- Given an input, a decision tree applies a test or asks a question about the input.
- Depending on the result of the test or the answer to the question, the input is channeled to the left or right node and the test represented by the new node is then applied to the input.
- The process continues recursively until a terminal node is reached.
- The input datapoint is thus “propagated” through the tree.
- At every step, the dataset is partitioned into two subsets such that similar points are in the same subset and different points are in different subsets.



- The training phase optimizes the parameters associated with the internal nodes and leaf nodes for the task of regression.

Decision Tree Regression

Tutorial on decision tree regression: https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/

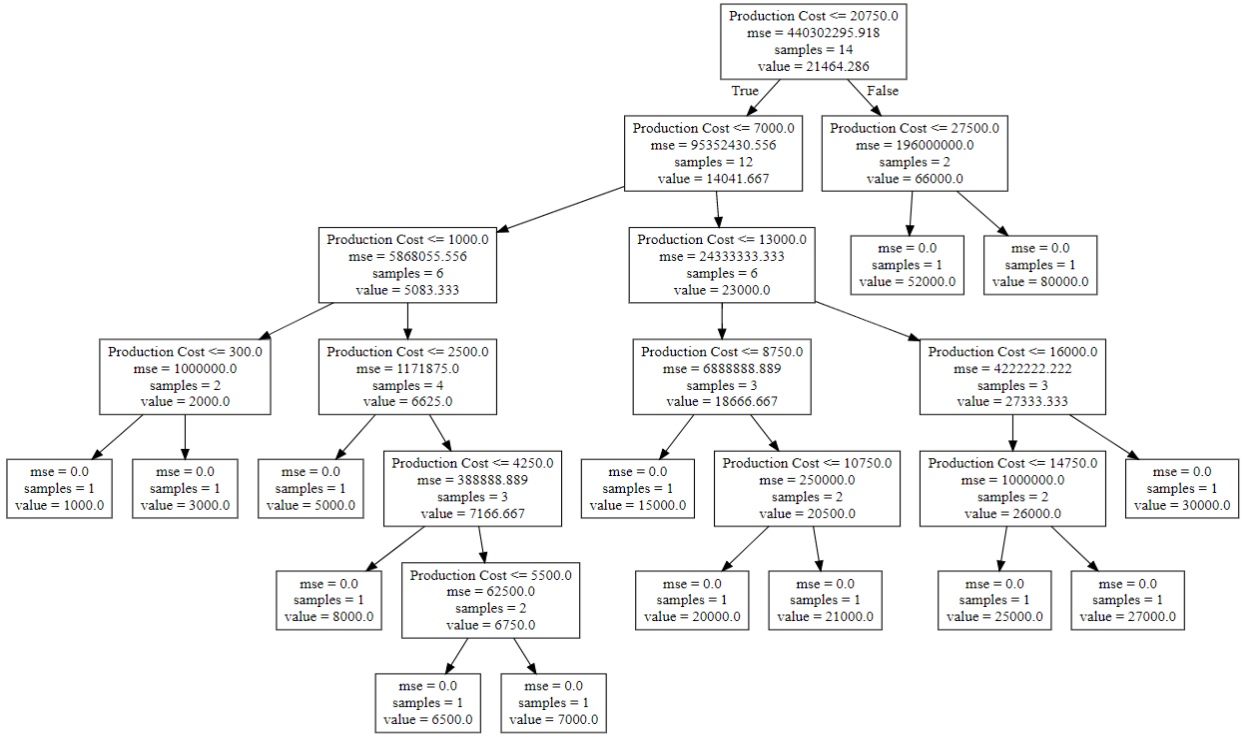


Decision Tree Regression

Problem with decision trees:

- Decision trees tend to overfit the training data.

To address this problem, an ***ensemble of randomly trained decision trees*** is often used.



Example from scikit-learn:

Face completion with linear and nonlinear regression models

Source: scikit-learn library