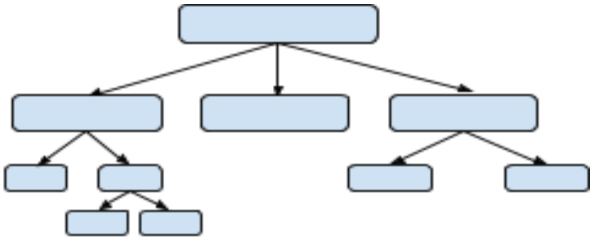
Vinoothna Bavireddy

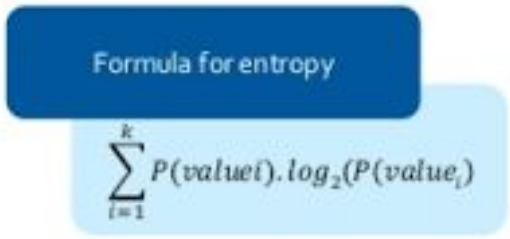
DECISION TREES



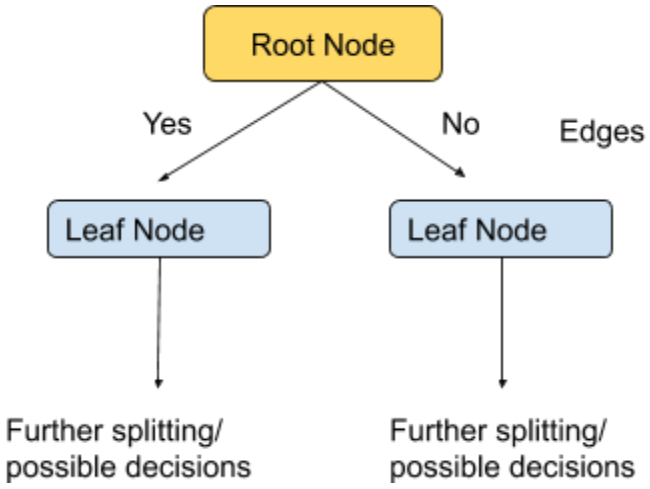
Decision Trees is a machine learning algorithm that works on a set of if-conditions. It is non-parametric and is trained on a training dataset fed with the correct outcomes for making predictions on the test data . Decision trees consist of nodes, edges and leaves where **nodes** are points at which the data is split, **edges** are the connections to the nodes (true or false to the if condition) and **leaves** are the outcomes at each node. The surface level (topmost) node of a decision tree is called the root node. Decision trees are capable of handling both categorical and numerical data.

Other important terms include:

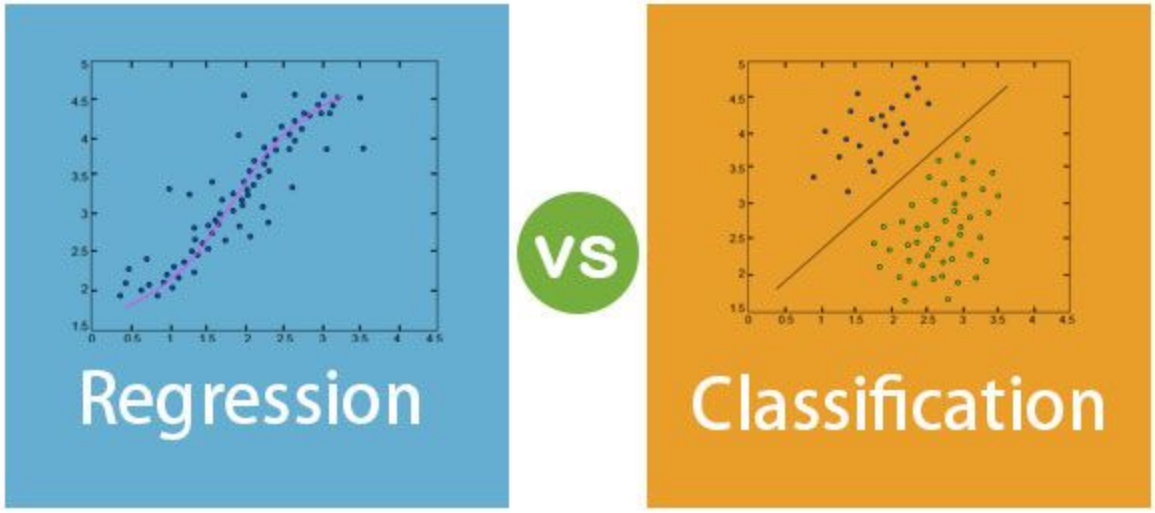
- ***Entropy:*** It is nothing but the measure of randomness or unpredictability in the dataset.



- ***Information Gain:*** It is defined as the decrease in entropy/ randomness after splitting the data.



Vinoothna Bavireddy



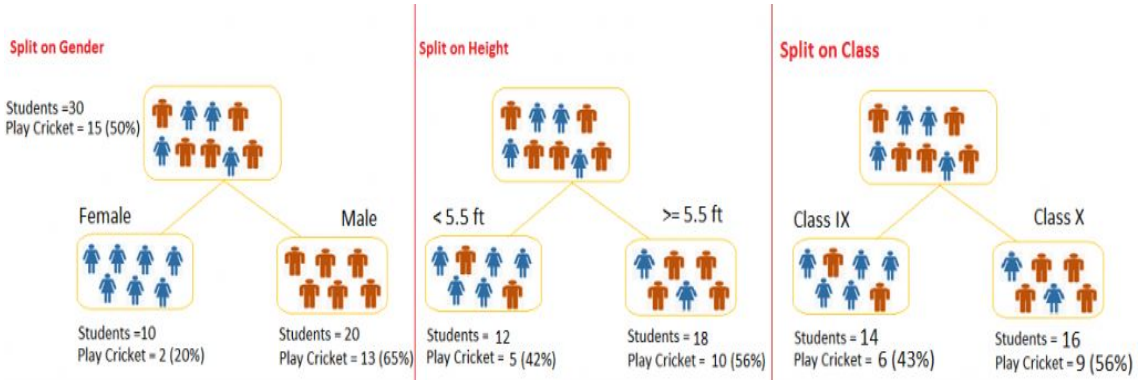
- Regression decision tree follows a regression rule where the target variable/ dependent variable is set and the splitting is based on the different Squared Errors calculated from the different independent variables.
- The dependent variables are continuous or numerical.

- The decision tree classifier works based on the set if-else conditions to classify or split data.
- Uses Unordered values (categorical)

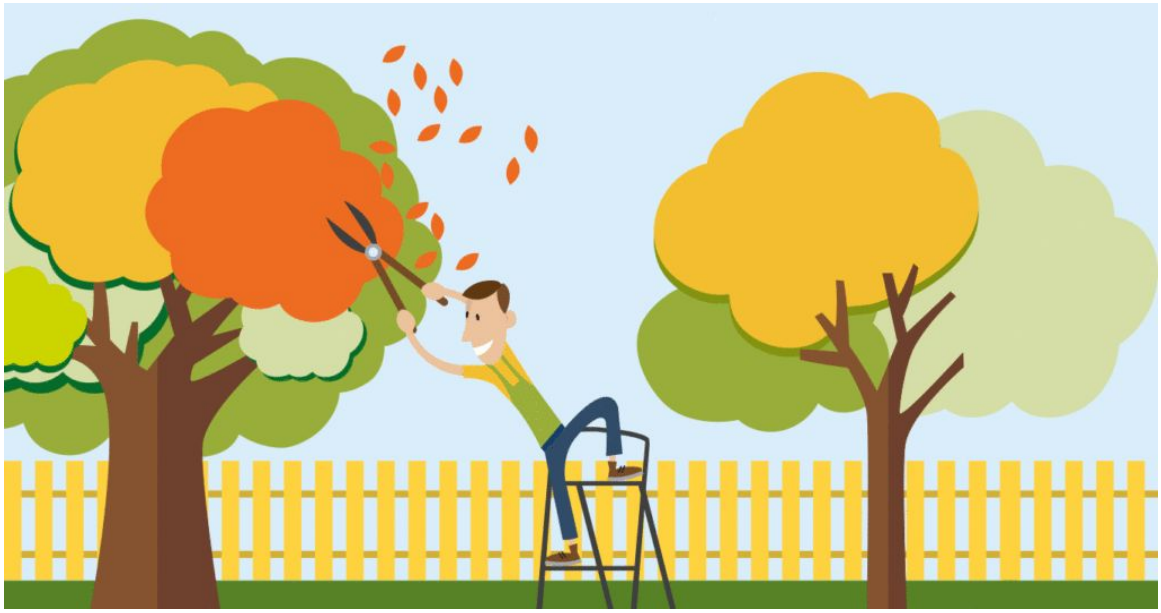
Working of Decision Trees

To build a decision tree you must follow the steps listed here:

1. ***Splitting:*** It is nothing but making subsets of the data based on the different variables. For example: Splitting students based on their gender, height or class.



2. ***Pruning:*** Decision trees are pruned meaning the number of branches are reduced by converting them into leaf nodes and reducing the number of leaf nodes from the original branches. This will help shortening the number of branches and classifying the data better to fit the training data. Simplifying the decision tree through pruning can help prevent overfitting of data but may result in poor classification of new values. To prune a decision tree one may join the adjacent nodes that have low information gain.



Advantages of Decision Trees:

1. Simple to understand,interpret and visualize
2. Data preparation is not intensive
3. Capable of handling both categorical and numerical data.
4. Performance is not affected by non-linear parameters.

Disadvantages of Decision Trees:

1. Overfitting is a concern when there is a lot of noise in the data.
2. High variance is an issue with decision trees (Even small variation can make the algorithm unstable).
3. Complicated decision trees have a low bias making it difficult to work with new data.

In Conclusion

Decision trees can be used to visualize data in the form of a flow chart for easy understanding. These are very useful in cases where the dataset is not very huge and has limited features to analyze. Steps involved in building decision trees are : Calculating the entropy, splitting the data based on the chosen features, decreasing the entropy (information gain), pruning the nodes for better performance.

PS: Decision trees tend to overfit and it is very important to choose features that bring value to the classification. In other words the steps of splitting and pruning are very important for the model's performance.

References:

1. Patel, S. (2017, May 14). Chapter 3 : Decision tree classifier — Theory. Medium. https://medium.com/machine-learning-101/chapter-3-decision-trees-theory-e7398adac567
2. Randerson. (2020, May 25). Python decision tree classifier example. Medium. https://medium.com/@randerson112358/python-decision-tree-classifier-example-d73bc3aeca6
3. Sehra, C. (2018, January 19). Decision trees explained easily. Medium. https://medium.com/@chiragsehra42/decision-trees-explained-easily-28f23241248
4. Borcan, M. (2020, March 21). Decision tree classifiers explained. Medium. https://medium.com/@borcandumitrumarius/decision-tree-classifiers-explained-e47a5b68477a