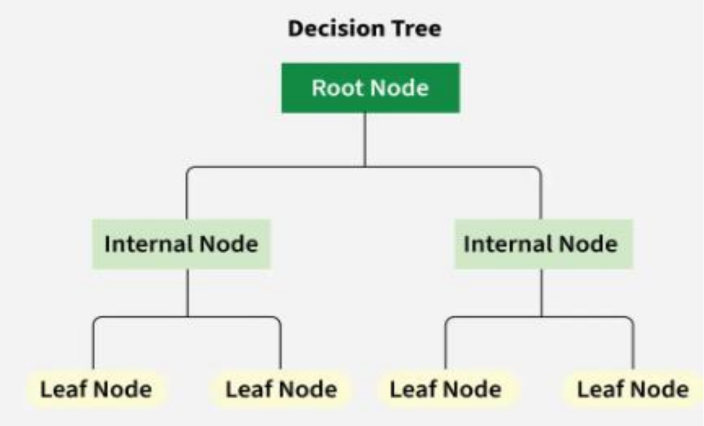
Decision Tree Algorithm in Data Analytics

→ A **decision tree** is a **supervised machine learning algorithm** used for **classification** and **regression** tasks.

→It works by breaking down a dataset into smaller subsets based on decision rules, forming a tree-like structure.

→Each internal node represents a decision based on an **attribute, branches indicate outcomes**, and **leaf nodes** represent **final decisions** or **predictions.**



Ex:

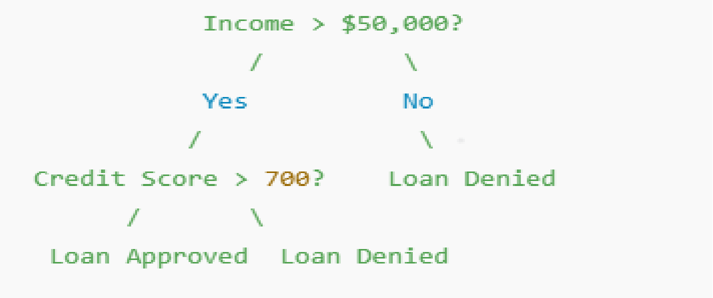
If a bank wants to approve loans based on customer profiles. A decision tree could use attributes such as **income, credit score, and debt-to-income ratio** to classify whether a loan should be **approved or denied**.

Income > $50,000?

Yes → **Credit Score > 700?**

- Yes → Loan Approved
- No → Loan Denied

- No → Loan Denied



```
                  Income > $50,000?
                       /         \
                    Yes           No
                   /               \
       Credit Score > 700?     Loan Denied
               /       \
       Loan Approved  Loan Denied
```

Advantages with Examples:

→Easy to Understand & Interpret :

Ex:

A hospital uses a decision tree to determine if a patient has the flu based on symptoms **like fever and cough**. Doctors can easily follow the tree structure to make quick decisions without needing complex calculations.

→Handles Both Numerical & Categorical Data :

Ex:

E-commerce company predicts whether a customer will buy a product based on **numerical data (age, income)** and **categorical data (gender, browsing behavior).**

→Useful for Feature Selection :

Ex:

A university wants to predict student dropout rates. A decision tree identifies that "attendance percentage" is the most important feature, helping the university focus on improving student engagement.

Disadvantages with Examples

→Prone to Overfitting :

Ex:

A stock market prediction model creates a highly detailed decision tree based on past trends. It works perfectly on historical data but fails to predict future market movements accurately.

→Unstable with Small Changes

Ex:

A decision tree for predicting employee performance changes drastically when a few records in the dataset are modified, making the model unreliable for long-term use.

→Biased if Data is Imbalanced

Ex:

A loan approval model trained mostly on high-income customers may predict that almost all applicants should get a loan, ignoring the potential of low-income but creditworthy applicants.

Less Effective for Large Datasets:

Ex:

A social media platform analyzing millions of user interactions finds that a single decision tree is too slow and inaccurate, so they switch to Random Forest for better performance.