Application of Big Data in Social Science

week 13

Heidi Hyeseung Choi
Fall 2022
HYSIS
heidichoi@hanyang.ac.kr

Announcements

~~9.1~~	~~1~~	~~Introduction~~
~~9.8~~	~~2~~	~~Web Scraping~~
~~9.15~~	~~3~~	
~~9.22~~	~~4~~	~~Natural Language Processing~~
~~9.29~~	~~5~~	
~~10.6~~	~~6~~	~~Text Analysis~~ ~~(recorded lecture on week 7)~~
~~10.13~~	~~7~~	
~~10.20~~	~~8~~	~~Mid term exam (as school schedule)~~
~~10.27~~	~~9~~	~~Midterm review & word cloud~~
~~11.3~~	~~10~~	~~Social Network Analysis~~
~~11.10~~	~~11~~	~~COVID-19 T-T~~
~~11.17~~	~~12~~	~~Machine Learning: Supervised Learning~~
11.24	**13**	**Supervised Learning**
12.1	14	Machine Learning: Unsupervised Learning
12.8	15	Data Visualization
12.15	16	Final Exam

Final Exam:

- Will be taken as exam.
  (no project for the finals)
- We still have hwk assignment 2.
- I will let you know about it next week.

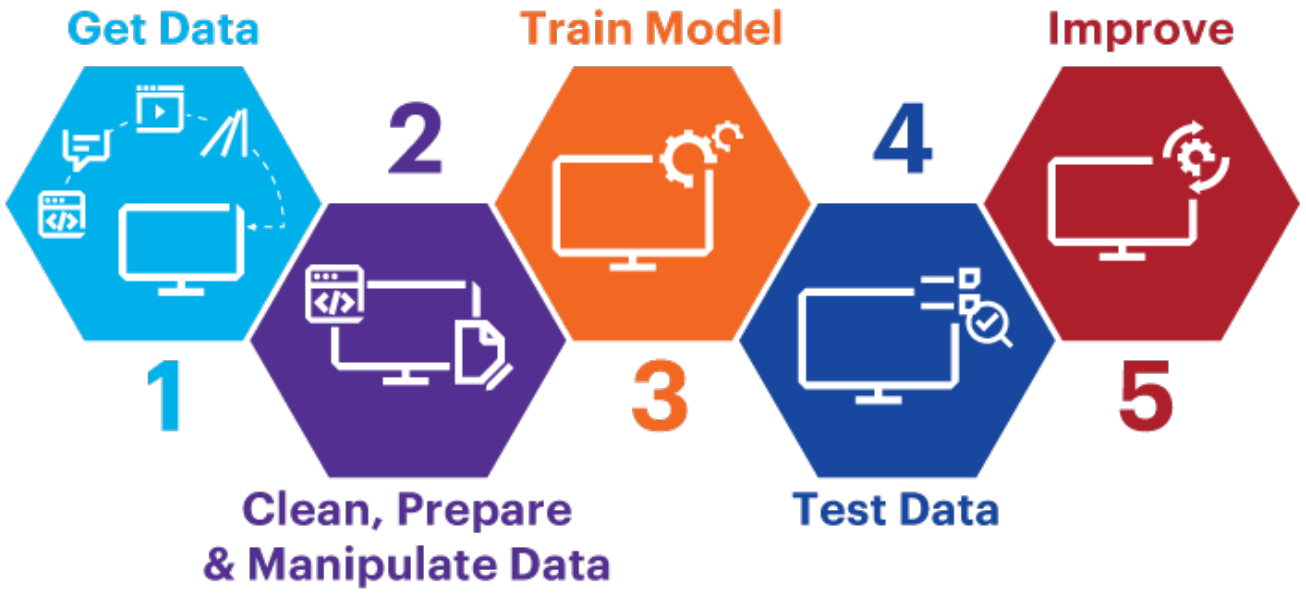
Pros and cons of k-Nearest Neighbors algorithm

- No assumptions about data — useful for nonlinear data.
- Simple algorithm — to explain and understand/interpret.
- High accuracy (relatively) — it is pretty high but not competitive in comparison to better supervised learning models.

- Does not work well with large dataset.
- Does not work well with high dimensions.
- Need feature scaling (standardization and normalization) before applying KNN algorithm to any dataset.
- Sensitive to noisy data, missing values and outliers.

Remember, training and testing data

- Scikit-learn: machine learning library.
  - classification, regression, clustering etc
  - easily used with pandas and numpy



3. Train Model:

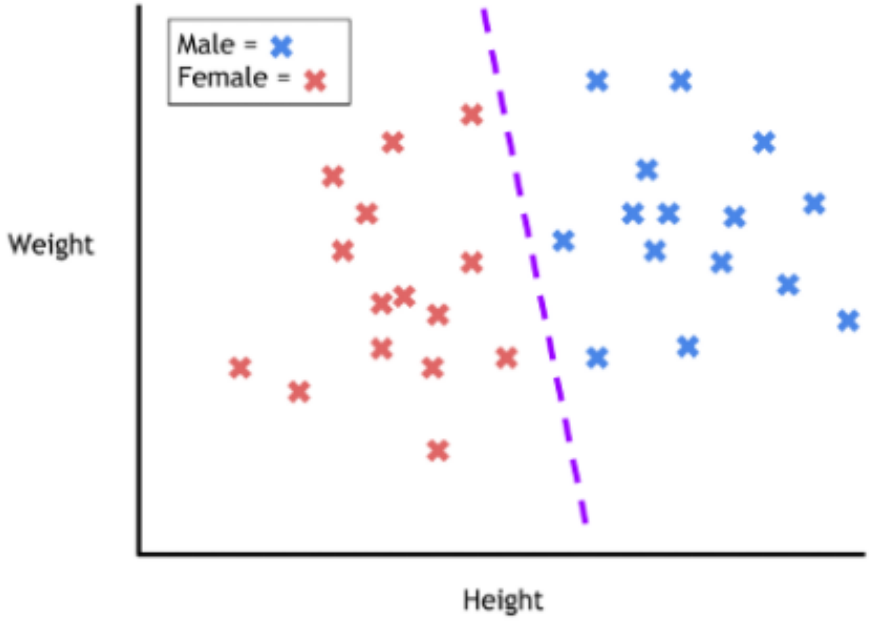
- Divide the dataset into a training set and a test set.
- we split dataset by using scikit-learn function
  train_test_split()

4. Test Data:

– Using the test set, we test our data and see the accuracy between actual and predicted category

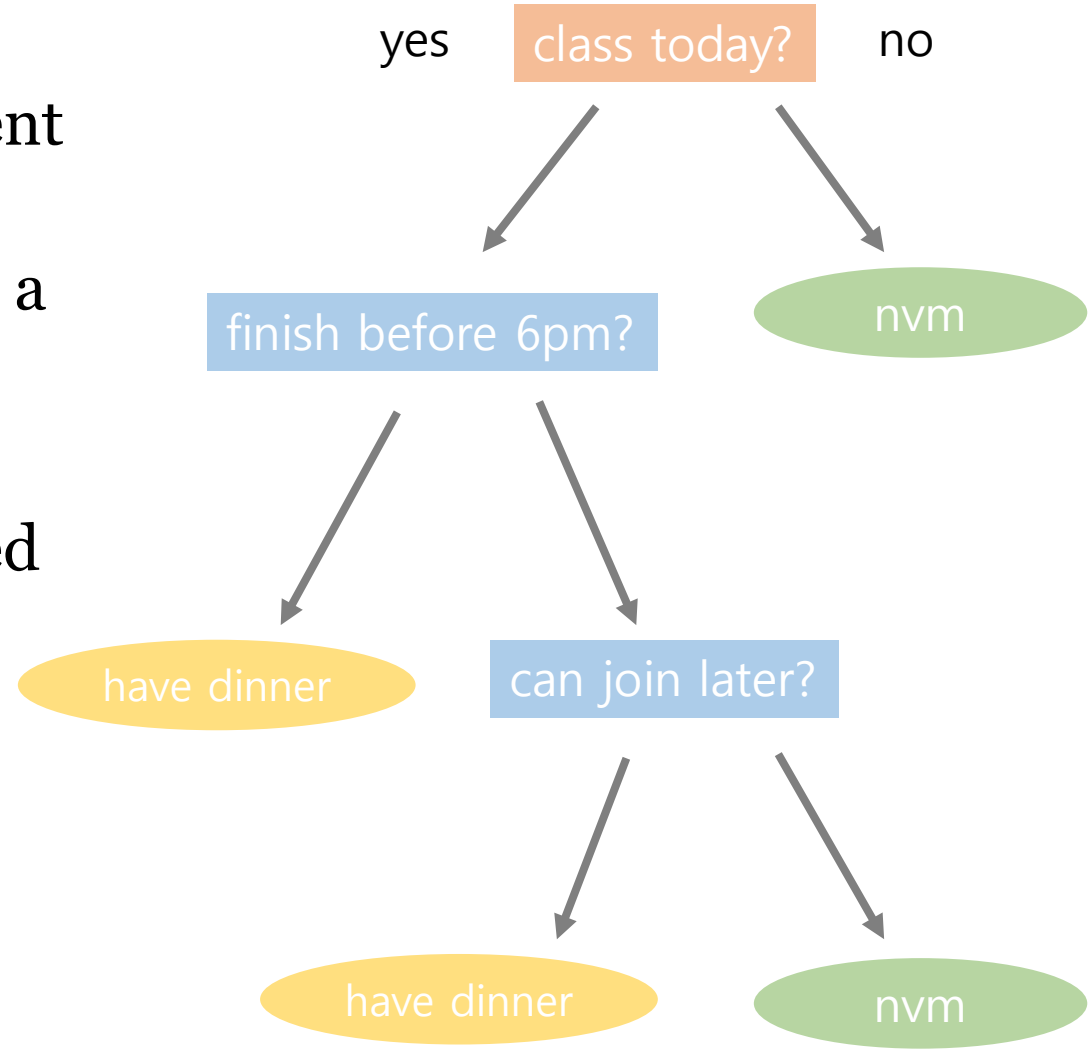
Supervised Learning

- Classification: the output variable is a category instead of values,
  - example 1: “red” or “blue”
  - example 2: “yes” or “no”.
- Take an input value and assign it a class (category) that it fits into based on the training data provided.
- eg) spam email
- Main algorithms:
  - Linear Classifiers
  - K-Nearest Neighbor
  - Decision Trees
  - Random Forest
  - Support Vector Machines



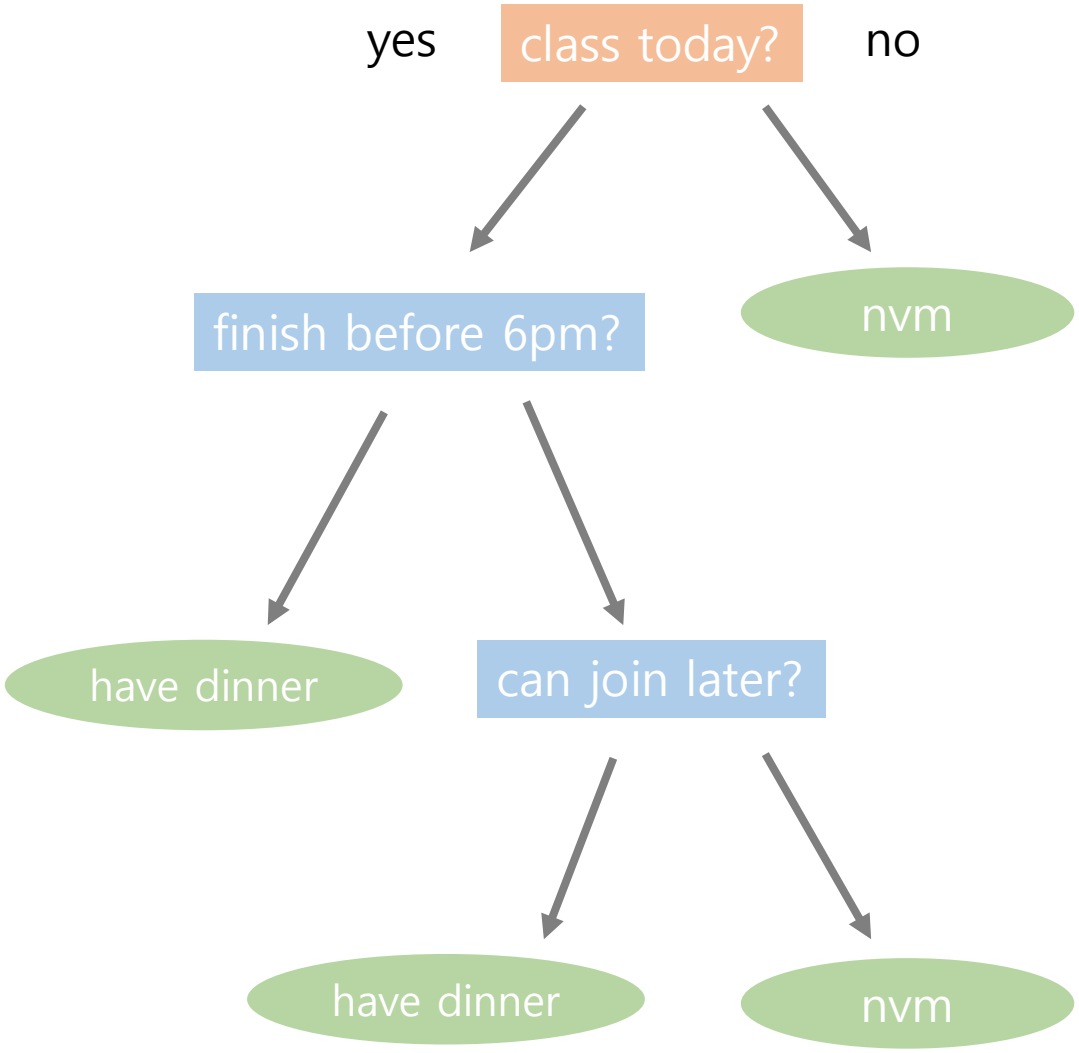
Decision Tree algorithm

- Another widely used algorithm in ML.
- It can be used to visually and explicitly represent decision making.
- Decision tree algorithm identifies ways to split a dataset based on different conditions.
- It conducts ‘feature-based’ splits.
- Decision Trees are a non-parametric supervised learning method used for classification and regression tasks.



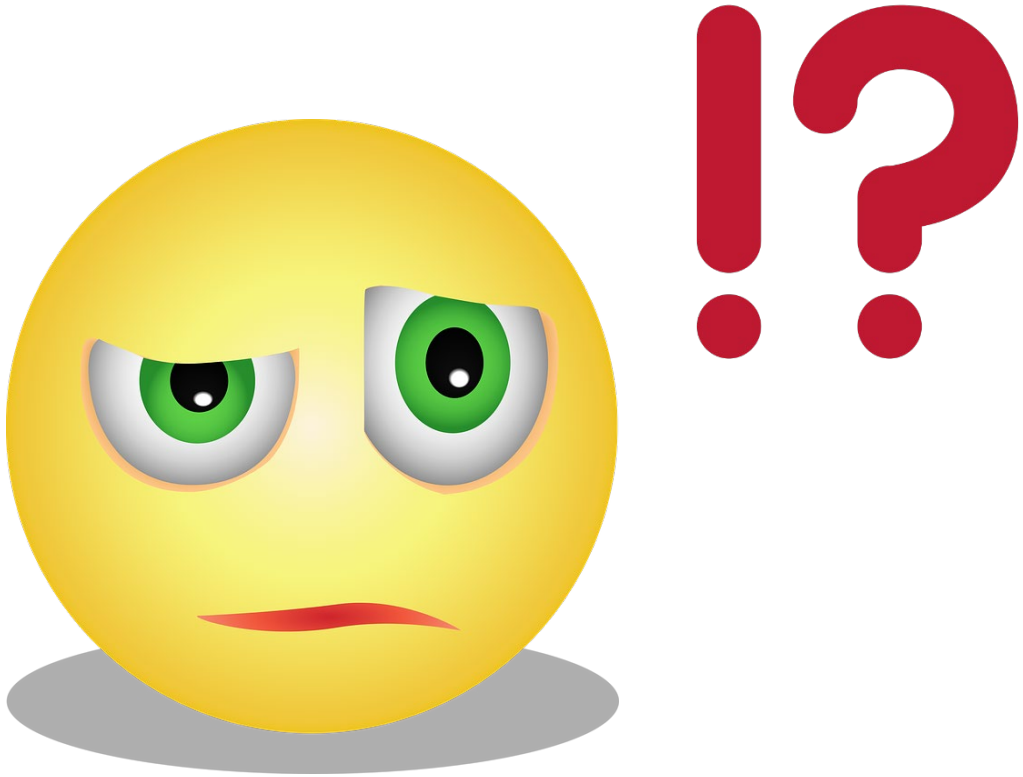
Decision Tree algorithm

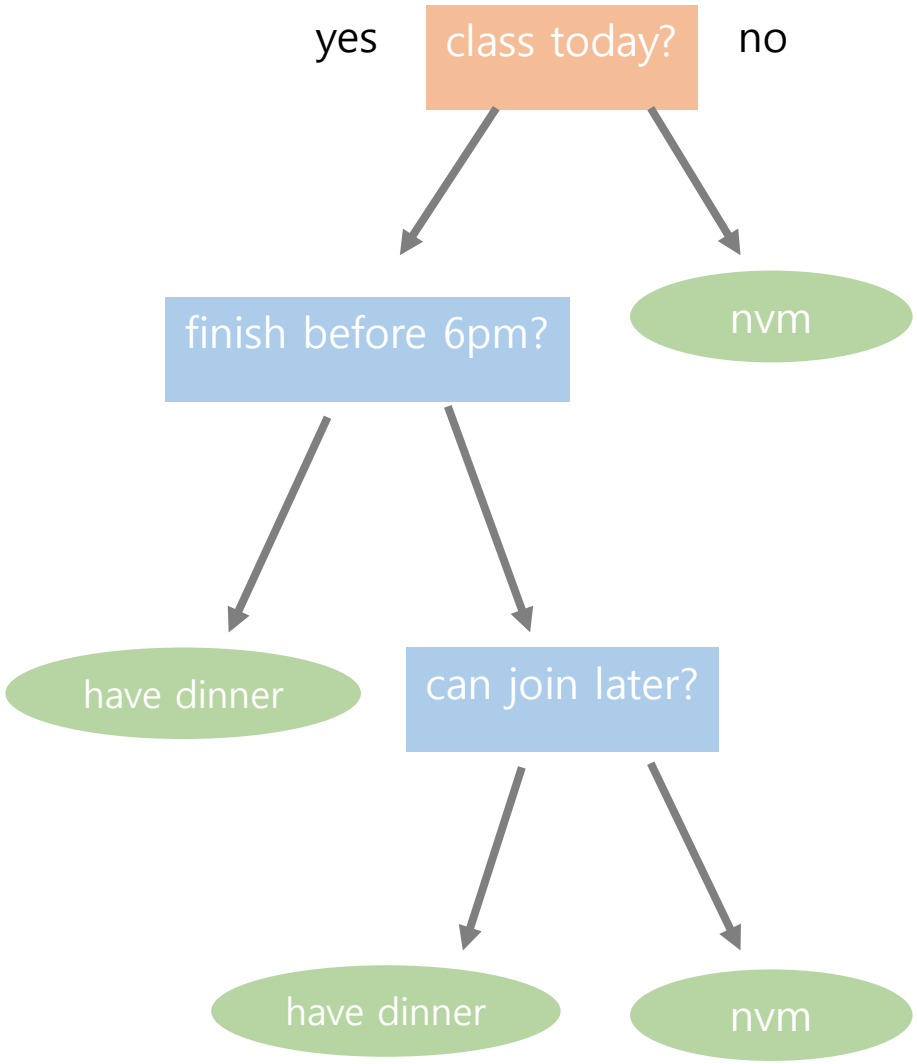
- **Root Nodes:** the node present at the beginning of a decision tree from this node the population starts dividing according to various features.
- **Decision Nodes:** the nodes we get after splitting the root nodes are called Decision Node.
- **Leaf Nodes:** the nodes where further splitting is not possible are called leaf nodes or terminal nodes.



Decision Tree algorithm

- What becomes the Root node?
- What becomes the decision node?
- When to stop splitting?





Level of Importance: 'Pure'ness

Entropy

- Entropy: the degree of disorder or uncertainty in a system.
  - i.e.: how messy(uncertain) our dataset is.
    - we do not know how to split into branches.
  - If we split our data, and it becomes 100:0, then it is pretty much ‘certain’.
  - But if we have 50:50 split, we are not certain about the final output.
  - In this case, we have some entropy in that node.
  - The entropy(uncertainty) should be minimized.
  - To minimize uncertainty, we need to choose the feature with most importance, which would make the tree to be as simple as possible.
  - The higher the importance of a feature, the more priority it would get for splitting.

Attribute Selection Measure (ASM)

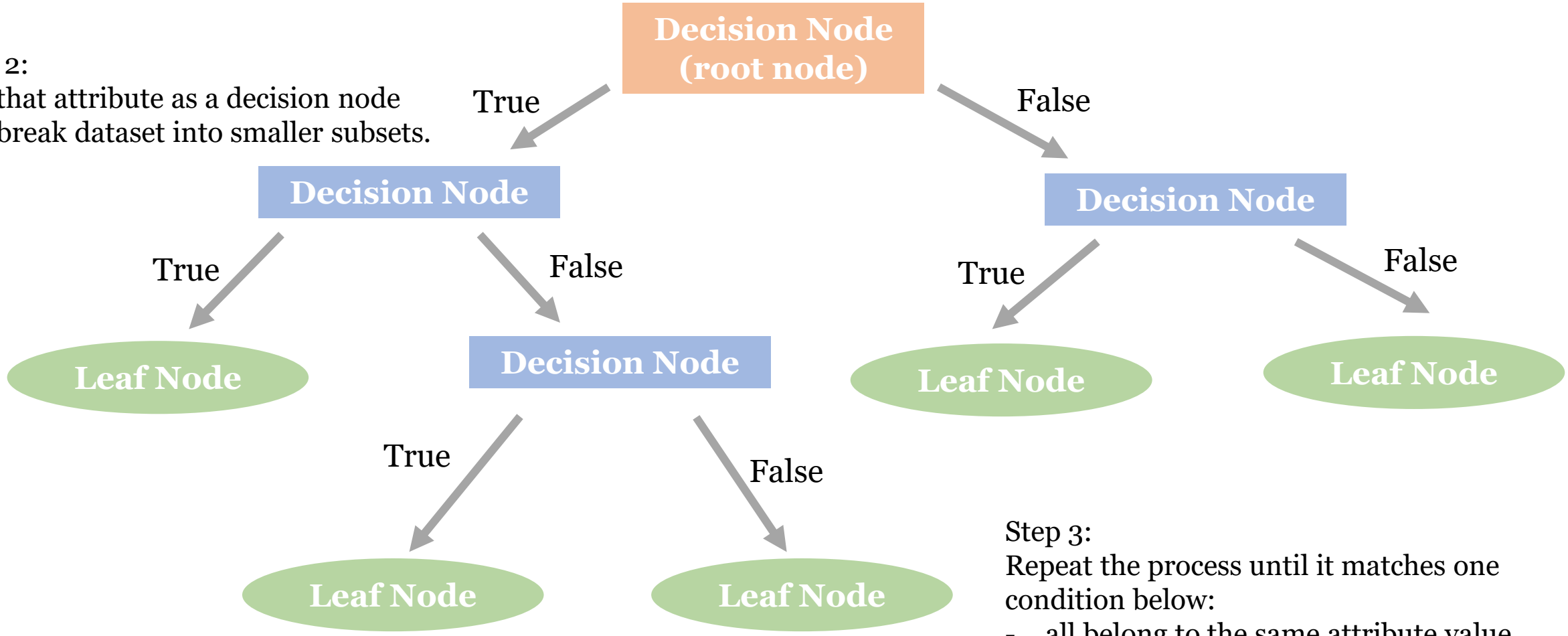
- ASM is the process by which we select the best feature to start splitting our tree into branches:
  1) Information Gain (Entropy)
     - the information value measures how much information a feature gives us about the class.
     - The split with the highest information gain will be taken as the first split and the process will continue until all children nodes are pure, or until the information gain is 0.

  2) Gini Index
- Gini impurity or Gini Index is another parameter on the basis of which splitting of decision tree is done. It is actually a measure of impurity while creating a decision tree in the CART algorithm.
- Pure: in a selected sample of dataset all data belongs to same class (gini=0)
- Impure: data is mixture of different classes.

Decision Tree

Step 1:
Select the best attribute using Attribute Selection Measures (ASM) to split.

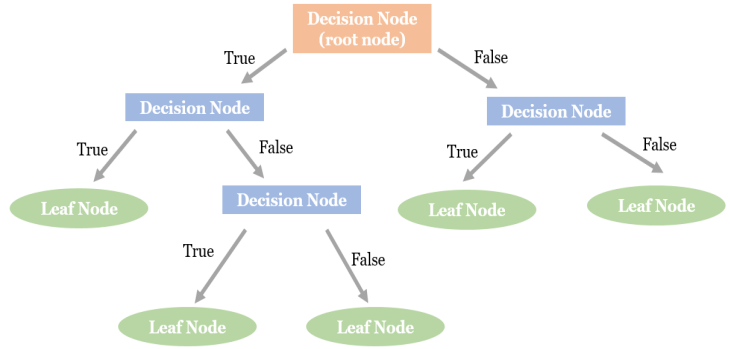
Step 2:
Use that attribute as a decision node and break dataset into smaller subsets.



Step 3:
Repeat the process until it matches one condition below:

- all belong to the same attribute value
- there are no more remaining attribute
- there are no more instances (features)

Decision Tree: how we will do it



Decision Tree Generator

Select best attribute using ASM: Information Gain, Gini Index or Gini Ratio

Break dataset into subsets

Repeat the process

Data

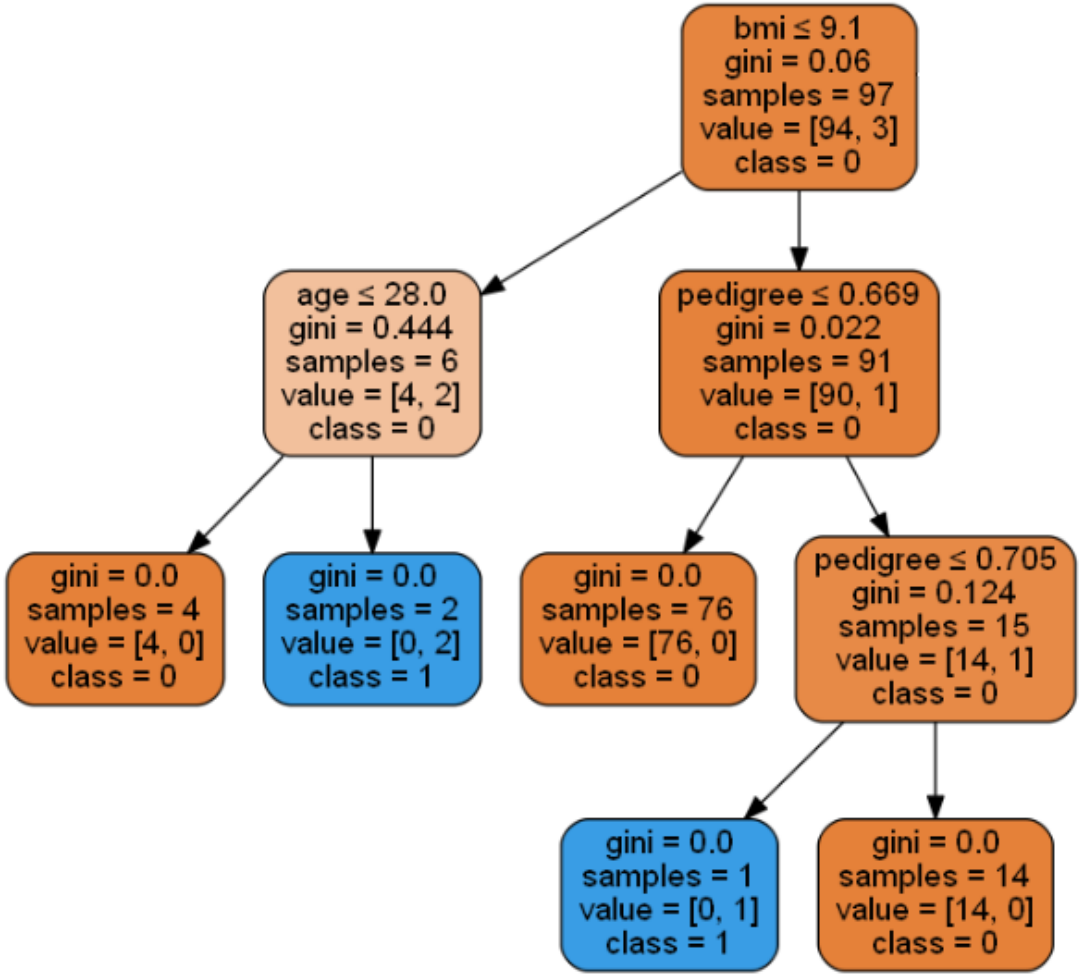
Training Data

Test Data

Model evaluation

Gini Index (aka Gini impurity)

- Each node has a decision rule that will split the data.
- “gini” refers to Gini ratio, and it measures the impurity of the node.
- It is considered ‘pure’ when all of its nodes belong to the same class.
- As it moves down the tree, it will lead towards the direction where it decreases the level of impurity (gini) to yield better classification.
- calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly. If all the elements are linked with a single class then it can be called pure.



Decision Tree

Pros

- Decision trees are easy to interpret and visualize.
- Can handle both categorical and numerical data
- Decision trees can easily capture non-linear patterns.
- It has no assumption on the distribution, due to its non-parametric nature of the algorithm.

Cons

- Sensitive to noisy data. It can overfit noisy data.
- The small variation(or variance) in data can result in the different decision tree.
- Need to be careful with parameter tuning.
- Can create biased learned trees if some classes dominate.

Random Forests

- To make results more robust, we create decision trees on randomly selected data samples.
- Then, we get predictions from each tree and select the best solution by means of ‘voting’.

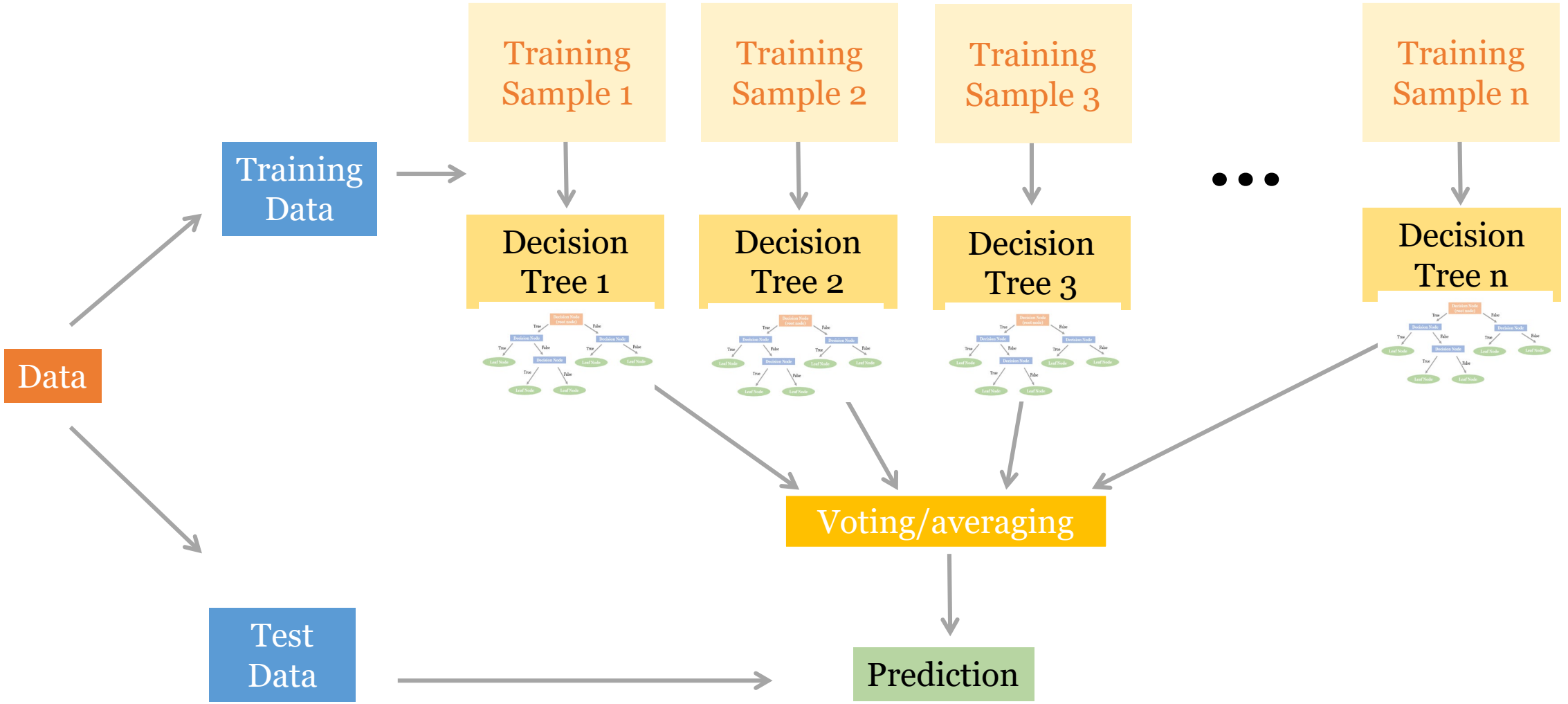
Decision Tree Classifier

eg) renting a car to go to Jeju

1. ask friends for recommendations
2. vote for the best rental car service

Random Forest Classifier

Random Forests



Random Forests

Pros

- It is considered to be highly accurate and robust method.
- It does not suffer from overfitting problem. It takes the average of the predictions, cancelling out the biases.

Cons

- It is slow in generating predictions for large data.
- It is more difficult to visually interpret as compared to decision trees.

Decision Tree or Random Forest?

Decision Tree	Random Forest
It is a tree-like decision-making diagram.	It is a group of decision trees combined together to give output.
Possibility of Overfitting.	Prevents Overfitting.
Gives less accurate result.	Gives accurate results.
Simple and easy to interpret.	Hard to interpret.
Less Computation	More Computation
Simple to visualize.	Complex Visualization.
Fast to process.	Slow to process.

Feature Importance

- Feature Importance refers to techniques that calculate a score for all the input features for a given model — the scores simply represent the “importance” of each feature.
- A higher score means that the specific feature will have a larger effect on the model that is being used to predict a certain variable.
- The scores are calculated on the weighted Gini indices
- Why Feature Importance?
  1) Data Understanding: understand what features are irrelevant for the model
  2) Model Improvement: to reduce the dimensionality of the model. This would make the model to be simpler.
  3) Model Interpretability: to interpret and communicate your model to other stakeholders.

Datasets

- Iris Dataset

Iris setosa

Iris virginica

Iris versicolor

- Diabetes Dataset

Libraries to install

>>> pip install pydotplus
>>> pip install dtreeviz
>>> pip install graphviz

And then, set PATH for graphviz
https://graphviz.org/download/
os.environ["PATH"] += os.pathsep + r'C:\Users\~~~~ your path‘

On mac
>>> sudo port install graphviz
>>> brew install graphviz

Graphviz Install

Or on conda prompt,
>>> conda install python-graphviz

Let’s launch our Jupyter notebook