APSSDC

Andhra Pradesh State Skill Development Corporation

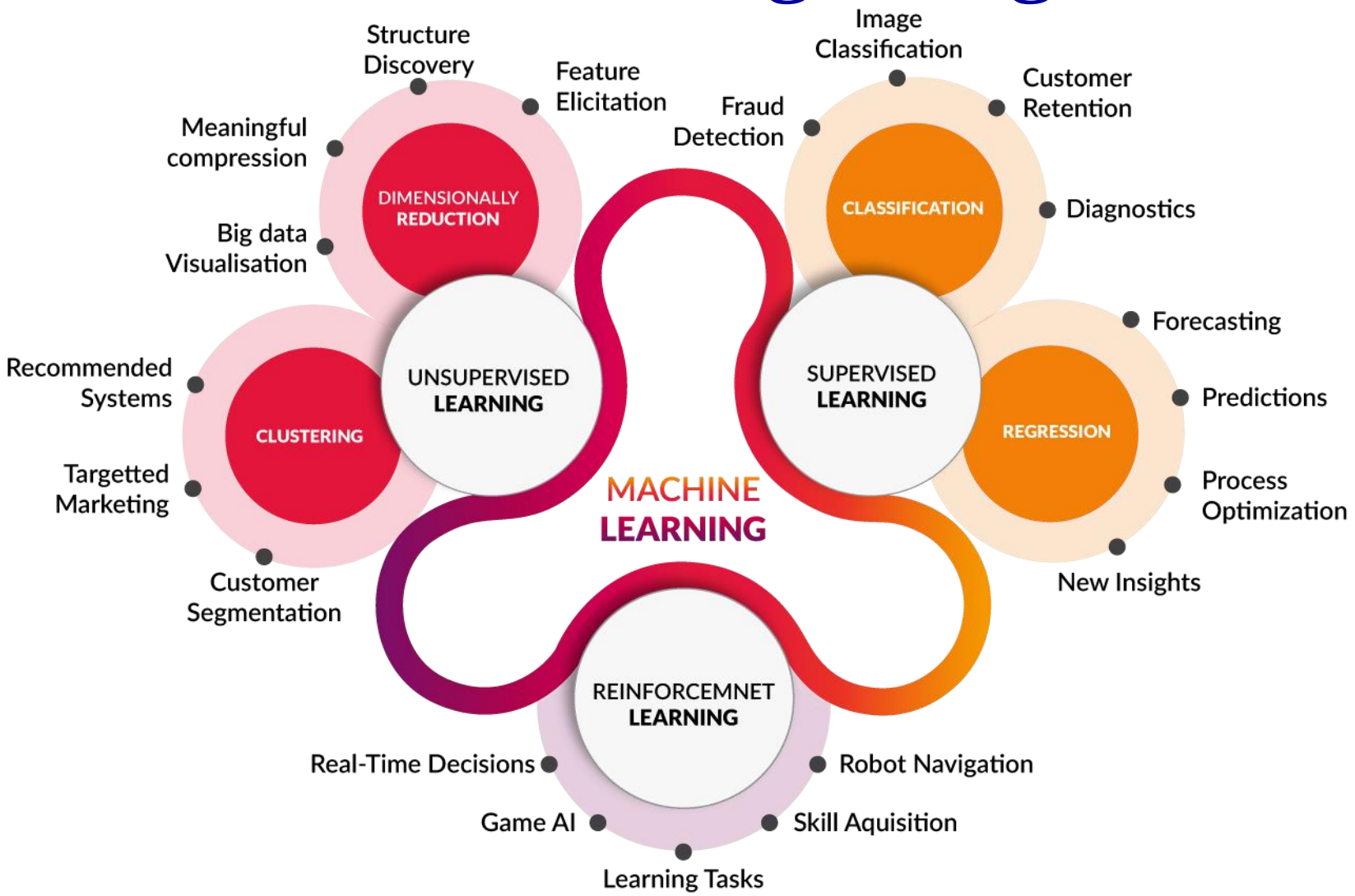
Random Forest Algorithm in

Machine Learning Using Sklearn

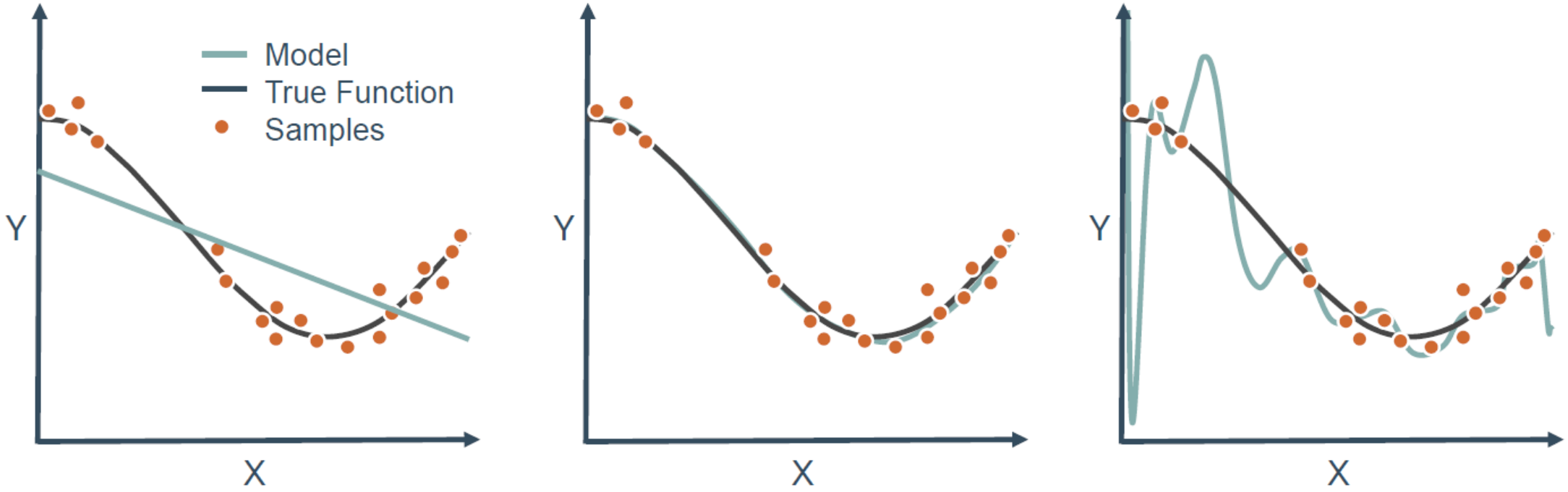
Day07 Agenda

- Decision Tree Regressor
- Random Forest Algorithm

Machine Learning Categories



Under Fitting Vs Best Fit vs Over Fitting



Over Fitting vs Under Fitting

- Under Fitting

- Over Fitting

Bias-Variance Trade-Off

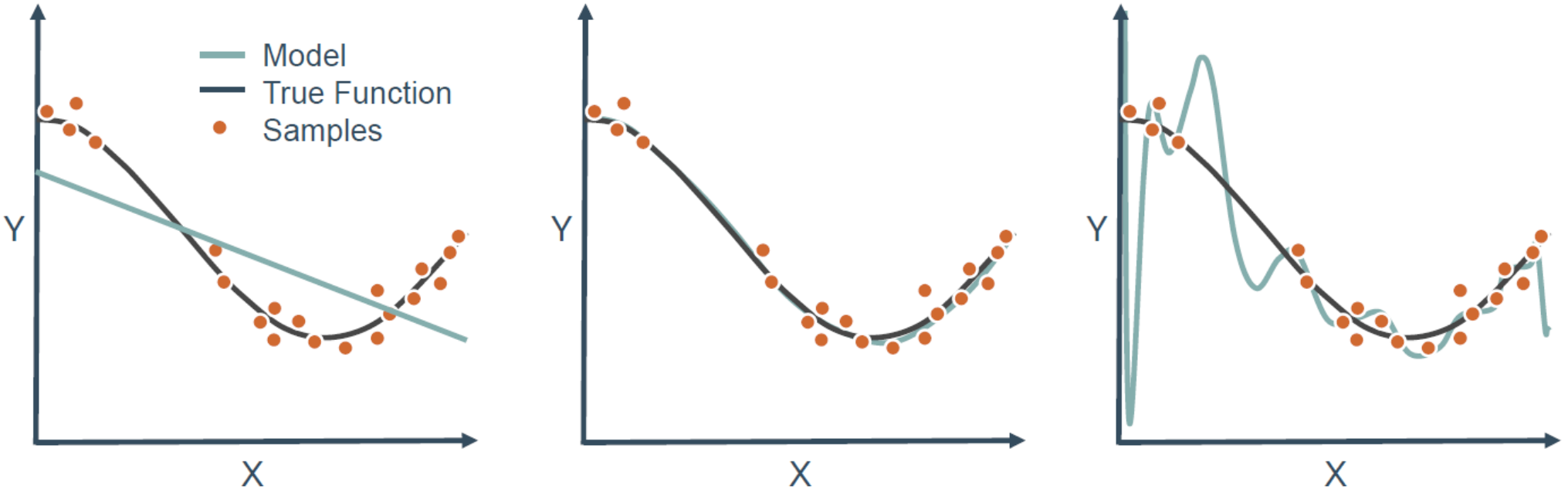
Bias are the simplifying assumptions made by a model to make the target function easier to learn.

- **Low Bias**: Suggests less assumptions about the form of the target function.
- **High-Bias**: Suggests more assumptions about the form of the target function.

Variance is the amount that the estimate of the target function will change if different training data was used.

- **Low Variance**: Suggests small changes to the estimate of the target function with changes to the training dataset.
- **High Variance**: Suggests large changes to the estimate of the target function with changes to the training dataset.

Bias-Variance Trade-Off



High Bias
Low Variance

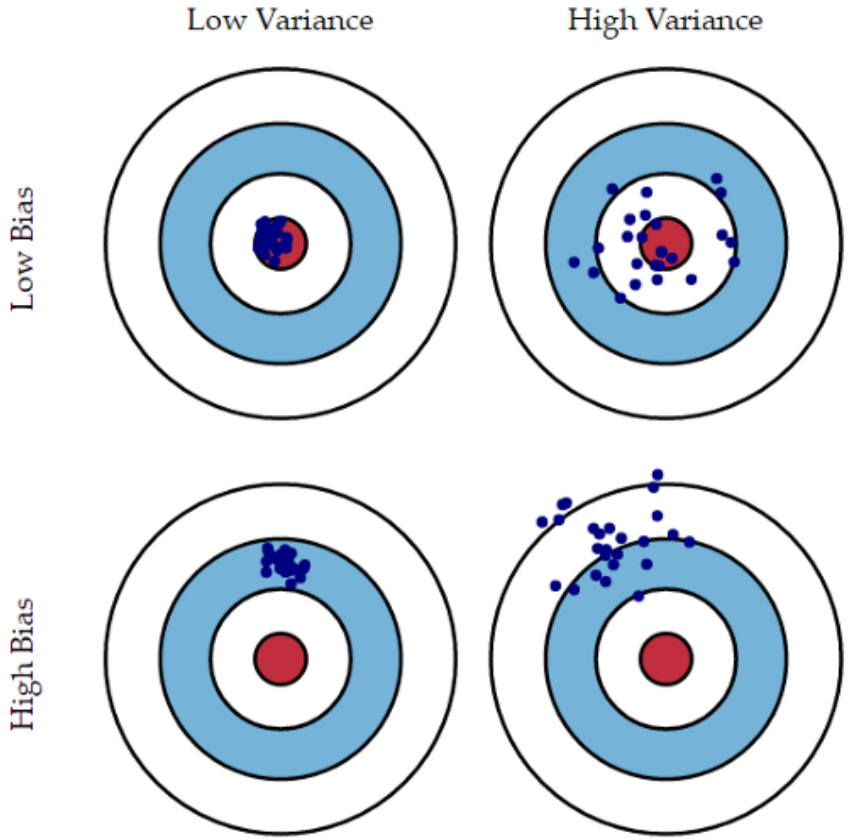
Just Right

Low Bias
High Variance

What is bias and variance?

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.

Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. As a result, such models perform very well on training data but has high error rates on test data.



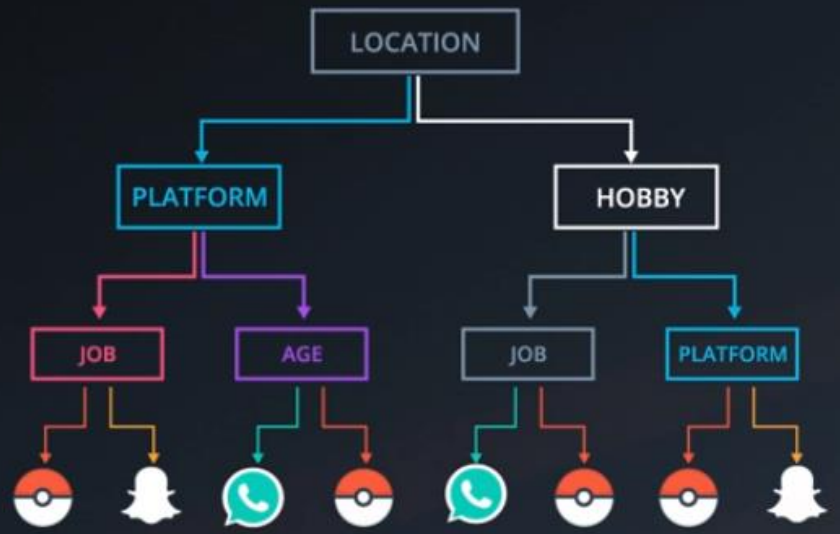
underfitting

In supervised learning, **underfitting** happens when a model unable to capture the underlying pattern of the data. These models usually have high bias and low variance. It happens when we have very less amount of data to build an accurate model or when we try to build a linear model with a nonlinear data.

Overfitting Problems in Decision Trees

Large Tables

Gender	Age	Location	Platform	Job	Hobby	App
F	15	US	iOS	School	Videogames	
F	25	France	Android	Work	Tennis	
M	32	Chile	iOS	Temp	Tennis	
F	40	China	iOS	Retired	Chess	
M	12	US	Android	School	Tennis	
M	14	Australia	Android	School	Videogames	



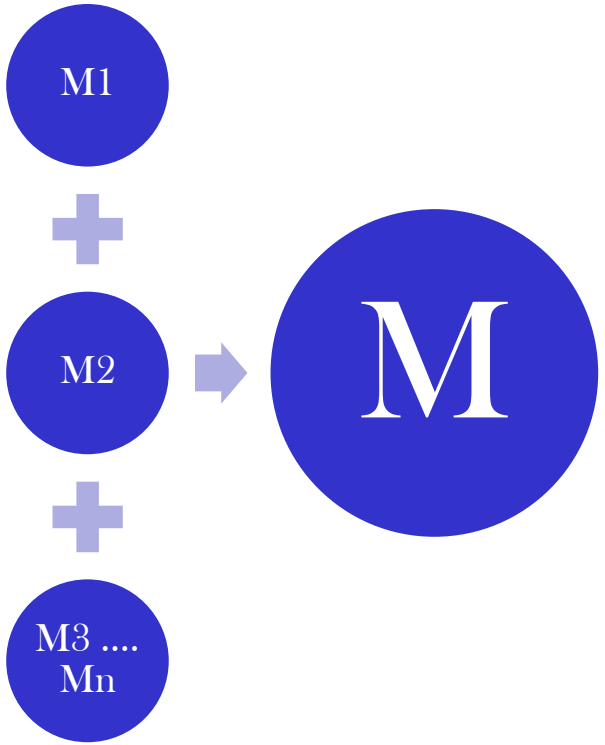
If client is male, between 15 and 25, in the US, on Android, in school, likes tennis, pizza, but does not like long walks on the beach, then they are likely to download Pokemon Go.

Ensemble

What are ensembles?

- When multiple models used together to make more powerful model. Combining the models.
- Actual Meaning of Ensemble is "Group of Musicians" in ML it is referred as combination of Multiple Models

1. Bagging
2. Boosting
3. Stacking



Random Forrest

Ensemble methods

- A single decision tree does not perform well
- But, it is super fast
- What if we learn multiple trees?

We need to make sure they do not all just learn the same

Bagging

If we split the data in random different ways, decision trees give different results, **high variance.**

Bagging: Bootstrap **agg**regat**ing** is a method that result in low variance.

If we had multiple realizations of the data (or multiple samples) we could calculate the predictions multiple times and take the average of the fact that averaging multiple onerous estimations produce less uncertain results

Bagging

- Reduces overfitting (variance)
- Normally uses one type of classifier
- Decision trees are popular
- Easy to parallelize

Random Forest

Random forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees.

- The term came from random decision forests that was first proposed by Tin Kam Ho of Bell Labs in 1995.
- The method combines Breiman's "bagging" idea and the random selection of features.

Features and Advantages

The advantages of random forest are:

- It is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier.
- It runs efficiently on large databases.
- It can handle thousands of input variables without variable deletion.
- It gives estimates of what variables are important in the classification.
- It generates an internal unbiased estimate of the generalization error as the forest building progresses.