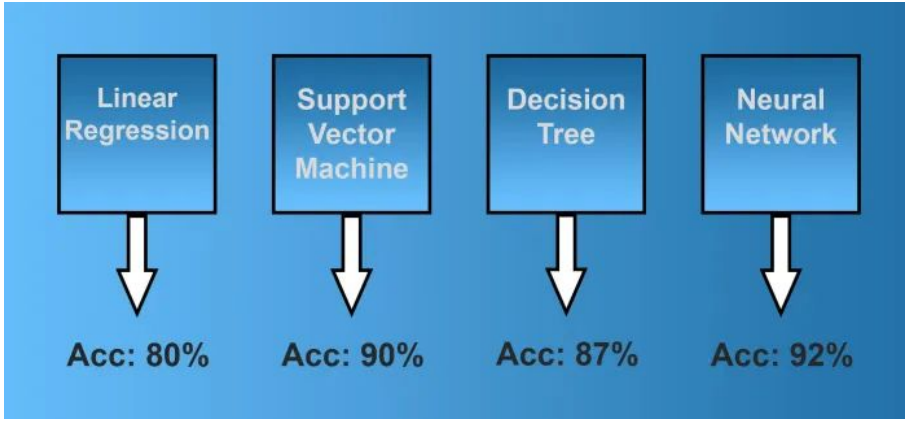
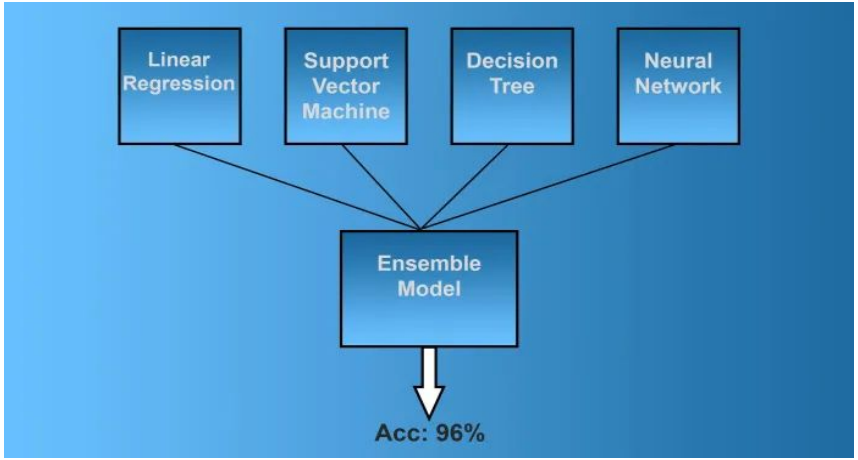
MODULE III

- **The principle of "the wisdom of the crowd".**
- In machine learning, crowd wisdom is achieved through ensemble learning. For many problems, the result obtained from an ensemble, a combination of machine learning models, can be more accurate than any single member of the group.
- Example : You want to develop a machine learning model that predicts inventory stock orders for your company based on historical data you have gathered from previous years.
- You use train four machine learning models using a different algorithms.



- These machine learning models are called **“weak learners”** because they fail to converge to the desired level.
- But weak doesn’t mean useless. You can combine them into an ensemble.
- For each new prediction, you run your input data through all four models, and then compute the average of the results.
- The reason ensemble learning is efficient is that your machine learning models work differently. Each model might perform well on some data and less accurately on others. When you combine all them, they cancel out each other’s weaknesses.



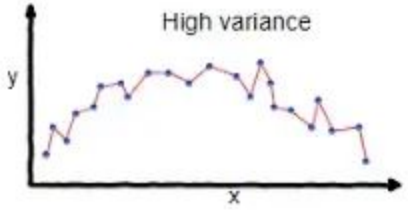
What is bias?

- Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.
- Error of the training data.

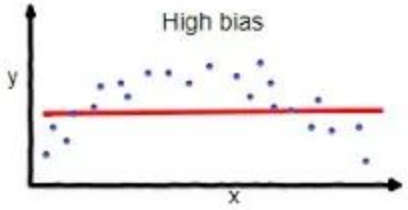
What is variance?

- Variance is the variability of model prediction for a given data point or a value which tells us spread of our data.
- Error of the testing data

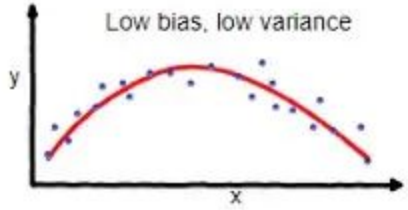
- In supervised learning, **underfitting** happens when a model unable to capture the underlying pattern of the data.
- These models usually have high bias and low variance. It happens when we have very less amount of data to build an accurate model or when we try to build a linear model with a nonlinear data.
- Also, these kind of models are very simple to capture the complex patterns in data like Linear and logistic regression.
- In supervised learning, **overfitting** happens when our model captures the noise along with the underlying pattern in data.
- It happens when we train our model a lot over noisy dataset. These models have low bias and high variance. These models are very complex like Decision trees which are prone to overfitting.



overfitting



underfitting



Good balance

Why is Bias Variance Tradeoff?

- If our model is too simple and has very few parameters then it may have high bias and low variance.
- On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data.
- This tradeoff in complexity is why there is a tradeoff between bias and variance.
- An algorithm can't be more complex and less complex at the same time.

Building an Ensemble System

Three strategies need to be chosen for building an effective ensemble system.

We have previously referred to these as the three pillars of ensemble systems:

(1) data sampling/selection;

(2) training member classifiers; and

(3) combining classifiers.

1. **Data Sampling and Selection: Diversity**

Making different errors on any given sample is of paramount importance in ensemble-based systems. After all, if all ensemble members provide the same output, there is nothing to be gained from their combination. Therefore, we need diversity in the decisions of ensemble members, particularly when they are making an error.

Diversity in ensembles can be achieved through several strategies, although using different subsets of the training data is the most common approach.

2. **Training Member Classifiers**

At the core of any ensemble-based system is the strategy used to train individual ensemble members.

3. **Combining Ensemble Members**

The last step in any ensemble-based system is the mechanism used to combine the individual classifiers.

Simple Ensemble Techniques

In this section, we will look at a few simple but powerful techniques, namely:

1. Max Voting
2. Averaging
3. Weighted Averaging

Max Voting

The max voting method is generally used for classification problems. In this technique, multiple models are used to make predictions for each data point. The predictions by each model are considered as a ‘vote’. The predictions which we get from the majority of the models are used as the final prediction.

For example, when you asked 5 of your colleagues to rate your movie (out of 5); we’ll assume three of them rated it as 4 while two of them gave it a 5. Since the majority gave a rating of 4, the final rating will be taken as 4. You can consider this as taking the mode of all the predictions.

The result of max voting would be something like this:

Colleague 1	Colleague 2	Colleague 3	Colleague 4	Colleague 5	Final rating
5	4	5	4	4	4

Averaging

Similar to the max voting technique, multiple predictions are made for each data point in averaging. In this method, we take an average of predictions from all the models and use it to make the final prediction. Averaging can be used for making predictions in regression problems or while calculating probabilities for classification problems.

For example, in the below case, the averaging method would take the average of all the values.

i.e. (5+4+5+4+4)/5 = 4.4

Colleague 1	Colleague 2	Colleague 3	Colleague 4	Colleague 5	Final rating
5	4	5	4	4	4.4

Weighted Average

This is an extension of the averaging method. All models are assigned different weights defining the importance of each model for prediction. For instance, if two of your colleagues are critics, while others have no prior experience in this field, then the answers by these two friends are given more importance as compared to the other people.

The result is calculated as [(5*0.23) + (4*0.23) + (5*0.18) + (4*0.18) + (4*0.18)] = 4.41.

	Colleague 1	Colleague 2	Colleague 3	Colleague 4	Colleague 5	Final rating
weight	0.23	0.23	0.18	0.18	0.18	4.41
rating	5	4	5	4	4	

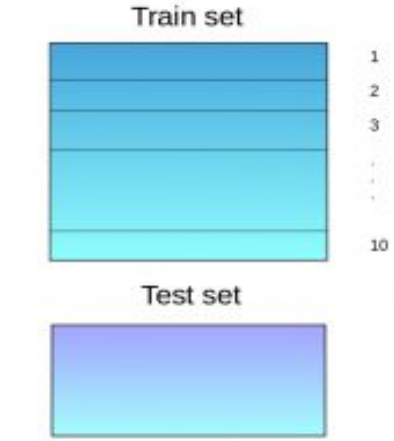
Advanced Ensemble techniques

1. Stacking
2. Blending
3. Bagging

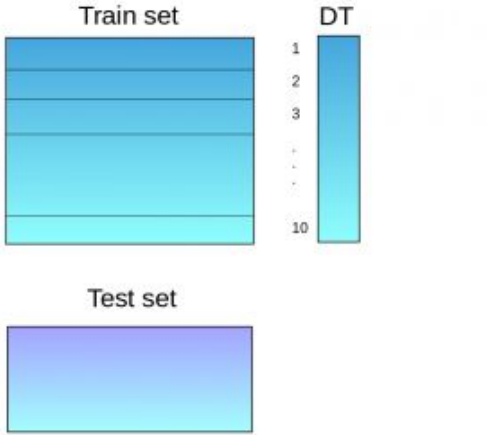
1. Stacking

- Stacking is an ensemble learning technique that uses predictions from multiple models (for example decision tree, knn or svm) to build a new model.
- This model is used for making predictions on the test set.

1. The train set is split into 10 parts

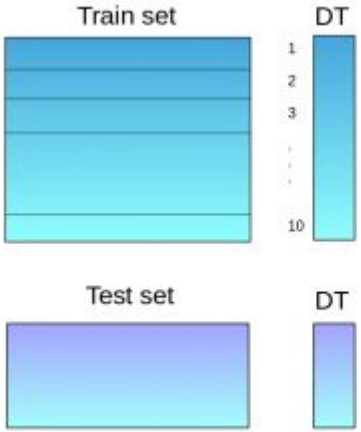


2. A base model (suppose a decision tree) is fitted on 9 parts and predictions are made for the 10th part. This is done for each part of the train set.

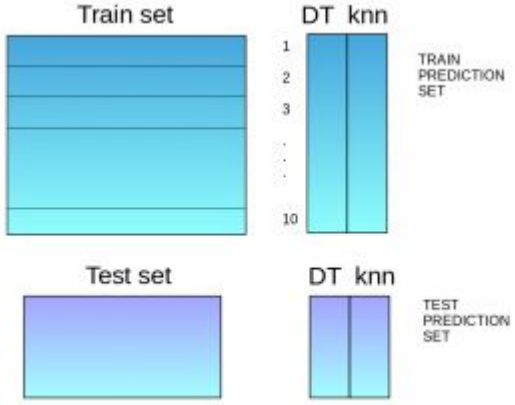


3. The base model (in this case, decision tree) is then fitted on the whole train dataset.

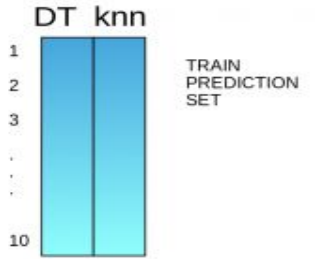
4. Using this model, predictions are made on the test set.



5. Steps 2 to 4 are repeated for another base model (say knn) resulting in another set of predictions for the train set and test set.



6. The predictions from the train set are used as features to build a new model.

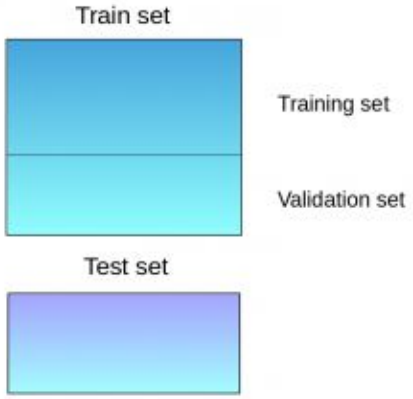


7. This model is used to make final predictions on the test prediction set.

2. Blending

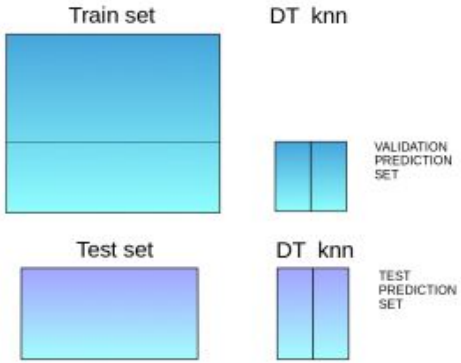
- Blending follows the same approach as stacking but uses only a holdout (validation) set from the train set to make predictions.
- In other words, unlike stacking, the predictions are made on the holdout set only.
- The holdout set and the predictions are used to build a model which is run on the test set.

1. The train set is split into training and validation sets



2. Model(s) are fitted on the training set.

3. The predictions are made on the validation set and the test set.

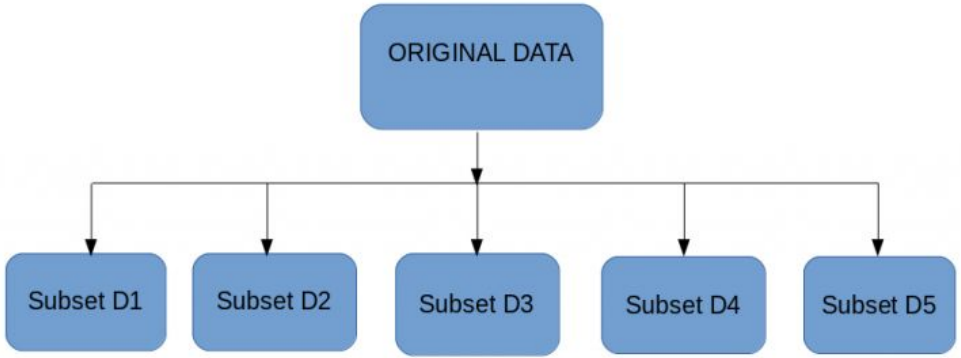


4. The validation set and its predictions are used as features to build a new model.

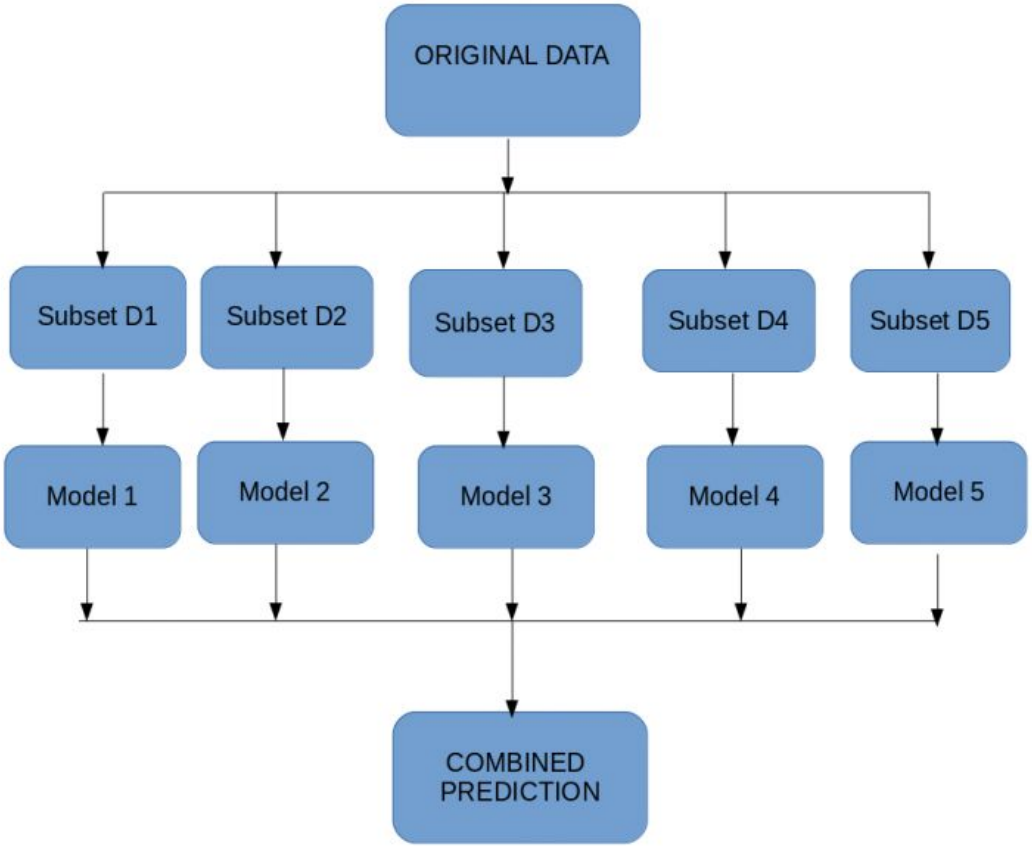
5. This model is used to make final predictions on the test and meta-features.

3. Bagging

- The idea behind bagging is combining the results of multiple models (for instance, all decision trees) to get a generalized result.
- Here's a question: If you create all the models on the same set of data and combine it, will it be useful?
- There is a high chance that these models will give the same result since they are getting the same input. So how can we solve this problem?
- One of the techniques is bootstrapping.
- Bootstrapping is a sampling technique in which we create subsets of observations from the original dataset, with replacement. The size of the subsets is the same as the size of the original set.
- Bagging (or Bootstrap Aggregating) technique uses these subsets (bags) to get a fair idea of the distribution (complete set). The size of subsets created for bagging may be less than the original set.



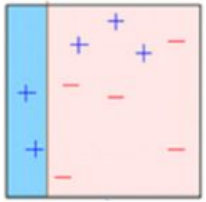
A. Multiple subsets are created from the original dataset, selecting observations with replacement.
B. A base model (weak model) is created on each of these subsets.
C. The models run in parallel and are independent of each other.
D. The final predictions are determined by combining the predictions from all the models.


ORIGINAL DATA
Subset D1
Subset D2
Subset D3
Subset D4
Subset D5
Model 1
Model 2
Model 3
Model 4
Model 5
COMBINED PREDICTION

4. Boosting

- Before we go further, here's another question for you: If a data point is incorrectly predicted by the first model, and then the next (probably all models), will combining the predictions provide better results?
- Such situations are taken care of by boosting.
- Boosting is a sequential process, where each subsequent model attempts to correct the errors of the previous model.
- The succeeding models are dependent on the previous model. Let's understand the way boosting works in the below steps.
    1. A subset is created from the original dataset.
    2. Initially, all data points are given equal weights.
    3. A base model is created on this subset.

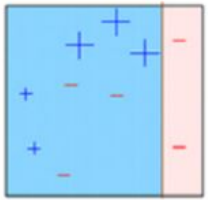
4. This model is used to make predictions on the whole dataset.



5. Errors are calculated using the actual values and predicted values.

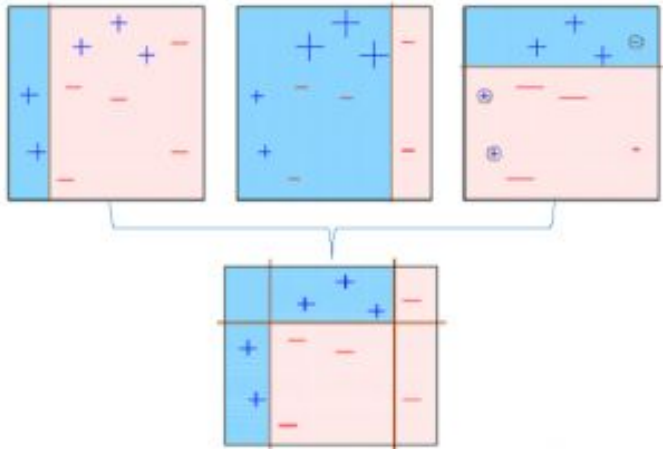
6. The observations which are incorrectly predicted, are given higher weights. (Here, the three misclassified blue-plus points will be given higher weights)

7. Another model is created and predictions are made on the dataset. (This model tries to correct the errors from the previous model)

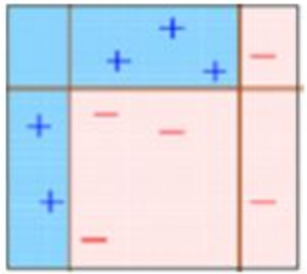


8. Similarly, multiple models are created, each correcting the errors of the previous model.

9. The final model (strong learner) is the weighted mean of all the models (weak learners).



- Thus, the boosting algorithm combines a number of weak learners to form a strong learner.
- The individual models would not perform well on the entire dataset, but they work well for some part of the dataset. Thus, each model actually boosts the performance of the ensemble.



Algorithms based on Bagging and Boosting

- Bagging and Boosting are two of the most commonly used techniques in machine learning. In this section, we will look at them in detail. Following are the algorithms we will be focusing on:
- Bagging algorithms:

    A. Bagging meta-estimator
    B. Random forest

- Boosting algorithms:

    A. AdaBoost
    B. GBM
    C. XGBM
    D. Light GBM
    E. CatBoost