1. Dataset

- A dataset is a particular instance of data that is used for analysis or model building at any given time.

TRS

- A dataset comes in different flavors such as numerical data, categorical data, text data, image data, voice data, and video data.

- For beginning data science projects, the most popular type of dataset is a dataset containing numerical data that is typically stored in a comma-separated values (CSV) file format

2. Data Wrangling

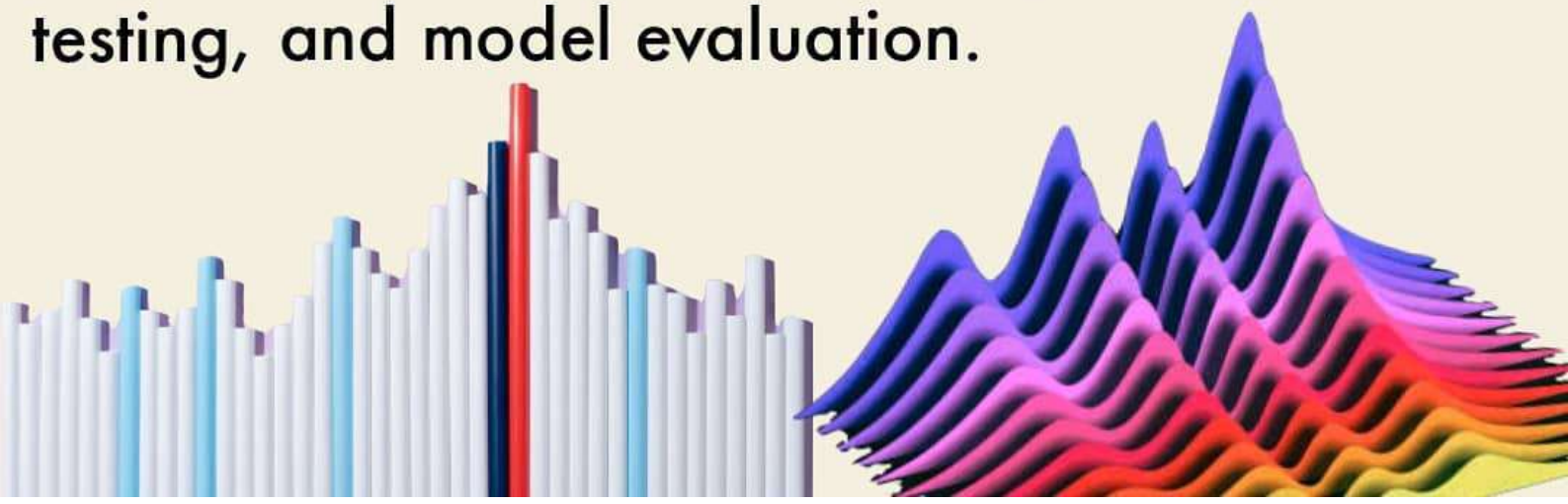
- Data wrangling is the process of converting data from its raw form to a tidy form ready for analysis.
- Data wrangling is an important step in data preprocessing and includes several processes like data importing, data cleaning, data structuring, string processing, HTML parsing, handling dates and times, handling missing data, and text mining.

3. Data Visualization

- It is one of the main tools used to analyze and study relationships between different variables.

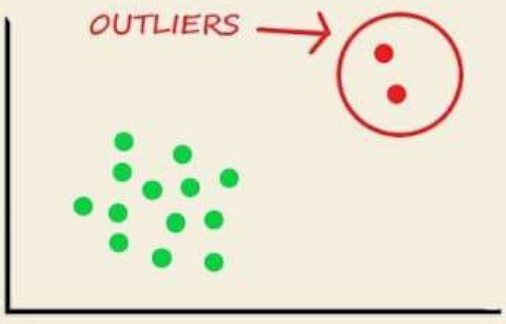
- Data visualization (e.g., scatter plots, line graphs, bar plots, histograms, qqplots, smooth densities, boxplots, pair plots, heat maps, etc.) can be used for descriptive analytics.

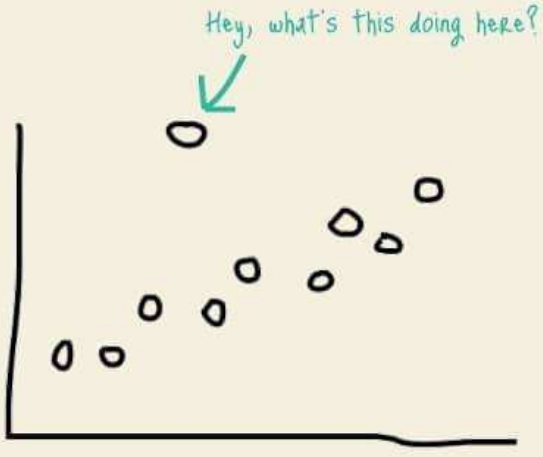
- Data visualization is also used in machine learning for data preprocessing and analysis, feature selection, model building, model testing, and model evaluation.



4. Outliers

- An outlier is a data point that is very different from the rest of the dataset.
- Outliers are very common and are expected in large datasets.
- One common way to detect outliers in a dataset is by using a box plot.
- Outliers can significantly degrade the predictive power of a machine learning model.
- Advanced methods for dealing with outliers include the RANSAC method.





5. Data Imputation

- Most datasets contain missing values. However, the removal of samples or dropping of entire feature columns is simply not feasible because we might lose too much valuable data.

- So, here we can use different interpolation techniques to estimate the missing values from the other training samples in our dataset.

- One of the most common interpolation techniques is mean imputation, where we simply replace the missing value with the mean value of the entire feature column.

6. Data Scaling

- Scaling your features will help improve the quality and predictive power of your model.

- Without scaling your features, the model will be biased towards a particular feature.

- In order to bring features to the same scale, we could decide to use either normalization or standardization of features.

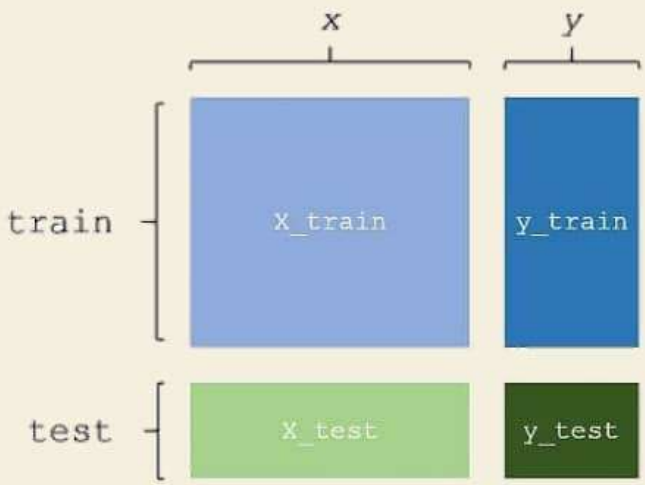
TRS

7. Data Partitioning

- In machine learning, the dataset is often partitioned into training and testing sets.

- The model is trained on the training dataset and then tested on the testing dataset.

- The testing dataset thus acts as the unseen dataset, which can be used to estimate a generalization error (the error expected when the model is applied to a real-world dataset after the model has been deployed).




```
#import
from sklearn.model_selection import
train_test_split
X_train,X_test,y_train,y_test = train_test_split
(X,y,test_size=0.3)
```

8. Supervised Learning

- These are machine learning algorithms that perform learning by studying the relationship between the feature variables and the known target variable.

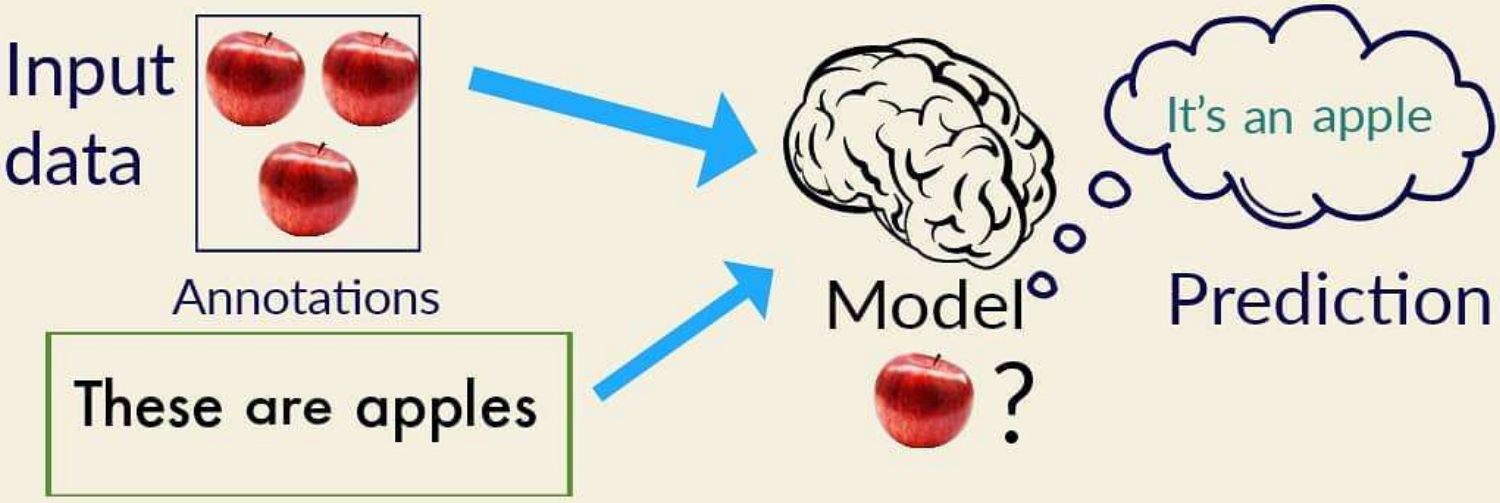
Supervised learning has two subcategories:

a) Continuous Target Variables

- Linear Regression, KNeighbors regression (KNR), and Support Vector Regression (SVR).

b) Discrete Target Variables

- Logistic Regression classifier, Support Vector Machines (SVM), Decision tree classifier

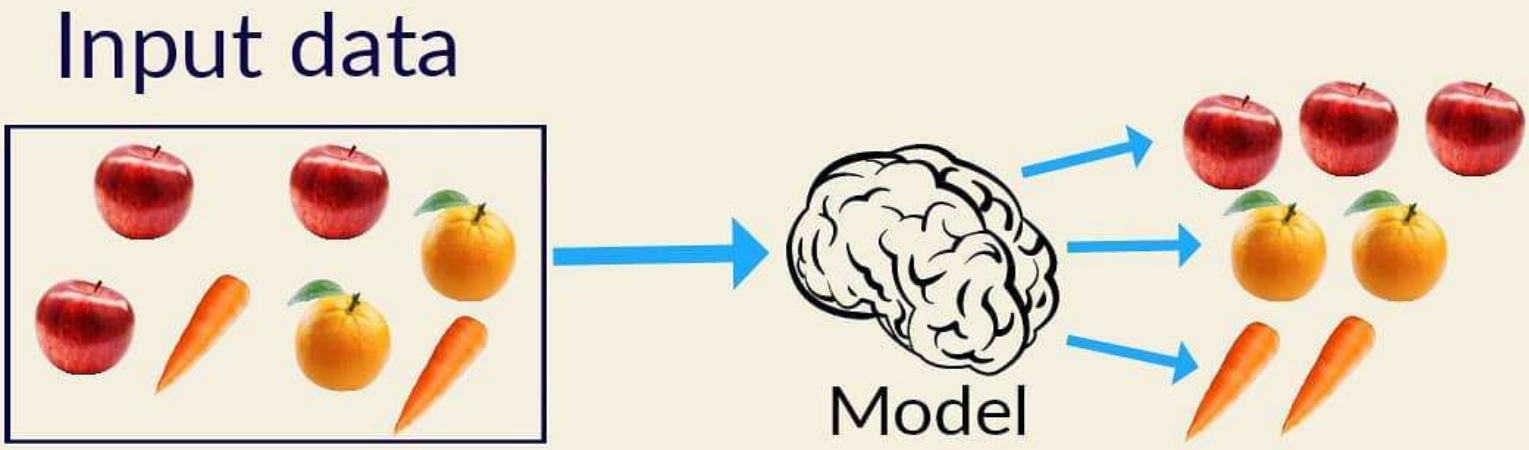


9. Unsupervised Learning

- In unsupervised learning, we deal with unlabeled data or data of unknown structure.

- Using unsupervised learning techniques, we are able to explore the structure of our data to extract meaningful information without the guidance of a known outcome variable or reward function.

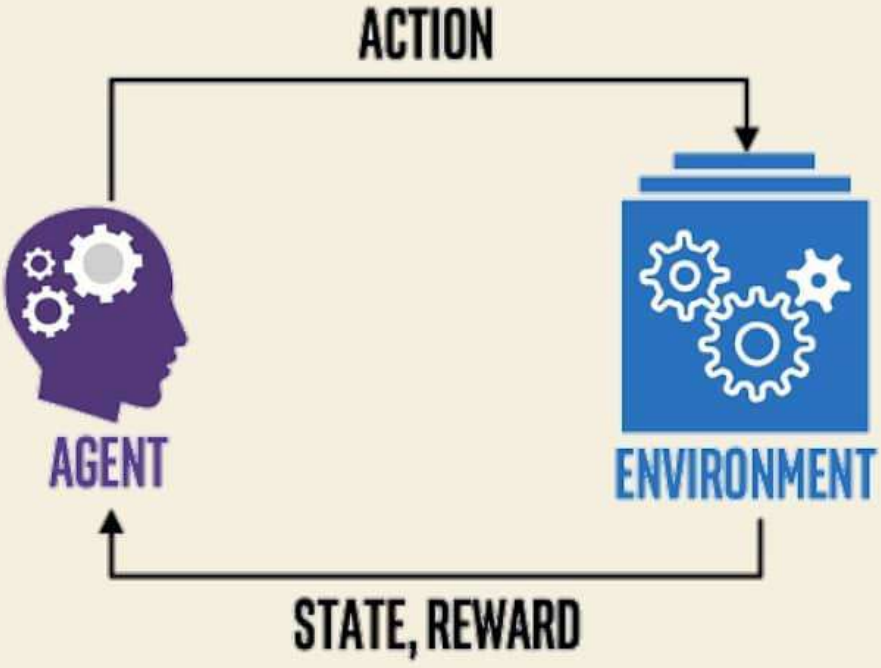
- K-means clustering is an example of an unsupervised learning algorithm.



10. Reinforcement Learning

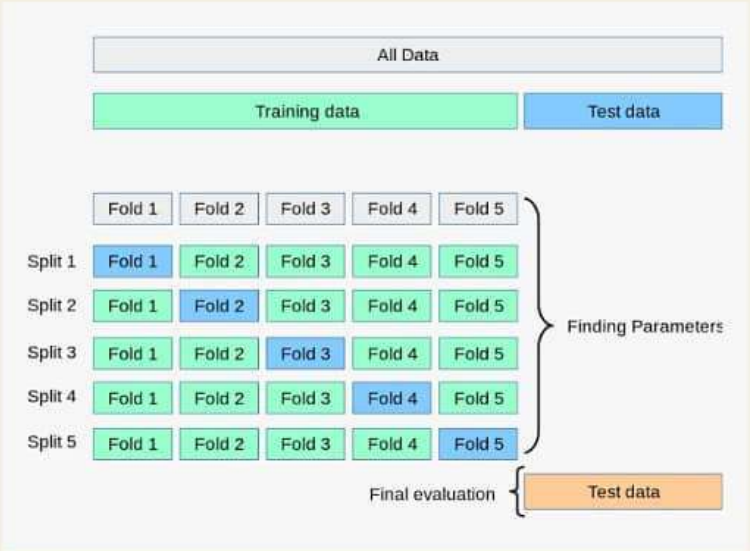
- Reinforcement Learning(RL) is a type of machine learning technique that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences.

- Reinforcement learning uses rewards and punishment as signals for positive and negative behavior.



11. Cross-validation

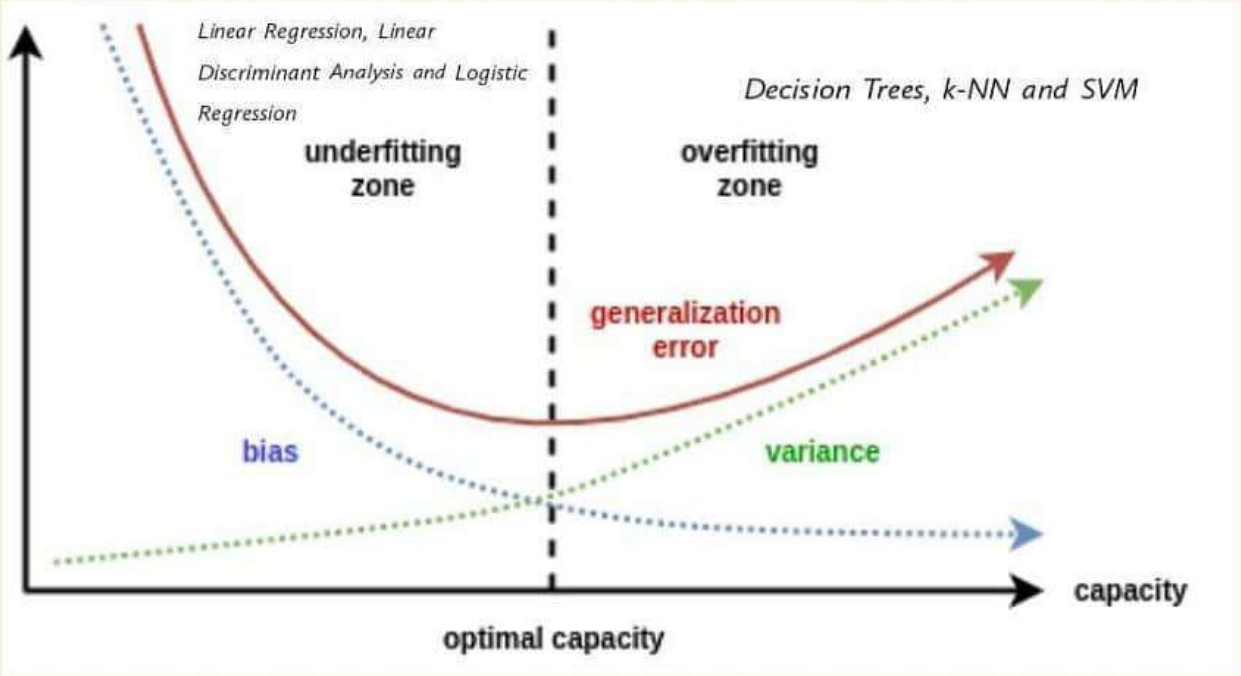
- Cross-validation is a method of evaluating a machine learning model's performance across random samples of the dataset.
- In k-fold cross-validation, the dataset is randomly partitioned into training and testing sets.
- The model is trained on the training set and evaluated on the testing set. The process is repeated k-times.
- The average training and testing scores are then calculated by averaging over the k-folds.



12. Bias-variance Tradeoff

- A model having high bias and low variance assumes more assumptions about the form of the target function, and a model having high variance and low bias over learns the training dataset.

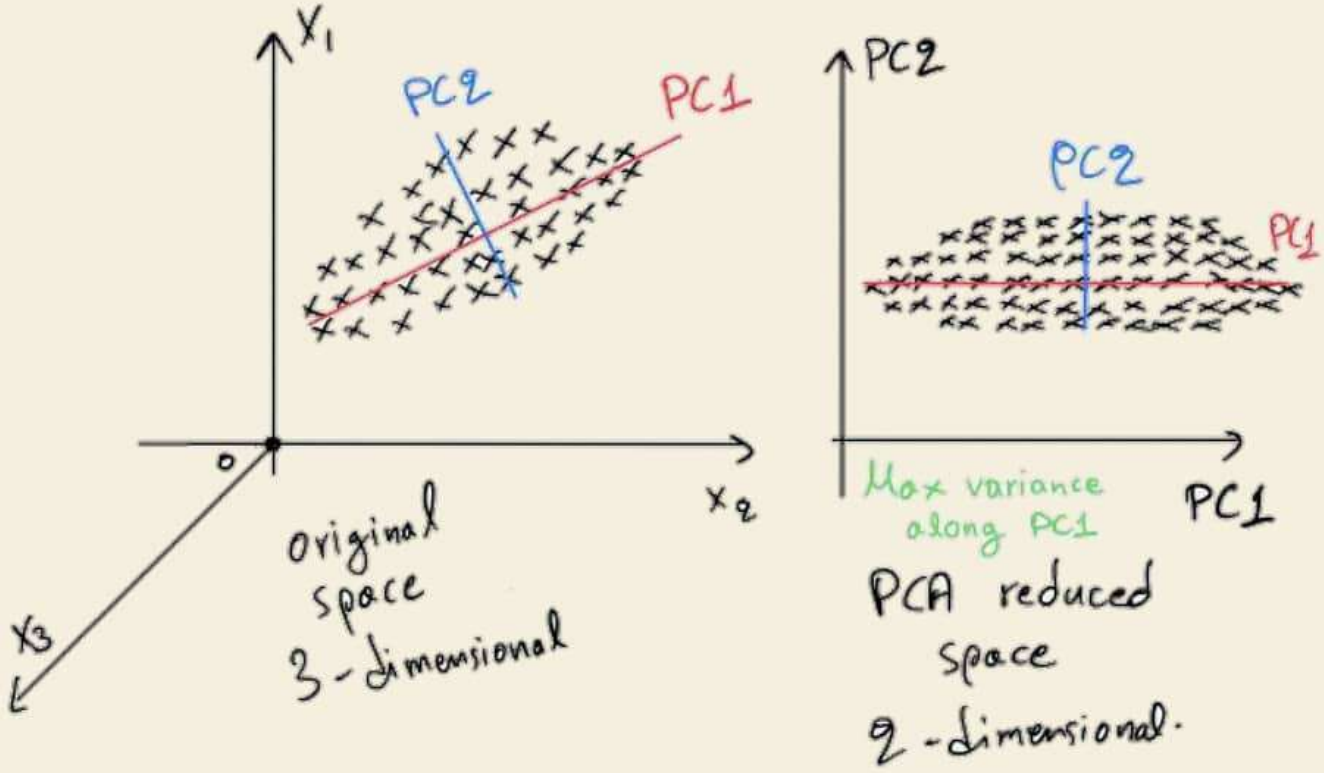
- The parameters of the model should be tuned to get the best fit model, that performs the best in production.



13. Principal Component Analysis (PCA)

- Principal Component Analysis (PCA) is a statistical method that is used for feature extraction. It is used for high-dimensional and correlated data.

- The basic idea of PCA is to transform the original space of features into the space of the principal component.



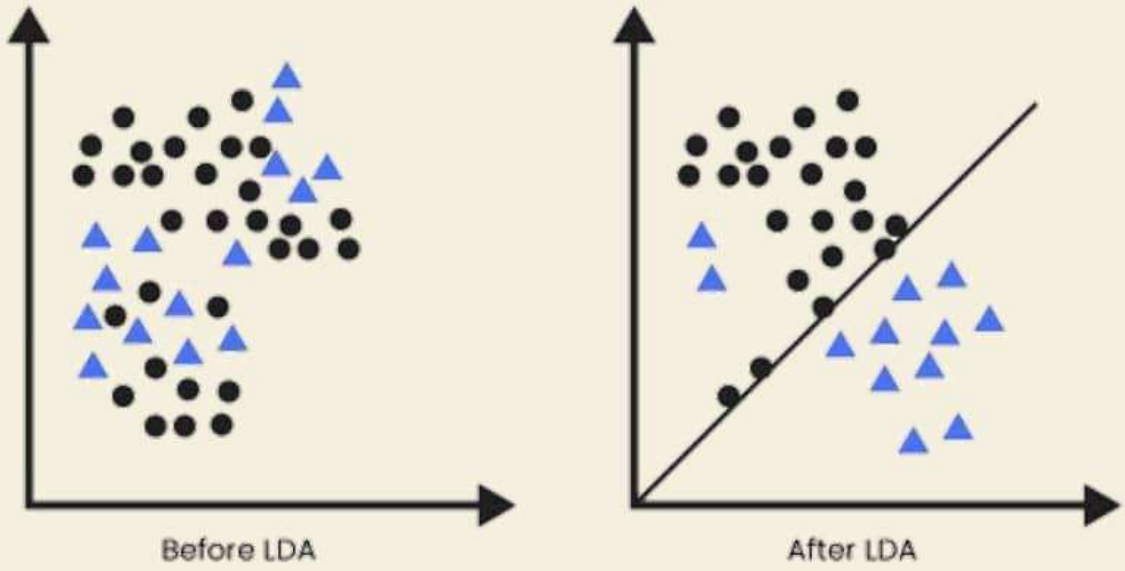
Credit : Serafeim Loukas

14. Linear Discriminant Analysis (LDA)

- Linear Discriminant Analysis is a dimensionality reduction technique which is commonly used for the supervised classification problems.

- It is used for modeling differences in groups i.e. separating two or more classes.

- Just like the PCA, It is used to project the features in higher dimension space into a lower dimension space.



15) Model Parameters and Hyperparameters

Model Parameters

- These are the parameters in the model that must be determined using the training data set. These are the fitted parameters.

- For example, suppose we have a model such as house price = a + b*(age) + c*(size).

- To estimate the cost of houses based on the age of the house and its size (square foot), then a, b, and c will be our model or fitted parameters.

Hyperparameters

- These are adjustable parameters that must be tuned to obtain a model with optimal performance.

Some examples of hyperparameters in machine learning:

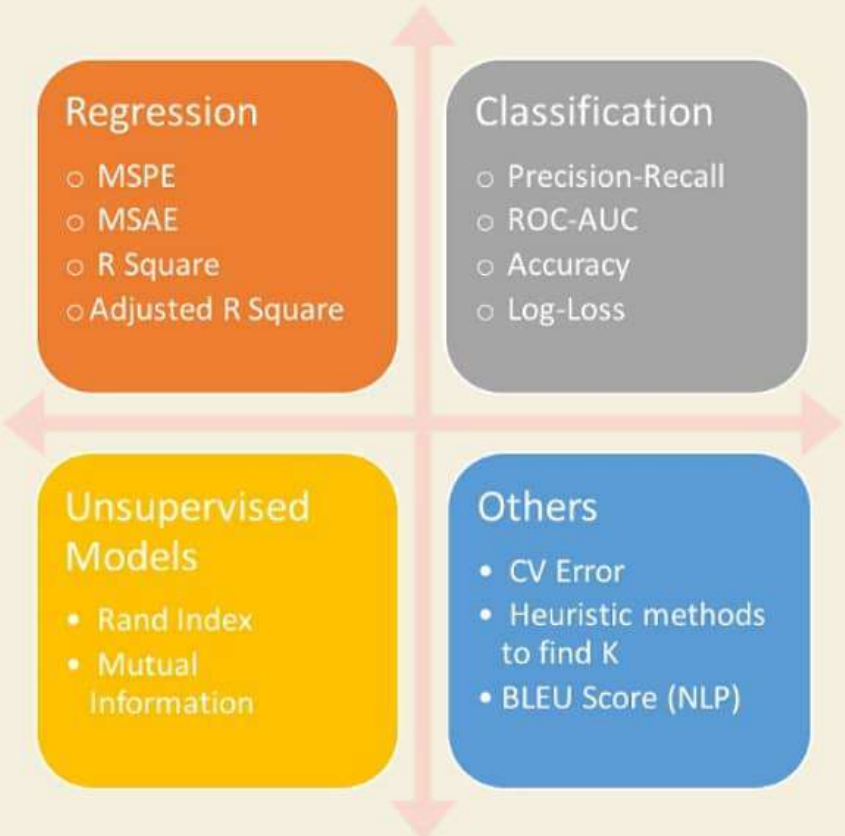
- Learning Rate
- Number of Epochs
- Regularization constant
- Number of branches in a decision tree
- Number of clusters in a clustering algorithm (like k-means)

- It is important that during training, the hyperparameters be tuned to obtain the model with the best performance (with the best-fitted parameters).

16) Evaluation Metrics

- In machine learning (predictive analytics), there are several metrics that can be used for model evaluations.

- A supervised learning (discrete target) model, also referred to as a classification model, can be evaluated using metrics such as accuracy, precision, recall, f1 score, and the area under ROC curve (AUC).



17) Math Concepts

a) Basic Calculus

- Most machine learning models are built with a dataset having several features or predictors. Hence, familiarity with multivariable calculus is extremely important for building a machine learning model. Here are the topics you need to be familiar with:

Functions of several variables; Derivatives and gradients; Step function, Sigmoid function, Logit function, ReLU (Rectified Linear Unit) function; Cost function; Plotting of functions; Minimum and Maximum values of a function

17) Math Concepts

b) Basic Linear Algebra

- Linear algebra is the most important math skill in machine learning. A data set is represented as a matrix. Linear algebra is used in data preprocessing, data transformation, dimensionality reduction, and model evaluation. Here are the topics you need to be familiar with:

Vectors; Norm of a vector; Matrices; Transpose of a matrix; The inverse of a matrix; The determinant of a matrix; Trace of a Matrix; Dot product; Eigenvalues; Eigenvectors

17) Math Concepts

c) Optimization Methods

- Most machine learning algorithms perform predictive modeling by minimizing an objective function, thereby learning the weights that must be applied to the testing data in order to obtain the predicted labels. Here are the topics you need to be familiar with:

Cost function/Objective function; Likelihood function; Error function; Gradient Descent Algorithm and its variants (e.g., Stochastic Gradient Descent Algorithm)

18) Statistics and Probability Concepts

- Statistics and Probability are used for visualization of features, data preprocessing, feature transformation, data imputation, dimensionality reduction, feature engineering, model evaluation, etc. Here are the topics you need to be familiar with:

Mean, Median, Mode, Standard deviation/variance, Correlation coefficient and the covariance matrix, Probability distributions (Binomial, Poisson, Normal), p-value, Bayes Theorem (Precision, Recall, Positive Predictive Value, Negative Predictive Value, Confusion Matrix, ROC Curve), Central Limit Theorem, R_2 score, Mean Square Error (MSE), A/B Testing, Monte Carlo Simulation

19) Regularization

- Regularisation is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.

The commonly used regularisation techniques are :

1) L1 regularisation LASSO (Least Absolute Shrinkage and Selection Operator)
2) L2 regularisation Ridge Regression
3) Dropout regularisation

- Lasso Regression adds "absolute value of magnitude" of coefficient as penalty term to the loss function(L).

$$||\mathbf{w}||_1 = |w_1| + |w_2| + ... + |w_N|$$

- Ridge regression adds "squared magnitude" of coefficient as penalty term to the loss function(L).

$$\|\mathbf{w}\|_2 = \left(|w_1|^2 + |w_2|^2 + \ldots + |w_N|^2\right)^{\frac{1}{2}}$$