Regularization

- One of the major aspects of training your machine learning model is avoiding overfitting. The model will have a low accuracy if it is overfitting. This happens because your model is trying too hard to capture the noise in your training dataset. By noise we mean the data points that don't really represent the true properties of your data, but random chance. Learning such data points, makes your model more flexible, at the risk of overfitting.
- In mathematics, statistics, and computer science, particularly in the fields of machine learning and inverse problems, regularization is a process of introducing additional information in order to solve an ill-posed problem or to prevent overfitting.This technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.

Normalization

- Normalization is an example of preprocessing data to remove or reduce the burden from machine learning (ML) to learn certain invariants, that is, things which make no difference in the meaning of the symbol, but only change the representation.
- Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.
- Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance.
- If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.
- The key difference between regularization and normalization is that regularization is used to solve the overfitting problem and normalization is used to solve the feature scaling problem

Sampling

- Oversampling and undersampling can be used to alter the class distribution of the training data and both methods have been used to deal with class imbalance.

- It is worth asking why anyone would use it rather than a cost-sensitive learning algorithm for dealing with data with a skewed class distribution and non-uniform misclassification costs. There are several reasons for this. The most obvious reason is there are not cost-sensitive implementations of all learning algorithms and therefore a wrapper-based approach using sampling is the only option.
- The reason that altering the class distribution of the training data aids learning with highly-skewed data sets is that it effectively imposes non-uniform misclassification costs. For example, if one alters the class distribution of the training set so that the ratio of positive to negative examples goes from 1:1 to 2:1, then one has effectively assigned a misclassification cost ratio of 2:1. This equivalency between altering the class distribution of the training data and altering the misclassification cost ratio is well known
- A second reason for using sampling is that many highly skewed data sets are enormous and the size of the training set must be reduced in order for learning to be feasible. In this case, undersampling seems to be a reasonable, and valid, strategy. If one needs to discard some training data, it still might be beneficial to discard some of the majority class examples in order to reduce the training set size to the required size, and then also employ a cost-sensitive learning algorithm, so that the amount of discarded training data is minimized.
- A final reason that may have contributed to the use of sampling rather than a cost-sensitive learning algorithm is that misclassification costs are often unknown. However, this is not a valid reason for using sampling over a cost sensitive learning algorithm, since the analogous issue arises with sampling—what should the class distribution of the final training data be? If this cost information is not known, a measure such as the area under the ROC curve could be used to measure classifier performance and both approaches could then empirically determine the proper cost ratio/class distribution.
- Key features:
    - It can increase the accuracy of the model
    - It can stimulate complex processes
    - It is lower cost

Types of Sampling

- Stratified Sampling :
    - Stratified sampling is a probability sampling technique wherein the researcher divides the entire population into different subgroups or strata, then randomly selects the final subjects proportionally from the different strata.

    - Stratified random sampling is used when the researcher wants to highlight a specific subgroup within the population. This technique is useful in such researches because it ensures the presence of the key subgroup within the sample.
- Simple Random Sampling:
    - A simple random sample is a subset of a statistical population in which each member of the subset has an equal probability of being chosen. An example of a simple random sample would be the names of 25 employees being chosen out of a hat from a company of 250 employees. In this case, the population is all 250 employees, and the sample is random because each employee has an equal chance of being chosen.
    - Simple random sampling is more commonly used when the researcher knows little about the population. If the researcher knew more, it would be better to use a different sampling technique, such as stratified random sampling, which helps to account for the differences within the population, such as age, race or gender.
- Systematic sampling :
    - Systematic sampling is a type of probability sampling method in which sample members from a larger population are selected according to a random starting point and a fixed, periodic interval. This interval, called the sampling interval, is calculated by dividing the population size by the desired sample size.
    - Despite the sample population being selected in advance, systematic sampling is still thought of as being random if the periodic interval is determined beforehand and the starting point is random.
- Cluster sampling :
    - Cluster sampling refers to a type of sampling method . With cluster sampling, the researcher divides the population into separate groups, called clusters. Then, a simple random sample of clusters is selected from the population. The researcher conducts his analysis on data from the sampled clusters.
    - Compared to simple random sampling and stratified sampling , cluster sampling has advantages and disadvantages. For example, given equal sample sizes, cluster sampling usually provides less precision than either simple random sampling or stratified sampling. On the other hand, if travel costs between clusters are high, cluster sampling may be more cost-effective than the other methods.
    - The main aim of cluster sampling can be specified as cost reduction and increasing the levels of efficiency of sampling. This specific technique can also be applied in integration with multi-stage sampling.

Probability methods

This is the best overall group of methods to use as you can subsequently use the most powerful statistical analyses on the results.

Method	**Best when**
Simple random sampling	Whole population is available.
Stratified sampling (random within target groups)	There are specific sub-groups to investigate (eg. demographic groupings).
Systematic sampling (every nth person)	When a stream of representative people are available (eg. in the street).
Cluster sampling (all in limited groups)	When population groups are separated and access to all is difficult, eg. in many distant cities.

Confusion Matrix

- A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.
- Let's now define the most basic terms, which are whole numbers (not rates):
    - true positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease.
    - true negatives (TN): We predicted no, and they don't have the disease.
    - false positives (FP): We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")
    - false negatives (FN): We predicted no, but they actually do have the disease. (Also known as a "Type II error.")
- Accuracy: Overall, how often is the classifier correct?
    - (TP+TN)/total
- Misclassification Rate: Overall, how often is it wrong?
    - (FP+FN)/total
    - equivalent to 1 minus Accuracy
    - also known as "Error Rate"
- True Positive Rate: When it's actually yes, how often does it predict yes?
    - TP/actual yes
    - also known as "Sensitivity" or "Recall"
- Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually

survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.
    - Precision = TP/(TP+FP)
- False Positive Rate: When it's actually no, how often does it predict yes?
    - FP/actual no
- Specificity: When it's actually no, how often does it predict no?
    - TN/actual no
    - equivalent to 1 minus False Positive Rate
- Precision: When it predicts yes, how often is it correct?
    - TP/predicted yes
- Prevalence: How often does the yes condition actually occur in our sample?
    - actual yes/total
- Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. The question recall answers is: Of all the passengers that truly survived, how many did we label? We have got recall of 0.631 which is good for this model as it’s above 0.5.
    - Recall = TP/TP+FN

Naive Bayes Classifier

- Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values.
- It is called naive Bayes or idiot Bayes because the calculation of the probabilities for each hypothesis are simplified to make their calculation tractable.

Decision Tree Classifier

- Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter.
- The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

K-Means

- K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to group points based on some feature, where the number of groups is represented by the variable K.

- The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.
- It assumes that
    - the variance of the distribution of each attribute (variable) is spherical
    - the variance of all variables are the same
    - the prior probability for all k clusters are the same, i.e. each cluster has roughly equal number of observations
    - performs poorly on categorical than numerical data
    - applicable for data whose variables are numerical

Random Forest

- Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees.
- In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results.
- The same random forest algorithm or the random forest classifier can use for both classification and the regression task.

Data Visualization

- Scatter plot, histogram, bar graph, matrix diagram, tableau are good visualization tools.
- Data preparation is a necessary step in data visualization. The amount of time needed for data preparation for a particular analysis problem ,directly depends on the health of the data i.e. how complete it is, how many missing values are there, how clean it is and what are the inconsistencies.
- R also has many good visualization packages similar to python

.

Scatter Plot

- Scatter plots are important in statistics for data visualization because they can show the extent of correlation, if any, between the values of observed quantities or phenomena (called variables).
- If no correlation exists between the variables, the points appear randomly scattered on the coordinate plane. If a large correlation exists, the points concentrate near a straight line. Scatter plots are useful data visualization tools for illustrating a trend.

Histogram

- A Histogram visualises the distribution of data over a continuous interval or certain time period. Each bar in a histogram represents the tabulated frequency at each interval/bin.
- Histograms help give an estimate as to where values are concentrated, what the extremes are and whether there are any gaps or unusual values. They are also useful for giving a rough view of the probability distribution.

Matrix Diagram

- A Matrix Diagram (MD) is a data visualization tool that allows a team to identify the presence and strengths of relationships between two or more lists of items.
- It provides a compact way of representing many-to-many relationships of varying strengths
- Relationships between things are often complex (many-to-many) and require us to think in more than one-dimension. The Matrix Diagram is a simple tool that allows relatively complex situations to be analysed in a simple straightforward way. They help us to expose interactions and dependencies between things that help us to understand complex causal relationships
- Essentially, the matrix chart is a table made up of rows and columns that present data visually and can be seen as the visual equivalent of a cross tabulation that divides data between the variables.

Cross Tabulation

- When conducting survey analysis, cross tabulations (also referred to as cross-tabs) are a quantitative research method appropriate for analyzing the relationship between two or more variables. Cross tabulations provide a way of analyzing and comparing the results for one or more variables with the results of another (or others).
- The axes of the table may be specified as being just one variable or formed from a number of variables. The resulting table will have as many rows and columns as there are codes in the corresponding axis specification.

Data Transformation

- Transformation is a mathematical operation that changes the measurement scale of a variable. This is usually done to make a set of usable with a particular statistical test or method.
- Many statistical methods require data that follow a particular kind of distribution, usually a normal distribution. All of the observations must

come from a population that follows a normal distribution. Groups of observations must come from populations that have the same variance or standard deviation. Transformations that normalize a distribution commonly make the variance more uniform and vice versa.

- If a population with a normal distribution is sampled at random then the means of the samples will not be correlated with the standard deviations of the samples. This partly explains why normalizing transformations also make variances uniform. The Central Limit Theorem (the means of a large number of samples follow a normal distribution) is a key to understanding this situation.
- Many biomedical observations will be a product of different influences, for example the resistance of blood vessels and output from the heart are two of the influences most closely related to blood pressure.
- In mathematical terms these influences usually multiply together to give an overall influence, so, if we take the logarithm of the overall influence then this is the sum of the individual influences [log(A * B) = log(A) + log(B)]. The Central Limit Theorem thus dictates that the logarithm of the product of several influences follows a normal distribution.

Machine Learning Applications

- Sale prediction based on history
- Spam mail detection.
- Real time product recommendation.
- Manual data entry.
- Medical Diagnosis.
- Customer segmentation and Lifetime value prediction.
- Financial analysis.
- Predictive maintenance.
- Image recognition (Computer Vision).

Bias and Variance

- Bias are the simplifying assumptions made by a model to make the target function easier to learn.
- Generally, parametric algorithms have a high bias making them fast to learn and easier to understand but generally less flexible. In turn, they have lower predictive performance on complex problems that fail to meet the simplifying assumptions of the algorithms bias.
    - 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.
- The target function is estimated from the training data by a machine learning algorithm, so we should expect the algorithm to have some variance. Ideally, it should not change too much from one training dataset to the next, meaning that the algorithm is good at picking out the hidden underlying mapping between the inputs and the output variables.
- Machine learning algorithms that have a high variance are strongly influenced by the specifics of the training data. This means that the specifics of the training have influences the number and types of parameters used to characterize the mapping function.
    - 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.
- Overfitting occurs when a statistical model or machine learning algorithm captures the noise of the data. Intuitively, overfitting occurs when the model or the algorithm fits the data too well. Specifically, overfitting occurs if the model or algorithm shows low bias but high variance.
- Overfitting is often a result of an excessively complicated model, and it can be prevented by fitting multiple models and using validation or cross-validation to compare their predictive accuracies on test data.
- Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data. Intuitively, underfitting occurs when the model or the algorithm does not fit the data well enough. Specifically, underfitting occurs if the model or algorithm shows low variance but high bias. Underfitting is often a result of an excessively simple model.
- Both overfitting and underfitting lead to poor predictions on new data sets. If you meet with an overfitting problem, you can use cross validation technique or increase the amount of regularization as well as training data.
- High Bias Techniques
    - Linear Regression, Linear Discriminant Analysis and Logistic Regression
- Low Bias Techniques
    - Decision Trees, K-nearest neighbours and Gradient Boosting
- Low Variance Techniques
    - Linear Regression, Linear Discriminant Analysis, Random Forest, Logistic Regression
- High Variance Techniques
    - Decision Trees, K-nearest neighbours and Support Vector Machine (SVM)

Classification Accuracy

- The skill of a classification machine learning algorithm is often reported as classification accuracy.
- This is the percentage of the correct predictions from all predictions made. It is calculated as follows:
    - classification accuracy = correct predictions / total predictions * 100.0
- If suppose we have performed a supervised classification on a dataset containing 100 test set instances out of which 80 are correctly classified. Then, 72% and 88% represents the 95% test set accuracy confidence boundary.

Classification Error

- You can calculate classification error as the percentage of incorrect predictions to the number of predictions made, expressed as a value between 0 and 1.
    - classification error = incorrect predictions / total predictions

Confidence Interval

- Rather than presenting just a single error score, a confidence interval can be calculated and presented as part of the model skill.
- A confidence interval is comprised of two things:
    - Range. This is the lower and upper limit on the skill that can be expected on the model.
    - Probability. This is the probability that the skill of the model will fall within the range.
- In general, the confidence interval for classification error can be calculated as follows:
    - (error +/-[const * sqrt((error * (1 - error)) / n)])*10
- In general, the confidence interval for classification accuracy can be calculated as follows:
    - (accuracy +/-[const * sqrt((error * (1 - error)) / n)])*100
- The values for const are provided from statistics, and common values used are:
    - 1.64 (90%)
    - 1.96 (95%)
    - 2.33 (98%)
    - 2.58 (99%)

Linear Regression

- Linear regression is an analysis that assesses whether one or more predictor variables explain the dependent (criterion) variable. The regression has five key assumptions:
    - Linear relationship between feature and target variables
    - The feature variables are normally distributed
    - Multivariate normality
    - No or little multicollinearity
    - No autocorrelation
    - Homoscedasticity

Data Mining

- Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers and develop more effective marketing strategies as well as increase sales and decrease costs. Data mining depends on effective data collection and warehousing as well as computer processing.
- Data mining programs analyze relationships and patterns in data based on what users request.In other cases, data miners find clusters of information based on logical relationships, or they look at associations and sequential patterns to draw conclusions about trends in consumer behavior.
- Data mining is all about finding patterns in data

Overfitting

- In other cases, data miners find clusters of information based on logical relationships, or they look at associations and sequential patterns to draw conclusions about trends in consumer behavior.
- A statistical model is said to be overfitted, when we train it with a lot of data (just like fitting ourselves in an oversized pants!). When a model gets trained with so much of data, it starts learning from the noise and inaccurate data entries in our data set.
- Then the model does not categorize the data correctly, because of too much of details and noise. The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models.
- A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.
- How to Prevent Overfitting:

    - Cross-validation
    - Train with more data
    - Remove features
    - Early stopping
    - Regularization
    - Ensembling

Deep Learning

- Deep learning is a particular kind of machine learning that achieves great power and flexibility by learning to represent the world as nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones.
- Eg. Shape Detection , Cat vs Dog Image Detection , etc

Machine Learning

- A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.
- Machine Learning is the learning in which machine can learn by its own without being explicitly programmed. It is an application of AI that provide system the ability to automatically learn and improve from experience. Here we can generate a program by integrating input and output of that program
- Eg. Machine Learning – Predicting weights based on height,Storm prediction System, etc

Methodological differences between machine learning and statistics

- The difference between the two is that machine learning emphasizes optimization and performance over inference which is what statistics is concerned about.
- This is how a statistician and machine learning practitioner will describe the outcome of the same model:
  ML professional: “The model is 85% accurate in predicting Y, given a, b and c.”
  Statistician: “The model is 85% accurate in predicting Y, given a, b and c; and I am 90% certain that you will obtain the same result.”

Prescriptive Analytics

- Prescriptive analytics is dedicated to finding the best course of action for a given situation.
- Prescriptive analytics is related to both descriptive and predictive analytics. While descriptive analytics aims to provide insight into what has happened

and predictive analytics helps model and forecast what might happen, prescriptive analytics seeks to determine the best solution or outcome among various choices, given the known parameters.
- Prescriptive analytics can also suggest decision options for how to take advantage of a future opportunity or mitigate a future risk, and illustrate the implications of each decision option

Diagnostic Analytics

- Diagnostic Analytics is a form of advanced analytics which examines data or content to answer the question “Why did it happen?”, and is characterized by techniques such as drill-down, data discovery, data mining and correlations.
- Identification of a condition, disease, disorder, or problem by systematic analysis of the background or history, examination of the signs or symptoms, evaluation of the research or test results, and investigation of the assumed or probable causes. Effective prognosis is not possible without effective diagnosis.

Descriptive Analytics

- Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of it. Descriptive statistics are broken down into measures of central tendency and measures of variability, or spread.
- Descriptive analytics aims to provide insight into what has happened

Predictive Analytics

- Predictive analytics is the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data. The goal is to go beyond knowing what has happened to providing a best assessment of what will happen in the future.
- Can be used for detecting fraud , optimizing marketing campaigns, improving operations, reducing risks etc

Statistical Tests

- Chi-squared :
    - Pearson's chi-squared test is a statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance. It is suitable for unpaired data from large samples.
    - It is the most widely used of many chi-squared tests (e.g., Yates, likelihood ratio, portmanteau test in time series, etc.) – statistical

procedures whose results are evaluated by reference to the chi-squared distribution. Its properties were first investigated by Karl Pearson in 1900.
    - In contexts where it is important to improve a distinction between the test statistic and its distribution, names similar to Pearson x-squared test or statistic are used.
- Pearson's correlation coefficient :
    - In statistics, the Pearson correlation coefficient (PCC, pronounced /ˈpɪərsən/), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC) or the bivariate correlation, is a measure of the linear correlation between two variables X and Y.
    - It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.
    - Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.
- Root Mean Squared Error :
    - The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) (or sometimes root-mean-squared error) is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed.
    - The RMSD represents the sample standard deviation of the differences between predicted values and observed values. These individual differences are called residuals when the calculations are performed over the data sample that was used for estimation, and are called prediction errors when computed out-of-sample.
    - The RMSD serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power.
    - RMSD is a measure of accuracy, to compare forecasting errors of different models for a particular dataset and not between datasets, as it is scale-dependent.
    - MSD is the square root of the average of squared errors. The effect of each error on RMSD is proportional to the size of the squared error; thus larger errors have a disproportionately large effect on RMSD. Consequently, RMSD is sensitive to outliers.

- ANOVA :
    - Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among group means in a sample.
    - In the ANOVA setting, the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether the population means of several groups are equal, and therefore generalizes the t-test to more than two groups.
    - ANOVA is useful for comparing (testing) three or more group means for statistical significance. It is conceptually similar to multiple two-sample t-tests, but is more conservative (results in less type I error) and is therefore suited to a wide range of practical problems.
    - It can be used to evaluate the relationship between a categorical value and a numerical variable

Cross Validation

- Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.
- In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data.
- The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation.
- The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.For classification problems, one typically uses stratified k-fold cross-validation, in which the folds are selected so that each fold contains roughly the same proportions of class labels.

Bootstrapping

- Bootstrapping is a sampling technique.In machine learning, the bootstrap method refers to random sampling with replacement. This sample is referred to as a resample.
- This allows the model or algorithm to get a better understanding of the various biases, variances and features that exist in the resample. Taking a

sample of the data allows the resample to contain different characteristics then it might have contained as a whole.
- Bootstrapping is also great for small size data sets that can have a tendency to overfit.The reason to use the bootstrap method is because it can test the stability of a solution. By using multiple sample data sets and then testing multiple models, it can increase robustness.

Stratification

- One common issue in data mining is the size of the data set. It is often limited. When this is the case, the test of the model is an issue. Usually, 2/3 of the data are used for training and validation and 1/3 for final testing.By chance, the training or the test set may not be representative of the overall data set.
- Stratification is one of the process which assures that each class is properly represented in both the training and testing data set while selecting data.
- One way to avoid doing stratification, regarding the training phase is to use k-fold cross-validation. Instead of having only one given validation set with a given class distribution, k different validation sets are used. However, this process doesn't guarantee a correct class distribution among the training and validation sets.

Artificial Intelligence

- The word Artificial Intelligence comprises of two words "Artificial" and "Intelligence". Artificial refers to something which is made by human or non natural thing and Intelligence means ability to understand or think. There is a misconception that Artificial Intelligence is a system, but it is not a system .
- AI is implemented in the system. There can be so many definition of AI, one definition can be "It is the study of how to train the computers so that computers can do things which at present human can do better."Therefore It is a intelligence where we want to add all the capabilities to machine that human contain.
- AI involves machines that can perform tasks that are characteristic of human intelligence. While this is rather general, it includes things like planning, understanding language, recognizing objects and sounds, learning, and problem solving.
- We can put AI in two categories, general and narrow. General AI would have all of the characteristics of human intelligence, including the capacities mentioned above. Narrow AI exhibits some facet(s) of human intelligence, and can do that facet extremely well, but is lacking in other areas. A machine

that’s great at recognizing images, but nothing else, would be an example of narrow AI.
- Machine learning involves the development of self-learning algorithms and artificial intelligence involves developing systems or software to mimic human to respond and behave in a circumstance.
- Machine learning, is a subset of AI. That is, all machine learning counts as AI, but not all AI counts as machine learning. For example, symbolic logic (rules engines, expert systems and knowledge graphs) as well as evolutionary algorithms and Baysian statistics could all be described as AI, and none of them are machine learning.

Statistics

- Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data. In applying statistics to, for example, a scientific, industrial, or social problem, it is conventional to begin with a statistical population or a statistical model process to be studied.
- Populations can be diverse topics such as "all people living in a country" or "every atom composing a crystal". Statistics deals with all aspects of data including the planning of data collection in terms of the design of surveys and experiments
- Statistics emphasizes interference whilst ML focuses on building systems

Reinforced Learning

- Reinforcement Learning is a type of Machine Learning, and thereby also a branch of Artificial Intelligence. It allows machines and software agents to automatically determine the ideal behaviour within a specific context, in order to maximize its performance. Simple reward feedback is required for the agent to learn its behaviour; this is known as the reinforcement signal.
- Reinforcement learning requires the agent to know the rewards for every action and is concerned with the problem of finding suitable actions to take in a given situation in order to maximize a reward
- There are three basic concepts in reinforcement learning: state, action, and reward. The state describes the current situation.Action is what an agent can do in each state.
- When a machine takes an action in a state, it receives a reward. Here the term “reward” is an abstract concept that describes feedback from the environment. A reward can be positive or negative. When the reward is positive, it is corresponding to our normal meaning of reward. When the reward is negative, it is corresponding to what we usually call “punishment.”

Machine Learning languages

- There are several languages which are used for machine learning and data science. R, SQL, Python, Spark, Java, C++, Scala, etc. are important ones. These languages offer ML capabilities with the required packages.

R Programming

- R is an integrated suite of software facilities for data manipulation, calculation and graphical display.
- Among other things it has :
    - an effective data handling and storage facility
    - a suite of operators for calculations on arrays, in particular matrices
    - a large, coherent, integrated collection of intermediate tools for data analysis
    - graphical facilities for data analysis and display either directly at the computer or on hard copy
    - a well developed, simple and effective programming language (called 'S') which includes conditionals, loops, user defined recursive functions and input and output facilities. (Indeed most of the system supplied functions are themselves written in the S language.)
- The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.
- R is very much a vehicle for newly developing methods of interactive data analysis. It has developed rapidly, and has been extended by a large collection of packages. However, most programs written in R are essentially ephemeral, written for a single piece of data analysis.

SQL

- SQL Server Machine Learning Services is an embedded, predictive analytics and data science engine that can execute R and Python code within a SQL Server database as stored procedures, as T-SQL script containing R or Python statements, or as R or Python code containing T-SQL.
- The key value proposition of Machine Learning Services is the power of its proprietary packages to deliver advanced analytics at scale, and the ability to bring calculations and processing to where the data resides, eliminating the need to pull data across the network.

Spark

- SPARK is a formally defined computer programming language based on the Ada programming language, intended for the development of high integrity software used in systems where predictable and highly reliable operation is

essential. It facilitates the development of applications that demand safety, security, or business integrity.

- SPARK aims to exploit the strengths of Ada while trying to eliminate all its potential ambiguities and insecurities. SPARK programs are by design meant to be unambiguous, and their behavior is required to be unaffected by the choice of Ada compiler.
- The combination of these approaches is meant to allow SPARK to meet its design objectives, which are:
    - logical soundness
    - rigorous formal definition
    - simple semantics
    - Security
    - expressive power
    - Verifiability
    - bounded resource (space and time) requirements
    - minimal runtime system requirements

SPSS

- SPSS is short for Statistical Package for the Social Sciences, and it's used by various kinds of researchers for complex statistical data analysis.It was was originally launched in 1968 by SPSS Inc., and was later acquired by IBM in 2009. Officially dubbed IBM SPSS Statistics, most users still refer to it as SPSS.
- The current versions (2015) are named IBM SPSS Statistics. The software name originally stood for Statistical Package for the Social Sciences (SPSS), reflecting the original market, although the software is now popular in other fields as well, including the health sciences and marketing.
- It doesn’t offer machine learning capabilities.

Classification

- Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).
- The output variables are often called labels or categories. The mapping function predicts the class or category for a given observation.
- It is common for classification models to predict a continuous value as the probability of a given example belonging to each output class. The probabilities can be interpreted as the likelihood or confidence of a given example belonging to each class. A predicted probability can be converted into a class value by selecting the class label that has the highest probability.

Regression

- Regression predictive modeling is the task of approximating a mapping function (f) from input variables (X) to a continuous output variable (y).
- A continuous output variable is a real-value, such as an integer or floating point value. These are often quantities, such as amounts and sizes.
- Because a regression predictive model predicts a quantity, the skill of the model must be reported as an error in those predictions.
- There are many ways to estimate the skill of a regression predictive model, but perhaps the most common is to calculate the root mean squared error, abbreviated by the acronym RMSE.

SVM

- Support Vector Machines(SVMs) have been extensively researched in the data mining and machine learning communities for the last decade and actively applied to applications in various domains.
- SVM can be used for both classification and regression
- SVMs are typically used for learning classification,regression, or ranking functions, for which they are called classifying SVM, support vector regression (SVR), or ranking SVM (or RankSVM) respectively.
- Two special properties of SVMs are that SVMs achieve (1) high generalization by maximizing the margin and (2) support an efficient learning of nonlinear functions by kernel trick.

Supervised Learning

- Supervised learning is one of the methods associated with machine learning which involves allocating labeled data so that a certain pattern or function can be deduced from that data.
- It is worth noting that supervised learning involves allocating an input object, a vector, while at the same time anticipating the most desired output value, which is mostly referred to as the supervisory signal. The bottom line property of supervised learning is that the input data is known and labeled appropriately.
- Consider there are two supervised learner models and we compare the test set accuracy of the models using the classical hypothesis testing paradigm using a 95% confidence setting. If the computed value of P is 2.53, then we can conclude that the models differ significantly in their performance.

Unsupervised Learning

- Unsupervised learning is the second method of machine learning algorithm where inferences are drawn from unlabeled input data.

- The goal of unsupervised learning is to determine the hidden patterns or grouping in data from unlabeled data. It is mostly used in exploratory data analysis. One of the defining characters of unsupervised learning is that both the input and output are not known.
- Unsupervised Learning is a type of ML from data without a target variable and labeled responses. Cluster Analysis is a most commonly used unsupervised learning method.

Key notes between Supervised and Unsupervised Learning

- Both require atleast one input attribute
- Supervised learning requires atleast one output attribute while unsupervised learning doesn’t

Data wrangling

- Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.