**Naive Bayes : This classifier assumes that every feature in a class in independent. This is better explained taking an example of a fruit**  apple. if it is red, round and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier would consider all of these properties to independently contribute to the probability that this fruit is an apple.

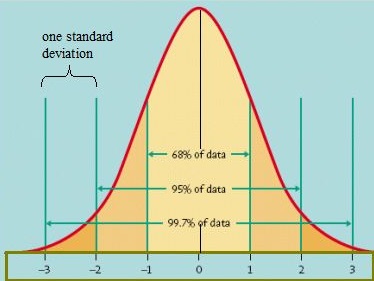
**Nan : Not an integer(Nan)** numeric data type value representing an undefined or unrepresentable value,if a dataset has Nan values it means that the data at that location is either missing or represented incorrectly.

**Natural Language Processing: This field aims to make computer system understand human speech** NLP is comprised of techniques to process, structure, categorize raw text and extract information. ChatBot is a classic example of NLP, where sentences are first processed, cleaned and converted to machine understandable format.

**Nosql**: usually tabular  relations used in relational databases whereas NoSQL database provides a mechanism for storage and retrieval of data that is modeled in means of Column Document ,Key-Value,Graph, Multi-model .NoSQL means Not only SQL

**Nominal Variable:** categorical variables having two or more categories without any kind of order to them. lets take a column called “name of cities” with values such as Delhi, Mumbai, Chennai there is no order between the variables(unless explicitly mentioned).

**Normal Distribution: It is also called**  the bell curve . binomial distribution is similar to normal distribution. difference between the two is normal distribution is continuous.



**Normalization:** Normalization is used when the attributes in our data have varying scales. normalization rescales your data so that they have the same scale.

For example, if you have a variable ranging from 0 to 1 and other from 0 to 1000, you can normalize the variable, such that both are in the range 0 to 1.

**Numpy:Numpy is a** package for scientific computing with Python. It contains among other things:

* a powerful N-dimensional array object
* sophisticated (broadcasting) functions
* tools for integrating C/C++ and Fortran code
* useful linear algebra, Fourier transform, and random number capabilities
* NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined.

Numpy speedliy integrates with wide variety of databases.

**One Hot Encoding: one hot encoding is a preprocess step in which** categorical variables are converted to numerical in an interpretable format. In this we create a Boolean column for each category of the variable.

EXAMPLE:

|  |  |  |  |
| --- | --- | --- | --- |
| Sr. No. | Name |  |  |
| 1 | Vivek |  |  |
| 2 | Akshat |  |  |
| 3 | Arshad |  |  |
| This is converted as |  |  |  |
| Sr. No. | Vivek | Akshat | Arshad |
| 1 | 1 | 0 | 0 |
| 2 | 0 | 1 | 0 |
| 3 | 0 | 0 | 1 |
|  |  |  |  |

**One Shot Learning:** It is a machine learning approach where the model is trained on a single example. One-shot Learning is generally used for object classification. This is performed to design effective classifiers from a single training example.

**Oozie:** Apache Oozie is the tool in which all sort of programs can be pipelined in a desired order to work in Hadoop’s distributed environment. it also provides a mechanism to run the job at a given schedule. It has two parts:

1. **Workflow engine:**  to store and run workflows composed of Hadoop jobs like MapReduce, Pig, Hive.
2. **Coordinator engine:** It runs workflow jobs based on predefined schedules and availability of data.

Features of Oozie:

* Oozie has client API and command line interface which can be used to launch, control and monitor job from Java application.
* Using its Web Service APIs one can control jobs from anywhere.
* Oozie has provision to execute jobs which are scheduled to run periodically.

**Ordinal Variable:** Ordinal variables are those variables which have discrete values but has some order involved.

**Outlier:** an observation that appears far away and diverges from an overall pattern in a sample.

**Overfitting: it** happens when the model is too sensitive and captures random patterns which are only in the training dataset.so it performs well on the train dataset but fails on the test set. We have two methods to overcome overfitting:

* Reduce the model complexity
* Regularization

**Pandas:** pandas is a easy-to-use data structure and data analysis library for the Python programming language. it’s a open source with high performance,fast and efficient DataFrame object for data manipulation with integrated indexing.

Using pandas we can read and write data in-memory data structures and different formats: CSV and text files, Microsoft Excel, SQL databases, and the fast HDF5 format.

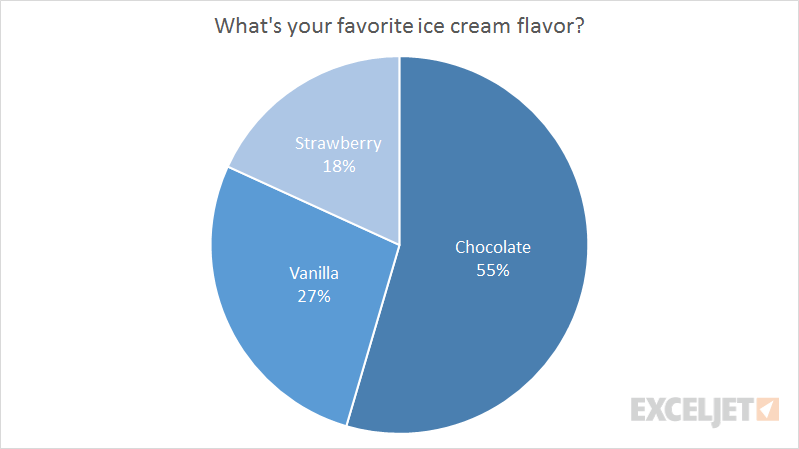
* Flexible reshaping and pivoting of data sets.
* Columns can be inserted and deleted from data structures for size mutability.
* High performance merging and joining of data sets.

**Parameters:** These are measurable factors that define a system. For machine learning models, model parameters are internal variables whose values can be determined from the data. For instance, the weights in linear and logistic regression fall under the category of parameters.

**Pattern Recognition:** focuses on the recognition of patterns and regularities in data. Classification is an example of pattern recognition where in each input value is assigned one of a given set of classes.

In computer vision, these techniques are used for optical character recognition (OCR), face detection, face recognition, object detection, and object classification.

**Pie Chart:**pie chart is a graphical representation where its divided into portions according to the inputs its easy to understand



**Pig:** it is a high level language for Hadoop used to write complex data transformations without knowing java, data manipulations can be done using this tool. Through the User Defined Functions(UDF) facility in Pig, Pig can invoke code in many languages like JRuby, Jython and Java.

Key features of Pig:

* It is able to store data at any point during a pipeline.
* It declares execution plans.
* Supports pipeline splits, thus allowing workflows to proceed along DAGs instead of strictly sequential pipelines.
* Users can create their own functions to do special-purpose processing.

**Polynomial Regression: i**n this technique, a curve fits into the data points. In a polynomial regression equation, the power of the independent variable is greater than 1. Although higher degree polynomials give lower error, they might also result in over-fitting.

**Pre-trained Model: here in this model we take a model that is pre designed for similar items instead of starting it from scratch.**

For example, if you want to build a self learning car. You can spend years to build a decent image recognition algorithm from scratch or you can take inception model (a pre-trained model) from Google which was built on ImageNet data to identify images in those pictures.

**Precision and Recall: precision is measure of** how many positives were predicted correctly represented as: Precision = TP / (TP + FP)

Whereas recall is described as the measured of how many of the positive predictions were correct ,represented as: Recall = TP / (TP + FN)

**Predictor Variable:** Predictor variable is used to make a prediction for dependent variables.

**Principal Component Analysis (PCA): it is an** approach to factor analysis that considers the total variance in the data, and transforms the original variables into a smaller set of linear combinations. PCA is sensitive to outliers; they should be removed.

it is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

PCA is used in data analysis and for making predictive models. It’s often used to visualize genetic distance and relatedness between populations.

|  |  |
| --- | --- |
| **P-Value:** P-value is the value of probability of getting a result equal to or greater  than the observed value, when the null hypothesis is true.  **Python:** Python is an open source programming language, widely used for various applications,  such as general purpose programming, data science and machine learning.  Usually preferred by beginners in these fields because of the following major advantages   * Easy to learn. * High-level language * Broadly used and supported   **PyTorch: it is a** open source machine learning library for python, based on Torch.  It is built to provide flexibility as a deep learning development platform.  It is extensively used because,its   * Easy to use API * Python support * Dynamic computation graphs   **Quartile:** Quartile divides a series into 4 equal parts. For any series, there are 4 quartiles  denoted by Q1, Q2, Q3 and Q4.  C:\Users\Reshma Sanikommu\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\B98646DB.tmp  **R**: R is an open-source programming language and a software environment for statistical computing,  machine learning, and data visualization.  Features of R:   1. It is platform independent, so it is compatible with multiple operating systems 2. R has a very strong and consistent online community support 3. The graphical capabilities of R are awesome 4. There is abundance of literature to learn R   **Range:** Range is the difference between the highest and the lowest value of the population.  It is used to measure the spread of the data. If we have a set of random numbers ,first arrange them into  Ascending order and take the difference between highest and lowest. Example: we have a list of  4,5,2,8,4,7,6,4,6,3.  So, first of all we will arrange these data points in ascending order:  2,3,4,4,4,5,6,6,7,8  Now the range of this set is the difference between the highest(8) and the lowest(2) value.  Range = 8-2 = 6  **Recommendation Engine:**  Recommendation engines basically are data filtering tools that  make use of algorithms and data to recommend the most relevant items to a particular user  it will create a positive effect on the user experience and they will visit more frequently  Nowadays in the digital age, any online shop you visit utilizes some sort of recommendation engine.  There are few types of recommendation engines:   * Content based filtering * Collaborative filtering   + User-User collaborative filtering   + Item-Item collaborative filtering * Hybrid recommendation systems   **Regression:** It is supervised learning method where the output variable is a real value,  such as “amount” or “weight”.  Example of Regression: Linear Regression, Ridge Regression, Lasso Regression  **Regression Spline:** this approach uses multiple bins and a separate model is built on each bin  instead of building one model for the entire dataset Regression Splines is a non-linear approach  that uses a combination of linear/polynomial functions to fit the data  **Regularization: regularization resovles the over fitting problem in** statistical models. In machine  learning, regularization penalizes the coefficients such that the model generalize better.  We have different types of regression techniques which uses regularization such as Ridge  regression and lasso regression.  **Reinforcement Learning: this is an example of machine learning where the machine is trained to**  Take its own decision based on the requriment and provide maximum efficiency.here the machine or the software  Trains itself continuously and applies it’s enriched knowledge to solve business problems.  This continual learning process ensures less involvement of human expertise which in turn  saves a lot of time .  Difference between Supervised Learning and Reinforcement Learning(RL) is  RL involves continuous learning by interacting with an environment An RL agent learns  from its past experience, rather from its continual trial and error learning process as against  supervised learning where an external supervisor  provides examples.  good example to understand the difference is self driving cars. Self driving cars use  Reinforcement learning to make decisions continuously like which route to take, what speed to  drive on, are some of the questions which are decided after interacting with the environment.  A simple manifestation for supervised learning would be to predict the total fare of a cab at the  end of a journey.  **Residual:** Residual of a value is the difference between the observed value and the  predicted value of the quantity of interest. Using the residual values, you can create  residual plots which are useful for understanding the model.  **Response Variable:** Response variable (or dependent variable) is that variable whose  variation depends on other variables.  Ridge Regression: Ridge regression performs ‘**L2 regularization**‘, i.e. it adds a factor of sum of  squares of coefficients in the optimization objective. Thus, ridge regression optimizes the following:  **Objective = RSS + α \* (sum of square of coefficients)**  Here, α (alpha) is the parameter which balances the amount of emphasis given to minimizing RSS  vs minimizing sum of squares of coefficients. α can take various values:   1. **α = 0:**    * The objective becomes same as simple linear regression.    * We’ll get the same coefficients as simple linear regression. 2. **α = ∞:**    * The coefficients will be zero. This is because of infinite weightage on square of    * coefficients, anything less than zero will make the objective infinite. 3. **0 < α < ∞:**    * The magnitude of α will decide the weightage given to different parts of objective.    * The coefficients will be somewhere between 0 and 1 for simple linear regression.   **ROC-AUC:** Let’s first understand what is ROC (Receiver operating characteristic) curve.  If we look at the confusion matrix, we observe that for a probabilistic model, we get different value  for each metric.  https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2018/05/Confusion_matrix.png  Hence, for each sensitivity, we get a different specificity. The two vary as follows:  https://www.analyticsvidhya.com/wp-content/uploads/2015/01/curves.png  The ROC curve is the plot between sensitivity and (1- specificity). (1- specificity) is also known  as false positive rate and sensitivity is also known as True Positive rate. Following is the ROC  curve for the case in hand.  https://www.analyticsvidhya.com/wp-content/uploads/2015/01/ROC.png  Let’s take an example of threshold = 0.5 (refer to confusion matrix). Here is the confusion matrix :  As you can see, the sensitivity at this threshold is 99.6% and the (1-specificity) is ~60%.  This coordinate becomes on point in our ROC curve. To bring this curve down to a single number,  we find the area under this curve (AUC).  Note that the area of entire square is 1\*1 = 1. Hence, AUC itself is the ratio under the curve and  the total area.  **Root Mean Squared Error (RMSE):** it is the difference between values predicted and actual values  It is the standard deviation of the residuals were Residuals are a measure of how far from the regression  line data points  The formula for RMSE is given by:  https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2018/05/rmse.png  Here,   * Predicted -> value predicted by the model * Actual -> observed values * N -> Total number of observations   **Rotational Invariance:** in mathematics, a function defined on an inner product space is said to  have rotational invariance if its value does not change when arbitrary rotations are applied to its argument.  For example, the function:  https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2018/06/RI.png  is invariant under rotations of the plane around the origin, because for a rotated set of coordinates  through any angle *θ.*  **Scala: this** combines concepts of object-oriented and functional programming languages.  Here are some key features of Scala   * Its an object-oriented language that supports many traditional design patterns * It supports functional programming which enables it to handle distributed programming at * fundamental level * It is designed to run on JVM platform that helps in directly using Java libraries * Scala can be easily implemented into existing java projects as Scala libraries can be used * within Java code * It supports first-class objects and anonymous functions   **Semi-Supervised Learning: problems were**  large amount of input data (X) and only  some of the data, is labeled (Y) are called semi-supervised learning problems.  These problems sit in between both supervised and unsupervised learning.  A good example is a photo archive where only some of the images are labeled,  (e.g. dog, cat, person) and the majority are unlabeled.  **Skewness:** skewness is a measure of symmetry. A distribution, or data set, is symmetric if it  looks the same to the left and right of the center point.  skewness - statistics  **SMOTE:** Synthetic Minority Over-Sampling Technique (SMOTE).it is an approach were classifiers  Are constructed from  imbalanced datasets is described. The idea is over-sampling the minority  (abnormal) class and under-sampling the majority (normal) class can achieve better classifier  performance (in ROC space) than only under-sampling the majority class. This is an over-sampling  approach in which the minority class is over-sampled by creating “synthetic” examples rather than  by over-sampling with replacement.  **Spatial-Temporal Reasoning:it is an area of** artificial intelligence drawn from the fields of  computer science, cognitive science, and cognitive psychology.it is the ability to mentally move  objects in space and time to solve multi-step problems. Three important things about Spatial-temporal  reasoning are:   1. It connects to mathematics at all levels, from kindergarten to calculus 2. It is innate in humans 3. Spatial-temporal reasoning abilities can be increased. This understanding of Spatial-temporal 4. reasoning forms the foundation of Spatial-temporal Math   **Standard Deviation:** Standard deviation signifies how dispersed is the data. It is the square root  of the variance of underlying data. Standard deviation is calculated for a population.  **Standardization:** Standardization is also called Z-score normalization it is the process where the features  are rescaled so that they’ll have the properties of a standard normal distribution with μ=0 and σ=1,  where μ is the mean (average) and σ is the standard deviation from the mean. Standard scores  (also called *z*scores) of the samples are calculated as follows:  https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2018/05/standardization.png  **Standard error: it is a statistical term which measure**  the accuracy of which a sample represents a population.  In statistics, a sample mean deviates from the actual mean of a population this deviation is known as  standard error.  **Statistics:** It is the study of the collection, analysis, interpretation, presentation, and organisation of data.  **Stochastic Gradient Descent: it is an** algorithm where we take a sample of data while computing  the gradient. update to the coefficients is performed for each training instance, rather than at the  end of the batch of instances. learning will be much faster for very large training datasets and  only need a small number of passes through the dataset to reach a good or good enough set of coefficients.  **Supervised Learning: it is an algorithm given with a set of independent variables and predicting the out come**  **Of dependent variable or output variable** Using these set of predictors, we generate a function that  map inputs to desired outputs. Like: y= f(x)  Examples of Supervised Learning algorithms: Regression, Decision Tree, Random Forest, KNN,  Logistic Regression etc.  **SVM: in this algorithm each datapoint is plotted as an item in n-**dimensional space .each point  has two coordinates called support vectors then we will find some line that splits the data between the  two differently classified groups of data. This will be the line such that the distances from the  closest point in each of the two groups will be farthest away. |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |