Data Science with R

Manav Madan Rawal

20PCM53

Introduction to R:

R is an object oriented programming (OOP) language developed by statisticians Ross Ihaka and Robert Gentleman for statistics. R is now widely used by Data Miners, bioinformaticians and statisticians for data analysis and developing statistical software.

One of the major pros of using R over languages like C++ is the large number of packages developed by the R community which considerably shorten the length and complexity of code needed to run a program.

R can be run in various development environments like Microsoft Visual Studio Code, JetBrains Data Spell and RStudio. For the purpose of this course, we will be using RStudio.

Packages:

R packages are a collection of R functions, complied code and sample data. They are stored under a directory called **"library"** in the R environment. By default, R installs a set of packages during installation. More packages are added later, when they are needed for some specific purpose. All the packages of R can be found and downloaded from CRAN (Comprehensive R Archive Network).

A package can be installed in R using the following command:

install.packages(“package name”, dependencies = TRUE)

Modules:

Modules act as an organizational unit for source code. Modules enforce to be more rigorous when defining dependencies and have a local search path. They can be used as a sub unit within packages or in scripts. For convenience sake let's think of them as being a collection of packages which serve similar purposes.

Given below are a few common packages along with their modules used in Data Science:

Module	**Package**	**Purpose**
R Programming	plyr	A set of tools that solves a common set of problems: you need to break a big problem down into manageable pieces, operate on each piece and then put all the pieces back together.
R Programming	dplyr	A fast, consistent tool for working with data frame like objects, both in memory and out of memory. It is a structure of data manipulation that provides a uniform set of verbs, helping to resolve the most frequent data manipulation hurdles.

R Programming	reshape2	Flexibly restructure and aggregate data using just two functions: melt and 'dcast' (or 'acast').
R Programming	sqldf	An R package for running SQL statements on R data frames, optimized for convenience
R Programming	ggplot2	ggplot2 is a plotting package also termed as Grammar of Graphics. It is a free, open-source, and easy-to-use visualization package widely used in R.
R Programming	ggmap	ggmap is an R package that makes it easy to retrieve raster map tiles from popular online mapping services like Google Maps and Stamen Maps and plot them using the ggplot2 framework.

R Programming	GGally	'GGally' extends 'ggplot2' by adding several functions to reduce the complexity of combining geometric objects with transformed data.
R Programming	gcookbook	Data sets used in the book "R Graphics Cookbook" by Winston Chang
R Programming	scales	Scale() Function in R, Scaling is a technique for comparing data that isn’t measured in the same way. The normalizing of a dataset using the mean value and standard deviation is known as scaling.
Statistical Modelling	visualize	Graphs the pdf or pmf and highlights what area or probability is present in user defined locations. Visualize is able to provide lower tail, bounded, upper tail, and two tail calculations. Supports strict and equal to inequalities.

Statistical Modelling	EnvStats	Graphical and statistical analyses of environmental data, with focus on analyzing chemical concentrations and physical parameters, usually in the context of mandated environmental monitoring.
Statistical Modelling	rMR	Analysis of oxygen consumption data generated by Loligo (R) Systems respirometry equipment.
Data Preparation	sampling	Functions to draw random samples using different sampling schemes are available. Functions are also provided to obtain (generalized) calibration weights, different estimators, as well some variance estimators.
Data Preparation	mice	Multivariate Imputation by Chained Equations
Data Preparation	pwr	Basic Functions for Power Analysis

Data Preparation	survival	Contains the core survival analysis routines, including definition of Surv objects, Kaplan-Meier and Aalen-Johansen (multi-state) curves, Cox models, and parametric accelerated failure time models.
Data Preparation	FrFr2	Regular and non-regular Fractional Factorial 2-level designs can be created.
Data Preparation	DoE.base	Creates full factorial experimental designs and designs based on orthogonal arrays for (industrial) experiments.
Data Preparation	MASS	Functions and datasets to support Venables and Ripley, "Modern Applied Statistics with S"
Data Preparation	fitdistrplus	Extends the fitdistr() function (of the MASS package) with several functions to help the fit of a parametric distribution to non-censored or censored data.

Data Preparation	car	Functions to Accompany J. Fox and S. Weisberg, An R Companion to Applied Regression
Data Preparation	predictmeans	Providing functions to diagnose and make inferences from various linear models, such as those obtained from 'aov', 'lm', 'glm', 'gls', 'lme', and 'lmer'. Inferences include predicted means and standard errors, contrasts, multiple comparisons, permutation tests and graphs.
Data Preparation	caret	Misc functions for training and plotting classification and regression models.

Data Preparation	e1071	Functions for latent class analysis, short time Fourier transform, fuzzy clustering, support vector machines, shortest path computation, bagged clustering, naive Bayes classifier, generalized k-nearest neighbour ...
Linear Algebra	lpSolve	Lp_solve is freely available (under LGPL 2) software for solving linear, integer and mixed integer programs.
Linear Algebra	pracma	Provides a large number of functions from numerical analysis and linear algebra, numerical optimization, differential equations, time series, plus some well-known special mathematical functions.

Linear Algebra	SparseM	Some basic linear algebra functionality for sparse matrices is provided: including Cholesky decomposition and backsolving as well as standard R subsetting and Kronecker products.
Linear Algebra	Matrix	A rich hierarchy of matrix classes, including triangular, symmetric, and diagonal matrices, both dense and sparse and with pattern, logical and numeric entries.
Linear Algebra	MatrixModels	Modelling with sparse and dense 'Matrix' matrices, using modular prediction and response module classes.

Predictive Modelling	pbkrtest	Parametric Bootstrap, Kenward-Roger and Satterthwaite Based Methods for Test in Mixed Models
Predictive Modelling	car	Functions to Accompany J. Fox and S. Weisberg, An R Companion to Applied Regression
Predictive Modelling	alr3	Data to Accompany Applied Linear Regression
Predictive Modelling	caTools	Contains several basic utility functions including: moving (rolling, running) window statistic functions, read/write for GIF and ENVI binary files, fast calculation of AUC, LogitBoost classifier, base64 encoder/decoder, round-off-error-free sum and cumsum, etc.
Predictive Modelling	leaps	Regression subset selection, including exhaustive search.

Machine Learning	e1071	Functions for latent class analysis, short time Fourier transform, fuzzy clustering, support vector machines, shortest path computation, bagged clustering, naive Bayes classifier, generalized k-nearest neighbour ...
Machine Learning	mice	Multivariate Imputation by Chained Equations
Machine Learning	caret	Misc functions for training and plotting classification and regression models.
Machine Learning	party	A computational toolbox for recursive partitioning.
Machine Learning	rpart	Recursive partitioning for classification, regression and survival trees.
Machine Learning	rpart.plot	Plot 'rpart' models. Extends plot.rpart() and text.rpart() in the 'rpart' package.
Machine Learning	tree	Classification and regression trees.

Machine Learning	MASS	Functions and datasets to support Venables and Ripley, "Modern Applied Statistics with S"
Machine Learning	lars	Efficient procedures for fitting an entire lasso sequence with the cost of a single least squares fit. Least angle regression and infinitesimal forward stagewise regression.
Machine Learning	stats	Statistics operations
Machine Learning	pls	Multivariate regression methods Partial Least Squares Regression (PLSR), Principal Component Regression (PCR) and Canonical Powered Partial Least Squares (CPPLS).
Machine Learning	randomForest	Classification and regression based on a forest of trees using random inputs.

Data Science:

Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns,

derive meaningful information, and make business decisions. Data science uses complex machine learning algorithms to build predictive models.

It can be divided into two sub fields.

i) **Data Analytics**
ii) **Data Analysis**

Data Analytics	**Data Analysis**
Data analytics is 'general' form of analytics that is used in businesses to make decisions from data that are data-driven.	Data analysis is a specialized form of data analytics used in businesses to analyze data and take some insights into it.
Data analytics consists of data collection and inspection in general and it has one or more users.	Data analysis consisted of defining data, investigating, cleaning, and transforming the data to give a meaningful outcome.
Data Analytics, in general, can be used to find masked patterns, anonymous correlations, customer preferences, market trends, and other necessary information that can help to make more notify decisions for business purposes.	Data analysis can be used in various ways one can perform analysis like descriptive analysis, exploratory analysis, inferential analysis, predictive analysis, and take useful insights from the data.
Let's say you have 1gb customer purchase-related data for the past 1 year, now one has to find what our customer's next possible purchases are, you will use data analytics for that.	Suppose you have 1gb customer purchase-related data of the past 1 year and you are trying to find what happened so far that means in data analysis we look into the past.

Challenges in Data Science:

i) **Data Authenticity**
ii) **Computational Complexity**
iii) **Data Security**

Q.) What is Data?

Ans.) The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media is known as data.

Data can be divided into two types:

i) Structured Data – Excel Files, JSON (JavaScript Object Notation) files, CSV (Comma Separated Value) files, etc.
ii) Unstructured Data – Words, sentences, noise, etc.

Machine Generated Data:

Machine-generated data is information automatically generated by a computer process, application, or other mechanism without the active intervention of a human.

Examples of this would be the data generated by a compiler or Interpreter after a program has been executed.

Image Data:

Image contains information made up of pixels.

Raw file converted to compresses file like .jpeg or .png

In Greyscale: 0 – Pure black

255 – Pure white

Image in matrix format:

$$\begin{matrix} 0 & 0 & 0 \\ 0 & 255 & 0 \\ 0 & 0 & 0 \end{matrix}$$

The above matrix represents a Greyscale image

In a coloured image, each position has 3 colours –RGB

$$\begin{matrix} 0,0,0 & 0,0,0 & 0,0,0 \\ 0,0,0 & 255,255,255 & 0,0,0 \\ 0,0,0 & 0,0,0 & 0,0,0 \end{matrix}$$

Video Data:

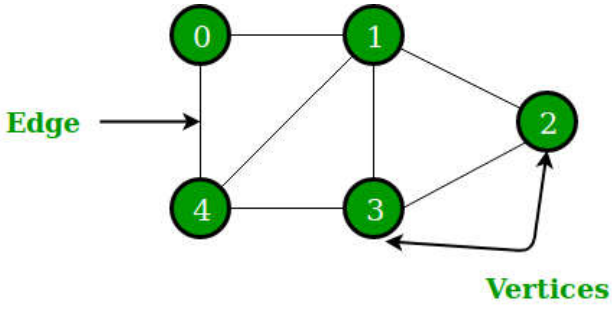
Collection of Frames is known as Video data.

Streaming Data:

Netflix, YouTube, etc.

Graph Data:

Facebook, LinkedIn, etc.



Natural Language:

A Special kind of unstructured data.

A topic of research in AI domains like sentiment analysis, spam filters, etc.

Machine Learning:

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed.

Machine Learning (ML) has three types of algorithms:

i) **Supervised Learning**
ii) **Unsupervised Learning**
iii) **Reinforcement Learning**

Supervised Learning:

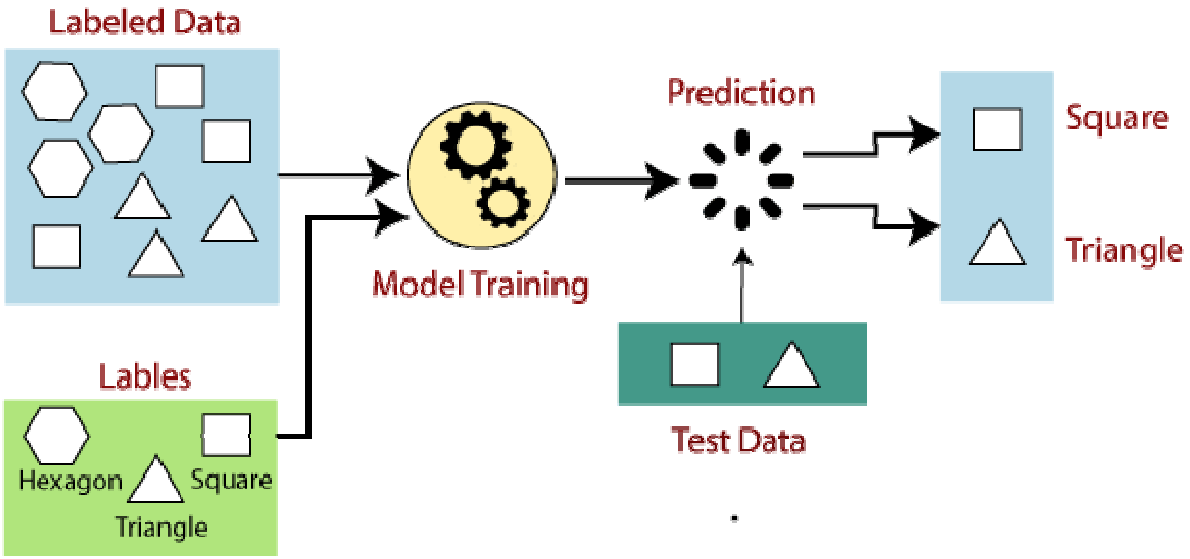
Uses Structured Data.

Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.

Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to **find a mapping function to map the input variable(x) with the output variable(y)**.

In the real-world, supervised learning can be used for **Risk Assessment, Image classification, Fraud Detection, spam filtering**, etc.

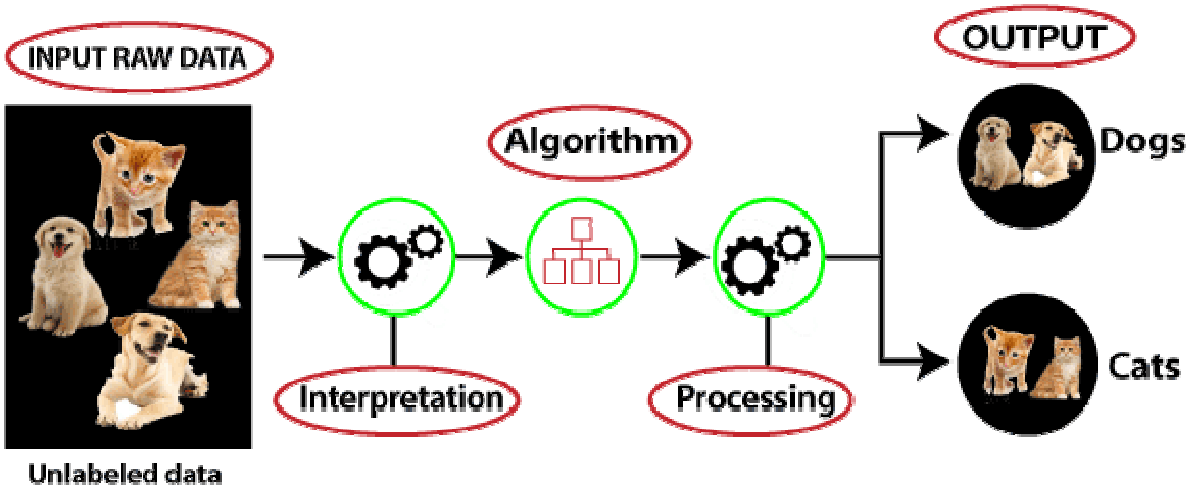


Unsupervised Learning:

Uses Unstructured data.

As the name suggests, unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead, models itself find the hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things.

Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to **find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format**.

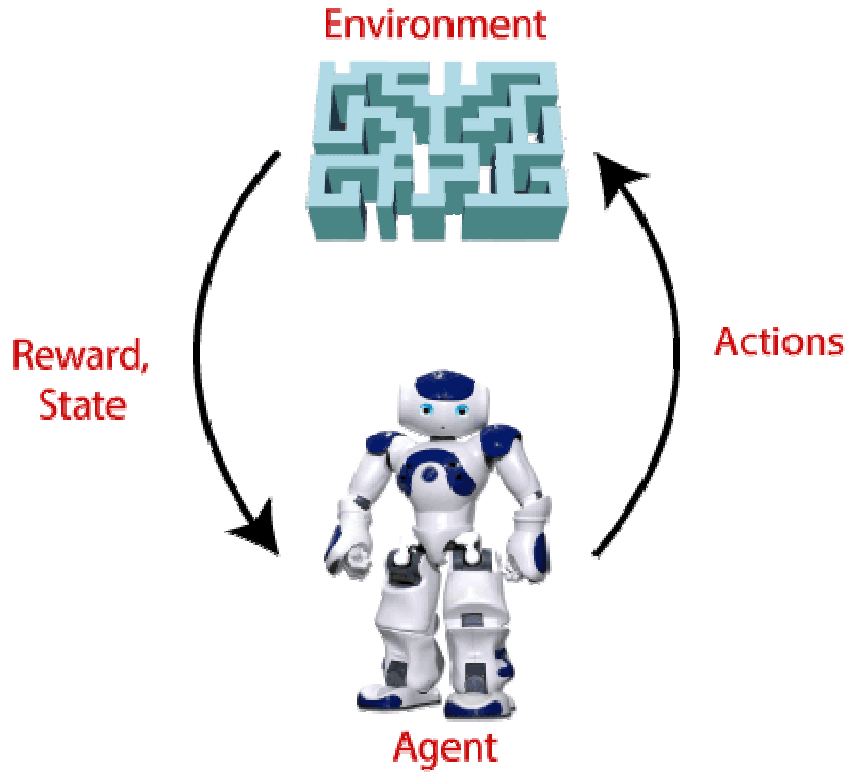


Reinforcement Learning:

Reinforcement Learning is a feedback-based Machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions. For each good action, the agent gets positive feedback, and for each bad action, the agent gets negative feedback or penalty.

In Reinforcement Learning, the agent learns automatically using feedbacks without any labeled data, unlike supervised learning.

An example of this would be driverless cars.



Data Science Process:

Data ---------> Business Problems --------> Data Cleaning --------> EDA

Where EDA stands for Exploratory Data Analysis.

Business problems can be divided into:

i) Continuous Problems
ii) Categorical Problems

Continuous Variable:

Continuous variables are numeric variables that have an infinite number of values between any two values. A continuous variable can be numeric or date/time. For example, the length of a part or the date and time a payment is received.

Categorical Variable:

Categorical variables contain a finite number of categories or distinct groups. Categorical data might not have a logical order. For example, categorical predictors include gender, material type, and payment method.

The dependent variable is the output, while the Independent variable is the Input.

If the Variable is Continuous, it's a Regression Problem; If the Variable is Categorical, it's a classification problem.

Primary Key:

A primary key is the column or columns that contain values that uniquely identify each row in a table.

Primary key should always be Unique. All Structured data should have a primary key.

Correlation:

In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data.

When we look at data, we come across three types of Correlations:

i) Positive Correlation
ii) Negative Correlation
iii) No Correlation

Positive Correlation:

A positive correlation is a relationship between two variables such that their values increase or decrease together.

Negative Correlation:

A negative correlation is a relationship between two variables such that as the value of one variable increases, the value of the other one decreases and vice versa.

No Correlation:

No Correlation is a relationship between two variables such that the increase or decrease in value of one variable has no effect on the other.

If the values of Correlation are very small and close to zero, we consider it to be No Correlation. For example Correlation values of 0.05 and 0.005 will be considered as No Correlation.

A Value of 0 is perfect No Correlation or Zero Correlation.

The Correlation values for any problem lie in the range [-1, 1].

If Cor = 1, It's a Perfect Positive Correlation.

If Cor = -1, It's a Perfect Negative Correltion.

Note: It's not always the case that a very small positive or negative correlation is No Correlation. It depends on the situation, therefore

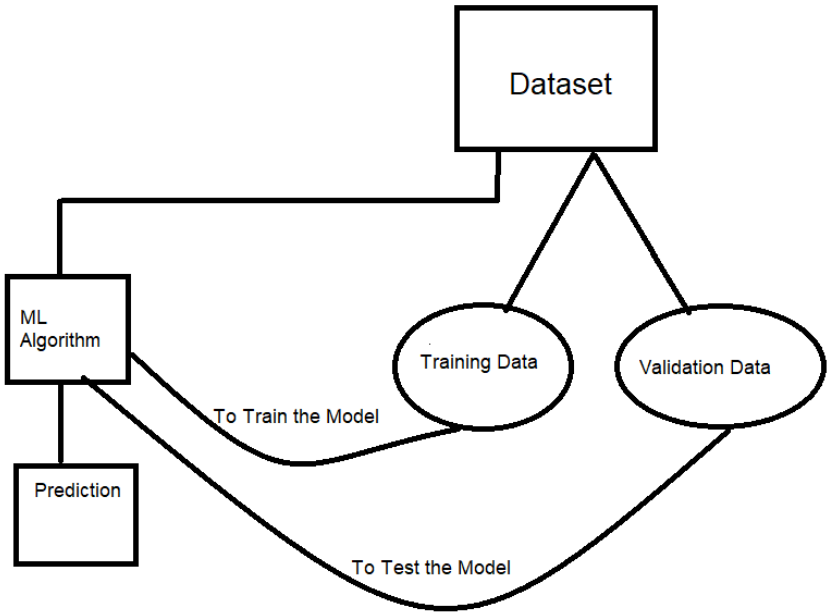
one should look at that particular dataset and it's conditions before deciding whether to consider a small number or not.

For example Cor = 0.2 is a positive Correlation but not always. If other statistical values are higher, this will be zero correlation and vice versa.

Model:

A Model is a Mathematical Equation. It may or may not be a Machine Learning Model.

An ML Model is a Mathematical equation to predict the future.



Validation Techniques:

Validation techniques in machine learning are used to get the error rate of the ML model.

Different Validation techniques are used depending on whether the data has Continuous or Categorical Variables.

Note: **Error = Actual Value – Predicted Value**

For Continuous Variables/ Regression Problems:

i) Mean Square Error(MSE)

Formula

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error
n = number of data points
Y_i = observed values
$\hat{Y}_i$ = predicted values

ii) Root Mean Square Error(RMSE)

Formula

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \hat{x}_i)^2}{N}}$$

RMSD = root-mean-square deviation
i = variable i
N = number of non-missing data points
x_i = actual observations time series
$\hat{x}_i$ = estimated time series

iii) Mean Absolute Error(MAE)

Formula

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

MAE = mean absolute error
y_i = prediction
x_i = true value
n = total number of data points

iv) Mean Absolute Percentage Error(MAPE)

$$M = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

M = mean absolute percentage error

n = number of times the summation iteration happens

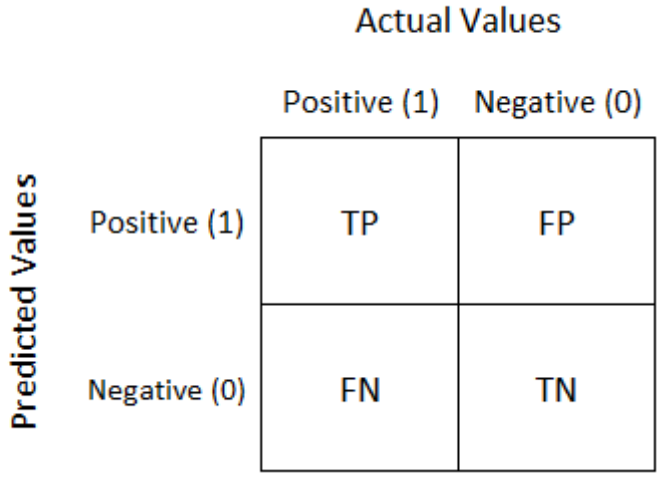
A_t = actual value

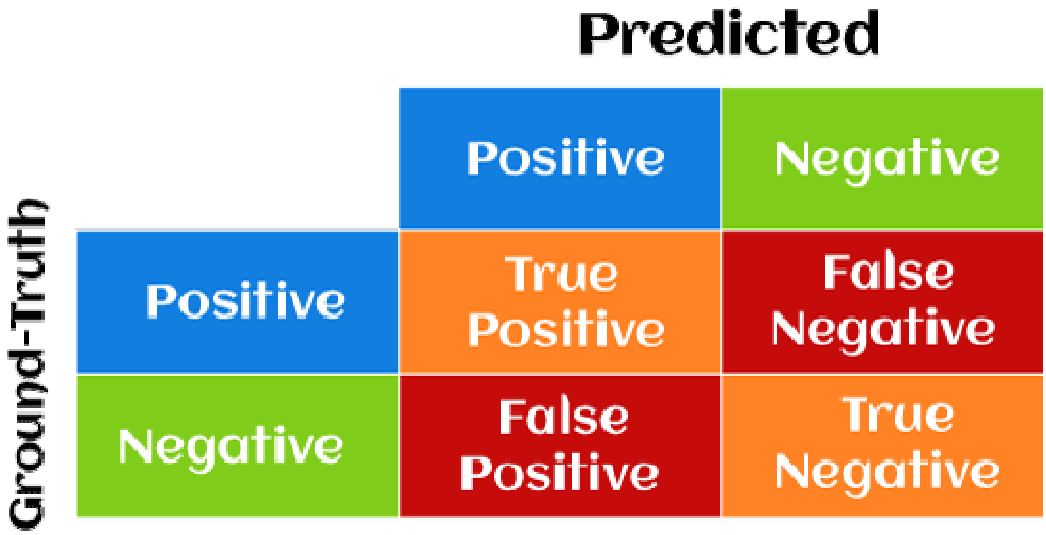
F_t = forecast value

For Categorical Variables/ Classification Problems:

i) Confusion Matrix

A way to evaluate the performance of a classifier is to look at the confusion matrix. The general idea is to count the number of times instances of class A are classified as class B.





Where,

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

Accuracy: $\frac{TP+TN}{Total\ no.\ of\ Observations}$

Example:

Calculate the accuracy from the table given below

Values	0	1
0	80	50
1	30	40

As we know, Accuracy = $\frac{TP+TN}{Total\ no.\ of\ Observations}$

= (80+40/200) X 100%

= 0.6 X 100% => 60%

TPR – True Positive Rate/ Recall:

TPR = $\frac{TP}{TP+FN}$

Specificity or TNR – True Negative Rate:

TNR = $\frac{TN}{FP+TN}$

Precision:

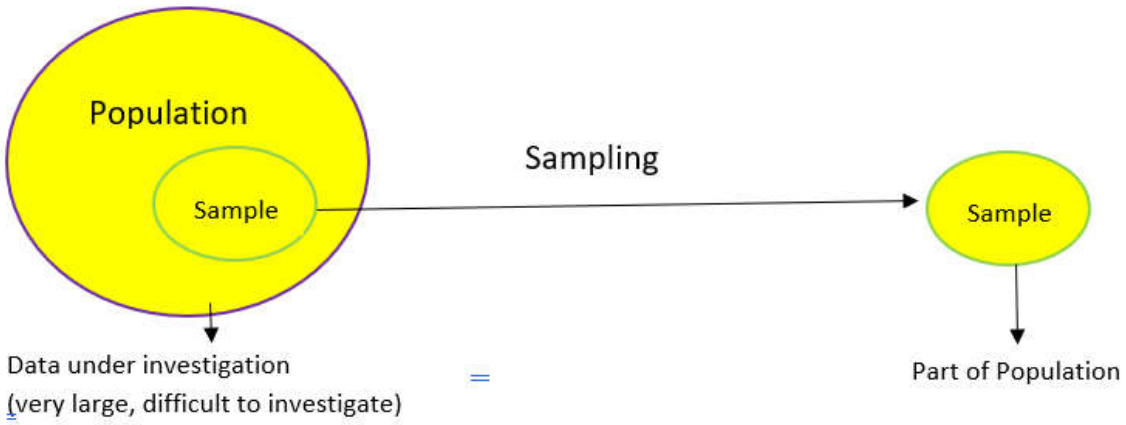
Precision = $\frac{TP}{FP+TN}$

F_1 Score: (can be from 0 to 1)

F_1 Score = $\frac{2\ X\ Recall\ X\ Precision}{Recall+Precision}$

Sampling Techniques:

Sampling is a method that allows us to get information about the population based on the statistics from a subset of the population (sample), without having to investigate every individual.



There are many Sampling techniques, let's look at some of them.

i) Random Sampling
ii) Systematic Sampling
iii) Stratified Sampling
iv) Purposive Sampling

Random Sampling:

Random sampling is a part of the sampling technique in which each sample has an equal probability of being chosen. A sample chosen randomly is meant to be an unbiased representation of the total population.

Disadvantage: Not enough samples for each group.

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 but with a fixed, periodic interval. This interval, called the sampling interval, is calculated by dividing the population size by the desired sample size.

You can increase or decrease the time interval but it should be common for all samples.

e.g: 1,2,3,4,5,6,7,8,9 – time interval of one.

Stratified Sampling:

Stratified random sampling is a method of sampling that involves the division of a population into smaller sub-groups known as strata. In stratified random sampling, or stratification, the strata are formed based on members' shared attributes or characteristics.

Advantage: You get more accurate results because you take samples from all groups.

Purposive Sampling:

Selecting samples according to our purpose is called Purposive Sampling.

Deployment/Automation/Presentation:

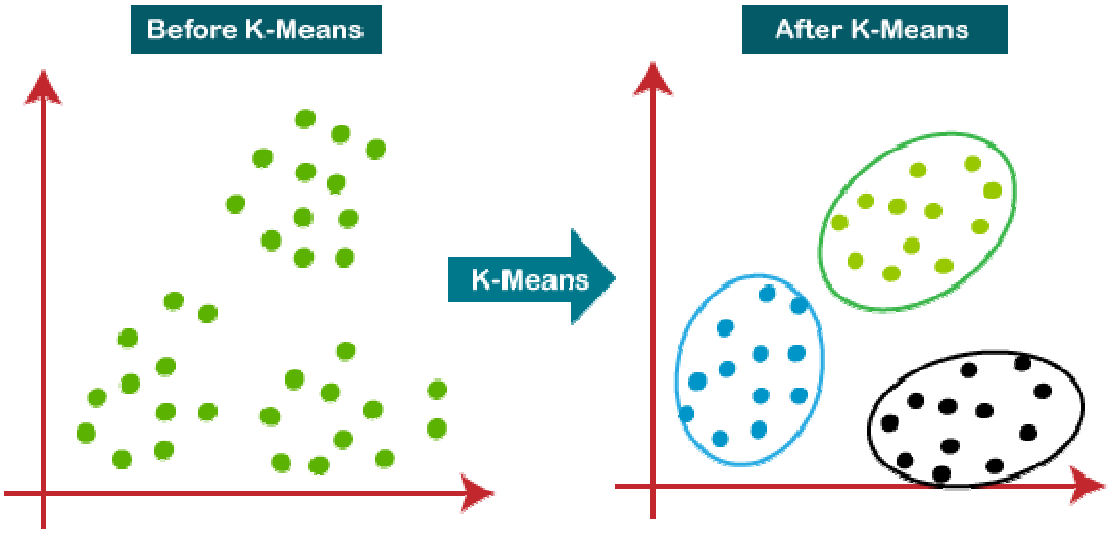
Deployment is the method by which you integrate a machine learning model into an existing production environment to make practical business decisions based on data. This is one of the last stages of a machine learning cycle.

Tests for Checking Correlation:

i) Correlation between Continuous and Continuous variables – Karl Pearson Test
ii) Correlation between Categorical and Categorical variables – Chi Square Test
iii) Correlation between Continuous and Categorical variables – Anova Test

K Means Clustering Algorithm:

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. It groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.



The working of the K-Means algorithm is explained in the below steps:

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other than the input dataset).

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

Step-4: Calculate the mean and place a new centroid of each cluster.

Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.

Step-7: The model is ready.

K means Clustering algorithm can be 1D or 2D. In 1D, you just have a tuple or a list of data and you have to cluster them into groups. Let's look at a few examples.

Eg.1) {2,4,6,8,15,20,25,35,50}

Now we randomly select the centroids or mid values.

Let's choose 8 as a centroid for group 1 and 25 as a centroid for group 2. Now we calculate the Euclidean distance between each element and the two centroids, whichever distance is lesser, that element is grouped into that group.

Note: In 1D K Means Clustering Euclidean Distance is just the absolute value of centroid-element.

Group 1	Group 2
8 (Randomly selected Centroid)	25 (Randomly selected Centroid)
2,4,6,8,15	20,25,35,50
7 – New Mid value as the mean	32 – New Mid value as the mean
2,4,6,8,15 – Final Cluster 1	20,25,35,50 – Final Cluster 2

You have to keep repeating the process until two successive Clusters are the same.

Here K =2 as two groups were formed

Eg.2) {7,49,-2,69,33,77,16,91,4,10}

G1	G2	G3
-2(Randomly selected Centroid)	77(Randomly selected Centroid)	4(Randomly selected Centroid)
No value fits	49,69,91	7,33,16,10
-2 – New Mid value as the mean	72 – New Mid value as the mean	14 – New Mid value as the mean
	49,69,91	7,3,16,10

In the above K=3 as three groups were formed.

2D K means Clustering:

In 2 Dimensional K means Clustering, you have data which is divided into columns or rows and then you have to cluster it. Here we use the Euclidean Distance Formula to find the distance.

Distance = $\sqrt{(x1 - x2)^2 + (y1 - y2)^2}$

Let's look at a few examples.

Eg.1)

Income	Spending
40	20
50	15
60	25
20	35
65	35
80	19

K =2

Since K is given as 2, we have to form two groups or clusters.

We form two clusters and then choose two values, one from each column as the centroid of that group. So let's take (40,20) as the centroid of group 1 and (20,35) as the centroid of group 2. Now we calculate the Euclidean distance between these centroids and some other pair like (50,15). After calculating the Euclidean distances, whichever is less, that pair will be grouped in that cluster. So in this case the Euclidean distance is lesser with the centroid of group 1, so (50,15) will be in group 1.

Now whenever, we add a pair to a group, we need to calculate a new centroid for that group, so we take the mean of 40 and 50 and 20 and 15. So (45,18) will be the new centroid of group 1 while group 2 remains unchanged and the process is repeated.

Linear Regression:

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

lm = Linear Regression (for Regression Problem)

glm = logistical Regression (for Classification Problem)

y = mx +c

m = $\sum \frac{(x-\bar{x})\,(y-\bar{y})}{(x-\bar{x})^2}$

Error	yp	x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
0.2	2.8	1	3	-2	-0.6	4	-1.2
0.8	3.2	2	4	-1	0.4	1	-0.4
1.6	3.6	3	2	0	-1.6	0	0
0	4.0	4	4	1	0.4	1	0.4
0.6	4.4	5	5	2	1.4	4	2.8
		$\bar{x}$=3	$\bar{y}$=3.6			∑ = 10	∑ = 4

m = 4/10 = 0.4

y = mx+c

3.6 = (0.4)(3) + C

c = 2.4

yp = (0.4)(1) + 2.4

y1 = 2.8

y2 = 3.2

y3 = 3.6

y4 = 4.0

Mean Square Error = 0.4096

Points to Remember:

Characteristics of Data:

i) Variety
ii) Velocity
iii) Volume

Applications of Data Science:

i) Forecasting
ii) Cancer Prediction
iii) Human Genome Sequencing
iv) Stock Price Prediction

Forward pipe operator : %>%

Types of Data Structures in R:

i) Vector
ii) Matrix
iii) Dataframe
iv) List

R fills values in a matrix Columnwise.

To fill missing values :

i) Categorical variables = mode
ii) Continuous variables = mean and median:

If outliers are present = median

Else use mean.

Function for predicting output: predict()

Function for dataframe: data.frame()

Checking the first six rows: head()

Checking the last six rows: tail()

Operator to use a particular column: %

Extension of R files: .R

Exercise:

1) What will the following code do?

```
getwd()
```

a) It will execute the file in Terminal
b) It will get the current working directory
c) It will display an error message
d) It will open the file with windows Explorer

Answer: b)

2) What will the following code do?

```
Setwd("C:/Users/ADMIN/Desktop/DataScience")
```

a) a) It will set the Working directory to C://Users/Admin/Desktop/DataScience
b) b) It will give an error
c) c) It will display C://Users/Admin/Desktop/DataScience inside RStudio
d) d) It will execute C://Users/Admin/Desktop/DataScience in Terminal

Answer: b) Reason is S is supposed to be lowercase in setwd()

3) Is the following Snippet of Code legal under the Syntax of R?

```
dim(comp_data)
```

a) Yes, it's syntactically legal
b) No, it's illegal as there's no identifier
c) Data Insufficient
d) R has a variable syntax so it doesn't matter

Answer: a)

4) Will the following code Execute?

```
comp_data=read.csv("Computer_Data")
```

a) Yes, it will
b) No, it won't
c) Data Insufficient
d) None of the above

Answer: b) Reason is file doesn't have an extension

5) Which Characteristic of data takes the most time to process?
   a) Variety
   b) Volume
   c) Volume and Variety take the same time but velocity takes less
   d) All take the same time

   Answer: a)

6) What do the following lines of code do?

```
results1 <- table(Prediction,
validation_data$Survived)
```

   a) It will give an error as the syntax is incorrect
   b) It takes the value of result1 and puts it into a function called table with parameters of Prediction and Validation_data$Survived
   c) It takes validation data from column survived, passes it to a function called Prediction, the output of which is stors in table form in result1
   d) It takes the value of table and puts it into a function called result1 with parameters of Prediction and Validation_data$Survived

Answer: c)

7) Is the following syntactically legal?

```
write.csv(output, file =
'C:/Users/Madan/OneDrive/Desktop/Titanic-
Survival-Prediction-Using-R-
master/Results/NaiveBayesResults.csv',
row.names = F)
```

a) Yes, it's syntactically legal

b) No, it's illegal as there's no identifier

c) Data Insufficient

d) No it's illegal as row.names isn't a valid function

Answer: a)

All the code can be found at
https://github.com/Manav02012002/DataScience_R