MALLA REDDY INSTITUTE OF TECHNOLOGY & SCIENCE
(SPONSORED BY MALLA REDDY EDUCATIONAL SOCIETY)
Permanently Affiliated to JNTUH & Approved by AICTE, New Delhi
NBA Accredited Institution, An ISO 9001:2015 Certified, Approved by UK Accreditation Centre
Granted Status of 2(f) & 12(b) under UGC Act. 1956, Govt. of India.

DATA SCIENCE

COURSE FILE

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

(CYBER SECURITY)

(2022-2023)

Faculty In-Charge
R. Sony

HOD-CSE
Dr. Madhusekhar

DATA SCIENCE (Professional Elective – II)

B.Tech. III Year I Sem.

L	T	P	C
3	0	0	3

Course Objectives

1. To learn concepts, techniques and tools they need to deal with various facets of data science practice, including data collection and integration
2. To exploring data analysis, predictive modeling, descriptive modeling, data product creation, evaluation, and effective communication
3. To understand the basic knowledge of algorithms and reasonable programming experience and some familiarity with basic linear algebra and basic probability and statistics
4. To identify the importance of recommendation systems and data visualization techniques

Course Outcomes

1. Understand basic terms what Statistical Inference means. Identify probability distributions commonly used as foundations for statistical modeling. Fit a model to data
2. Discuss the significance of exploratory data analysis (EDA) in data science and to apply basic tools (plots, graphs, summary statistics) to carry out EDA
3. Apply basic machine learning algorithms and to identify common approaches used for Feature Generation
4. Analyze fundamental mathematical and algorithmic ingredients that constitute a Recommendation Engine and to Build their own recommendation system using existing components

UNIT - I:
Introduction: What is Data Science? - Big Data and Data Science hype – and getting past the hype - Why now? – Datafication - Current landscape of perspectives - Skill sets needed - Statistical Inference - Populations and samples - Statistical modeling, probability distributions, fitting a model - Intro to R

UNIT - II:
Exploratory Data Analysis and the Data Science Process - Basic tools (plots, graphs and summary statistics) of EDA - Philosophy of EDA - The Data Science Process - Case Study: Real Direct (online real estate firm) - Three Basic Machine Learning Algorithms, Linear Regression - k-Nearest Neighbors (k-NN) - k-means

UNIT - III:
One More Machine Learning Algorithm and Usage in Applications - Motivating application: Filtering Spam - Why Linear Regression and k-NN are poor choices for Filtering Spam - Naive Bayes and why it works for Filtering Spam

UNIT - IV:
Data Wrangling: APIs and other tools for scrapping the Web - Feature Generation and Feature Selection (Extracting Meaning From Data) - Motivating application: user (customer) retention - Feature Generation (brainstorming, role of domain expertise, and place for imagination) - Feature Selection algorithms – Filters; Wrappers; Decision Trees; Random Forests

UNIT - V:
Data Visualization - Basic principles, ideas and tools for data visualization 3 - Examples of inspiring (industry) projects - Exercise: create your own visualization of a complex dataset - Data Science and Ethical Issues - Discussions on privacy, security, ethics - A look back at Data Science - Next-generation data scientists

TEXT BOOKS:

1. Doing Data Science, Straight Talk From The Frontline. Cathy O'Neil and Rachel Schutt, O'Reilly, 2014
2. Mining of Massive Datasets v2.1, Jure Leskovek, Anand Rajaraman and Jeffrey Ullman, Cambridge University Press, 2014
3. Machine Learning: A Probabilistic Perspective, Kevin P. Murphy, 2013 (ISBN 0262018020)

REFERENCE BOOKS:

1. Elements of Statistical Learning, Trevor Hastie, Robert Tibshirani and Jerome Friedman, 2nd Edition, 2009 (ISBN 0387952845)
2. Foundations of Data Science, Avrim Blum, John Hopcroft and Ravindran Kannan
3. Data Mining and Analysis: Fundamental Concepts and Algorithms, Mohammed J. Zaki and Wagner Miera Jr. Cambridge University Press, 2014
4. Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber and Jian Pei, 3rd Edition, 2011 (ISBN 0123814790)

What is DATA SCIENCE?

Data science is an inter disciplinary field that uses scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from noisy, structured and unstructured data, and apply knowledge from data across a broad range of application domains. Data science is related to data mining, machine learning and big data.

Data science is a "concept to unify statistics, data analysis, informatics, and their related methods" in order to "understand and analyses actual phenomena" with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge.

Data science is a deep study of the massive amount of data, which involves extracting meaningful insights from raw, structured, and unstructured data that is processed using the scientific method, different technologies, and algorithms.

It is a multidisciplinary field that uses tools and techniques to manipulate the data so that you can find something new and meaningful.

In short, we can say that data science is all about:

- Asking the correct questions and analyzing the raw data.
- Modeling the data using various complex and efficient algorithms.
- Visualizing the data to get a better perspective.
- Understanding the data to make better decisions and finding the final result.

Example:

Let suppose we want to travel from station A to station B by car. Now, we need to take some decisions such as which route will be the best route to reach faster at the location, in which route there will be no traffic jam, and which will be cost-effective. All these decision factors will act as input data, and we will get an appropriate answer from these decisions, so this analysis of data is called the data analysis, which is a part of data science.

HOW THIS DATA SCIENCE WORKS?

Problem Statement

To build a data science model or utilize a machine learning algorithm, you will need to understand what the problem is. This step can also be called something more along the lines of a '*business use case*'. In this step, you will most likely experience working with stakeholders the most, anyone from data analysts, business analysts, product managers, to your company's senior executives.

Here is an example of a bad problem statement:

- *"We wanna predict how many people will buy our product in the year 2022"*

Here is an example of a good problem statement:

- *"The current way of predicting sales is inaccurate"*

While the first example makes sense, it does not highlight the problem, it highlighted a possible solution instead. The focus first should always be to understand the problem in its simplest form. From there, we can then present possible solutions using data science techniques and models.

Another part of the problem statement can be the process of defining goals. For example, it will be useful to ask what the current sales prediction accuracy is, the goal accuracy, and hopefully, if the model can or cannot reach that goal accuracy, and what it means to not reach it exactly.

Data Collection

Regarding the holistic data science process described in this article, the data collection process is perhaps the furthest removed step from academia to professional environments. As an example, in an educational course, you may be given a dataset right away that is already processed and explored. For working environments or professional settings, you will have to learn how to acquire that data from an outside source or an internal source within a data table. This step can take quite a bit of time as you will need to explore nearly all data tables in your database, or across databases. The data that you ultimately acquire might use different data from various sources. The final data will eventually be read into a data frame so that it can then be analyzed, trained, and predicted on.

Here are some possible ways of acquiring data:

- from CSV files
- from Google Sheets
- from Salesforce
- JSON files
- database tables
- from other websites
- and much more

Cleaning Data

- The next step is to clean the data, referring to the scrubbing and filtering of data. This procedure requires the conversion of data into a different format. It is necessary for processing and analyzing of information. If the files are web locked, then it is also needed to filter the lines of these files. Moreover, cleaning data also constitute withdrawing and replacing values. In case of missing data sets, the replacement must be done properly, since they could look like non-values. Additionally, columns are split, merged, and withdrawn as well.

- The data we use will determine our model's reliability, so this phase is time-consuming but is also the most important. One can effortlessly use the data from this phase moving forward.

Exploratory Data Analysis

This step in the data science process can generally follow the same format. At this point, you will have your main, single dataframe. For the sake of the data science problem, you will need to separate your X features versus your y target variable — what you are trying to predict. These features can span from one to hundreds, or even more, but it is best to start off simple and analyze the main features of your dataset first (*the ones you intuitively expect to be significant to the model's prediction*), and then get a glimpse of all of the features next.

You can look at a variety of descriptive statistics that can help to define your data, here are some of the easier and more common ways to describe your data — oftentimes with the pandas library:

- df[['feature_1', 'feature_n']].head() — first 5 rows of your data
- df[['feature_1', 'feature_n']].tail() — last 5 rows of your data
- df[['feature_1', 'feature_n']].describe() — count, mean, std, min, 25%, 50%, 75%, max — giving you a good idea of the distribution of your data and specific features
- analyzing missing data — sometimes it is expected
- data anomalies

- erroneous data — negative values that should not be negative, etc.

Model Comparison

As you can see, we have performed several steps before starting to discuss the main '*data science*' part. In this section, whether you are performing something like regression or classification, it is always best to compare several models before choosing one to update and enhance as your final model(*s*).

For example, although it might seem obvious to pick a specific machine learning algorithm for your use case, it is best to remove your bias, and obtain a baseline for say, 5 to 10 common algorithms. From there, you can compare the benefits of each—not just the accuracy. For example, you might want to compare the time it takes to train the model, or how expensive it could be, to the requirements of transforming your data.

Results Discussion

Before implementing your model into production, you will need to discuss the results with your stakeholder. You will look at what your accuracy means or your loss metric, like RMSE — Root Mean Square Error. These results are oftentimes confusing to people who do not study or employ data science, so it is your job to make them as simple as possible so that stakeholders can make decisions from your results — to move on or not for example (*sometimes a complex machine learning algorithm is not the answer to the problem*).

What is a Data Scientist

Data scientists are big data wranglers, gathering and analyzing large sets of structured and unstructured data. A data scientist's role combines computer science, statistics, and mathematics. They analyze, process, and model data then interpret the results to create actionable plans for companies and other organizations.

Data scientists are analytical experts who utilize their skills in both technology and social science to find trends and manage data. They use industry knowledge, contextual understanding, skepticism of existing assumptions – to uncover solutions to business challenges.

A data scientist's work typically involves making sense of messy, unstructured data, from sources such as smart devices, social media feeds, and emails that don't neatly fit into a database.

2. BIG DATA AND DATA SCIENCE HYPE

HYPE: intensive publicity or promotion

Data science - extension of statistics that deals with large datasets with the help of computer science technologies. Machine Learning is subset of data science.

Big data – deals with the vast collection of heterogeneous data from different sources and is not available in standard formats that we are aware of. This data won't be tabulated with a table or chart or graph.

It classifies into

1) Structured

Structured is one of the types of big data and By structured data, we mean data that can be processed, stored, and retrieved in a fixed format. It refers to highly organized information that can be readily and seamlessly stored and accessed from a database by simple search engine algorithms. For instance, the employee table in a company database will be structured as the employee details, their job positions, their salaries, etc., will be present in an organized manner.

Examples – RDBMS, OLTP(online transaction processing)

2) Unstructured

Unstructured data refers to the data that lacks any specific form or structure whatsoever. This makes it very difficult and time-consuming to process and analyze unstructured data. Email is an example of unstructured data. Structured and unstructured are two important types of big data.

Examples- social networks, digital images

3) **Semi-structured**

Semi structured is the third type of big data. Semi-structured data pertains to the data containing both the formats mentioned above, that is, structured and unstructured data. To be precise, it refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data. Thus we come to the end of types of data.

Examples – XML files, text files etc.

Refer [Big Data and Data Science – is this hype? - Saama](#).

Data Science is an area.	Big Data is a technique to collect, maintain and process huge information.
It is about the collection, processing, analyzing, and utilizing of data in various operations. It is more conceptual.	It is about extracting vital and valuable information from a huge amount of data.
It is a field of study just like Computer Science, Applied Statistics, or Applied Mathematics.	It is a technique for tracking and discovering trends in complex data sets.
The goal is to build data-dominant products for a venture.	The goal is to make data more vital and usable i.e. by extracting only important information from the huge data within existing traditional aspects.
Tools mainly used in Data Science include SAS, R, Python, etc	Tools mostly used in Big Data include Hadoop, Spark, Flink, etc.

It is a superset of Big Data as data science consists of Data scrapping, cleaning, visualization, statistics, and many more techniques.	It is a sub-set of Data Science as mining activities which is in a pipeline of Data science.
It is mainly used for scientific purposes.	It is mainly used for business purposes and customer satisfaction.
It broadly focuses on the science of the data. Example- digital advertisement, internet search	It is more involved with the processes of handling voluminous data. Example – gaming sector, health care sector.

3. Getting Past the Hype

Rachel's experience going from getting a PhD in statistics to working at Google is a great example to illustrate why we thought, in spite of the aforementioned reasons to be dubious; there might be some meat in the data science sandwich. In her words:

It was clear to me pretty quickly that the stuff I was working on at Google was different than anything I had learned at school when I got my PhD in statistics. This is not to say that my degree was useless; far from it—what I'd learned in school provided a framework and way of thinking that I relied on daily, and much of the actual content provided a solid theoretical and practical foundation necessary to do my work.

But there were also many skills I had to acquire on the job at Google that I hadn't learned in school. Of course, my experience is specific to me in the sense that I had a statistics background and picked up more computation, coding, and visualization skills, as well as domain expertise while at Google. Another person coming in as a computer scientist or a social scientist or a physicist would have different gaps and would fill them in *accordingly. But what is important here is that, as individuals, we each had different strengths and gaps, yet we were able to solve problems by putting ourselves together into a data team well-suited to solve the data problems that came our way.*

Here's a reasonable response you might have to this story. It's a general truism that, whenever you go from school to a real job, you realize there's a gap between what you learned in school and what you do on the job. In other words, you were simply facing the difference between academic statistics and industry statistics.

A couple replies to this:

- Sure, there's is a difference between industry and academia. But does it really have to be that way? Why do many courses in school have to be so intrinsically out of touch with reality?
- Even so, the gap doesn't represent simply a difference between industry statistics and academic statistics. The general experience of data scientists is that, at their job, they have access to a *larger body of knowledge and methodology*, as well as a process, which we now define as the *data science process* that has foundations in both statistics and computer science.

4) Why Now?

We have massive amounts of data about many aspects of our lives, and, simultaneously, an abundance of inexpensive computing power. Shopping, communicating, reading news, listening to music, searching for information, expressing our opinions—all this is being tracked online, as most people know.

What people might not know is that the "datafication" of our offline behavior has started as well, mirroring the online data collection revolution (more on this later). Put the two together, and there's a lot to learn about our behavior and, by extension, who we are as a species.

It's not just Internet data, though—it's finance, the medical industry, pharmaceuticals, bioinformatics, social welfare, government, education, retail, and the list goes on. There is a growing influence of data in most sectors and most industries. In some cases, the amount of data collected might be enough to be considered "big" (more on this in the next chapter); in other cases, it's not.

But it's not only the massiveness that makes all this new data interesting (or poses challenges). It's that the data itself, often in real time, becomes the building blocks of data *products*. On the Internet, this means Amazon recommendation systems, friend recommendations on Facebook, film and music recommendations, and so on. In finance, this means credit ratings, trading algorithms, and models. In education, this is starting to mean dynamic personalized learning and assessments coming out of places like Knewton and Khan Academy. In government, this means policies based on data.

We're witnessing the beginning of a massive, culturally saturated feedback loop where our behavior changes the product and the product changes our behavior. Technology makes this possible: infrastructure for large-scale data processing, increased memory, and bandwidth, as well as a cultural acceptance of technology in the fabric of our lives. This wasn't true a decade ago.

5) DATAFICATION

1. Taking all aspects of life and turning them into data.
2. **Datafication is** a technological trend turning many aspects of our life into data which is subsequently transferred into information realized as a new form of value.

Simply put, datafication is a set of tools, processes, and technologies used to create a data-driven organization or team. Data-fed enterprises use data logging. This is a method of collecting real (or system, in many IT companies) data over a period of time and converting it into a digital format that can be reported and manipulated to provide a comprehensive view.

Datafication is an interesting concept and led us to consider its importance with respect to people's intentions about sharing their own data. We are being datafied, or rather our actions are, and when we "like" someone or something online, we are intending to be datafied, or at least we should expect to be. But when we merely browse the Web, we are unintentionally, or at least passively, being datafied through cookies that we might or might not be aware of. And when we walk around in a store, or even on the street, we are being datafied in a completely unintentional way, via sensors, cameras, or Google glasses.

This spectrum of intentionality ranges from us gleefully taking part in a social media experiment we are proud of, to all-out surveillance and stalking. But it's all datafication. Our intentions may run the gamut, but the results don't.

Benefits of Datafication:

- Datafication helps you manage your data
- Datafication Accelerates Data Processing.
- Quick access to related data.

Reference for Datafication: https://griffinnet.com/what-is-datafication-and-how-can-it-impact-your-business/#:~:text=Datafication allows you to easily instruct

6. The Current Landscape (with a Little History)

So, what is data science? Is it new, or is it just statistics or analytics rebranded? Is it real, or is it pure hype? And if it's new and if it's real, what does that mean?

This is an ongoing discussion, but one way to understand what's going on in this industry is to look online and see what current discussions are taking place. This doesn't necessarily tell us what data science is, but it at least tells us what other people think it is, or how they're perceiving it. For example, on Quora there's a discussion from 2010 about "What is Data Science?" and here's Metamarket CEO Mike Driscoll's answer:

Data science, as it's practiced, is a blend of Red-Bull-fueled hacking and espresso-inspired statistics.

But data science is not merely hacking—because when hackers finish debugging their Bash one-liners and Pig scripts, few of them care about non-Euclidean distance metrics.

And data science is not merely statistics, because when statisticians finish theorizing the perfect model, few could read a tab-delimited file into R if their job depended on it.

Data science is the civil engineering of data. Its acolytes possess a practical knowledge of tools and materials, coupled with a theoretical understanding of what's possible.

Driscoll then refers to Drew Conway's Venn diagram of data science from 2010, shown in Figure 1-1.

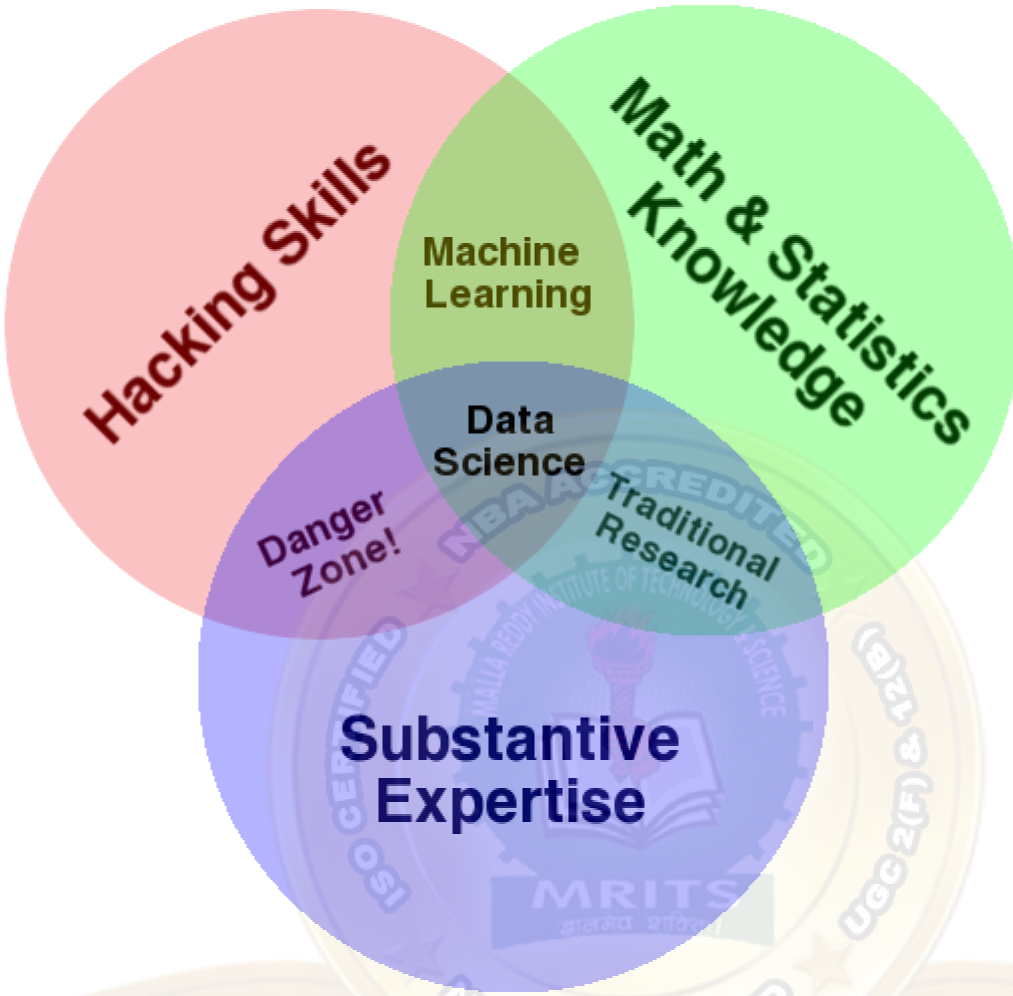


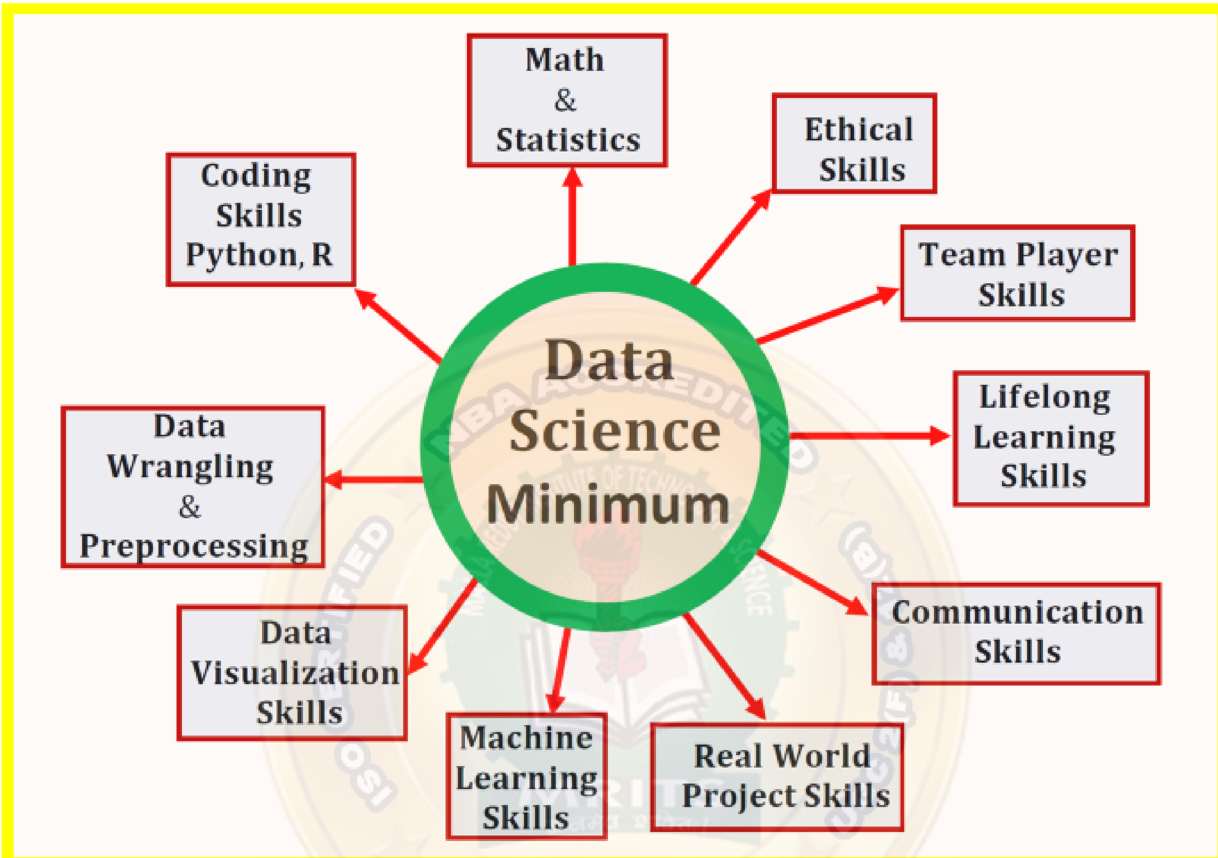
Figure 1-1. Drew Conway's Venn diagram of data science

- Statistics (traditional analysis you're used to thinking about)
- Data munging (parsing, scraping, and formatting data)
- Visualization (graphs, tools, etc.)

But wait, is data science just a bag of tricks? Or is it the logical extension of other fields like statistics and machine learning?.

Reference: [1. Introduction: What Is Data Science? - Doing Data Science [Book] (oreilly.com)](oreilly.com).

7. SKILLS SETS NEEDED:



Data Science is such a broad field that includes several subdivisions like data preparation and exploration; data representation and transformation; data visualization and presentation; predictive analytics; machine learning, etc. For beginners, it's only natural to raise the following question: **What skills do I need to become a data scientist?**

This article will discuss 10 essential skills that are necessary for practicing data scientists. These skills could be grouped into 2 categories, namely, **technological skills** (Math & Statistics, Coding Skills, Data Wrangling & Preprocessing Skills, Data Visualization Skills, Machine Learning

Skills,and Real World Project Skills) and **soft skills** (Communication Skills, Lifelong Learning Skills, Team Player Skills and Ethical Skills).

1. Mathematics and Statistics Skills

(I) Statistics and Probability

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

a) Mean

b) Median

c) Mode

d) Standard deviation/variance

e) Correlation coefficient and the covariance matrix

f) Probability distributions (Binomial, Poisson, Normal)

g) p-value

h) MSE (mean square error)

i) R2 Score

j) Baye's Theorem (Precision, Recall, Positive Predictive Value, Negative Predictive Value, Confusion Matrix, ROC Curve)

k) A/B Testing

l) Monte Carlo Simulation

(II) Multivariable Calculus

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

a) Functions of several variables

b) Derivatives and gradients

c) Step function, Sigmoid function, Logit function, ReLU (Rectified Linear Unit) function

d) Cost function

e) Plotting of functions

f) Minimum and Maximum values of a function

(III) Linear Algebra

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

a) Vectors

b) Matrices

c) Transpose of a matrix

d) The inverse of a matrix

e) The determinant of a matrix

f) Dot product

g) Eigenvalues

h) Eigenvectors

(IV) Optimization Methods

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

a) Cost function/Objective function

b) Likelihood function

c) Error function

d) Gradient Descent Algorithm and its variants (e.g. Stochastic Gradient Descent Algorithm)

Find out more about the gradient descent algorithm here: **Machine Learning: How the Gradient Descent Algorithm Works**.

2. Essential Programming Skills

Programming skills are essential in data science. Since Python and R are considered the 2 most popular programming languages in data science, essential knowledge in both languages are crucial. Some organizations may only require skills in either R or Python, not both.

(I) Skills in Python

Be familiar with basic programming skills in python. Here are the most important packages that you should master how to use:

a) Numpy

b) Pandas

c) Matplotlib

d) Seaborn

e) Scikit-learn

f) PyTorch

(ii) Skills in R

a) Tidyverse

b) Dplyr

c) Ggplot2

d) Caret

e) Stringr

(iii) Skills in Other Programming Languages

Skills in the following programming languages may be required by some organizations or industries:

a) Excel

b) Tableau

c) Hadoop

d) SQL

e) Spark

3. Data Wrangling and Preprocessing Skills

Data is key for any analysis in data science, be it inferential analysis, predictive analysis, or prescriptive analysis. The predictive power of a model depends on the quality of the data that was used in building the model. Data comes in different forms such as text, table, image, voice or video. Most often, data that is used for analysis has to be mined, processed and transformed to render it to a form suitable for further analysis.

i) **Data Wrangling**: The process of data wrangling is a critical step for any data scientist. Very rarely is data easily accessible in a data science project for analysis. It's more likely for the data to be in a file, a database, or extracted from documents such as web pages, tweets, or PDFs. Knowing how to wrangle and clean data will enable you to derive critical insights from your data that would otherwise be hidden.

ii) **Data Preprocessing**: Knowledge about data preprocessing is very important and include topics such as:

a) Dealing with missing data

b) Data imputation

c) Handling categorical data

d) Encoding class labels for classification problems

e) Techniques of feature transformation and dimensionality reduction such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA).

4. Data Visualization Skills

Understand the essential components of a good data visualization.

a) **Data Component**: An important first step in deciding how to visualize data is to know what type of data it is, e.g. categorical data, discrete data, continuous data, time series data, etc.

b) **Geometric Component:** Here is where you decide what kind of visualization is suitable for your data, e.g. scatter plot, line graphs, barplots, histograms, qqplots, smooth densities, boxplots, pairplots, heatmaps, etc.

c) **Mapping Component:** Here you need to decide what variable to use as your x-variable and what to use as your y-variable. This is important especially when your dataset is multi-dimensional with several features.

d) **Scale Component:** Here you decide what kind of scales to use, e.g. linear scale, log scale, etc.

e) **Labels Component:** This include things like axes labels, titles, legends, font size to use, etc.

f) **Ethical Component**: Here, you want to make sure your visualization tells the true story. You need to be aware of your actions when cleaning, summarizing, manipulating and producing a data visualization and ensure you aren't using your visualization to mislead or manipulate your audience.

5. Basic Machine Learning Skills

Machine Learning is a very important branch of data science. It is important to understand the machine learning framework: Problem Framing; Data Analysis; Model Building, Testing &Evaluation; and Model Application. Find out more about the machine learning framework from here: **The Machine Learning Process**.

The following are important machine learning algorithms to be familiar with.

i) Supervised Learning (Continuous Variable Prediction)

a) Basic regression

b) Multiregression analysis

c) Regularized regression

ii) Supervised Learning (Discrete Variable Prediction)

a) Logistic Regression Classifier

b) Support Vector Machine Classifier

c) K-nearest neighbor (KNN) Classifier

d) Decision Tree Classifier

e) Random Forest Classifier

iii) Unsupervised Learning

a) Kmeans clustering algorithm

6. Skills from Real World Capstone Data Science Projects

Skills acquired from course work alone will not make your a data scientist. A qualified data scientist must be able to demonstrate evidence of successful completion of a real world data science project that includes every stages in data science and machine learning process such as problem framing, data acquisition and analysis, model building, model testing, model evaluation, and deploying model. Real world data science projects could be found in the following:

a) Kaggle Projects

b) Internships

c) From Interviews

7. Communication Skills

Data scientists need to be able communicate their ideas with other members of the team or with business administrators in their organizations. Good communication skills would play a key role here to be able to convey and present very technical information to people with little or no understanding of technical concepts in data science. Good communication skills will help foster an atmosphere of unity and togetherness with other team members such as data analysts, data engineers, field engineers, etc.

8. Be a Lifelong Learner

Data science is a field that is ever-evolving, so be prepared to embrace and learn new technologies. One way to keep in touch with developments in the field is to network with other data scientists. Some platforms that promote networking are LinkedIn, github, and medium (**Towards Data**

Science and **Towards AI** publications). The platforms are very useful for up-to-date information about recent developments in the field.

9. Team Player Skills

As a data scientist, you will be working in a team of data analysts, engineers, administrators, so you need good communication skills. You need to be a good listener too, especially during early project development phases where you need to rely on engineers or other personnel to be able to design and frame a good data science project. Being a good team player world help you to thrive in a business environment and maintain good relationships with other members of your team as well as administrators or directors of your organization.

10. Ethical Skills in Data Science

Understand the implication of your project. Be truthful to yourself. Avoid manipulating data or using a method that will intentionally produce bias in results. Be ethical in all phases from data collection, to analysis, to model building, analysis, testing and application. Avoid fabricating results for the purpose of misleading or manipulating your audience. Be ethical in the way you interpret the findings from your data science project.

Reference: Data Science Minimum: 10 Essential Skills You Need to Know to Start Doing Data Science | by Benjamin Obi Tayo Ph.D. | Towards Data Science

8. STATISTICAL INFERENCE

Inference: a conclusion reached on the basis of evidence and reasoning. a guess that you make or an opinion that you form based on the information that you have.

Statistical inference is a vast area which includes many statistical methods from analyzing data to drawing inferences or conclusions in research or business problems. It plays a vital role in the application of data science across industries.

Statistical inference is the process of drawing conclusions about unknown population properties, using a sample drawn from the population. Unknown population properties can be, for example, mean, proportion or variance. These are also called parameters.

Statistical inference is broadly divided into 2 parts:

1. Estimation and

2. Hypothesis Testing.

Estimation is further divided into point estimation and interval estimation.

In point estimation, we estimate an unknown parameter using a single number that is calculated from the sample data. For example, the average salary of junior data scientists based on a sample is 55,000 euros

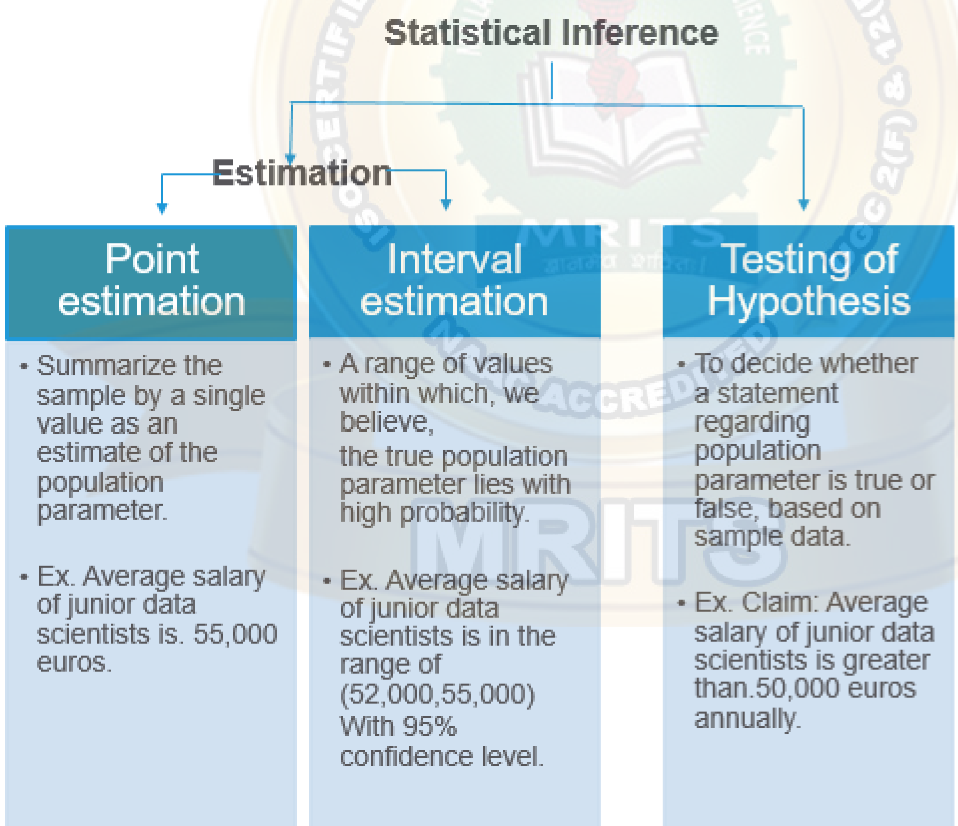
In Interval estimation, we find a range of values within which we believe the true population parameter lies with high probability. Here, the average salary of junior data scientists is between 52,0000 and 58,000, with a 95% confidence level.

In hypothesis testing we need to decide whether a statement regarding a population parameter is true or false, based on sample data. For example, a claim that the average salary of junior data scientists is greater than 50,0000 euros annually can be tested using sample data.

Importance of Statistical Inference

Inferential Statistics is important to examine the data properly. To make an accurate conclusion, proper data analysis is important to interpret the research results. It is majorly used in the future prediction for various observations in different fields. It helps us to make inference about the data. The statistical inference has a wide range of application in different fields, such as:

- Business Analysis
- Artificial Intelligence
- Financial Analysis
- Fraud Detection
- Machine Learning
- Share Market
- Pharmaceutical Sector



Statistical Inference Examples

An example of statistical inference is given below.

Question: From the shuffled pack of cards, a card is drawn. This trial is repeated for 400 times, and the suits are given below:

Suit	Spade	Clubs	Hearts	Diamonds
No.of times drawn	90	100	120	90

While a card is tried at random, then what is the probability of getting a

1. Diamond cards
2. Black cards
3. Except for spade

Solution:

By statistical inference solution,

Total number of events = 400

i.e.,90+100+120+90=400

(1) The probability of getting diamond cards:

Number of trials in which diamond card is drawn = 90

Therefore, P(diamond card) = 90/400 = 0.225

(2) The probability of getting black cards:

Number of trials in which black card showed up = 90+100 =190

Therefore, P(black card) = 190/400 = 0.475

(3) Except for spade

Number of trials other than spade showed up = 90+100+120 =310

Therefore, P(except spade) = 310/400 = 0.775

Reference: 1 Statistical Inference - Definition, Types, Procedure, and Example (byjus.com)

2. What is statistical inference - an introduction to inferential statistics (digitaschools.com)

9. POPULATION AND SAMPLES

A population is the entire group that you want to draw conclusions about.

A sample is the specific group that you will collect data from.

The size of the sample is always less than the total size of the population. In research, a population doesn't always refer to people.

In research, a population doesn't always refer to people. It can mean a group containing elements of anything you want to study, such as objects, events, organizations, countries, species, organisms, etc.

Collecting data from a population

Populations are used when your research question requires, or when you have access to, data from every member of the population.

Usually, it is only straightforward to collect data from a whole population when it is small, accessible and cooperative.

Example: Collecting data from a population. A high school administrator wants to analyze the final exam scores of all graduating seniors to see if there is a trend. Since they are only interested in applying their findings to the graduating seniors in this high school, they use the whole **population** dataset.

For larger and more dispersed populations, it is often difficult or impossible to collect data from every individual. For example, every 10 years, the federal US government aims to count every person living in the country using the US Census. This data is used to distribute funding across the nation.

However, historically, marginalized and low-income groups have been difficult to contact, locate and encourage participation from. Because of non-responses, the population count is incomplete and biased towards some groups, which results in disproportionate funding across the country.

In cases like this, sampling can be used to make more precise inferences about the population.

Collecting data from a sample

When your population is large in size, geographically dispersed, or difficult to contact, it's necessary to use a sample. With statistical analysis, you can use sample data to make estimates or test hypotheses about population data.

Example: Collecting data from a sampleYou want to study political attitudes in young people. Your population is the 300,000 undergraduate students in the Netherlands. Because it's not practical to collect data from all of them, you use a **sample** of 300 undergraduate volunteers from three Dutch universities who meet your inclusion criteria. This is the group who will complete your online survey.

Ideally, a sample should be randomly selected and representative of the population. Using probability sampling methods (such as simple random sampling or stratified sampling) reduces the risk of sampling bias and enhances both internal and external validity.

For practical reasons, researchers often use non-probability sampling methods. Non-probability samples are chosen for specific criteria; they may be more convenient or cheaper to access. Because of non-random selection methods, any statistical inferences about the broader population will be weaker than with a probability sample.

Reasons for sampling

- **Necessity**: Sometimes it's simply not possible to study the whole population due to its size or inaccessibility.
- **Practicality**: It's easier and more efficient to collect data from a sample.
- **Cost-effectiveness**: There are fewer participant, laboratory, equipment, and researcher costs involved.
- **Manageability**: Storing and running statistical analyses on smaller datasets is easier and reliable.

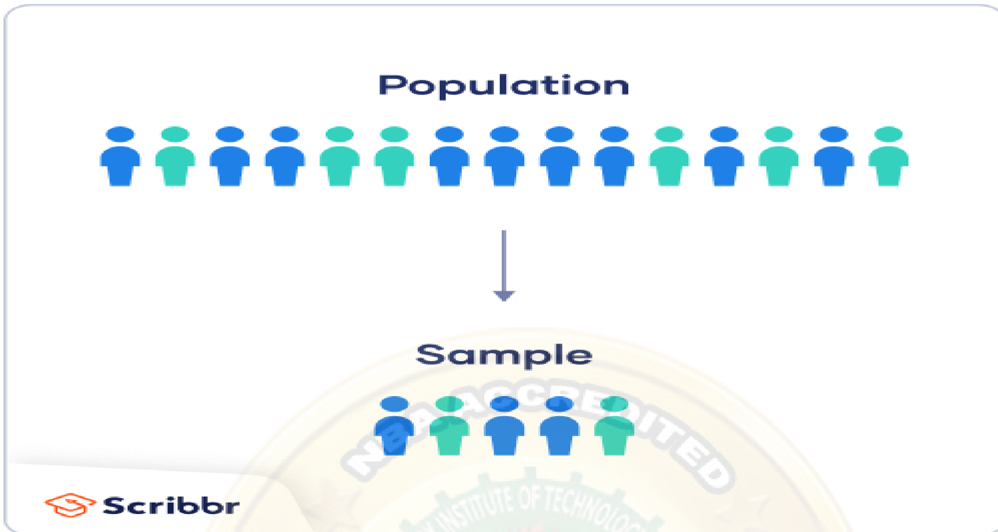
Population and Sample Examples.

All the people who have the ID proofs is the population and a group of people who only have voter id with them is the sample.

All the students in the class are population whereas the top 10 students in the class are the sample.

All employees in an office would be population.

Out of all employees all the managers in the office is sample.



Reference: Population vs. Sample | Definitions, Differences & Examples (scribbr.com)

10. STATISTICAL MODELING

Statistical modeling is the use of mathematical models and statistical assumptions to generate sample data and make predictions about the real world. A statistical model is a collection of probability distributions on a set of all possible outcomes of an experiment. Statistical modeling is the process of applying statistical analysis to a dataset. A statistical model is a mathematical representation (or mathematical model) of observed data.

The application of statistical modeling to raw data helps data scientists approach data analysis in a strategic manner, providing intuitive visualizations that aid in identifying relationships between variables and making predictions.

Common data sets for statistical analysis include Internet of Things (IoT) sensors, census data, public health data, social media data, imagery data, and other public sector data that benefit from real-world predictions.

Statistical Modeling Techniques

The first step in developing a statistical model is gathering data, which may be sourced from spreadsheets, databases, data lakes, or the cloud. The most common statistical modeling methods for analyzing this data are categorized as either supervised learning or unsupervised learning. Some popular statistical model examples include logistic regression, time-series, clustering, and decision trees.

Supervised learning techniques include regression models and classification models:

- **Regression model**: a type of predictive statistical model that analyzes the relationship between a dependent and an independent variable. Common regression models include logistic, polynomial, and linear regression models. Use cases include forecasting, time series modeling, and discovering the causal effect relationship between variables.
- **Classification model**: a type of machine learning in which an algorithm analyzes an existing, large and complex set of known data points as a means of understanding and then appropriately classifying the data; common models include models include decision trees, Naive Bayes, nearest neighbor, random forests, and neural networking models, which are typically used in Artificial Intelligence.

Unsupervised learning techniques include clustering algorithms and association rules:

- **K-means clustering**: aggregates a specified number of data points into a specific number of groupings based on certain similarities.
- **Reinforcement learning**: an area of deep learning that concerns models iterating over many attempts, rewarding moves that produce favorable outcomes and penalizing steps that produce undesired outcomes, therefore training the algorithm to learn the optimal process.

fig 1: before applying k-means clustering

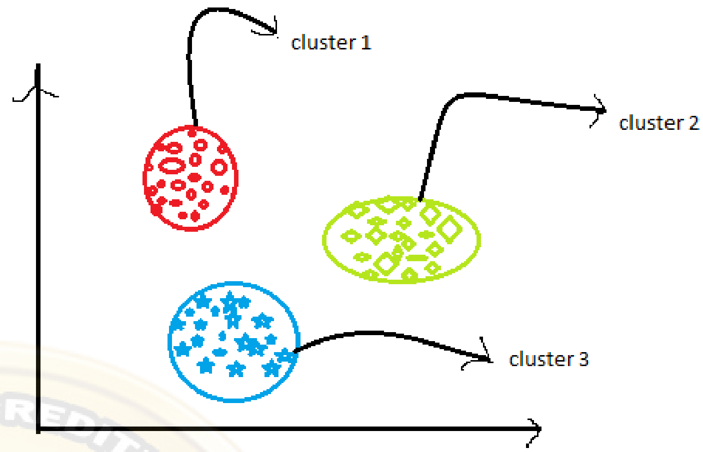


fig 2: After applying K-means clustering

There are three main types of statistical models: parametric, nonparametric, and semi parametric:

- **Parametric**: a family of probability distributions that has a finite number of parameters.
- **Nonparametric**: models in which the number and nature of the parameters are flexible and not fixed in advance.
- **Semi parametric**: the parameter has both a finite-dimensional component (parametric) and an infinite-dimensional component (nonparametric).

What Is Statistical Analysis?

Statistical analysis is the process of collecting and analyzing data in order to discern patterns and trends. It is a method for removing bias from evaluating data by employing numerical analysis. This technique is useful for collecting the interpretations of research, developing statistical models, and planning surveys and studies.

Types of Statistical Analysis

Given below are the 6 types of statistical analysis:

- **Descriptive Analysis** Descriptive statistical analysis involves collecting, interpreting, analyzing, and summarizing data to present them in the form of charts, graphs, and tables. Rather than drawing conclusions, it simply makes the complex data easy to read and understand.

- **Inferential Analysis** The inferential statistical analysis focuses on drawing meaningful conclusions on the basis of the data analyzed. It studies the relationship between different variables or makes predictions for the whole population.

- **Predictive Analysis** Predictive statistical analysis is a type of statistical analysis that analyzes data to derive past trends and predict future events on the basis of them. It uses machine learning algorithms, data mining, data modelling, and artificial intelligence to conduct the statistical analysis of data.

- **Prescriptive Analysis** The prescriptive analysis conducts the analysis of data and prescribes the best course of action based on the results. It is a type of statistical analysis that helps you make an informed decision.

- **Exploratory Data Analysis** Exploratory analysis is similar to inferential analysis, but the difference is that it involves exploring the unknown data associations. It analyzes the potential relationships within the data.

- **Causal Analysis:** The causal statistical analysis focuses on determining the cause and effect relationship between different variables within the raw data. In simple words, it determines why something happens and its effect on other variables. This methodology can be used by businesses to determine the reason for failure**.**

Benefits of Statistical Analysis

Statistical analysis can be called a boon to mankind and has many benefits for both individuals and organizations. Given below are some of the reasons why you should consider investing in statistical analysis:

- It can help you determine the monthly, quarterly, yearly figures of sales profits, and costs making it easier to make your decisions.
- It can help you make informed and correct decisions.
- It can help you identify the problem or cause of the failure and make corrections. For example, it can identify the reason for an increase in total costs and help you cut the wasteful expenses.
- It can help you conduct market analysis and make an effective marketing and sales strategy.
- It helps improve the efficiency of different processes.

Statistical Analysis Methods

Although there are various methods used to perform data analysis, given below are the 5 most used and popular methods of statistical analysis:

- Mean

Mean or average mean is one of the most popular methods of statistical analysis. Mean determines the overall trend of the data and is very simple to calculate. Mean is calculated by summing the numbers in the data set together and then dividing it by the number of data points. Despite the ease of calculation and its benefits, it is not advisable to resort to mean as the only statistical indicator as it can result in inaccurate decision making.

- Standard Deviation

Standard deviation is another very widely used statistical tool or method. It analyzes the deviation of different data points from the mean of the entire data set. It determines how data of the data set is spread around the mean. You can use it to decide whether the research outcomes can be generalized or not.

- Regression

Regression is a statistical tool that helps determine the cause and effect relationship between the variables. It determines the relationship between a dependent and an independent variable. It is generally used to predict future trends and events.

- Hypothesis Testing

Hypothesis testing can be used to test the validity or trueness of a conclusion or argument against a data set. The hypothesis is an assumption made at the beginning of the research and can hold or be false based on the analysis results.

- Sample Size Determination

Sample size determination or data sampling is a technique used to derive a sample from the entire population, which is representative of the population. This method is used when the size of the population is very large. You can choose from among the various data sampling techniques such as snowball sampling, convenience sampling, and random sampling.

Statistical Analysis Examples

Look at the standard deviation sample calculation given below to understand more about statistical analysis.

The weights of 5 pizza bases in cms are as follows:

Particulars (Weight in cms)	Mean Deviation	Square of Mean Deviation
9	9-6.4 = 2.6	(2.6)2 = 6.76
2	2-6.4 = - 4.4	(-4.4)2 = 19.36
5	5-6.4 = - 1.4	(-1.4)2 = 1.96
4	4-6.4 = - 2.4	(-2.4)2 = 5.76
12	12-6.4 = 5.6	(5.6)2 = 31.36

Calculation of Mean = (9+2+5+4+12)/5 = 32/5 = 6.4

Calculation of mean of squared mean deviation = (6.76+19.36+1.96+5.76+31.36)/5 = 13.04

Sample Variance = 13.04

Standard deviation = √13.04 = 3.611

Reference: What is Statistical Analysis? Types, Methods and Examples | Simplilearn

What is Statistical Modeling? Definition and FAQs | HEAVY.AI

PROBABILITY DISTRIBUTIONS:

In probability theory and statistics, a **probability distribution** is the mathematical function that gives the probabilities of occurrence of different possible **outcomes** for an experiment. It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events (subsets of the sample space).

The number of times a value occurs in a sample is determined by its **probability** of occurrence. Probability is a number between 0 and 1 that says how likely something is to occur:

- 0 means it's impossible.
- 1 means it's certain.

The higher the probability of a value, the higher its frequency in a sample.

Reference: Introduction to Probability Distributions for Data Science (analyticsvidhya.com)

FITTING A MODEL:

Model fitting is a measure of how well a machine learning model generalizes to similar data to that on which it was trained. A model that is well-fitted produces more accurate outcomes. A model that is overfitted matches the data too closely. A model that is underfitted doesn't match closely enough.

Each machine learning algorithm has a basic set of parameters that can be changed to improve its accuracy. During the fitting process, you run an algorithm on data for which you know the target variable, known as "labeled" data, and produce a machine learning model. Then, you compare the outcomes to real, observed values of the target variable to determine their accuracy.

Next, you use that information to adjust the algorithm's standard parameters to reduce the level of error, making it more accurate in uncovering patterns and relationships between the rest of its features and the target.

You repeat this process until the algorithm finds the optimal parameters that produce valid, practical, applicable insights for your practical business problem.

Why is Model Fitting Important?

Model fitting is the essence of machine learning. If your model doesn't fit your data correctly, the outcomes it produces will not be accurate enough to be useful for practical decision-making. A properly fitted model has hyper parameters that capture the complex relationships between known variables and the target variable, allowing it to find relevant insights or make accurate predictions.

Fitting is an automatic process that makes sure your machine learning models have the individual parameters best suited to solve your specific real-world business problem with a high level of accuracy.

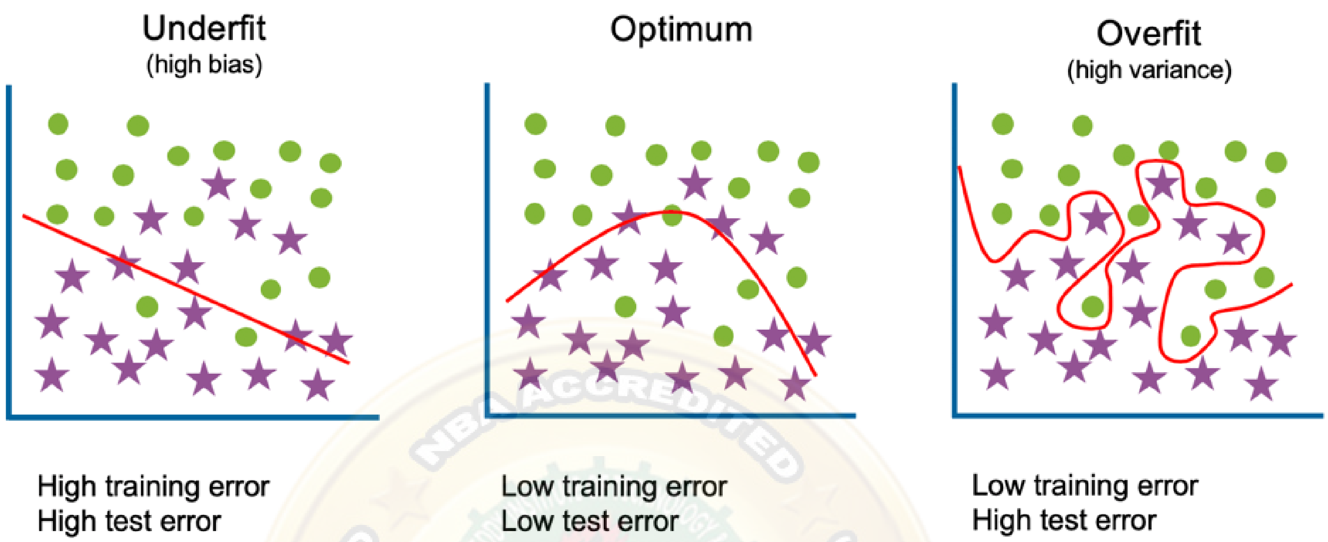
Overfitting in Machine Learning

In the real world, the dataset present will never be clean and perfect. It means each dataset contains impurities, noisy data, outliers, missing data, or imbalanced data. Due to these impurities, different problems occur that affect the accuracy and the performance of the model. One of such problems is Overfitting in Machine Learning. *Overfitting is a problem that a model can exhibit.*

- Overfitting & underfitting are the two main errors/problems in the machine learning model, which cause poor performance in Machine Learning.
- Overfitting occurs when the model fits more data than required, and it tries to capture each and every datapoint fed to it. Hence it starts capturing noise and inaccurate data from the dataset, which degrades the performance of the model.
- An overfitted model doesn't perform accurately with the test/unseen dataset and can't generalize well.

An overfitted model is said to have low bias and high variance

What is Overfitting?



Underfitting: A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data, i.e., it only performs well on training data but performs poorly on testing data. *(It's just like trying to fit undersized pants!)*

Underfitting destroys the accuracy of our machine learning model. Its occurrence simply means that our model or the algorithm does not fit the data well enough.

It usually happens when we have fewer data to build an accurate model and also when we try to build a linear model with fewer non-linear data. In such cases, the rules of the machine learning model are too easy and flexible to be applied to such minimal data and therefore the model will probably make a lot of wrong predictions. Underfitting can be avoided by using more data and also reducing the features by feature selection.

Reasons for Underfitting:

1. High bias and low variance
2. The size of the training dataset used is not enough.
3. The model is too simple.
4. Training data is not cleaned and also contains noise in it.

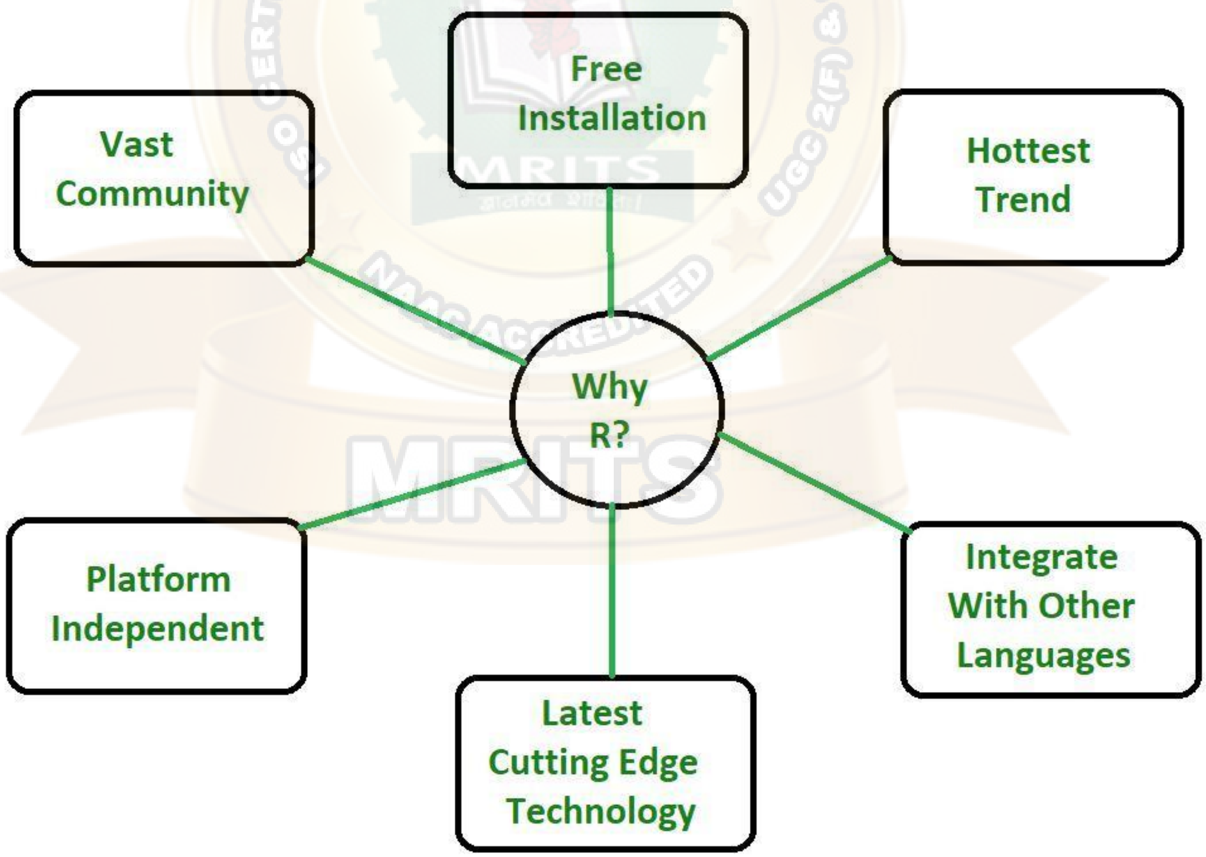
Reasons for Overfitting are as follows:

1. High variance and low bias
2. The model is too complex
3. The size of the training data

Introduction to R Language:

R is an open-source programming language that is widely used as a statistical software and data analysis tool. R generally comes with the Command-line interface. R is available across widely used platforms like Windows, Linux, and macOS. Also, the R programming language is the latest cutting-edge tool.

Why R Programming Language?



- R programming is used as a leading tool for machine learning, statistics, and data analysis. Objects, functions, and packages can easily be created by R.
- It's a platform-independent language. This means it can be applied to all operating system.
- It's an open-source free language. That means anyone can install it in any organization without purchasing a license.
- R programming language is not only a statistic package but also allows us to integrate with other languages (C, C++). Thus, you can easily interact with many data sources and statistical packages.
- The R programming language has a vast community of users and it's growing day by day.
- R is currently one of the most requested programming languages in the Data Science job market that makes it the hottest trend nowadays.

Features of R Programming Language

Statistical Features of R:

- **Basic Statistics:** The most common basic statistics terms are the mean, mode, and median. These are all known as "Measures of Central Tendency." So using the R language we can measure central tendency very easily.
- **Static graphics:** R is rich with facilities for creating and developing interesting static graphics. R contains functionality for many plot types including graphic maps, mosaic plots, biplots, and the list goes on.
- **Probability distributions:** Probability distributions play a vital role in statistics and by using R we can easily handle various types of probability distribution such as Binomial Distribution, Normal Distribution, Chi-squared Distribution and many more.
- **Data analysis:** It provides a large, coherent and integrated collection of tools for data analysis.

Programming Features of R:

- **R Packages:** One of the major features of R is it has a wide availability of libraries. R has CRAN(Comprehensive R Archive Network), which is a repository holding more than 10, 0000 packages.
- **Distributed Computing:** Distributed computing is a model in which components of a software system are shared among multiple computers to improve efficiency and performance. Two new packages **ddR and multidplyr** used for distributed programming in R were released in November 2015.

Advantages of R:

- R is the most comprehensive statistical analysis package. As new technology and concepts often appear first in R.
- As R programming language is an open source. Thus, you can run R anywhere and at any time.
- R programming language is suitable for GNU/Linux and Windows operating system.
- R programming is cross-platform which runs on any operating system.
- In R, everyone is welcome to provide new packages, bug fixes, and code enhancements.

Disadvantages of R:

- In the R programming language, the standard of some packages is less than perfect.
- Although, R commands give little pressure to memory management. So R programming language may consume all available memory.
- In R basically, nobody to complain if something doesn't work.
- R programming language is much slower than other programming languages such as Python and MATLAB.

Applications of R:

- We use R for Data Science. It gives us a broad variety of libraries related to statistics. It also provides the environment for statistical computing and design.

- R is used by many quantitative analysts as its programming tool. Thus, it helps in data importing and cleaning.
- R is the most prevalent language. So many data analysts and research programmers use it. Hence, it is used as a fundamental tool for finance.
- Tech giants like Google, Facebook, bing, Twitter, Accenture, Wipro and many more using R nowadays.

Data types in R

Data Type	Example	Verify

Logical	TRUE, FALSE	`v <- TRUE`
		`print(class(v))`
		it produces the following result −
		[1] "logical"

Numeric	12.3, 5, 999	`v <- 23.5`
		`print(class(v))`
		it produces the following result −
		[1] "numeric"

Integer	2L, 34L, 0L	`v <- 2L`
		`print(class(v))`
		it produces the following result −
		[1] "integer"

Complex	3 + 2i	`v <- 2+5i`
		`print(class(v))`
		it produces the following result −

		[1] "complex"

Character	'a' , '"good", "TRUE", '23.4'	`v <- "TRUE"`
		`print(class(v))`
		it produces the following result −
		[1] "character"

Raw	"Hello" is stored as 48 65 6c 6c 6f	`v <- charToRaw("Hello")`
		`print(class(v))`
		it produces the following result −
		[1] "raw"

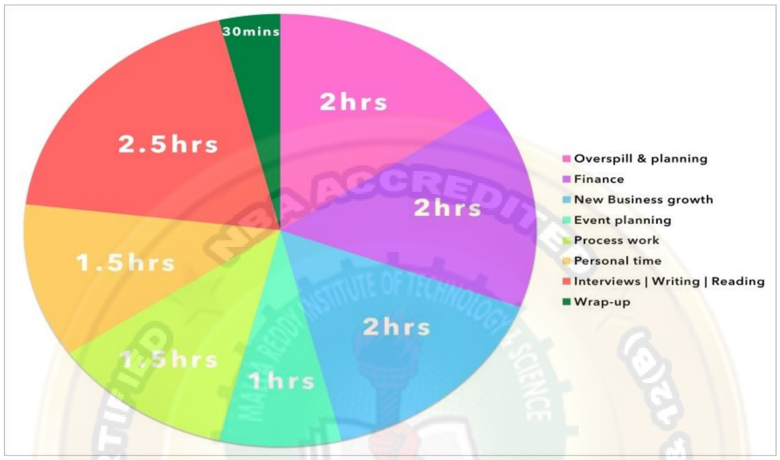
In R programming, the very basic data types are the R-objects called **vectors** which hold elements of different classes as shown above. Please note in R the number of classes is not confined to only the above six types. For example, we can use many atomic vectors and create an array whose class will become array.

ENVIRONMENTAL SETUP:

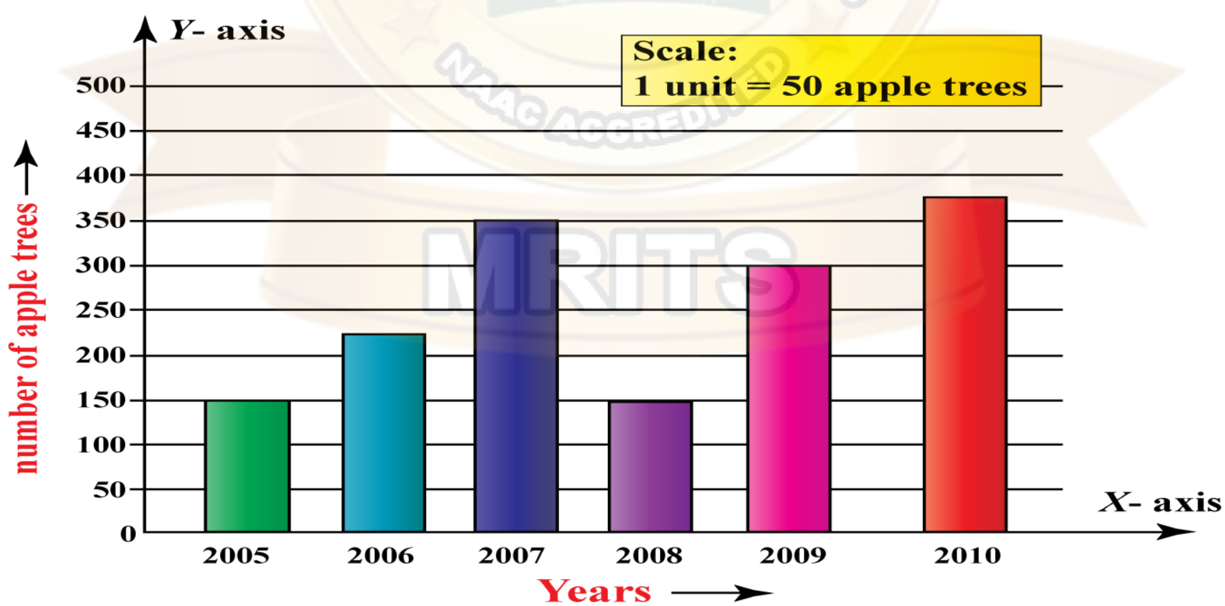
Reference : R Programming Environment setup in Windows - bbminfo

Different types of plots in python and SVM(Support vector machine)

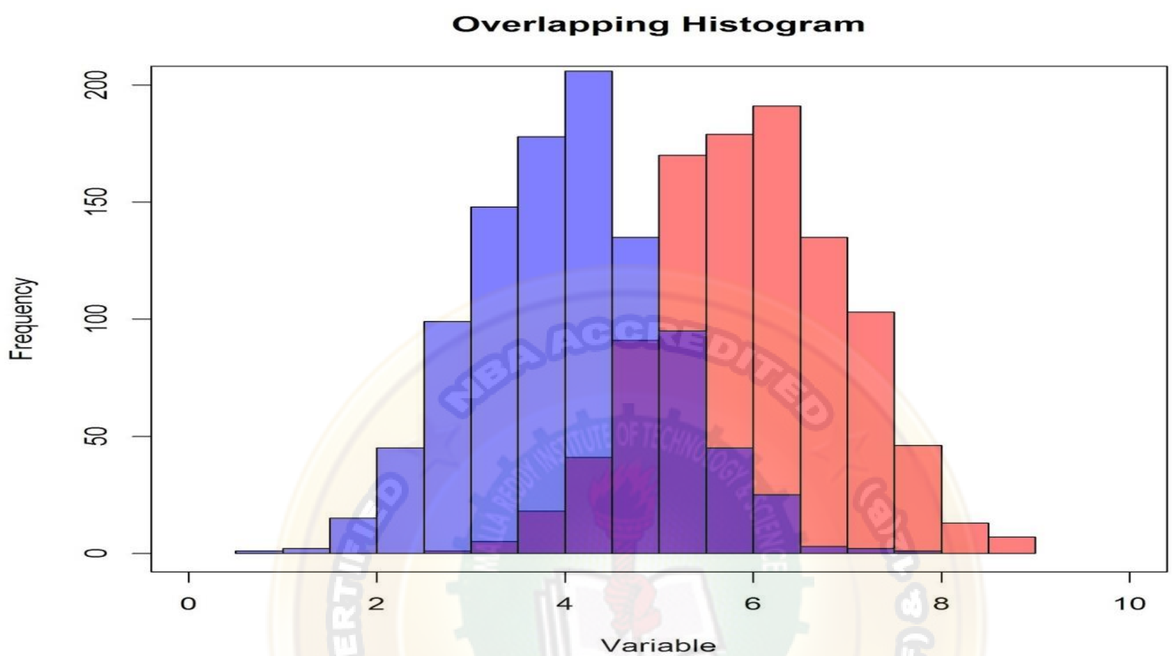
Pie chart:



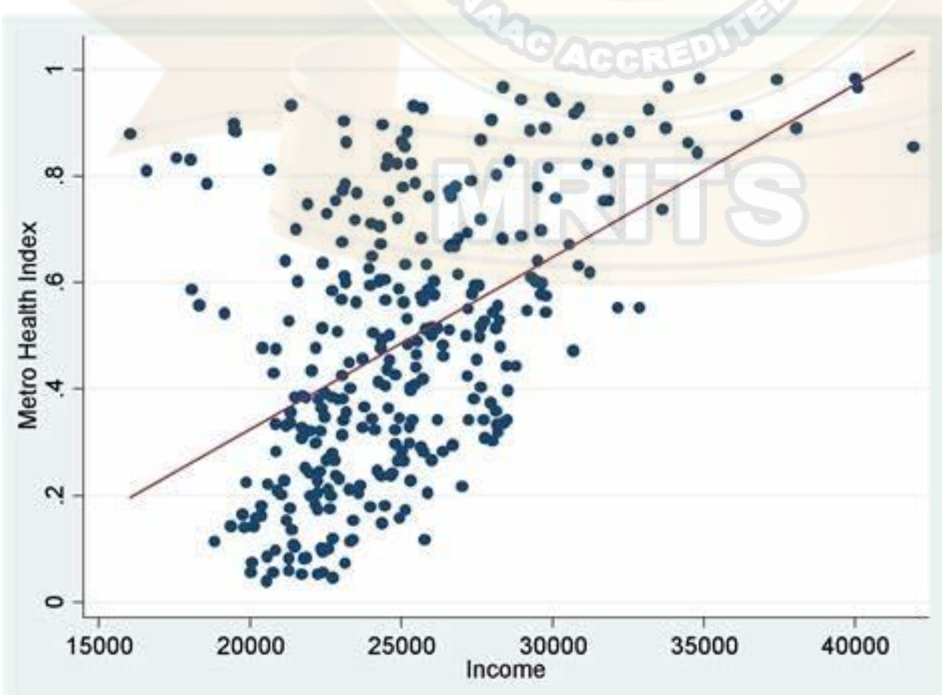
Bar graph



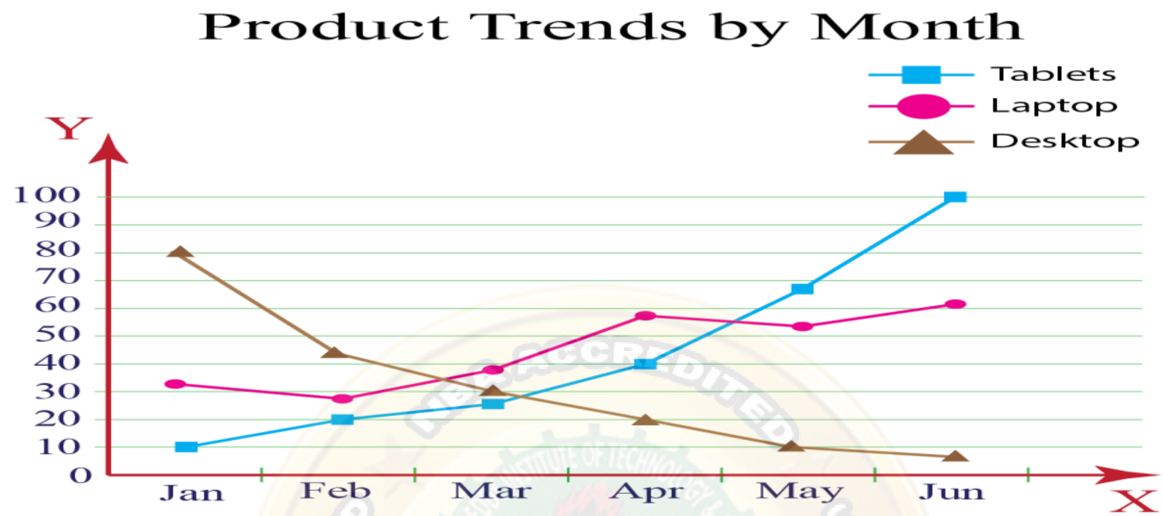
Histograms



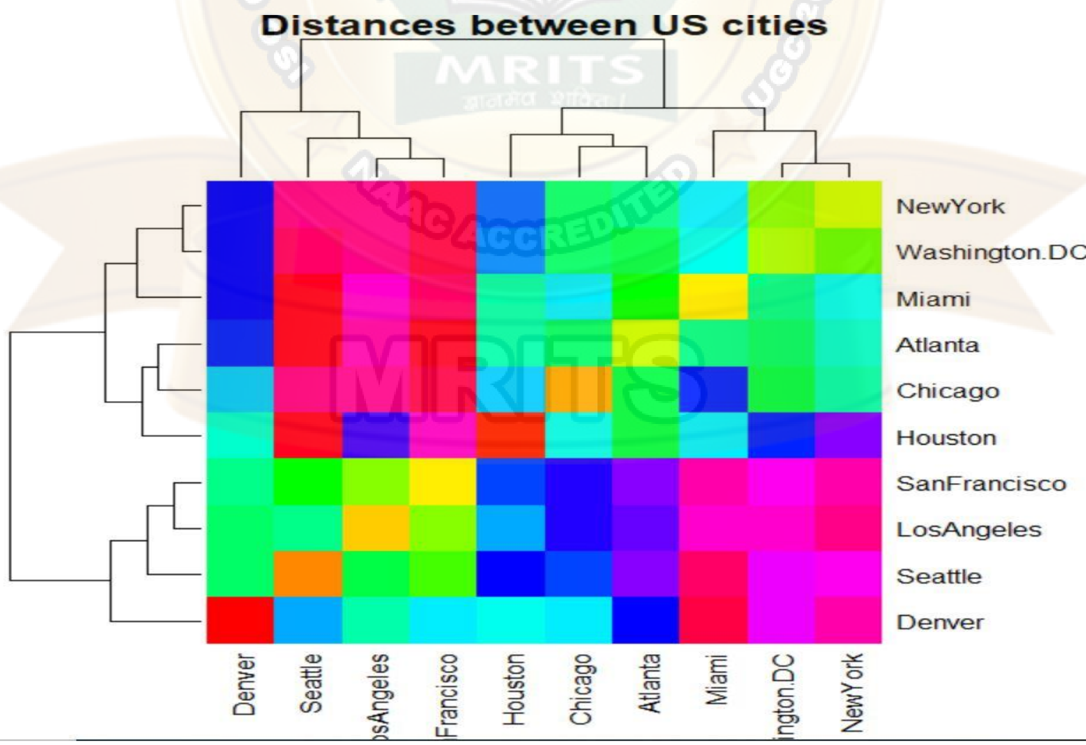
Scatterplot



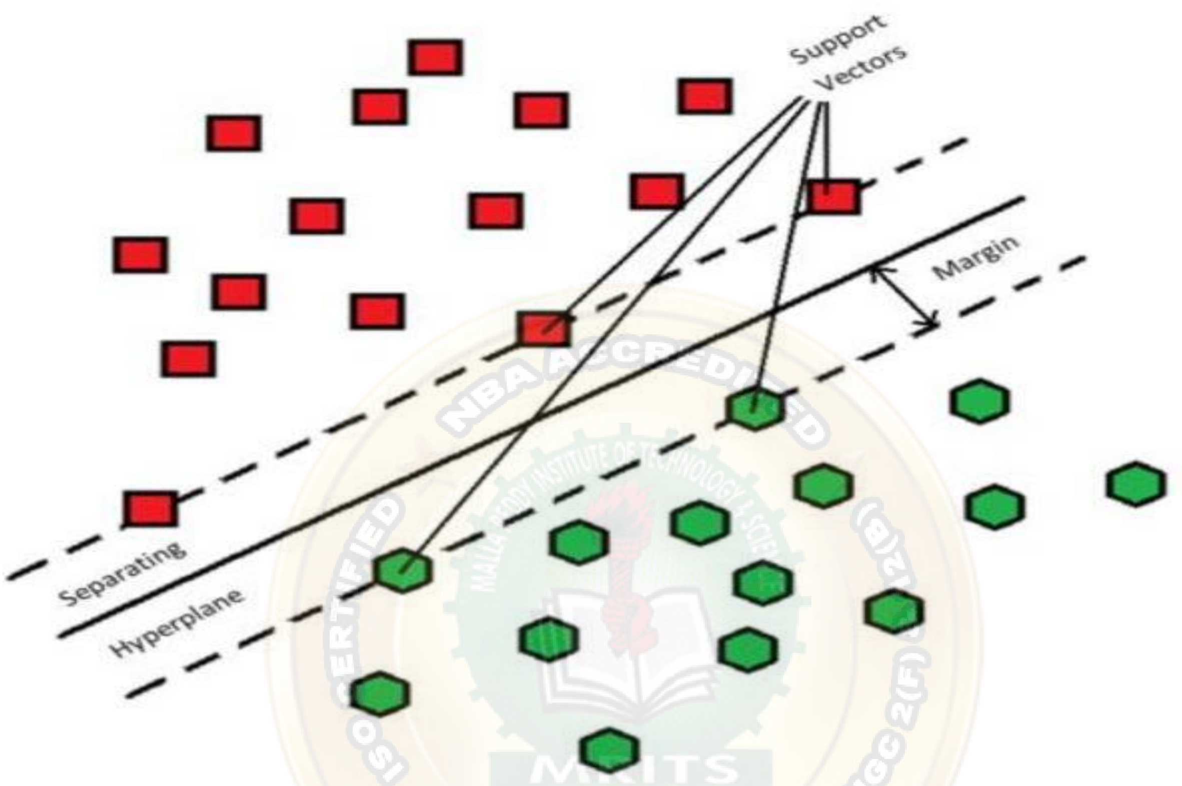
Line graphs



Heatmaps



SVM(Support Vector Machine)



27/9/2022

UNIT-II

Exploratory Data Analysis:

It is an approach to analyze the data using visual techniques.

- It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

Examples

Buy (or) Sellers 100 diff kind of shoes

dress shoes, hiking boots, sandals etc.

→ 1-3 different types of shoes

Sneakers, dress shoes and sandals

→ Data collection is an essential part of EDA.

→ Data cleaning

Significance

It allows data scientists to analyze the data before coming to any assumption.

It ensures that the results produced are valid and applicable to business outcomes.

Ex: Employee data

8 columns

name, Gender, start Date, Last login, Salary, Bonus%, Senior mngr, Team

To point first five rows we will use head() fn

```
import pandas as pd
import numpy as np
(dataframe) df = pd.read_csv('employees.csv')
df.head()
```

Numpy - Numerical Python [open Source module]

It provides mathematical computation on arrays and matrices.

```
>>> a=np.array([1,2,3])
>>> type(a) <type 'numpy.ndarray'>
```

Python Pandas package providing fast, flexible and expressive data structures designed to make working with relational (or) labeled data.

Dataframes

```
import pandas as pd
lst = ['java', 'python', 'Js']
dframe = pd.Dataframe(lst)
Print(dframe)
```

o/p:

```
0   java
1   python
2   Js
3   c++
```

→ df.shape

o/p: (1000, 8)

↓ rows ↓ columns

→ df.describe() → basic stastical computations on the dataset like Extreme values, count of data points, Sd, etc... Missing values are automatically skipped.

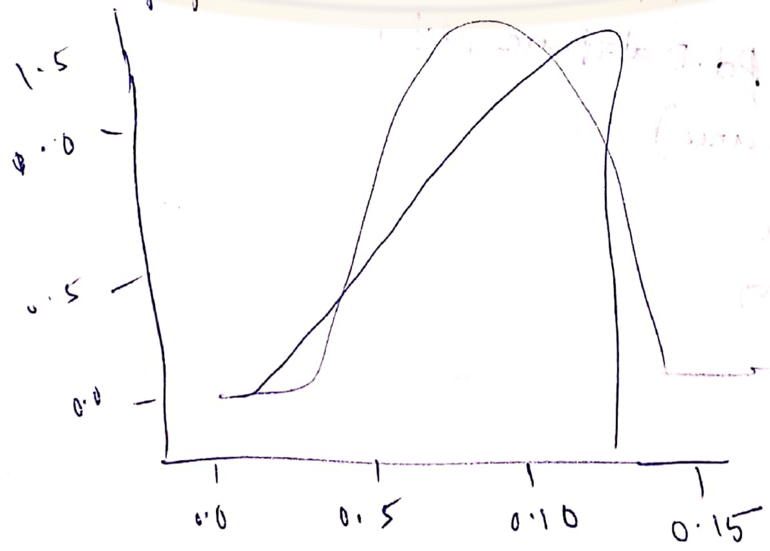
→ EDA helps us to

- give insight into a data set
- understand the underlying structure
- Extract important parameters and relationships that holds b/w them.
- Test underlying assumptions.

Types of EDA

1) Univariate Non-graphical - Don't provide a full picture of the data.

we use just one variable to research the info. The standard goal of univariate non-graphical EDA is to know the underlying sample distribution/ data, and make observations about the population.



2) Univariate Graphical-

Graphical Methods are required to provide a full picture of Data.

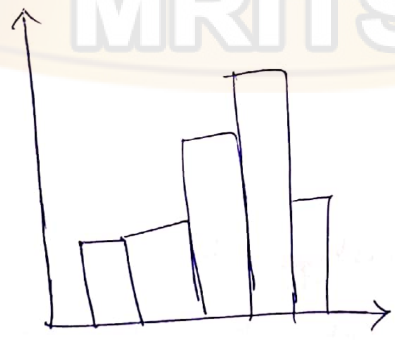
- stem- and- leaf plots

Frequency diagram in which the raw data is displayed together with its frequency.

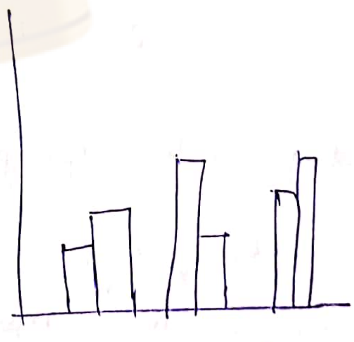
28, 38, 42, 5, 13, 23, 14, 36, 56, 20, 3

Stem	Leaf
0	3 5
1	3 4
2	0 3 8
3	6 8
4	2
5	6

- Histograms, a bar plot in which each bar represents the frequency (count) or proportion (count / total count) of cases for a range of values

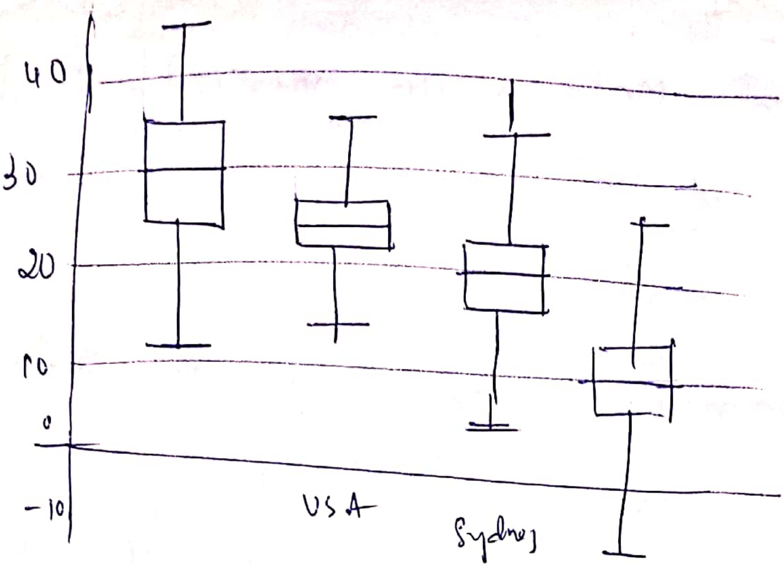


Histograms



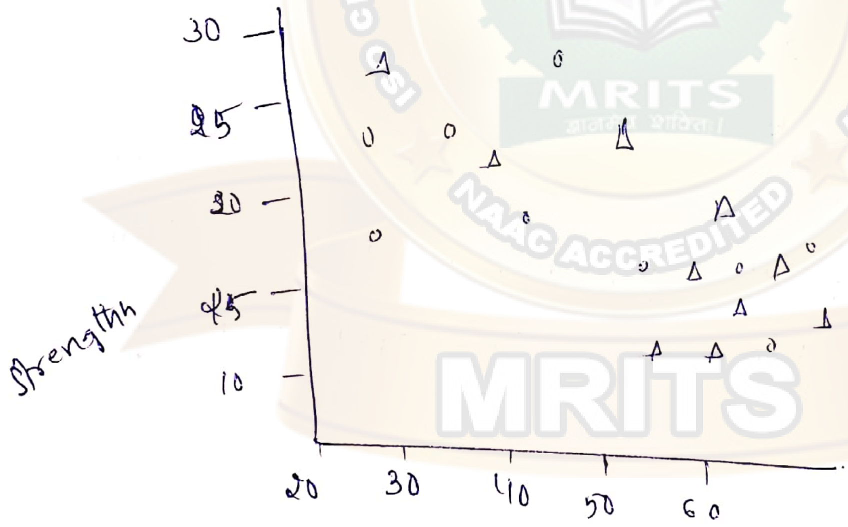
Bar plot

- Box plots - Which graphically depict the five - number Summary of min, first Quartile, median, third Quartile, max.



3) multivariate non-graphical

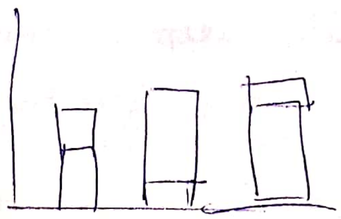
Data arises from more than one variable. shows the relationship b/w two or more variables of the data through cross-tabulation or statistics.



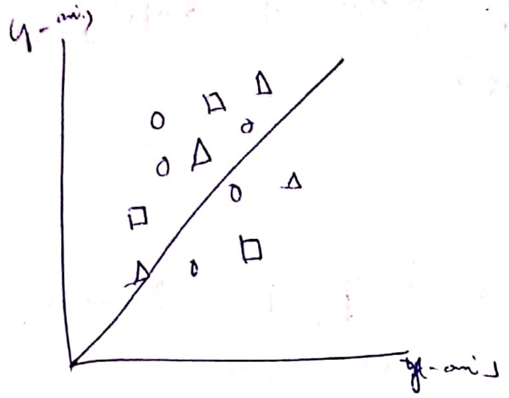
4) multivariate Graphical

Displays relationships b/w two or more sets of data. The most used graphic is

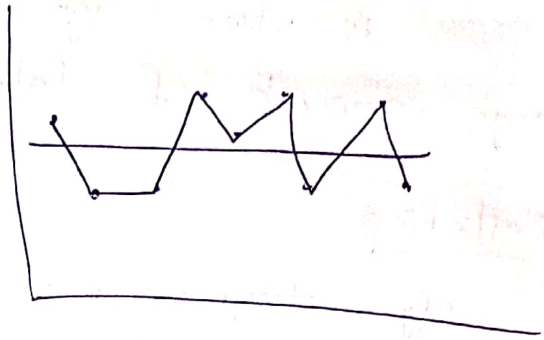
- Bar plot



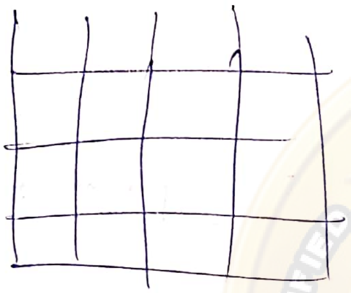
• Scatter plot



• Run chart

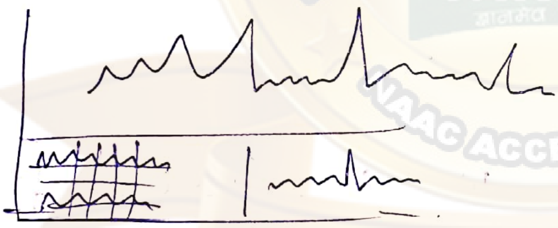


• heat map



• Bubble chart

multivariate



Basic Tools.

R — S/W Env, statistical Computing

Python - interpreted, OOPs with dynamic Semantics.

Real World Examples of EDA:-

• Professional sports
    Most successful players and teams

• History -
    To create new data about past Events.

Healthcare:

To store large stores of Medical data. large amount of data in EMR's.

- Marketing-

Why customers are no longer buying a product. or why a particular Campaign is successful.

- Retail
  - fraud detection-
  - space travels
  - food Industry- what is the most popular ~~food~~ food is in each state.
  - Why some customers prefer getting burgers from McDonald's rather than Burger king.

Summary Statistics

In the previous Section we saw ways of Visualizing attributes using plots to start understanding properties of how data is distributed, in data Analysis.

In this section we start discussing statistical, numerical, Summaries of data to Quantify properties.

That one purpose of EDA is to to spot problems in data (as part of data wrangling).

It is the measurements meant to describe data. Examples of summary statistics for a single numerical variable is the mean, median, mode, max, min, range, variance, SD, Skewness etc.

- For categorical variables is the no of distinct counts. The most basic summary statistic for text data is term frequency and inverse document frequency.

- For bivariate data, the SS is linear correlation p value based z-test, t-test anaylisis (or) Analysis of variance.

<u>Visualization</u>

It can be used to explore and describe data.

- Examples of visualizations for numeric data are line charts with error bars, histograms, box and whisker plots for categorical data bar charts. For bivariate data are scatter charts or combinatio charts.

- Tools and libraries which can be used for plotting visualizations - Excel | libre office - weka - matplotlib (python) - Seaborn (python) - grammer of graphics (ggplot2).

Range - Subtract the lowest value from the highest value

$R = H - L$

Summarize (min_depth = min (depth), max_depth = max(depth))

When paired with measures of central tendency, the range can tell you about the span of the distribution.

Centeral Tendency.

Centeral or typical value for a probability distribution Measures of centeral Tendency are often called 'averages'

Refering to the centeral location of the distribution.

Mean, mode, median.

↓ ↓ L Middle value of observations.

$\mu = \frac{\Sigma x}{n}$ Most frequently occurred value in the dataset.

ggplot (aes (x = depth)) + geom_histogram (bins = 100) +
geom_vline (aes (xintercept = median (depth)),
color = "red")

→ philosophy of EDA:-

The study of a fundamental nature of knowledge, reality and existence, especially when considered as an acedemic discipline.

Systematized study of general and fundamental Questions, those about Existence, reason, knowoledge, values, mind and language.

EDA is an philosophy for data analysis that employs a variety of techniques. (mostly graphical to maximize insight into a dataset.

Philosophy of EDA

It is not identical (similar) to statistical graphics although the two terms are used almost interchangeably.

Statistical Graphics is a collection of techniques.

and data Analysis
↳ procedures generally yield their output in numeric (or) tabular form.

Graphical techniques allow such results to be displayed in some sort of pictorial form.

EDA is relies heavily on such techniques.

They can also provide insight into a data set to help with testing assumptions,

Model selection - Variance b/w observed and implied data using Correlation & Covariance

Estimator selection

relationship Indentification [investments collections]

outlier detection

Covariance → Measure of joint variability of two random variables

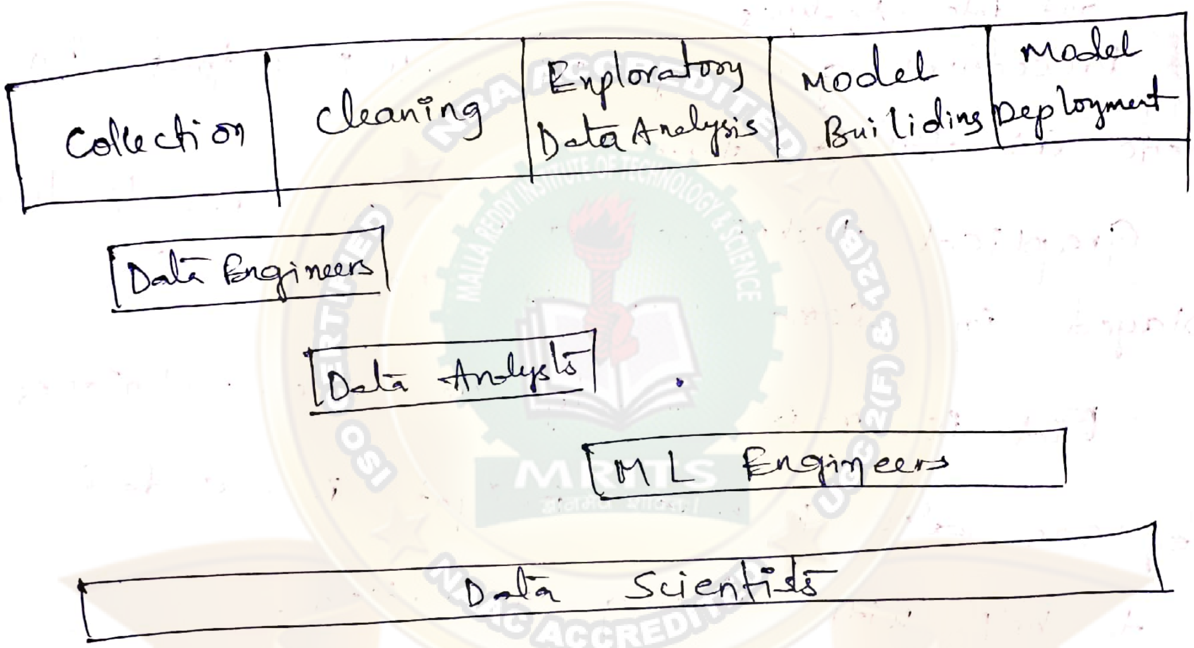
Correlation → A mutual r/s b/w two or more things.

→
- EDA is not a mere collection of techniques.
- philosophy as to how we dissect (Analyse) a dataset, what we look for, how we look, and how we interpret.

→ The Data Science Process

The data science process is a systematic approach to solving a data problem.

It provides a structured framework for articulating your problem as a question, deciding how to solve it, and then presenting the solⁿ to stakeholders.



Framing the problem
understanding and framing the problem is the first step of DS life cycle.
— The framing will help to build an effective model that will have a positive impact on the organisation.

Collecting the data (CRM - customer relationship Management)
Interaction with customers.
— The next step is to collect the right set of data. Roughly 2.5 quintillion bytes of data created everyday come in unstructured formats.

we will Extract the data and export it into a usable format, such as a CSV or JSON file.

Cleaning Data

Most of the data we collect during the collection of phase will be unstructured, irrelevant and unfiltered.

Bad data produces Bad results,

cleaning data Eliminates duplicates, and null values, Corrupt data, inconsistend data types, invalid entries, missing data and improper formatting.

EDA:-

We have large amount of organized, high-Quality data; we can uncover valueable insights that will be useful in the nextphase of DS Lifecycle.

Model Building and Deployment:-

[Highlights the value in Strategic business planning & operation).

Modelling data — we will use ML, Statistical models and algorithms to extract high value insights and predictions.

lastly we will communicate your findings to stakeholders. Every data scientists needs to build their report of visualization skills to do this.

Stakeholders are mainly intreested in what your results means for their organisation, they won't care about the complex back end work that was used to build your model.

Example:- (we can make better decisions).

solving a problem for the VP sales of your company. you should ask Questions like

- who are the customers?
- why are they buying our product
- How do we predict if a custome is gng to buy a product?
- How much money will we lose if we don't actively sell the product to these groups?

By above Questions the sales reveals that they want to understand why certain segments of customers bought less than expected.

Their end goal might be to determine whether to continue to invest in these segments.

Examples

1. Google ~~Analytics~~
2. Demand prediction for the manufacturing industry. (optimizing supply chains & delivering orders).
3. Recommendation systems in Marketing & advertising
4. credit Scoring for financial Institutions.

(NPL - Non performing loans - The loans that haven't been settled for at least 90 days.

European Banking federation informs an avg 3.74% are NPLs. [Estimate the loan debtor's creditworthiness and predict which loans can become NPL's in future)

5. Predictive Analysis in health care

- Improve patient care
- Improve supply chain Efficiency and pharamaceutical logistics.
- Initiate therapy at an early stage.

6. Weather predictions in Agriculture sector.

- The direction and speed of the wind
- Humidity relative
- Max, Min dew point temperatures.

Adv:

- We can make better decisions

ML can analyze millions of bytes of given data within seconds.

- Increasing Sales.

Case study:

Online Real Estate firm | price.

Predicting or estimating selling price of a property can be of great help when making important decisions such as purchase of a home or real estate as an investment vehicle.

→ It can increase the negotiation capacity both for the buyer and for the seller when they having an estimate of the value of the property.

In this case study we will go through the following points.

1) Defining a problem:

- Importance of being able to predict the price of the property.
- The problem we want to face in this case is Regression task.
- We have a list of properties that have been sold in the past like area, location, no. of washrooms etc.
- we must predict continuous value.

2) Acquiring the data

- ID - unique property identifier
- Country
- city
- province_department
- operation_type - Type of business
- property_type - apartment / villa
- Rooms.
- bedrooms, Surface_total, currency, price

3) Competition configuration and requirements.

- Competition start date
- End of the public phase - First subdivision
- End of the private phase - Second & last
- Description of the details of the competition.

5) Selection of a winning Model

6) Deploying an API in production
↓
Application programming Interface.

→ Company has a web platform, to connect these predictions.

The form we fill in our web page directly communicates and via API with the server of company and immediately returns the predicted sale price.

7) Deploy the model in visual app (streamlit)

8) Conclusion

→ As we can see, with the competitions we cover the ML process as we do it hand in hand with you trying to solve a problem with data science.

Three Basic Machine Learning Algorithms:-

Linear Regression

It is a machine learning algorithm based on supervised learning. It performs a Regression task.
↓
Train the machine using data which is labelled. It means some data is already tagged with the correct answer. It takes place in the presence of a supervisor

Unsupervised learning:-

Where you do not need to supervise the model. Instead you need to allow the model to work on its own to discover information.

It mainly deals with the unlabelled data.

Regression task:- [Gives a continuous o/p. If we are used to build a model that predicts the future outcome where the o/p will be continuous.

It helps in Qualifying the relationship between the interrelated economic variables.

It is the prediction of the state of an outcome variable at a particular timepoint with the help of other correlated independent variables.

data set divided into { dependent } - output ; independent } - Input — the model data which is trained.

Check that whether these is no correlation b/w these two.

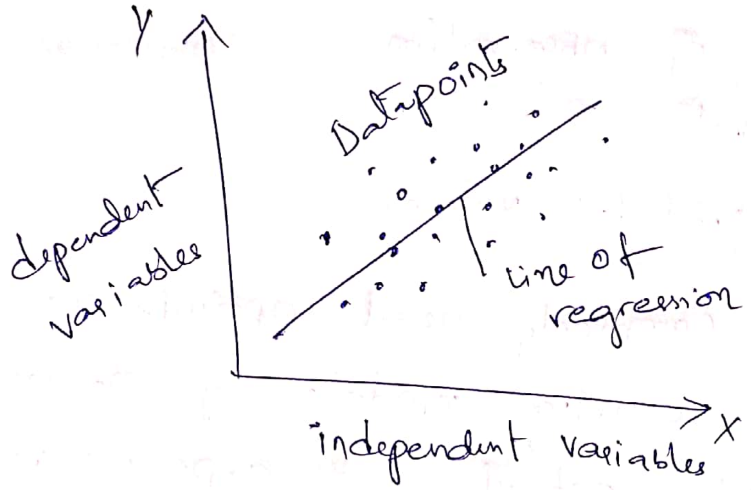
SLR - we should draw a line point b/w the datapoints the distance b/w the points and the line should be low. Then it can predicts.

Linear Regression in ML

It is one of the easiest and most popular ML algorithms. It is a statistical method that is used for predictive analysis.

It makes predictions for continuous/ real or numeric (value) variables such as sales, salary, age, product price, etc.

- It shows the linear r/s b/w dependent (Y) and 1 or more independent variables (X).



Mathematically, we can represent a linear regression as

$$y = a_0 + a_1 * X + \varepsilon \text{ (epsilon)}$$

Y - Dependent Variable (Target)

X - Independent Var (predictor)

a_0 = intercept of the line

a_1 = linear regression Coefficient (slope)

ε = random error.

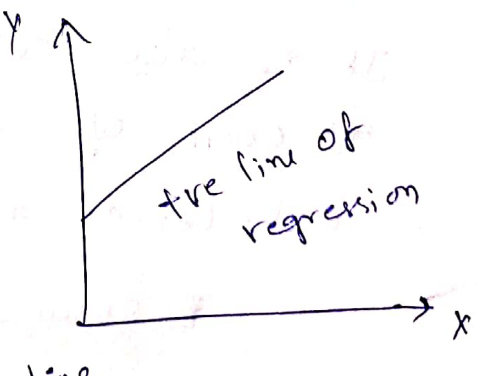
Linear Regression Line:-

showing r/s b/w dependent & independent.

- Positive - $y = a_0 + a_1 x$

dependent ↑ Y-axis

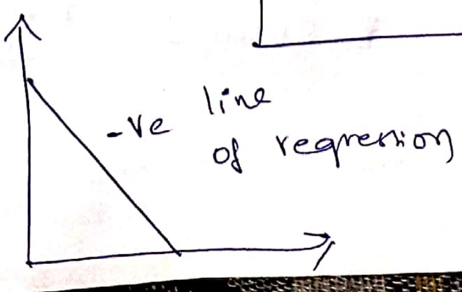
independent ↑ X-axis



- Negative - $y = -a_0 + a_1 x$

dependent ↓, Y-axis

independent ↑ - X-axis



MSE - Mean squared Error.

$$MSE = \frac{1}{N} \sum_{i=1}^{n} (y_i - (a_1 x_i + a_0))^2$$

N - Total No. of observation

y_i = actual value

$(a_1 x_i + a_0)$ = predicted value.

Gradient Descent - commonly used optimization algorithm to train machine learning models by means of minimizing errors b/w actual and expected results.

Model performance:

The Goodness of fit determines how the line of regression fits the set of observations.

The process of finding the best model out of various models is called optimization.

R-squared Method

- Statistical method that determines the goodness of fit
- It measures the strength of the r/s b/w the dependent and independent variable.
- It is also called a coefficient of determination (or) coeff of multiple determination
- Calculated as

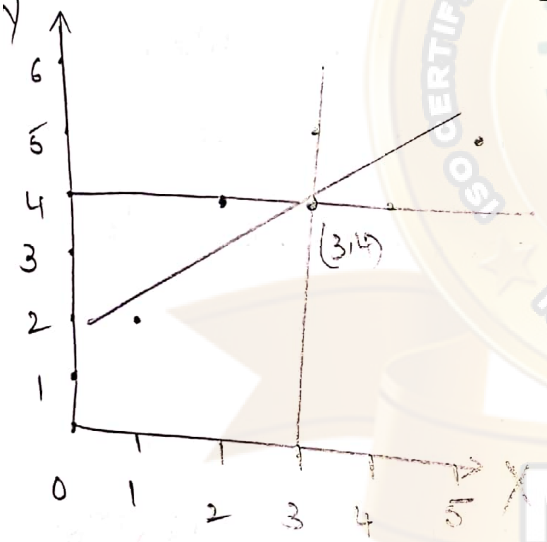
$$R\text{-Squared} = \frac{\text{Explained variation}}{\text{Total variation}}$$

Eg:- $Y = b_0 + b_1 X$

Independent variable(x)	dependent Variable(y)	$\hat{y}$	$(x-\bar{x})$	$\hat{y}-y$	$(y-\bar{y})$	$(\hat{y}-y)^2$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
1	2	2.8	-2	0.8	-2	0.64	4	4
2	4	3.4	-1	0.6	0	0.36	1	0
3	5	4	0	-1	1	1	0	0
4	4	4.6	1	0.6	0	0.36	1	0
5	5	5.2	2	0.2	1	0.04	4	2
Mean 3	4					2.4	10	6

Slope of the line will be 1.

$$b_1 = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{6}{10} = 0.6$$



$y = b_0 + b_1 X$

$4 = b_0 + 0.6(3)$

$4 = b_0 + 1.8$

$-1.8 \quad -1.8$

$2.2 = b_0$

$y = b_0 + b_1 X$

$b_0 = 2.2$

$b_1 = 0.6$

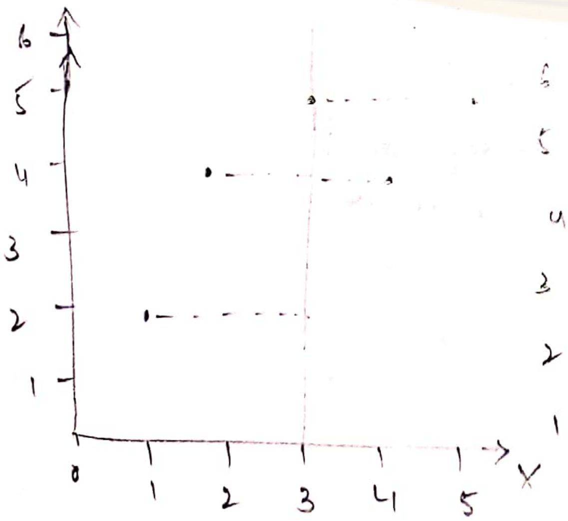
$\hat{y} = 2.2 + 0.6x$

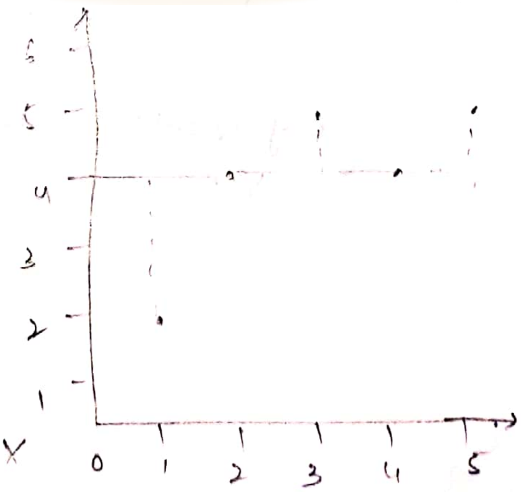
$= 4$

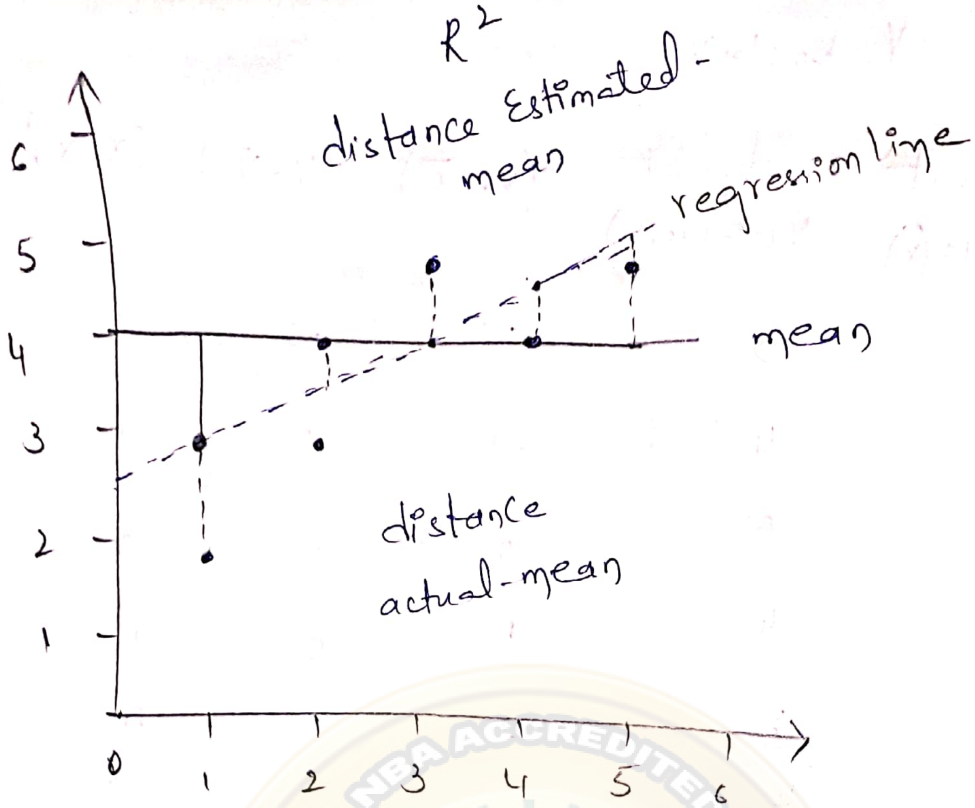
$$\text{standard Error of estimation} = \sqrt{\frac{\Sigma(\hat{y}-y)^2}{n-2}}$$

$$= \sqrt{\frac{2.4}{5-2}}$$

$$= 0.89$$







Simple LR and Multiple LR:-

dependent variable must be continuous | real value.

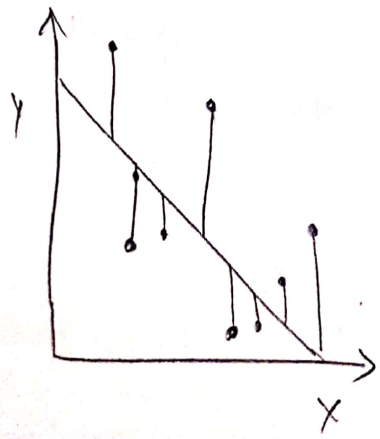
If contains only one predictor, or X variable, predicting the response or Y-variable.

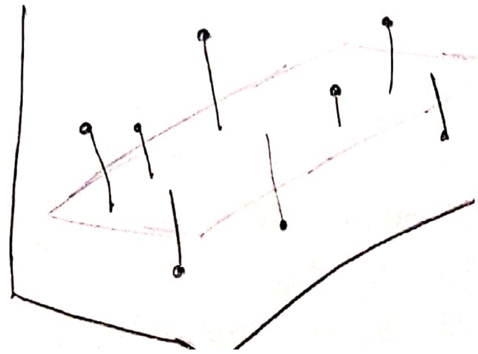
– ML algorithm to predict a dependent variable with two or more predictors.

Simple	Multiple
$y = b_0 + b_1 X_1$	$y = b_0 + b_1 X_1 + b_2 X_2 - - - - - + b_n + X_n$

y → dependent var

$b_1 X_1 + b_2 X_2$ → Independent variables.





K-Nearest Neighbors (KNN):- [Advantages of Linear Regression]

- predictions of umbrellas sold based on the rain happened in the area.
- predictions of AC sold based on the temperature in summer.

— It Measures the Similarity.

- KNN is a supervised learning algorithm that can be used for both classification and regression problems.
- It is one of the simplest ML algorithms based on supervised learning technique.
- It assumes the similarity b/w the new case/data and available cases and put the new case into the category is most similar to available categories.
- It is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- It is also called a lazy learner algorithm
↳ not learn in training set, directly performs in classification

Because it does not learn from the training set immediately instead it stores the dataset and the time of classification.

KNN is a non-parametric algorithm. means Not having any predictions / assumptions.

eg:- classification

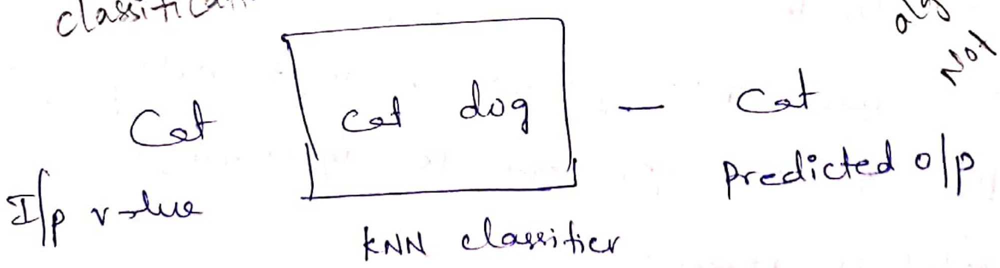


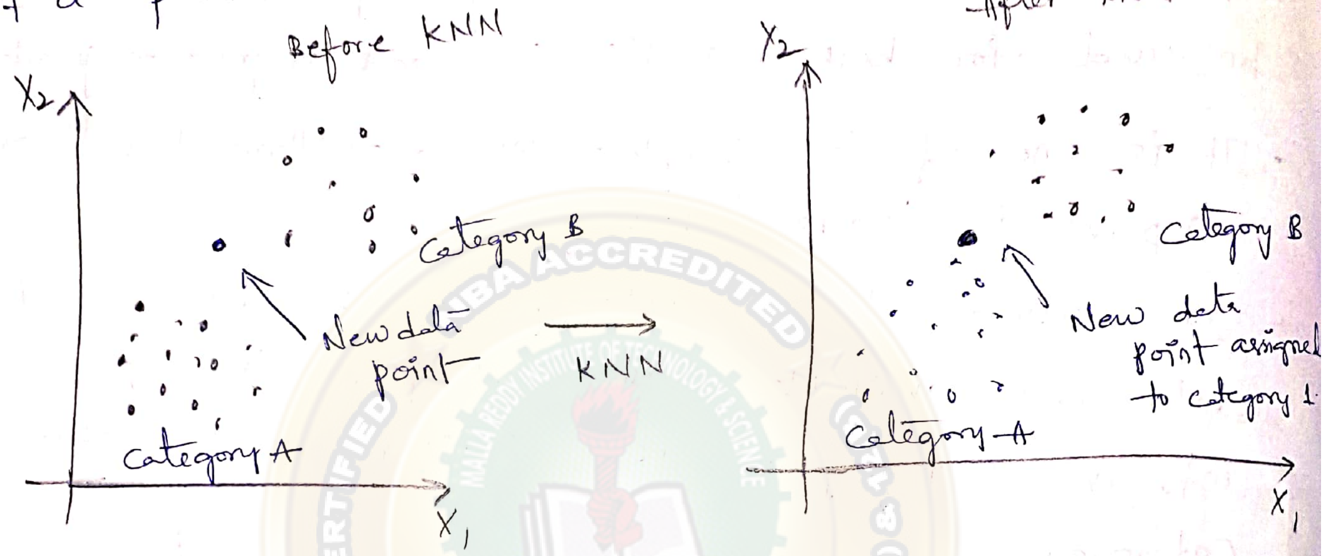
Image of a creature that looks similar to cat and dog. we want to know whether it is a cat or dog.

It works on similarity measure.

Why do we need a KNN algorithm?

There are two categories i.e. category A and B. we have a new data point X_1, so this data point will lie in which of these categories.

we can easily identify the category of class of a particular dataset.



How does KNN work?

- Select the no ~~of~~ k of the neighbors
- Calculate the Euclidean distance of k no. of neighbors.
- Take the k nearest neighbors as per the Euclidean distance.
- Among these k neighbors, count the no of data points in the each category.
- Assign the new data points to that category for which the no of neighbor is maximum.
- Our model is ready.

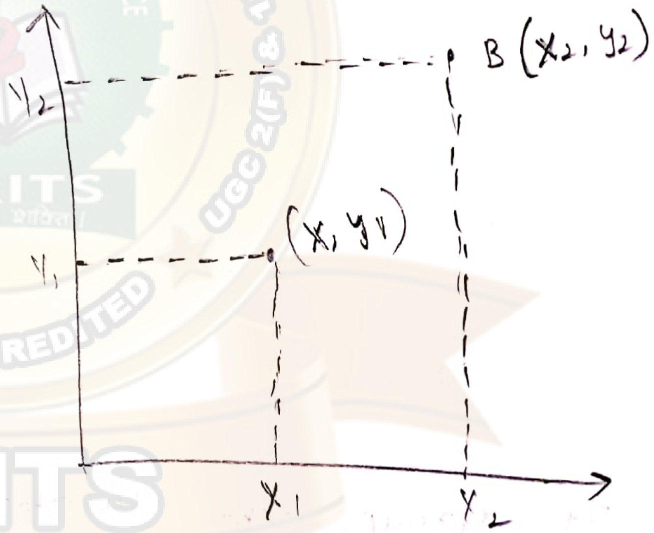
eg:

Training Instance	x_1	x_2	o/p	Distance (d)	Neighbor Rank	d^2	vote	Rank
I_1	7	7	0	$\sqrt{(7-3)^2+(7-7)^2} = 4$	3	16	$1/16 = 0.06$	3
I_2	7	4	0	$\sqrt{(7-3)^2+(4-7)^2} = 5$	4	25	0.04	4
I_3	3	4	1	$\sqrt{(3-3)^2+(4-7)^2} = 3$	1	9	0.11	1
I_4	1	4	1	$\sqrt{(1-3)^2+(4-7)^2} = 3.6$	2	12.96	0.08	2.

test instance $t_1 = (3, 7) - 1$

three nearest neighbors (k=3)

Eucledian distance b/w A_1 and $B_2 = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$



Adv -

- Simple to implement
- It is robust to the noisy training data
  strong
- It can be more effective if the training data is large.

disadv

- Determine the value of k which may be complex some time
- The computation cost is high because of calculating

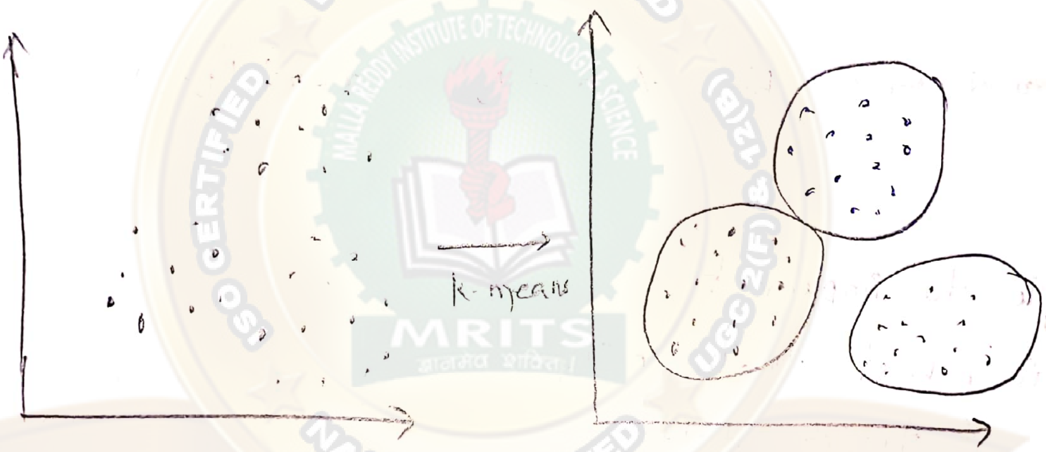
Applications

- Text Mining
- Agriculture
- finance
- Medical
- Facial Recognition
- Recommendation Systems (Amazon, Netflix) etc.

Scanned by CamScanner

K- Means algorithm:-

- It is an clustering algorithm.
  ↓
  A group of similar things that are close together.
- It is an unsupervised learning algorithm.
- It is used to find groups which have not been explicitly labeled in the data.
- It is used to solve the clustering problem in ML
- Algorithm that segments data into clusters to study similarities.



Before k Means   After k-means

- It groups the unlabeled dataset into different clusters.

Here k defines the no. of pre-defined clusters that need to be created in the process.

if k=2, there will be two clusters
k=3, " three "

- It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset without any training.

- It is a centroid based alg, where each cluster is associated with a centroid.

The Main aim of this algorithm is to minimize the sum of distances b/w the datapoint and their corresponding clusters.

eg:-

Step 1: - Take the Mean value

step 2: find the nearest no. of Mean and put in cluster.

step 3: Repeat the ① & ② until we get Same Mean.

Datapoints - $k = \{2, 4, 6, 9, 12, 16, 20, 24, 26\}$

No. of clusters = 2 $\{4, 12\}$ — Random no's

$k_1 = \{2, 4, 6\}$ $k_2 = \{9, 12, 16, 20, 24, 26\}$

$= \frac{2+4+6}{3}$ $= \frac{9+12+16+20+24+26}{6}$

$= 4$ $= 18 \ (17.8)$

$\rightarrow \{4, 18\}$

$k_1 = \{2, 4, 6, 9\}$ $k_2 = \{12, 16, 20, 24, 26\}$

$= \frac{21}{4} = 5.25$ $= \frac{98}{5}$

$= 5$ $= 19.6$

$= 20$

$\rightarrow \{5, 20\}$

$k_1 = \{2, 4, 6, 9, 12\}$ $k_2 = \{16, 20, 24, 26\}$

$= 21.5$

$= 6.6$ $= 22$

$= 7$

Scanned by CamScanner

$\rightarrow (7, 22)$

$k_1 = \{2, 4, 6, 9, 12\}$ $\quad$ $k_2 = \{16, 20, 24, 26\}$

$= 6.6$ $\quad$ $= 21.5$

$= 7$ $\quad$ $= 22$

The Mean value is equal in previous and present iterations.

By Euclidean Distance $\rightarrow$ How to form clusters

$$= \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2} = \sqrt{(x - x_i)^2 + (y - y_i)^2}$$

SN	X	Y
1	170	60
2	160	55
3	180	75
4	150	50
5	175	65
6	190	85

clusters $\Rightarrow k = 2$

C	X	Y
C_1	170	60
C_2	160	55

Take 3

Initial centroids

For Row 3

Ed from C_1 to 3

$$Ed = \sqrt{(180-170)^2 + (75-60)^2}$$
$$= \sqrt{(10)^2 + (15)^2}$$
$$= \sqrt{100 + 225}$$
$$= \sqrt{325} = 18.02$$

Ed from C_2 to 3

$$Ed = \sqrt{(180-160)^2 + (75-55)^2}$$
$$= \sqrt{20^2 + 20^2}$$
$$= \sqrt{400 + 400}$$
$$= 28.28$$

New centroids

C	X	Y
C_1	175	67.5
C_2	160	55

$$\frac{\overset{C_1}{170} + \overset{C_3}{180}}{2} = \frac{75 + 60}{2}$$

C_1 to ④

$$Ed = \sqrt{(175-150)^2 + (67.5-50)^2}$$

$$= \sqrt{25^2 + (17.5)^2}$$

$$= \sqrt{931.25} = 30.51$$

C_2 to ④

$$Ed = \sqrt{(160-150)^2 + (55-50)^2}$$

$$= \sqrt{10^2 + 5^2}$$

$$= \sqrt{100+25} = \sqrt{125}$$

$$= 11.18$$

Row 4 is nearest position to cluster 2.

New Centroid

C	X	Y
C_1	175	67.5
C_2	155	52.5

$$\frac{160+150}{2} = 155 \qquad \frac{55+50}{2} = \frac{105}{2} = 52.5$$

C_1 to ④

$$Ed = \sqrt{(175-175)^2 + (67.5-65)^2}$$

$$= \sqrt{0 + (2.5)^2}$$

$$= 2.5$$

C_2 to ⑤

$$Ed = \sqrt{(175-155)^2 + (65-52.5)^2}$$

$$= \sqrt{(20)^2 + (12.5)^2}$$

$$= 23.58$$

New centroids

C_1	X	Y
C_1	175	66.25
C_2	155	52.5

$$= \frac{175+175}{2} \qquad \frac{67.5+65}{2}$$

C_1 to ⑥

$$Ed = \sqrt{(175-190)^2 + (66.25-85)^2}$$

$$= \sqrt{(15)^2 + (18.75)^2}$$

$$= 24.01$$

C_2 to ⑥

$$Ed = \sqrt{(155-190)^2 + (52.5-85)^2}$$

$$= \sqrt{(35)^2 + (32.5)^2}$$

$$= 47.76$$

C	X	Y
C_1	182.5	75.6
C_2	155	52.5

$$\frac{175+190}{2} = 182.5$$

$$\frac{66.25+85}{2} =$$

$C_1 = \{1, 3, 5, 6\}$

$C_2 = \{2, 4\}$

01/11/2022

UNIT-III

One more Machine learning Algorithm :-

Support Vector Machine Algorithm :-

SVM is one of the most popular supervised learning algorithms. which is used for classification and regression problems. process of categorizing a given set of data into classes.

Primarly, it is used for classification problems in Machine learning.

The goal of SVM alg is to create the best line or decision boundary ~~that~~ can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.

This best decision boundary is called a hyperplane.

SVM chooses the entreme points / vectors that help in creating the hyperplane.

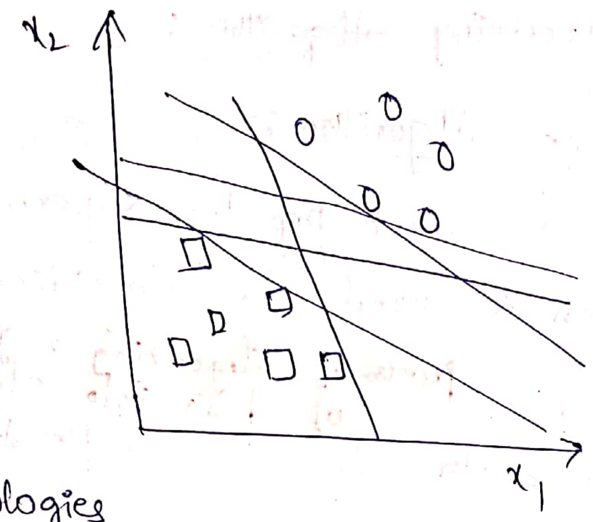
In this algorithm, we plot each data item as a point in n dimensional space (where n is the number of features) with the ~~(fea)~~ value of each feature being the value of a particular coordinate.

Hyperplane :-

Defined as an n-1 dimensional Eucledian space that separates as n- dimensional Eucledian space that into two disconnected parts or classes.

The optimal hyperplane - Distance of hyperplane is equal from both the nearest data points of two classes.

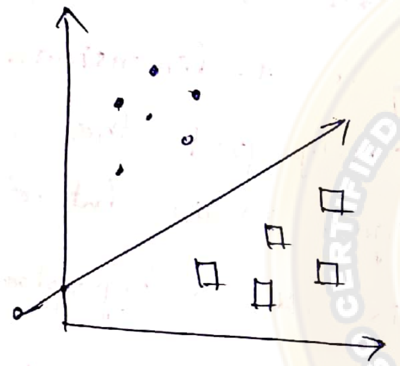
If data is perfectly separable,



The selection of the best hyperplane is "optimal".

Terminologies

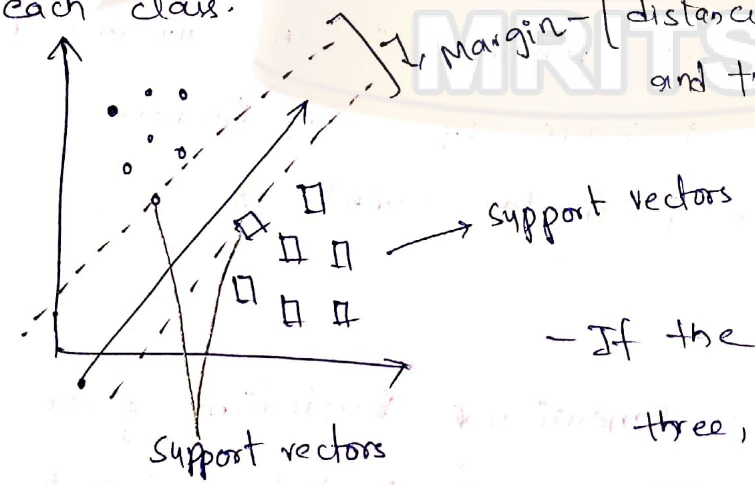
1. Maximal Margin classifier



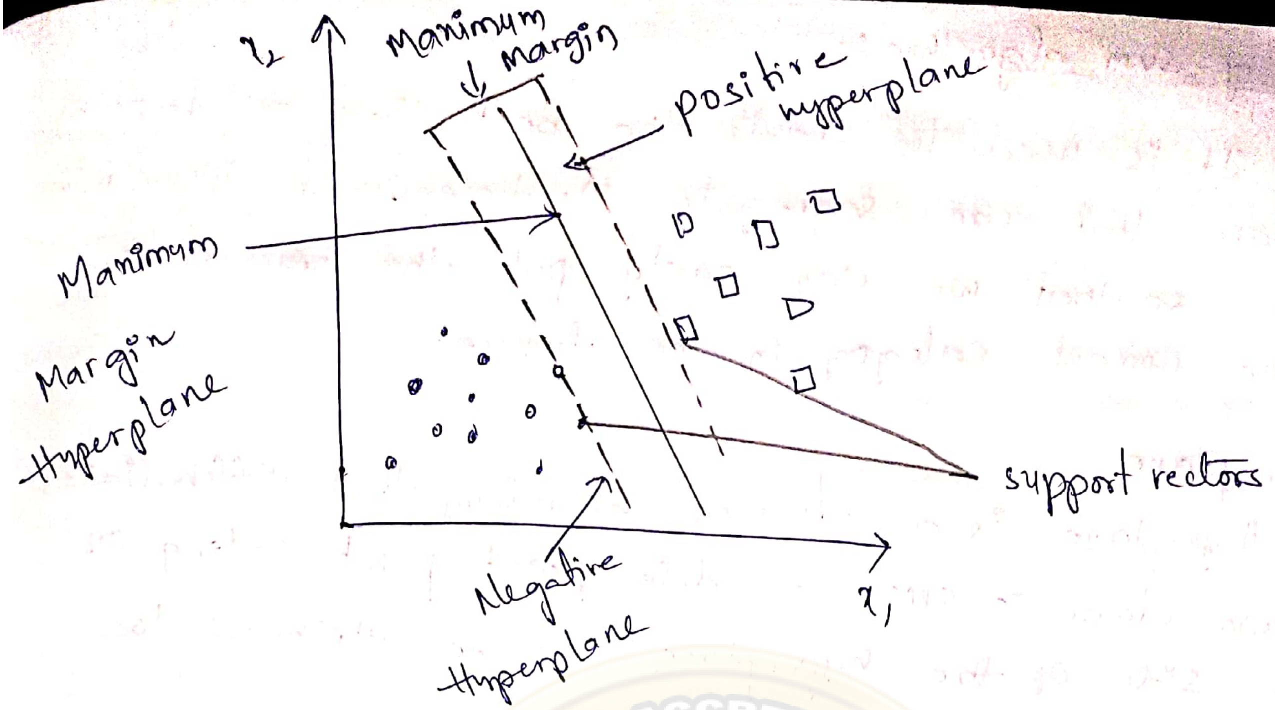
- It looks like a straight line when we draw a decision boundary b/w two classes.

But how do we know where to draw a hyperplane?

- Draw a random hyperplane and then you check the distance b/w the plane and the closest data points from each class.



- If the no. of i/p features is three, then hyperplane becomes a 2-D plane.



eg:-

New data

Cat → Model Training → Prediction → o/p

Past Labelled dog

↓

we are giving the i/p [like how the cat looks like, how the dog looks like]

Applications : Face Recognition

- Hand Writing Recognition. Validating signs on documents
- Email classification
- classification of images
- Bioinformatics - classification of genes

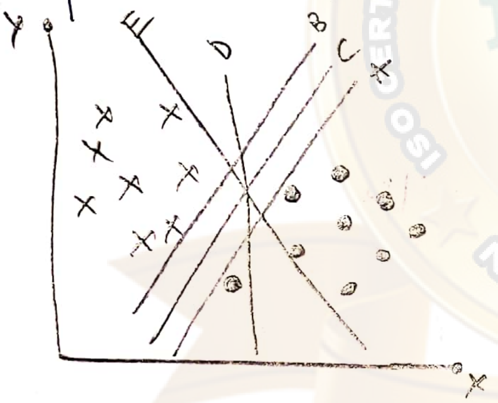
Use of SVM algorithm:-

It is used to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data in the correct category in the future.

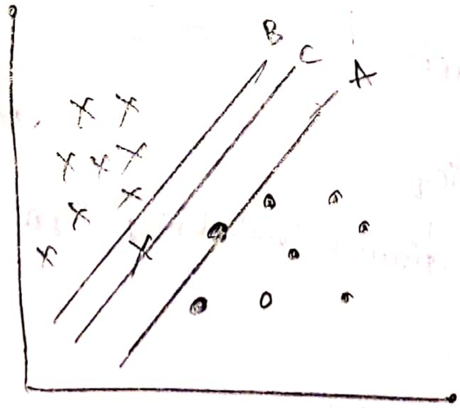
Hyper-plane

Hyperplane is a decision boundary that differentiates the two classes in SVM. A data point point calling on either side of the hyperplane can be attributed to different classes.

· Maximum classification - The selected line must be able to successfully segregate all the data points into the respective classes.



Best Separation:- It means, we must choose a line such that it perfectly able to separate the points.

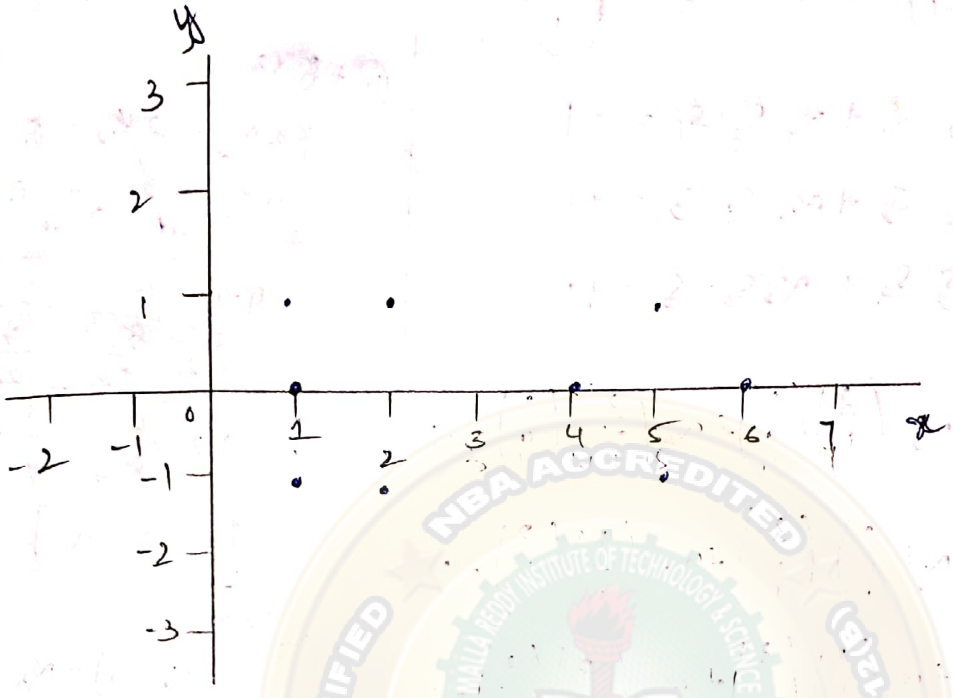


In both cases line 'c' is successfully classifying the all data points why?

Example

Plot hyperplane of following points

(1, 1) (2, 1) (1, −1) (2, −1) (4, 0) (5, 1) (5, −1) (6, 0)

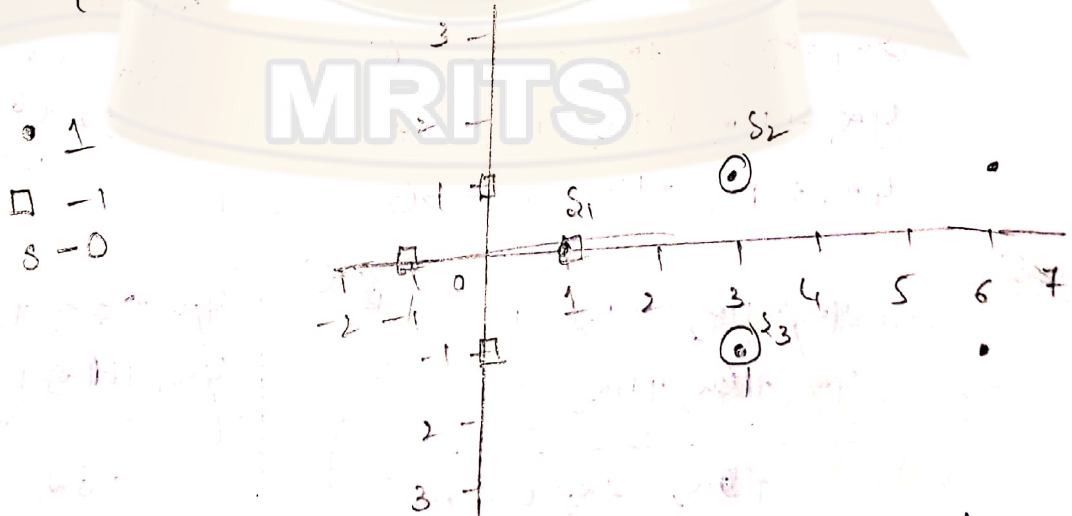


Linear Example Solved

Positively labeled data points

$$\left\{ \begin{pmatrix}3\\1\end{pmatrix}, \begin{pmatrix}3\\-1\end{pmatrix}, \begin{pmatrix}6\\1\end{pmatrix}, \begin{pmatrix}6\\-1\end{pmatrix} \right\}$$

negatively - $\left\{ \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}0\\-1\end{pmatrix}, \begin{pmatrix}-1\\0\end{pmatrix} \right\}$



$S_1 = \begin{pmatrix}1\\0\end{pmatrix}$

$S_2 = \begin{pmatrix}3\\1\end{pmatrix}$

$S_3 = \begin{pmatrix}3\\-1\end{pmatrix}$

Each vector is augmented with a 1 as a bias point.

↓

Make (something) greater by adding to it. Increase.

So $S_1 = \begin{pmatrix}1\\0\end{pmatrix}$ then $\tilde{S}_1 = \begin{pmatrix}1\\0\\1\end{pmatrix}$

$S_2 = \begin{pmatrix}3\\1\end{pmatrix}$ then $\tilde{S}_2 = \begin{pmatrix}3\\1\\1\end{pmatrix}$

$S_3 = \begin{pmatrix}3\\-1\end{pmatrix}$ then $\tilde{S}_3 = \begin{pmatrix}3\\-1\\1\end{pmatrix}$

$$\alpha_1 S_1\cdot S_1 + \alpha_2 S_2\cdot S_1 + \alpha_3 S_3\cdot S_1 = -1$$
$$\alpha_1 S_1\cdot S_2 + \alpha_2 S_2\cdot S_2 + \alpha_3 S_3\cdot S_2 = +1$$
$$\alpha S_1\cdot S_3 + \alpha_2 S_2 S_3 + \alpha_3 S_3\cdot S_3 = +1$$

$$\alpha_1\begin{pmatrix}1\\0\\1\end{pmatrix}\begin{pmatrix}1\\0\\1\end{pmatrix} + \alpha_2\begin{pmatrix}3\\1\\1\end{pmatrix}\begin{pmatrix}1\\0\\1\end{pmatrix} + \alpha_3\begin{pmatrix}3\\-1\\1\end{pmatrix}\begin{pmatrix}1\\0\\1\end{pmatrix} = -1$$
$$\alpha_1\begin{pmatrix}1\\0\\1\end{pmatrix}\begin{pmatrix}3\\1\\1\end{pmatrix} + \alpha_2\begin{pmatrix}3\\1\\1\end{pmatrix}\begin{pmatrix}3\\1\\1\end{pmatrix} + \alpha_3\begin{pmatrix}3\\-1\\1\end{pmatrix}\begin{pmatrix}3\\1\\1\end{pmatrix} = 1$$
$$\alpha_1\begin{pmatrix}1\\0\\1\end{pmatrix}\begin{pmatrix}3\\-1\\1\end{pmatrix} + \alpha_2\begin{pmatrix}3\\1\\1\end{pmatrix}\begin{pmatrix}3\\-1\\1\end{pmatrix} + \alpha_3\begin{pmatrix}3\\-1\\1\end{pmatrix}\begin{pmatrix}3\\-1\\1\end{pmatrix} = 1$$

$$\alpha_1(1+0+1) + \alpha_2(3+0+1) + \alpha_3(3-0+1) = -1$$
$$\alpha_1(3+0+1) + \alpha_2(9+1+1) + \alpha_3(9-1+1) = 1$$
$$\alpha_1(3+0+1) + \alpha_2(9-1+1) + \alpha_3(9+1+1) = 1$$

$$2\alpha_1 + 4\alpha_2 + 4\alpha_3 = -1 \longrightarrow (1)$$
$$4\alpha_1 + 11\alpha_2 + 9\alpha_3 = 1 \longrightarrow (2)$$
$$4\alpha_1 + 9\alpha_2 + 11\alpha_3 = 1 \longrightarrow (3)$$

(2) − (3)

$$\cancel{4\alpha_1} + 11\alpha_2 + 9\alpha_3 = 1$$
$$\cancel{4\alpha_1} + 9\alpha_2 + 11\alpha_3 = 1$$
$$- \quad - \quad - \quad -$$
$$+2\alpha_2 - 2\alpha_3 = 0 \longrightarrow (4)$$

(4) × 4

(1) × 1 − (2)

$$\cancel{4\alpha_1} + 8\alpha_2 + 8\alpha_3 = -2$$
$$\cancel{4\alpha_1} + 11\alpha_2 + 9\alpha_3 = 1$$
$$- \quad - \quad - \quad -$$
$$-3\alpha_2 - \alpha_3 = -3 \rightarrow (5)$$

(4) + (5) × 2

$$+2\alpha_2 - \cancel{2\alpha_3} = 0$$
$$-6\alpha_2 - \cancel{2\alpha_3} = -6$$
$$+ \quad + \quad +$$
$$8\alpha_2 = 6$$
$$\boxed{\alpha_2 = \frac{\cancel{8}\,3}{4\,2} = \frac{3}{2}}$$
$$\alpha_2 = \frac{\cancel{6}\,3}{\cancel{8}\,4} = \frac{3}{4} = 0.75$$
$$\boxed{\alpha_2 = 0.75}$$

Sub α_2 in (5)

$$-3(0.75) - \alpha_3 = -3$$
$$-2.25 - \alpha_3 = -3$$
$$-\alpha_3 = -3 + 2.25$$
$$+\alpha_3 = +0.75$$
$$\boxed{\alpha_3 = 0.75}$$

Sub α_2, α_3 in (1)

$$2\alpha_1 + 4(0.75) + 4(0.75) = -1$$
$$2\alpha_1 + 3 + 3 = -1$$
$$2\alpha_1 + 6 = -1$$
$$2\alpha_1 = -1 - 6$$
$$2\alpha_1 = -7$$
$$\alpha_1 = -\frac{7}{2}$$
$$\boxed{\alpha_1 = -3.5}$$

(2) × (1)

$$\cancel{4\alpha_1} + 3\alpha_2 + 4\alpha_3 = -1$$
$$\cancel{4\alpha_1} + 11\alpha_2 + 9\alpha_3 = 1$$
$$- \quad - \quad -$$
$$-8\alpha_2 - 5\alpha_3 = -2 \rightarrow (5)$$

$$\cancel{8\alpha_2} - 8\alpha_3 = 1$$
$$-\cancel{8\alpha_2} - 5\alpha_3 = -2$$
$$-13\alpha_3 = -1$$
$$\alpha_3 = \frac{13}{13} = 0.759$$

$0.75 \times 3 = 4.25$; 3 ; 2 ; 3.00 ; 4.25 ; .75 ; $-3 + 4.25$; 3.00 ; 0.75

$$2\alpha_2 - 2\alpha_3 = 1$$

$$2\alpha_2 - 2(0.75) = 1$$

$$2\alpha_2 - 1.5 = 1$$

$$2\alpha_2 = 1 + 1.5$$

$$2\alpha_2 = 2.5$$

$$\alpha_2 = \frac{2.5}{2}$$

$$= 1.25$$

$\alpha_1 = -3.5$

$\alpha_2 = 0.75$

$\alpha_3 = 0.75$

weight vector) $\tilde{w} = \sum_i \alpha_i \tilde{s}_i$

$$= -3.5\begin{pmatrix}1\\0\\1\end{pmatrix} + 0.75\begin{pmatrix}3\\1\\1\end{pmatrix} + 0.75\begin{pmatrix}3\\-1\\1\end{pmatrix}$$

$$= \begin{pmatrix}1\\0\\-2\end{pmatrix}$$

Hyperplane Eqn. $y = wx + b$

$w = \begin{pmatrix}1\\0\end{pmatrix}$ and $b = -2$

Big problems that SVMs (support vector Machines) be used:

Image classification:

The idea behind this classification is to find the best separating hyperplane b/w. different classes of objects such that it can classify images from various categories in a simplified form.

(i) Face detection - There are seven classes (A to G) with respect to the different types of faces that it can detect.

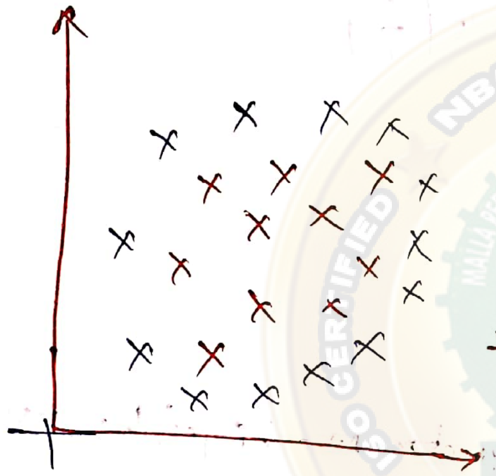
(ii) Detection of Tumors in medical images such as CT Scan and MRI. This is the problem that arises out in practice.

* Speech Recognition :-

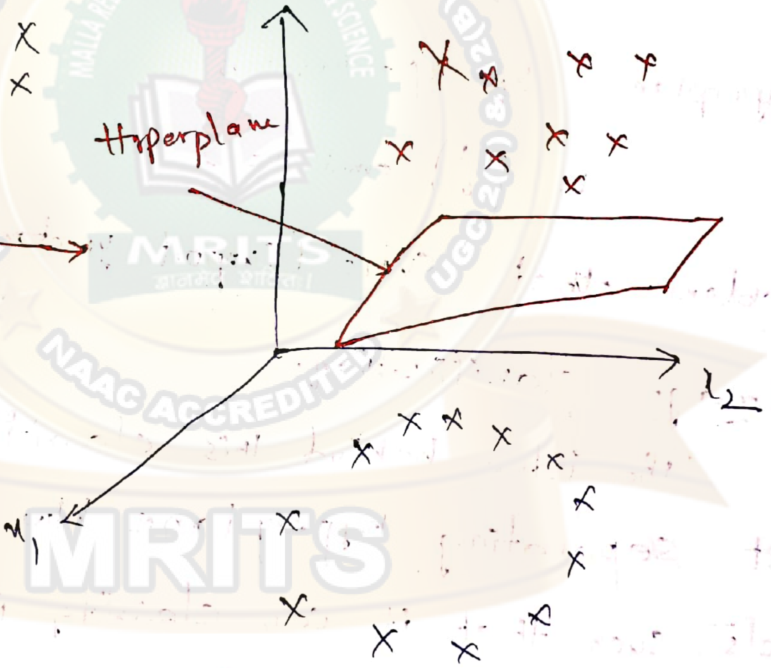
The task of an SVM is to distinguish b/w different words that are being said.

In this application, a set of training data that contain sounds along with their transcriptions into the same words or different words is given as input to the SVM.

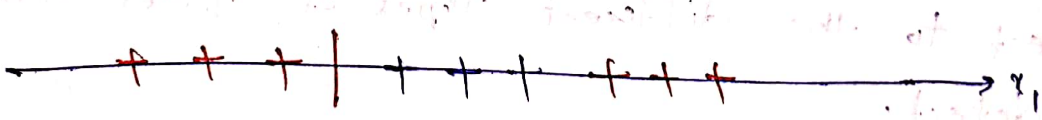
Non-linear data (2-Dimensional space)



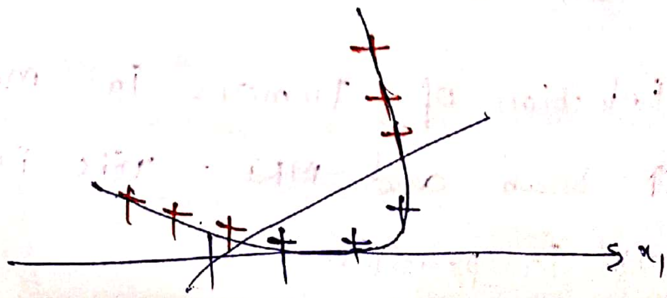
→ By using three-dimensional space



1. Dimensional space



2-Dimensional space



Filtering spam :- [Mail Tester to test the Spammyness of email]

In ML, spam filtering protocols use instance-
based or memory based learning methods to identify
and classify incoming spam emails based on their
resemblance to stored training examples of spam emails.

Filters block unsolicited (not asked for, given or
done voluntarily).
a threat to the
or suspicious emails that are
security of n/w from getting to the computer System.
Also, at the email level, the user can have a
customized spam filter that will block spam emails
in accordance with some set conditions.

Email addresses and phones are targeted by these
spams and sometimes dangerous messages.

In the literature, filtering junk mail has been
tackled in different ways. ML based techniques are
largely used, and allow good performance in general.

When we considered the spam filtering issue,
the first thing we did was to read too many
messages we received lately.

The spam message (french) has a random unique offer
that will expire soon, with a lot of exclamation marks.

What makes the human analysis quick is its
Simplicity. The KISS (Keep it Simple stupid) principle
at it its finest.

Spam filters detect unsolicited, unwanted and virus-infested email (spam) and stop it from getting into email inboxes.

Internet Service providers (ISPs) use spam filters to make sure they aren't distributing spam.

Small-to-medium-size businesses (SMBs) also use spam filters to protect their employees and networks.

Spam filters are applied to both inbound email (email entering the n/w) and outbound email (email leaving the network).

Spam filters can be hosted in the "cloud", on computer servers, or integrated into email s/w such as Microsoft outlook.

Spam filters use "heuristics" methods which means that each mail message is subjected to thousands of predifined rules (algorithms). Each rule assigns a numerical score to the probability of the message being spam, and if the score passes a certain threshold the email is flagged as spam and blocked from going further.

Different spam filters are:

- Content filters - parse the content of messages, scanning for words that are commonly used in spam emails.
- Header filters - examine the email header source to look for suspicious information (such as spammer email addresses)

. Blocklist filters :- Stop emails that come from a blocklist of suspicious IP addresses.

. Rules-based filters - apply customized rules designed by the organization to exclude emails from specific senders, or emails containing specific words in their Subject line.

→ Why is linear Regression and KNN are poor choice for spam filtering?

Linear Regression and K-NN for filtering spam :-

. Spamming has become a time-consuming and expensive issue for which a variety of new directions have recently been Explored.

The solⁿ developed is an offline program that uses the K-Nearest Neighbor (KNN) algorithm and a pre-classified e-mail data set for the learning process.

Why won't linear Regression work for filtering spam?

- In order to use linear Regression, we need a training set of emails where messages have already been labelled with some output variable.

In this case, the tests are either spam or not.

- We could do this by making spam messages labelled by human elevators, which is a logical, but time-consuming, solution.

- If we create a setup, email messages will come through without a mark, so we would yo use the model to predict the labels.

The first thing to remember is that our goal is binary (0 if not spam, 1 if spam) - In case we will not get 0 (or) 1 using linear regression, we would get a number.

Strictly speaking, this choice is not ideal, linear regression is built to model a continuous output, and this is binary.

Why KNN a poor choice for spam filtering:-

What does it means for a spam to be a similar to another? we can compare the strings and count the matches, or number of similar words or anything like that, but that's only going to measure 'similarity' in a very specific sense.

KNN classifier are good whenever there is a really meaningful distance metric.

In the spam case KNN classifiers are going to label as spam things that are "close" to known spams being 'close' in the sense of your distance metric (which will likely be poor).

Therefore, KNN classifiers are only going to filter spams that are really similar to what you are already know. It won't really genearlize properly.

Finally we need a lot of data for a good KNN classifier and running KNN queries for millions or emails a day can be really costly.

There, KNN doesn't work well in this case because there is no good distance metric and it is pretty costly to run.

Naive Bayes Algorithm :-

- It is an supervised learning algorithm, which is based on Baye's theorem and used for solving classification problems.

- It is mainly used in text classification that includes a high dimensional training datasets.

- It helps in building the fast ML models that can make quick predictions.

- It is a probabilistic classifier, which means it predicts on the basis of probability of an object.

Why is it called Naive Bayes ?

· Naive - Because it assumes that the occurrence of a certain feature is independent of the occurrence of other features.

eg. If the fruit is identified on the bases of color, shape, and taste, then red, spherical and

Sweet fruit is recognized as an apple.

Hence each feature individually contributes to identify that it is an apple without depending on each other.

Bayes – It depends on Baye's theorem.

It depends on the conditional probability.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Working of Naive Baye's classifier

eg:- dataset of weather conditions and corresponding target variable play.

steps

- Convert the given dataset into frequency tables
- Generate likelihood table by finding the probabilities
- use Baye's theorem to calculate the posterior probability.

Problem:- If the weather is sunny, then the player should play or not?

outlook	play	outlook	play
Rainy	Yes	Overcast	Yes
Sunny	Yes	Rainy	~~Yes~~ No
Overcast	Yes	Sunny	No
Overcast	Yes	Sunny	~~No~~ Yes
Sunny	No	Rainy	No
Rainy	Yes	Overcast	Yes
Sunny	Yes	Overcast	Yes

Frequency table

weather	Yes	No	
overcast	5	0	– 5/14 = 0.35
Rainy	2	2	– 4/14 = 0.29
sunny	3	2	– 5/14 = 0.35
Total	10	~~5~~ 4	
	10/14 = 0.71	4/14 = 0.29	

Applying Baye's theorem

$$P(Yes \mid sunny) = \frac{P(sunny \mid Yes) * P(Yes)}{P(sunny)}$$

$$= \frac{(3/10)(0.71)}{0.35}$$

$$= 0.60$$

$$P(No \mid sunny) = \frac{P(sunny \mid No) * P(No)}{P(sunny)}$$

$$= \frac{(2/4)(0.29)}{0.35}$$

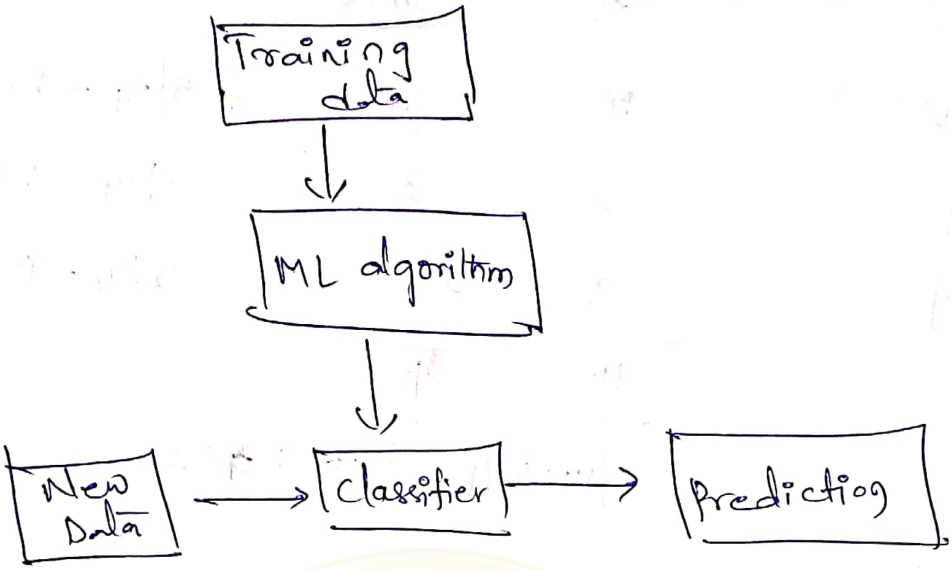
$$= 0.41$$

∴ From above calculation

$$P(Yes \mid sunny) > P(No \mid Sunny).$$

So, if the weather is sunny, the player should play.

Why Naive Bayes works for filtering spam?



Naive Bayes Classifiers work by correlating the use of tokens (typically words, or sometimes other things)

- It is a popular statistical technique of e-mail filtering. They typically use bag of words features to identify email spam, an approach commonly used in text classification.

process

Particular words have particular probabilities of occuring in spam email and legitmate email.

For instance, most email users will frequently encounter the word 'Viagra' in spam mail, but will seldom see it in other email.

The filter doesn't know these probalities in advance and must be trained so it can build them up.

M

Mathematical foundation

(i) Compute the probability that the msg is spam, knowing that a given word appears in the message.

ii, To compute the probability that the message is spam, taking into consideration all of its words

iii, Sometimes a third time, to deal with rare words.

→ Compute the probability that a message containing a given word is spam.

Suppose message contains word "replica".

Most people who are used to receiving email know that the message is like to be spam.

$$Pr(S|W) = \frac{Pr(W|S) \cdot Pr(S)}{Pr(W|S) \cdot Pr(S) + Pr(W|H) \cdot Pr(H)}$$

$Pr(S|W)$ - Probability that a msg is spam. [knowing word replica in it].

$Pr(S)$ - overall probability that any given msg is spam.

$Pr(W|S)$ - probability that the word "replica" appears in spam msg's.

$Pr(H)$ - overall probability that any given msg is not spam ("ham").

$Pr(W|H)$ - probability that the word "replica" appears in ham messages.

↓

E-mail that is generally desired and isn't considered as spam." Desired? you must be saying to yourself "I do not desire this mail, how this is ham and why am I getting it? the answer is you requested it.

Spam - you did not ask for message's from this Source - So, it is spam.

ham [directly] - signing up for a new online Service

UNIT- IV

26/11/22

Data Wrangling :-

It is the transformation of raw data into a format that is easier to use.

Scrapping data from the web, Carrying out statistical analyses, creating dashboards and visualizations all these tasks involve manipulating data in one way or another. But before we can do any of these things, we need to ensure that our data are in a format we can use. This is where the most important form of data manipulation Comes in data wrangling.

Data wrangling involves transforming and mapping data from one format into another.

The aim is to make data more accessible for things like business analytics or machine learning.

The data wrangling process can involve a variety of tasks. These includes things like data collection, exploratory analysis, data cleansing, creating datastructures.

API's and other tools for web scraping :-

↳ Application programming Interface (API) is a way for two or more computer programs to communicate with each other.

Applications : Google Maps API
Youtube API
Twitter API

Web Scraping :- [web harvesting, web data extraction].

It refers to the extraction of data from a website. The information is collected and then exported into a format that is more useful for the user. Be it a spreadsheet or an API.

Web - A network of fine threads. It is an info system enabling documents and other web resources to be accessed over the Internet.

Scrapping - The action or sound of something scraping.

Extracting data from websites.

WebScraping Sfw may directly accesses the www directly, using http protocol or a web browser.

It is a form of copying in which specific data is gathered and copied from the web,

Scraping a web page involves fetching it and extracting from it.

Fetching - downloading of a page (which a browser does when a user views a page). Web crawling is a main component of web scraping, to fetch pages for later processing. Once fetched, Extraction can take place. The content of page may be parsed, searched and reformatted, and its data copied into a spreadsheet or loaded in a database.

Web Scrapers typically take something out of a page, to make use of it for another purpose somewhere else.

An example would be finding and copying names and telephone numbers, companies and their URL's or email addresses to a list (Contact Scraping).

Web scraping is useful technique for finding and utilizing information by collecting data from any online source. It refers to using or creating computer s/w to collect all of this data.

Web Scraping API's - tools that carry out the heavylifting for you and bring you closer to web data.

1. WebScrappingAPI

It is a tool that allows you to scrape any online source without getting blocked. It collects the HTML from any webpage using a simple API.

It provides ready to process data whether you want to use it to extract price and product information, gather and analyze real estate, HR and financial data,

Suitable for : web developers, data scientists.

2. ScraperApi [web developers]

It is a tool for developers building web scrapers. The web services handles proxies, browsers, and CAPTCHAs so that developers can get the raw HTML from any website.

3. Proxy - The authority to represent someone else, especially in voting

• A person authorized to act on behalf of another.

3. Scraping Bee :-

It offers the opportunity to web scrape without getting blocked, using both classic and premium proxies. It focuses on extracting any data you need rendering web pages inside a real browser (chrome).

Thanks to their large proxy pool, developers and companies to handle the scraping technique without taking care of proxies and headless browsers.

4. ZenScrape :-

It is a web scraping API that (refers) returns the HTML of any website and ensures developers collect information fast and efficiently.

5. ScrapingBot :-

It is an excellent tool for developers who cannot dedicate as much time developing their scraper.

It helps extract precise data from any website.

It is developed mainly for collecteing data such as product descriptions, price, costs, images, etc.

6. Scrapingdog :-

It is the web scraper API that handles million of proxies, browsers and CAPTCHA's to provide you with any web pages HTML data. The tool rotates IP addresses with each request from a list of millions of proxies.

7. ScrapingAnt
8. Scraperstack
9. ScraperBox [online data without any impediments] → obstruction
10. Apify. [robotic process Automation].

Feature Generation and feature Selection:-

What exactly is a "feature" and why would you want to generate a new one? A feature is another term for an 'Attribute' (RapidMiner's term of a column)

RapidMiner - It is an integrated Enterprise AI framework that offers AI sol's to positively impact businesses.

It is used as a data science software platform for data extraction, data mining, deep learning, ML and predictive analytics.

It is widely used in a (rapi) no of business and commercial applications as well as various field's such as research, training, education, application development.

Feature generation is used to take one or more attributes from your dataset and create a new "feature" from them.

A typical examples: Calculating the rate of change over time, calculating the percentage of an observed value, or even a simple extraction of a prefix value of a string.

There are many feature Generation operators in RapidMiner which are found under Blending > attributes > Generation

Generate attribute operator can do simple mathematical calculations like add, sub, mult etc as well as really advanced logical, string, and date calculations.

Feature selection:-

After you've generated new features, you might want to tidy up some of your data. Perhaps you calculated a rate of change and want to remove the attributes you used to create the calculation and only keep the new feature Attribute.

In ML and statistics, feature selection also known as variable, attribute (or) variable subset selection, it is the process of selecting a subset of relevant features for use in model construction.

Feature Selection, Techniques are used for several reasons:

- Simplication of models to make them easier to interpret by reaschers/users.
- Shorter training times
- to avoid the curse of dimensionality

↓

Refers to various phenomena that arise when analysing and organizing data in high-dimensional spaces.

It is used for Feature optimization.

Feature optimization uses machine learning to test and measure the performances of your features. It helps to identify the optimal group of features required to build the best model. often to remove the unnecessary ones.

- There are several ways to do feature optimization in RapidMiner. Those are Backward Elimination and Forward selection.

Starts with all attributes and then drops out attributes that aren't use.

This is usually done by embedding a machine learning algorithm [i.e. Decision Tree - Internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. Decision Node and Leaf Node are two nodes in Decision Tree.

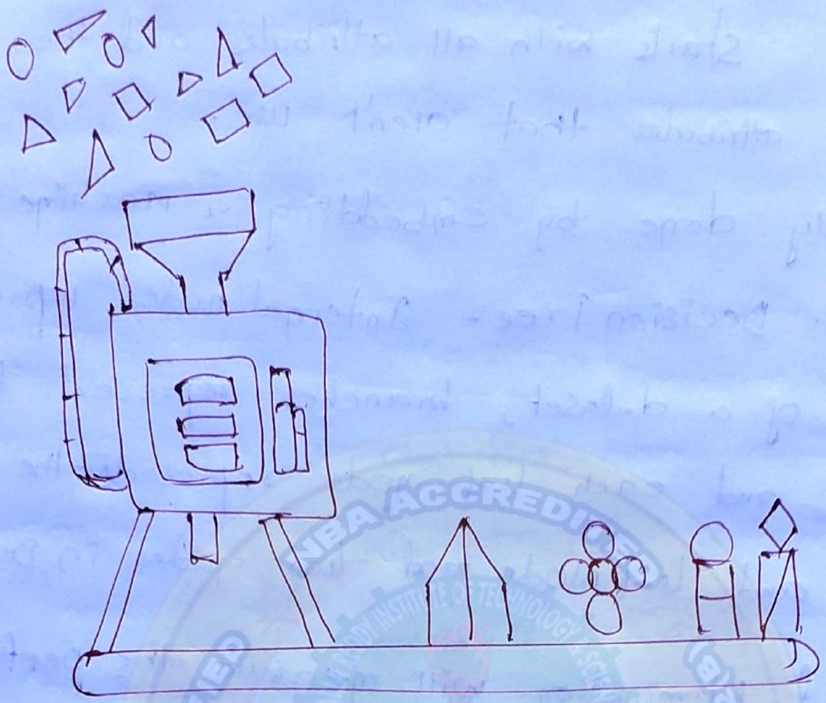
Backward Elimination will measure the performance and keep only the ones that add to the performance of the dataset.

Forward [Feature] Selection is the same concept, except it starts with an empty dataset and add the features.

- Feature selection is to reduce over-fitting by removing extraneous data, it allows the model to focus only on the important features of data.

In the real word, dataset collection is loosely controlled, noisy, unreliable, redundant, and incomplete. In the previous blogs, we introduced how to improve data quality and how to handle imbalanced data.

Feature generation is the process of transforming features into new features that better relate to the target. This can involve mapping a feature into a new feature using a function like log, or creating a new feature from one or multiple features using multiplication or addition.



Feature Generation can improve model performance when there is a feature interaction. Two or more features interact if the combined effect is (greater or less) than the sum of their individual effects.

Feature selection:-

not all features are relevant. Moreover, too many features may adversely affect the model performance. This is because as the number of features increases, it becomes more difficult for the model to learn mappings b/w features and target (It is known as the curse of dimensionality).

eg:- the right transformation depends on the type and structure of the data, data size and the goal.

This can involve transforming single feature into a new feature using standard operators like log, square power, exponential, reciprocal, add, division, multiplication etc.

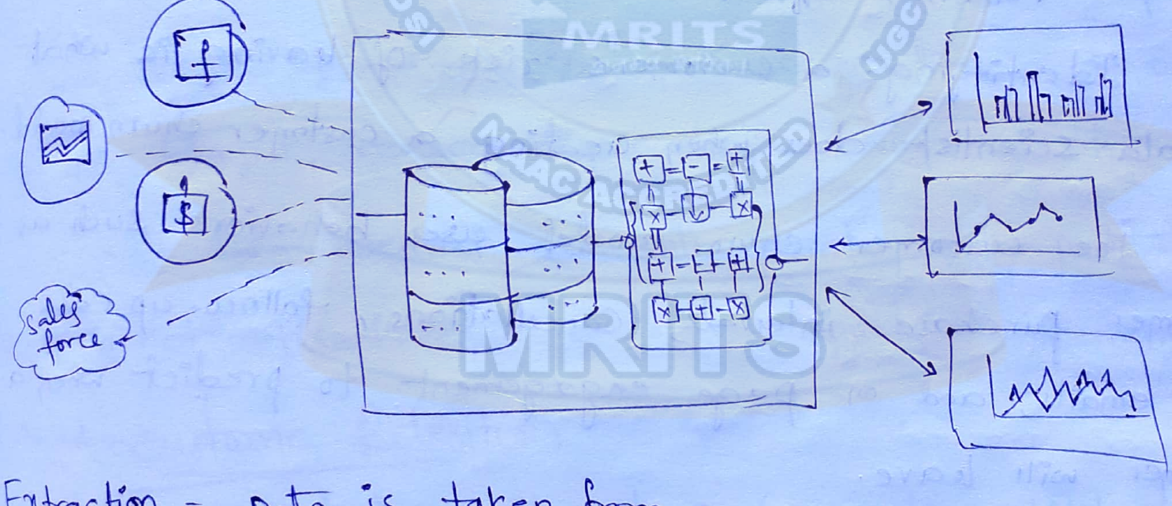
Extracting Meaning from Data:-

Data Extraction is the process of collecting or retrieving disparate types of data from a variety of sources, many of which may be poorly organized or completely understand.

Data Extraction makes its possible to consolidate, process and refine data so that it can be stored in a centralized location in order to be transformed.

[ETL]

Extract Load Transform Analyze



Extraction - Data is taken from one or more source or systems. The extraction locates and identifies relevant data, then prepares it for processing or transformation.

Transformation: once the data has been successfully extracted, it is ready to be refined. During the transformation phase, data is stored, organized and cleansed.

for example, duplicate entries will be deleted, missing values removed or enriched, and audits will be performed to produce data that is reliable, Consistent, and usable.

Loading: The transformed, high Quality data is then delivered to a single, unified target location for storage and analysis.

User (customer) Retention Analysis:- [The continued use, existence, possession of something] someone)

Knowing your revenue loss from customer churn (is a measure of the ~~[number of the]~~ number of individuals or items moving out of a collective group over a specific period) can enable a data science consultant to better identify your problem and ultimately improve your customer retention efforts.

Identifying a customer's risk of leaving is what a data scientist does when creating a customer churn model.

The customer churn model uses behaviors such as customer purchase intervals, cancellations, follow-up calls and emails, and on-page engagement to predict when a customer will leave.

How can we stop customers churning?

the customer churn score can notify a person or system of the possibility of a customer leaving.

then that person or system can then respond with a call or an offer that is tested to prevent a customer from leaving.

Examples :-

Company A connects its churn model to Hubspot with a special column. Their salesperson / people see the score daily. If a customer enters the dangers zone a call is given to that customer within 24 hours. customer churn rate is dropped by 2%. saving the company from losing 10% of its revenue that Quarter.

Company B is an eCommerce store. With their data scientist, they developed a system that sends a personalized offer to customer (churn) who entered the churn danger zone. Their customer retention increased by 25% and their revenue by 10%..

Lack of cross-department cooperation can be one of the biggest reasons why customer churn mode fail.

A data scientist can only point to the problem, they cannot fix internal customer support or Experience issues.

customer Retention

Customer retention refers to a Company's ability to turn customers into repeat buyers and prevent them from switching to a competitor.

It indicates whether your product and the quality of your service will please your existing customers. ~~reward~~

- reward programs (cc companies)
- Zomato pro / Swiggy super.
- wallet cashback (Paytm / gpay)

It indicates the quality of a product or service and the degree of customer loyalty.

Retention is best achieved by overcoming barriers to switching, maximizing the value of products and services, meeting customer expectations, and enriching the customer experience.

eg:-

	order_id	cust_id	order_date	amount
1	1	1	15/1/2020	150
2	2	1	10/2/2020	″
3	3	2	16/1/2020	″
4	4	2	25/2/2020	″
5	5	3	10/1/2020	″
6	6	3	20/2/2020	″
7	7	4	20/1/2020	″
8	8	5	20/2/2020	″

-- jan 0

- feb 1, 2, 3 → 3

we have to check each month that the particular customer ordered in last month also.

Code

```
select * from transaction;
select * from transaction;

select month (this_month.order_date) as month_date,
       count (distinct last_month. cust_id) from
            transaction this_month
  left join transaction last_month
```

on this_month.cust_id = last_month.cust_id and
DATEDIFF (month, last_month.order_date, this_month.order_date) = 1
group by month (this_month.order_date)

o/p

	month_date	No column name
1	1	0
2	2	3

Feature Generation (Brainstorming)

It is the first ever technique of idea generation.

It is an individual or group idea Generation Technique to find a solution for a particular problem by generating multiple solutions; In fact, importance is attached to the quantity of ideas and not quality at the generation stage.

Ideas may be blended to create a single great ideas as implied by the motto "1+1=3".

History -

Alex OsBorn Gives birth brainstorming.

↓

1940, Advertising executive came up with the technique of brainstorming following his frustration at the inability of the employees to come up with innovative ideas for advertising campaigns.

The technique was the result of his attempt to

Scanned by CamScanner

fix rules that would provide people with the freedom of action and mind to trigger and reveal fresh ideas.

The original name was "think up".

Brainstorming is a Conference technique through the practice of which a group endevors to come up with a solⁿ for a particular problem by collecting all the ideas spontaneously contributed by the participating members.

Obsorn's argument was included by the principles named as

(i) put the emphasis on quantity of ideas (Quality)

(ii) ~~Hold~~ back criticism or judgment;

(iii) Be open to strange idea;

(iv) Blend ideas to enhance them (1+1=3)

Steps for effective Brainstorming:-

(i) Decide on a suitable place and facilitator

(ii) Decide on the participants

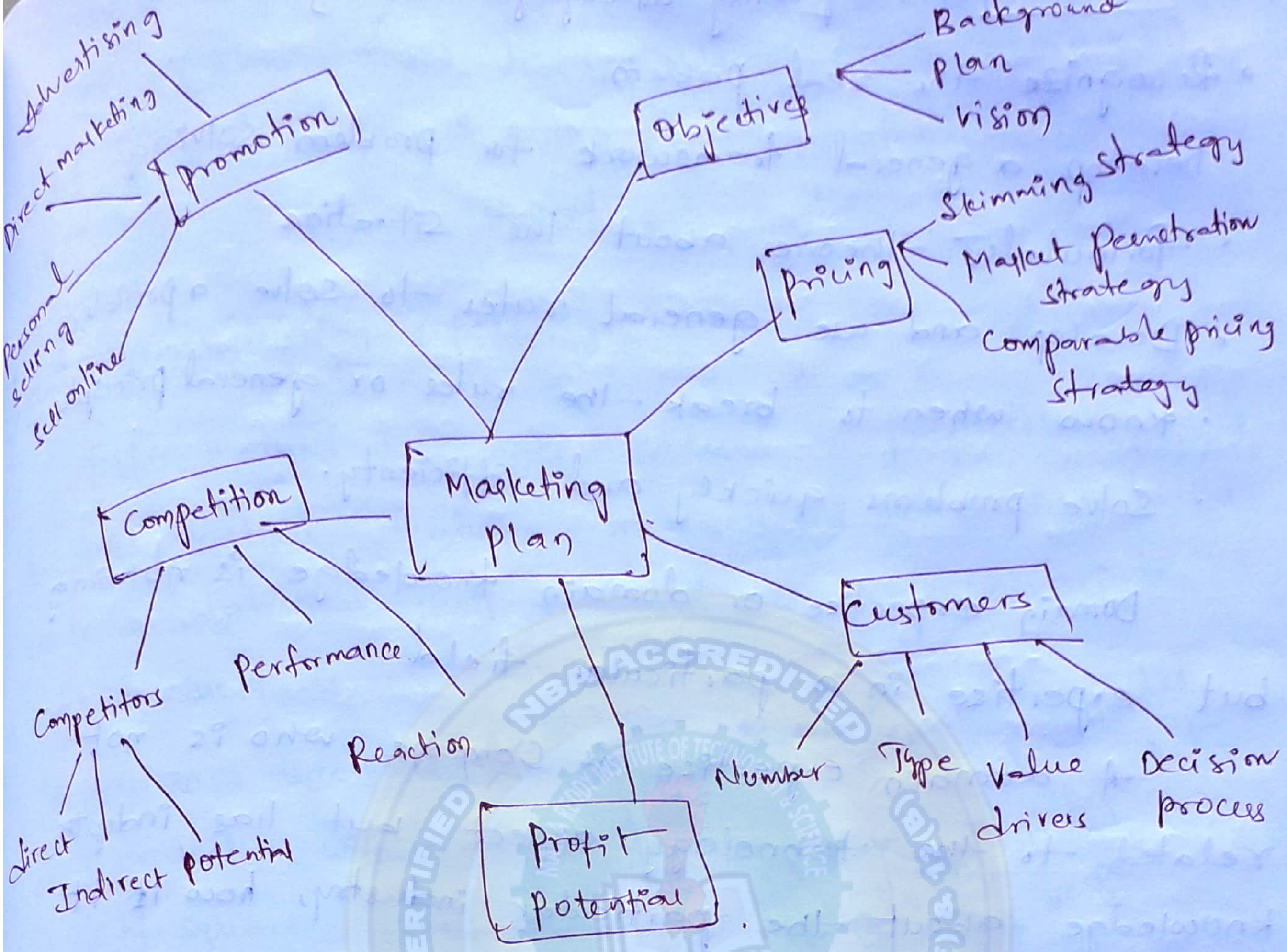
(iii) Specify the problem for which possible sol's are to be found and the goal

(iv) Set a time limit

(v) Diverge prior to converging (Everyone to pen down the ideas)

(vi) Let the brainstorming begin

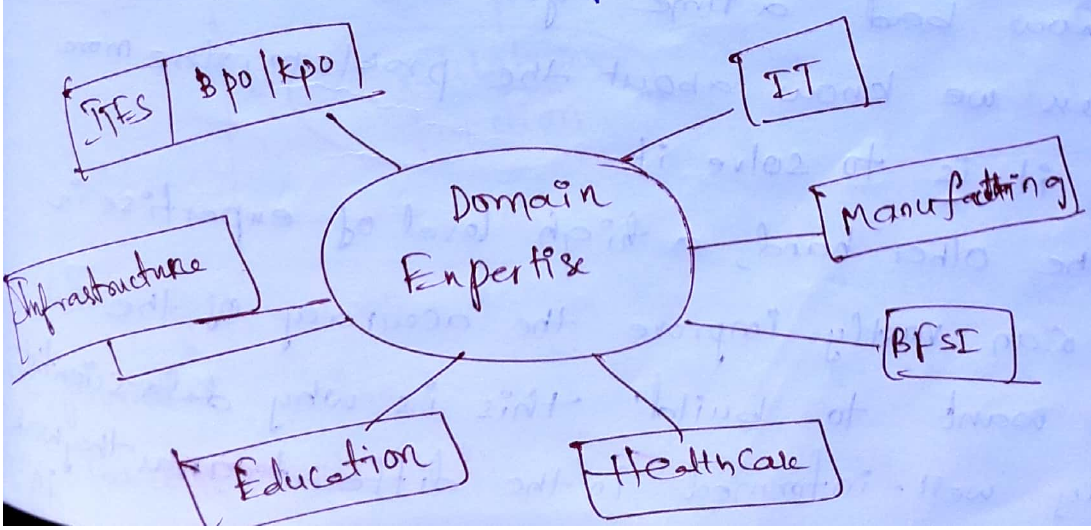
(vii) Choose the best Idea.



Role of Domain Expertise and place for Imagination

Domain expertise is the knowledge and understanding of a particular field.

As a data scientists, we may be working in a wide variety of industries, each of which has its own intricacies that can only be learned gradually over time.



Domain Experts [Hiring an employee in s/w]

- Recognize the real problem
- Develop a general framework for problem solving
- formulate theories about the situation
- Develop and use general rules to solve a problem
- know when to break the rules or general principles
- Solve problems quickly and efficienty.

Domain expertise or domain knowledge is nothing but expertise in a particular field.

A domain expertise is someone, who is not related to the technology aspect but has indepth knowledge about the particular industry, how is it shaping up, the trends, what are the things that will impact the industry.

for example, you cannot unlock the full power of an algorithm without proper knowledge about the field where the data comes from.

Try to build a complex data model in an industry that you don't know anything about and tell us how bad a time you had.

The less we know about the problem, the more difficult it is to solve it.

On the other hand, a high level of expertise in the area can vastly improve the accuracy of the model you want to build. this is why data scientists are usually well-informed in the different areas they work in.

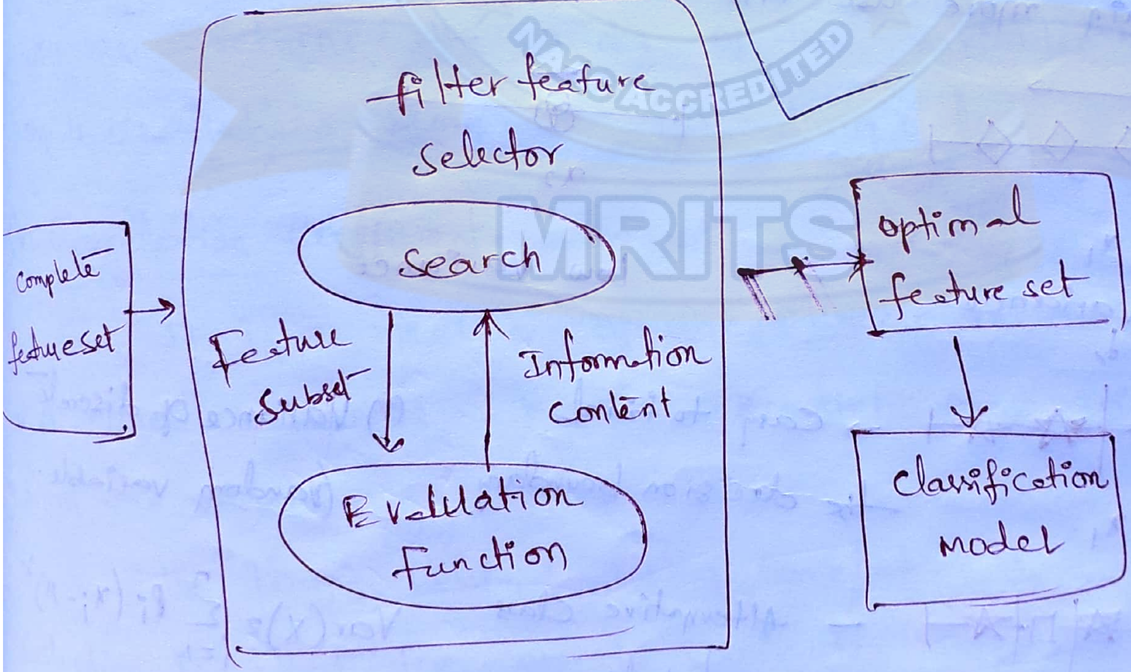
they may not be experts in everything there is to know (who would be capable of such a thing?) but a good data scientist usually focuses on more than one area of expertise.

Feature selection (filters):-

Filter method is generally used as preprocessing step. for this method, features are selected according to the various statistical tests or based on the univariate metrics such as

- variance threshold
- Correlation with target
- Chi-square
- Mutual information
- Information gain

correlation with target
- Correlation with other variables
- Mutual info of independent variable with respect to target.



When you use the filter Based feature selection component, we provide a dataset and identify the column that contains the label or dependent variable. we then specify a single method to use in measuring feature importance. The component outputs a dataset that contains the best feature columns, as ranked by predictive power.

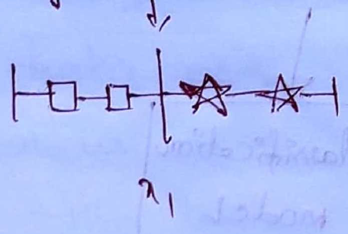
The filters method looks at individual feature for identifying it's relative importance. A feature may not be useful on its own but may be an important influencer when combined with other features.

The filter method does not remove multicollinearity.
↓
Refers to the statistical phenomenon where two or more independent variables are strongly correlated.

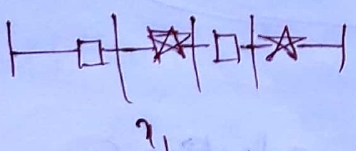
(1) Variance Threshold

- Compute the variance of each feature
- Assume that features with a higher variance may contain more use information.

high variance ·· low variance



→ easy to find the decision boundary



— Alternative class labels

(i) Variance of discrete random variable:

$$Var(X) = \sum_{i=1}^{n} P_i (x_i - \mu)^2$$

(ii) with n datapoints

$$Var(X) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

(ii) Information Gain

It calculates the reduction in entropy from the transformation of a dataset. It can be used for feature selection by evaluating the information gain of each variable in the content of the target variable

(iii) chi-square Test

It is used for categorical features in a dataset. we calculate chi-square test b/w each feature and the target and select the desired no of features with the best chi-square scores.

(iv) fisher's score

It is one of the most widely used supervised feature selection methods. The algorithm which we will use returns the ranks of the variables based on the fisher's score in descending order.

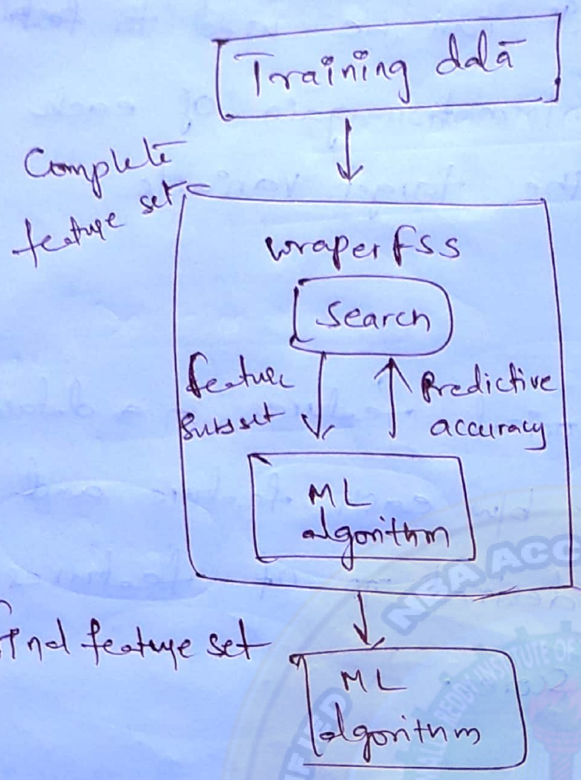
(v) Correlation Coefficient

It is a measure of the linear relationship of 2 or more variables. we can predict one variable from the other.

the logic behind using correlation for feature selection is that the good variables are highly correlated with the target.

Feature Selection (Wrapper Method):-

Wrapper methods a Supervised Methods



- Use a subset of features
- Train the model
- Evaluate performance
- Add/ remove features
- Repeat the process

Methods

- forward feature selection
- Backward feature selection / Backward feature Elimination

} Iterative Methods

- start with having no feature at all / empty set / fs
- In every iteration we keep on adding which best improve our model.
- Addition of a new variable doesn't not improve the performance of the model.

initially $F' = \phi$ [Most significant feature

- If there is no improving of the performance of the model. At that point we messed up.

Backward

- we start with all elements

- feature subset is the entire feature vector.
- During the iteration, we can remove the least significance feature $F' = F$

Decision Tree Algorithm:-

It is a supervised learning technique that can be used for both classification and Regression problems. But mostly it is preferred for solving classification problems.

It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the Decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and leaf Node.

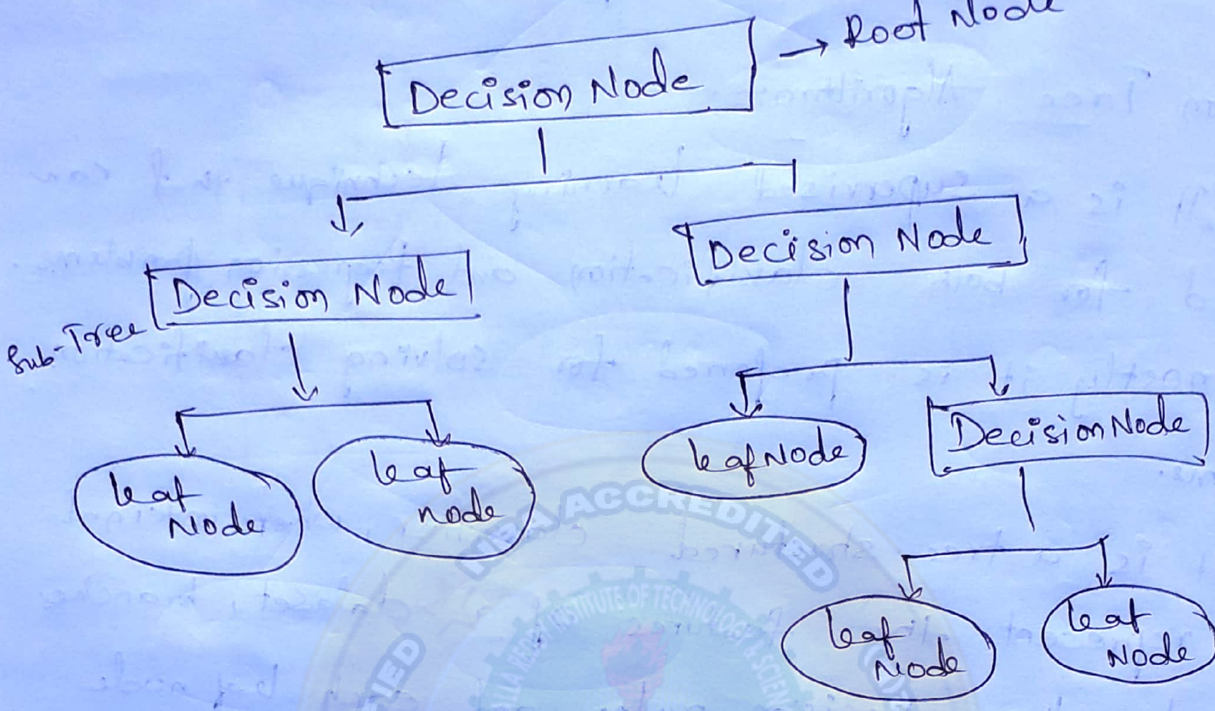
Decision Node ↓ used to make any decision and have multiple branches.

Leaf Node ↓ These are the outputs of those decisions and do not contain any further branches.

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

In order to Build a tree, we use the CART algorithm, which stands for 'classification and Regression Tree algorithm'.

A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtree.



Why we use Decision Trees?

- These are usually mimic human thinking ability while making a decision, so it is easy to understand.
- The logic behind the decision Tree can be easily understood because it shows a tree-like structure.

Terminologies

- Root Node - where the DT starts
- Leaf Node - find o/p
- Splitting - process of dividing decision node
- Branch/SubTree - A tree formed by splitting the tree
- Pruning - process of removing unwanted branches
- Parent/Child node.

How Does Decision Tree Algorithm work ?

For predicting the class of the given data set the alg starts from the root node of the tree. It compares the values of root (variable) attribute with the record (real dataset) attribute and based on the comparision, follows the branch and jumps to the next node.

step1:- Begin the tree with root node, Say 's' which Contains the complete dataset.

step 2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).

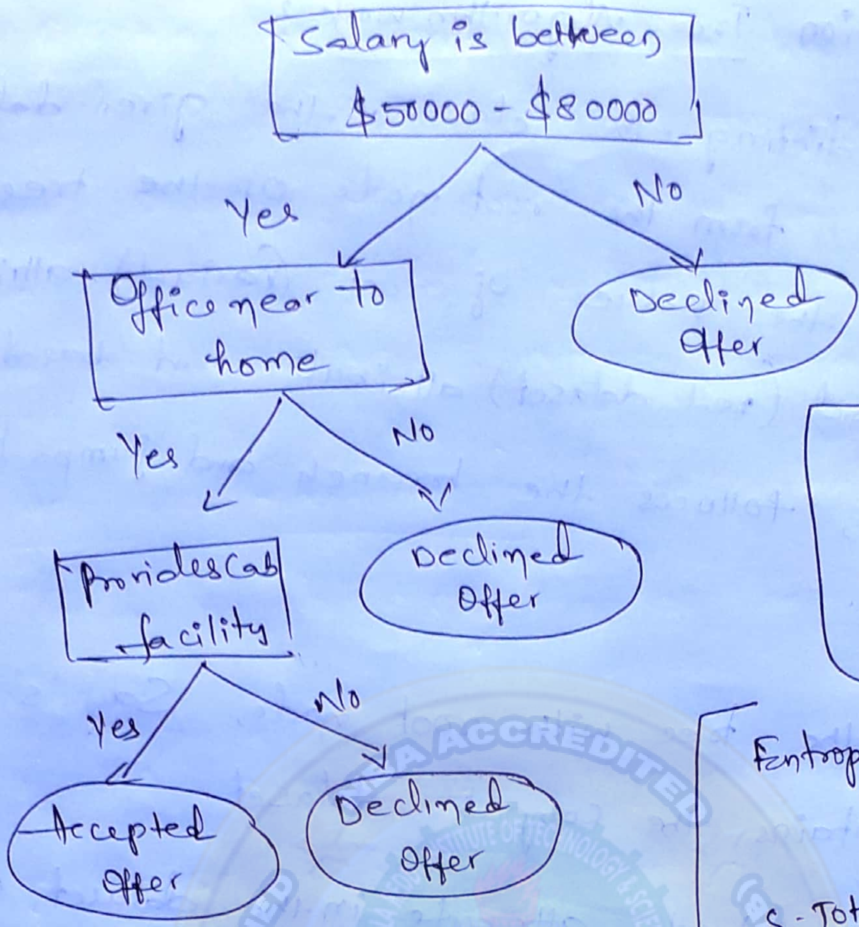
step 3: Divide the 'S' into Subsets that contains possible values for the best attributes.

step 4: Generate the Decision Tree node, Contains best attribute.

Step 5. Recursively make new decision trees using the substes of the dataset created in step-3.

Ex:- Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or Not.

so, to solve this problem, decision tree Starts with the root node.



Entropy - Randomness / disorder of a system

$$Entropy(S) = -P(yes)\log_2 P(yes) - P(no)\log_2 P(No)$$

S - Total no. of samples
P(yes) - probability of Yes
P(No) - probability of No

ASM

(i) Information Gain - Measurement of changes in entropy after the segmentation of a dataset based on a attribute

- It calculates how much info a feature provides us about a class.

(ii) Gini Index -

It is a measure of impurity or purity while creating a DT in the CART algorithm.

$$\text{Gini Index} = 1 - \sum_j P_j^2$$

- It is easy to understand
- Complexity is high.

Random Forest Algorithm:-

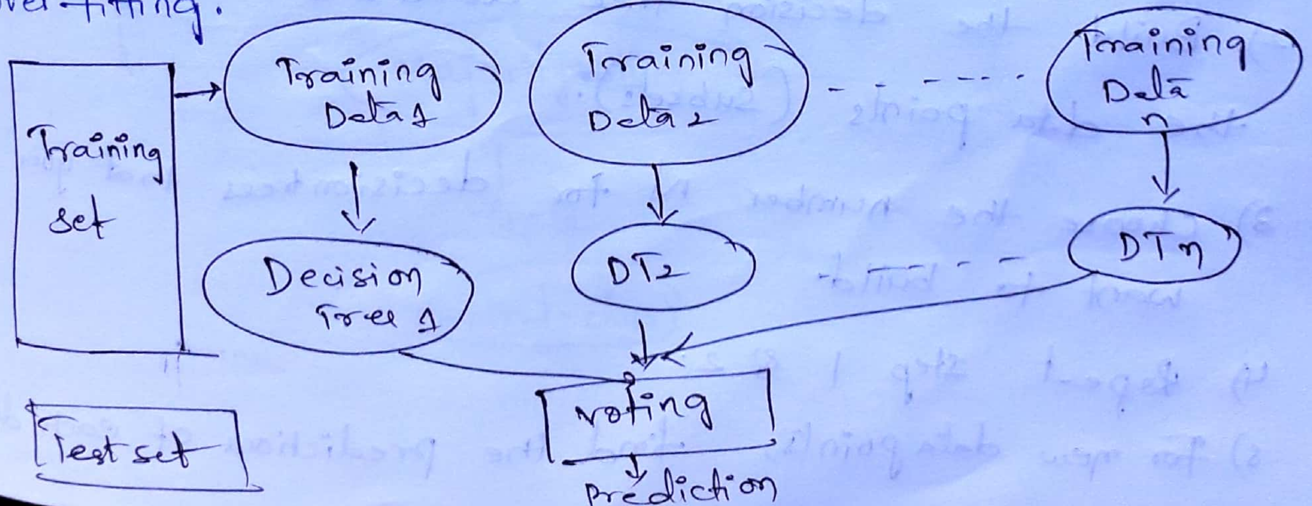
It is one of the supervised learning technique. It can be used for both classification and Regression problems in Machine learning.

It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

Random forest is a classifier that contains a no of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greatest no. of trees in the forest leads to higher accuracy and prevents the problems of over-fitting.



It is possible that some decision trees may predict the correct output, while others may not.

There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.

Why we use Random forest?

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

How does Random forest algorithm work?

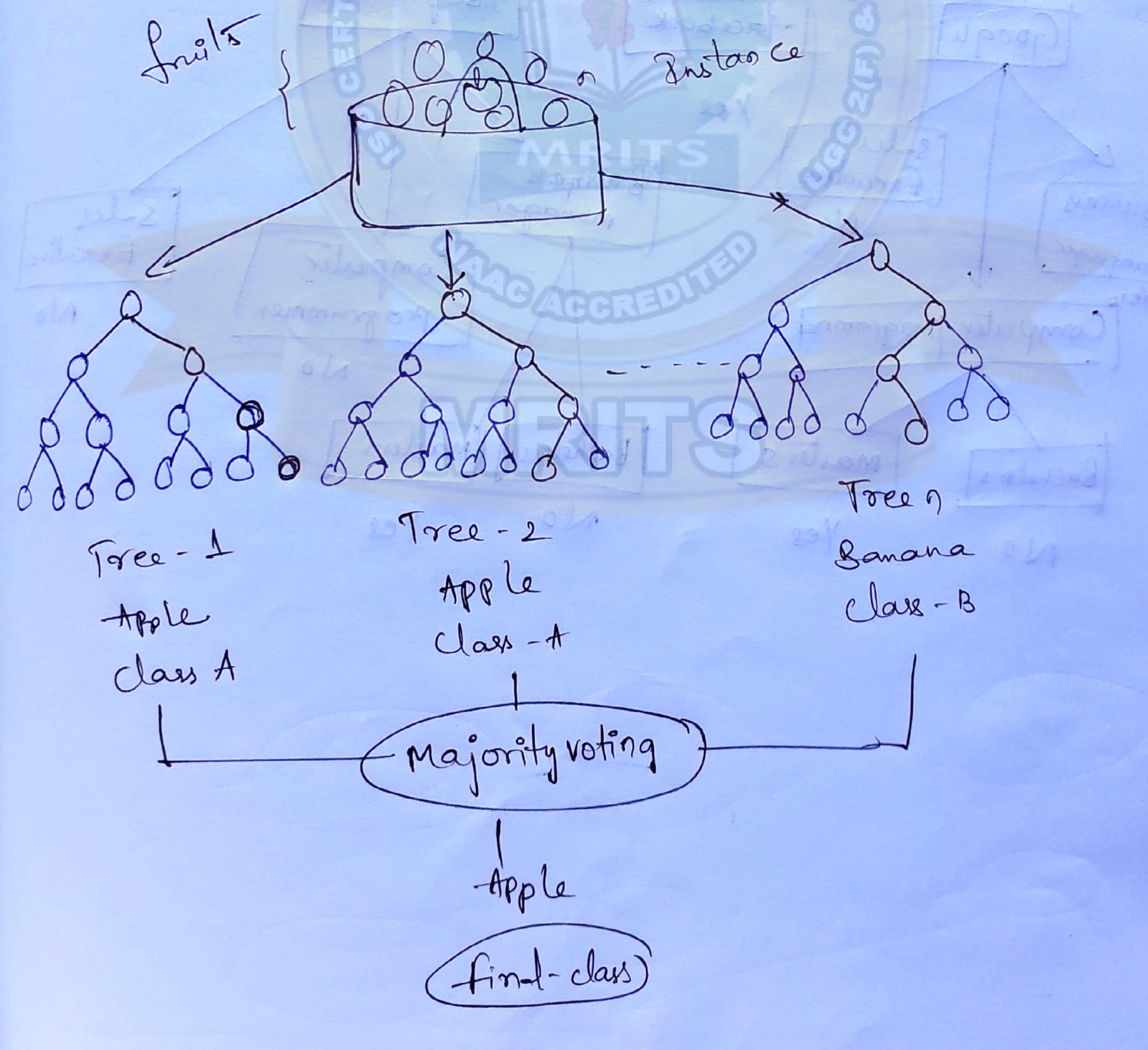
Random forest works in two-phase first is to create the random forest by combining N decision tree and second is to make predictions for each tree created in the first phase.

1) Select random k data points from the training set
2) Build the decision tree associated with the selected the data points (subsets).
3) Choose the number N for decision trees that you want to build.
4) Repeat step 1 & 2.
5) For new data points, find the predictions of each decision

tree, and assign the new data points to the category that wins the majority votes.

eg:- suppose there is a dataset that contains multiple fruit images. so, this dataset is given to the Random forest classifier. The dataset is divided into subsets and given to each decision tree.

During the training phase, each decision tree produces a prediction result, and when a new data point occurs, then based on the majority of results, the Random forest classifier predicts the final decision.

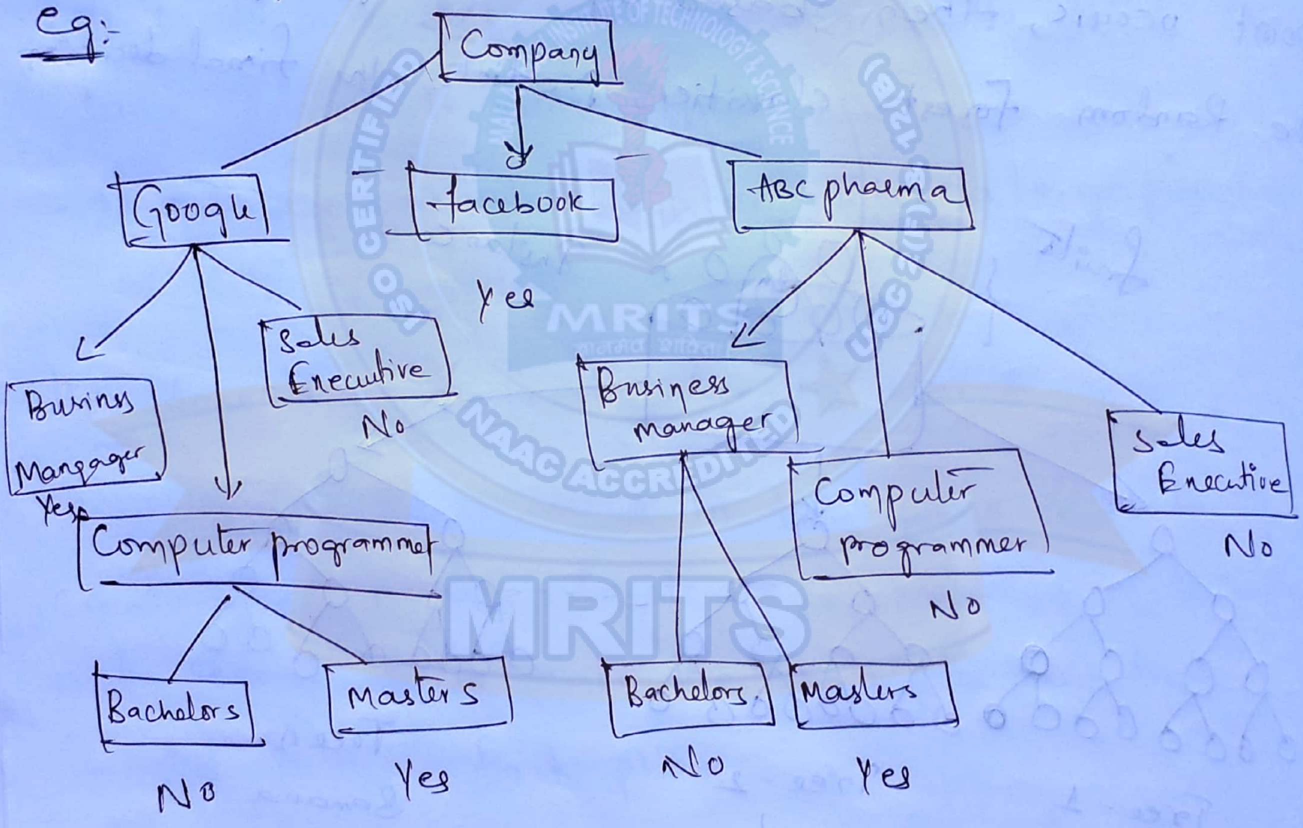


Applications:-

- Banking - Mostly use for identification of loan risk
- Medicine - Disease trends and risks of disease can be identified.
- Land use - Identify the areas of Similar land use by this algorithm.
- Marketing - Marketing trends can be identified using this algorithm.

Salary >100k$

eg:-



UNIT-V

DATA VISUALIZATION

Data visualization:

Data visualization is a graphical representation of quantitative information and data by using visual elements like graphs, charts, and maps.

Data visualization converts large and small data sets into visuals, which is easy to understand and process for humans.

Data visualization tools provide accessible ways to understand outliers, patterns, and trends in the data.

In the world of Big Data, the data visualization tools and technologies are required to analyze vast amounts of information.

Data visualizations are common in your everyday life, but they always appear in the form of graphs and charts. The combination of multiple visualizations and bits of information are still referred to as Infographics.

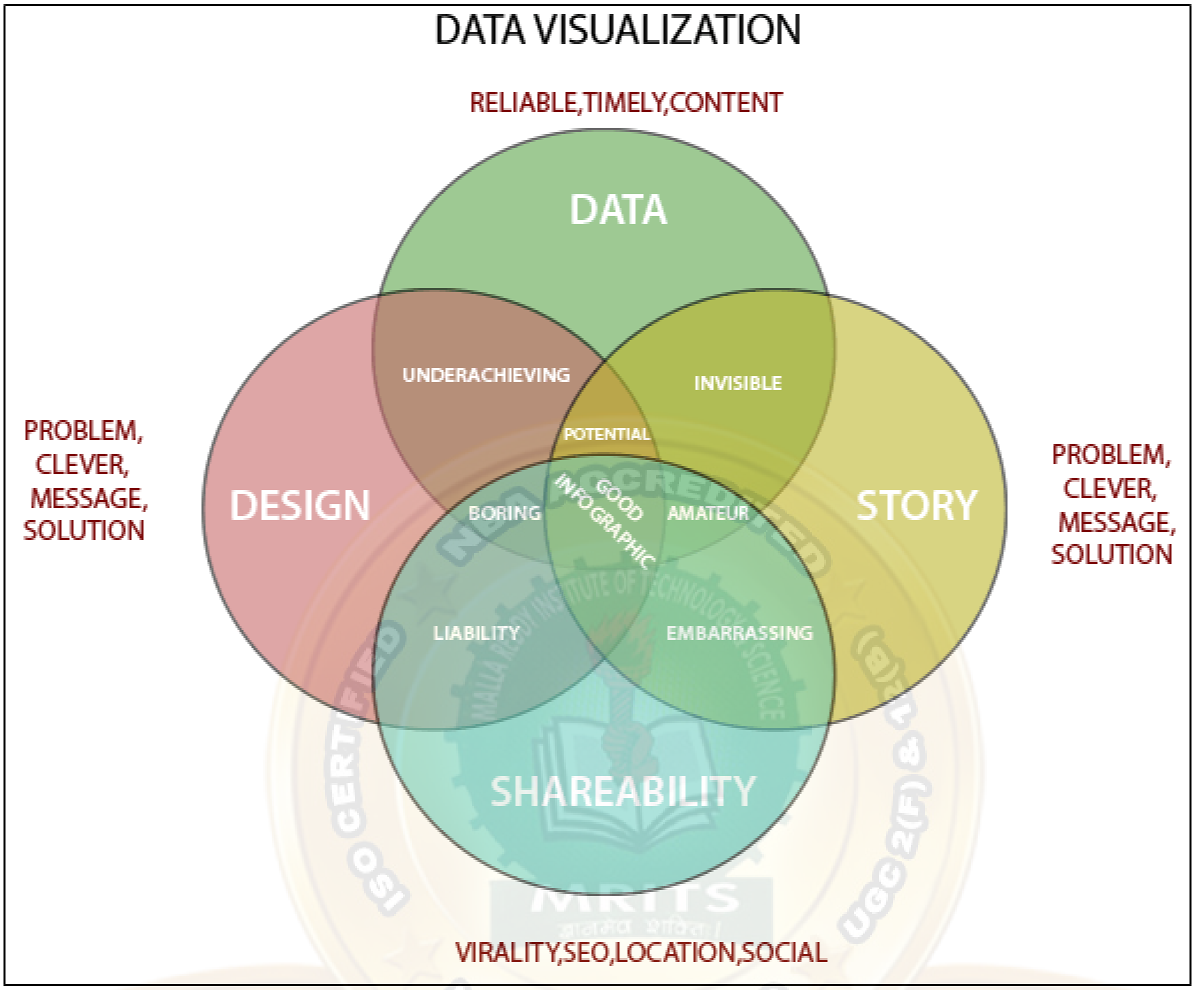
Data visualizations are used to discover unknown facts and trends. You can see visualizations in the form of line charts to display change over time. Bar and column charts are useful for observing relationships and making comparisons. A pie chart is a great way to show parts-of-a-whole. And maps are the best way to share geographical data visually.

Today's data visualization tools go beyond the charts and graphs used in the Microsoft Excel spreadsheet, which displays the data in more sophisticated ways such as dials and gauges, geographic maps, heat maps, pie chart, and fever chart.

What makes Data Visualization Effective?

Effective data visualization are created by communication, data science, and design collide. Data visualizations did right key insights into complicated data sets into meaningful and natural.

American statistician and Yale professor **Edward Tufte** believe useful data visualizations consist of ?complex ideas communicated with clarity, precision, and efficiency.



To craft an effective data visualization, you need to start with clean data that is well-sourced and complete. After the data is ready to visualize, you need to pick the right chart.

After you have decided the chart type, you need to design and customize your visualization to your liking. Simplicity is essential - you don't want to add any elements that distract from the data.

Why Use Data Visualization?

1. To make easier in understand and remember.
2. To discover unknown facts, outliers, and trends.
3. To visualize relationships and patterns quickly.
4. To ask a better question and make better decisions.
5. To competitive analyze.
6. To improve insights.

Here are some noteworthy numbers, based on research, that confirm the importance of visualization:

- People get 90% of information about their environment from the eyes.
- 50% of brain neurons take part in visual data processing.
- Pictures increase the wish to read a text up to 80%.
- People remember 10% of what they hear, 20% of what they read, and 80% of what they see.
- If a package insert doesn't contain any illustrations, people will remember 70% of the information. With pictures added, they'll remember up to 95%.

Relevant data visualization brings lots of advantages for your business:

- **Fast decision-making.** Summing up data is easy and fast with graphics, which let you quickly see that a column or touchpoint is higher than others without looking through several pages of statistics in Google Sheets or Excel.
- **More people involved.** Most people are better at perceiving and remembering information presented visually.
- **Higher degree of involvement.** Beautiful and bright graphics with clear messages attract readers' attention.
- **Better understanding of data**. Perfect reports are transparent not only for technical specialists, analysts, and data scientists but also for CMOs and CEOs, and help each and every worker make decisions in their area of responsibility.

Principles of successful data visualization

The first thing to do before creating any graphic is to check all data for accuracy and consistency. For example, if the scaling factor is 800%, whereas the average is 120–130%, you should check where this number comes from. Maybe it's some kind of outlier that you need to delete from the graph so it doesn't skew the overall picture: 800% downplays the difference between 120% and 130%. This kind of outlying data in a report can lead to an incorrect decision. In real life, we're accustomed to the fact that the right message should be delivered to the right person at the right time. There are three similar principles for data visualization:

1. Choose the right graphic depending on your goal.
2. Confirm that the message of your graphic suits the audience.
3. Use an appropriate design for the graphic.

If your message is timely but the graphic isn't dynamic or there's an incorrect insight or a difficult design, then you won't get the result you hoped for.

Types of graphs and how to choose

If you choose the wrong graph, your readers will be confused or interpret the data incorrectly. That's why before creating a graph, it's important to decide what data you want to visualize and for what purpose:

- To compare different data points
- To show data distribution: for instance, which data points are frequent and which are not
- To show the structure of something with the help of data
- To follow the connections between data points

Let's have a look at the most popular types of charts and the goals they can help you achieve.

1. Line chart

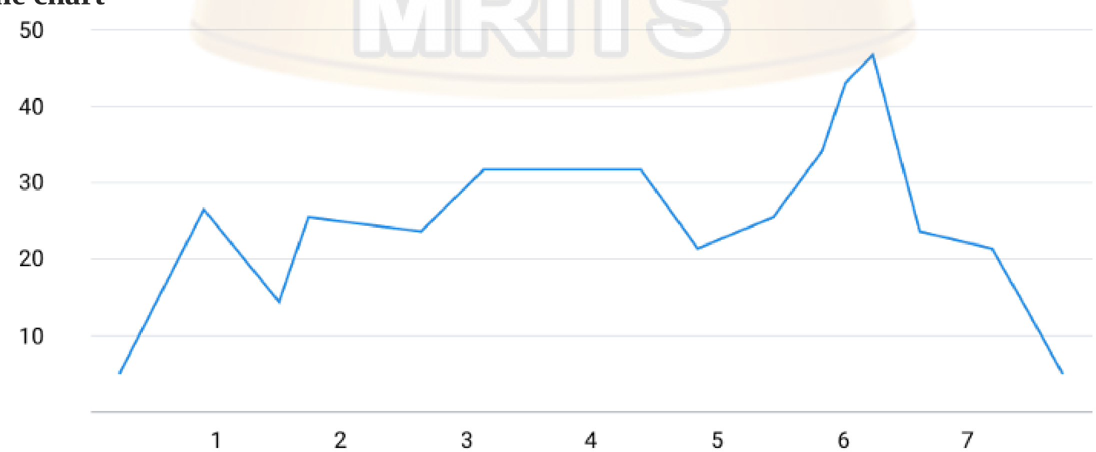


Image courtesy of the author

A line chart shows how one or more variables change across data points. This type of chart is useful for comparing changes within data sets over time — for instance, traffic statistics for three landing pages by month over a one-year period.

2. Bar chart

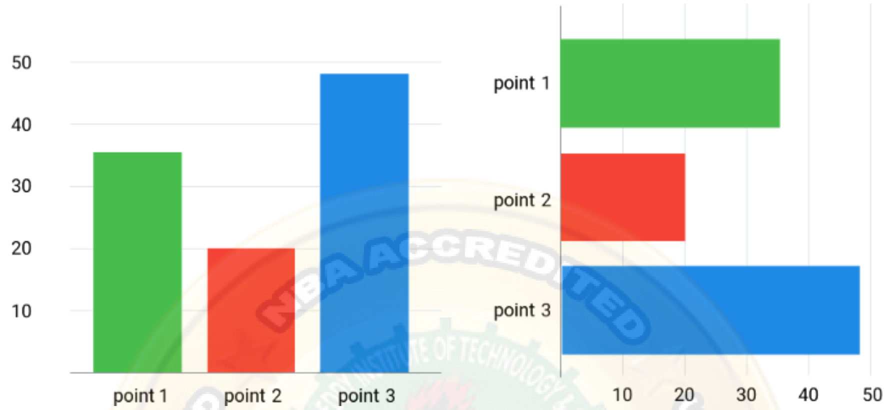


Image courtesy of the author

The bar chart is another diagram that's perfectly suited for comparing data sets. Horizontal bar charts are often used when you need to compare lots of data sets or to visually emphasize the distinct advantage of one of the data sets. Vertical bar charts illustrate how data points change over time — for example, how the annual company profit has changed over the past few years.

3. Histogram

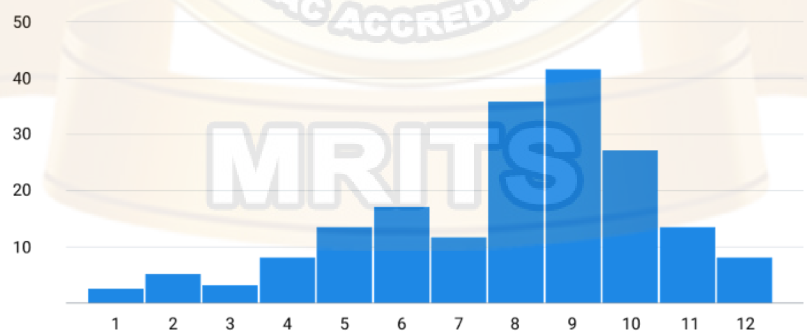


Image courtesy of the author

A histogram is often mistaken for a bar chart due to their visual similarities, but the goals of these charts are different. A histogram shows the distribution of a data set across a continuous interval or a definite time period. On the vertical axis of this chart, you can see frequency, whereas on the horizontal you can see time intervals.

Unlike a histogram, a bar chart doesn't show any continuous interval; each column displays a category of its own. It's easier to demonstrate the number of purchases in different years with the help of a bar chart. If you want to know the order values ($10–100, $101–200, $201–300, etc.) of purchases, it's better to choose a histogram.

4. Pie chart

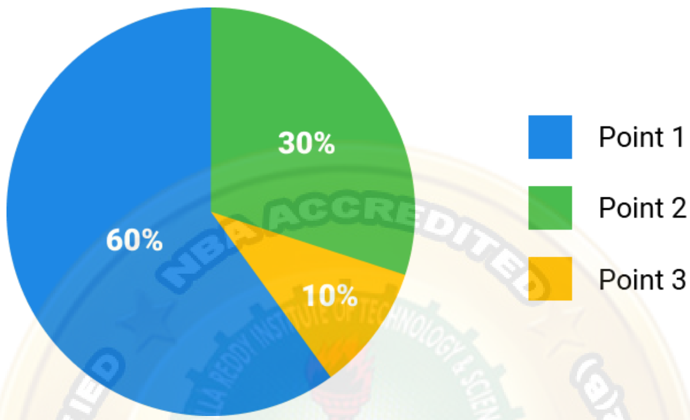


Image courtesy of the author

The pie chart displays shares of each value in a data set. It's used to show the components of any data set. For instance, what percentage of general sales is attributed to each product category?

5. Scatter plot

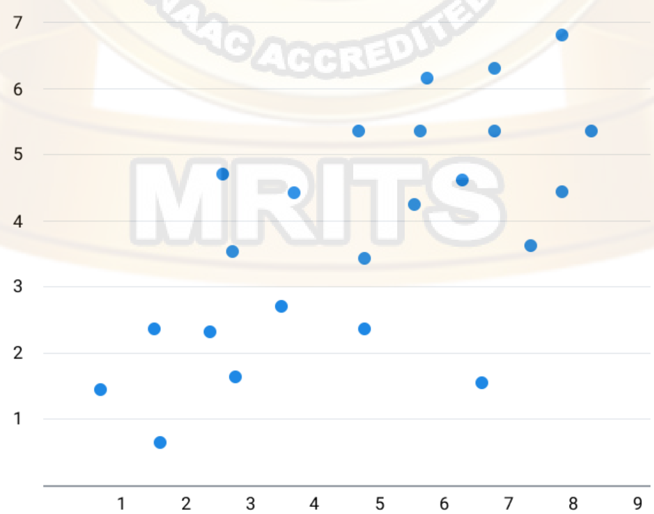


Image courtesy of the author

The scatter plot shows the connection between data points. For example, with the help of a scatter plot, you can find out how the conversion rate changes depending on the size of the product discount.

6. Bubble chart: This is an interesting chart that allows you to compare two parameters by means of a third. Let's take the conversion rate and discount size from the previous example, add to them revenue (indicated by circle size), and we'll get something like the following chart.

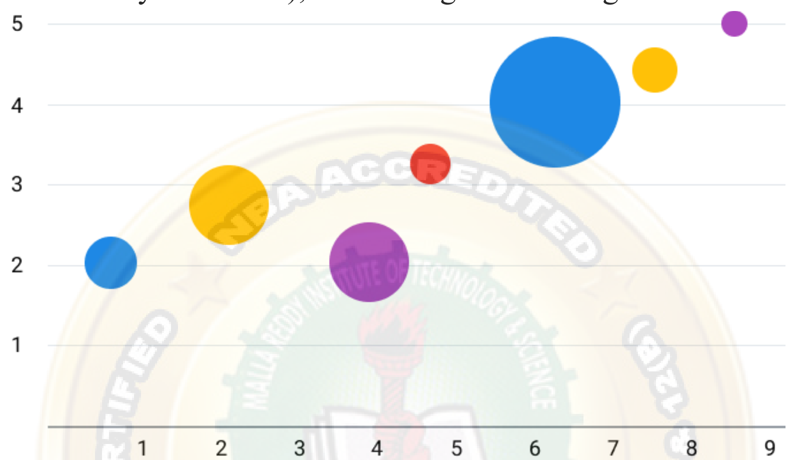


Image courtesy of the author

Looking at this chart, it's easy to notice that products with a 30% discount have the highest conversion rate, while products with no discount or a 5% discount bring in the most revenue.

7. Geo chart

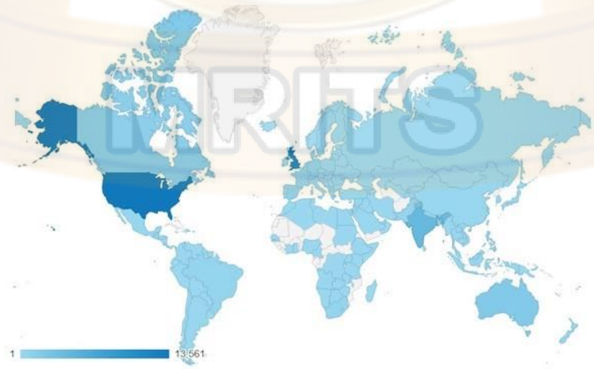


Image courtesy of the author

The geo chart is a simple one. It's used when you need to demonstrate a certain distribution across regions, countries, and continents.

We've mentioned some of the most popular charts but not all of them. You can find other types of graphs in the Data Visualization Catalogue. Also, we recommend this handy infographic that helps you choose the right type of chart for your goal(s).

Comparing reporting software

Nowadays, there are lots of data visualization tools on the market. Some of them are paid, others are free. Some of them work fully on the web, others can be installed on a desktop but work online, and others are offline only. We've made a list of 10 popular tools for data visualization:

1. Excel/Google Spreadsheets
2. Data Studio
3. Tableau
4. Power BI
5. QlikView
6. R Studio
7. Visual.ly
8. Tangle
9. iCharts
10. Smart Data

The first five tools and services are produced by companies specializing in data visualization. Numbers six through ten are quite interesting tools, mostly free and online. They offer non-standard types of visualization and may offer new ways of approaching your data.

What to look for when choosing a reporting tool:

- **Start from the tasks you want to accomplish.** For example, a major trend on the market nowadays is dynamic reports. If a tool cannot work with dynamic reports, that's a strike against it.

- **Consider the amount of money you're ready to pay.** If your team is big enough and every employee has to work with the visualization tool, then the cost per user may be a stop sign.

- **Decide who will use the tool and how.** Is there a possibility for group editing? How simple is it to start working with the tool? Is the interface friendly? Is there a possibility to create a report without any knowledge of programming? For example, R Studio is a great service, especially for searching for trends and building attribution and correlation models. But if you don't know any programming languages, can't connect any specific libraries, and aren't a technical specialist, it will be difficult for you to start working with R Studio.

We've chosen five services and prepared a table comparing their advantages, disadvantages, and main characteristics. Before we start, let's explain how *dynamic data visualization* and *dynamic reports* differ.

Dynamic reports refer to the possibility to import data from different sources in real time. Google Data Studio doesn't have dynamic reports. Let's say we've connected a Data Studio request from Google BigQuery and then changed something in this request. To record these changes in the report, we at least need to refresh the Data Studio page. However, if in Google BigQuery we add or delete some field (not just change the logic of the calculation but change the table structure), then Data Studio will close the report with an error. You'll have to redo it.

Dynamic data visualization refers to the possibility to look at summary statistics over different dates during one session. For example, in Google Analytics you can change the time interval and get statistics for the dates you need.

Key characteristics of the top five visualization tools

	Google Data Studio	R Studio	Microsoft Power BI	OWOX BI Smart Data	Tableau
Price	Free	$50+ for the licensed version	Freemium	$60+ per month	$70+ per user per month
Number of suites	1	5	2 Free for 1 GB Paid (Pro) for 10 GB (approximately $10 per month)	3	4 One for individual users and 3 for corporate users
Several data sources in one widget	Yes	Yes	Yes	Yes	Yes
Dynamic reports	No	Yes	Yes	Yes	Yes
Number of connectors	More than 50	ODBS data connector, SQLServer, Amazon Redshift, Tableau	More than 100	CRM, Google BigQuery, cost pipelines	More than 30
Dynamic data representation	Yes	No	No	Yes	Yes
Ability to clarify and change data before launching	No	Yes	Yes	No	Yes
Real-time group report editing	Yes	No	No	No	Yes
Interface	Easy, user-friendly	Special skills required	Need time to adjust	User-friendly, automatic	Need time to adjust
Peculiarities	Data updates at launch Convenient to share dashboards	Scope of processed data is limited by RAM Different software integrations	Support for web, mobile, and desktop	Works via natural language interface Possible to import data into Data Studio and Google Spreadsheets	Widely used; high level of data safety; expensive; support for several platforms

We want to discuss in detail three tools that are actively used alongside OWOX BI: Google Data Studio, Google Sheets, and OWOX BI Smart Data.

Inspiring Examples of Industrial Design

Looking for inspiring examples of industrial design? The following products solve design dilemmas, inspire creativity, and bring beauty and fun to everyday life.

Some of the designs below are available for purchase, and others are prototypes for potential products.

1. Morgan Felt Folding Stool by Brett Mellor

The Felt Folding Stool brings origami and flat pack together in a piece of furniture. The felt is saturated with resin in a specific pattern to create rigid panels with flexible non-resined seams like a folded sheet of paper. This allows the felt to fold up into a stool or collapse flat for easy storage and transportation.

2)[Armstrong Light Trap]

The LEDs in the Armstrong light trap turn on when the lamp is uncorked, and off when the corks are put in place. The amount of light emanating from the lamp can be controlled by removing multiple corks.

3. “In the Fog” by Dmitry Kozinenko

The “In the Fog” metal furniture project combines an industrial aesthetic with more ethereal, organic textures. The various pieces look like they’ve materialized out of thin air, or that they’re gradually disappearing into the fog.

4. Nessie Ladle by Jenny Pokryvailo

The Nessie Ladle is a great example of form meets function meets FUN. Nessie’s legs allow the ladle to stand upright on its own (great for preventing soup spills), and the handle gives the appearance of a mythical culinary interloper. Beware!

5. Vool. The Wooden Laptop Stand.

From the project description on Industrial Design Served:
Vool is a wooden laptop stand brought to you by a member of Dopludo collective that will turn long working hours in front of a laptop into a pleasure. Using your laptop with Vool stand provides healthy ergonomics and proper posture. Add external keyboard and mouse/tablet and you got yourself fully functional desktop computer, when done with it, there's a space inside of a stand you can put it in (also perfect for A4 paper/documents). It also can be used as an independent laptop stand. Being placed on a lap it provides posture and protects legs from laptop's heat.

6. Rotary Mechanical by Richard Clarkson

The Rotary Mechanical smartphone might not be practical, but the concept is pretty cool. The steampunk and minimalism-inspired design is a “harmonious combination of mechanical parts and digital technologies.” It has interchangeable brass dials (a rotary dial and a button dial) and an electroplated copper body that’s designed to look even **better** with wear.
All pieces can be replaced modularly as new technology becomes available, thus helping to reduce “digital rot.”

DATA SCIENCE AND ETHICAL ISSUES

Despite the numerous possibilities and advantages of data science to solve complex problems and gain new insights, the appropriate way of using and analyzing data, especially in today’s technologically dependent society, continues to face ethical questions and challenges. Although ethics in relation to computer science has been a topic of discussion since the 1950s, the topic has only recently joined the data science debate. Nonetheless, an overall consent or a common conceptual framework for ethics in data science is still nonexistent. In particular, privacy rights, data validity, and algorithm fairness in the areas of Big Data, Artificial Intelligence, and Machine Learning are the most important ethical challenges in need of a more thorough investigation. Thus, this chapter contributes to the overall discussion by providing an overview of current ethical challenges that are not only crucial for data science in general but also for the tourism industry in the future.

Ethics in Data Science and Proper Privacy and Usage of Data

Data may be utilized to make decisions and have a large influence on businesses. However, this valuable resource is not without its drawbacks. How can businesses acquire, keep, and use data in an ethical manner? What are the rights that must be protected? Some ethical practices must be followed by data-handling business personnel. Data is someone’s personal information and there must be a proper way to use the data and maintain privacy.

What is Ethics?

The term “ethics” comes from the Greek word Ethos, which means “habit” or “custom.” Ethics instructs us on what is good and wrong. Philosophers have pondered this crucial topic for a long time and have a lot to say about it. Most people associate ethics with morality: a natural sense of what is “good.” We as humans live in a society, and society has rules and regulations. We must be able to decide what is right and what is wrong. Ethics deals with feelings, laws, and social norms which determine right from wrong. Our ways of life must be reasonable and live up to the standards of society.

Why Ethics in Data Science is important?

Today, data science has a significant impact on how businesses are conducted in disciplines as diverse as medical sciences, smart cities, and transportation. Whether it’s the protection of personally identifiable data, implicit bias in automated decision-making, the illusion of free choice in psychographics, the social impacts of automation, or the apparent divorce of truth and trust in virtual communication, the dangers of data science without ethical considerations are as clear as ever. The need for a focus on data science ethics extends beyond a balance sheet of these potential problems because data science practices challenge our understanding of what it means to be human.

Algorithms, when implemented correctly, offer enormous potential for good in the world. When we employ them to perform jobs that previously required a person, the benefits may be enormous: cost savings, scalability, speed, accuracy, and consistency, to name a few. And because the system is more precise and reliable than a human, the outcomes are more balanced and less prone to social prejudice.

When are public data useful to a data scientist?

- Public data are by default anonymized (census data)
- By its nature there is no privacy concern (imagenet)
- Public data come with an identifier that allows user to join them with private data (census)
- Public data can semantically join without an id (imagenet)

ImageNet Challenge

3

Privacy

- Dramatic increase in scale of information collected and stored
  - in interest of law enforcement, national security, economic incentives
- Users know more about access and use of personal information and their private details
- Privacy advocates have raised concerns about extent of privacy violations. Different Legal and technical approaches have been taken to reinforce privacy rights

Next generation data scientist

1. Data Scientist A data scientist collects, analyzes, and interprets large volumes of data, in many cases, to improve a company's operations

2. 3. Ideally the generation of data scientists-in- training are seeking to do more than become technically proficient and land a comfy salary in a nice city—although those things would be nice. We'd like to encourage the next-gen data scientists to become problem solvers and question askers, to think deeply about appropriate design and process, and to use data responsibly and make the world better, not worse. Let's explore those concepts in more detail in the next sections. The best minds of my generation are thinking about how to make people click ads… That sucks.
3. 4. BEING PROBLEM SOLVERS First, let's discuss the technical skills. Next gen data scientists should strive to have a variety of hard skills including coding, statistics, machine learning, visualization, communication, and math. Also, a solid foundation in writing code, and coding practices such as paired programming, code reviews, debugging, and version control are incredibly valuable. 1
4. 5. It's never too late to emphasize exploratory data analysis and conduct feature selection as Will Cukierski emphasized. Brian Dalessandro emphasized the infinite models a data scientist has to choose from—constructed by making choices about which classifier, features, loss function, optimization method, and evaluation metric to use. Huffaker discussed the construction of features or metrics: transforming the variables with logs, constructing binary variables (e.g., the user did this action five times), and aggregating and counting. As a result of perceived triviality, all this stuff is often overlooked, when it's a critical part of data science. It's what Dalessandro called the "Art of Data Science."
5. 6. Another caution: many people go straight from a dataset to applying a fancy algorithm. But there's a huge space of important stuff in between. It's easy to run a piece of code that predicts or classifies, and to declare victory when the algorithm converges. That's not the hard part. The hard part is doing it well and making sure the results are correct and interpretable.
6. 7. WHAT WOULD A NEXT-GEN DATA SCIENTIST DO? Next-gen data scientists don't try to impress with complicated algorithms and models that don't work. They spend a lot more time trying to get data into shape than anyone cares to admit maybe up to 90% of their time. Finally, they don't find religion in tools, methods, or academic departments. They are versatile and interdisciplinary.
7. 8. CULTIVATING SOFT SKILLS Tons of people can implement k- nearest neighbors, and many do it badly. In fact, almost everyone starts out doing it badly. What matters isn't where you start out, it's where you go from there. It's important that one cultivates good habits and that one remains open to continuous learning. Some habits of mind that we believe might help solve problems are persistence, thinking about thinking, thinking flexibly, striving for accuracy, and listening with empathy.
8. WHAT WOULD A NEXT-GEN DATA SCIENTIST DO? Next-gen data scientists remain skeptical about models themselves, how they can fail, and the way they're used or can be misused. Next gen data scientists understand the implications and consequences of the models they're building. They think about the feedback loops and potential gaming of their models.