COST PREDICTION
By
ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR

Four Weeks Training

at

Guru Nanak Dev Engineering College, Ludhiana
(From 16th July 2021 to 13th August 2021)

SUBMITTED BY

Abhay Tiwari
Information Technology
1921002
1905294

Information Technology Department
GURU NANAK DEV ENGINEERING COLLEGE
LUDHIANA, INDIA

COST PREDICTION
By
ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR

Four Weeks Training

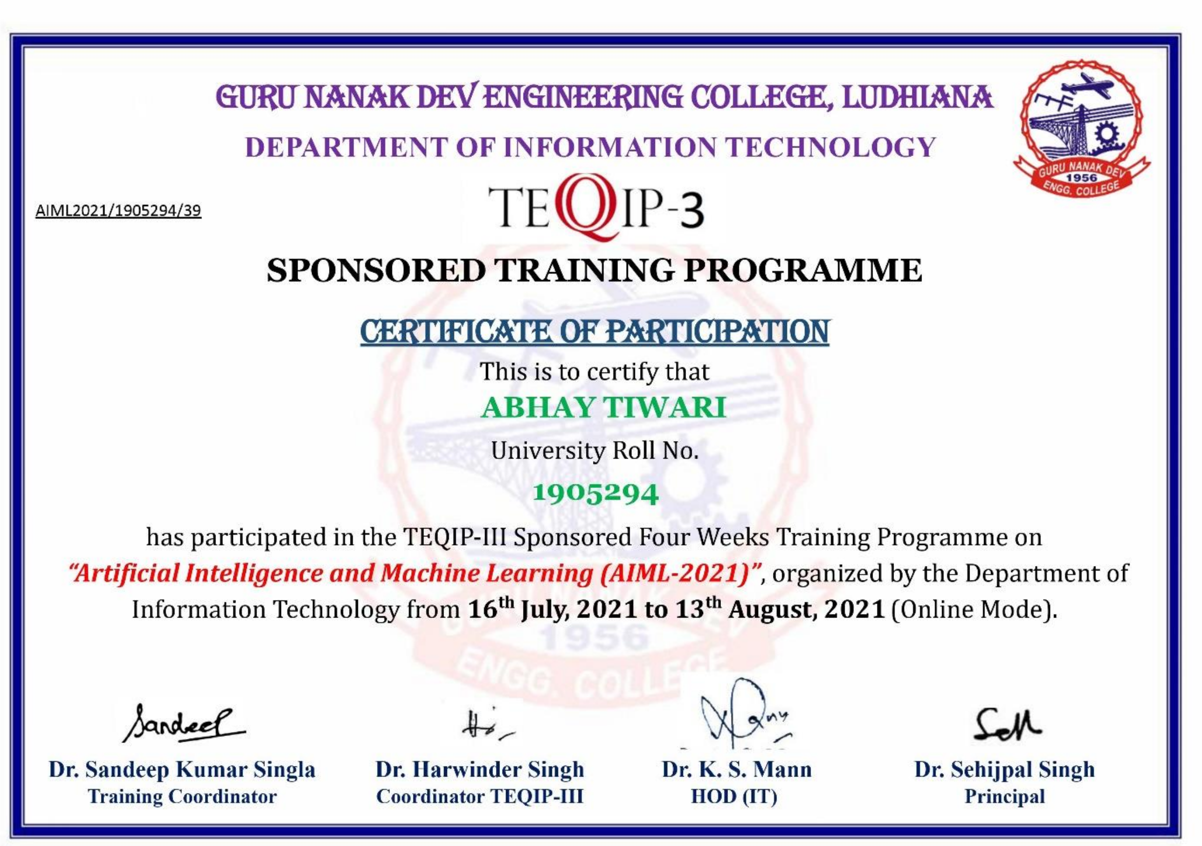
at

Guru Nanak Dev Engineering College, Ludhiana
(From 16th July 2021 to 13th August 2021)

SUBMITTED BY

Abhay Tiwari
Information Technology
1921002
1905294

Information Technology Department
GURU NANAK DEV ENGINEERING COLLEGE
LUDHIANA, INDIA



GURU NANAK DEV ENGINEERING COLLEGE, LUDHIANA

DEPARTMENT OF INFORMATION TECHNOLOGY

AIML2021/1905294/39

TEQIP-3

SPONSORED TRAINING PROGRAMME

CERTIFICATE OF PARTICIPATION

This is to certify that

ABHAY TIWARI

University Roll No.

1905294

has participated in the TEQIP-III Sponsored Four Weeks Training Programme on *"Artificial Intelligence and Machine Learning (AIML-2021)"*, organized by the Department of Information Technology from **16th July, 2021 to 13th August, 2021** (Online Mode).

Sandeep

Dr. Sandeep Kumar Singla
Training Coordinator

Dr. Harwinder Singh
Coordinator TEQIP-III

Dr. K. S. Mann
HOD (IT)

Dr. Sehijpal Singh
Principal

ACKNOWLEDGEMENT

WE are highly grateful to the Dr. Sehajpal Singh, Principal, Guru Nanak Dev Engineering College (GNDEC), Ludhiana, for providing this opportunity to carry out the four weeks industrial training at Guru Nanak Dev Engineering college.

The constant guidance and encouragement received from Dr. Kulvinder Singh Mann, HoD, IT Department, GNDEC Ludhiana has been of great help in carrying out the project work and is acknowledged with reverential thanks.

We would like to express a deep sense of gratitude and thanks profusely to Dr. Sandeep Kumar Singla, without her wise counsel and able guidance, it would have been impossible to complete the project in this manner.

We express gratitude to other faculty members of the computer science and engineering department of GNDEC for their intellectual support throughout the course of this work.

Finally, we are indebted to all whosoever have contributed in this report work.

Abhay Tiwari

1905294

Abstract

The term 'machine learning' is one of the most popular and frequently used terms of today. There is a non trivial possibility that you have heard this term at least once if you have some sort of familiarity with technology, no matter what domain you work in. The mechanics of machine learning, however, are a mystery to most people. Therefore, it is important to understand what machine learning actually is, and to learn about it step by step, through practical examples.

Machine learning, from a systems perspective, is defined as the creation of automated systems that can learn hidden patterns from data to aid in making intelligent decisions.

Artificial intelligence is the simulation of human intelligence processes by machines, especially computer systems. Specific applications of AI include expert systems, natural language processing, speech recognition and machine vision.

Health insurance today is compulsory and almost everyone has a connection with a public or private insurance company. The factors that determine the amount of insurance vary from company to company. Moreover, the villagers are unaware that the Indian government provides free health insurance to people below the poverty line. This is a very complicated method, and some villagers either have private health insurance or don't invest at all. It is also possible to deceive the insurance amount and purchase unnecessary expensive health insurance.

The project does not provide the exact amount required by health insurance companies, but provides a sufficient indication of the amount related to an individual for health insurance.

This should not be the only criterion when choosing health insurance because the prognosis is premature and does not meet the requirements of a specific company. Estimating the amount of your health plan early can help you better understand how much you will need

List of Figures

TABLE OF CONTENTS

Chapter 1:- Introduction To AI, ML and DL

In today's world, technology is growing very fast, and we are getting in touch with different new technologies day by day.

Here, one of the booming technologies of computer science is Artificial Intelligence, Machine Learning and Deep Learning which is ready to create a new revolution in the world by making intelligent machines.

1.1. Introduction to Artificial Intelligence (AI) :-

Artificial Intelligence is composed of two words **Artificial** and **Intelligence**, where Artificial defines *"man-made"* and intelligence defines *"thinking power"*, hence AI means *"a man-made thinking power."*

So, we can define AI as: -

It is a branch of computer science by which we can create intelligent machines which canbehave like a human, think like humans, and able to make decisions.

1.1.1. Importance of Artificial Intelligence :-

Before Learning about Artificial Intelligence, we should know that what is the importance of AI and why should we learn it. Following are some main reasons to learn about AI:

With the help of AI, you can create such software or devices which can solve real-worldproblems very easily and with accuracy such as health issues, marketing, traffic issues, etc.

With the help of AI, you can create your personal virtual Assistant, such as Cortana, GoogleAssistant, Siri, etc.

With the help of AI, you can build such Robots which can work in an environment wheresurvival of humans can be at risk..

- **AI automates repetitive learning and discovery through data.** Instead of automating manual tasks, AI performs frequent, high-volume, computerized tasks. And it does so reliably and without fatigue. Of course, humans are still essential toset up the system and ask the right questions.
- **AI adapts through progressive learning algorithms** to let the data do the programming. AI finds structure and regularities in data so that algorithms can acquire skills. Just as an algorithm can teach itself to play chess, it can teach itself what product to recommend next online. And the models adapt when given new data.

- **AI analyzes more and deeper data** using neural networks that have many hidden layers. Building a fraud detection system with five hidden layers used to be impossible. All that has changed with incredible computer power and big data. You need lots of data to train deep learning models because they learn directly from the data.

- **AI gets the most out of data.** When algorithms are self-learning, the data itself is an asset. The answers are in the data. You just have to apply AI to find them. Since the role of the data is now more important than ever, it can create a competitive advantage. If you have the best data in a competitive industry, even if everyone is applying similar techniques, the best data will win.

- **AI achieves incredible accuracy** through deep neural networks. For example, your interactions with Alexa and Google are all based on deep learning. And theseproducts keep getting more accurate the more you use them.

1.1.2. Advantages of Artificial Intelligence :-

Following are some main advantages of Artificial Intelligence:

- **High Accuracy with less errors:** AI machines or systems are prone to less errors and high accuracy as it takes decisions as per pre-experience or information.
- **High-Speed:** AI systems can be of very high-speed and fast-decision making, because of that AI systems can beat a chess champion in the Chess game.
- **High reliability:** AI machines are highly reliable and can perform the same actionmultiple times with high accuracy.
- **Digital Assistant:** AI can be very useful to provide digital assistant to the users such as AI technology is currently used by various E-commerce websites to show the products as per customer requirement.
- **Useful as a public utility:** AI can be very useful for public utilities such as a self- driving car which can make our journey safer and easy, facial recognition for security purpose, Natural language processing to communicate with the human inhuman-language, etc.

1.1.3. Disadvantage of Artificial Intelligence :-

Everything is coming with advantages and disadvantages even technologies also the same goes for artificial intelligence. Being so advantageous technology still, it has some disadvantages which we need to keep in our mind while creating an AI system. Following arethe disadvantages of AI:

- **High Cost:** The hardware and software requirement of AI is very costly as it requires lots of maintenance to meet current world requirements.
- **Can't think out of the box:** Even we are making smarter machines with AI, but still they cannot work out of the box, as the robot will only do that work for which they are trained, or programmed.

- **No Original Creativity:** As humans are so creative and can imagine some new ideas but still AI machines cannot beat this power of human intelligence and cannot be creative and imaginative.
- **Increase dependency on machines:** With the increment of technology, people aregetting more dependent on devices and hence they are losing their mental capabilities.
- **No feelings and emotions:** AI machines can be an outstanding performer, but stillit does not have the feeling so it cannot make any kind of emotional attachment with human, and may sometime be harmful for users if the proper care is not taken.

1.2. Introduction to Machine Learning (ML) :-

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

So, we can define ML as :-

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

The term 'machine learning' is one of the most popular and frequently used terms of today. There is a nontrivial possibility that you have heard this term at least once if you have some sort of familiarity with technology, no matter what domain you work in. The mechanics of machine learning, however, are a mystery to most people.

1.2.1. Importance of Machine Learning :-

The importance of machine learning can be easily understood by its uses cases, Currently, machine learning is used in self-driving cars, cyber fraud detection, face recognition, and friend suggestion by Facebook, etc. Various top companies such as Netflix and Amazon have build machine learning models that are using a vast amount of data to analyze the user interest and recommend product accordingly.

We can train machine learning algorithms by providing them the huge amount of data and let them explore the data, construct the models, and predict the required output automatically.

Advantage of Machine Learning :-

No Human Intervention Needed (Automation) :- With ML, you don’t need to babysit yourproject on every step of the way. As it means providing the machines the ability to learn, it lets them make predictions and also improve the algorithms on their own.

Continuous Improvement: - As Machine Learning algorithms gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions.

Handling Multi-Dimensional and Multi-Variety Data :- Its algorithms are excellent at handling data that are multi-dimensional and multi-variety, and they can do this in dynamic or random uncertain environments.

Wide Applications :- It has many wide applications such as banking, financial sector, healthcare, retail, publishing, etc.

1.2.2. Disadvantage of Machine Learning :-

Data Acquisition :- It needs huge data to train on, and these should be unbiased and of good quality. There can also be times where the algorithm must wait for new data to be generated and fetched.

Algorithm Selection :- The selection of an algorithm in Machine Learning is still a manual job. We have to run and test our data in all the algorithms. After that only we can decide what algorithm we want. We choose them on the basis of result accuracy. The process is very much time-consuming.

Time and Resources :- It needs much time to let the algorithms adapt, learn and develop in order to fulfill their purpose with a considerable amount of accuracy and relevancy. It also needs huge resources to function.

High error-susceptibility :- It is autonomous but highly susceptible to errors. If you train an algorithm with data sets small enough to not be inclusive, you end up with biased predictions coming from a biased training set. This leads to irrelevant advertisements being displayed to customers. In the case of Machine Learning, such blunders can set off a chain of errors that can go undetected for long periods of time. And when they do get noticed, it takes quite sometime to recognize the source of the issue, and even longer to correct it.

1.3. Introduction to Deep Learning(DL) :-

Deep learning is based on the branch of machine learning, which is a subset of artificial intelligence. Since neural networks imitate the human brain and so deep learning will do. In deep learning, nothing is programmed explicitly. Basically, it is a machine learning class that makes use of numerous nonlinear processing units so as to perform feature extraction as well as transformation. The output from each preceding layer is taken as input by each one of the successive layers.

Deep learning models are capable enough to focus on the accurate features themselves by requiring a little guidance from the programmer and are very helpful in solving out the problemof dimensionality. Deep learning algorithms are used, especially when we have a huge no of inputs and outputs.

Since deep learning has been evolved by the machine learning, which itself is a subset of artificial intelligence and as the idea behind the artificial intelligence is to mimic the human behavior, so same is "the idea of deep learning to build such algorithm that can mimic the brain".

Deep learning is implemented with the help of Neural Networks, and the idea behind themotivation of Neural Network is the biological neurons, which is nothing but a brain cell.

So, it can define as :-

Deep learning is a collection of statistical techniques of machine learning for learning feature hierarchies that are actually based on artificial neural networks.

So basically, deep learning is implemented by the help of deep networks, which are nothing but neural networks with multiple hidden layers.

Generally speaking, deep learning is a machine learning method that takes in an input X, and usesit to predict an output of Y. Given a large dataset of input and output pairs, a deep learning algorithm will try to minimize the difference between its prediction and expected output. By doingthis, it tries to learn the association/pattern between given inputs and outputs — this in turn allowsa deep learning model to generalize to inputs that it hasn't seen before.

Deep Learning Algorithms use something called a neural network to find associations between a set of inputs and outputs. The basic structure is seen below:

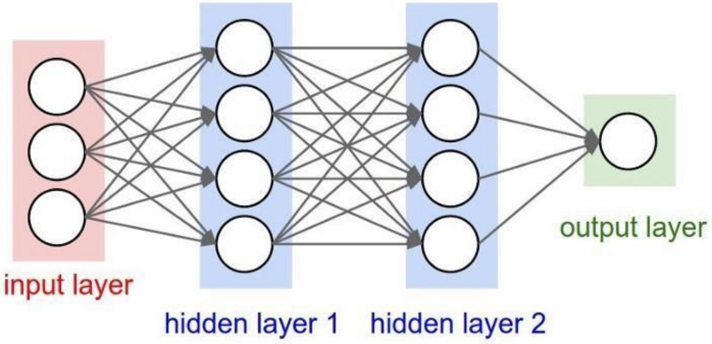


Fig 1.1 : Neural Network Structure

A neural network is composed of input, hidden, and output layers — all of which are composed of"nodes". Input layers take in a numerical representation of data (e.g. images with pixel specs), output layers output predictions, while hidden layers are correlated with most of the computation.

1.3.1. <u>Importance of Deep Learning</u> :-

The ability to process large numbers of features makes deep learning very powerful when dealing with unstructured data. However, deep learning algorithms can be overkill for less complex problems because they require access to a vast amount of data to be effective. For instance, Image Net, the common benchmark for training deep learning models for comprehensive image recognition, has access to over 14 million images.

If the data is too simple or incomplete, it is very easy for a deep learning model to become over fitted and fail to generalize well to new data. As a result, deep learning models are not as effective as other techniques (such as boosted decision trees or linear models) for most practical business problems such as understanding customer churn, detecting fraud transactions and other cases with smaller datasets and fewer features. In certain caseslike multiclass classification, deep learning can work for smaller, structured datasets.

1.3.2. **Advantages of Machine Learning** :-

The main advantages of Machine Learning is :-

- Features are automatically deduced and optimally tuned for desired outcome. Featuresare not required to be extracted ahead of time. This avoids time consuming machine learning techniques.

- Robustness to natural variations in the data is automatically learned.

- The same neural network based approach can be applied to many differentapplications and data types.

- Massive parallel computations can be performed using GPUs and are scalable for large volumes of data. Moreover it delivers better performance results when amountof data are huge.

- The deep learning architecture is flexible to be adapted to new problems in the future.

1.3.3. **Disadvantage of Deep Learning** :-

Some disadvantage of Deep Learning are :-

- It requires very large amount of data in order to perform better than other techniques.

- It is extremely expensive to train due to complex data models. Moreover deep learning requires expensive GPUs and hundreds of machines. This increases cost tothe users.

- There is no standard theory to guide you in selecting right deep learning tools as it requires knowledge of topology, training method and other parameters. As a result itis difficult to be adopted by less skilled people.

- It is not easy to comprehend output based on mere learning and requires classifiers todo so. Convolutional neural network based algorithms perform such tasks.

1.4. AI vs ML vs DL and Applications :-

In this new era of technology, companies and developers around the world are talking about artificial intelligence (AI), machine learning (ML), and deep learning (DL). Let us understand some basic difference and relation between these:

1.4.1. Artificial Intelligence vs Machine Learning vs Deep Learning :-

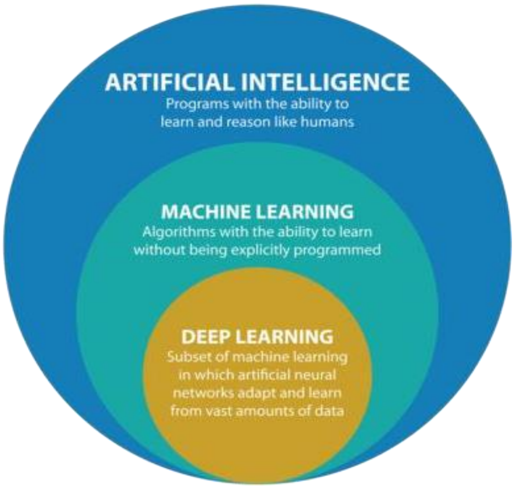

ARTIFICIAL INTELLIGENCE
Programs with the ability to learn and reason like humans
MACHINE LEARNING
Algorithms with the ability to learn without being explicitly programmed
DEEP LEARNING
Subset of machine learning in which artificial neural networks adapt and learn from vast amounts of data

Fig 1.2 : Relation of AI , ML and DL

Artificial Intelligence :-

AI stands for Artificial Intelligence, and is basically the study/process which enablesmachines to mimic human behaviour through particular algorithm.

AI is the broader family consisting of ML and DL as it's components.

AI is a computer algorithm which exhibits intelligence through decision

making.Search Trees and much complex math is involved in AI.

The aim is to basically increase chances of success and not accuracy.

Three broad categories/types Of AI are: Artificial Narrow Intelligence (ANI), ArtificialGeneral Intelligence (AGI) and Artificial Super Intelligence (ASI).

The efficiency Of AI is basically the efficiency provided by ML and DL respectively.

Examples of AI applications include: Google's AI-Powered Predictions, Ridesharing AppsLike Uber , Commercial Flights Use an AI Autopilot, etc.

Machine Learning :-

ML stands for Machine Learning, and is the study that uses statistical methods enablingmachines to improve with experience.

ML is the subset of AI.

ML is an AI algorithm which allows system to learn from data.

If you have a clear idea about the logic(math) involved in behind and you can visualize the complex functionalities like K-Mean, Support Vector Machines, etc., then it defines the ML aspect.

The aim is to increase accuracy not caring much about the success ratio.

Three broad categories/types Of ML are: Supervised Learning, Unsupervised Learning and Reinforcement Learning

Less efficient than DL as it can’t work for longer dimensions or higher amount of data.

Examples of ML applications include: Virtual Personal Assistants: Siri, Alexa, Google, etc., Email Spam and Malware Filtering.

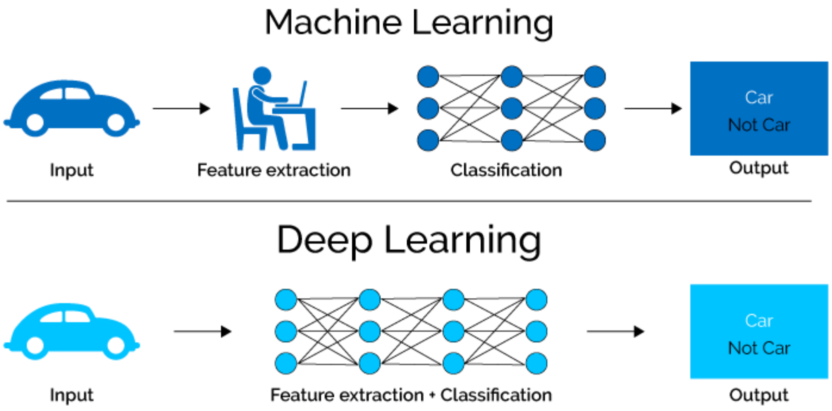


Fig 1.3 : difference between ML and DL

Deep Learning :-

DL stands for Deep Learning, and is the study that makes use of Neural Networks(similarto neurons present in human brain) to imitate functionality just like a human brain.

Examples of DL applications include: Sentiment based news aggregation, Image analysis andcaption generation, etc.

1.4.2. <u>Applications of AI,ML and DL</u> :-

AI, ML and DL has various applications in today’s society.

Applications of AI :-

Following are some sectors which have the application of Artificial Intelligence:-

- **AI in Healthcare :**

In the last, five to ten years, AI becoming more advantageous for the healthcare industry and going to have a significant impact on this industry.

Healthcare Industries are applying AI to make a better and faster diagnosis than humans. AI can help doctors with diagnoses and can inform when patients are worsening so that medical help can reach to the patient before hospitalization.

- **AI in Gaming :**

  AI can be used for gaming purpose. The AI machines can play strategic games like chess, where the machine needs to think of a large number of possible places.

- **AI in Finance :**

  AI and finance industries are the best matches for each other. The finance industry is implementing automation, chatbot, adaptive intelligence, algorithm trading, and machine learning into financial processes.

- **AI in Data Security :**

  The security of data is crucial for every company and cyber-attacks are growing very rapidly in the digital world. AI can be used to make your data more safe and secure. Some examples such as AEG bot, AI2 Platform,are used to determine software bug and cyber-attacks in a better way.

- **AI in Social Media :**

  Social Media sites such as Facebook, Twitter, and Snapchat contain billions of user profiles, which need to be stored and managed in a very efficient way. AI can organize and manage massive amounts of data. AI can analyze lots of data to identify the latest trends, hashtag, and requirement of different users.

- **AI in Automotive Industry :**

  Some Automotive industries are using AI to provide virtual assistant to their user for better performance. Such as Tesla has introduced TeslaBot, an intelligent virtual assistant.

  Various Industries are currently working for developing self-driven cars which can make your journey more safe and secure.

- **AI in Robotics :**

  Artificial Intelligence has a remarkable role in Robotics. Usually, general robots are programmed such that they can perform some repetitive task, but with the help of AI, we can create intelligent robots which can perform tasks with their own experiences without pre-programmed.

Humanoid Robots are best examples for AI in robotics, recently the intelligent Humanoid robot named as Erica and Sophia has been developed which can talk and behave like humans.

- **AI in Entertainment :**

  We are currently using some AI based applications in our daily life with some entertainment services such as Netflix or Amazon. With the help of ML/AI algorithms, these services show the recommendations for programs or shows.

- **AI in Agriculture :**
  Agriculture is an area which requires various resources, labor, money, and time for best result. Now a day's agriculture is becoming digital, and AI is emerging in this field. Agriculture is applying AI as agriculture robotics, solid and crop monitoring, predictive analysis. AI in agriculture can be very helpful for farmers.

- **AI in E-commerce :**
  AI is providing a competitive edge to the e-commerce industry, and it is becoming more demanding in the e-commerce business. AI is helping shoppers to discover associated products with recommended size, color, or even brand.

- **AI in education :**
  AI can automate grading so that the tutor can have more time to teach. AI chatbot cancommunicate with students as a teaching assistant.
  AI in the future can be work as a personal virtual tutor for students, which will beaccessible easily at any time and any place.

Application of ML :-

Machine learning is a buzzword for today's technology, and it is growing very rapidly day byday. We are using machine learning in our daily life even without knowing it such as Google

Maps, Google assistant, Alexa, etc. Below are some most trending real-world applications ofMachine Learning:

- **Image Recognition :**

  Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc. The popular use case of image recognition and face detection is, Automatic friend tagging suggestion

  Facebook provides us a feature of auto friend tagging suggestion. Whenever we upload a photo with our Facebook friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning's face detection and recognition algorithm.

- **Speech Recognition :**

  While using Google, we get an option of "Search by voice," it comes under speech recognition, and it's a popular application of machine learning.

  Speech recognition is a process of converting voice instructions into text, and it is alsoknown as "Speech to text", or "Computer speech recognition**.**" At present, machine learning algorithms are widely used by various applications of speech recognition. Google assistant**,** Siri**,** Cortana**,** and Alexa are using speech recognition technology to follow the voice instructions.

- **Email Spam and Malware Filtering :**

  Whenever we receive a new email, it is filtered automatically as important, normal, and spam. We always receive an important mail in our inbox with the important symbol and spam emails in our spam box, and the technology behind this is Machine learning. Below are some spam filters used by Gmail:

  Content Filter, Header filter, General blacklists filter, Rules-based filters, Permission filters

- **Virtual Personal Assistant :**

  We have various virtual personal assistants such as Google assistant**,** Alexa**,** Cortana**,** Siri. As the name suggests, they help us in finding the information using our voice instruction. These assistants can help us in various ways just by our voice instructions such as Play music, call someone, Open an email, Scheduling an appointment, etc. These virtual assistants use machine learning algorithms as an important part.

- **Online Fraud Detection :**

  Machine learning is making our online transaction safe and secure by detecting fraud transaction. Whenever we perform some online transaction, there may be various ways that a fraudulent transaction can take place such as fake accounts, fake ids, and steal money in the middle of a transaction. So to detect this, Feed Forward Neural network helps us by checking whether it is a genuine transaction or a fraud transaction.

  For each genuine transaction, the output is converted into some hash values, and thesevalues become the input for the next round. For each genuine transaction, there is a specific pattern which gets change for the fraud transaction hence, it detects it and makes our online transactions more secure.

- **Automatic Language Translation :**

  Nowadays, if we visit a new place and we are not aware of the language then it is nota problem at all, as for this also machine learning helps us by converting the text into our known languages. Google's GNMT (Google Neural Machine Translation) providethis feature, which is a Neural Machine Learning that translates the text into our familiar language, and it called as automatic translation.

- **Stock Market trading :**

  Machine learning is widely used in stock market trading. In the stock market, there is always a risk of up and downs in shares, so for this machine learning's long short termmemory neural network is used for the prediction of stock market trends.

Applications of DL :-

- **Automatic Colorization of Black and White Images :**

  Image colorization is the problem of adding colour to black and white photographs. Traditionally this was done by hand with human effort because it is such a difficult task. Deep learning can be used to use the objects and their context within the photograph to colour the image, much like a human operator might approach the problem. Generally the approach involves the use of very large convolutional neural networks and supervised layers that recreate the image with the addition of colour.

- **Automatic Machine Translation :**

  This is a task where given words, phrase or sentence in one language, automatically translate it into another language. Automatic machine translation has been around fora long time, but deep learning is achieving top results in two specific areas:

  Automatic Translation of Text. Automatic Translation of Images.

  Text translation can be performed without any preprocessing of the sequence, allowing the algorithm to learn the dependencies between words and their mapping to a new language. Stacked networks of large LSTM recurrent neural networks are used to perform this translation.

  As you would expect, convolutional neural networks are used to identify images that have letters and where the letters are in the scene. Once identified, they can be turned into text, translated and the image recreated with the translated text. This is often called instant visual translation.

- **Automatic Game Playing :**

  This is a task where a model learns how to play a computer game based only on the pixels on the screen. This very difficult task is the domain of deep reinforcement models and is the breakthrough that DeepMind (now part of google) is renown for achieving.

1.5. Traditional Programming VS Machine Learning :-

Traditional programming :- Traditional programming is a manual process—meaning a person (programmer) creates the program. But without anyone programming the logic, one has to manually formulate or code rules.

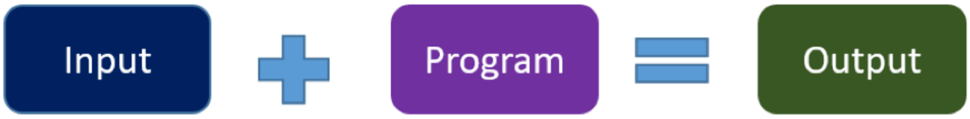


Fig 1.4 : Traditional Programming

In machine learning, on the other hand, the algorithm automatically formulates the rules from thedata.

Machine learning :- Unlike traditional programming, machine learning is an automated process. It can increase the value of your embedded analytics in many areas, including data prep, natural language interfaces, automatic outlier detection, recommendations, and causality and significancedetection. All of these features help speed user insights and reduce decision bias.

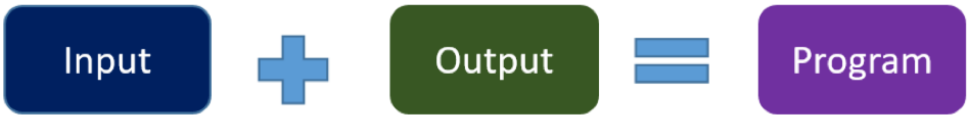


Fig 1.5 : Machine Learning

For example, if you feed in customer demographics and transactions as input data and use historical customer churn rates as your output data, the algorithm will formulate a program that can predict if a customer will churn or not. That program is called a predictive model**.**

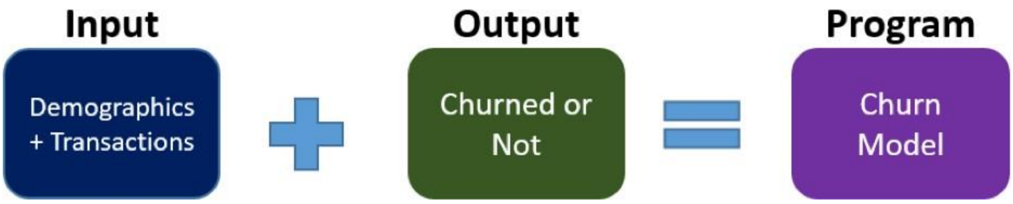


Fig 1.6 : Machine Learning Program

1.6. Programming Language Support :-

If you are interested in working on your own AI projects, then you will need to know what the most popular AI programming languages are.

There are quite a few AI programming languages, and there is none of them that can be called “the best AI programming language.” They all have their pros and cons, and today we will show some of popular languages.

- LISP
- Python
- C++
- Java
- Prolog
- R
- Scala

We can use all these languages for AI and Machine Learning.

Why we use most python for AI and Machine Learning ?

Python is an AI programming language that has gained huge popularity. The main reasonsare the simple syntax, less coding and a large number of available libraries ready for use. Simple syntax means you can focus on the core value of programming, thinking, or problem- solving.

The earlier mentioned libraries include NumPy, SciPy, matplotlib, nltk, SimpleAI. Python is an open-source AI programming language. That’s why it has a huge fan base among programmers. Because it can be used broadly, to make small scripts and up to enterprise applications, it’s suitable for AI.

Where other AI programming languages use punctuation, Python uses English keywords. It’s designed to be readable. It has only a few keywords and has a clearly defined syntax. If you are a student, you will pick up the language quickly. The libraries are portable across platforms such as UNIX, Windows, and Macintosh.

It also provides interfaces for all major commercial databases. When it comes to scalability, itprovides a better structure and support for large enterprise programs than it does for simple shell scripts.

2. Chapter 2 :- Workflow of ML Project

2.1. Workflow

The workflow of the Machine Learning project includes the defining the problem and Collection of data, Analyzation of data, Preparation of data, Evaluation of Algorithms, Resultsimprovisation and Present Results.

- **Define Problem OR Collect Data :-**

Investigation and characterization of problem is mandatory to better understand the goals of the project. Gathering data is one of the most important stages of machine learning workflows. During data collection, you are defining the potential usefulness and accuracy of your project with the quality of the data you collect. To collect data, you need to identify yoursources and aggregate data from those sources into a single dataset. This could mean streaming data from Internet of Things sensors, downloading open source data sets, or constructing a data lake from assorted files, logs, or media.

- **Analyzing data :-**

After getting the dataset, the next step in the model building workflow is almost always data visualization. Specifically, we’ll perform exploratory data analysis on the data to accomplish several tasks:

1. View data distributions
2. Identify skewed predictors
3. Identify outliers

Use descriptive statistics and visualization to better understand the data you have available
i.e. Plotting and finding features.

- **Prepare Data :-**

Machine Learning algorithms are completely dependent on data because it is the most crucialaspect that makes model training possible. On the other hand, if we won’t be able to make sense out of that data, before feeding it to ML algorithms, a machine will be useless. In simple words, we always need to feed right data i.e. the data in correct scale, format and containing meaningful features, for the problem we want machine to solve.

This makes data preparation the most important step in ML process. Data preparation may bedefined as the procedure that makes our dataset more appropriate for ML process.

After selecting the raw data for ML training, the most important task is data pre-processing. In broad sense, data preprocessing will convert the selected data into a form we can work with or can feed to ML algorithms. We always need to preprocess our data so that it can beas per the expectation of machine learning algorithm.

- **Evaluate Algorithms :-**

One of the core tasks in building any machine learning model is to evaluate its performance. While data preparation and training a machine learning model is a key step in the machine learning pipeline, it’s equally important to measure the performance of this trained model. How well the model generalizes on the unseen data is what defines adaptive v/snon-adaptive machine learning models. Without doing a proper evaluation of the ML model using different metrics, and depending only on accuracy, can lead to a problem when the respective model is deployed on unseen data and can result in poor predictions. This happens because, in cases like these, our models don’t learn but instead memorize; hence, they cannot generalize well on unseen data.

- **Improve Results :-**

Use algorithm tuning and ensemble methods to get the most out of well-performing algorithms on your data. Having one or two algorithms that perform reasonably well on a problem is a good start, but sometimes you may be incentivised to get the best result you can given the time and resources you have available. When tuning algorithms you must have a high confidence in the results given by your test harness. This means that you should be usingtechniques that reduce the variance of the performance measure you are using to assess algorithm runs.

- **Present Results :-**

Finalize the model, make predictions and present results i.e. use the model with field data.

2.2. Data Preprocessing

A simple definition could be that data preprocessing is a data mining technique to turn the raw data gathered from diverse sources into cleaner information that’s more suitable for work. In other words, it’s a preliminary step that takes all of the available information to organize it, sortit, and merge it.

Raw data can have missing or inconsistent values as well as present a lot of redundant information. The most common problems you can find with raw data can be divided into 3 groups:

- **Missing data :-**

You can also see this as inaccurate data since the information that isn’t there creates gaps thatmight be relevant to the final analysis. Missing data often appears when there’s a problem in the collection phase, such as a glitch that caused a system’s downtime, mistakes in data entry,or issues with biometrics use, among others.

- **Inconsistent data :-**

Inconsistencies happen when you keep files with similar data in different formats and files. Duplicates in different formats, mistakes in codes of names, or the absence of data constraintsoften lead to inconsistent data, introduces deviations that you have to deal with before analysis.

If you didn't take care of those issues, the final output would be plagued with faulty insights. That's especially true for more sensitive analysis that can be more affected by small mistakes, like when it's used in new fields where minimal variations in raw data can lead to wrong assumptions.

Machine learning algorithms make assumptions about the dataset you are modelling. Often, raw data is comprised of attributes with varying scales. For example, one attribute may be in kilograms and another may be a count. Although not required, you can often get a boost in performance by carefully choosing methods to rescale your data. So the data should be normalized or standardized or normalized before any further proceeding, depending on the situation.

⇒ **Scaling** :-

Most probably our dataset comprises of the attributes with varying scale, but we cannot provide such data to ML algorithm hence it requires rescaling. Data rescaling makes sure that attributes are at same scale. Generally, attributes are rescaled into the range of 0 and 1. ML algorithms like gradient descent and k-Nearest Neighbours requires scaled data. We can rescale the data with the help of MinMaxScaler class of scikit-learn Python library.

⇒ **Normalization** :-

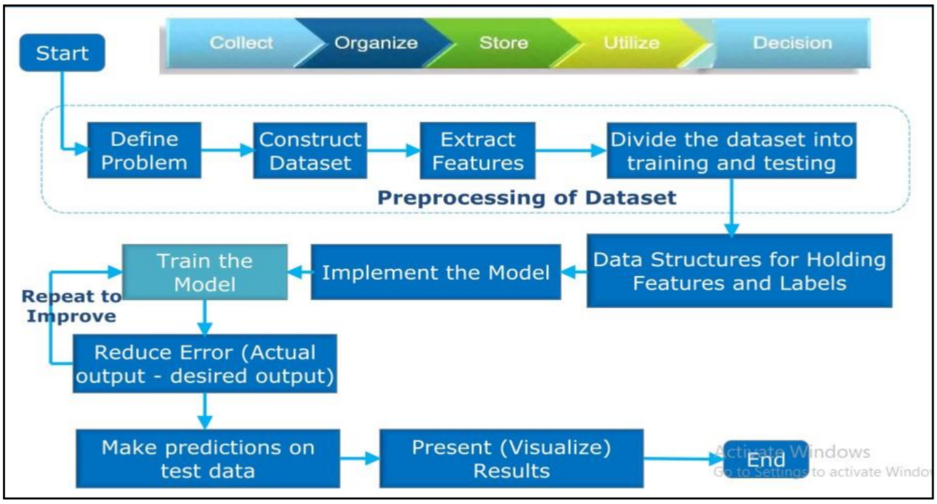
Another useful data preprocessing technique is Normalization. This is used to rescale each row of data to have a length of 1. It is mainly useful in Sparse dataset where we have lots of zeros. We can rescale the data with the help of Normalizer class of scikit-learn Python library.

⇒ **Binarization** :-

As the name suggests, this is the technique with the help of which we can make our data binary. We can use a binary threshold for making our data binary. The values above that threshold value will be converted to 1 and below that threshold will be converted to 0. For example, if we choose threshold value = 0.5, then the dataset value above it will become 1 and below this will become 0. That is why we can call it binarizing the data or thresholding the data. This technique is useful when we have probabilities in our dataset and want to convert them into crisp values.

⇒ **Standardization** :-

Another useful data preprocessing technique which is basically used to transform the data attributes with a Gaussian distribution. It differs the mean and SD (Standard Deviation) to a standard Gaussian distribution with a mean of 0 and a SD of 1. This technique is useful in ML algorithms like linear regression, logistic regression that assumes a Gaussian distribution in input dataset and produce better results with rescaled data. We can standardize the data (mean = 0 and SD =1) with the help of StandardAero class of scikit-learn Python library.



Collect
Organize
Store
Utilize
Decision
Start
Define Problem
Construct Dataset
Extract Features
Divide the dataset into training and testing
Preprocessing of Dataset
Data Structures for Holding Features and Labels
Implement the Model
Train the Model
Repeat to Improve
Reduce Error (Actual output - desired output)
Make predictions on test data
Present (Visualize) Results
End
Activate Windows
Go to Settings to activate Windo

Fig 2.1 : Workflow of project Machine Learning

3. Chapter 3 :- Machine Learning And Its Types

In this Chapter we have to focus on detail content about machine Learning. According to the definition that is :-

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

Machine learning is a method of data analysis that automates analytical model building. It is a branchof artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

Types of Machine Learning :-

Machine Learning is basically divided into four categories :-

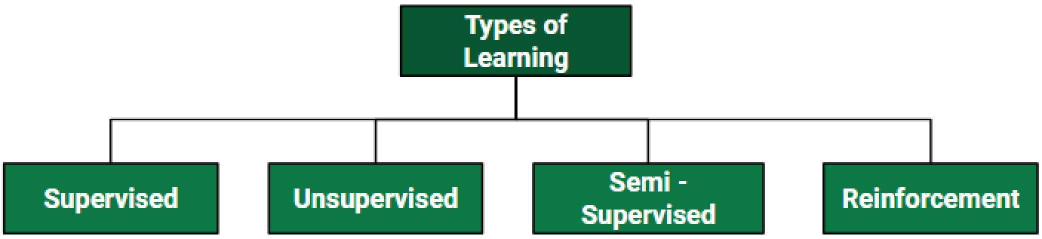
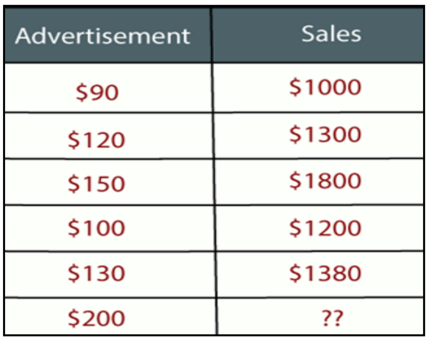


Fig 3.1 : Types of Machine Learning

- **Supervised Machine Learning**
- **Unsupervised Machine Learning**
- **Semi-Supervised Machine Learning**
- **Reinforcement Machine Learning**

3.1. Supervised Machine Learning :-



Advertisement	Sales
$90	$1000
$120	$1300
$150	$1800
$100	$1200
$130	$1380
$200	??

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

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

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

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

Supervised learning can be further divided into two types of problems:

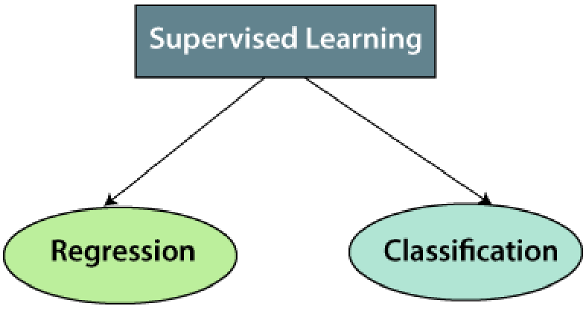


Fig 3.2 : Types of Supervised Machine Learning

- Regression
- Classification

3.1.1. __Regression__ :-

Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc.

Example: Suppose there is a marketing company A, who does various advertisement every year and get sales on that. The below list shows the advertisement made by the company in the last 5 years and the corresponding sales:

Fig 3.3 : Example of Regression

Now, the company wants to do the advertisement of $200 in the year 2019 and wants to knowthe prediction about the sales for this year. So to solve such type of prediction problems in machine learning, we need regression analysis.

In Regression, we plot a graph between the variables which best fits the given data points, using this plot, the machine learning model can make predictions about the data. In simple words, "Regression shows a line or curve that passes through all the data points on target- predictor graph in such a way that the vertical distance between the data points and the regression line is minimum." The distance between data points and line tells whether a model has captured a strong relationship or not.

3.1.2. Classification :-

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, Yes or No, 0 or 1, Spam or Not Spam, cat or dog, etc. Classes can be called as targets/labels or categories.

Unlike regression, the output variable of Classification is a category, not a value, such as "Green or Blue", "fruit or animal", etc. Since the Classification algorithm is a Supervised learning technique, hence it takes labelled input data, which means it contains input with the corresponding output.

The best example of an ML classification algorithm is Email Spam Detector.

The main goal of the Classification algorithm is to identify the category of a given dataset, and these algorithms are mainly used to predict the output for the categorical data.

Classification algorithms can be better understood using the below diagram. In the below diagram, there are two classes, class A and Class B. These classes have features that are similar to each other and dissimilar to other classes.

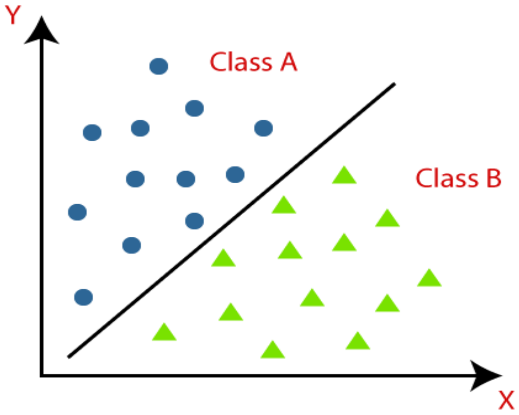


Fig 3.4 : Classification

3.2. Unsupervised Machine Learning :-

Unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead, models itself find the hidden patterns and insights from the given data. Itcan be compared to learning which takes place in the human brain while learning new things. It can be defined as :-

Unsupervised learning is a type of machine learning in which models are trained using unlabelled dataset and are allowed to act on that data without any supervision.

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

Below are some main reasons which describe the importance of Unsupervised Learning:

- Unsupervised learning is helpful for finding useful insights from the data.
- Unsupervised learning is much similar as a human learns to think by their ownexperiences, which makes it closer to the real AI.
- Unsupervised learning works on unlabelled and uncategorized data which makeunsupervised learning more important.
- In real-world, we do not always have input data with the corresponding output so to solvesuch cases, we need unsupervised learning.

The unsupervised learning algorithm can be further categorized into two types of problems:

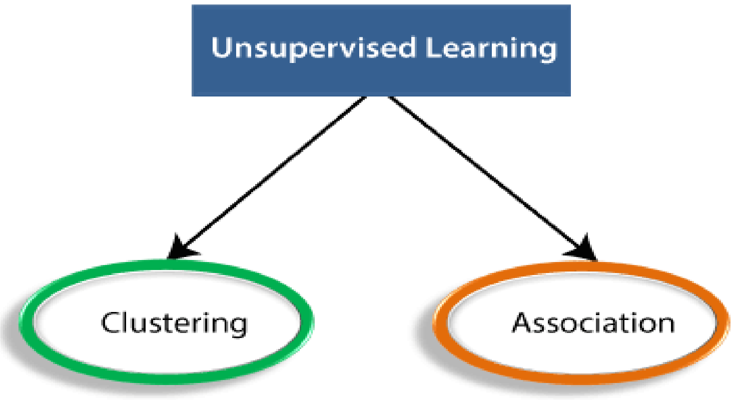


Fig 3.5 : Types of Unsupervised Learning

- Clustering
- Association

3.2.1. Clustering :-

Clustering or cluster analysis is a machine learning technique, which groups the unlabelled dataset. It can be defined as :

A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group.

It does it by finding some similar patterns in the unlabelled dataset such as shape, size, colour, behaviour, etc., and divides them as per the presence and absence of those similar patterns. It is an unsupervised learning method, hence no supervision is provided to the algorithm, and it deals with the unlabelled dataset. After applying this clustering technique, each cluster or group is provided with a cluster-ID. ML system can use this id to simplify the processing of large and complex datasets. The clustering technique is commonly used for statistical data analysis.

Example: Let's understand the clustering technique with the real-world example of Mall: When we visit any shopping mall, we can observe that the things with similar usage are grouped together. Such as the t-shirts are grouped in one section, and trousers are at other sections, similarly, at vegetable sections, apples, bananas, Mangoes, etc., are grouped in separate sections, so that we can easily find out the things. The clustering technique also works in the same way.

The clustering methods are broadly divided into **Hard clustering** (data point belongs to only one group) and **Soft Clustering** (data points can belong to another group also). But there are also other various approaches of Clustering exist. Below are the main clustering methods used in Machine learning:

- **Partitioning Clustering :-**

  It is a type of clustering that divides the data into non-hierarchical groups. It is also known as the centroid-based method. The most common example of partitioning clustering is the K-Means Clustering algorithm.

  In this type, the dataset is divided into a set of k groups, where K is used to define the number of pre-defined groups. The cluster centre is created in such a way that the distance between the data points of one cluster is minimum as compared to another cluster centroid.

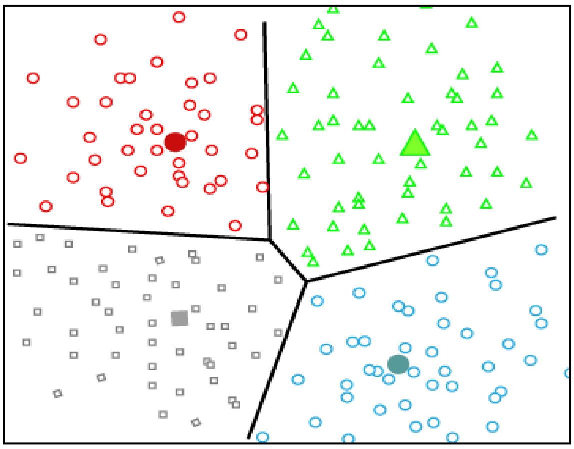


Fig 3.6 : Partitioning Clustering

- **Density-Based Clustering :-**

The density-based clustering method connects the highly-dense areas into clusters, and the arbitrarily shaped distributions are formed as long as the dense region can be connected. This algorithm does it by identifying different clusters in the dataset and connects the areas of high densities into clusters. The dense areas in data space aredivided from each other by sparser areas.

These algorithms can face difficulty in clustering the data points if the dataset has varyingdensities and high dimensions.

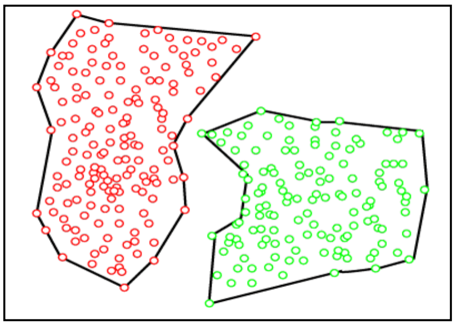


Fig 3.7 : Density-Based Clustering

- **Grid-based Methods Clustering :-**

In this method the data space is formulated into a finite number of cells that form a grid- like structure. All the clustering operation done on these grids are fast and independent ofthe number of data objects example STING (Statistical Information Grid), wave cluster, CLIQUE (CLustering In Quest) etc.

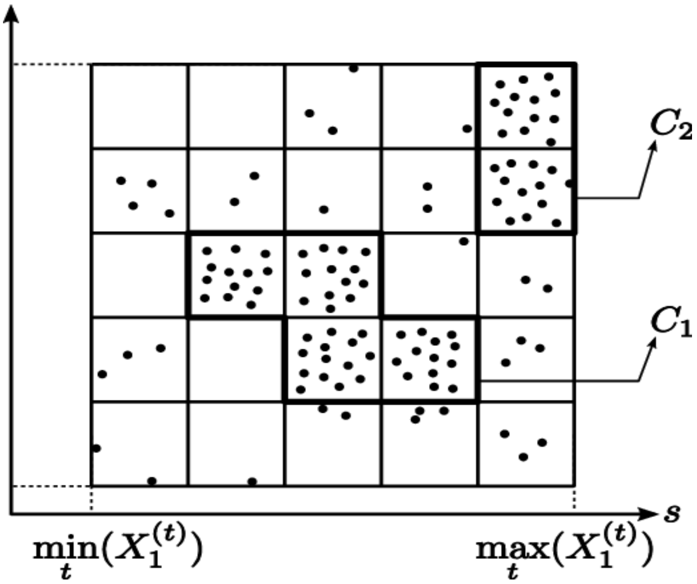


Fig 3.8 : Grid-based Methods Clustering

- **Hierarchical Clustering :-**

Hierarchical clustering can be used as an alternative for the partitioned clustering as there is no requirement of pre-specifying the number of clusters to be created. In this technique,the dataset is divided into clusters to create a tree-like structure, which is also calleda dendrogram. The observations or any number of clusters can be selected by cutting the tree at the correct level. The most common example of this method is the Agglomerative Hierarchical algorithm.

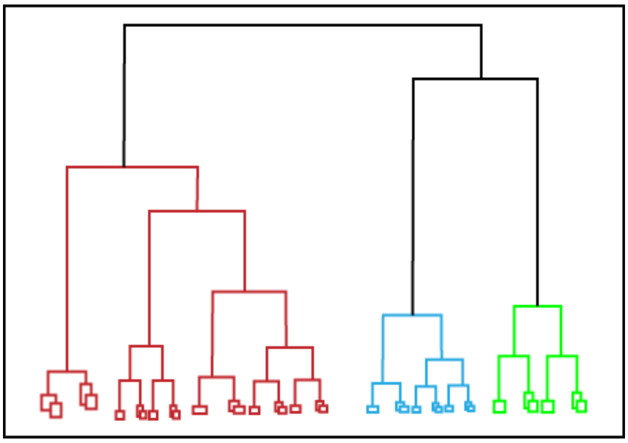


Fig 3.9 : Hierarchical Clustering

3.2.2. <u>Association</u> :-

Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable. It tries to find some interesting relations or associations among the variables of dataset. It is based on different rules to discover the interesting relations between variables in the database.

The association rule learning is one of the very important concepts of machine learning, andit is employed in Market Basket analysis, Web usage mining, continuous production, etc. Here market basket analysis is a technique used by the various big retailer to discover theassociations between items. We can understand it by taking an example of a supermarket, as in a supermarket, all products that are purchased together are put together.

For example, if a customer buys bread, he most likely can also buy butter, eggs, or milk, so these products are stored within a shelf or mostly nearby. Consider the below diagram:

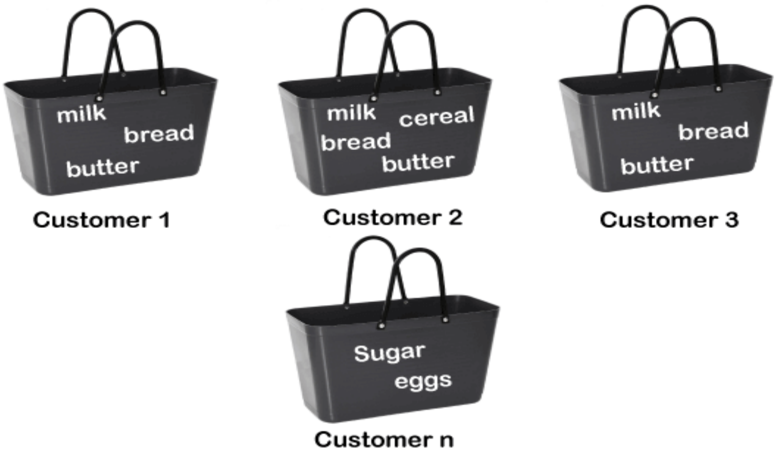


Fig 3.10 : Association Example

3.3. Semi-Supervised Machine Learning :-

Before understanding the Semi-Supervised learning, you should know the main categories of Machine Learning algorithms. Machine Learning consists of three main categories: SupervisedLearning, Unsupervised Learning, and Reinforcement Learning. Further, the basic difference between Supervised and unsupervised learning is that supervised learning datasets consist of an output label training data associated with each tuple, and unsupervised datasets do not consist thesame. Semi-supervised learning is an important category that lies between the Supervised and Unsupervised machine learning. so we can define as :

Semi-Supervised learning is a type of Machine Learning algorithm that represents the intermediate ground between Supervised and Unsupervised learning algorithms. It uses thecombination of labelled and unlabelled datasets during the training period.

3.4. Reinforcement Machine Learning :-

Reinforcement Learning is a feedback-based Machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions. For each good action, the agent gets positive feedback, and for each bad action, the agent gets negative feedback or penalty. In Reinforcement Learning, the agent learns automatically using feedbacks without any labelled data, unlike supervised learning. Since there is no labelled data,so the agent is bound to learn by its experience only. RL solves a specific type of problem where decision making is sequential, and the goal is long-term, such as game-playing, robotics, etc.

The agent learns with the process of hit and trial, and based on the experience, it learns to perform the task in a better way. Hence, we can say that

Reinforcement learning is a type of machine learning method where an intelligent agent (computer program) interacts with the environment and learns to act within that.

It is a core part of Artificial intelligence, and all AI agent. works on the concept of reinforcementlearning. Here we do not need to pre-program the agent, as it learns from its own experience without any human intervention.

Example: Suppose there is an AI agent present within a maze environment, and his goal is to find the diamond. The agent interacts with the environment by performing some actions, and based on those actions, the state of the agent gets changed, and it also receives a reward or penalty as feedback.

Types of Reinforcement learning :- There are mainly two types of reinforcement learning, which are:

- Positive Reinforcement
- Negative Reinforcement

3.4.1. Positive Reinforcement :-

The positive reinforcement learning means adding something to increase the tendency that expected behaviour would occur again. It impacts positively on the behaviour of the agent and increases the strength of the behaviour. This type of reinforcement can sustain the changes for a long time, but too much positive reinforcement may lead to an overload of states that can reduce the consequences.

3.4.2. Negative Reinforcement :-

The negative reinforcement learning is opposite to the positive reinforcement as it increases the tendency that the specific behaviour will occur again by avoiding the negative condition. It can be more effective than the positive reinforcement depending on situation and behaviour, but it provides reinforcement only to meet minimum behaviour.

4. Chapter 4 :- Supervised Machine Learning Algorithms

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

The most widely used learning algorithms are:

- Linear Regression
- Multiple Linear Regression (MLR)
- Polynomial Regression
- Logistic Regression
- Support Vector Machines (SVMs)
- Decision Trees
- Naive Bayes Classifier
- Random Forests
- Neural networks

4.1. Linear Regression :-

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

A linear regression line equation is written in the form of :- **Y = B0 + B1 X**

where X is the independent variable and plotted along

the x-axisY is the dependent variable and plotted

along the y-axis

The slope of the line is B1, and B0 is the intercept (the value of y when x = 0).

Technically, B0 is called the intercept because it determines where the line intercepts the y-axis. In machine learning we can call this the bias, because it is added to offset all predictions that we make. The B1 term is called the slope because it defines the slope of the line or how x translates into a (y) value before we add our bias. The goal is to find the best estimates for the coefficients to minimize the errors in predicting y from x.

B1 can be estimated as :

$$B1 = \frac{\sum_{i=1}^{n}(x_i - mean(x)) \times (y_i - mean(y))}{\sum_{i=1}^{n}(x_i - mean(x))^2}$$

Fig 4.1 : (B1) Slope of Line

$$B0 = mean(y) - B1 \times mean(x)$$

Fig 4.2 : (B0) Intercept

Root Mean Square Error : Known as RMSE

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - y_i)^2}{n}}$$

Fig 4.3 : RMSE

p is the predicted value and y is the actual value, i is the index for a specific instance, because wemust calculate the error across all predicted values.

4.2. <u>Multiple Linear Regression</u> :-

Multiple Linear Regression is one of the important regression algorithms which models the linearrelationship between a single dependent continuous variable and more than one independent variable.

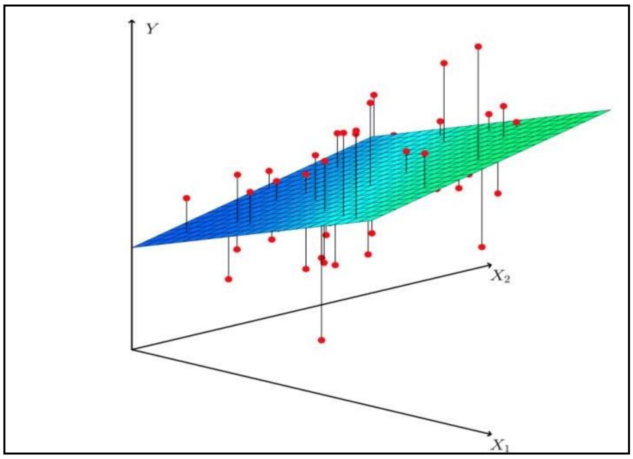


Fig 4.4 : Multiple Linear Regression

For MLR, the dependent or target variable(Y) must be the continuous/real, but the predictor orindependent variable may be of continuous or categorical form.

Each feature variable must model the linear relationship with the dependent

variable.MLR tries to fit a regression line through a multidimensional space

of data-points.

In Multiple Linear Regression, the target variable(Y) is a linear combination of multiple predictor variables x_1, x_2, x_3, ...,x_n. Since it is an enhancement of Simple Linear Regression, so the same is applied for the multiple linear regression equation, the equation becomes :

y = B0 + B1 * xi1 + B2 * xi2 + B3 * xi3…….. Bn *xin + e

where, y is a

dependent variable.xi

are independent

variable.

B0 is a constant and B1 …Bn are coefficients that we need

to estimate.e is the model’s random error (residual) term.

4.3. Polynomial Regression :-

Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. The Polynomial Regressionequation is given below:

y= b0+b1x1+ b2x12+ b2x13+...... bnx1n

It is also called the special case of Multiple Linear Regression in ML. Because we add somepolynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression.

It is a linear model with some modification in order to increase the accuracy. The dataset used inPolynomial regression for training is of non-linear nature. It makes use of a linear regression model to fit the complicated and non-linear functions and datasets.

Hence, "In Polynomial regression, the original features are converted into Polynomial features ofrequired degree (2,3,..,n) and then modelled using a linear model."

The need of Polynomial Regression in ML can be understood in the below points:

- If we apply a linear model on a linear dataset, then it provides us a good result as we have seen in Simple Linear Regression, but if we apply the same model without any modification on a non-linear dataset, then it will produce a drastic output. Due to which loss function will increase, the error rate will be high, and accuracy will be decreased.
- So for such cases, where data points are arranged in a non-linear fashion, we need the
  Polynomial Regression model. We can understand it in a better way using the below comparison diagram of the linear dataset and non-linear dataset.

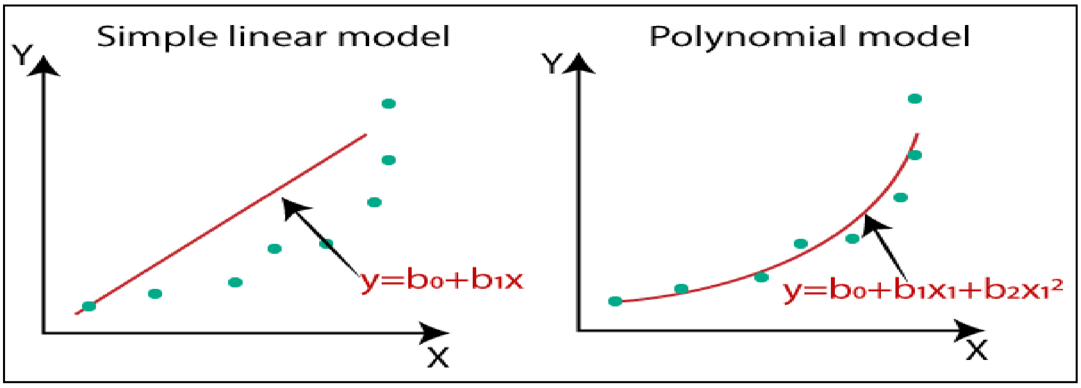


Fig 4.5 : Need of Polynomial model

In the above image, we have taken a dataset which is arranged non-linearly. So if we try to cover it with a linear model, then we can clearly see that it hardly covers any data point. On the other hand, a curve is suitable to cover most of the data points, which is of the Polynomial model.

Hence, if the datasets are arranged in a non-linear fashion, then we should use the Polynomial Regression model instead of Simple Linear Regression.

4.4. Logistic Regression :-

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for given set of features(or inputs), X.

Contrary to popular belief, logistic regression is a regression model. The model builds aregression model to predict the probability that a given data entry belongs to the categorynumbered as “1”. Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function.

$$y = \frac{1}{1+e^{-z}} + \varepsilon \text{ where } z = w \cdot x$$

W are the regression

coefficientZ is linear

in x

Response y is a sigmoid function of linear

combination of xi.y Is obtained within 0 and 1

Essentially a single layer perception/ ANN with sigmoid activation function

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:

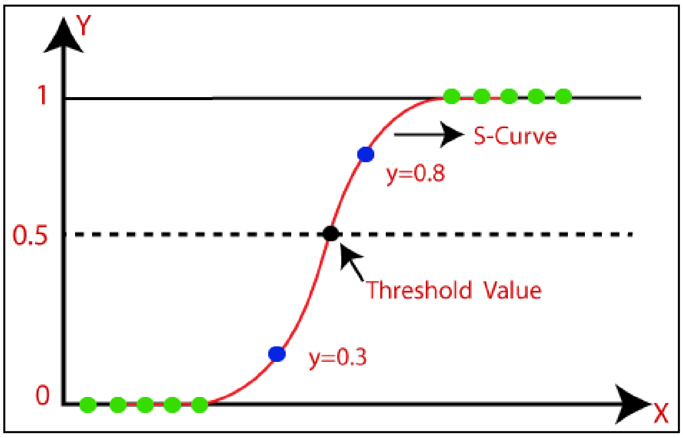


Fig 4.6 : Logistic Function Graph

Assumptions for Logistic Regression:

1. The dependent variable must be categorical in nature.
2. The independent variable should not have multi-collinearity

4.5. <u>Support Vector Machines(SVMs)</u> :-

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm 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 extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane

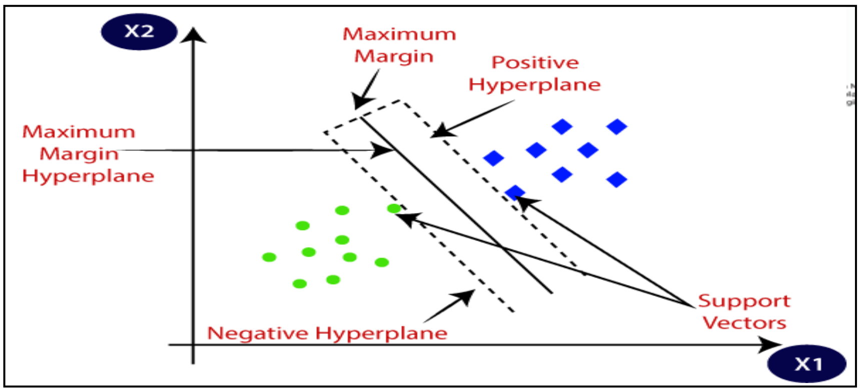


Fig 4.7 : SVM

Example: SVM can be understood with the example that we have used in the KNN classifier. Suppose we see a strange cat that also has some features of dogs, so if we want a model that can accurately identify whether it is a cat or dog, so such a model can be created by using the SVM algorithm. We will first train our model with lots of images of cats and dogs so that it can learn about different features of cats and dogs, and then we test it with this strange creature. So as support vector creates a decision boundary between these two data (cat and dog) and choose extreme cases (support vectors), it will see the extreme case of cat and dog. On the basis of the support vectors, it will classify it as a cat. Consider the below diagram:

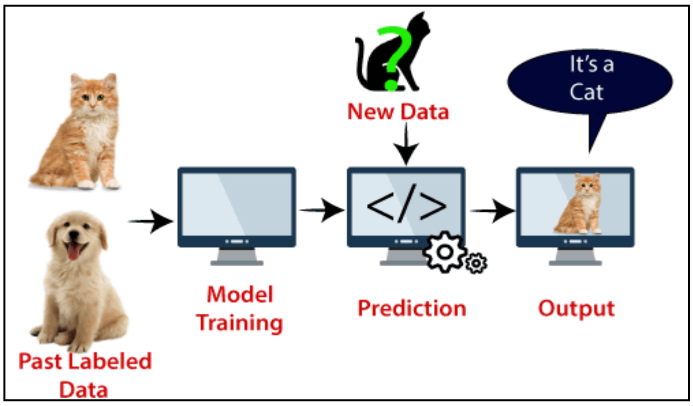


Fig 4.8 : SVM Example

SVM can be of two types:

Linear SVM: Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

Non-linear SVM: Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

4.5.1. Working

of SVM :-

Linear SVM:

The working of the SVM algorithm can be understood by using an example. Suppose we have a dataset that has two tags (green and blue), and the dataset has two features x1 and x2. We want a classifier that can classify the pair(x1, x2) of coordinates in either green or blue. Consider the below image:

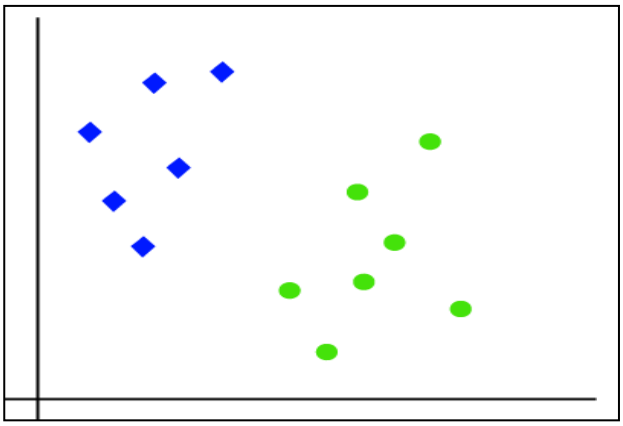
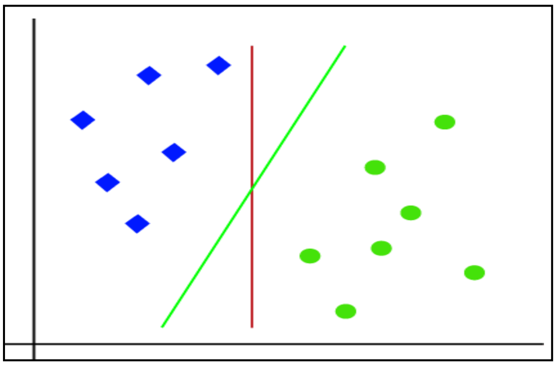


Fig 4.9 : Dataset Graph

So as it is 2-d space so by just using a straight line, we can easily separate these two classes. But there can be multiple lines that can separate these classes. Hence, the SVM algorithm helps to find the best line or decision boundary; this best boundary or region is called as a hyper plane. SVM algorithm finds the closest point of the lines from both the classes. Thesepoints are called support vectors. The distance between the vectors and the hyper plane is called as margin. And the goal of SVM is to maximize this margin. The hyper plane with maximum margin is called the optimal hyper plane. Consider the below images:



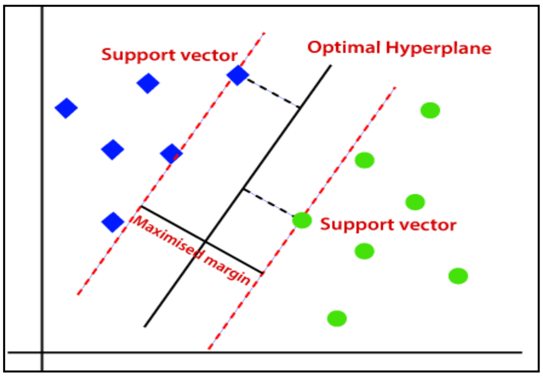


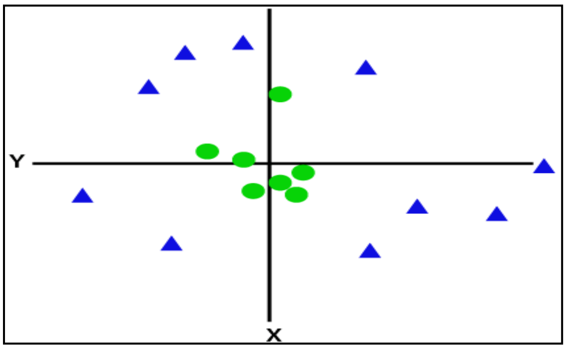
Fig 4.10 : Separate Dataset **Fig 4.11 : Linear Hyper plane**

Non-Linear SVM:

If data is linearly arranged, then we can separate it by using a straight line, but for non-linear data, we cannot draw a single straight line. So to separate these data points, we need to add one more dimension. For linear data, we have used two dimensions x and y, so for non-linear data, we will add a third dimension z. It can be calculated as:

z=x2 +y2

By adding the third dimension, the sample space will become as below image:



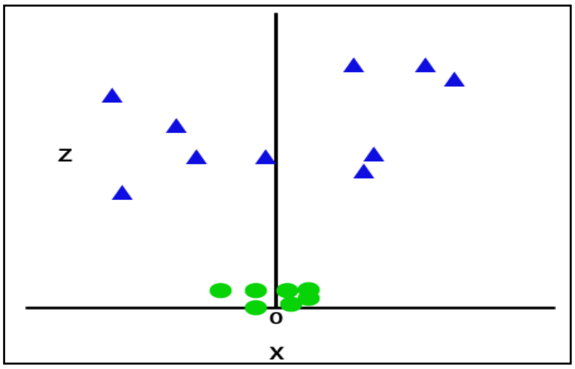
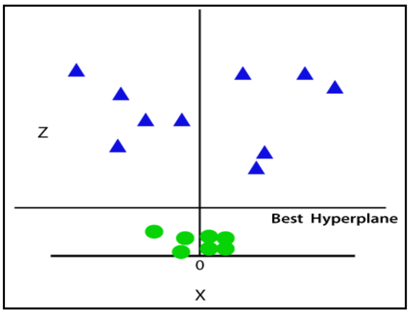


Fig 4.12 : Non-Linear Dataset

Fig 4.13 : Non-Linear Dataset

So now, SVM will divide the datasets into classes in the following way. Since we are in 3-dSpace, hence it is looking like a plane parallel to the x-axis. If we convert it in 2d space withz=1, then it will become like the below images:



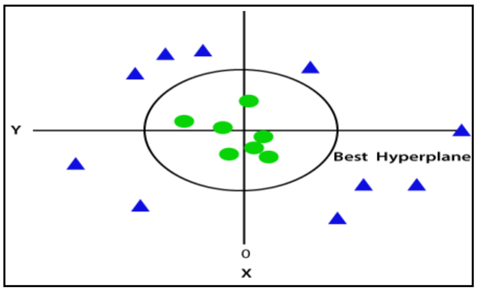


Fig 4.14 : Best Hyper plane(3-d)

Fig 4.15 : Best Hyper plane(2-d)

Hence we get a circumference of radius 1 in case of non-linear data.

4.6. <u>Decision Tree</u> :-

Decision Tree is a supervised learning method used in data mining for classification and regression methods. It is a tree that helps us in decision-making purposes. The decision tree creates classification or regression models as a tree structure. It separates a data set into smaller subsets, and at the same time, the decision tree is steadily developed. The final tree is a tree with the decision nodes and leaf nodes. A decision node has at least two branches. The leaf nodes show a classification or decision. We can't accomplish more split on leaf nodes-The uppermost decision node in a tree that relates to the best predictor called the root node. Decision trees can deal with both categorical and numerical data.

The decisions or the test are performed on the basis of features of the given dataset. 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 sub trees.

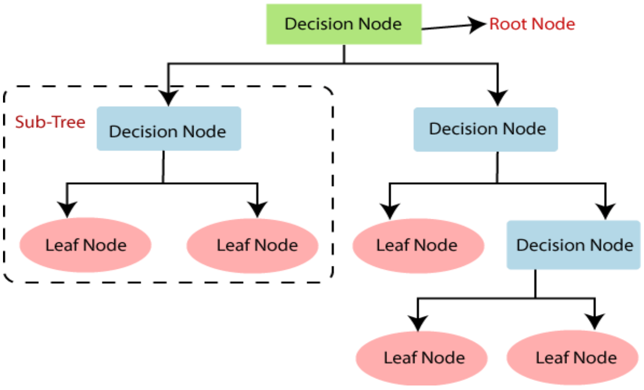


Fig 4.16 : General Structure of Decision Tree

4.6.1. <u>Working of Decision Tree Algorithm</u> :-

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.

For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

- **Step-1:** Begin the tree with the root node, says 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, which contains the best attribute.
- **Step-5:** Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

Example: 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, the decision tree starts with the root node (Salary attribute by ASM). The root node splits further into the next decision node(distance from the office) and one leaf node based on the corresponding labels. The next decision node further gets split into one decision node (Cab facility) and one leaf node. Finally, the decision node splits into two leaf nodes (Accepted offers and Declined offer). Consider the below diagram:

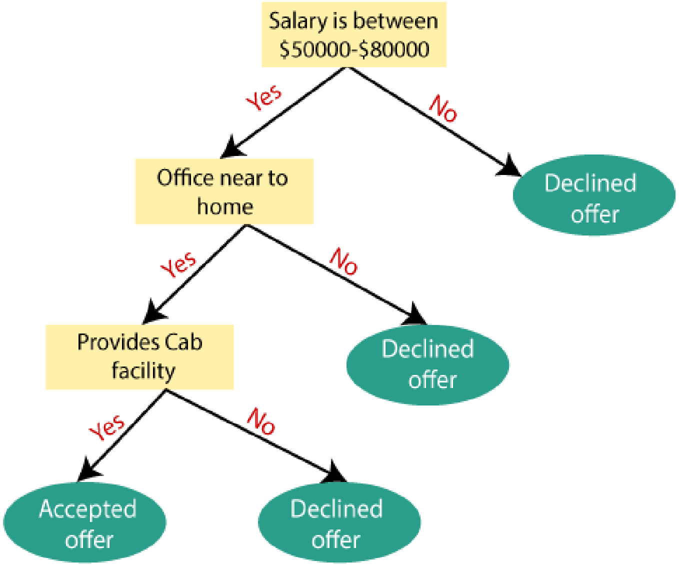


Fig 4.17 : Decision Tree Example

4.7. Naive Bayes Classifier :-

The Naive Bayes classifier separates data into different classes according to the Bayes' Theorem, along with the assumption that all the predictors are independent of one another. It assumes that aparticular feature in a class is not related to the presence of other features. It is mainly usedin text classification that includes a high-dimensional training dataset. Naive Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. Some popular examples of Naive Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. For example, you can consider a fruit to be a watermelon if it is green, round and has a 10-inch diameter. These features could depend on each other for their existence, but each one of them independently contributes to the probability that the fruit under consideration is a watermelon. That's why this classifier has the term 'Naive' in its name.

Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability. The formula for Bayes' theorem is given as:

$$P(C|O) = \frac{P(O|C)P(C)}{P(O)}$$

Where,

P(C|O) is Posterior probability: Probability of hypothesis C on the observed event O.

P(O|C) is Likelihood probability: Probability of the evidence given that the probability of ahypothesis is true.

P(C) is Prior Probability: Probability of hypothesis before observing

the evidence.P(O) is Marginal Probability: Probability of Evidence.

4.8. Random Forest :-

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. 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.

As the name suggests,

Random Forest is a classifier that contains a number of decision trees on various subsets ofthe 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 andbased on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem ofover fitting.

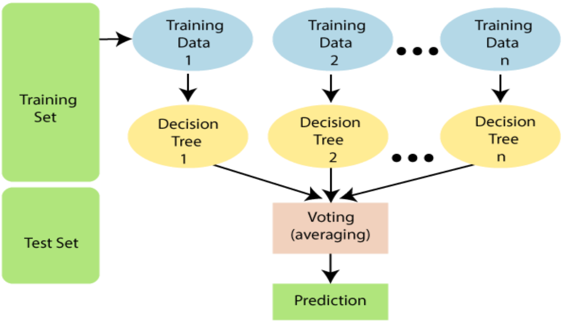


Fig 4.18 : Random Forest algorithm

4.8.1. Working of Random Forest Algorithm :-

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. The Working process can be explained in the below steps and diagram:

- **Step-1:** Select random K data points from the training set.

- **Step-2:** Build the decision trees associated with the selected data points (Subsets).
- **Step-3:** Choose the number N for decision trees that you want to build.
- **Step-4:** Repeat Step 1 & 2.
- **Step-5:** For new data points, find the predictions of each decision tree, and assign thenew data points to the category that wins the majority votes.

The working of the algorithm can be better understood by the below example:

Example: 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. Consider the below image:

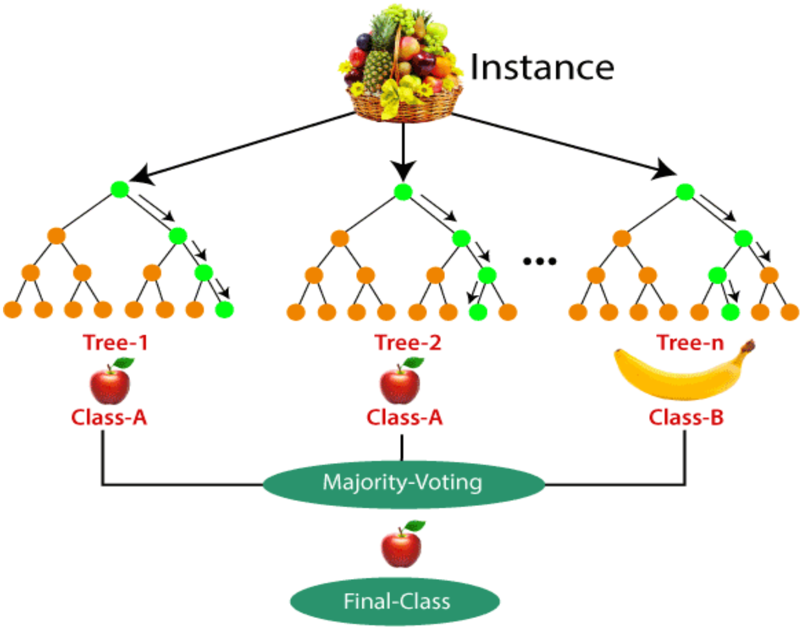


Fig 4.19 : Random Forest Example

5. Chapter 5 :- Unsupervised Machine Learning Algorithms

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

Some popular unsupervised learning algorithms are

- K-means clustering
- KNN (k-nearest neighbours)
- Hierarchal clustering
- Apriori algorithm

5.1. K-Means Clustering :-

K-Means Clustering is an Unsupervised machine learning algorithm, which groups the unlabelleddataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

It is an iterative algorithm that divides the unlabelled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The algorithm takes the unlabelled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K centre points or centroids by an iterative process.
- Assigns each data point to its closest k-centre. Those data points which are near to theparticular k-centre, create a cluster.

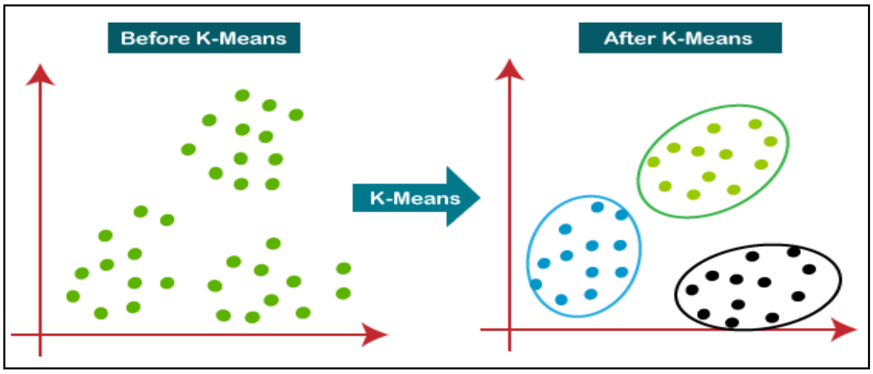


Fig 5.1 : K-Means Clustering

5.1.1. <u>Working of K-Means Algorithm</u> :-

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

- **Step-1:** Select the number K to decide the number of clusters.
- **Step-2:** Select random K points or centroids. (It can be other from the input dataset).
- **Step-3:** Assign each data point to their closest centroid, which will form thepredefined K clusters.
- **Step-4:** Calculate the variance and place a new centroid of each cluster.
- **Step-5:** Repeat the third steps, which means reassign each data point to the newclosest centroid of each cluster.
- **Step-6:** If any reassignment occurs, then go to step-4 else go to FINISH.
- **Step-7:** The model is ready.

<u>Let's understand the above steps by considering the visual plots:</u>

Suppose we have two variables M1 and M2. The x-y axis scatter plot of these two variables isgiven below:

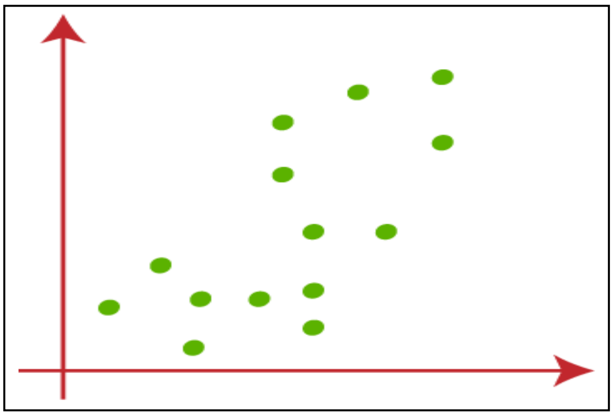


Fig 5.2 : Scatter Dataset

Let's take number k of clusters, i.e., K=2, to identify the dataset and to put them into differentclusters. It means here we will try to group these datasets into two different clusters.

We need to choose some random k points or centroid to form the cluster. These points can be either the points from the dataset or any other point. So, here we are selecting the below two points as k points, which are not the part of our dataset.

Now we will assign each data point of the scatter plot to its closest K-point or centroid. We will compute it by applying some mathematics that we have studied to calculate the distance between two points. So, we will draw a median between both the centroids. Consider the below images:

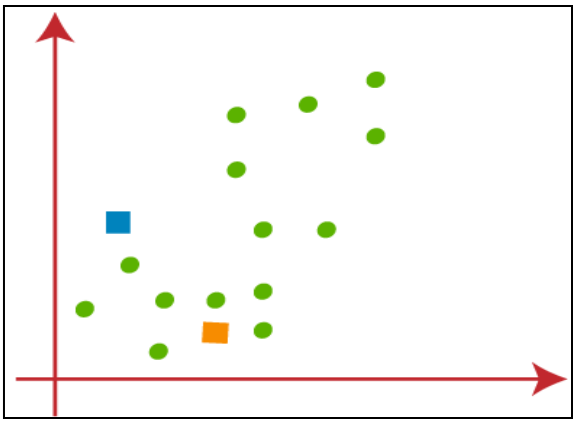


Fig 5.3 : Taking Centroid Points

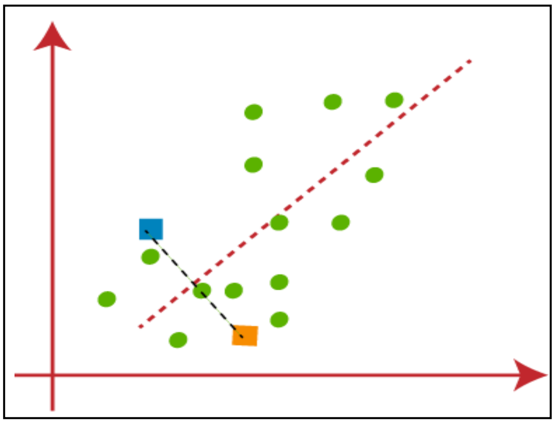


Fig 5.4 : Draw Median

From the above image, it is clear that points left side of the line is near to the K1 centroid, and points to the right of the line are close to the K2 centroid. Let's divide them for clear visualization.

As we need to find the closest cluster, so we will repeat the process by choosing a new centroid. To choose the new centroids, we will compute the centre of gravity of these centroids, and will find new centroids as below:

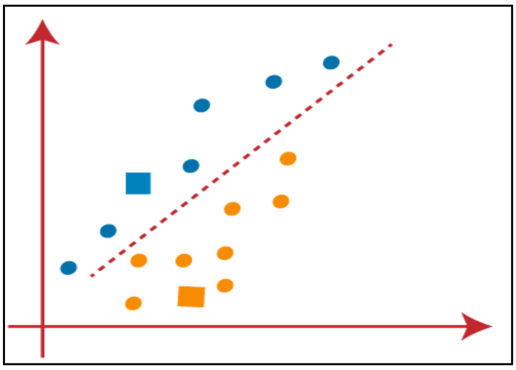


Fig 5.5 : Dividing clusters

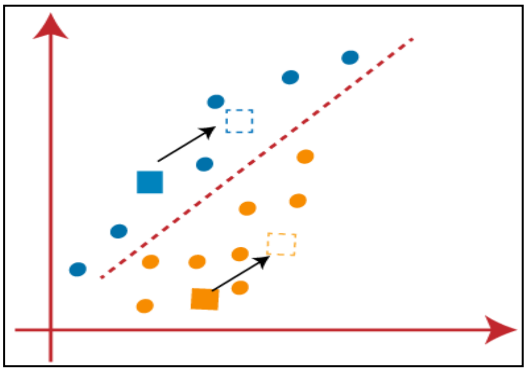
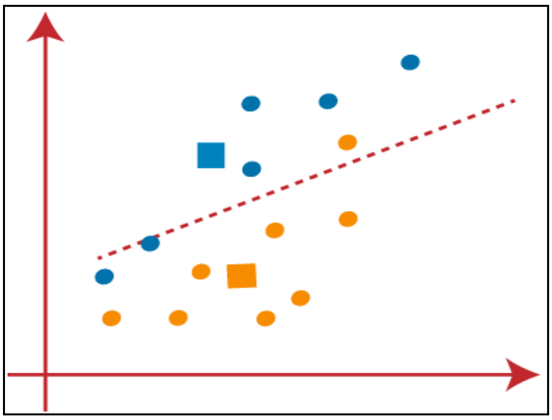


Fig 5.6 : New Centroids

Next, we will reassign each data point to the new centroid. For this, we will repeat the same process of finding a median line. We will repeat the process by finding the centre of gravity of centroids, so the new centroids will be as shown in the below image:



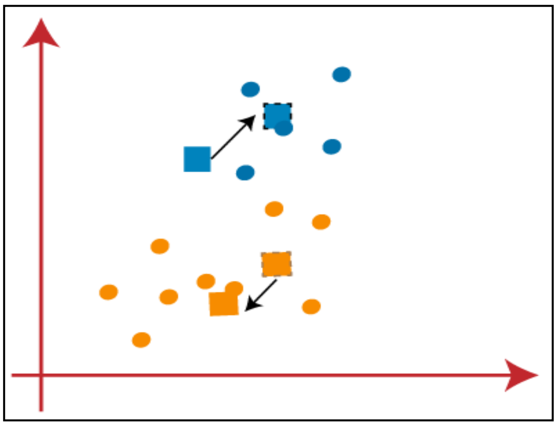
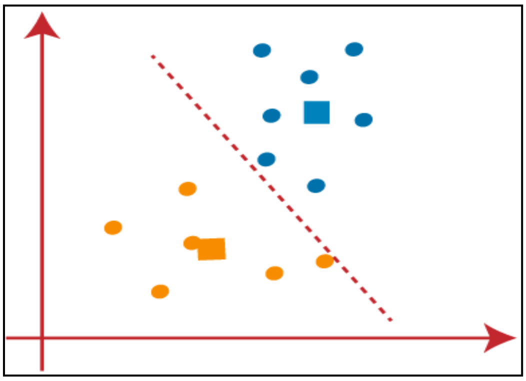


Fig 5.7 : New Centroid Median **Fig 5.8 : Again New Centroids**

As we got the new centroids so again will draw the median line and reassign the data points. So, We can see in image; there are no dissimilar data points on either side of the line, which means our model is formed. Consider the below image:



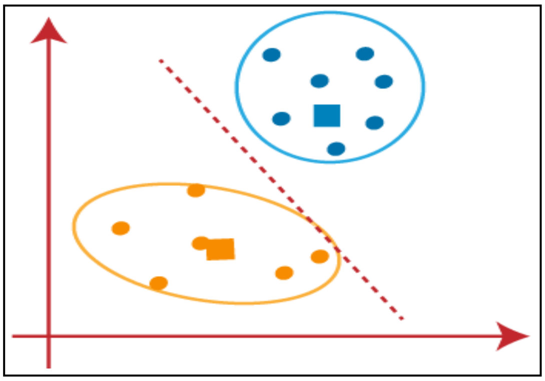


Fig 5.9 : Again Divide Points **Fig 5.10 : Final Modal**

As our model is ready, so we can now remove the assumed centroids, and the two final clusters will be as shown in the below image:

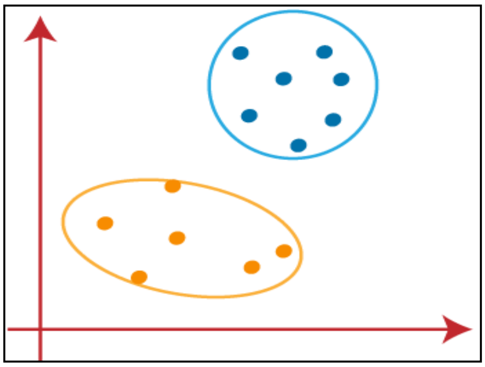


Fig 5.11 : Final Clusters

6. Chapter 6 :- **Project On Health Insurance Pricing Prediction In ML**

6.1. Introduction :-

The goal of this project is to allow a person to get an idea about the necessary amount required according to their own health status. Later they can comply with any health insurance company and their schemes and benefits keeping in mind the predicted amount. This can help a person in focusing more on the health aspect of an insurance rather than the futile part.

Health insurance is a must today and almost everyone has a connection with a public or private insurance company. The factors that determine the amount of insurance vary from company to company. Moreover rural residents are unaware that the Indian government provides free health insurance to people below the poverty line. This is a very complicated method and some villagerseither have private health insurance or don't invest health insurance at all. In addition it is be deceived about the insurance amount and you can purchase unnecessary expensive health insurance.

The project does not provide the exact amount required by health insurance companies, but provides a sufficient indication of amount associated with an individual for his health insurance.

It should not be the only criterion when choosing health insurance as the prognosis is premature and does not meet the requirements of any particular company. Estimating the amount of your health insurance plan early can help you better understand how much you will need.

6.2. Dataset Used :-

The main data source for this project was user Kaggle Dmarco. The data set consists of 1338 records with 6 attributes. Attributes are age, gender, body mass index, children, smokers, and rates. The data was saved as a csv file in a structured format. Many factors affect the amount of a health plan, such as previous physical condition, family medical history, body mass index (BMI),marital status, location, past coverage, and more.

According to the data set, age and smoking status have the greatest influence on sheep prediction,with smokers being the only attributes with the greatest effect. Since the child property had little effect on the prediction, this property was removed from the input to the regression model, providing a more accurate calculation in a shorter time.

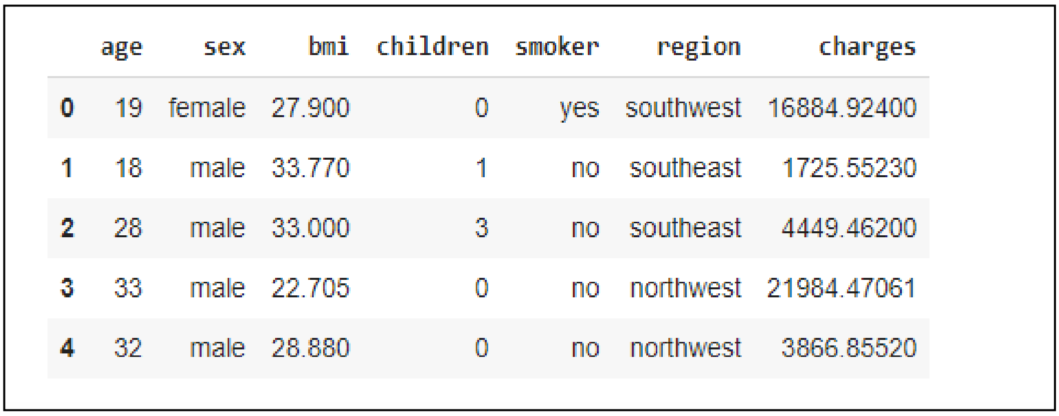
Dataset Link: Medical Cost Personal Datasets | Kaggle

6.3. Design and Implementation :-

6.3.1. Data Cleaning and Preparation :-

Data was taken from the kaggle website. There is a variety of data on the website, and the data used for the project is data on the amount of insurance. The data included various attributes such as age, gender, body mass index, smokers, and board attributes that served as labels for the project. The data was in a structured format and stored in csv file format.

```
import pandas as pd
insurance =
pd.read_csv("/insurance_dataset.csv")
insurance.head()
```

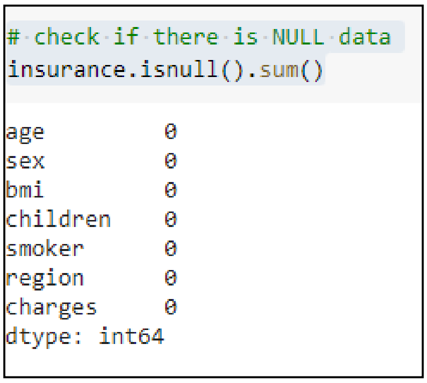


	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Fig 6.1 : Sample Dataset

Check if there is NULL data

```
# check if there is
NULL data
insurance.isnull().su
m()
```



```
# check if there is NULL data
insurance.isnull().sum()

age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

Fig 6.2 : Null Check

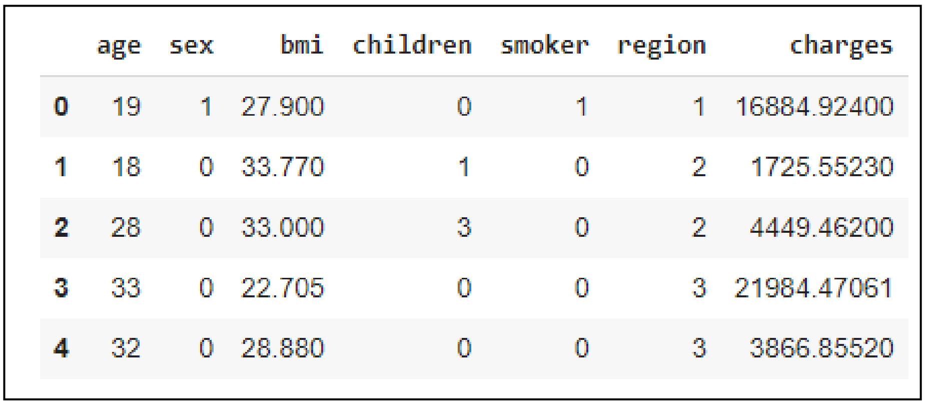
The dataset contained string values. Regression Algorithms seem to be working on features represented as numbers only. By looking at the dataset the columns ‘sex’, ‘smoker’ and ‘region’ are in string format. So they have to be converted into numerical values.

Sex column can be converted into numerical by replacing male with 0 and female

with 1.Smoker is converted by replacing Yes with 1 and no with 0

And region will be converted by setting southwest as 1 southeast as 2 northeast as 3 andnorthwest as 4.

```
# converting 'sex', 'smoker' and 'region' to numerical
 values as theregression algorithms only work with numbers
 so
# sex- male will be 0 and female
will be 1# smoker- if yes then 1
else 0
# region- southwest - 1 southeast - 2 northwest - 3
northeast - 4insurance['sex'] = insurance['sex'].apply({
      'male' : 0, 'female' : 1
}.get)
insurance['smoker'] =
      insurance['smoker'].apply({'yes' :
      1, 'no' : 0
}.get)
insurance['region'] =
      insurance['region'].apply({
      'southwest' : 1, 'southeast' : 2,
      'northwest' : 3, 'northeast' : 4,
}.get)
insuranc
e.head()
```



	age	sex	bmi	children	smoker	region	charges
0	19	1	27.900	0	1	1	16884.92400
1	18	0	33.770	1	0	2	1725.55230
2	28	0	33.000	3	0	2	4449.46200
3	33	0	22.705	0	0	3	21984.47061
4	32	0	28.880	0	0	3	3866.85520

Fig 6.3 : Sample Dataset after Preparation

6.3.2. <u>Finding Correlation Between Different Columns Of Dataset</u> :-

- Finding correlation between charges and age

```
import seaborn as sb
sb.jointplot(x=insurance['age'],y=insurance['ch
arges'])
```

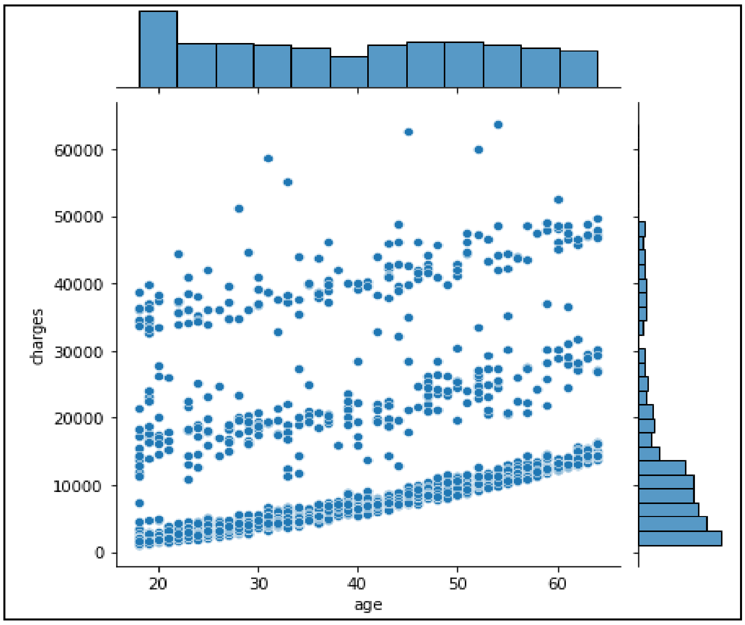


Fig 6.4 : Visualization of Charges and Age

Here we can see that health insurance premiums also increase with age.

- Finding correlation between charges and smoker

```
import seaborn as sb
sb.jointplot(x=insurance['smoker'],y=insurance['ch
arges'])
```

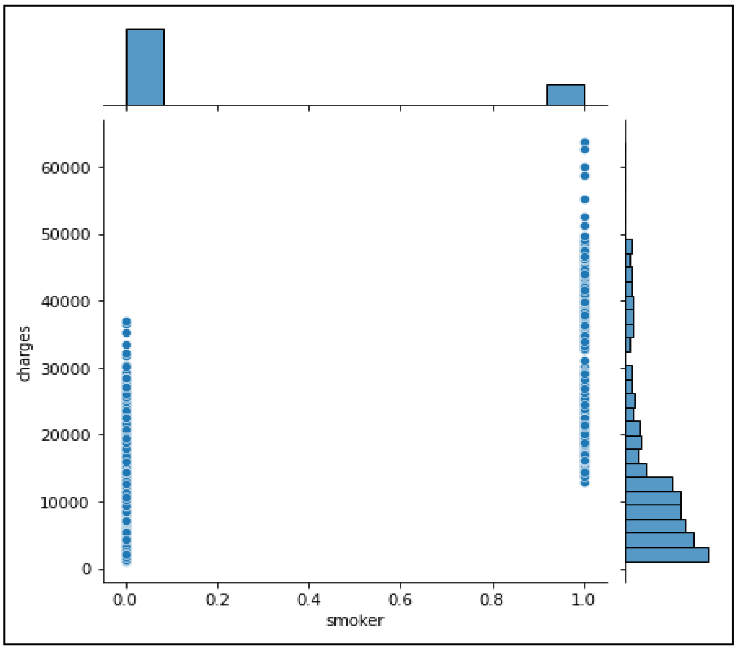


Fig 6.5 : Visualization of Charges and Smoker

Here we see that charges for smokers are higher than non-smokers**.**

6.4. <u>Training the dataset</u> :-

Once training data is in a suitable form to feed to the model, the training and testing phase of the model can proceed. During the training phase, the primary concern is the model selection. This involves choosing the best modelling approach for the task, or the best parameter settings for a given model. Multiple Linear Regression was selected as there are many components on which the charges are dependent like age, sex, bmi, smoker, region etc.

Step 1 - The Dataset is Split Into ‘X’ Array That Contains The Features And a ‘Y’ Array With The Target Variable.

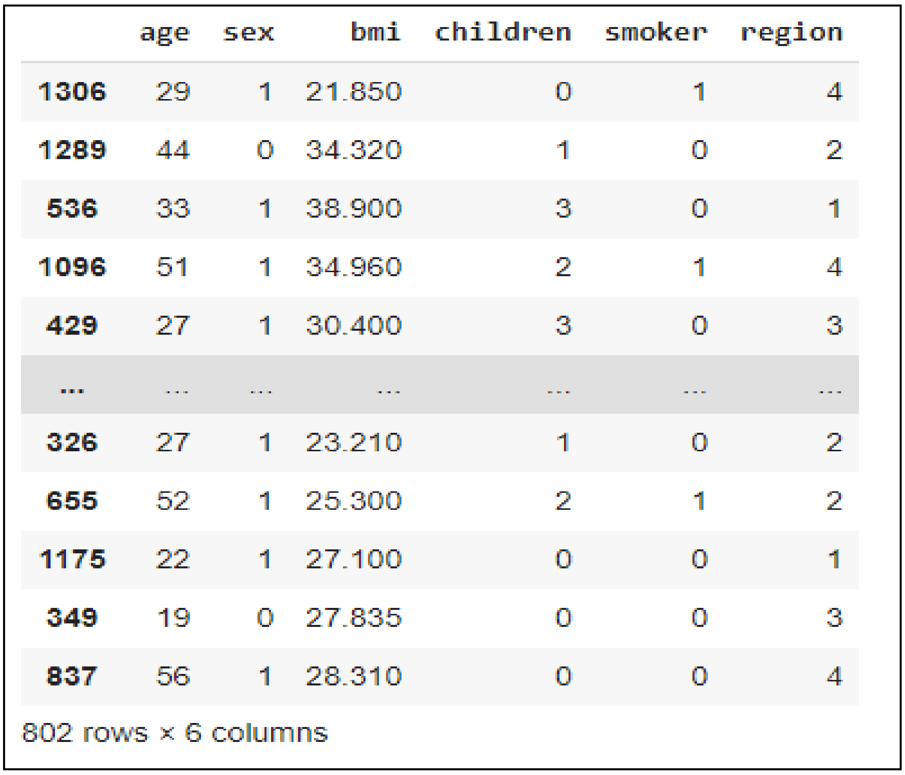
X contains age, sex, bmi, children, smoker and region and y contains charges (predicted value)

```
x =
insurance[['age','sex','bmi','children','smoker','regio
n']]y = insurance['charges']
```

Step 2 - Split The Dataset Into a Training And Testing Dataset

```
from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test =
train_test_split(x,y,test_size = 0.4)X_train
```

The dataset is divided into Training and testing dataset the testing dataset is ¼ th of the original dataset and training dataset is ¾th of the original dataset.



	age	sex	bmi	children	smoker	region
1306	29	1	21.850	0	1	4
1289	44	0	34.320	1	0	2
536	33	1	38.900	3	0	1
1096	51	1	34.960	2	1	4
429	27	1	30.400	3	0	3
...	...	...	...	...	...	...
326	27	1	23.210	1	0	2
655	52	1	25.300	2	1	2
1175	22	1	27.100	0	0	1
349	19	0	27.835	0	0	3
837	56	1	28.310	0	0	4

802 rows × 6 columns

Fig 6.6 : Dividing Dataset For Testing

and TrainingStep 3 - Training And Testing The Model

```
from sklearn.linear_model import
LinearRegressionmodel =
LinearRegression()
model.fit(X_train,Y_train)
```

Step 4 - Prediction From The

Model Here we pas ‘X_test’ to

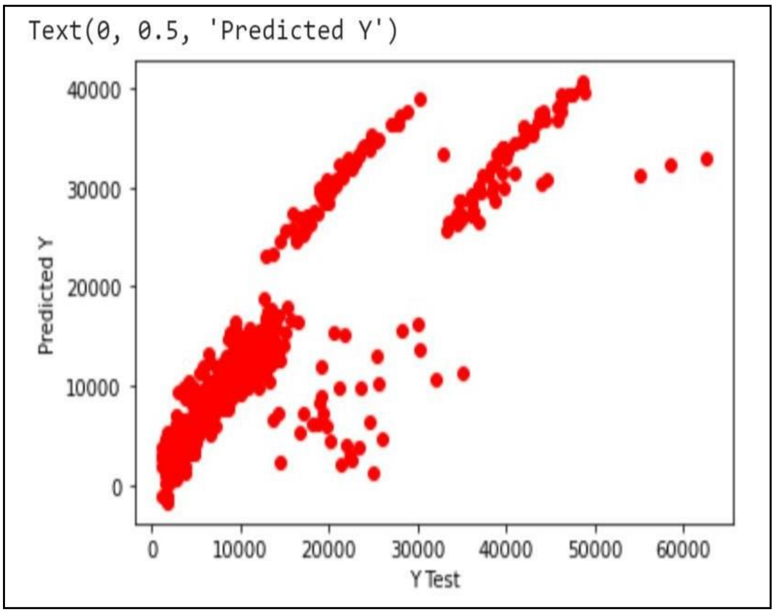
predict ‘Y_test’

```
predictions =
model.predict(X_test)
predictions[0:5]
```

Step 5 - Comparing The Results

Lets compare these predictions with actual results by plotting a graph

```
import matplotlib.pyplot as plt
plt.scatter(Y_test,predictions,
c="red")plt.xlabel('Y Test')
plt.ylabel('Predicted Y')
```



```
Text(0, 0.5, 'Predicted Y')
```

Fig 6.7 :

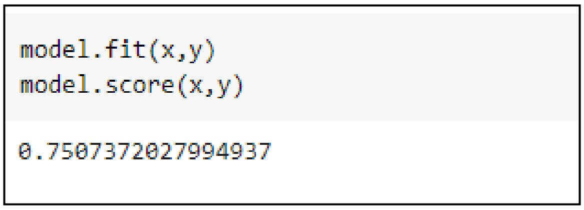
Comparing ResultsStep 6 – Finding the regression

model

Finding the regression value or R-square value:

```
model.fit
(x,y)
```

```
model.sco
re(x,y)
```



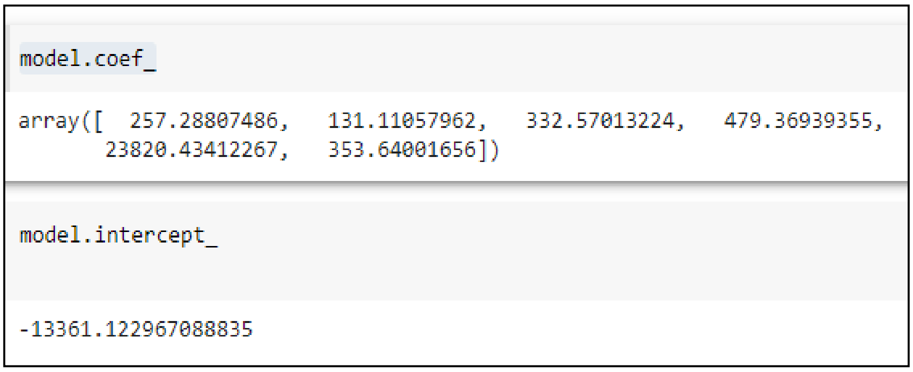
```
model.fit(x,y)
model.score(x,y)
```

```
0.7507372027994937
```

Fig 6.8 : R-square value

The R-square value obtained is 0.7507372027994937 which means that from the five independent variables namely age, sex, bmi, smoker, region and children it affects the dependent variable (charges) of 75%. The smaller the R-square value the weaker the influence of the independent variable on the dependent variable and vice versa.

Coefficient and intercept values



```
model.coef_
```

```
array([  257.28807486,   131.11057962,   332.57013224,   479.36939355,
        23820.43412267,   353.64001656])
```

```
model.intercept_
```

```
-13361.122967088835
```

Fig 6.9 : Coefficient and Intercept

Based on the output, the intercept value is -13361.122967088835. and the coefficients are257.2880786, 131.11057962, 332.57013224, 479.36939955, 23820.43412267, 353.64001656.

The regression model can be written as

y = -13361.122967088835 + 257.2880786X1 + 131.11057962X2 + 332.57013224X3 +479.36939955X4 + 23820.43412267X5 + 353.64001656X6

6.5. <u>Prediction</u> :-

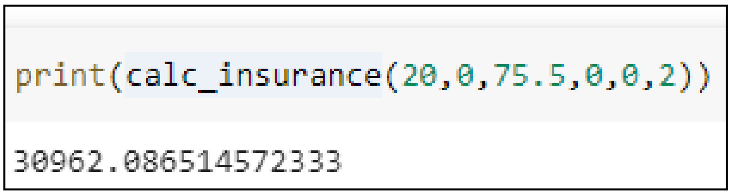
After getting the regression model, prediction can be made using the

regression model.Function for returning the insurance cost :

```
def calc_insurance(age,sex,bmi,smoker,children,region):
y = ((age * model.coef_[0]) + (sex * model.coef_[1]) + (bmi *
model.coe f_[2]) + (smoker * model.coef_[3]) + (children *
model.coef_[4]) + (region * model.coef_[5]))
return y
```

This function for the input 20 as age 0- male for sex 75.5 for bmi, 0 for smoker (no) 0 for numberof children and 2 as region (northeast) will return

```
print(calc_insurance(20,0,75.5,0,0,2))
```



```
print(calc_insurance(20,0,75.5,0,0,2))

30962.086514572333
```

Fig 6.10 : Output Prediction

7. Chapter 7 :- Conclusion

Health insurance is a must today and almost everyone has a connection with a public or private insurance company. The factors that determine the amount of insurance vary from company to company. What's more, rural residents are unaware that the Indian government provides free health insurance to people below the poverty line. This is a very complicated method, and some villagers either have private health insurance or don't invest in health insurance at all. In addition, it is easy to be deceived about the insurance amount, and you can purchase unnecessary expensive health insurance.

Our project does not give health insurance companies the exact amount they need, but does provide asufficient indication of the amount that is relevant to individuals for their health insurance.

Because the prognosis is premature and does not meet the requirements of a specific company, it should not be the sole criterion for health insurance selection. Estimating the amount of your health insurance plan early can help you better understand how much you will need. A place where a personcan be sure that the amount they are going to choose is justified. It can also give you insight into howto get extra benefits from your health insurance.

Premium forecasting focuses on your own health and not on other companies' insurance terms. Thesemodels can be applied to data collected over the next several years to predict insurance premiums. This can help people as well as insurance companies work together to improve health-focused insurance.

8. References

- **https://www.kaggle.com/mirichoi0218/insurance**
- **10 Factors That Affect Your Health Insurance Premium Costs(iffcotokio.co.in)**
- **https://en.wikipedia.org/wiki/Healthcare_in_India**
- **https://www.javatpoint.com/machine-learning**