1.INTRODUCTION

Data science & Business analytics enables businesses to process huge amounts of structured and unstructured big data to detect patterns. This in turn allows companies to increase efficiencies, manage costs, identify new market opportunities, and boost their market advantage. Asking a personal assistant like Alexa or Siri for a recommendation demands data science. So does operating a self driving car, using a search engine that provides useful results, or talking to a chatbot for customer service. These are all real-life applications for data science. Data science is the practice of mining large data sets of raw data, both structured and unstructured, to identify patterns and extract actionable insight from them. This is an interdisciplinary field, and the foundations of data science include statistics, inference, computer science, predictive analytics, machine learning algorithm development, and new technologies to gain insights from big data.

To define data science and improve data science project management, start with its life cycle. The first stage in the data science pipeline workflow involves capture: acquiring data, sometimes extracting it, and entering it into the system. The next stage is maintenance, which includes data warehousing, data cleansing, data processing, data staging, and data architecture.Data processing follows, and constitutes one of the data science fundamentals. It is during data exploration and processing that data scientists stand apart from data engineers. This stage involves data mining, data classification and clustering, data modeling, and summarizing insights gleaned from the data—the processes that create effective data.Next comes data analysis, an equally critical stage. Here data scientists conduct exploratory and confirmatory work, regression, predictive analysis, qualitative analysis, and text mining. This stage is why there is no such thing as cookie cutter data science—when it’s done properly.During the final stage, the data scientist communicates insights. This involves data visualization, data reporting, the use of various business intelligence tools, and assisting businesses, policymakers, and others in smarter decision making.

Both data science and business analytics focus on solving business problems, and both involve collecting data, modeling it, and then gleaning insights from the data. The main difference is that business analytics is specific to business-related problems such as profit and costs.In contrast, data science methods explore how a wide range of factors—anything from customer preferences to the weather—might affect a business. Data science combines data with technology and algorithm building to answer many questions. Business analytics is a narrower field, analyzing data from the business itself with statistical traditional theory to generate insights and business solutions. Learn more about Customer Analytics. By 2020, there will be around 40 zettabytes of data—that's 40 trillion gigabytes. The amount of data that exists grows exponentially. At any time, about 90 percent of this huge amount of data gets generated in the most recent two years, according to sources like IBM and SINTEF.In fact, internet users generate about 2.5 quintillion bytes of data every day. By 2020, every person on Earth will be generating about 146,880 GB of data every day, and by 2025, that will be 165 zettabytes every year.This means there is a huge amount of work in data science—much left to uncover. According to The Guardian, in 2012 only about 0.5 percent of all data was analyzed.

2. REQUIREMENTS SPECIFICATION

2.1 SOFTWARE REQUIREME

PYTHON 3.5.0

It is an interpreted high-level programming language for general purpose programming. Python has a design philosophy that emphasizes code readability and a syntax that allows programmers to express concepts in fewer lines of code, notably using significant white space. It provides constructs that enable clear programming on both small and large scales. Python features a dynamic type system and automatic memory management. It supports multiple programming, paradigms, including object oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

Python Interpreters are available for many OS. Mostly Python implementation includes a read-eval-print loop, permitting to function as a command line interpreter for which the user enters statements sequentially and receives results immediately.

Some things that Python is often used for are:

- Web development.
- Scientific programming.
- Desktop GUIs.
- Network programming.

Fig 2.3 Python

This is a small example of a Python program. It shows "Hello World!" on the screen.

```
print("Hello World!")
# This code does the same thing, only it is longer: ready = True
if ready:
print ("Hello World!")
```

3.TASKS

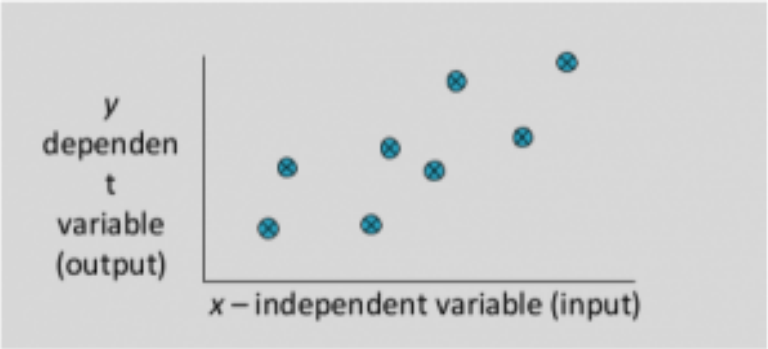
3.1 PREDICTION USING SUPERVISED MACHINE LEARNING

Data scientists use many different kinds of machine learning algorithms to discover patterns in big data that lead to actionable insights. At a high level, these different algorithms can be classified into two groups based on the way they "learn" about data to make predictions: supervised and unsupervised learning.

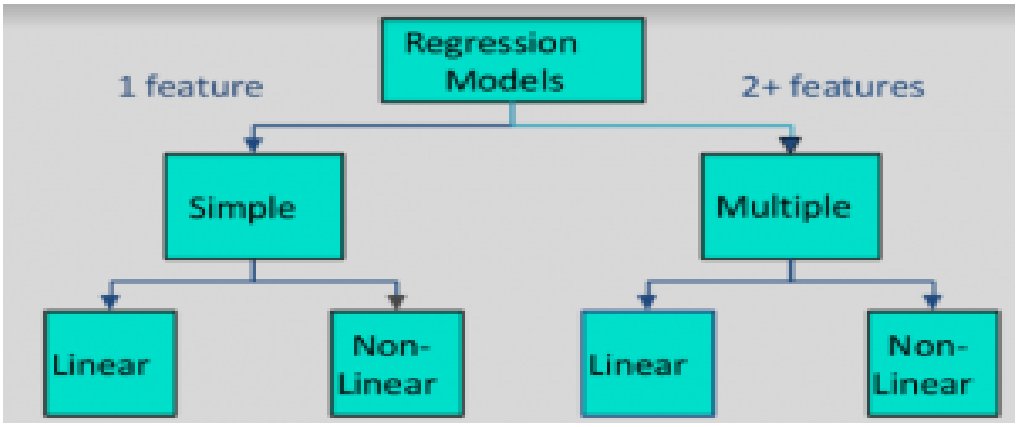
Supervised Machine Learning: The majority of practical machine learning uses supervised learning. Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output Y = f(X) . The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.Techniques of Supervised Machine Learning algorithms include linear and logistic regression**,** multi- class classification**,** Decision Trees and support vector machines. Supervised learning requires that the data used to train the algorithm is already labeled with correct answers. For example, a classification algorithm will learn to identify animals after being trained on a dataset of images that are properly labeled with the species of the animal and some identifying characteristics. Supervised learning problems can be further grouped into Regression and Classification problems. Both problems have as goal the construction of a succinct model that can predict the value of the dependent attribute from the attribute variables. The difference between the two tasks is the fact that the dependent attribute is numerical for regression and categorical for classification.

Regression

A regression problem is when the output variable is a real or continuous value, such as "salary" or "weight". Many different models can be used, the simplest is the linear regression. It tries to fit data with the best hyper-plane which goes through the points.



Regression Types



For Examples:

- Predicting age of a person
- Predicting nationality of a person
- Predicting whether stock price of a company will increase tomorrow
- Predicting whether a document is related to sighting of UFOs

Solution **:** Predicting age of a person (because it is a real value, predicting nationality is categorical, whether stock price will increase is discrete-yes/no answer, predicting whether a document is related to UFO is again discrete- a yes/no answer). Let's take an example of linear regression. We have a Housing data set and we want to predict the price of the house.

Classification

A classification problem is when the output variable is a category, such as "red" or "blue" or "disease" and "no disease". A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes.
For example, when filtering emails "spam" or "not spam", when looking at transaction data, "fraudulent", or "authorized". In short Classification either predicts categorical class labels or classifies data (construct a model) based on the training set and the values (class labels) in classifying attributes and uses it in classifying new data. There are a number of classification models. Classification models include logistic regression, decision tree, random forest, gradient-boosted tree, multilayer perceptron, one-vs-rest, and Naive Bayes.

- Predicting the gender of a person by his/her handwriting style
- Predicting house price based on area
- Predicting whether monsoon will be normal next year

- Predict the number of copies a music album will be sold next month

Predicting the gender of a person Predicting whether monsoon will be normal next year. The other two are regression. As we discussed classification with some examples. Now there is an example of classification in which we are performing classification on the iris dataset using RandomForestClassifier in python

Classification is of two types:

- **Binary Classification**: When we have to categorize given data into 2 distinct classes. Example – On the basis of given health conditions of a person, we have to determine whether the person has a certain disease or not.
- **Multiclass Classification**: The number of classes is more than 2. For Example – On the basis of data about different species of flowers, we have to determine which specie does our observation belongs to.

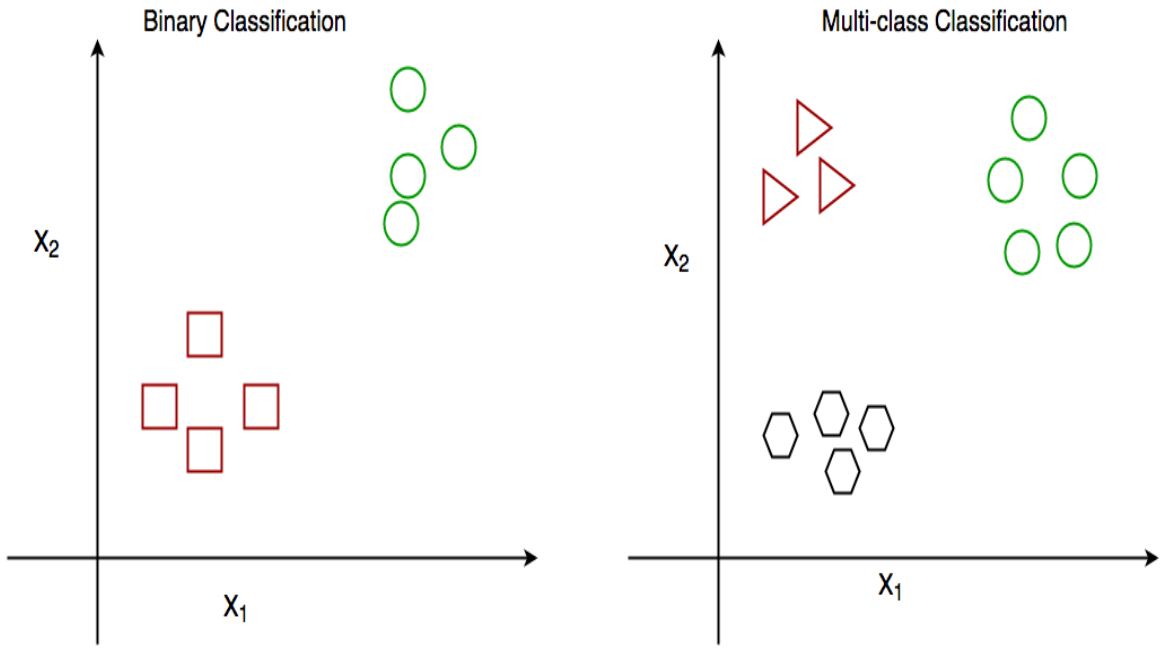
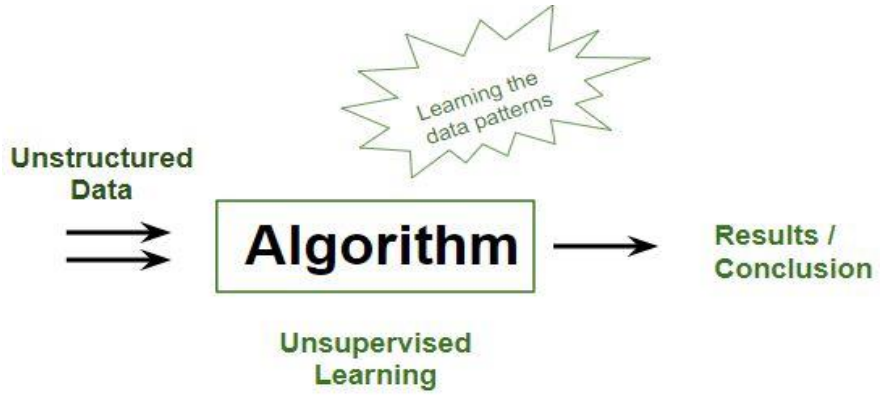


Fig: Binary and Multiclass Classification.

classification algorithms with gender classification using hair length (by no means am I trying to stereotype by gender, this is only an example). To classify gender (target class) using hair length as feature parameter we could train a model using any classification algorithms to come up with some set of boundary conditions which can be used to differentiate the male and female genders using hair length as the training feature. In gender classification case the boundary condition could the proper hair length value.

3.2 PREDICTION USING UNSUPERVISED MACHINE LEARNING



It's a type of learning where we don't give a target to our model while training i.e. training model has only input parameter values. The model by itself has to find which way it can learn. Data-set in Figure A is mall data that contains information of its clients that subscribe to them. Once subscribed they are provided a membership card and so the mall has complete information about the customer and his/her every purchase. Now using this data and unsupervised learning techniques, the mall can easily group clients based on the parameters we are feeding in.

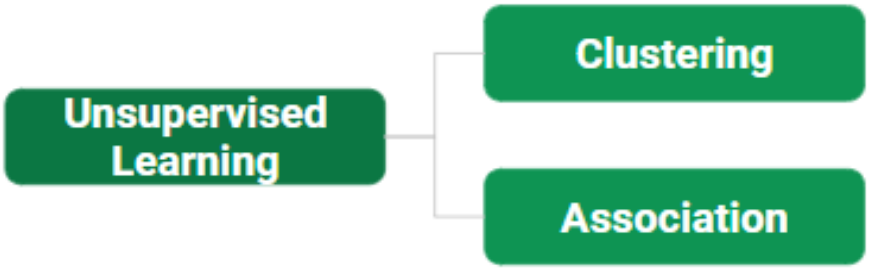
CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72
11	Male	67	19	14
12	Female	35	19	99
13	Female	58	20	15
14	Female	24	20	77
15	Male	37	20	13
16	Male	22	20	79
17	Female	35	21	35

Figure A

Training data we are feeding is

- **Unstructured data**: May contain noisy(meaningless) data, missing values, or unknown data

- **Unlabeled data**: Data only contains a value for input parameters, there is no targeted value(output). It is easy to collect as compared to labelled one in the Supervised approach.



TypesofUnsupervisedLearning

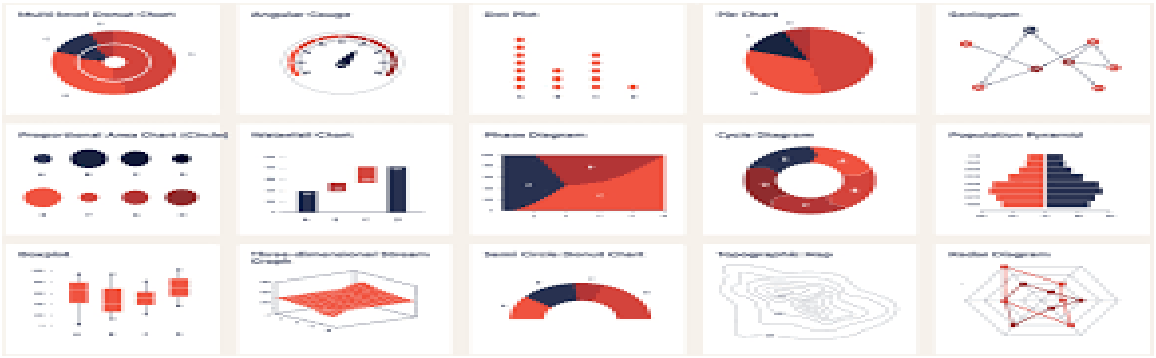
- **Clustering:** Broadly this technique is applied to group data based on different patterns, our machine model finds. For example, in the above figure, we are not given an output parameter value, so this technique will be used to group clients based on the input parameters provided by our data.
- **Association:** This technique is a rule-based ML technique that finds out some very useful relations between parameters of a large data set. For e.g. shopping stores use algorithms based on this technique to find out the relationship between the sale of one product w.r.t to others sales based on customer behavior. Once trained well, such models can be used to increase their sales by planning different offers.

Some algorithms:

- K-Means Clustering
- DBSCAN – Density-Based Spatial Clustering of Applications with Noise
- BIRCH – Balanced Iterative Reducing and Clustering using Hierarchies
- Hierarchical Clustering

3.3 EXPLOTORY DATA ANALYSIS

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including machine learning.



There are four primary types of EDA:

Univariate non-graphical. This is simplest form of data analysis, where the data being analyzed consists of just one variable. Since it’s a single variable, it doesn’t dealwith causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.

Univariate graphical: Non-graphical methods don't provide a full picture of the data. Graphical methods are therefore required. Common types of univariate graphics include:Stem-and-leaf plots, which show all data values and the shape of the distribution.Histograms, a bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values.Box plots, which graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum.

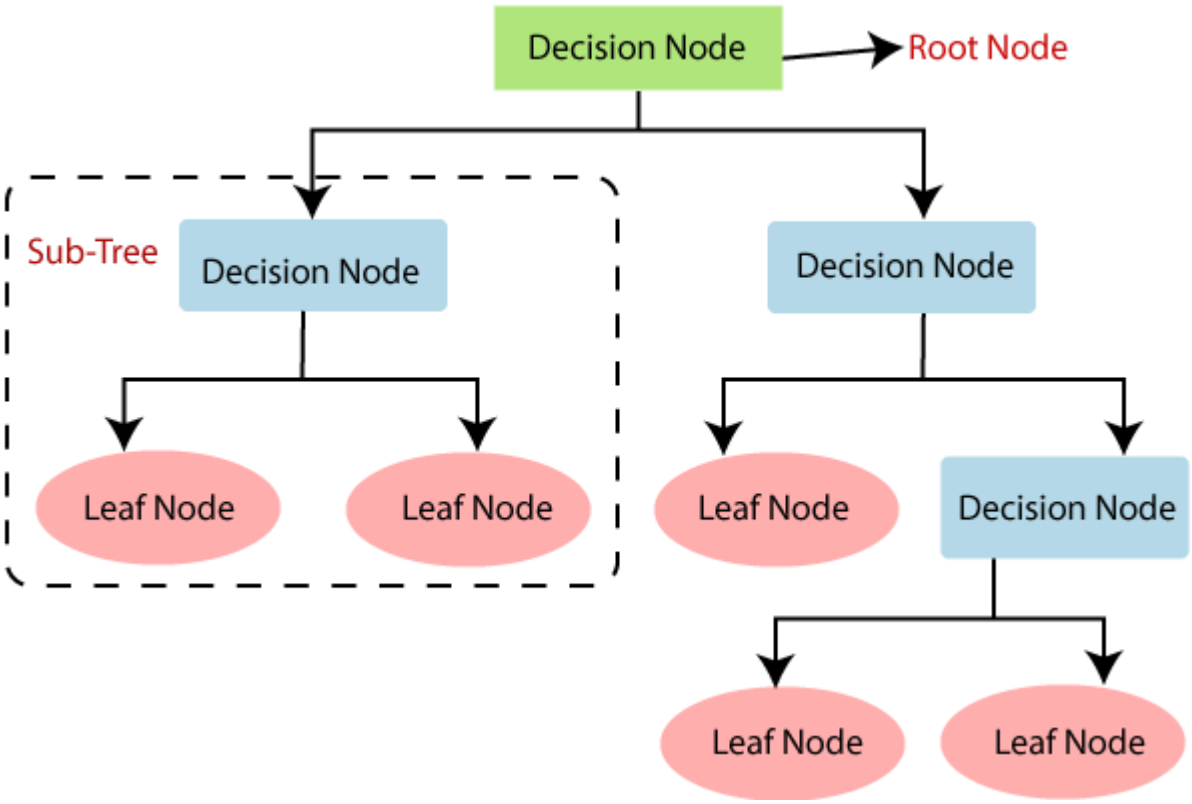
Multivariate nongraphical: Multivariate data arises from more than one variable. Multivariate non-graphical EDA techniques generally show the relationship between two or more variables of the data through cross-tabulation

Multivariate graphical: Multivariate data uses graphics to display relationships between two or more sets of data. The most used graphic is a grouped bar plot or bar chart with each group representing one level of one of the variables and each bar within a group representing the levels of the other Specific statistical functions and techniques you can perform with EDA tools include:

- Clustering and dimension reduction techniques, which help create graphical displays of high-dimensional data containing many variables.
- Univariate visualization of each field in the raw dataset, with summary statistics.
- Bivariate visualizations and summary statistics that allow you to assess the relationship between each variable in the dataset and the target variable you're looking at.
- Multivariate visualizations, for mapping and understanding interactions between different fields in the data.
- K-means Clustering is a clustering method in unsupervised learning where data points are assigned into K groups, i.e. the number of clusters, based on the distance from each group's centroid. The data points closest to a particular centroid will be clustered under the same category. K-means Clustering is commonly used in market segmentation, pattern recognition, and image compression.
- Predictive models, such as linear regression, use statistics and data to predict outcomes.

3.4 PREDICTION USING DECISION TREES

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.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.It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.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 subtrees.



There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine

learning model. Below are the two reasons for using the Decision treeDecision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.The logic behind the decision tree can be easily understood because it shows a tree-like structure.

Decision Tree Terminologies

Root Node: Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.

Leaf Node: Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.

Splitting: Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.

Branch/Sub Tree: A tree formed by splitting the tree.

Pruning: Pruning is the process of removing the unwanted branches from the tree.

Parent/Child node: The root node of the tree is called the parent node, and other nodes are called the child nodes.

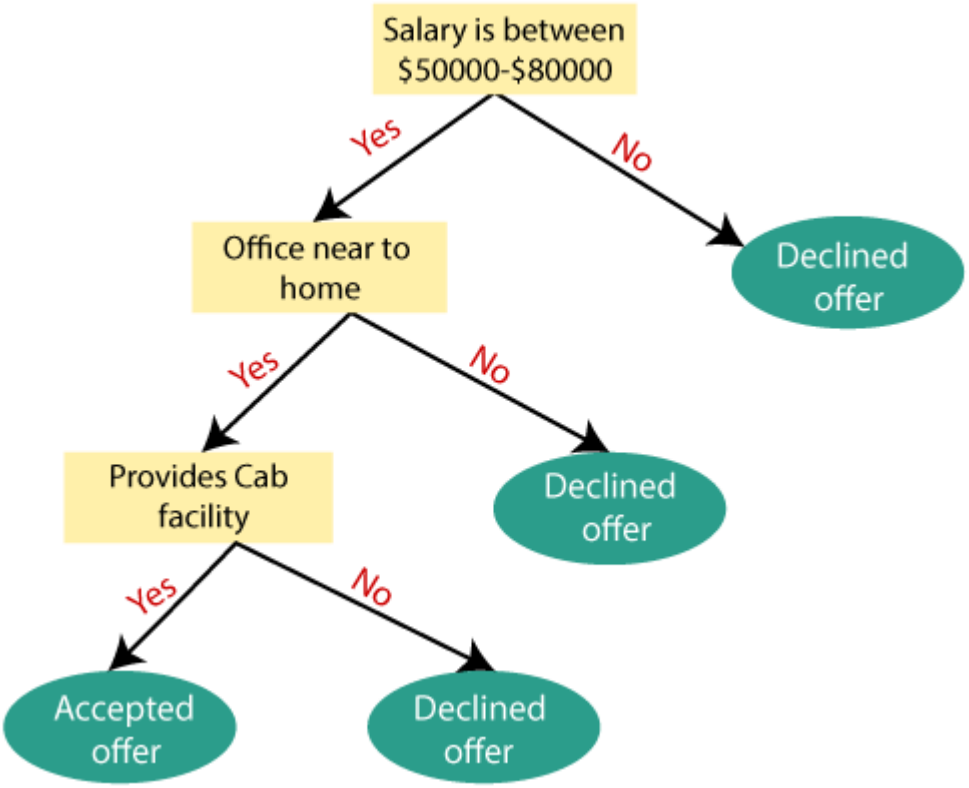
How does the Decision Tree algorithm Work

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:



Attribute Selection MeasuresWhile implementing a Decision tree, the main issue arises that how to select the best attribute for the root node and for sub-nodes. So, to solve such problems there is a technique which is called as Attribute selection measure or ASM**.** By this measurement, we can easily select the best attribute for the nodes of the tree. There are two popular techniques for ASM.

1. Information Gain

- Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.
- It calculates how much information a feature provides us about a class.
- According to the value of information gain, we split the node and build the decision tree.
- A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula:

Information Gain= Entropy(S)- [(Weighted Avg) *Entropy(each feature)

Entropy: Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)

Where,

S= Total number of samples P(yes)= probability of yes P(no)= probability of no

2. Gini Index

- Gini index is a measure of impurity or purity used while creating a decision tree in the CART(Classification and Regression Tree) algorithm.
- An attribute with the low Gini index should be preferred as compared to the high Gini index.
- It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits.
- Gini index can be calculated using the below formula:

Gini Index= $1- \sum_j P_j^2$

Pruning: Pruning is a process of deleting the unnecessary nodes from a tree in order to get the optimal decision tree.A too-large tree increases the risk of overfitting, and a small tree may not capture all the important features of the dataset. Therefore, a technique that decreases the size of the learning tree without reducing accuracy is known as Pruning.

3.5 TIME LINE ANALYSIS

Time series analysis is a statistical technique dealing in time series data, or trend analysis.A time-series contains sequential data points mapped at a certain successive time duration, it incorporates the methods that attempt to surmise a time series in terms of understanding either the underlying concept of the data points in the time series or suggesting or making predictions. Forecasting data using time-series analysis comprises the use of some significant model to forecast future conclusions on the basis of known past outcomes.An objective of time series analysis is to explore and understand patterns in changes over time where these patterns signifies the components of a time series including trends, cycles, and irregular movements. When such components reside in a time series, the data model must be considered for these patterns for generating accurate forecasts, such as future sales, GDP, and global temperatures.Consider an example of a restaurant in which prediction is made on the number of customers as when will more customers appear in the restaurant at a specified time duration based on the previous appearance of customers with time. We can use Time Series for multiple investigations to predict future as circadian rhythms, seasonal behaviours, trends, changes, etc. to interrogate the questions like predicted values, what is leading and lagging behind, connections and association, control, repetitions, and hidden pattern, etc.

Time series analysis is basically the recording of data at a regular interval of time, which could lead to taking a versed decision, crucial for trade and so have multiple applications such as Stock Market and Trends Analysis, Financial Analysis and forecasting, Inventory analysis, Census Analysis, Yield prediction, Sales forecasting, etc.

Multiple applications of the Time-Series Analysis

Broadly specified time-series models are Autoregressive (AR), Integrated (I), Moving Average(MA), and some other models are the combination of these models such as Autoregressive Moving Average (ARMA), and Autoregressive Integrated Moving Average (ARIMA) models.These models reflect measurements near concurrently in time will be more closely relevant as compared to measurements distant apart.

Examples of Time-Series Analysis

Consider an example In the financial domain, the main objective is to recognize trends, seasonal behaviour, and correlation through the usage of time series analysis technique and producing filters based on the forecasts, this includes;

1. To predict expected utilities- For the perfect and successively trading, it is necessary to have accurate and reliable future predictions such as asset prices, variation in usage, products in demand in statistical form through market research, and time-series dataset.
2. Simulate series- After getting statistical output data of financial time series, that can be used for creating simulations of future events. It helps us to determine the count of trades, expected trading costs and returns, required financial and technical investment, several risks in trading, etc.
3. Presume relationship- Recognition of the relationship between the time series and other quantities gives us trading signs to improve the existing fashion of trading. For example, to know the spreading of foreign exchange pair and its variation with a proposal, estimated trades can be inferred for a certain period for forecasting a widespread to reduce transaction costs.

4. CONCLUSION

data science education is well into its formative stages of development; it is evolving into a self-supporting discipline and producing professionals with distinct and complementary skills relative to professionals in the computer, information, and statistical sciences. However, regardless of its potential eventual disciplinary status, the evidence points to robust growth of data science education that will indelibly shape the undergraduate students of the future. In fact, fueled by growing student interest and industry demand, data science education will likely become a staple of the undergraduate experience. There will be an increase in the number of students majoring, minoring, earning certificates, or just taking courses in data science as the value of data skills becomes even more widely recognized. The adoption of a general education requirement in data science for all undergraduates will endow future generations of students with the basic understanding of data science that they need to become responsible citizens. Continuing education programs such as data science boot camps, career accelerators, summer schools, and incubators will provide another stream of talent. This constitutes the emerging watershed of data science education that feeds multiple streams of generalists and specialists in society; citizens are empowered by their basic skills to examine, interpret, and draw value from data.Today, the nation is in the formative phase of data science education, where educational organizations are pioneering their own programs, each with different approaches to depth, breadth, and curricular emphasis (e.g., business, computer science, engineering, information science, mathematics, social science, or statistics). It is too early to expect consensus to emerge on certain best practices of data science education. However, it is not too early to envision the possible forms that such practices might take. Nor is it too early to make recommendations that can help the data science education community develop strategic vision and practices. The following is a summary of the findings and recommendations discussed in the preceding four chapters of this report..

REFERENCES

1. “Abdulhakam.AM.Assid“iq, Othman O. Khalifa, Md. Rafiqul Islam, Sheroz Khan Department of Electrical & Computer Faculty of Engineering,”- International Islamic University Malaysia.
2. https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825.
3. htttp://www.kaggle.com/c/walmart-recruiting-stores-sales-forecasting.
4. www.kdnuggets.com
5. http://www.themalaysian.blogspot.com/2006/08/fatal-roadaccidents-ranking-malaysia.html, August.2006.
6. https://www.ibm.com/in-en/analytics/data-science
7. https://www.edureka.co/blog/what-is-data-science/
8. https://www.simplilearn.com/big-data-and-analytics/senior-data-scientist-masters-program-training