Royal Education Society's

COCSIT LATUR

FACULTY OF COMPUTER STUDIES

[NEW CBCS PATTERN]

BSC CS TY(V SEM)

Model Question Paper

Data Science(BCS 503)

Date: */ /* ***Time:***

Time: *Three Hours* ***Maximum Marks*** *- 75*

Instructions to the candidates:

1. All questions are Compulsory.

2. Figures to the right indicate full marks.

3. Assume suitable data, if required.

Q.1 Attempt any FIVE of the following (3 Marks each) 15

1. Explain Data Visualization
2. Explain Data Mining
3. What is Artificial Intelligence
4. What is scalable data?
5. Importance of data science
6. Importance of research methodology
7. What is predictive analytics

Q. 2 Attempt any three of the following (5 Marks each) 15

1. Explain classification
2. Explain essential of algorithms
3. Explain hypothesis technique
4. Differentiate descriptive and inferential statistics

Q. 3 Attempt any three of the following (5 Marks each) 15

1. Explain data computational techniques
2. Explain parallel computing algorithm
3. Write various applications of data science
4. Explain experimentation and evaluation in data science

Q. 4 Attempt any three of the following (5 Marks each) 15

1. Write a note on Big Data
2. Explain segmentation using Clustering
3. Explain data science life cycle
4. Explain different machine learning algorithms (any three)

Q .5 Short notes on any three of the following (5 Marks each) 15

1. EDA
2. Regression Analysis
3. data structure
4. Machine Learning
5. AI

******

.

1Q Attempt any FIVE of the following (3 Marks each) 15

1. **Explain Data Visualization**

Data visualization provides a quick and effective way to communicate information in a universal manner using visual information. The practice can also help businesses identify which factors affect customer behavior; pinpoint areas that need to be improved or need more attention; make data more memorable for stakeholders; understand when and where to place specific products and predict sales volumes.

Other benefits of data visualization include the following:

- the ability to absorb information quickly, improve insights and make faster decisions;
- an increased understanding of the next steps that must be taken to improve the organization;
- an improved ability to maintain the audience interest with information they can understand;
- an easy distribution of information that increases the opportunity to share insights with everyone involved;
- eliminate the need for data scientists since data is more accessible and understandable; and
- an increased ability to act on findings quickly and, therefore, achieve success with greater speed and less mistakes.

2. **Explain Data Mining**

The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data Mining.

Data mining is the act of automatically searching for large stores of information to find trends and patterns that go beyond simple analysis procedures. Data mining utilizes complex mathematical algorithms for data segments and evaluates the probability of future events. Data Mining is also called Knowledge Discovery of Data (KDD).

3. **What is Artificial Intelligence**

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

Artificial neural networks and deep learning artificial intelligence technologies are quickly evolving, primarily because AI processes large amounts of data much faster and makes predictions more accurately than humanly possible.

4. **What is scalable data?**

The ability of a computer application or product (hardware or software) to continue to function well when it (or its context) is changed in size or volume in order to meet a user need. Typically,

the rescaling is to a larger size or volume. The rescaling can be of the product itself (for example, a line of computer systems of different sizes in terms of storage

Scalable structures refer to structures that can grow with the input grows without concern for memory. Imagine scalable structures such as a tree structure that can grow and search elements quickly, generally O(log(n)) even though they are a bit more memory intensive.

5. Importance of data science

- ✓ In the healthcare industry, physicians use Data Science to analyze data from wearable trackers to ensure their patients' well-being and make vital decisions. Data Science also enables hospital managers to reduce waiting time and enhance care.
- ✓ Data Science is widely used in the banking and finance sectors for fraud detection and personalized financial advice.
- ✓ Data Science facilitates firms to leverage social media content to obtain real-time media content usage patterns. This enables the firms to create target audience-specific content, measure content performance, and recommend on-demand content.
- ✓ Data Science helps study utility consumption in the energy and utility domain. This study allows for better control of utility use and enhanced consumer feedback.
- ✓ Data Science applications in the public service field include health-related research, financial market analysis, fraud detection, energy exploration, environmental protection, and more.

6. Importance of research methodology

- Research inculcates scientific and inductive thinking and promotes the development of logical habits of thinking and organization.
- Research plays a dynamic role in several fields and it has increased significantly in recent times, it can be related to small businesses and also to the economy as a whole.
- Most of the Government Regulations and Policies are based on and are a result of intensive research.
- Its significance lies in solving various planning and operational problems.
- It aids in decision making.
- It involves the study of cause and effect relationships between various variables and helps to identify behavior/patterns/trends in certain variables.

7. What is predictive analytics

Predictive analytics is a category of data analytics aimed at making predictions about future outcomes based on historical data and analytics techniques such as statistical modelling and machine learning. The science of predictive analytics can generate future insights with a significant degree of precision. With the help of sophisticated predictive analytics tools and models, any organization can now use past and current data to reliably forecast trends and behaviors milliseconds, days, or years into the future.

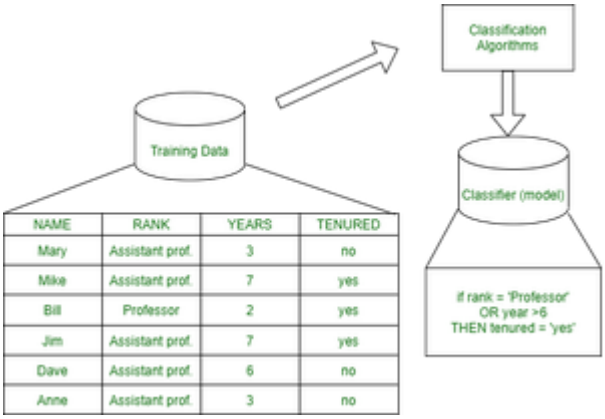
Predictive analytics tools give users deep, real-time insights into an almost endless array of business activities. Tools can be used to predict various types of behaviour and patterns,

such as how to allocate resources at particular times, when to replenish stock or the best moment to launch a marketing campaign, basing predictions on an analysis of data collected over a period of time.

Q. 2 Attempt any three of the following (5 Marks each) 15

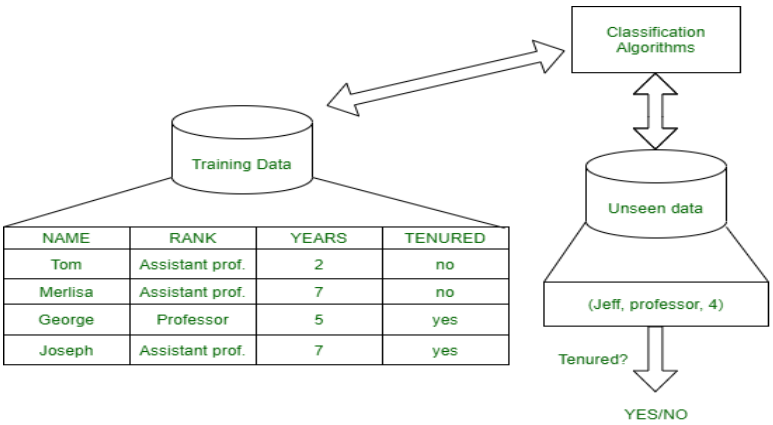
1) Explain classification

Classification: It is a data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.
Example: Before starting any project, we need to check its feasibility. In this case, a classifier is required to predict class labels such as 'Safe' and 'Risky' for adopting the Project and to further approve it. It is a two-step process such as:
1. Learning Step (Training Phase): Construction of Classification Model
Different Algorithms are used to build a classifier by making the model learn using the training set available. The model has to be trained for the prediction of accurate results.



NAME	RANK	YEARS	TENURED
Mary	Assistant prof.	3	no
Mike	Assistant prof.	7	yes
Bill	Professor	2	yes
Jim	Assistant prof.	7	yes
Dave	Assistant prof.	6	no
Anne	Assistant prof.	3	no

2. Classification Step: Model used to predict class labels and testing the constructed model on test data and hence estimate the accuracy of the classification rules.



NAME	RANK	YEARS	TENURED
Tom	Assistant prof.	2	no
Merlisa	Assistant prof.	7	no
George	Professor	5	yes
Joseph	Assistant prof.	7	yes

Training and Testing:
Suppose there is a person who is sitting under a fan and the fan starts falling on him, he should

get aside in order not to get hurt. So, this is his training part to move away. While Testing if the person sees any heavy object coming towards him or falling on him and moves aside then the system is tested positively and if the person does not move aside then the system is negatively tested.
Same is the case with the data, it should be trained in order to get the accurate and best results. There are certain data types associated with data mining that actually tells us the format of the file (whether it is in text format or in numerical format).
Attributes – Represents different features of an object. Different types of attributes are:

1. Binary: Possesses only two values i.e. True or False
Example: Suppose there is a survey evaluating some products. We need to check whether it's useful or not. So, the Customer has to answer it in Yes or No.
Product usefulness: Yes / No
 Symmetric: Both values are equally important in all aspects
 Asymmetric: When both the values may not be important.
2. Nominal: When more than two outcomes are possible. It is in Alphabet form rather than being in Integer form.
Example: One needs to choose some material but of different colors. So, the color might be Yellow, Green, Black, Red.
Different Colors: Red, Green, Black, Yellow
 Ordinal: Values that must have some meaningful order.
Example: Suppose there are grade sheets of few students which might contain different grades as per their performance such as A, B, C, D
Grades: A, B, C, D
 Continuous: May have an infinite number of values, it is in float type
Example: Measuring the weight of few Students in a sequence or orderly manner i.e. 50, 51, 52, 53
Weight: 50, 51, 52, 53
 Discrete: Finite number of values.
Example: Marks of a Student in a few subjects: 65, 70, 75, 80, 90
Marks: 65, 70, 75, 80, 90
Syntax:
 Mathematical Notation: Classification is based on building a function taking input feature vector "X" and predicting its outcome "Y" (Qualitative response taking values in set C)
 Here Classifier (or model) is used which is a Supervised function, can be designed manually based on expert's knowledge. It has been constructed to predict class labels (Example: Label – "Yes" or "No" for the approval of some event).

2) Explain essential of algorithms

In data science, computer science and statistics converge. As data scientists, we use statistical principles to write code such that we can effectively explore the problem at hand.
This necessitates at least a basic understanding of data structures, algorithms, and time-space complexity so that we can program more efficiently and understand the tools that we use. With larger datasets, this

becomes particularly important. The way that we write our code influences the speed at which our data is analyzed and conclusions can be reached accordingly.

Algorithms are everywhere in programming, and with good reason. They provide a set of instructions for solving all kinds of software problems, making life easier for developers. There are thousands of programming algorithms out there today, so good software developers and engineers need to know the different ones that are available and when it is most appropriate to use them. A good algorithm will find the most efficient way to perform a function or solve a problem, both in terms of speed and by minimizing the use of computer memory.
Many problems can be resolved by starting with some of the most popular algorithms according to the function required. For example, sorting algorithms are instructions for arranging the items of an array or list into a particular order, while searching algorithms are used to find and retrieve an element from wherever it is stored in a data structure. Here are 11 algorithms that we think every programmer should know about:

Here are some principles that are important to understand before discussing some of the common algorithms.

Sorting Algorithms:
Sorting raw data sets is a simple but crucial step in computer/data science, and increasingly important in the age of big data. Sorting typically involves finding numerical or alphabetical orders (ascending or descending).

Searching Algorithms
Searching is such a basic IT function, but so important to get right when programming. It could involve searching for something within an internal database or trawling virtual spaces for a specific piece of information, and there are two standard approaches used today.

Recursion: Recursion is when a function calls itself. Perhaps the quintessential example of recursion is in implementation of a factorial function:

```
def factorial(n):
    if n < 1:          #base case
        return 1
    else:               #recursive case
        return n * factorial(n-1)
```

The function is called within the function itself and will continue calling itself until the base case (in this case, when n is 1) is reached.

Divide and Conquer (D&C): A recursive approach for problem-solving, D&C (1) determines the simplest case for the problem (AKA the base case) and (2) reduces the problem until it is now the base case. That is, a complex problem is broken down into simpler sub-problems. These sub-problems are solved and their solutions are then combined to solve the original, larger problem

3) Explain hypothesis technique

Hypothesis testing is generally used when you are comparing two or more groups.

For example, you might implement protocols for performing intubation on pediatric patients in the pre-hospital setting. To evaluate whether these protocols were successful in improving intubation rates, you could measure the intubation rate over time in one group randomly assigned to training in the new protocols, and compare this to the intubation rate over time in another control group that did not receive training in the new protocols.

When you are evaluating a hypothesis, you need to account for both the variability in your sample and how large your sample is. Based on this information, you'd like to make an assessment of whether any differences you see are meaningful, or if they are likely just due to chance. This is formally done through a process called hypothesis testing.
Five Steps in Hypothesis Testing:

1. Specify the Null Hypothesis
2. Specify the Alternative Hypothesis
3. Set the Significance Level (a)
4. Calculate the Test Statistic and
Corresponding P-Value
5. Drawing a Conclusion

Step 1: Specify the Null Hypothesis
The null hypothesis (H 0) is a statement of no effect, relationship, or difference between two or more groups or factors. In research studies, a researcher is usually interested in disproving the null hypothesis.

Step 2: Specify the Alternative Hypothesis
The alternative hypothesis (H 1) is the statement that there is an effect or difference. This is usually the hypothesis the researcher is interested in proving. The alternative hypothesis can be one-sided (only provides one direction, e.g., lower) or two-sided. We often use two-sided tests even when our true hypothesis is one-sided because it requires more evidence against the null hypothesis to accept the alternative hypothesis.

Step 3: Set the Significance Level (a)
The significance level (denoted by the Greek letter alpha— a) is generally set at 0.05. This means that there is a 5% chance that you will accept your alternative hypothesis when your null hypothesis is actually true. The smaller the significance level, the greater the burden of proof needed to reject the null hypothesis, or in other words, to support the alternative hypothesis.

Step 4: Calculate the Test Statistic and Corresponding P-Value
In another section we present some basic test statistics to evaluate a hypothesis. Hypothesis testing generally uses a test statistic that compares groups or examines associations between variables. When describing a single sample without establishing relationships between variables, a confidence interval is commonly used.

The p-value describes the probability of obtaining a sample statistic as or more extreme by

chance alone if your null hypothesis is true. This p-value is determined based on the result of your test statistic. Your conclusions about the hypothesis are based on your p-value and your significance level.

4) Differentiate descriptive and inferential statistics

Descriptive vs Inferential Statistics: Key Differences

Descriptive Statistics	**Inferential Statistics**
It is concerned with describing the population under study. Sampling is not required.	It focuses on drawing conclusions about the populations, based on sample analysis.
Collects, organizes, analyses and presents the data in a meaningful way.	Compares data, test hypotheses and make predictions of the future outcome.
The form of result is charts, Graphs, and tables.	The result is displayed in the form of probability.
It describes a situation.	It explains the likelihood of the occurrence of an event.
It explains the data (already known) to summarize sample.	It attempts to reach the conclusions to learn about the population; that extends beyond the data available.

Both, Descriptive and Inferential Statistics methods are equally critical to advancements across scientific fields like data science. That's why it becomes extremely important for statisticians and data scientists to understand that both methods have their own advantages and limitations.

3Q Attempt any three of the following (5 Marks each) 15

1. Explain data computational techniques

In order to obtain valuable insight from the vast amounts of data that connected devices generate on a daily basis, data scientists must be capable of applying data science models to IoT datasets in order to extract, store and analyse said data, make real time predictions and detect anomalies. The sheer volume of IoT datasets requires traditional data science technologies to be improved and enhanced, so that we are able to discover critical insight, improve processes and make smarter decisions. So, although traditional data science paved the way for the data analytics that we know today, the analytic techniques we use for IoT are far more sophisticated and advanced, due to the nature of the data they deal with.

1) Basic Statistics:

This class includes basic statistical tasks. Examples include computing the mean, variance, and other moments; estimating the number of distinct elements in a data set; counting the number

of elements and finding frequently occurring ones; and calculating order statistics such as the median.

These tasks typically require O(N) calculations for N data points. Some other calculations that arguably fall into this class include sorting and basic forms of clustering.

Such simple statistical computations are widely used in and of themselves, but they also appear inside myriad more complex analyses. For example, multidimensional counts are important in count-based methods such as association rules and in probabilistic inference in graphical models with discrete variables.

2. Linear Algebraic computation:

This class includes all the standard problems of computational linear algebra, including linear systems, eigenvalue problems, and inverses. A large number of linear models, including linear regression, PCA, and their many variants, result in linear algebraic computations. Many of these are well-solved by generic linear algebra approaches.

3. Explain parallel computing algorithm

Parallel Computing:
It is the use of multiple processing elements simultaneously for solving any problem. Problems are broken down into instructions and are solved concurrently as each resource that has been applied to work is working at the same time.

Advantages of Parallel Computing over Serial Computing are as follows:

1.It saves time and money as many resources working together will reduce the time and cut potential costs.
2.It can be impractical to solve larger problems on Serial Computing.
3.It can take advantage of non-local resources when the local resources are finite.
4.Serial Computing ‘wastes’ the potential computing power, thus Parallel Computing makes better work of the hardware.

Types of Parallelism:

1.Bit-level parallelism –
It is the form of parallel computing which is based on the increasing processor’s size. It reduces the number of instructions that the system must execute in order to perform a task on large-sized data.
Example: Consider a scenario where an 8-bit processor must compute the sum of two 16-bit integers. It must first sum up the 8 lower-order bits, then add the 8 higher-order bits, thus requiring two instructions to perform the operation. A 16-bit processor can perform the operation with just one instruction.

2.Instruction-level parallelism –
A processor can only address less than one instruction for each clock cycle phase. These instructions can be re-ordered and grouped which are later on executed concurrently without affecting the result of the program. This is called instruction-level parallelism.

3.Task Parallelism –
Task parallelism employs the decomposition of a task into subtasks and then allocating each of the subtasks for execution. The processors perform the execution of sub-tasks concurrently.

4. Data-level parallelism (DLP) –
Instructions from a single stream operate concurrently on several data – Limited by non-regular data manipulation patterns and by memory bandwidth

4. Write various applications of data science

1. Fraud and Risk Detection
The earliest applications of data science were in Finance. Companies were fed up of bad debts and losses every year. However, they had a lot of data which use to get collected during the initial paperwork while sanctioning loans. They decided to bring in data scientists in order to rescue them out of losses.
Over the years, banking companies learned to divide and conquer data via customer profiling, past expenditures, and other essential variables to analyze the probabilities of risk and default. Moreover, it also helped them to push their banking products based on customer's purchasing power.

2. Medical Image Analysis
Procedures such as detecting tumors, artery stenosis, organ delineation employ various different methods and frameworks like Map Reduce to find optimal parameters for tasks like lung texture classification. It applies machine learning methods, support vector machines (SVM), content-based medical image indexing, and wavelet analysis for solid texture classification.

3. Genetics and Genomics
Data Science applications also enable an advanced level of treatment personalization through research in genetics and genomics. The goal is to understand the impact of the DNA on our health and find individual biological connections between genetics, diseases, and drug response. Data science techniques allow integration of different kinds of data with genomic data in the disease research, which provides a deeper understanding of genetic issues in reactions to particular drugs and diseases. As soon as we acquire reliable personal genome data, we will achieve a deeper understanding of the human DNA. The advanced genetic risk prediction will be a major step towards more individual care.

4. Drug Development

The drug discovery process is highly complicated and involves many disciplines. The greatest ideas are often bounded by billions of testing, huge financial and time expenditure. On average, it takes twelve years to make an official submission.
Data science applications and machine learning algorithms simplify and shorten this process, adding a perspective to each step from the initial screening of drug compounds to the prediction of the success rate based on the biological factors. Such algorithms can forecast how the compound will act in the body using advanced mathematical modeling and simulations instead of the "lab experiments". The idea behind the computational drug discovery is to create computer model simulations as a biologically relevant network simplifying the prediction of future outcomes with high accuracy.

5. Virtual assistance for patients and customer support
Optimization of the clinical process builds upon the concept that for many cases it is not actually necessary for patients to visit doctors in person. A mobile application can give a more effective solution by bringing the doctor to the patient instead.
The AI-powered mobile apps can provide basic healthcare support, usually as chatbots. You simply describe your symptoms, or ask questions, and then receive key information about your medical condition derived from a wide network linking symptoms to causes. Apps can remind you to take your medicine on time, and if necessary, assign an appointment with a doctor.
This approach promotes a healthy lifestyle by encouraging patients to make healthy decisions, saves their time waiting in line for an appointment, and allows doctors to focus on more critical cases.

6. Targeted Advertising
If you thought Search would have been the biggest of all data science applications, here is a challenger – the entire digital marketing spectrum. Starting from the display banners on various websites to the digital billboards at the airports – almost all of them are decided by using data science algorithms.
This is the reason why digital ads have been able to get a lot higher CTR (Call-Through Rate) than traditional advertisements. They can be targeted based on a user's past behavior.
This is the reason why you might see ads of Data Science Training Programs while I see an ad of apparels in the same place at the same time.

7. Advanced Image Recognition
You upload your image with friends on Facebook and you start getting suggestions to tag your friends. This automatic tag suggestion feature uses face recognition algorithm.

In their latest update, Facebook has outlined the additional progress they've made in this area, making specific note of their advances in image recognition accuracy and capacity.
"We've witnessed massive advances in image classification (what is in the image?) as well as object detection (where are the objects?), but this is just the beginning of understanding the most relevant visual content of any image or video. Recently we've been designing techniques that identify and segment each and every object in an image, a key capability that will enable entirely new applications."

In addition, Google provides you with the option to search for images by uploading them. It uses image recognition and provides related search results.

5. Explain experimentation and evaluation in data science

Experimentation:

1. A/B Testing

AB testing is an extremely common method of experimentation used in industry to understand the impact changes we make to our product. It could be as simple as changing the layout on a web page or the color of a button and measuring the effect this change has on a key metric such as click-through rates. In general, we can take two different approaches to AB testing. We can use the Frequentist approach and the Bayesian approach, each of which has its own advantages and disadvantages.

Frequentist:
I would say frequentist AB testing is by far the most common type of AB testing done and follows directly from the principles of frequentist statistics. The goal here is to measure the causal effect of our treatment by seeing if the difference between our metric in the A and B groups is statistically significant at some significance level, 5 or 1 per cent is typically chosen. More specifically, we will need to define a null and alternate hypothesis and determine if we can or cannot reject the null. Depending on the type of metric we choose we might use a different statistical test but chi-square and t-tests are commonly used in practice. A key point about the frequentist approach is that the parameter or metric we compute is a constant. Therefore, there is no probability distribution associated with it.

Bayesian
The key difference in the Bayesian approach is that our metric is a random variable and therefore has a probability distribution. This is quite useful as we can now incorporate uncertainty about our estimates and make probabilistic statements which are often much more intuitive to people than the frequentist interpretation. Another advantage of using a Bayesian approach is that we may reach a solution faster compared to AB testing as we do not necessarily need to assign equal numbers of data to each variant. This means that a Bayesian approach may converge to a solution faster using fewer resources.

2. Regression Discontinuity Design (RDD)

RDD is another technique available taken from economics and is particularly suited when we have a continuous natural cut-off point. Those just below the cut-off do not get the treatment and those just above the cut-off do get the treatment. The idea is that these two groups of people are very similar so the only systematic difference between them is whether they were treated or not.

Because the groups are considered very similar these assignments essentially approximates randomized selection. For example is that certificates of merit were only given to individuals who

scored above a certain threshold on a test. Using this approach we can now just compare the average outcome between the two groups at the threshold to see if there is a statistically significant effect.

It may help to visualise RDD. Below is a graph which shows the average outcome below and above the threshold. Essentially all we are doing is measuring the difference between the two blue lines beside the cutoff point.

Source: Example: RDD cutoff: Threshold at 50 on the x-axis

3. Difference in Differences (Diff in Diff)

I will go into a bit more detail on Diff in Diff as I recently used this technique on a project. The problem was one that many data scientists might come across in daily work and was related to preventing churn. It is very common for companies to try and identify customers who are likely to churn and then to design interventions to prevent it. Now identifying churn is a problem for machine learning. What we care about is whether we can come up
with a way to measure the effectiveness of our churn intervention. Being able to empirically measure the effectiveness of our decisions is incredibly important as we want to quantify the effects of our features (usually in monetary terms) and it is a vital part of making informed decisions for the business. One such intervention would be to send an email to those at risk of churning reminding them of their account and in effect to try and make them more engaged with our product. This is the basis of the problem we will be looking at here.

Evaluation:

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and over fitted models. There are two methods of evaluating models in data science, Hold-Out and Cross-Validation. To avoid overfitting, both methods use a test set (not seen by the model) to evaluate model performance.

1.Hold-Out

In this method, the mostly large dataset is randomly divided to three subsets:

1. Training set is a subset of the dataset used to build predictive models.
2. Validation set is a subset of the dataset used to assess the performance of model built in the training phase. It provides a test platform for fine tuning model's parameters and selecting the best-performing model. Not all modeling algorithms need a validation set.
3. Test set or unseen examples is a subset of the dataset to assess the likely future performance

of a model. If a model fit to the training set much better than it fits the test set, overfitting is probably the cause.

2.Cross-Validation

When only a limited amount of data is available, to achieve an unbiased estimate of the model performance we use k-fold cross-validation. In k-fold cross-validation, we divide the data into k subsets of equal size. We build models k times, each time leaving out one of the subsets from training and use it as the test set. If k equals the sample size, this is also called leave-one-out.

4Q Attempt any three of the following (5 Marks each) 15

1. **Write a note on Big Data**

'Big Data' is the application of specialized techniques and technologies to process very large sets of data. These data sets are often so large and complex that it becomes difficult to process using on-hand database management tools. Examples include web logs, call records, medical records, military surveillance, photography archives, video archives and large-scale e-commerce
Big data management is a broad concept that encompasses the policies, procedures and technology used for the collection, storage, governance, organization, administration and delivery of large repositories of data. It can include data cleansing, migration, integration and preparation for use in reporting and analytics.
Machine Learning Techniques
Three types of Machine Learning Algorithms

1. Supervised Learning How it works: This algorithm consist of a target / outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables). Using these set of variables, we generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data. Examples of Supervised Learning: Regression, Decision Tree, Random Forest, KNN, Logistic Regression etc.
2. Unsupervised Learning How it works: In this algorithm, we do not have any target or outcome variable to predict / estimate. It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention. Examples of Unsupervised Learning: Apriori algorithm, K-means.

3. Reinforcement Learning: How it works: Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it trains itself continually using trial and error. This machine learns from past experience and tries to capture the best possible knowledge to make accurate business decisions. Example of Reinforcement Learning: Markov Decision Process

2. Explain segmentation using Clustering

There is not one process for clustering and segmentation. However, we have to start somewhere, so we will use the following process:
Clustering and Segmentation in 9 steps
1. Confirm data is metric
2. Scale the data
3. Select Segmentation Variables
4. Define similarity measure
5. Visualize Pair-wise Distances
6. Method and Number of Segments
7. Profile and interpret the segments
8. Robustness Analysis

Step 1: Confirm data is metric

While one can cluster data even if they are not metric, many of the statistical methods available for clustering require that the data are so: this means not only that all data are numbers, but also that the numbers have an actual numerical meaning, that is, 1 is less than 2, which is less than 3 etc. The main reason for this is that one needs to define distances between observations (see step 4 below), and often ("black box" mathematical) distances (e.g. the "Euclidean distance") are defined only with metric data.
However, one could potentially define distances also for non-metric data. In general, a "best practice" for segmentation is to creatively define distance metrics between our observations.

Step 2: Scale the data

This is an optional step. Note that for this data, while 6 of the "survey" data are on a similar scale, namely 1-7, there is one variable that is about 2 orders of magnitude larger: the Income variable.
Having some variables with a very different range/scale can often create problems: most of the "results" may be driven by a few large values, more so than we would like. To avoid such issues, one has to consider whether or not to standardize the data by making some of the initial raw attributes have, for example, mean 0 and standard deviation

Step 3: Select Segmentation Variables

The decision about which variables to use for clustering is a critically important decision that will have a big impact on the clustering solution. So we need to think carefully about the variables we will choose for clustering. Good exploratory research that gives us a good sense of what variables may distinguish people or products or assets or regions is critical. Clearly this is a step where a lot of contextual knowledge, creativity, and experimentation/iterations are needed. Moreover, we often use only a few of the data attributes for segmentation (the segmentation attributes) and use some of the remaining ones (the profiling attributes) For example, in market research and market segmentation, one may use attitudinal data for segmentation (to segment the customers based on their needs and attitudes towards the products/services) and then

demographic and behavioral data for profiling the segments found.

Step 4: Define similarity measure
Remember that the goal of clustering and segmentation is to group observations based on how similar they are. It is therefore crucial that we have a good understanding of what makes two observations (e.g. customers, products, companies, assets, investments, etc.) “similar”.
If the user does not have a good understanding of what makes two observations (e.g. customers, products, companies, assets, investments, etc.) “similar”, no statistical method will be able to discover the answer to this question.
Most statistical methods for clustering and segmentation use common mathematical measures of distance. Typical measures are, for example, the Euclidean distance or the Manhattan distance

Step 5: Visualize Pair-wise Distances
Having defined what we mean “two observations are similar”, the next step is to get a first understanding of the data through visualizing for example individual attributes as well as the pairwise distances (using various distance metrics) between the observations. If there are indeed multiple segments in our data, some of these plots should show “mountains and valleys”, with the mountains being potential segments.
Visualization is very important for data analytics, as it can provide a first understanding of the data.

Step 6: Method and Number of Segments
There are many statistical methods for clustering and segmentation. In practice one may use various approaches and then eventually select the solution that is statistically robust interpretable, and actionable - among other criteria.
Two widely used methods: the Kmeans Clustering Method, and the Hierarchical Clustering Method. Like all clustering methods, these two also require that we have decided how to measure the distance/similarity between our observations. Explaining how these methods work is beyond our scope. The only difference to highlight is that Kmeans requires the user to define how many segments to create, while Hierarchical Clustering does not.

Step 7: Profile and interpret the segments
Having decided (for now) how many clusters to use, we would like to get a better understanding of who the customers in those clusters are and interpret the segments.
Data analytics is used to eventually make decisions, and that is feasible only when we are comfortable (enough) with our understanding of the analytics results, including our ability to clearly interpret them.
To this purpose, one needs to spend time visualizing and understanding the data within each of the selected segments. For example, one can see how the summary statistics (e.g. averages, standard deviations, etc) of the profiling attributes differ across the segments.

Step 8: Robustness Analysis
The segmentation process outlined so far can be followed with many different approaches, for example:
 using different subsets of the original data

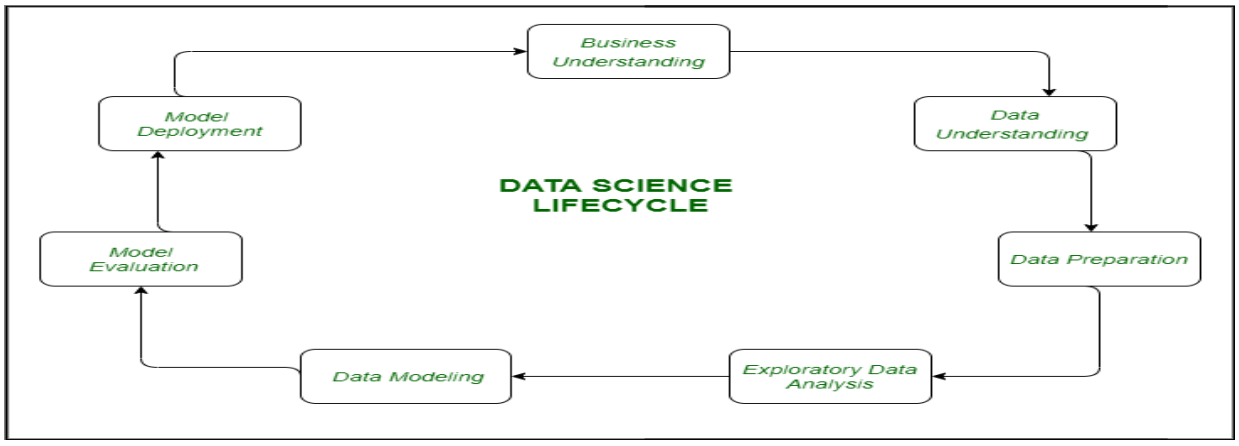
- using variations of the original segmentation attributes
- using different distance metrics
- using different segmentation methods
- using different numbers of clusters

Much like any data analysis, segmentation is an iterative process with many variations of data, methods, number of clusters, and profiles generated until a satisfying solution is reached.

Data Analytics is an iterative process; therefore, we may need to return to our original raw data at any point and select new raw attributes as well as new clusters.

3. Explain data science life cycle

Data Science Lifecycle revolves around the use of machine learning and different analytical strategies to produce insights and predictions from information in order to acquire a commercial enterprise objective. The complete method includes a number of steps like data cleaning, preparation, modelling, model evaluation, etc.



The lifecycle outlines the major stages that projects typically execute, often iteratively:

1. **Business understanding:** The complete cycle revolves around the enterprise goal. After desirable perception only we can set the precise aim of evaluation that is in sync with the enterprise objective.
2. **Data Understanding:** After enterprise understanding, the subsequent step is data understanding. This includes a series of all the reachable data. This step includes describing the data, their structure, their relevance, their records type. Explore the information using graphical plots. Basically, extracting any data that you can get about the information through simply exploring the data.
3. **Preparation of Data:** Next comes the data preparation stage. This consists of steps like choosing the applicable data, integrating the data by means of merging the data sets, cleaning it, treating the lacking values through either eliminating them. Format the data into the preferred structure, eliminate undesirable columns and features.
4. **Exploratory Data Analysis:** This step includes getting some concept about the answer and elements affecting it, earlier than constructing the real model. Distribution of data inside distinctive variables of a character is explored graphically the usage of bar-graphs, Relations between distinct aspects are captured via graphical representations like scatter plots and warmth maps.
5. **Data Modeling:** Data modelling is the coronary heart of data analysis. A model takes the organized data as input and gives the preferred output. This step consists of selecting the suitable kind of model,

whether the problem is a classification problem, or a regression problem or a clustering problem. After deciding on the model family, amongst the number of algorithms amongst that family, we need to cautiously pick out the algorithms to put into effect and enforce them. We need to tune the hyper parameters of every model to obtain the preferred performance

6. **Model Evaluation:** Here the model is evaluated for checking if it is geared up to be deployed. The model is examined on an unseen data, evaluated on a cautiously thought out set of assessment metrics. We additionally need to make positive that the model conforms to reality. If we do not acquire a quality end result in the evaluation, we have to re-iterate the complete modelling procedure until the preferred stage of metrics is achieved.
7. **Model Deployment:** The model after a rigorous assessment is at the end deployed in the preferred structure and channel.

4. Explain different machine learning algorithms

Here is the list of commonly used machine learning algorithms. These algorithms can be applied to almost any data problem:

1. Decision Tree
2. KNN
3. K-Means

1. Decision Tree

This is one of my favorite algorithm and I use it quite frequently. It is a type of supervised learning algorithm that is mostly used for classification problems. Surprisingly, it works for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible

In the image above, you can see that population is classified into four different groups based on multiple attributes to identify 'if they will play or not'. To split the population into different heterogeneous groups, it uses various techniques like Gini, Information Gain, Chi-square, entropy.

2. KNN (K- Nearest Neighbors) It can be used for both classification and regression problems. However, it is more widely used in classification problems in the industry. K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The case being assigned to the class is most common amongst its K nearest neighbors measured by a distance function. These distance functions can be Euclidean, Manhattan, Minkowski and Hamming distance. First three functions are used for continuous function and fourth one (Hamming) for categorical variables. If K = 1, then the case is simply assigned to the class of its nearest neighbor. At times, choosing K turns out to be a challenge while performing KNN modeling.

KNN can easily be mapped to our real lives. If you want to learn about a person, of whom

you have no information, you might like to find out about his close friends and the circles he moves in and gain access to his/her information!
Things to consider before selecting KNN:
 KNN is computationally expensive
 Variables should be normalized else higher range variables can bias it
 Works on pre-processing stage more before going for KNN like outlier, noise removal.

3. K-Means It is a type of unsupervised algorithm which solves the clustering problem. Its procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). Data points inside a cluster are homogeneous and heterogeneous to peer groups. Remember figuring out shapes from ink blots? k means is somewhat similar this activity. You look at the shape and spread to decipher how many different clusters / population are present!

How K-means forms cluster:
1. K-means picks k number of points for each cluster known as centroids.
2. Each data point forms a cluster with the closest centroids i.e. k clusters.
3. Finds the centroid of each cluster based on existing cluster members. Here we have new centroids.
4. As we have new centroids, repeat step 2 and 3. Find the closest distance for each data point from new centroids and get associated with new k-clusters. Repeat this process until convergence occurs i.e. centroids does not change.
How to determine value of K:
In K-means, we have clusters and each cluster has its own centroid. Sum of square of difference between centroid and the data points within a cluster constitutes within sum of square value for that cluster. Also, when the sum of square values for all the clusters are added, it becomes total within sum of square value for the cluster solution. We know that as the number of cluster increases, this value keeps on decreasing but if you plot the result you may see that the sum of squared distance decreases sharply up to some value of k, and then much more slowly after that. Here, we can find the optimum number of cluster.

5Q

2. EDA

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. 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.

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 deal with 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:
o Stem-and-leaf plots, which show all data values and the shape of the distribution.
o Histograms, a bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values.

o 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 or statistics.
 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 variable.
Other common types of multivariate graphics include:
 Scatter plot, which is used to plot data points on a horizontal and a vertical axis to show how much one variable is affected by another.
 Multivariate chart, which is a graphical representation of the relationships between factors and a response.
 Run chart, which is a line graph of data plotted over time.
 Bubble chart, which is a data visualization that displays multiple circles (bubbles) in a two-dimensional plot.
 Heat map, which is a graphical representation of data where values are depicted by color.

3. Regression Analysis

The process of identifying the relationship and the effects of this relationship on the outcome of future values of objects is defined as regression. Regression helps in identifying the behavior of a variable when other variable(s) are changed in the process. Regression analysis is used for prediction and forecasting applications.

Linear Regression
Linear regression is such a useful and established algorithm, that it is both a statistical model and a machine learning model. Linear regression tries a draw a best fit line that is close to the data by finding the slope and intercept.

Linear regression equation is,
$Y=a+bx$

In this equation:

 y is the output variable. It is also called the target variable in machine learning or the dependent variable.

 x is the input variable. It is also referred to as the feature in machine learning or it is called the independent variable.

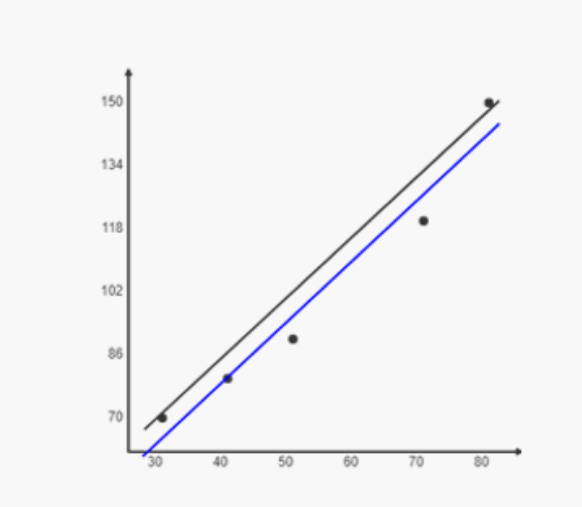
 a is the constant

 b is the coefficient of independent variable

Electricity bill (Y)	Weather(X)
70	30
80	40
90	50
120	70
150	80

Here the weather is given in Fahrenheit and Electricity bill is in dollars...

Lets plot in graph and find the best fit line



Multiple linear regression

Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables. It is sometimes known simply as multiple regression, and it is an extension of linear regression. The variable that we want to predict is known as the dependent variable, while the variables we use to predict the value of the dependent variable are known as independent or explanatory variables.

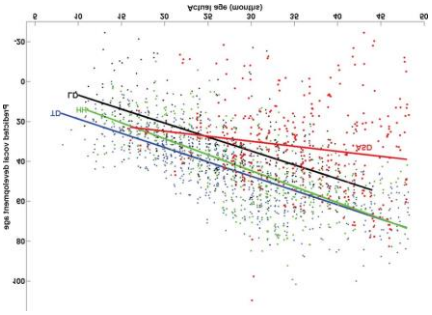


Figure 1: Multiple linear regression model predictions for individual observations (Source)
 Multiple linear regression refers to a statistical technique that uses two or more independent variables to predict the outcome of a dependent variable.
 The technique enables analysts to determine the variation of the model and the relative contribution of each independent variable in the total variance.
 Multiple regression can take two forms, i.e., linear regression and non-linear regression.
Multiple Linear Regression Formula

Where:
 yi is the dependent or predicted variable
 β0 is the y-intercept, i.e., the value of y when both xi and x2 are 0.
 β1 and β2 are the regression coefficients representing the change in y relative to a one-unit change in xi1 and xi2, respectively.
 βp is the slope coefficient for each independent variable
 ϵ is the model's random error (residual) term.
Understanding Multiple Linear Regression

Simple linear regression enables statisticians to predict the value of one variable using the available information about another variable. Linear regression attempts to establish the relationship between the two variables along a straight line.

Multiple regression is a type of regression where the dependent variable shows a linear relationship with two or more independent variables. It can also be non-linear, where the dependent and independent variables do not follow a straight line.
Both linear and non-linear regression track a particular response using two or more variables graphically. However, non-linear regression is usually difficult to execute since it is created from assumptions derived from trial and error.

4. Data structure

Data structures are the organizational tools data scientists use to update, manage and index internet services efficiently. Data structures are also used as the basis for many algorithms, due to the way they can filter and sort large quantities of data. There are three main parts to a data structure that make it work:

- **The Memory Address:** The fixed raw data element of any desired feature or function.
- **The Pointer:** A reference tool that represents the location of a memory address.
- **The Procedure:** A written code that manipulates or creates different functions inside the structure, either automatically or manually.

Types of Data Structures

How the above metrics are applied to data will determine what type of data structure any given database is using. There are several types of data structures, which include:

Arrays

Arrays are a more basic data structure. They are defined by any number of the same type of raw data element being present in a specific order. Arrays can be fixed-length or resizable—permitting that the data elements remain the same—and use an integer index as a pointer, and a mathematical formula procedure to compute specified data. Array data structures are great for accessing randomly generated data.

Linked Lists

Linked list structures are defined by a linear arrangement of raw data elements of any type, called nodes. Each node has a specific data value, and always points to the next node in the linear arrangement. Linked list data structures are great for efficiently inserting or removing individual data elements without having to restructure the rest of the list.

Stacks

Stack data structures are categorized by the data elements following a specific procedure in perpetuity. Stack procedure generally follows these steps:

- **Push:** Adds new items to the stack.
- **Pop:** Removes items from the stack.
- **Peek (or top):** Returns the top element of the stack.
- **Isempty:** Returns true when the stack is empty, otherwise returns false.

You can use stacks in tandem with a linked list or array data structure by substituting a stack for the procedure.

Trees

Tree graphs are a hierarchical data structure for the organization of larger or more complex abstract data types (ADT). Tree data structures are organized by three elements:

- **Root value**
- **Parent nodes**
- **Child nodes**

The root value is the beginning, or the root, of all the other raw data elements in the structure, and the cornerstone the rest of the data relates to. The parent nodes are larger, less specific data elements that smaller or more specific data can fall under as child nodes. There are several types of tree data structures, including:

- **Binary Trees:** Where each parent node has exactly zero or two children.
- **Binary Search Trees:** Where the left child value of a parent node is less than or equal to the parent value and the right child value is greater than or equal to the parent value.
- **N-ary Trees:** Where the maximum number of children a parent node can have is zero or N. For example, a 3-ary tree has no more than three children nodes.

Graphs

A graph data structure consists of nodes—also called vertices in some cases—and edges—also called lines or arcs. Due to the edges, which can connect any two nodes, graphs are a nonlinear data structure. There is a finite set of nodes in a graph, which are often used to represent data networks.

5. **Machine Learning**

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.

Evolution of machine learning

Because of new computing technologies, machine learning today is not like machine learning of the past. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks; researchers interested in artificial intelligence wanted to see if computers could learn from data. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results. It's a science that's not new – but one that has gained fresh momentum.

While many machine learning algorithms have been around for a long time, the ability to automatically apply complex mathematical calculations to big data – over and over, faster and faster – is a recent development. Here are a few widely publicized examples of machine learning applications you may be familiar with:

- The heavily hyped, self-driving Google car? The essence of machine learning.
- Online recommendation offers such as those from Amazon and Netflix? Machine learning applications for everyday life.
- Knowing what customers are saying about you on Twitter? Machine learning combined with linguistic rule creation.
- Fraud detection? One of the more obvious, important uses in our world today

5. **AI**

AI is just a computer that is able to mimic or simulate human thought or behavior. Within that, there's a subset called machine learning that is now the underpinning of the most exciting part of AI. By allowing computers to learn how to solve problems on their own, machine learning has made a series of breakthroughs that once seemed nearly impossible. It's the reason that computers can spot a friend's face in a photo or steer a car. It's the reason people are actively talking about the arrival of human-like AI.

In general, AI systems work by ingesting large amounts of labeled training data, analyzing the data for correlations and patterns, and using these patterns to make predictions about future states. In this way, a chatbot that is fed examples of text chats can learn to produce lifelike exchanges with people, or an image recognition tool can learn to identify and describe objects in images by reviewing millions of examples.

AI programming focuses on three cognitive skills: learning, reasoning and self-correction.
AI is important because it can give enterprises insights into their operations that they may not have been aware of previously and because, in some cases, AI can perform tasks better than humans. Particularly when it comes to repetitive, detail-oriented tasks like analyzing large numbers of legal documents to ensure relevant fields are filled in properly, AI tools often complete jobs quickly and with relatively few errors.