USN | | | | | | | | | | |

Internal Assessment Test 1 – December 2023

Sub:	Artificial Intelligence and Machine Learning - Set 2					Sub Code:	21CS54	Branch:	ISE
Date:	19-12-2023	Duration:	90 Minutes	Max Marks:	50	Sem / Sec:	**5 A,B,C**		**OBE**

	Answer any FIVE FULL Questions	MARKS	CO	RBT
1a	Define Machine Learning, Data, Information, Knowledge, and Intelligence with an example. Explain and Sketch the Knowledge pyramid representation.	8	CO2	L2
1b	List out the factors that drive the popularity of machine learning	2	CO2	L1
2a	List out the types of Data available and explain each with an example	7	CO2	L1
2b	Write a note on Elements of Big Data.	3	CO2	L1
3a	Explain Supervised, Unsupervised Learning & Semi-supervised Learning.	7	CO2	L2
3b	List out the major applications of machine learning in detail	3	CO2	L1
4a	Explain the machine-learning process model with the flow diagram	7	CO2	L2
4b	What are the differences between classification and regression with an example?	3	CO2	L1
5a	Explain Reinforcement Learning in detail with an example	6	CO2	L1
5b	How the data source can be classified, explain its types in detail	4	CO2	L2
6a	Explain Big Data Analytics and Types of Analytics with examples.	6	CO2	L2
6b	List out the challenges of Machine Learning in detail.	4	CO2	L2

CI **CCI** **HOD**

USN | | | | | | | | | | |

Internal Assessment Test 1 – December 2023

Sub:	Artificial Intelligence and Machine Learning - Set 2					Sub Code:	21CS54	Branch:	ISE
Date:	19-12-2023	Duration:	90 Minutes	Max Marks:	50	Sem / Sec:	**5 A,B,C**		**OBE**

	Answer any FIVE FULL Questions	MARKS	CO	RBT
1a	Define Machine Learning, Data, Information, Knowledge, and Intelligence with an example. Explain and Sketch the Knowledge pyramid representation.	8	CO2	L2
1b	List out the factors that drive the popularity of machine learning	2	CO2	L1
2a	List out the types of Data available and explain each with an example	7	CO2	L1
2b	Write a note on Elements of Big Data.	3	CO2	L1
3a	Explain Supervised, Unsupervised Learning & Semi-supervised Learning.	7	CO2	L2
3b	List out the major applications of machine learning in detail	3	CO2	L1
4a	Explain the machine-learning process model with the flow diagram	7	CO2	L2
4b	What are the differences between classification and regression with an example?	3	CO2	L1
5a	Explain Reinforcement Learning in detail with an example	6	CO2	L1
5b	How the data source can be classified, explain its types in detail	4	CO2	L2
6a	Explain Big Data Analytics and Types of Analytics with examples.	6	CO2	L2
6b	List out the challenges of Machine Learning in detail.	4	CO2	L2

CI **CCI** **HOD**

USN | | | | | | | | | | |

Internal Assessment Test 1 – December 2023

Sub:	Artificial Intelligence and Machine Learning Set 2					Sub Code:	21CS54	Branch:	ISE
Date:	19-12-023	Duration:	90 Minutes	Max Marks:	50	Sem / Sec:	**5 A,B,C**		**OBE**

Answer any FIVE FULL Questions			MARKS	CO
1a	• Machine Learning • Data • Information • Knowledge • Intelligence • Sketch the Knowledge pyramid representation.	1M 1M 1M 1M 1M 3M	8	CO2
1b	Drive the popularity of machine learning	2M	2	CO2
2a	Types of Data available and explain each with an example • Structured • Unstructured • Semistructured	1M 2M 2M 2M	7	CO2
2b	Write a note on Elements of Big Data. • Volume • Velocity • Variety • Veracity • Validity • Value	1M 1M 1M	3	CO2
3a	• Supervised • Unsupervised Learning • Semi-supervised Learning.	3M 3M 1M	7	CO2
3b	Major applications of machine learning in detail • Sentiment analysis • Recommendation systems • Voice assistants • Google maps	1M 1M 1M	3	CO2
4a	Explain the machine-learning process model with the flow diagram • Understanding the business • Understanding the data • Preparation of data	1M 1M 1M	7	CO2

	• Modelling • Evaluate • Deployment • Flow diagram	1M 1M 1M 1M		
4b	What are the differences between classification and regression with an example? • Classification • Regression	1M 1M	3	CO2
5a	Explain Reinforcement Learning in detail with an example • Definition • Example	3M 3M	6	CO2
5b	Data source can be classified as, • Open or public data source • Social media • Multimodal data	1M 1M 1M 1M	4	CO2
6a	Big Data Analytics and Types of Analytics with examples • Descriptive analytics • Diagnostic analytics • Predictive analytics • Prescriptive analytics	1.5M 1.5M 1.5M 1.5M	6	CO2
6b	Challenges of Machine Learning • 'ill-posed' problems • Huge data • High computation power • Complexity of the algorithms	1M 1M 1M 1M	4	CO2

1a. Processed data is called information. This includes patterns, associations, or relationships among data.

- For example, sales data can be analyzed to extract information like which is the fast selling product.
- Condensed information is called knowledge.
- For example, the historical patterns and future trends obtained in the above sales data can be called knowledge.
- Unless knowledge is extracted, data is of no use. Similarly, knowledge is not useful unless it is put into action.
- Intelligence is the applied knowledge for actions. An actionable form of knowledge is called intelligence.
- Computer systems have been successful till this stage.
- The ultimate objective of knowledge pyramid is wisdom that represents the maturity of mind that is, so far, exhibited only by humans.
-

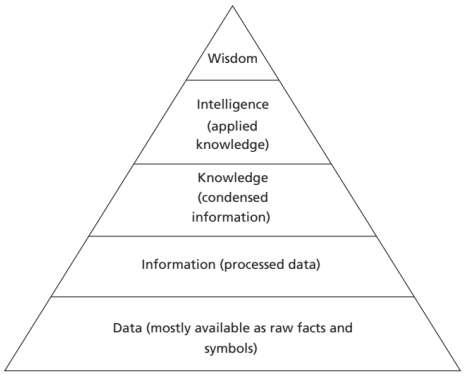


Figure 1.1: The Knowledge Pyramid

- Here comes the need for machine learning. The objective of machine learning is to process these archival data for organizations to take better decisions to design new products, improve the business processes, and to develop effective decision support systems.

1b. Drive the popularity of machine learning

1. High volume of available data to manage- Facebook,Twitter, Youtube, Data Doubled every year
2. The cost of storage has been reduced. The hardware cost has also dropped.
3. The availability of complex algorithms

2a. Types of Data available and explain each with an example

• **In Big Data, there are three kinds of data.**
• **They are structured data, unstructured data, and semi-structured data.**

Structured Data

• In structured data, data is stored in an organized manner such as a database where it is available in the form of a table. The data can also be retrieved in an organized manner using tools like SQL.
• The structured data frequently encountered in machine learning are listed below:

Record Data

A dataset is a collection of measurements taken from a process. We have a collection of objects in a dataset and each object has a set of measurements.
• The measurements can be arranged in the form of a matrix.
• Rows in the matrix represent an object and can be called as entities, cases, or records.
• The columns of the dataset are called attributes, features, or fields. The table is filled with observed data. Also, it is better to note the general jargons that are associated with the dataset.
• Label is the term that is used to describe the individual observations.

Data Matrix

• It is a variation of the record type because
it consists of numeric attributes
• The standard matrix operations can be applied on these data.
The data is thought of as points or vectors in the multidimensional space where every attribute is a dimension describing the object

Graph Data

• It involves the relationships among objects.
• For example, a web page can refer to another web page. This can be modeled as a graph. The modes are web pages and the hyperlink is an edge that connects the nodes.
• Ordered Data
• Ordered data objects involve attributes that have an implicit order among them. The examples of ordered data are:

- Temporal data – It is the data whose attributes are associated with time.
- For example, the customer purchasing patterns during festival time is sequential data.
- Time series data is a special type of sequence data where the data is a series of measurements over time.
- Sequence data – It is like sequential data but does not have time stamps.
- This data involves the sequence of words or letters. For example, DNA data is a sequence of four characters – A T G C.
- Spatial data – It has attributes such as positions or areas. For example, maps are spatial data where the points are related by location.

Unstructured Data

Unstructured data includes video, image, and audio. It also includes textual documents, programs, and blog data.

It is estimated that 80% of the data are unstructured data.

Semi-Structured Data

Semi-structured data are partially structured and partially unstructured. These include data like XML/JSON data, RSS feeds, and hierarchical data.

2b. Elements of Big Data

- Data whose volume is less and can be stored and processed by a small-scale computer is called 'small data'.
- These data are collected from several sources, and integrated and processed by a small-scale computer.
- Big data, on the other hand, is a larger data whose volume is much larger than 'small data' and is characterized as follows:

6V's of Big Data:

- Volume
- Velocity
- Variety
- Veracity
- Validity
- Value
- **Volume –** Since there is a reduction in the cost of storing devices, there has been a tremendous growth of data.
- Small traditional data is measured in terms of gigabytes (GB) and terabytes (TB),
- Big Data is measured in terms of petabytes (PB)(one petabyte is 1,024 terabytes) and exabytes (EB). One exabyte is 1 million terabytes.
- **2. Velocity –** The fast arrival speed of data and its increase in data volume is noted as velocity.
- The availability of IoT devices and Internet power ensures that the data is arriving at a faster rate.
- Velocity helps to understand the relative growth of big data and its accessibility by users, systems, and applications.
- **3. Variety :**
- The variety of Big Data includes:
- • Form – There are many forms of data. Data types range from text, graph, audio, video, to maps.
- There can be composite data too, where one media can have many other sources of data, for example, a video can have an audio song.
- •Function – These are data from various sources like human conversations, transaction records, and old archive data.

- • Source of data – This is the third aspect of variety. There are many sources of data. The data source can be classified as open/public data, social media data and multimodal data.

4.Veracity of data :

- Veracity of data deals with aspects like conformity to the facts, truthfulness, believability, and confidence in data.
- There may be many sources of error such as technical errors, typographical errors, and human errors.
- So, veracity is one of the most important aspects of data.

5. Validity :

- Validity is the accuracy of the data for taking decisions or for any other goals that are needed by the given problem.

6. Value :

- Value is the characteristic of big data that indicates the value of the information that is extracted from the data and its influence on the decisions that are taken based on it.
- The data quality of the numeric attributes is determined by factors like precision, bias, and accuracy.
- Precision is defined as the closeness of repeated measurements.
- Often, standard deviation is used to measure the precision.
- Bias is a systematic result due to erroneous assumptions of the algorithms or procedures.
- Accuracy is the degree of measurement of errors that refers to the closeness of measurements to the true value of the quantity.
- Normally, the significant digits used to store and manipulate indicate the accuracy of the measurement.

3a. Explain Supervised, Unsupervised Learning & Semi-supervised Learning.

- Supervised algorithms use labelled dataset. As the name suggests, there is a supervisor or teacher component in supervised learning.
- A supervisor provides labelled data so that the model is constructed and generates test data.
- In supervised learning algorithms, learning takes place in two stages.
- In layman terms, during the first stage, the teacher communicates the information to the student that the student is supposed to master.
- The student receives the information and understands it.
- During this stage, the teacher has no knowledge of whether the information is grasped by the student.
- This leads to the second stage of learning.
- The teacher then asks the student a set of questions to find out how much information has been grasped by the student.
- Based on these questions, the student is tested, and the teacher informs the student about his assessment.
- This kind of learning is typically called supervised learning.

Supervised learning has two methods:

1. Classification

2. Regression

Classification

- Classification is a supervised learning method.
- The input attributes of the classification algorithms are called independent variables.

- The target attribute is called label or dependent variable.

- The relationship between the input and target variable is represented in the form of a structure which is called a classification model.
- So, the focus of classification is to predict the 'label' that is in a discrete form (a value from the set of finite values).
- An example a classification algorithm takes a set of labelled data images such as dogs and cats to construct a model that can later be used to classify an unknown test image data.

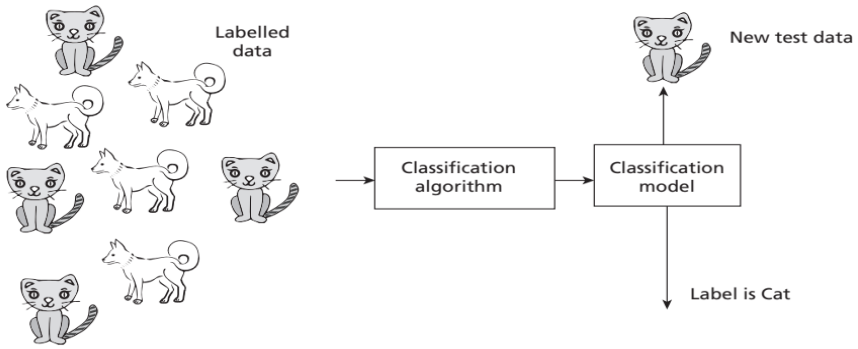
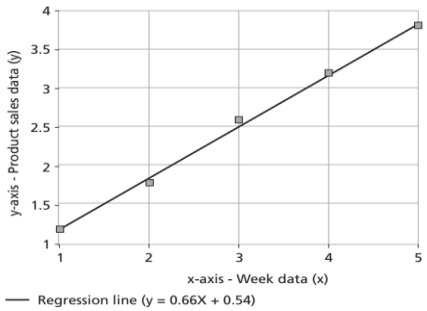


Figure 1.7: An Example Classification System

- In classification, learning takes place in two stages.
- During the first stage, called training stage, the learning algorithm takes a labelled dataset and starts learning.
- After the training set, samples are processed and the model is generated.
- In the second stage, the constructed model is tested with test or unknown sample and assigned a label.
- This is the classification Process
- Similarly, in the case of Iris dataset, if the test is given as (6.3, 2.9, 5.6, 1.8,?), the classification will generate the label for this. This is called classification.
- One of the examples of classification is – Image recognition, which includes classification of diseases like cancer, classification of plants, etc.
- The classification models can be categorized based on the implementation technology like decision trees, probabilistic methods, distance measures, and soft computing methods

Regression Models

- Regression models, unlike classification algorithms, predict continuous variables like price.
- In other words, it is a number.
- A fitted regression model is shown in Figure for a dataset that represent weeks input x and product sales y.
- The regression model takes input x and generates a model in the form of a fitted line of the form y = f(x). Here, x is the independent variable that may be one or more attributes and y is the dependent variable.



- In fig, linear regression takes the training set and tries to fit it with a line – product sales = 0.66 × Week + 0.54
- Here, 0.66 and 0.54 are all regression coefficients that are learnt from data.
- The advantage of this model is that prediction for product sales (y) can be made for unknown week data (x).
- For example, the prediction for unknown eighth week can be made by substituting x as 8 in that regression formula to get y.
- Both regression and classification models are supervised algorithms.
- Both have a supervisor and the concepts of training and testing are applicable to both.

The main difference is that

- **Regression models predict continuous variables such as product price.**
- **Classification concentrates on assigning labels such as class.**

Unsupervised Learning

- The second kind of learning is by self-instruction.
- There are no supervisor or teacher components.
- In the absence of a supervisor or teacher, self-instruction is the most common kind of learning process.
- This process of self-instruction is based on the concept of trial and error.
- Here, the program is supplied with objects, but no labels are defined.
- The algorithm itself observes the examples and recognizes patterns based on the principles of grouping.
- Grouping is done in ways that similar objects form the same group.
- Cluster analysis and Dimensional reduction algorithms are examples of unsupervised algorithms.
- Cluster analysis is an example of unsupervised learning.
- It aims to group objects into disjoint clusters or groups.
- Cluster analysis clusters objects based on its attributes.
- All the data objects of the partitions are similar in some aspect and vary from the data objects in the other partitions significantly.
- Some of the examples of clustering processes are — segmentation of a region of interest in an image, detection of abnormal growth in a medical image, and determining clusters of signatures in a gene database.

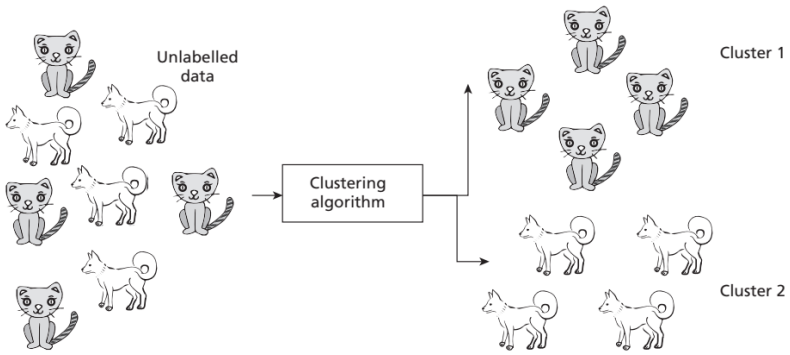


Figure 1.9: An Example Clustering Scheme

Some of the key clustering algorithms are:

- **k-means algorithm**
- **Hierarchical algorithms**

Dimensionality Reduction

- Dimensionality reduction algorithms are examples of unsupervised algorithms.
- It takes a higher dimension data as input and outputs the data in lower dimension by taking advantage of the variance of the data.
- It is a task of reducing the dataset with few features without losing the generality.

Semi-supervised Learning

- There are circumstances where the dataset has a huge collection of unlabelled data and some labelled data.
- Labelling is a costly process and difficult to perform by the humans.
- Semi-supervised algorithms use unlabelled data by assigning a pseudo-label.
- Then, the labelled and pseudo-labelled dataset can be combined.

3b. Major applications of machine learning

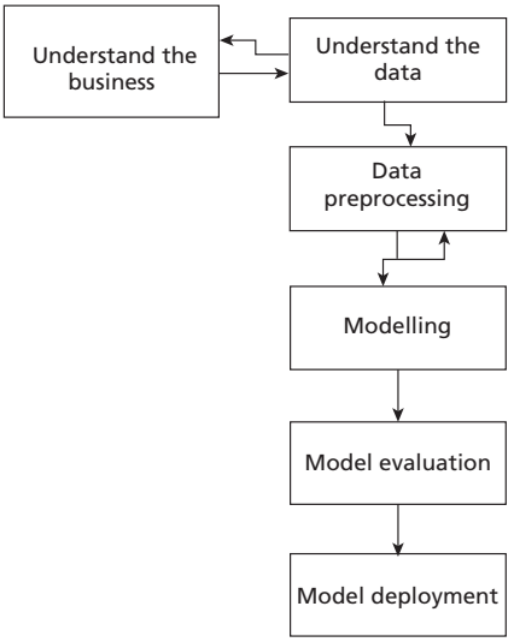
Some applications are listed below:

- **1. Sentiment analysis –** This is an application of natural language processing (NLP) where the words of documents are converted to sentiments like happy, sad, and angry which are captured by emoticons effectively.
- **For movie reviews or product reviews, five stars or one star are automatically attached using sentiment analysis programs.**
- **2. Recommendation systems –** These are systems that make personalized purchases possible. For example, Amazon recommends users to find related books or books bought by people who have the same taste like you, and Netflix suggests shows or related movies of your taste.
- **The recommendation systems** are based on machine learning.

- **3. Voice assistants –** Products like Amazon Alexa, Microsoft Cortana, Apple Siri, and Google Assistant are all examples of voice assistants.
- They take speech commands and perform tasks. These chatbots are the result of machine learning technologies.

4. Technologies like Google Maps and those used by Uber are all examples of machine learning which offer to locate and navigate shortest paths to reduce time.

4a. Machine-learning process model with the flow diagram



- **1. Understanding the business –** This step involves understanding the objectives and requirements of the business organization. Generally, a single data mining algorithm is enough for giving the solution. This step also involves the formulation of the problem statement for the data mining process.
- **2. Understanding the data –** It involves the steps like data collection, study of the characteristics of the data, formulation of hypothesis, and matching of patterns to the selected hypothesis.
- **3. Preparation of data –** This step involves producing the final dataset by cleaning the raw data and preparation of data for the data mining process. The missing values may cause problems during both training and testing phases. Missing data forces classifiers to produce inaccurate results. This is a perennial problem for the classification models. Hence, suitable strategies should be adopted to handle the missing data
- **4. Modelling –** This step plays a role in the application of data mining algorithm for the data to obtain a model or pattern.
- **5. Evaluate –** This step involves the evaluation of the data mining results using statistical analysis and visualization methods. The performance of the classifier is determined by evaluating the accuracy of the classifier. The process of classification is a fuzzy issue. For example, classification of emails requires extensive domain knowledge and requires domain experts. Hence, performance of the classifier is very crucial.
- **6. Deployment –** This step involves the deployment of results of the data mining algorithm to improve the existing process or for a new situation.

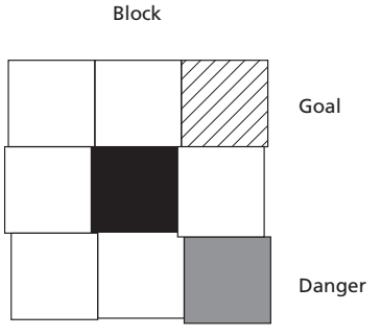
4b. Differences between classification and regression with an example

- **Regression models predict continuous variables such as product price.**
- **Classification concentrates on assigning labels such as class.**

5a. Reinforcement Learning in detail

- Reinforcement learning mimics human beings.
- Like human beings use ears and eyes to perceive the world and take actions, reinforcement learning allows the agent to interact with the environment to get rewards.

- The agent can be human, animal, robot, or any independent program.
- The rewards enable the agent to gain experience.
- The agent aims to maximize the reward. The reward can be positive or negative (Punishment). When the rewards are more, the behavior gets reinforced and learning becomes possible.
- Consider the following example of a Grid game
- In this grid game, the gray tile indicates the danger, black is a block, and the tile with diagonal lines is the goal.
- The aim is to start, say from bottom-left grid, using the actions left, right, top and bottom to reach the goal state.



- To solve this sort of problem, there is no data.
- The agent interacts with the environment to get experience.
- In the above case, the agent tries to create a model by simulating many paths and finding rewarding paths.
- This experience helps in constructing a model. It can be said in summary, compared to supervised learning, there is no supervisor or labelled dataset.
- Many sequential decisions need to be taken to reach the final decision.
- Therefore, reinforcement algorithms are reward-based, goal-oriented algorithms.

5b. Data source can be classified, explain its types in detail

The data source can be classified as open/public data, social media data and multimodal data.

Open or public data source – It is a data source that does not have any stringent copyright rules or restrictions.

Its data can be primarily used for many purposes.

Government census data are good examples of open data:

Digital libraries that have huge amount of text data as well as document images

Scientific domains with a huge collection of experimental data like genomic data and biological data

Healthcare systems that use extensive databases like **patient databases, health insurance data, doctors' information, and bioinformatics information**

2. Social media – It is the data that is generated by various social media platforms like Twitter, Facebook, YouTube, and Instagram.

An enormous amount of data is generated by these platforms.

3. Multimodal data – It includes data that involves many modes such as text, video, audio and mixed types.

6a. Big Data Analytics and Types of Analytics with examples

- The primary aim of data analysis is to assist business organizations to take decisions.
- For example, a business organization may want to know which is the fastest selling product, in order for them to market activities.

- Data analysis is an activity that takes the data and generates useful information and insights for assisting the organizations.
- Data analysis and data analytics are terms that are used interchangeably to refer to the same concept. However, there is a subtle difference.
- Data analytics is a general term and data analysis is a part of it.
- Data analytics refers to the process of data collection, preprocessing and analysis. It deals with the complete cycle of data management.
- Data analysis is just analysis and is a part of data analytics. It takes historical data and does the analysis.
- It takes historical data and does the analysis.
- Data analytics, instead, concentrates more on future and helps in prediction.

There are four types of data analytics:

- **1. Descriptive analytics**
- **2. Diagnostic analytics**
- **3. Predictive analytics**
- **4. Prescriptive analytics**

Descriptive Analytics

- It is about describing the main features of the data.
- After data collection is done, descriptive analytics deals with the collected data and quantifies it.
- It is often stated that analytics is essentially statistics.
- There are two aspects of statistics – Descriptive and Inference. Descriptive analytics only focuses on the description part of the data and not the inference part.
- What was our overall productivity?

Diagnostic Analytics

- It deals with the question – 'Why?'. This is also known as causal analysis, as it aims to find out the cause and effect of the events.
- For example, if a product is not selling, diagnostic analytics aims to find out the reason. There may be multiple reasons and associated effects are analyzed as part of it.
- Why did our company sales decrease in the previous quarter?

Predictive Analytics

- It deals with the future. It deals with the question – 'What will happen in future given this data?'.
- This involves the application of algorithms to identify the patterns to predict the future.
- The entire course of machine learning is mostly about predictive analytics and forms the core of this book.
- Predicting maintenance issues, Predicting article popularity

Prescriptive Analytics

- It is about the finding the best course of action for the business organizations.
- Prescriptive analytics goes beyond prediction and helps in decision making by giving a set of actions.
- It helps the organizations to plan better for the future and to mitigate the risks that are involved.
- Automatic adjustment of product pricing based on customer demand and external factors.

6b. Challenges of Machine Learning

Computers are better than humans in performing tasks like computation.
For example, while calculating the square root of large numbers, an average human may blink but computers can display the result in seconds.

Computers can play games like chess, GO, and even beat professional players of that game.

- **Problems –** Machine learning can deal with the 'well-posed' problems where specifications are complete and available. Computers cannot solve 'ill-posed' problems.

Input (x_1, x_2)	Output (y)
1, 1	1
2, 1	2
3, 1	3
4, 1	4
5, 1	5

Can a model for this test data be multiplication? That is, y = x1 × x2. Well! It is true! But, this is equally true that y may be y = x1 ÷ x2, or y = x1^ x2. So, there are three functions that fit the data. This means that the problem is ill-posed. To solve this problem, one needs more example to check the model. Puzzles and games that do not have sufficient specification may become an ill-posed problem and scientific computation has many ill-posed problems.

2. Huge data – This is a primary requirement of machine learning. Availability of a quality data is a challenge.

A quality data means it should be large and should not have data problems such as missing data or incorrect data.

3. High computation power – With the availability of Big Data, the computational resource requirement has also increased.

Systems with Graphics Processing Unit (GPU) or even Tensor Processing Unit (TPU) are required to execute machine learning algorithms.

Also, machine learning tasks have become complex.

4.Complexity of the algorithms – The selection of algorithms, describing the algorithms, application of algorithms to solve machine learning task, and comparison of algorithms have become necessary for machine learning or data scientists now.

Algorithms have become a big topic of discussion, and it is a challenge for machine learning professionals to design, select, and evaluate optimal algorithms.

5. Bias/Variance – Variance is the error of the model. This leads to a problem called bias/ variance tradeoff.

A model that fits the training data correctly but fails for test data, in general lacks generalization, is called overfitting.

The reverse problem is called underfitting where the model fails for training data but has good generalization.