1.Explain the term machine learning, and how does it work? Explain two machine learning applications in the business world. What are some of the ethical concerns that machine learning applications could raise?

Machine learning is the field if study where computers are given the ability to “learn” which broadly means giving them the ability to understand data and make predictions and decisions based on that.

ML uses complex algorithms that enable the machine to detect patterns and relationships in the input data. In case of Supervised learning techniques, sample input and output data is provided from which an approximation of the relationship is derived, this is then used to predict an approximate output of new input data. For unsupervised learning where only raw data is available, patterns and similarities within the datapoints are detected by the algorithm which is then used to classify new data into the existing classes in case of problems such as classification.

In the current business world, ML holds great importance as it is used for various purposes.

1. Consumer recommendation Systems: E-commerce marketplaces are common today and these rely heavily on recommendation systems to provide their customers personalized recommendations. These recommendations allow the customer the ability to find things of interest much more easily than having to search manually through millions of products and thus persuade them to spend more on the platform. This directly leads to more sales and hence drives business growth of the platform
2. Fraud detection systems: ML can be used to monitor the usual behaviour of customers like credit card usage patterns, and notify in case of any unusual behaviour which can signal to a fraudulent transaction. These systems can be very quick to detect any unusual behaviour and are used by a variety of businesses to prevent frauds

ETHICAL CONCERNS

Machine learning is a powerful tool that gives the user the ability to gain insights on large amounts of data in a very easy manner. With this great power, comes great ethical concern about its responsible use. Many companies nowadays require a lot of personal information in order to use their services. Many online services that command a large userbase like e-commerce and social media platforms have a lot of data about the behaviour of a user like their spending habits, hobbies, daily life, travel locations etc. Here are a couple common examples misuse:

Big social media platforms now have the ability to influence a large number of people through their recommendation engines. This can easily be used to deliver a specific type of content to a target demographic which can influence their views on certain topics. These can and are regularly used by political organizations and governments to manipulate a large group of users quite easily.

E-commerce sites can push specific products more than others which leads unfair playing field for companies – mostly smaller ones. As an example, Amazon has been accused of this behaviour often as they push their own amazon branded products as higher in search results regardless of their quality and reviews. This leads to smaller companies, that may actually have a better product, not getting proper recognition.

2. Describe the process of human learning:

I. Under the supervision of experts

Human learning under the supervision of experts is essentially the education system that involves the traditional schools where a teacher shares their knowledge with students. They help them learn and provide guidance on the process. Students can learn from books, practical and other means and the teacher is there guide them along the process helping them in overcoming difficulties and providing and advanced or different perspective to a problem

ii. With the assistance of experts in an indirect manner

Indirect assistance of experts can be equated to the self-paced learning that is quite popular these days. A student can have access to learning resources and go about the learning at their own pace. Resources like expert video lectures, articles and notes can help a student overcome difficulties in this scenario. Here, instead of a teacher directly helping and guiding a student, this indirect help is provided

iii. Self-Education

Self-education in terms of human learning can mean learning through experimentation and experience. A person has to develop their own understanding without external help. An example of this can be scientific research where many new areas are explored for which prior information can simply not exist. In such a case we develop our won understanding of something by performing experiments and studying it to improve our understanding.

3. Provide a few examples of various types of machine learning.

a. Supervised Learning:

Classification problem: Outlier Detection, Image classification, Optical Character Recognition

Regression problems: Housing prices Prediction, predicting weather temperatures, determining travel charges

b. Unsupervised Learning: Astronomical data analysis for detection star clusters or entirety new entities, organization of computing clusters, Image recognition

c. Reinforcement Learning: Autonomous Driving, Image Processing, Natural language processing

4. Examine the various forms of machine learning.

a. Supervised Learning: This type of machine learning deals with labelled data. Here sample input data is provided along with the output for the algorithm to determine the relationship between the input features and the output. This can further be classified into two – Classification and regression

Classification problems require datapoints to be separated into classes. This can be binary or multiclass depending on the problem

Regression problems require prediction of continuous values like prediction of housing prices, predicting weather temperatures or determining travel charges

b. Unsupervised Learning: When dealing with unlabelled input data, this type technique is helpful in extracting similarities and hidden patterns in the data which can allow us to create classes for the datapoints. We can further classify unsupervised learning into clustering and association.

Clustering is the grouping of datapoints into clusters containing similar data. It finds common attributes from the data and groups the ones with most commonalities together.

Association works by finding relationships between variables and determines those that appear common often. This can be explained by the example of e-commerce sites like amazon that provide recommendations for “Frequently bought together” items. It recognizes a trend of people buying specific products together and suggests other buyers of the same.

c. Reinforcement Learning: This type of learning is reminiscent of human learning and how human psychology works. The algorithm works based on a hit and trial method and a positive or negative feedback is given on an action. The algorithm takes this feedback and tried to maximize the positive feedback and the algorithm is forced to reiterate in case of a negative feedback. This type of learning is seen solving problems like calculation of shortest route on a map, or training a robot to walk on varying terrain and providing positive and negative feedback in case of a desired or undesired behaviour and natural language processing.

5. Can you explain what a well-posed learning problem is? Explain the main characteristics that must be present to identify a learning problem properly.

By definition, A well posed learning problem is – A computer program is said to learn from experience E, while performing a task T with some performance metric P. If it performs T with a performance measure P, then it upgrades with E.

From this we can, extract three most important components of a well posed learning problem:

* Task
* Performance Metric
* Experience

This can be better visualized using some examples:

1. Facial Recognition

This system is capable of recognizing different faces and identify them based on the database. It works by locating and measuring facial characteristics.

Here,  
T – Recognizing a face and identifying it

P – Ability to differentiate between different types of faces and accurately identifying it against the available database

E – Training with as many different variations of faces as possible

1. Email Spam Filter

This system analyses the incoming emails by going through its contents and identifying any components that point to it being a spam mail

T – Identifying an incoming email as spam or not

P – Ability successfully detects an email as spam with lease amount of false positives and spam slippage into inbox

E – Training against various kinds of spam mail and improving with manual identification of an email as spam or not by the user

6. Is machine learning capable of solving all problems? Give a detailed explanation of your answer.

For most people, Machine learning seems like THE solution to all problems of the mankind. The perception on AI/ML of general public is formed largely by sci-fi movies, viral videos and articles. An interaction with someone in my own family came as a surprise to me where they stated that in the next 50 years AI will just take over humans like we see in the Movies and they are convinced of it.

However, ML, as of now, is not a universal solution to all problems. There are many inherent limitations to the methods we use for ML that prevent it from being a universal solution.

Ethics:

One of the most widely talked about limitations of ML are Ethics. A ML model can learn, it can detect patterns in data, predict future outcomes based on previous data, make decisions and get better at it based on feedback, all of this is based on strictly binary logic. It strictly does what the instructions tell it to do. A navigation program will choose the shortest route to your destination with what information it has, it can be programmed to look for traffic conditions, blockages etc. But it cannot decide the best route according to you. People have driven to obscure locations, lost in the maze of small roads etc while blindly trusting the navigation systems.

Self-driving vehicle systems like tesla’s have had multiple instances of accidents, even harming children due to inability to judge. In cases like this who is to blame in case this. The company advertises this as “self-driving” but the driver is to be attentive and in control at all times but in case of the fully self-driving systems of the future we imagine, it’s a dilemma of responsibility in case of accidents. The algorithm does not feel bad when it may hit a tree, another person or a child.

Privacy:

In this age where we use cloud storage and social media as much as we do, the data collection is enormous and the algorithms just do what they are programmed to do. An AI does not know what content in your feed is damaging or untrue. The only thing it focuses on is what you want to watch and increase your watch time and retention on the platform. This is what it is designed to do. Short from content is on the rise now with videos spanning only a few seconds that evoke instant emotions. The AI sees that this is what interests you and feed more of it to you. People complain that their social feeds now are irrelevant and garbage. However, in reality most people watch these videos more than any long form informative educational video, this is what the AI saw and pushed more for.

One of the big controversies in the past year was Apple announcing that all photos, public or private in a person’s apple cloud account will be scanned for child pornography and authorities will be notified in case of any suspicions. This was a good cause but comes with so many ethical questions. People were not given a choice to opt in to it, it would be redundant anyway, makes sense. However, can we allow a large organization like apple to have access to all our private photos, it seems to be for a good cause but we can be certain that this invasion of privacy have a high chance of increasing. A false positive in case of a system like this has the possibility of ruining the life of a person. They can be forced to reveal information that doesn’t necessarily need to be revealed to prove their innocence. The company claims that all this is done on device and only suspicious photos will be flagged and sent for review. Again, the case of false positives arises, how many of your private images will be sent to the company for review, the owner has no control over all this. After much backlash this was rolled back as of now, but it started the conversation, how can we trust the AI to make good decisions on such delicate matters where the context of the moment and feelings need to be considered.

Data:

Lack of data is an issue with many ML algorithms that require large amounts of data to train before they are of use in the real world. Neural networks are examples of algorithms that require enormous amounts of training data and with increasing architecture size the required data also increases.

Even if you have a large amount of data, the quality of it is also of great importance. If the data quality is not good, the model trained on it, despite having enough amount of training data, will not yield good result in real world conditions. Similarly, lack of good features, lack of good real data can cause a model to perform poorly.

Another issue is of a situation, a model trained with data from one situation will not necessarily give good results in another. An example of this is the breast cancer prediction models. Most of the widely used models were trained on a large number of x-rays, but the problem was that most of these were from white women. It was found that black women were more likely to be suffering from breast cancer and the models used for detection were not as effective on this demographic.

Regulatory Restrictions:

With the growing use of AI, combined with the privacy and ethical concerns as we saw before, many governing bodies across the world have begun making and enforcing laws that determine fair use of data. At a large scale, a large amount of data is required and here the various laws can interfere. This is necessary for the protection of privacy of users but is limiting for the algorithms in terms of the data that can be used and hence cause them to be not as effective as they potentially can be.

In conclusion, ML has a lot of potential and has significantly accelerated the development of our world. The amount of data processing that is possible now would be simply be humanly impossible due to the amount of time and human error that would otherwise be involved. But it is not yet the ultimate problem solver that many believe it to be.

7. What are the various methods and technologies for solving machine learning problems? Any two of them should be defined in detail.

Machine learning has a variety of uses and various methods for different scenarios. Here are some of the methods used for an ML problem:

1. Regression
2. Classification
3. Clustering
4. Neural networks
5. Reinforcement Learning
6. Natural Language processing
7. Deep learning
8. Dimensionality reduction
9. Ensemble methods

Creating an ML model from scratch is a huge task and requires a great deal of mathematical knowledge. To make this process faster and more accessible, over the years various technologies have been created like:

1. TensorFlow
2. Keras
3. Scikit-Learn
4. Microsoft Cognitive Toolkit
5. Torch
6. Theano

These technologies have been created with the contribution on large tech companies as well as independent developers in case of open-source ones. These have helped tremendously in accelerating the growth of ML in recent years by making them much more accessible. Here is a detailed description of some of them:

1. Scikit-Learn:

Also Known as “sklearn” this is an open-source machine learning library for Python programming language. It contains various methods for tackling ML problems like classification, regression, clustering, model selection, pre-processing and dimensionality reduction. It is distributed under the New BSD License. It integrates well with many other python libraries and is also the foundation of other more advanced and focused libraries.

1. TensorFlow

TensorFlow is an open-source software library for machine learning and artificial intelligence. It has a particular focus on training and inference of deep neural networks. This was developed the Google Brain team for internal use initially and was released publicly in 2015 under the Apache License 2.0. It has support for various programming languages with stable Python and C++ APIs.

8. Can you explain the various forms of supervised learning? Explain each one with an example application.

Supervised learning can be mainly divided into two major types: Classification and regression.

Classification is used for dividing datapoints into classes. Such algorithms work on categorical variables where defined categories like male-female, true-false, age group, income group like variables are the output. Some of the widely used classification algorithm are KNN, Naïve Bayes, SVM and Decision trees

An example of classification problem can be taken from the iris database that contains various features of flowers like petal length, petal width, sepal length and sepal width. This database can be used to teach an algorithm to differentiate between the various flower types based on the features of the flower and identify the type of flower by matching the new values to its learning from the test database.

Regression on the other hand is useful for continuous variables. Here the predicted values need to be more precise. This is used for problems like predicting weather, market trends, potential business expenses etc. Algorithms like Linear Regression, Regression Trees, Non-Linear Regression, Bayesian Linear Regression and Polynomial Regression are some of the widely used for solving regression problems.

The housing price prediction is a good example of a regression where the price of a property is to be determined based on the various features like Area, Number of rooms, Number of stories, Location of the property and available facilities like air conditioning and water supply. Here the estimated values are continuous as we are to estimate the price of a house which will inherently be different for different houses.

9. What is the difference between supervised and unsupervised learning? With a sample application in each region, explain the differences.

The fundamental difference between a supervised and an unsupervised learning problem is the kind of data used. Supervised requires labelled data with distinct features and an output variable so the algorithm can establish a relation between them. Unsupervised on the ither hand utilizes unlabelled data and attempts to find hidden patterns in the data and categorize them according to the specified parameters

Considering the example of housing price prediction, we have data with clearly labelled features and a price which is our output variable. Here the requirement is to find the best possible representation of the house prices based on the various factors like the location, number of rooms, condition and age of the house etc. Here the goal is to find a relation between the various features and the output variable and not the features themselves.

If we remove the prices from consideration and instead try to determine the kinds of houses we have based on those features, this can turn into an unsupervised learning problem where we try to find the types of hoses we have and categorize them into different levels of luxury. Here the goal of the algorithm is to determine the similarities between various datapoints and how with another variable. Unsupervised learning methods like K-mean can classify this data and do the desired task.

10. Describe the machine learning process in depth.

The process of machine learning can be divided majorly in to seven steps:

1. Data Gathering:

This is the first step in the machine learning process. Data needs to be gathered from various sources and identify any problems with it. The data can be from various different sources like varying databases and devices. The quality and quantity of the data determine the efficiency of the model after training which makes this a very important step.

We need to identify the various data sources, collect data from the said sources and integrate the data gathered into a dataset

1. Data Preparation:

In this step we need to analyse our data for any problems like missing, null or outlier values. We try to understand the features present and understand their nature like the skewness of various features, collinearity, difference in value ranges between the features etc.

The raw data we have gathered can contain bad data like missing or null values, outlier values, invalid data and noise that can reduce the effectiveness of the model. If required, we also need to split the data into training, testing and validation for the respective purposes.

We can then use techniques like standardization, dimensionality reduction to prepare the data better for the next step.

1. Choosing A Model:

After we have prepared the data, we need to choose a model that the data will be fed to. There are many different kinds of ML models available to choose from depending on the task we need to perform.

Depending on the problem, the task can be related to Classification, Regression, Clustering association etc. We need to determine the model we will be using based on the type of data, desired efficiency of the model, the constraints like time, computing hardware and power required for it etc

1. Training:

This step involves the training the model with the training data the so it can learn from. This step can be expensive in terms of time and computing power depending on conditions like the amount of data and computing power available.

1. Evaluation:

Once the model training is done, we need to evaluate its performance on unseen data. A model will probably give great results on the data it was trained with but with training data we evaluate its performance with new values. The training data already had an expected known output value and the evaluation is done by comparing how frequently the model gives a right prediction.

The results we obtain from this step determine how we move forward with our process.

1. Hyperparameter Tuning:

Here we try to find the best values for the hyperparameters of the model. This method is a bit of a hit and trial method and the outcomes are seen by trying out various values of the hyperparameters and re-do the training and evaluation step.

Once satisfied with the outcome of the evaluation, we can move to the final deployment of the model.

1. Deployment:

The final step in the process is Deployment where we put out trained model to work with real world data.

In this stage we determine if the accuracy and speed of the system is to our requirement. This monitoring is done at regular intervals and, if needed, the model can be retrained with new data, retuned and re-evaluated and re-deployed

a. Make brief notes on any two of the following:

i. MATLAB is one of the most widely used programming languages.

MATLAB is an abbreviation for Matrix Laboratory. It is a proprietary multi-paradigm programming language that is matrix manipulations, function plotting, algorithm implementation, creating user interfaces and interfacing with other programming Languages. The operations performed in MATLAB are similar to the ones performed by the NumPy library for python.

MATLAB is used in a variety of use cases like chemical engineering where it can be used to simulate a chemical reaction, visualize relationship between the various reactions, simulation of a machine under operating conditions etc. It has various uses in science, engineering and economics with over 4 million users over the world.

ii. Deep learning applications in healthcare

Deep learning is widely increasing in popularity for healthcare applications. Healthcare provides some very complex and highly sensitive problems for AI to tackle. DL has proven to be of great help in different areas of healthcare from patient care, research as well as for insurance purposes.

Image recognition and object detection are used for diagnosis using MR and CT scans for image segmentation, disease detection and prediction. These algorithms are capable of tissue analysis and can flag important images for further analysis. These are also useful for early detection of an issue in certain cases which would have been undiagnosed until a much later stage using previous techniques. For example, DL is used for early detection of Alzheimer’s or ultrasound breast nodules

Human body can produce a lot of data and complex data analysis is made much easier using Convoluted Neural Networks (CNNs). According to a 2018 study, these CNNs diagnosed melanoma with 10% to 7% more accuracy than experts even with patient history.

The increasing popularity of health measurement wearables is another area where the algorithms can use the electronic health records and, in many cases, have provided the users early warnings of a conditions like heart attacks. These also now come with features like automatic alert to the emergency services in case of an accident which can save lives with a faster notification

For research and development, researchers use DL to analyse complex combinations of data for purposes of drug discovery where it can analyse the genomic, clinical and population data much faster and more efficiently which has accelerated the development times.

Better understanding on highly complex areas like DNA sequencing, genome based development is also now possible and many previously unknown insights are new being discovered which can improve the overall understanding of biology.

Deep learning has also been of great use in the recent Covid-19 outbreak where it has been used for analysis of chest x-rays and population data which has provided on field benefits like early detection of the disease, predicting the ICU admission, finding potential high risk individuals etc

11. Make a comparison between: -

* 1. Generalization and abstraction:

Generalization in ML is the ability of a trained model to give good results with previously unseen data. This requires prevention of overfitting and underfitting of data which will lead to poor performance with unseen data.

Abstraction on the other hand is a fundamental object-oriented programming concept as seen across different languages. This is used to make the program more modular and increase its security. It can easily allow for further expansion of the program without having to expose sensitive data to other parts of a program

* 1. Learning that is guided and unsupervised:

Guided or Supervised learning requires using labelled input and output data. Here the feature and output variables of the data are clearly defined and the algorithm’s job is to approximate the relation between the features and the output variable in order to predict the outcome of unseen data. This is broadly divided into Classification and Regression for problems relating to Categorical and Continuous variables respectively

Unsupervised learning, in contrast, is used for working with unlabelled data where the goal is to gain insights about the data and not prediction of a certain output. This is further divided into Clustering and Association type problems where the goal of the algorithm is to find insights from the data like hidden patterns and similarities in order to divide it into clusters or find relationships between various features.

* 1. Regression and classification

Regression problems deal with data containing continuous values for the output variable. This can be for example House prices, temperatures of a certain region etc.

Classification deals with categorical output variables where the data is to be divided into distinct categories for example types of flower, true or false, type of car etc