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 a subset of Artificial Intelligence which provides mathematical tools to analyze data, do Exploratory Data Analysis, make predictive models and exposed APIs and more. Machine Learning works by providing the computer capability to make predictions or help analysis of data by creating complex algorithmic models that are trained to work within constraints that have to be specified in training data. Ethical concerns that machine learning applications can raise are data related. As more and more data hungry applications are being created, the fine line between ethical handling and collection of data from various sources may blur. For this reason, we need sanity checks in place.\***

**Applications:**

**a. Learning the overall sentiment of consumers regarding a product launched a while ago using their reviews on an e-commerce website using Natural Language Processing to understand how the product is faring in the market**

**b. Using machine learning algorithms to cluster similar customers and provide them recommendations based on their history, on a streaming service platform.**

2. Describe the process of human learning:

i. Under the supervision of experts

**Under supervision, learning is defined by guidance provided to the learner. This guidance or supervision trains the learner to solve certain type of problems. Problems can range from formal communication, to mathematical calculations. Pre-calculated answers provide for rectifications and course correction during learning.**

ii. With the assistance of experts in an indirect manner

**Under indirect assistance, learning is defined by guidance provided to the learner through indirect means. These indirect means can range from books, to periodic advise from experts. This guidance or supervision trains the learner to solve certain type of problems but not as easily as under supervision. Problems can range from formal communication, to mathematical calculations. Expert indirect guidance provide for only limited rectifications and course correction during learning.**

iii. Self-education

**Under no supervision, learning is defined by self guidance provided to the learner by himself. The learner has the entire burden of finding ways to solve certain type of problems. Problems can range from formal communication, to mathematical calculations. Only self-calculated answers provide for rectifications and course correction during learning.**

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

**a. Supervised machine learning- Logistic Regression, Random Forest Classifier, Naive Bayes Classifier**

**b. Unsupervised machine learning- Heirarchichal clustering, DBSCAN algorithm, K means and K medoids algorithms .**

**c. Semi-supervised machine learning- Associative reinforcement learning, deep reinforcement learning.**

4. Examine the various forms of machine learning.

**Supervised Machine Learning is a form of machine learning which involves target variable as labeled data, and independent variables. Using labeled dataset, training is supervised and predictions are made.**

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.

**If a machine learning model learns from certain type of actions that it can record, to do some specific task after evaluating itself using an evaluation metric, then it would be called a well posed learning problem**

**The main characteristics that must be present are:**

**a. A task that has to be done by the machine learning model**

**b. An action that has to be recorded and learned from by the model**

**c. Evaluation metric using which the efficacy of the machine learning model can be measured.**

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

**No. Machine learning cannot solve all problems. It can only be used to solve well defined objectives that are not vague. Objective solving capability is also limited by the statistical tools and mathematical algorithms that exist today.**

**A problem like predicted the chances of terminal cancer in a patient or classifying a spam mail is still in the domain of well defined objectives. Trying to solve objectives like eradicating world wide hunger, creating actual medicine for all kinds of cancers or, reanimating a cryogenically frozen animal through machine learning is not possible. Some portions of the bigger problems however, can fall in the range of machine learning.**

**Predicting the food shortage world might face in the next 5 years can be done using machine learning. Using this prediction, world leaders can formulate a strategy to counter food shortage.**

**Diving down to the cellular level and using computer vision to predict which kind of cell may turn cancerous using a sample of different cells, based on lifestyle, current health and other factors might be possible using machine learning.**

**So, Machine Learning is not the solution to every problem. But it can certainly ease some of the constraints by solving them for us, or at least a portion of the problem.**

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

**The various methods and technologies for solving machine learning problems are->**

**a. Ensemble methods**

**b. Regression**

**c. Classification**

**d. Reinforcement learning**

**Models that incorporate the use of multiple base learner models that work simultaneously to produce results are ensemble models. A key aspect of ensemble models is that more different the base learners are, more powerful is the combination and hence more accurate predictions. There are 4 types of ensemble model**

**Bagging which is also called Bootstrap aggregation. Multiple base learners learn on subsets of data. A majority vote based on the results of the individual base learners is then used to output the final result.**

**Boosting is mainly used to reduce bias in model. Multiple base learners are trained in a sequence in which each learner covers up the weakness of the previous learner to ultimately produce a powerful model.**

**Stacking is an ensemble model that combines multiple base learner models constructed parallel to each other and independent. All the models produce predictions that are used creating a meta classifier which is trained on the predictions made by base models. The resultant prediction is then taken as the final prediction. Space and Time complexity issue is huge for stacking models.**

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

**The Various forms of Supervised Learning are->**

**Logistic Regression->**

**It is a binary classification algorithm based on linear regression. A hyperplane is fit using a mathematical equation such that most of training data points are correctly classified into two class labels. Lambda is used as the hyperparameter. Lambda=0 would lead to overfitting and large lambda would lead to underfitting. Logistic Regression can be easily applied for classifying whether a voter would vote for BJP or Congress based on factors like number of visits by politicians, work done in their neighborhood etc.**

**Decision Tree->**

**It is a multi class classifier that creates multiple hyperplanes to funnel down the data points based on criteria and classify data points that pass certain threshold/criteria and do not. In a different sense, it creates up side down trees and keeps splitting data points for each criteria they pass and do not pass. Decision Tree is based on the concept of maximising the information gain from splitting the data points using criterias. Depth is the hyperparameter for Decision Trees and needs to be tuned. Large depths can lead to overfitting while small depths lead to underfitting. Classifying the data points using each passed criteria for Petal length, petal width, sepal length and sepal width can be done for IRIS dataset.**

**K-NN->**

**K nearest neighbors requires nearest instances or similar instances for a data point to be successfully classified. Normally, training data is learned for model generation. But, here training examples are stored. Distance function is used to find closest examples to an instance so that successful prediction can be made. K-NN can be used for recommendation systems where the input instance is checked for closest training example and then data point is classified into a group for which recommendations can be made.**

**Linear Regression->**

**It is a regression algorithm that tries to fit a straight line over data points in a 2-D plane where the two variables involved are highly correlated with each other and are independent and not collinear. Main idea is to keep the root mean squared error values(distances between the line to be fit and the training data points) minimum. An HR department of an organization can use the current experience of a professional they wish to interview, to get an idea about how much salary can be given to him/her as salaries are highly positively correlated with experience.**

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

**Supervised machine learning is done using labeled data whereas unsupervised learning has no labeled data.**

**Classification of flowers in IRIS dataset is a very common application of supervised machine learning. The dataset has 4 features: Petal length, Petal width, Sepal length and Sepal width. Target variable contains labels Virginica, Versicolor and Setosa. If we use Decision Trees algorithm, then labeled data is necessary for criterias or thresholds to be built that have to be either passed or failed by values of independent variables. The class labels are known beforehand and come from a finite set.**

**Clustering of customers based on their previous purchases on e-commerce websites requires unlabeled data. If k means algorithm is used then based on calculations of centroids, data points are classified into clusters. Since the class labels were not known beforehand but did come from a finite set, number of clusters initialized becomes highly important.**

10. Describe the machine learning process in depth.

**First step is to specify the problem statement we wish to solve**

**Next, we collect data relevant to our problem statement. Data collection can be through webscraping, surveys, company records**

**Data has to be cleaned and manipulated, null values removed or imputed. Cleaned data also needs to be split into training and test set so that it is usable by machine learning model for learning phase.**

**We train the machine learning model over the collected and cleaned data. Training can be supervised or unsupervised, depending upon the problem statement and the availability of labeled data.**

**Next, we test the trained machine learning model using performance metrics based on the type of problem we are solving and the algorithm we used. For Regression, we may use the R-squared value. For classification, we can use log losses, recall and precision scores etc. If unsupervised model is built, we might use the DB Index.**

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

MATLAB is one of the most widely used programming languages.

ii. Deep learning applications in healthcare

iii. Study of the market basket

iv. Linear regression (simple)

**MATLAB short for Matrix Laboratory is a widely used programming language which finds its most usage in writing mathematically intensive, technical programs that have a very low space and time complexity issue. MATLAB has graph plotting capabilities, fast array-based computations, and large mathematical functions library.**

**Simple linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line. Both variables should be quantitative.Regression analysis makes use of mathematical models to describe relationships. For example, suppose that height was the only determinant of body weight. Weight = 80 + 2 x (70) = 220 lbs. ... In this simple linear regression, we are examining the impact of one independent variable on the outcome**

11. Make a comparison between:-

1.Generalization and abstraction

2. Learning that is guided and unsupervised

3. Regression and classification

**1.Generalization is the characteristic of a machine learning model that dictates how much variations in the input data regarding problem statement can a model handle. Low generalization would mean that the model will not be able to give correct results over varied data, meaning overfitting or high variance. High generalization would mean that the model will be able to give correct results over varied data.**

**Abstraction is a step in machine learning process where the simplification of the problem is done to make the process more streamlined and more efficient while using less memory. Abstraction can constitute one or more steps. Feature selection, dimensionality reduction are some of the examples.**

**2. Supervised machine learning requires labeled data. Performance metrics are different for regression and classification problems. Number of class labels are known beforehand. Computationally taxing on the machine.**

**Unsupervised machine learning is done on unlabeled data. Unsupervised machine learning performance metrics are applicable for all problems being solved by this approach. Number of class labels is not known beforehand. Computationally not as taxing as supervised machine learning.**

**3. Regression involves predicting the value of target variable using labeled data, based on the values of independent variables. The values of target variable belong to an open set.**

**Classification involves predicting the value of target variable using either labeled or unlabeled data, based on values of independent variables. The values of target variable belong to a finite set**