Sheet

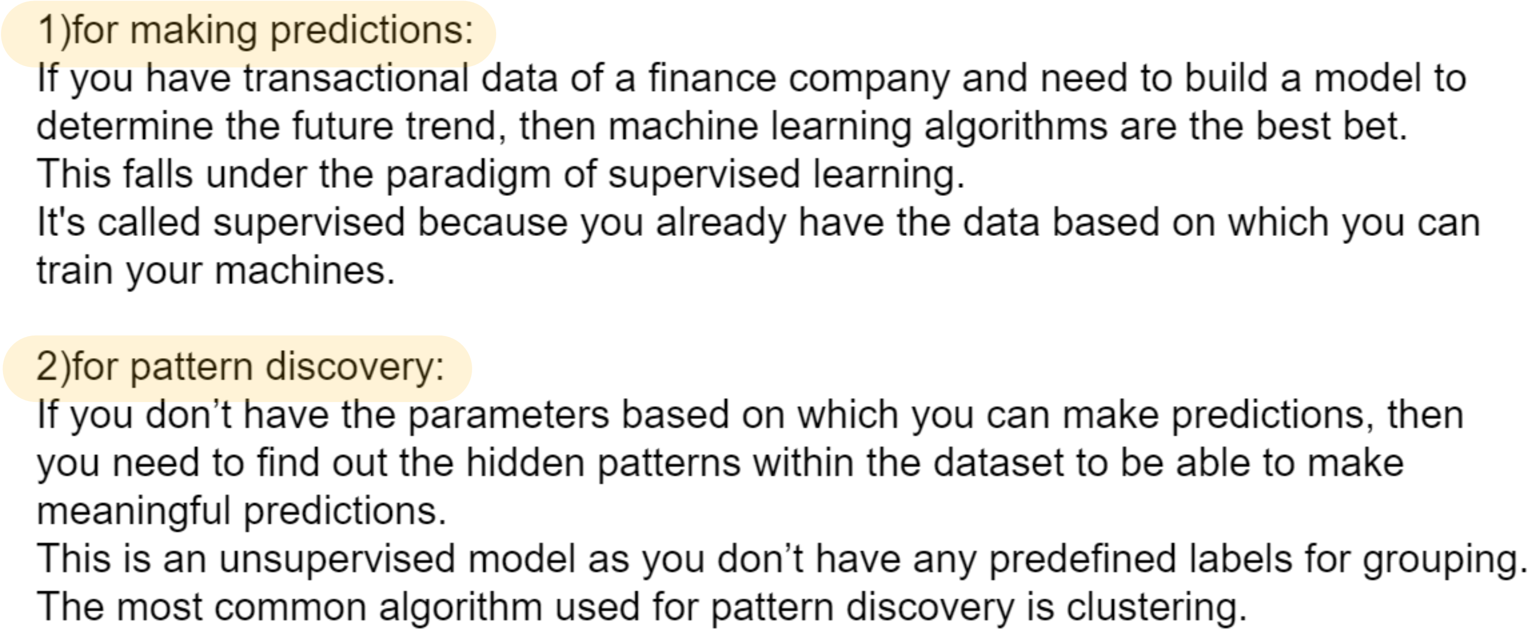
Data Science

Student Name: منة الله محمد عبدالفتاح …………………………………….

Student Number: 200900064 ………………………………….

Level: Second level ……………………………………………..

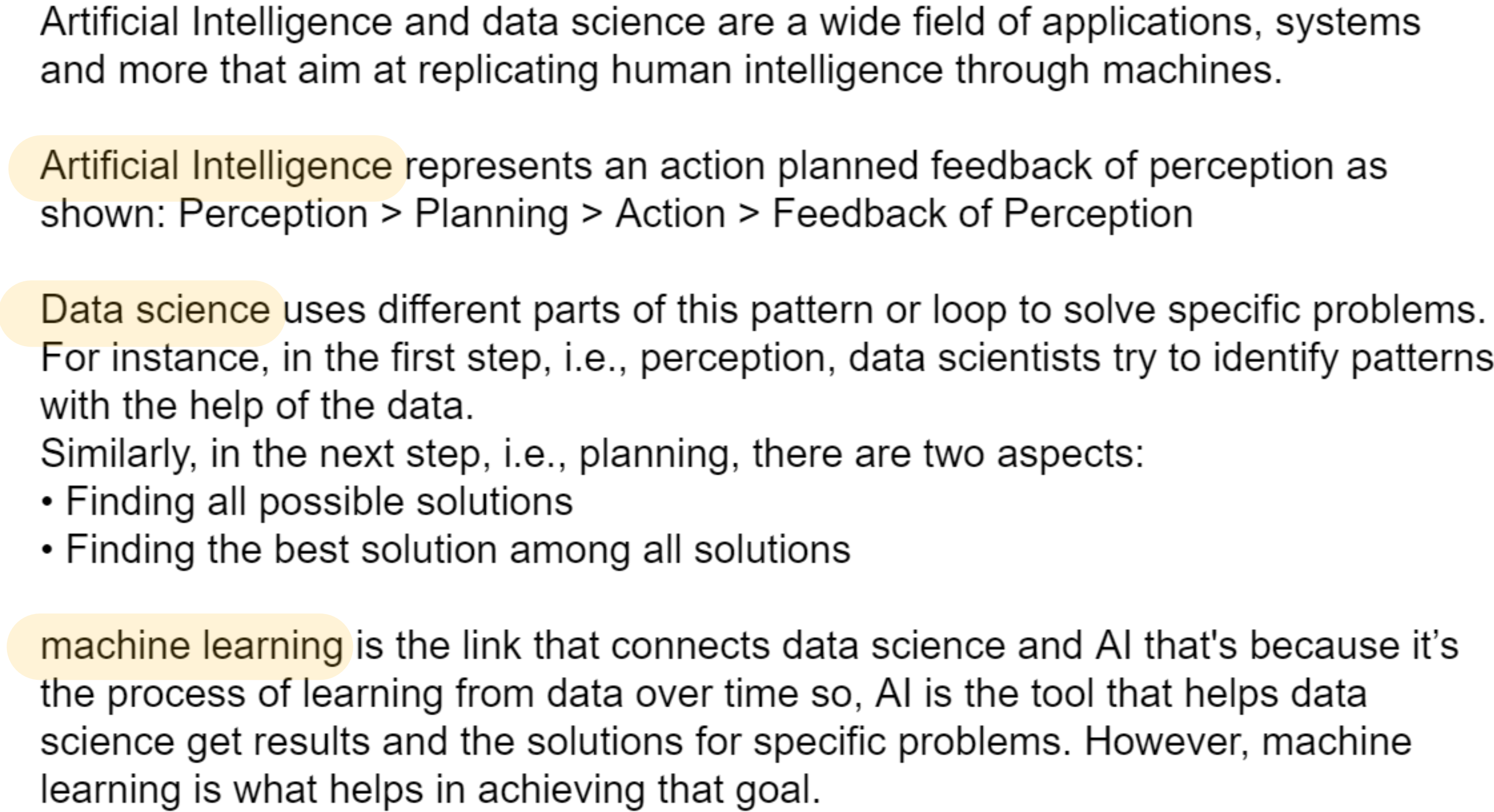
1- Why do data scientists need machine learning?



1)for making predictions:
If you have transactional data of a finance company and need to build a model to determine the future trend, then machine learning algorithms are the best bet.
This falls under the paradigm of supervised learning.
It's called supervised because you already have the data based on which you can train your machines.

2)for pattern discovery:
If you don’t have the parameters based on which you can make predictions, then you need to find out the hidden patterns within the dataset to be able to make meaningful predictions.
This is an unsupervised model as you don’t have any predefined labels for grouping.
The most common algorithm used for pattern discovery is clustering.

2- What is the relation between data science, machine learning and artificial intelligence?



Artificial Intelligence and data science are a wide field of applications, systems and more that aim at replicating human intelligence through machines.

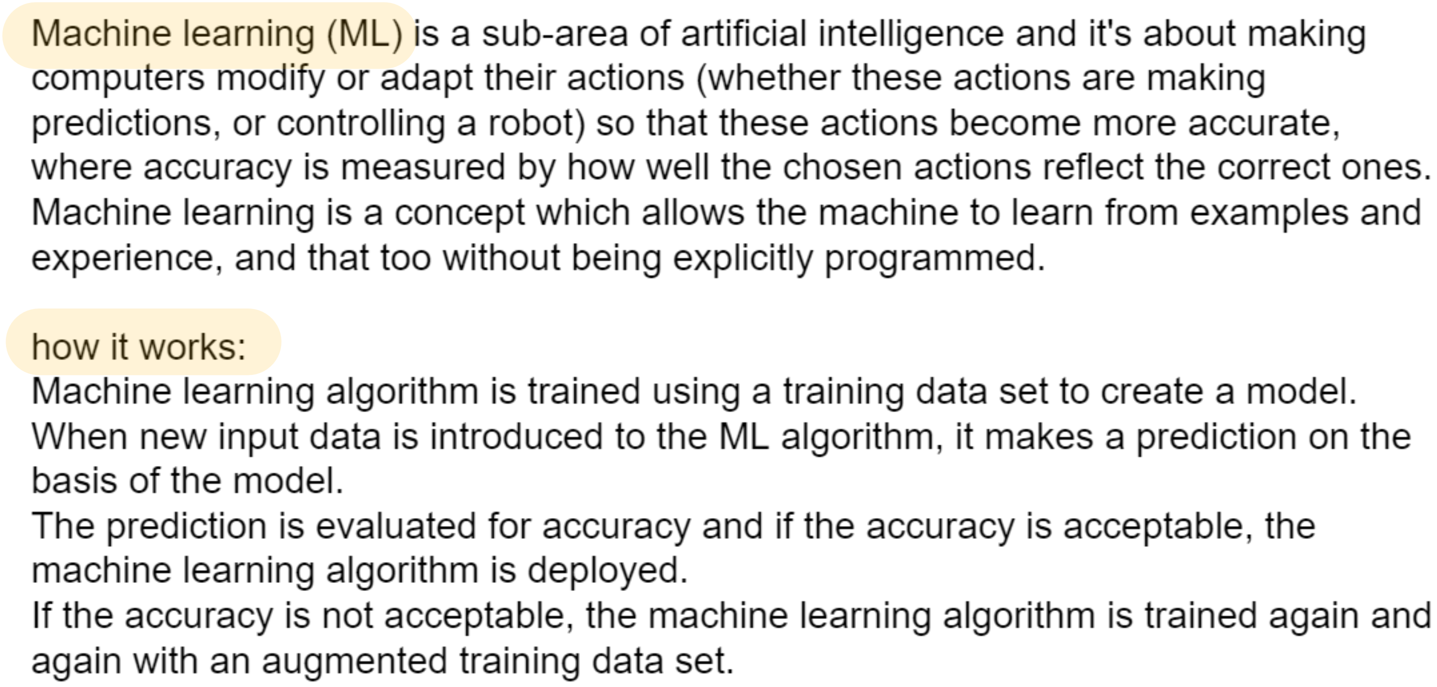
Artificial Intelligence represents an action planned feedback of perception as shown: Perception > Planning > Action > Feedback of Perception

Data science uses different parts of this pattern or loop to solve specific problems.
For instance, in the first step, i.e., perception, data scientists try to identify patterns with the help of the data.
Similarly, in the next step, i.e., planning, there are two aspects:

- Finding all possible solutions
- Finding the best solution among all solutions

machine learning is the link that connects data science and AI that's because it’s the process of learning from data over time so, AI is the tool that helps data science get results and the solutions for specific problems. However, machine learning is what helps in achieving that goal.

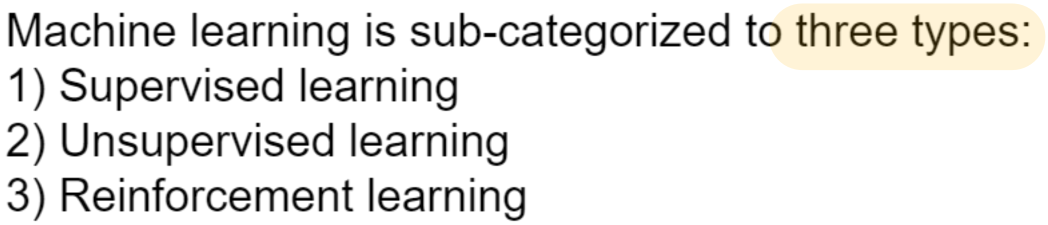
3- Define machine learning and explain how does it work?



Machine learning (ML) is a sub-area of artificial intelligence and it's about making computers modify or adapt their actions (whether these actions are making predictions, or controlling a robot) so that these actions become more accurate, where accuracy is measured by how well the chosen actions reflect the correct ones. Machine learning is a concept which allows the machine to learn from examples and experience, and that too without being explicitly programmed.

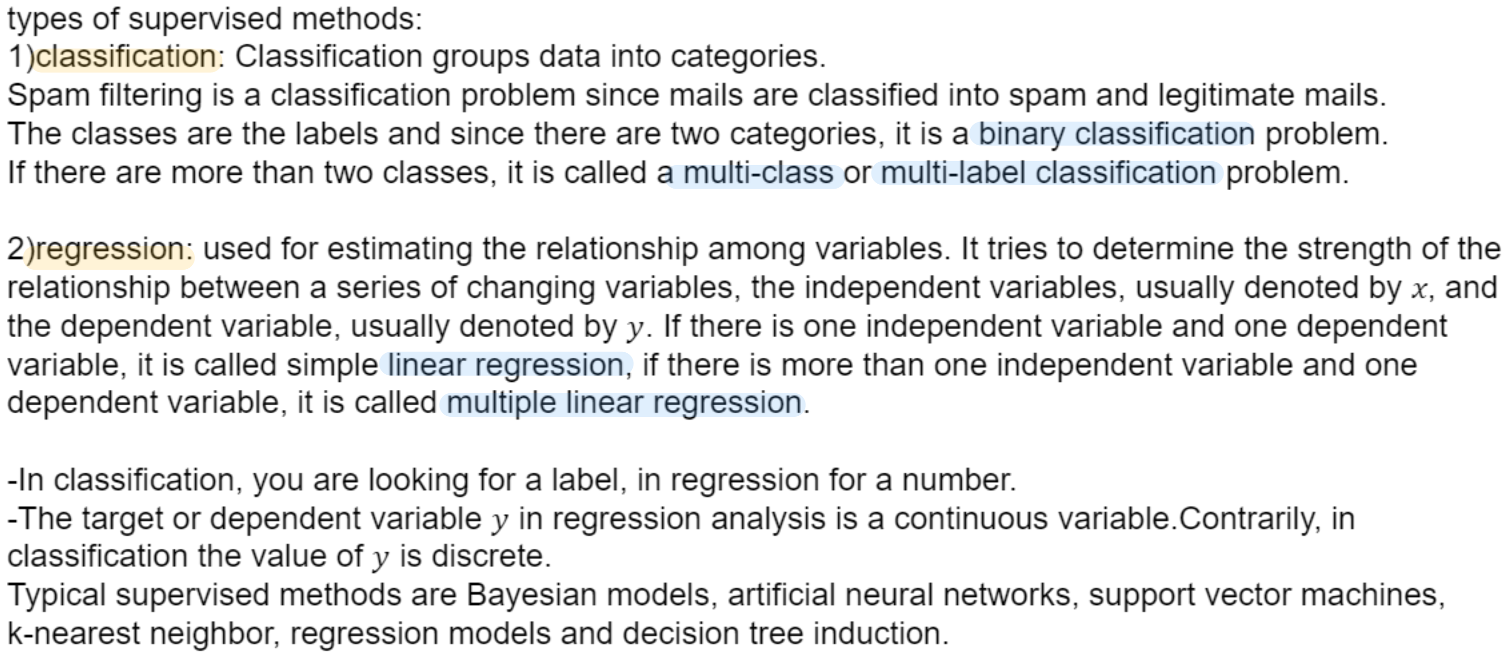
how it works:
Machine learning algorithm is trained using a training data set to create a model.
When new input data is introduced to the ML algorithm, it makes a prediction on the basis of the model.
The prediction is evaluated for accuracy and if the accuracy is acceptable, the machine learning algorithm is deployed.
If the accuracy is not acceptable, the machine learning algorithm is trained again and again with an augmented training data set.

4- What are the types of machine learning?



Machine learning is sub-categorized to three types:
1) Supervised learning
2) Unsupervised learning
3) Reinforcement learning

5- Explain the types of supervised learning?

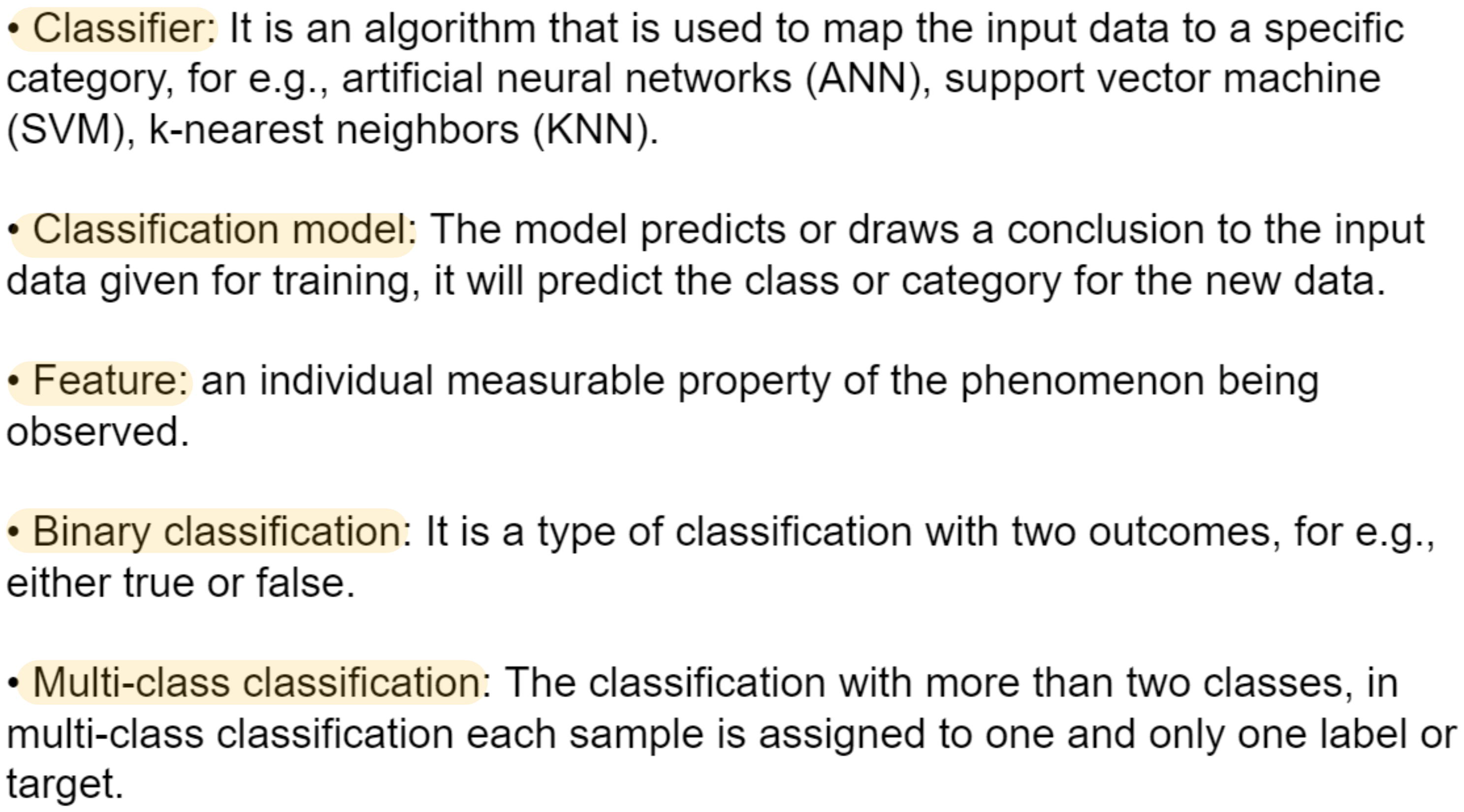


types of supervised methods:
1)classification: Classification groups data into categories.
Spam filtering is a classification problem since mails are classified into spam and legitimate mails.
The classes are the labels and since there are two categories, it is a binary classification problem.
If there are more than two classes, it is called a multi-class or multi-label classification problem.

2)regression: used for estimating the relationship among variables. It tries to determine the strength of the relationship between a series of changing variables, the independent variables, usually denoted by x, and the dependent variable, usually denoted by y. If there is one independent variable and one dependent variable, it is called simple linear regression, if there is more than one independent variable and one dependent variable, it is called multiple linear regression.

-In classification, you are looking for a label, in regression for a number.
-The target or dependent variable y in regression analysis is a continuous variable.Contrarily, in classification the value of y is discrete.
Typical supervised methods are Bayesian models, artificial neural networks, support vector machines, k-nearest neighbor, regression models and decision tree induction.

6- Define: classifier, classification model, feature, binary classification, and multi-class classification?



• Classifier: It is an algorithm that is used to map the input data to a specific category, for e.g., artificial neural networks (ANN), support vector machine (SVM), k-nearest neighbors (KNN).

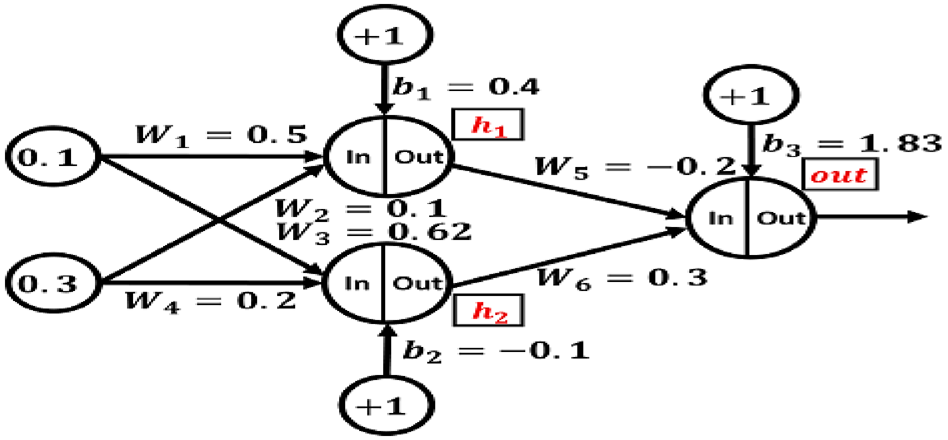
• Classification model: The model predicts or draws a conclusion to the input data given for training, it will predict the class or category for the new data.

• Feature: an individual measurable property of the phenomenon being observed.

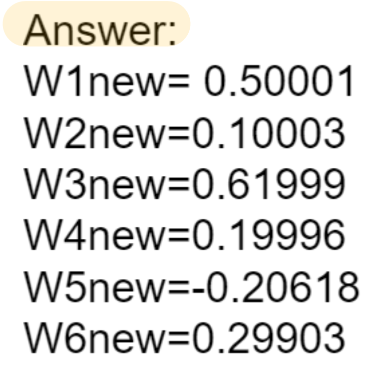
• Binary classification: It is a type of classification with two outcomes, for e.g., either true or false.

• Multi-class classification: The classification with more than two classes, in multi-class classification each sample is assigned to one and only one label or target.

7-

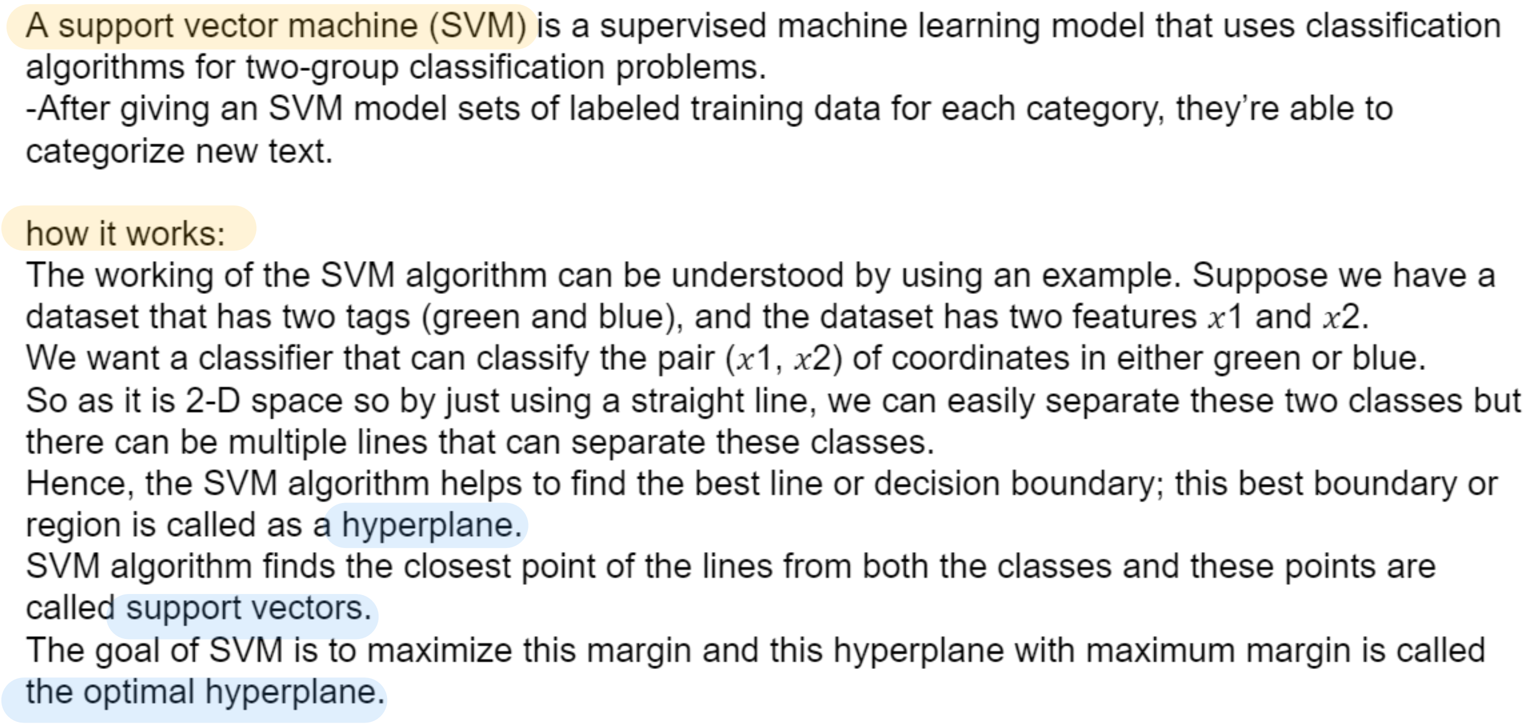


From the above figure, use backpropagation algorithm to calculate: w_{1new}, w_{2new}, w_{3new}, w_{4new}, w_{5new}, w_{6new}. Use $\eta = 0.01$ and the desired output = 0.03.



Answer:
W1new= 0.50001
W2new=0.10003
W3new=0.61999
W4new=0.19996
W5new=-0.20618
W6new=0.29903

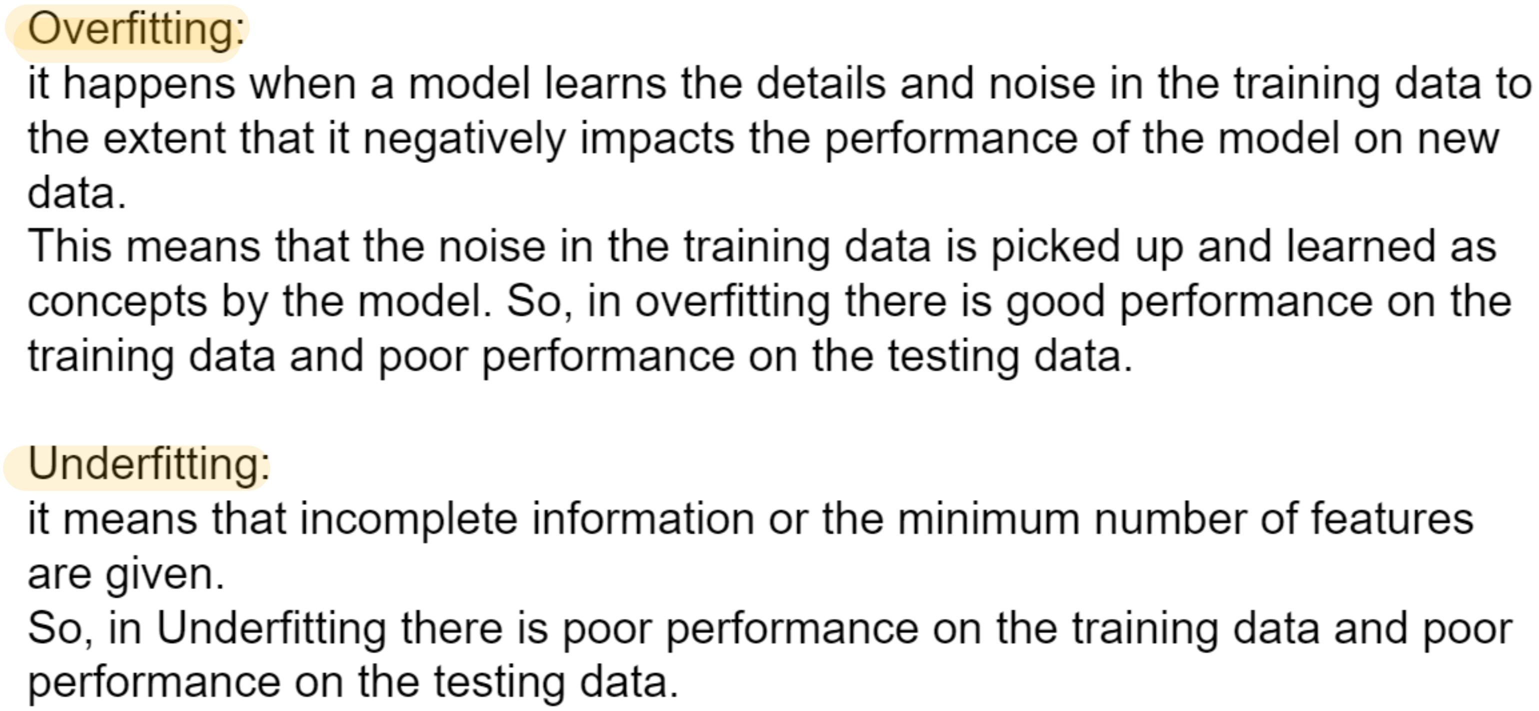
8- Define support vector machine and how does it work?



A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems.
-After giving an SVM model sets of labeled training data for each category, they’re able to categorize new text.

how it works:
The working of the SVM algorithm can be understood by using an example. Suppose we have a dataset that has two tags (green and blue), and the dataset has two features x1 and x2.
We want a classifier that can classify the pair (x1, x2) of coordinates in either green or blue.
So as it is 2-D space so by just using a straight line, we can easily separate these two classes but there can be multiple lines that can separate these classes.
Hence, the SVM algorithm helps to find the best line or decision boundary; this best boundary or region is called as a hyperplane.
SVM algorithm finds the closest point of the lines from both the classes and these points are called support vectors.
The goal of SVM is to maximize this margin and this hyperplane with maximum margin is called the optimal hyperplane.

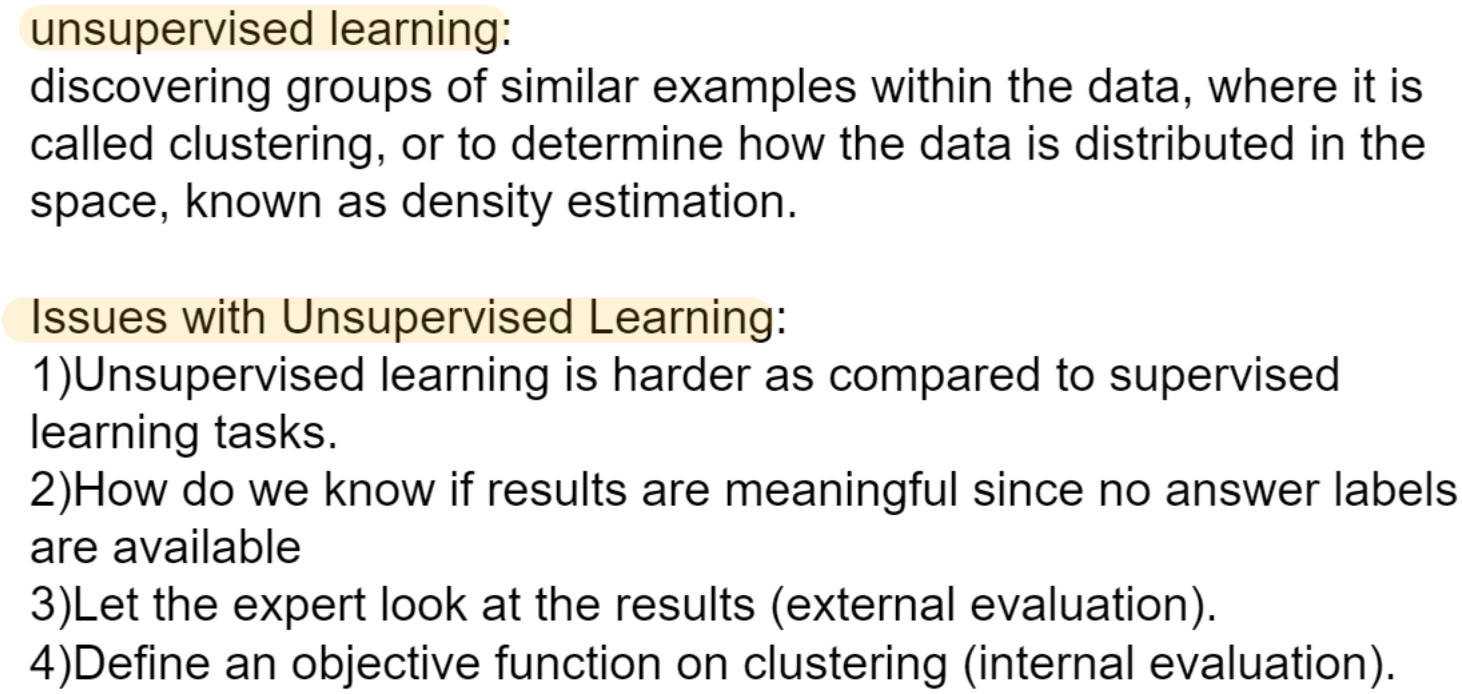
9- what is the difference between overfitting and underfitting?



Overfitting:
it happens when a model learns the details and noise in the training data to the extent that it negatively impacts the performance of the model on new data.
This means that the noise in the training data is picked up and learned as concepts by the model. So, in overfitting there is good performance on the training data and poor performance on the testing data.

Underfitting:
it means that incomplete information or the minimum number of features are given.
So, in Underfitting there is poor performance on the training data and poor performance on the testing data.

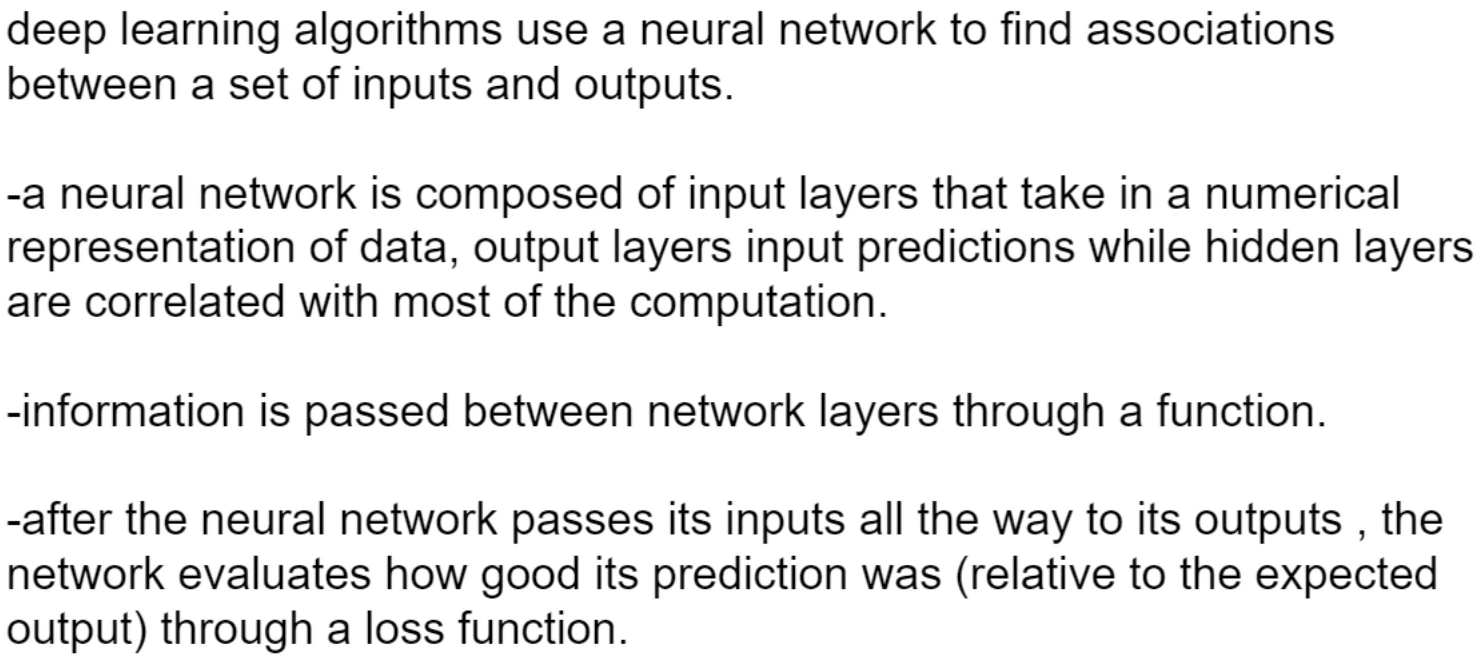
10- Define unsupervised learning and what are the issues with unsupervised learning?



unsupervised learning:
discovering groups of similar examples within the data, where it is called clustering, or to determine how the data is distributed in the space, known as density estimation.

Issues with Unsupervised Learning:
1)Unsupervised learning is harder as compared to supervised learning tasks.
2)How do we know if results are meaningful since no answer labels are available
3)Let the expert look at the results (external evaluation).
4)Define an objective function on clustering (internal evaluation).

11- How does deep learning algorithms “learn”?



deep learning algorithms use a neural network to find associations between a set of inputs and outputs.

-a neural network is composed of input layers that take in a numerical representation of data, output layers input predictions while hidden layers are correlated with most of the computation.

-information is passed between network layers through a function.

-after the neural network passes its inputs all the way to its outputs , the network evaluates how good its prediction was (relative to the expected output) through a loss function.

12- Using confusion matrix, what is the formula of: classification accuracy, misclassification rate, precision, recall and f-measure?

classification accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

misclassification rate (error rate):

- $\mathrm{ERR} = \dfrac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}}$

precision:

$$\textbf{\textit{Precision}} = \frac{\text{TP}}{\text{TP + FP}}$$

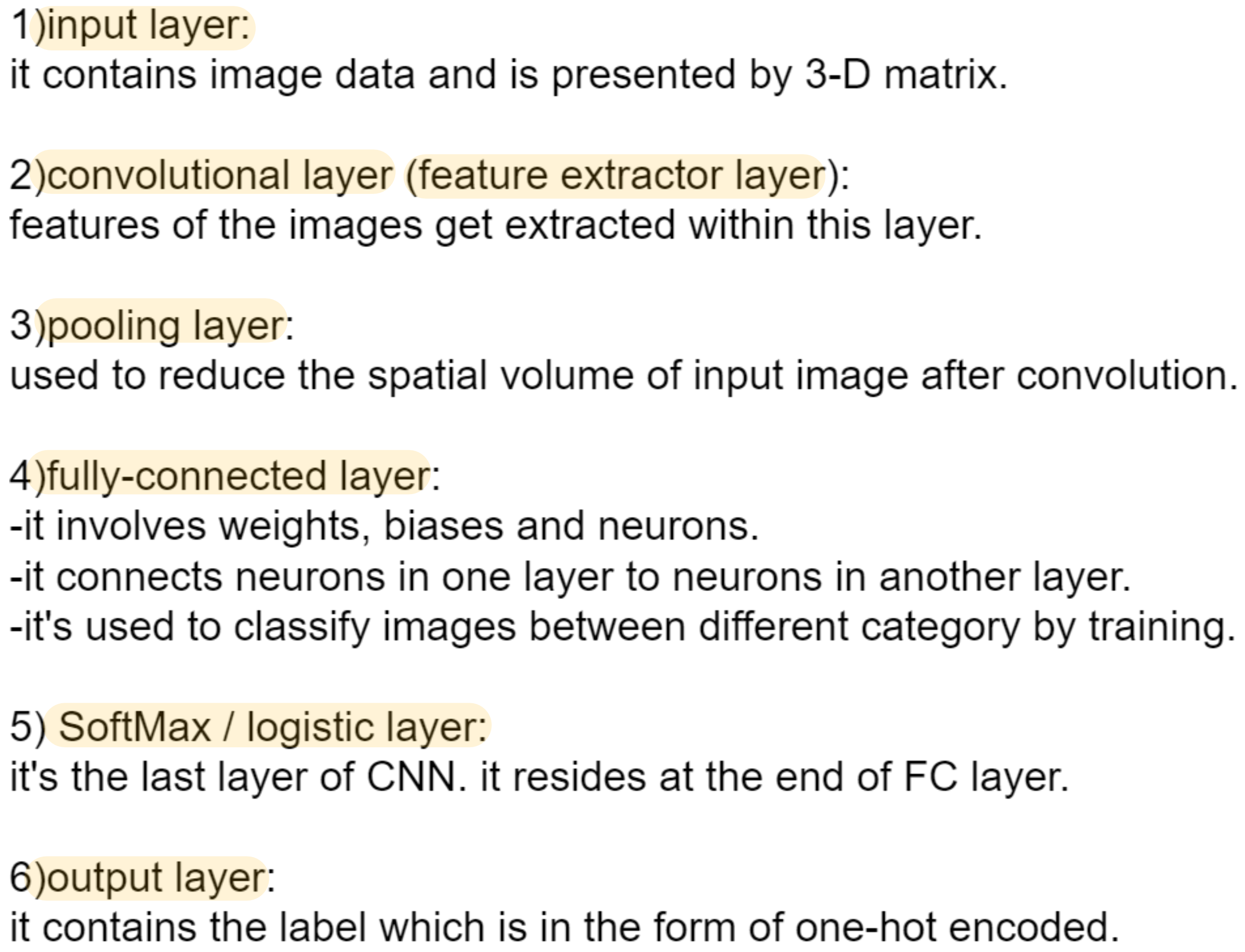
recall:

$$\textbf{\textit{Recall}} = \frac{\text{TP}}{\text{TP + FN}}$$

f-measure:

- $\mathrm{F}_1 = \dfrac{2 \cdot \mathrm{PREC} \cdot \mathrm{REC}}{\mathrm{PREC} + \mathrm{REC}}$

13- Explain the layers in convolutional neural networks?



1)input layer:
it contains image data and is presented by 3-D matrix.

2)convolutional layer (feature extractor layer):
features of the images get extracted within this layer.

3)pooling layer:
used to reduce the spatial volume of input image after convolution.

4)fully-connected layer:
-it involves weights, biases and neurons.
-it connects neurons in one layer to neurons in another layer.
-it's used to classify images between different category by training.

5) SoftMax / logistic layer:
it's the last layer of CNN. it resides at the end of FC layer.

6)output layer:
it contains the label which is in the form of one-hot encoded.