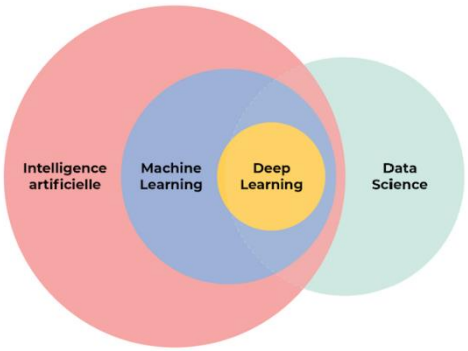
1. Differentiate Machine Learning, artificial intelligence, and data science.



As the image above represents, there are some common aspects of these topics. AI is the way of mimicking human intelligence. ML is a subfield of AI that solve problems by assessing the data that we already have. Lastly, Data Science is a bit different domain than AI and ML which deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. However, the meaning of data science is a so much **controversial topic.**

2. What is the difference between linear regression and logistic regression?

Linear regression is a model to predict a value depending on data that is given to the model. It can give any value as output between **-infinity** to **+infinity,** possibly**.**

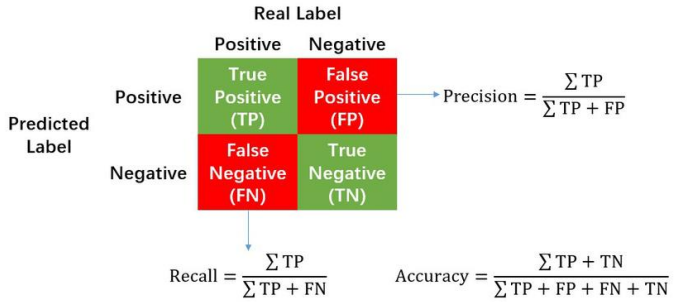
Logistic Regression is a **classification** model. It can give a value between 0 and 1 which demonstrates the similarity of the inputs to a class or to an output.

3. Explain the curse of dimensionality?

If we have too many dimensions in our data, then curse of dimensionality can be observed. Curse of dimensionality means the exponential amount of combination possibility that a dataset may have with too many columns. Then, we may need so much training data to get a reliable output. The solution to this problem can be dimensionality reduction techniques such as **PCA and t-SNE.**

4. What are precision, recall, f-measure, and roc? Explain what they are and when

we use each one.



As represented in the image above, precision score refers to the **certainty of the positive classification task.** And recall refers to the **certainty of the negative classification task.**

> Precision can also be seen as the probability that a randomly selected item which is labeled as "relevant" is a true positive.
>
> For the individual element, the recall percentage gives the probability that a randomly selected relevant item from the dataset will be detected.

F-Measure=

$$\text{F1-score} = 2.\frac{\text{Precision}.\,\text{Recall}}{\text{Precision} + \text{Recall}}$$

ROC (receiver operating characteristics) =

It is a **curve** that represents different precision and recall scores at **different thresholds.**

5. What is the difference between train set, test set, and validation set?

Train set is used only for training model,

Test set for only testing at the end,

And validation set is the sample of data used to provide an unbiased evaluation of a model fit on the training dataset **while tuning model hyperparameters.**

6. What is the p-value? It's a reliable measurement? How can we be sure?

A p-value, or probability value, is a number describing how likely it is that your data would have occurred by random chance (i.e. that the null hypothesis is true). A high p-value means the result of the study can be derived by placing random data and a low p-value means that the result is not a result of luck.

7. What is PCA?

If we have a data with so many columns or dimension, we can use PCA (Principal Component Analysis) to reduce the number of dimensions with **minimum loss in information.** However, minimum loss does not mean that there will be no loss. The aim of PCA is to **minimize this information loss.** After implementing PCA, we can observe how many percent of the info is gone.