Theoretical Test

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

While these 3 terms include many common aspects, and there are some controversial about the differences; we can say that **AI** is a research area to mitigate human intelligence by solving complex problems.
Machine learning is a method to accomplish this by evaluating the existing data, which also means that machine learning is a subfield of AI
Lastly, **data science** derives meaning out of data but also maintains the pipelines of obtaining and cleaning of it.

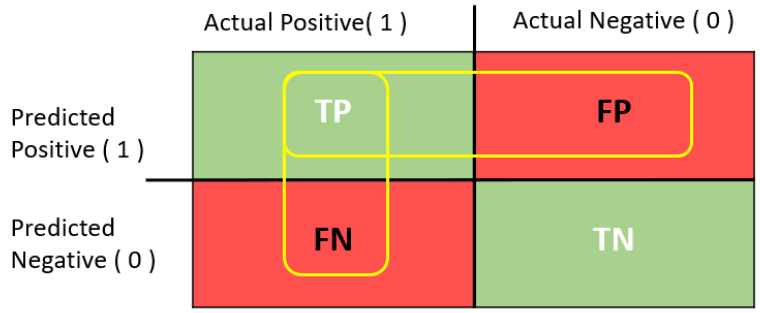
2. What is the difference between linear regression and logistic regression?

In brief, while linear regression is used to solve **regression** problems, logistic regression is for classification problems. However, both of them use 'linear' methods to solve them. In this way, we can say that logistic regression is also a 'linear' method like linear regression. The reason why the curve of the logistic regression does not look like a straight line is the **sigmoid activation function**.

3. Explain the curse of dimensionality?

Curse of dimensionality refers to the problems that occur with a **high number of dimensions** (columns) in a data.
The idea here is that, (for example) if we have too many binary features then we will have t**oo many combinations** for these values which are **exponents of 2.**
Then, to obtain the max amount of performance, **we will need too much data** to observe all the behaviors of the data on different coordinates.
There are some methods such as dimensionality reduction (PCA, t-SNE..) or forward-feature selection to handle this problem.

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



Precision and Recall are two different metrics to handle the accuracy dilemma of a model.

=> precision answers the question of: What proportion of positive identifications was actually correct?
Its formula is TP / (TP+FP)
For example, if you have a security system and you do not want to make a false warning to your customers about a misunderstood robbery event, you may want this metric to be high.

=> And **recall** answers: What proportion of actual positives was identified correctly?
Its formula: TP / (TP+FN)
If you have a security system and you do not want to miss any single robbery attempt even if it is not 100% confirmed, you may want this metric to be high.

F-Measure is a metric to combine both of them:
F-Measure = (2 * Precision * Recall) / (Precision + Recall)
And the **roc curve** is the plot of different precision recall scores at different classification thresholds. To evaluate this graph, we can use AUC (area under curve) score to find the optimum point for precision and recall.

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

Train set and **validation set** is used while training, but validation is used just to fine-tune the parameters and to see the effect of the training. **Test set** is used after all the training process is done to make an unbiased performance measurement for the **model.**

6. What is the p-value? Is it a reliable measurement? How can we be sure?

The p-value is used to determine if the outcome of an experiment is statistically significant. A low p-value means that, assuming the null hypothesis is true, there is a **very low likelihood** that this outcome was a result of luck. A high p-value means that, assuming the null hypothesis is true, this **outcome was very likely.**

7. What is PCA?

PCA is a dimensionality reduction technique to squeeze a data with high dimension into **lower dimensions** to overcome problems such as the **curse of dimensionality** as we have mentioned. To give an analogy, it is like **taking the picture of a 3D world and printing it to a 2D paper.** The picture below tries to explain that idea. But the critical point here is that we need to find the **‘right**

angle’ to represent that 3D world in a 2D paper, which is the task of a dimensionality reduction algorithm.