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

* Artificial intelligence is a simulation of human intelligence in machines. AI enables machines to think, learn, and find solutions (solve problems) just like human brains do.
* Machine learning is a subset of artificial intelligence. It essentially gives machines the ability to learn and improve through experiences – **without the need to program them explicitly**. ML aims to develop programs that can access data and utilise it to learn for themselves.
* Data science is basically a multidisciplinary field that essentially focuses on extracting insights from large data sets – both raw and structured. The data scientists use statistics, machine learning, and predictive analysis to establish solutions of questions.

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

Linear Regression is used in problem that you have predict value in a range ( a continuous values). A very famous example of this is predict the price house base on characteristics. Logistic Regression d is a [classification](https://www.sciencedirect.com/topics/computer-science/classification) model rather than regression model, which can be applied when the problem has discrete outcome. Example of problem the are Credit Analyses, when to have to decide to loan money or not for some one base on the historical.

1. Explain the curse of dimensionality?

The curse of dimensionality is how call some phenomena that occur when analyzing [high-dimensional spaces](https://en.wikipedia.org/wiki/High-dimensional_space) that do not occur in low-dimensional.

The most famous phenomena is the sparsity. The high dimensional datasets are at risk of being very sparse: most training instances are likely to be far away from each other. Of course, this also means that a new instance will likely be far away from any training instance, making predictions much less reliable than in lower dimensions, since they will be based on much larger extrapolations. In short, the more dimensions the training set has, the greater the risk of overfitting it.

This mean solutions for that is use dimensionality reduction algorithm

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

Those are method to measure the performance of classification models and they are used when it is necessary to measure the model performance.

* Precision is ratio of the total instances that are correctly detected by the model.



* Recall is the ratio of positive instances that are correctly detected by the classifier



* The ROC or the ROC curve plots the true positive rate (another name for recall) against the false positive rate. One way to compare classifiers is to measure the area under the ROC curve (AUC). A per‐ fect classifier will have a ROC AUC equal to 1, whereas a purely random classifier will have a ROC AUC equal to 0.5
* f-measure is is the [harmonic mean](https://en.wikipedia.org/wiki/Harmonic_mean) of the precision and recall

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

Train set is being used to fit the model. In other words, training set are used to learn the parameters of the model of interest.

Validantion set used to tune the hyper-parameter and improve the model selection. This happen once the initial is model is trained (or fit)

Test set is a part of the date used to assess the performance of the final model

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

P-value helps us determine how likely it is to get a particular result when the [null hypothesis](https://en.wikipedia.org/wiki/Null_hypothesis) is assumed to be true.

It's a reliable measurement, because it is give the probability of getting a sample like ours or more extreme than ours if the null hypothesis is correct.

If the p-value is very small (<0.05 is considered generally), then our sample is “strange,” and this means that our assumption that the null hypothesis is correct is most likely to be false. Thus, we reject it.

1. What is PCA?

Principal Component Analysis (PCA) is the most popular dimensionality reduction algorithm. In simply word, it is identifies the hyperplane that lies closest to the data, and then it projects the data onto it. This helps to reduce the quantity of features the you have to use is our models.