Exercise 2

1- What is overfitting and underfitting?

Overfitting: refers to a model that models the training data, happens when a model learns the detail 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 or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the model's ability to generalize. Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function.

Underfitting: refers to a model that can neither model the training data nor generalize to new data. An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data. Underfitting is often not discussed as it is easy to detect given a good performance metric. The remedy is to move on and try alternate machine learning algorithms. Nevertheless, it does provide a good contrast to the problem of overfitting.

2- Why we test the model on both train set and test set?

In applied machine learning, we seek a model that learns the relationship between the input and output variables using the training dataset. The goal is that we learn a relationship that generalizes to new examples beyond the training dataset. In the case of machine learning competitions, like those on Kaggle, we are given access to the complete training dataset and the inputs of the test dataset and are required to make predictions for the test dataset,

this leads to a possible situation where we may accidentally or choose to train a model to the test set.

Exercise 3

3- What are the common techniques of regularization?

Regularization is a technique which makes slight modifications to the learning algorithm such that the model generalizes better. This in turn improves the model's performance on the unseen data as well.