Hassan Issa-Cisse

10180100188

# ARTIFICIAL INTELIGENCE MIDSEM – REPORT

We have been tasked to work on a dataset to predict the future price of houses in Boston with the Boston Dataset. Usually for analysis purpose like this, a regression model is used and depending on the type of data you have at disposal, then the correct choice of regression can be made.

First of all, lets define a regression: it is an analysis technique used in machine learning in order to predict independent variable called target. There are two types of regression, we have the linear regression and the multiple regression. For the linear regression, we want to find the correlation between the independent variable and the dependent variable. We will have something like this:

$$Y= a.x + b$$

For the multiple regression, we have more than one dependent variable and it can be put as a matrice of dependent variable X:

$$Y=a.X + b$$

The Boston Housing Dataset consists of price of houses in various places in Boston. Alongside with price, the dataset also provides information such as Crime (CRIM), areas of non-retail business in the town (INDUS), the age of people who own the house (AGE), and there are many other attributes. In the Boston scenario, we used a linear regression model because only the price attribute was needed for the analysis.

Before analysis, we need to clean the data by removing outliers or checking for missing values. An outlier can be described as an anomaly within the dataset. For instance, it can be a 10 in a group of 10,000. Therefore, we look for values that doesn't match our dataset standings.

During the analysis, I came across some exiting tools used for the machine learning, one is the sklearn modules. Sklearn is a library for python used for vector machine, random forest. But in our case, we used sklearn.linear_model for our work. It features a robust and efficient set of algorithms to plot and analyze dataset.

At the end the MSE can be used to check the robustness of the model you have trained, if your MSE is too far from your predicted values then it means you model is poor.