MEMORY

1. Data collection

The datasets were obtained by me from Kaggle and downloaded from Glassdoor via web scraping by a third party.

Each of them contains information about job offers and they differ in the type of profile for which the offer is for, of which we highlight: data engineer, data analyst and data scientist.

1. Data cleaning

Some of the cleaning tasks I have had to carry out are:

* Removing and creating columns
* Replacing empty or erroneous values
* Removing special characters
* Edit outliers

1. Code

The project code is developed in Python in a notebook file. Headed by an index containing all the steps followed throughout this project.

1. Difficulties

* Escoger modelos.
* Codificar variables.
* Reemplazar valores erróneos
* Decidir que variables utilizar para predecir.

1. Documentation

* <https://scikit-learn.org/stable/user_guide.html>
* <https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/>
* <https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/>
* <https://towardsdatascience.com/4-machine-learning-techniques-for-outlier-detection-in-python-21e9cfacb81d>
* <https://analyticsindiamag.com/a-complete-guide-to-categorical-data-encoding/>
* <https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/>

1. Conclusions

The best model is by far Decision tree as it gives us a prediction of 0.94, which in my opinion is quite high considering that it is real data.