Assignment 2 (Jupyter NoteBook)

**You can form groups of 5-6 people.**

**Please form your teams and pick 2 topics. (Like plan A and plan B for the next lecture)**

**Assignment 1: Build a performing machine learning Regression (OR CLASSIFICATION )algorithm**

To build a performing machine learning model from A-Z. Check this resource. It has all of the instructions together with the scripts to build your model. Go through it to see which parts will apply to the model you are trying to build and the dataset you are working with:

<https://www.slideshare.net/CharlesVestur/building-a-performing-machine-learning-model-from-a-to-z>

**Project banks and Dataset repository**:

You must pick a REGRESSION or classification problem for this assignment.

There are hundreds of projects and data sets in the following resources. Have a look and pick at least two topics. You will only do one of the projects for this assignment though!

* [UCI Machine Learning Repository](http://archive.ics.uci.edu/ml/) – 350+ searchable datasets spanning almost every subject matter. You’ll definitely find datasets that interest you.       <http://archive.ics.uci.edu/ml/index.php>
* [Kaggle Datasets](https://www.kaggle.com/datasets) – 100+ datasets uploaded by the Kaggle community. There are some really fun datasets here, <https://www.kaggle.com/datasets>
* [data.gov](https://www.data.gov/) – Open datasets released by the U.S. government. Great place to look if you’re interested in social sciences. <https://www.data.gov/>

Every team leader will share their topic 1, with the rest of the class  during the lecture to make sure that their topic is not taken by any other teams. So, if 2, or more than 2 teams picked a same project, then they have to change their topic and move to their topic # 2 (or possibly Topic # 3)

**Reports (70 marks)**:

* Write a report which is no longer than 10 pages.
* Explain the dataset you are using, explain all the steps in your solution pipeline, including but not limited to :

-Preprocessing steps (DATA procuring, cleaning, imputation),

- Make sure you are using the proper columns (features)

-What the features are and determine the target labels,

- Splitting data into train/test

- Building a Regressoron this dataset using any Machine learning algorithm you prefer ( e.g., Random Forest,  XG boost, etc …), explain why you are choosing it and how it works.

-Evaluating your best result, analyze the result and write a summary and conclusion. (use the appropriate evaluation tool, such as MSE,MAE, R square  etc )

Note : Although this is valuable to get an excellent/accurate model, the magnitude of the evaluation metric you use, for example the magnitude of your MSE is not going to influence your mark at this time…as we are trying to practice using the tools.

**Presentation (30 marks):**

Each group member should present the work during the class time. The presentation **should be ~20 minutes long per group. So make enough slides for this duration of time** and all students should contribute in the presentation.

**What you will need:**

* **Install Anaconda**
* **Install Python**
* **Install Jupyter Notebook through Anaconda**
* **Install and import the following libraries:**
* **Pandas**
* **Numpy**
* **Matplotlib**
* **Scikit-learn**

**Late submissions will not be accepted and will receive NO MARK for the assignment.**

**Submissions:**

**1-    Data Set**

**2-    Your code (Jupyter notebook) + (html)**

**3-    The output file**

**4-    Your report in PDF format**

**5-    Your presentation slides**

**Please upload all these 5 files in a Zipped file**

**Good luck,**

**Ali**