**CPSC 480 – Project#1 Regression models – 50 points**

In chapter2 we used the California census dataset to build a model of housing prices in the state. Here are the main steps we discussed:

1. Look at the big picture.

2. Get the data.

3. Discover and visualize the data to gain insights.

4. Prepare the data for Machine Learning algorithms.

5. Select a model and train it.

6. Fine-tune your model.

7. Present your solution.

8. Launch, monitor, and maintain your system.

The project can be found in the following Jupyter Notebook link: <https://github.com/ageron/handson-ml/blob/master/02_end_to_end_machine_learning_project.ipynb>

For your project you will be building a regression model that learns from a real dataset, that you are interested in, and be able to predict a real value. Select a real data set, the list below will help you find one, and try to go through the whole process from step 1 to 8, as we did for housing prices model. Feel free to adapt these steps to your needs. You should use a Jupyter Notebook for this project.

Submit 1) your Jupyter Notebook for the code and comments, and 2) the dataset you worked on (an online link to the dataset would work too)

Here are a few places you can look to get data:

* **Popular open data repositories**
* UC Irvine Machine Learning Repository (<http://archive.ics.uci.edu/ml/index.php>)
* Kaggle datasets (<https://www.kaggle.com/datasets>)
* Amazon’s AWS datasets (<https://registry.opendata.aws/>)
* **Meta portals (they list open data repositories)**
* Data Portals (<http://dataportals.org/>)
* Quandl( <https://www.quandl.com/>)
* **Other pages listing many popular open data repositories**
* Wikipedia’s list of Machine Learning datasets (<https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research>)
* Quora.com (<https://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public>)
* The datasets subreddit (<https://www.reddit.com/r/datasets/>

**Grading rubric:**

2 points –Project description

3 points –Using a dataframe with any useful functions such as info() and describe()

3 points – Visualizing the dataset

4 points – Preprocessing

10 points – Splitting the data into training and testing

10 points- Selecting a model and train it

3 points- Accuracy of the model (such as root mean square error (RMSE))

5 points – Using cross validation technique

5 points – Link for a dataset

5 points - Jupyter Notebook with the code