Group Proposal by Hsueh Yi Lu , Ali Can Mutlu, Zhonghang Yu

After you have selected a topic, a model, and a data set, submit a proposal of what you plan to do for the project. The proposal should be a few hundred words, and should address the following items.

- What problem did you select and why did you select it?

  We selected the topic of pollution for our project, specifically death and the air pollution pollution from production. We selected this topic because we feel strongly about human caused pollution and its effects on society. We believe that through the skills that we picked up in our class, we can clearly lay out the seriousness of the situation to our class.

- What database/dataset will you use? Does it need to be cleaned?

  We will use the “Death Due to Air Pollution” dataset found on kaggle.

  Link: https://www.kaggle.com/akshat0giri/death-due-to-air-pollution-19902017

  We also use the world health situation

  https://www.kaggle.com/utkarshxy/who-worldhealth-statistics-2020-complete?select=airPollutionDeathRate.csv

  https://www.kaggle.com/theworldbank/world-bank-gdp-ranking

  https://www.kaggle.com/danevans/world-bank-wdi-212-health-systems

  https://www.kaggle.com/prateekmaj21/electricity-production-by-source-world

  https://www.kaggle.com/ruchi798/global-environmental-indicators

  Yes it needs to be cleaned. We will definitely get rid of the null values and format the data.

- What data mining algorithm will you use? Will it be a standard form, or will you have to customize it?

  We will use the standard form of linear regression.

- What packages will you use to implement the model? Why?

We will use packages such as Numpy, Panda, matplot, seaborn. These are the packages that we picked up in our class, the simple usage will look like the following; Panda to read the data. Numpy to clean the data and show basic statistics of the data. MAtplot and Seborn to plot visually appealing graphs.

- What reference materials will you use to obtain sufficient background on applying the chosen model to the specific problem that you selected?

  We will use the machine learning prospective as reference material, specifically a chapter called “Linear Regression”(starting at page 207).

- How will you judge the performance of your results? What metrics will you use? Provide a rough schedule for completing the project.

We will cut a portion of the data set(training data) to make a prediction model. After building the model we will make assumptions. Looking at the findings that we get through our model, we will compare them with the portion of the data we didn't use. According to the final comparison we will give our evaluation of the project.

Rough Schedule:

1. Cleaning the data, using pandas package
2. Dividing the data and picking the training data
3. Using the testing part to build the prediction model
4. Testing our prediction model with the testing data
5. Evaluating the findings