**DataGood Mentor Project**

**Overview**

The goal of this project is to utilize a medical insurance cost dataset to learn important techniques in data science, and educate students on how to provide clear, presentable data. The dataset here will be on medical costs and insurance for individuals, and we will be taking a look at how various factors can contribute to this, as well as exploring the relationships between characteristics and traits of individuals in our dataset. You will ultimately create a final deliverable that summarizes the work that you've done, and interesting conclusions you have found over the course of a few weeks.

**Technology Stack**

* NumPy
* Pandas
* Scikit-learn

**Goals**

**#1**

* Analyze data through the use of common exploratory data analysis techniques, as well as modelling methods in order to get a better understanding of how to create meaningful visualizations that help others extrapolate information from datasets.

**#2**

* Utilize regression tools to predict insurance costs or other features based on provided data such as age, BMI, or other factors for our given dataset.

**#3**

* Create a final deliverable that summarizes the work done and the information found over the course of a few weeks of work.

**Fall Semester Timeline**

|  |  |
| --- | --- |
| **Week** | **Plan** |
| Week 1: 10/12 - 10/18 | * EDA + Visualization |
| Week 2: 10/19 - 10/25 | * Outlier Detection |
| Week 3: 10/26 - 11/01 | * Modelling |
| Week 4: 11/2 - 11/08 | * Linear Regression |
| Week 5: 11/09 - 11/15 | * Regression Improvements and Random Forests |
| Week 6: 11/16 - 11/22 | * YOUR FOCUS HERE |
| Week 7: 11/23 - 11/24  **Thanksgiving Holiday (11/25-11/27)** | * Final Deliverable |

**Resources**

* Medical Cost Personal Datasets:

<https://www.kaggle.com/mirichoi0218/insurance>

This dataset is relatively quite small, but is really clear and concise. It is a great way to showcase and try many data science techniques. Note that the original source comes from a book on Machine Learning in R, so the data actually comes from simulations on the basis of demographic statistics from the US Census Bureau. For the purposes of this project, it is still enough to use to test tools for manipulating data and structuring it concisely.