**Machine Learning Project Proposal**

By Denys Ladden and Emil B. Berglund

CS-320-01

Dr. Nathan Johnson

25.10.2021

For this project we will use a publicly available dataset to construct a linear regression model that predicts the value of one of the features(variables) in the dataset

Overview:

For this project proposal we chose to use data on insurance charges and different factors leading to these charges. The values included in this dataset are 7 different variables: Age, Bmi, sex, smoker, children, region, and charges. The age is a naturally a numerical value indicating the persons age, bmi is the body mass index which can indicate good/poor health in some cases, the sex and smoker variables are Boolean values, true or false for smoking (1 or 0) and male or female for the sex. Children is also a numerical value indicating the number of children the person has and there are 4 regions, northeast, southeast, southwest, and northwest. Finally, the ‘charges’ variable indicates the individual medical cost billed by the insurance company.

To construct the linear regression model, we will use the BMI variable and the age variable with the independent variable being charges to compare how each predictor is affected by those two variables. An AI program would use the BMI and age of a person and try to determine the charges that would be billed by the insurance company of that person.

**In the second section, Data Sources, list the database name and source, the independent variable and the predictors. Show ten rows of the database**.

**Data sources:**

**The database name:** Medical Cost Personal Dataset

**The source:** https://www.kaggle.com/mirichoi0218/insurance/version/1

**Output variable:** Charges billed by insurance company

**Predictors:** BMI and age of the person

Table

Description automatically generated

In the third section, Model Details, use a scatter plot and explain how there is likely a linear relationship between the predictor variables and the outcome variable.

List any predictor or outcome variables that are categorical.

Will you need to create dummy variables?

Chart, scatter chart

Description automatically generated**Model details:**Age versus charges

Looking at this scatterplot we see three distinct linear relationships, which is not exactly what we expected. However, for each of these three distinct “sections” we can see that there is a linear relationship between the age and charges, and as the age increases the charges increase, so it’s a positive correlation.

This is actually a quite interesting result because it shows that even though the age is a factor, there are other important factors as well, which leads to this divide between the three distinct sections.

Chart, scatter chart

Description automatically generated**Model details:**BMI versus charges

For this scatterplot we can also kind of see that there are two types of correlations. They are both linear, but one follows way closer to the x-axis and the other one follows and incremental path as BMI increases, charges increase as well.

This is actually quite interesting, because it shows how BMI is not a very accurate measure of good or bad health. In some cases, for example severe obesity and cases similar to that, BMI can indicate if a person is unhealthy, this would be the linear relationship on the scatterplot where the charges are increasing steadily along with the BMI, however, for the general population, BMI cannot be used accurately to indicate the health of a person, i.e how much should be charged by the insurance company.

Even though the results for the two scatterplots weren’t exactly what we were expecting, they still showed us some very interesting and useful relationships between the predictors and the output variable.