Problem Statement

A client’s requirement is, he wants to predict the insurance charges based on the several parameters. The Client has provided the dataset of the same. As a data scientist, you must develop a model which will predict the insurance charges.

1. Identify your problem statement
   1. Requirements are clear so its comes under supervised learning
   2. Output column is number so its comes under regression
   3. Many inputs are numbers so its comes under Machine learning
2. basic info about the dataset
   1. 6 – columns and 1338- rows
   2. Inputs columns – 5 and output column – 1
   3. Input – 3 number columns and 2 – string columns
   4. Output -1 number columns
3. pre-processing method
   1. Input has 2 string columns, so we need to convert this two column into numbers
   2. Sex columns is converted into 2 column as sex\_male, sex\_female.
   3. We can identify the sex data with one column. Example if sex\_male is 1 then its original value is male, if sex\_male is 0 then original value is female. So we can drop the second column(sex\_female)
   4. Smoker column is converted into 2 column as smoker\_yes, smoker\_no. same as the previous process we can drop the second column(smoker\_no)
4. I am using 4 algorithm for this problem statement
   1. Multiple Linear regression
   2. Support Vector Machine
   3. Decision tree
   4. Random Forest
5. All the research values
   1. R2 score for different algorithm are documented and available in same github repository
6. Best model for this problem statement is Random Forest, because its R2 score is higher then any other model
   1. R2 score – 0.87 (criterion='absolute\_error', max\_features='log2', n\_estimators = 100, random\_state = 0)
   2. This R2 score also not a good model, but this is the highest from other models