Requirement:

To predict the insurance charges based on the several parameters.

Dataset:

The client has provided a dataset comprising the following:

* Age
* Sex
* BMI
* Children
* Smoker
* Charges

The “Charges” is the target which is to be learned and predicted based on other parameter fields.

The total number of records are 1338 and having considered the parameters, it seems to be reasonable.

Three steps:

* Domain:
  + Since the dataset type is number, Machine learning is the domain.
* Learning:
  + Both the input and outputs are clearly defined, and the data are available. Hence it is Supervised learning.
* Fundamental type:
  + The prediction is to find the continuous numerical value, the fundamental type for the supervised learning is “Regression”.

Phases

Learning

* Collection of data
  + The dataset “Test.csv” file is available.
  + This file contains both the parameter fields and the result field with values.
* Preprocessing data
  + The parameter fields “age”, “bmi” and “children” fields are numerical data, and they don’t need to be reprocessed.
  + The fields “sex” and “smoker” are categorical data, and these are to be converted to numerical data.
  + The field “sex” contains two values – male and female. Hence this is ordinal data.
  + The field “smoker” contains two values -yes and no. Hence this is also ordinal data.
  + For the ordinal data, label encoder function is the choice. However, we can go with one hot encoding for these two fields sex and smoker also. Both gives the same results.
* Train and test splitting
  + The total records are 1338. Having considered the input parameters, it is a decent one.
  + We can split the records for training and testing purpose with the proportion of 70:30.
* Train and test, evaluate and save the best model.
  + Though many machine learning algorithms are available, here let us have the following algorithm for the model creation:
    - Multi Linear Regression.
    - Support Vector Regression.
    - Decision Tree Regressor
    - Random Forest Regressor.
  + Let us arrive the r2 score for each algorithm with different hyper tuning parameters if any.

Deployment

* Load the saved model.
* Give input
* Predict
* Call to action.