Problem Statement or Requirement:

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:**

We are going to predict the Insurance amount of a particular person based on the age,sex,BMI,smoker and if he/she has children or not.On checking the dataset, the datas are in numerical values so we can choose the Machine Learning domain. As we going to predict the insurance amount (input and output is labelled clearly) we can proceed with Supervised learning and the output has series of insurance amount we are proceeding with Regression Algorithm

2.) Tell basic info about the dataset (Total number of rows, columns)

Total number of rows=1338

Total number of Columns=6

3.) Mention the pre-processing method if you’re doing any (like converting string to number – nominal data)

Data preprocessing is done for the columns (sex and smoker). Since it is a ordinal data we can easily segregate the data as below

Sex-Male/Female (true,false)

Smoker- (yes/no) (true,false)

4.) Develop a good model with r2\_score. You can use any machine learning algorithm; you can create many models. Finally, you have to come up with a final model.

5.) All the research values (r2\_score of the models) should be documented. (You can make tabulation or screenshot of the results.)

| **Multilinear** |  | 0.789479035 |  |
| --- | --- | --- | --- |
|  |  |  |  |
| **Decision Tree** |  |  |  |
| **criterion** | **splitter** | **max\_depth** | **R2\_Score** |
| squared\_error | best | 1 | 0.6713300243 |
| squared\_error | best | 10 | 0.7363835139 |
| squared\_error | best | 100 | 0.7138641118 |
| squared\_error | *random* | 1 | 0.6713300243 |
| squared\_error | *random* | 10 | 0.7853690931 |
| squared\_error | *random* | 100 | 0.6841949455 |
| *friedman\_mse* | best | 1 | 0.6713300243 |
| *friedman\_mse* | best | 10 | 0.7165162992 |
| *friedman\_mse* | best | 100 | 0.6885718438 |
| *friedman\_mse* | *random* | 1 | 0.6713300243 |
| *friedman\_mse* | *random* | 10 | 0.7852920906 |
| *friedman\_mse* | *random* | 100 | 0.7262066884 |
| *absolute\_error* | best | 1 | 0.6785647099 |
| *absolute\_error* | best | 10 | 0.7433069673 |
| *absolute\_error* | best | 100 | 0.6632394094 |
| *absolute\_error* | *random* | 1 | 0.6785647099 |
| ***absolute\_error*** | ***random*** | **10** | **0.807796695** |
| *absolute\_error* | *random* | 100 | 0.6910207121 |
| *poisson* | best | 1 | 0.6713300243 |
| *poisson* | best | 10 | 0.7540240338 |
| *poisson* | best | 100 | 0.7298263875 |
| *poisson* | *random* | 1 | 0.6713300243 |
| *poisson* | *random* | 10 | 0.8221072806 |
| *poisson* | *random* | 100 | 0.6944355794 |
|  |  |  |  |
|  |  |  |  |
| SVM |  |  |  |
| Hyper parameter | RBF | POLY | SIGMOID |
| C10 | -0.1446789063 | -0.569764757 | -0.09827741296 |
| C100 | -0.4270277175 | -0.6247945892 | -0.2270079322 |
| C1000 | -0.5206048977 | -0.6497742558 | -16.42134225 |
| C2000 | -0.6088377227 | -0.6555626785 | -69.06312413 |
| C3000 | -0.6920757069 | -0.6593669308 | -176.2796292 |
|  |  |  |  |
|  |  |  |  |
| **Random Forest** |  |  |  |
| criterion | n\_estimators | max\_features | Rsquare |
| squared\_error | 1 | sqrt | 0.7356318688 |
| squared\_error | 100 | sqrt | 0.8520006347 |
| squared\_error | 1000 | sqrt | 0.8710271903 |
| squared\_error | 1 | log2 | 0.7356318688 |
| squared\_error | 100 | log2 | 0.8710271903 |
| squared\_error | 1000 | log2 | 0.871922689 |
| absolute\_error | 1 | sqrt | 0.7482332242 |
| absolute\_error | 100 | sqrt | 0.8710685856 |
| **absolute\_error** | **1000** | **sqrt** | **0.8727386925** |
| absolute\_error | 1 | log2 | 0.7482332242 |
| absolute\_error | 100 | log2 | 0.8710685856 |
| absolute\_error | 1000 | log2 | 0.8727386925 |
| friedman\_mse | 1 | sqrt | 0.7356502377 |
| friedman\_mse | 100 | sqrt | 0.8710544016 |
| friedman\_mse | 1000 | sqrt | 0.8718801965 |
| friedman\_mse | 1 | log2 | 0.7356502377 |
| friedman\_mse | 100 | log2 | 0.8710544016 |
| friedman\_mse | 1000 | log2 | 0.8718801965 |
| poisson | 1 | sqrt | 0.7386561586 |
| poisson | 100 | sqrt | 0.8680156985 |
| poisson | 1000 | sqrt | 0.8711905794 |
| poisson | 1 | log2 | 0.7386561586 |
| poisson | 100 | log2 | 0.8680156985 |
| poisson | 1000 | log2 | 0.8711905794 |

6.) Mention your final model, justify why u have chosen the same.

Kindly create Repository in the name Regression Assignment.

The final model I am going to use is Random Forest algorithm and it has the best accuracy(Rsquare value=**0.8727386925**) which is close to 1

Upload all the ipynb and final document in the pdf

Communication is important (How you are representing the document.)