1. Identify your problem statement.

We need to predict the **insurance charges** the requirement is very clear

Stage 1 Domain selection: - This problem statement will come under Machine learning.   
Stage 2: -Input and output is already available and It will come under **supervised learning**

stage 3: Output as number it will false under regression

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

1338 rows × 5 columns

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

Converting categorical data into meaning full numerical data.

Nominal – One hot-encoding “column expansion”

dataset=pd.get\_dummies(dataset,drop\_first=True)

1. 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 final model.

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

1 .Multiple Linear Regression.

R\_score value :- 0.78913454847886

1. Support Vector Machine

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S. No | Hyper Parameter | linear (R Value) | poly (R Value) | sigmoid (R Value) | rdf (Default) (R Value) |
| 2 | C10 | 0.462426338 | 0.038625187 | 0.039440121 | 0.0323806 |
| 3 | C100 | 0.628963203 | 0.616469835 | 0.52684154 | 0.319664545 |
| 4 | C500 | 0.763031105 | 0.828561191 | 0.442956062 | 0.661622633 |
| 5 | C1000 | 0.76483948 | 0.854651559 | 0.212045419 | 0.810719571 |

1. Decision Tree

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | criterion | max\_features | splitter | R Value |
| 1 | squared\_error (default) | default=None | best (default) | 0.703465017 |
| 2 | friedman\_mse | sqrt | random | 0.729071515 |
| 3 | absolute\_error | log2 | random | 0.643444109 |
| 4 | poisson | sqrt | best (default) | 0.719747448 |
| 5 | friedman\_mse | log2 | best (default) | 0.670136582 |
| 6 | absolute\_error | sqrt | best (default) | 0.724470915 |
| 7 | friedman\_mse | log2 | random | 0.680738094 |

1. Random Forest

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | criterion | Hyper Paramter | max\_features | R Value |
| 1 | squared\_error (default) | n\_estimators =100 | log2 | 0.866171806 |
| 2 | friedman\_mse | n\_estimators =100 | sqrt | 0.866171806 |
| 3 | absolute\_error | n\_estimators =500 | sqrt | 0.870266763 |
| 4 | poisson | n\_estimators =1000 | log2 | 0.865902125 |
| 5 | friedman\_mse | n\_estimators =2000 | log2 | 0.866650181 |
| 6 | friedman\_mse | n\_estimators =3000 | log2 | 0.866383892 |
| 7 | friedman\_mse | n\_estimators =5000 | log2 | 0.866763448 |

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