1. Multiple Linear Regression (R-values=0.789479034)

2. Support Vector Machine

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| c | poly | rbf | sigmoid | linear |
| 1 | -0.0756996 | -0.08338238 | -0.0754292 | -0.01010 |
| 10 | 0.038716222 | -0.03227 | 0.039307 | 0.46246 |
| 100 | 0.617956 | 0.3200317 | 0.52761035 | 0.62887 |
| 500 | 0.8263683 | 0.6642984 | 0.444606103 | 0.7631058 |
| 1000 | 0.8566487 | 0.810206485 | 0.2874706 | 0.764931 |
| 2000 | 0.860557 | 0.85477664 | -0.59395 | 0.7440418 |
| 3000 | 0.859893 | 0.866339 | -2.124419478 | 0.789479 |

The SVM Regression use R-value (rbfand(c: 3000) =0.866339).

3. Decision Tree

|  |  |  |  |
| --- | --- | --- | --- |
| Criterion | Max Features | Splitter | R-values |
| squared error | sqrt | best | 0.640659 |
|  | sqrt | random | 0.675400 |
|  | log2 | best | 0.59058 |
|  | log2 | random | 0.5871189 |
|  | None | best | 0.6975829 |
|  | None | random | 0.70436446 |
| friedman\_mse | sqrt | best | 0.71590 |
|  | sqrt | random | 0.694292 |
|  | log2 | best | 0.7109069 |
|  | log2 | random | 0.6606759 |
|  | None | best | 0.700821 |
|  | None | random | 0.7143761823 |
| absolute error | sqrt | best | 0.7202352 |
|  | sqrt | random | 0.704753 |
|  | log2 | best | 0.7252246 |
|  | log2 | random | 0.6750806 |
|  | None | best | 0.66909 |
|  | None | random | 0.68066760 |
| poisson | sqrt | best | 0.7487391 |
|  | sqrt | random | 0.6938914 |
|  | log2 | best | 0.719600 |
|  | log2 | random | 0.62326 |
|  | None | best | 0.7198393 |
|  | None | random | 0.7677767 |

The Decision Tree use R-value is 0.7677767for the ‘Poisson ‘criterion with None and ‘random’.

4. Random Forest

|  |  |  |  |
| --- | --- | --- | --- |
| Criterion | Max Features | N\_Estimators | R-values |
| squared error | sqrt | 50 | 0.8696151 |
|  | sqrt | 100 | 0.8708117957 |
|  | log2 | 50 | 0.8696151 |
|  | log2 | 100 | 0.8708117 |
|  | None | 50 | 0.8509112 |
|  | None | 100 | 0.8549545 |
| friedman\_mse | sqrt | 50 | 0.8702337 |
|  | sqrt | 100 | 0.87086180 |
|  | log2 | 50 | 0.8702337 |
|  | log2 | 100 | 0.870861800 |
|  | None | 50 | 0.853186 |
|  | None | 100 | 0.85514339 |
| absolute error | sqrt | 50 | 0.871684 |
|  | sqrt | 100 | 0.87197643703 |
|  | log2 | 50 | 0.871684 |
|  | log2 | 100 | 0.8719764 |
|  | None | 50 | 0.8541233 |
|  | None | 100 | 0.85318 |
| poisson | sqrt | 50 | 0.86320820 |
|  | sqrt | 100 | 0.8677524 |
|  | log2 | 50 | 0.86320820 |
|  | log2 | 100 | 0.8677524 |
|  | None | 50 | 0.8503250 |
|  | None | 100 | 0.853588 |

The highest R-values in your table are achieved with the following criterion and parameters:

* **Criterion**: absolute error
* **Max Features**: sqrt
* **N\_Estimators**: 100
* **R-value**: 0.87197643703

This combination provides the highest R-value of 0.87197643703.

BEST METHOD OF REGRESSION:

Random Forest:

The highest R-value of 0.87197643703.

OR

Support Vector Machine:

The R-value of 0.866339.

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

Machine Learning

Supervised Learning

Regression

2.) Tell basic info about the dataset (Total number of rows, columns=1338 rows × 6 columns)

 Total **number of rows& columns**: [independent=1338 rows × 5 columns][ dependent= 1338 rows × 1 columns].

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

Convert categorical data (e.g., gender, smoker) into numerical format using techniques like one-hot encoding or label encoding.

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 final model.

Multiple Linear Regression

* R2 Score: 0.789479034

Support Vector Machine Regression

* Kernel: Poly, C: 3000: R2 Score = 0.859893
* Kernel: RBF, C: 3000: R2 Score = 0.866339
* Kernel: Sigmoid, C: 3000: R2 Score = -2.124419478
* Kernel: Linear, C: 3000: R2 Score = 0.789479

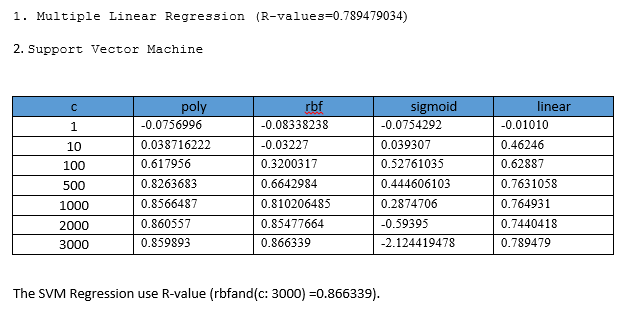
Decision Tree Regression

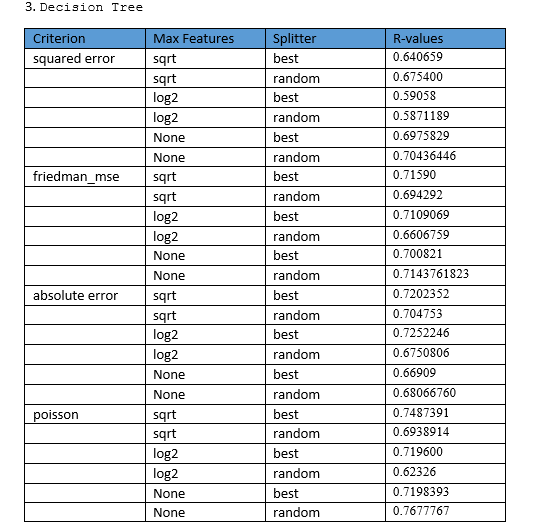
* Criterion: Poisson, Max Features: None, Splitter: Random: R2 Score = 0.7677767

Random Forest Regression

* Criterion: Absolute Error, Max Features: sqrt, N\_Estimators: 100: R2 Score = 0.8719764370

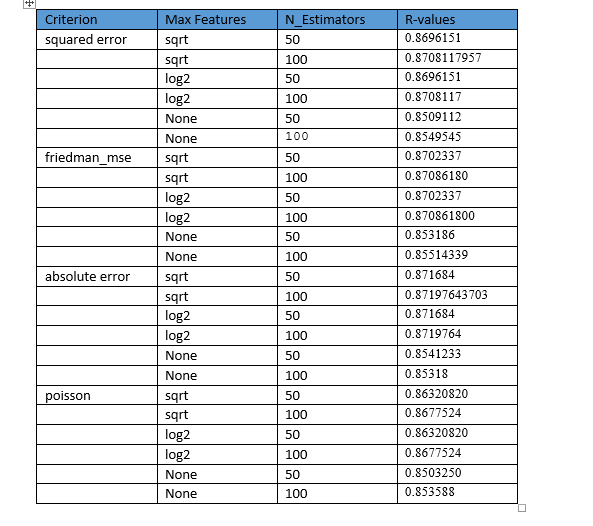
5.) All the research values (r2\_score of the models) should be documented. (You can make tabulation or screenshot of the results.)

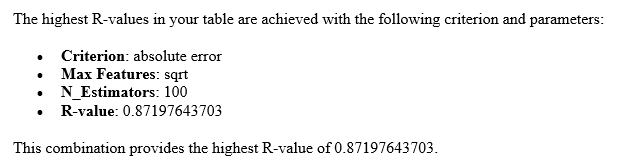












6.) Mention your final model, justify why u have chosen the same.

#### Final Model and Justification

The **Random Forest Regression** model with the following parameters:

* **Criterion**: Absolute Error
* **Max Features**: sqrt
* **N\_Estimators**: 100

This model was chosen because it achieved the highest R2 Score of 0.87197643703 among all the models tested.