1.) Identify your problem statement

Ans:

* Machine Learning
* Supervised Learning
* Regression

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

Ans:

* No of rows: 1338
* No of columns: 6

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

Ans:

Splitting the data into multiple columns as one Hot Encoding for the columns sex and smoker.

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.

Ans : Models are attached in the github

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 Algorithm:**

|  |  |  |
| --- | --- | --- |
| **S.No** | **Model Name** | **R2\_Score** |
| 1. | Mul\_linear\_model | 0.78 |

2. **SVM**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Model Name** | **Kernel** | **R2\_Score** |
| 1. | SVM\_model | rbf | 0.81 |
| 2. | SVM\_model\_linear | Linear | 0.76 |
| 3. | SVM\_model\_poly | poly | 0.85 |
| 4. | SVM\_model\_sigmoid | sigmoid | 0.28 |

**3. Decision Tree:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No** | **Model Name** | **creterion** | **splitter** | **R2\_Score** |
| 1. | d\_tree\_1 | squared\_error | best | 0.70 |
| 2. | d\_tree\_2 | friedman\_mse | best | 0.69 |
| 3. | d\_tree\_3 | absolute\_error | best | 0.66 |
| 4. | d\_tree\_4 | Poisson | best | 0.71 |
| 5. | d\_tree\_5 | squared\_error | random | 0.73 |
| 6. | d\_tree\_6 | friedman\_mse | random | 0.67 |
| 7. | d\_tree\_7 | absolute\_error | random | 0.73 |
| 8. | d\_tree\_8 | Poisson | random | 0.71 |

4. **Random Forest:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **S.No** | **Model Name** | **creterion** | **N\_estimators** | **Max\_leaf\_nodes** | **Max\_samples** | **R2\_Score** |
| 1. | r\_forest\_1 | squared\_error | 100 | 10 | 100 | 0.88 |
| 2. | r\_forest\_2 | friedman\_mse | 100 | 10 | 100 | 0.88 |
| 3. | r\_forest\_3 | absolute\_error | 100 | 10 | 100 | 0.88 |
| 4. | r\_forest\_4 | Poisson | 100 | 10 | 100 | 0.87 |

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

* Final Model are **r\_forest\_1, r\_forest\_2,r\_forest\_3**
* These 3 models provided r2\_score as 0.88 compare to other models
* Compare to other models Random forest algorithm worked well for this dataset.