**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**

Stage 1: Machine Learning (Numeric data)

Stage 2: Supervised Learning (Requirement is clear)

Stage 3: Regression

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

Columns: age, sex, bmi, children, smoker, charges

Total Number of rows: 1339

Input: age, sex, bmi, children, smoker

Output: charges

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

Sex (male, female) and Children (Yes, No) columns are categorical data. Since those are Nominal data, I have used One – hot encoding algorithm to convert into binary column

**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.**

Created below 4 models:

1. Multiple Linear Regression
2. Support Vector Regression
3. Decision Tree Regression
4. Random Forest Regression

**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 Algorithm r2 value: 0.789**

**2. Support Vector Regression**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.No** | **C** | **rbf** | **poly** | **sigmoid** | **linear** |
| **1** | **0.1** | -0.089 | -0.088 | -0.088 | -0.08 |
| **2** | **1** | -0.0833 | -0.075 | -0.075 | -0.01 |
| **3** | **10** | -0.032 | 0.038 | 0.039 | 0.46 |
| **4** | **100** | 0.32 | 0.617 | 0.527 | 0.62 |
| **5** | **500** | 0.66 | 0.82 | 0.44 | 0.763 |
| **6** | **1000** | 0.81 | 0.856 | 0.287 | 0.764 |
| **7** | **2000** | 0.854 | 0.86 | -0.59 | 0.744 |
| **8** | **5000** | 0.874 | 0.859 | -7.5 | 0.741 |

**In SVR, best r2 value is (Kernel=’rbf’,c=5000) = 0.874**

**3. Decision Tree**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No** | **criterion** | **splitter** | **Max\_feature** | **r2** |
| 1 | squared\_error | best | None | 0.69 |
| 2 | random | None | 0.74 |
| 3 | random | sqrt | 0.61 |
| 4 | best | sqrt | 0.322 |
| 5 | best | log2 | 0.72 |
| 6 | random | log2 | 0.66 |
| 7 | friedman\_mse | best | None | 0.697 |
| 8 | random | None | 0.731 |
| 9 | random | sqrt | 0.69 |
| 10 | best | sqrt | 0.633 |
| 11 | best | log2 | 0.747 |
| 12 | random | log2 | 0.504 |
| 13 | absolute\_error | best | None | 0.678 |
| 14 | random | None | 0.746 |
| 15 | best | sqrt | 0.732 |
| 16 | random | sqrt | 0.744 |
| 17 | best | log2 | 0.67 |
| 18 | random | log2 | 0.684 |
| 19 | poisson | best | None | 0.729 |
| 20 | random | None | 0.65 |
| 21 | best | sqrt | 0.758 |
| 22 | random | sqrt | 0.651 |
| 23 | best | log2 | 0.705 |
| 24 | random | log2 | 0.651 |

In Decision Tree, best r2 value is **0.758**(criterion=poisson, Splitter=best, max\_features=sqrt)

1. **Random Forest**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No** | **criterion** | **n\_estimators** | **Max\_feature** | **r2** |
| **1** | squared\_error | 10 | None | 0.833 |
| **2** | 10 | sqrt | 0.852 |
| **3** | 10 | log2 | 0.852 |
| **4** | 50 | None | 0.849 |
| **5** | 50 | sqrt | 0.869 |
| **6** | 50 | log2 | 0.869 |
| **7** | 100 | None | 0.853 |
| **8** | 100 | sqrt | 0.87102 |
| **9** | 100 | log2 | 0.87102 |
| **10** | absolute\_error | 10 | None | 0.835 |
| **11** | 10 | sqrt | 0.857 |
| **12** | 10 | log2 | 0.857 |
| **13** | 50 | None | 0.852 |
| **14** | 50 | sqrt | 0.87 |
| **15** | 50 | log2 | 0.87 |
| **16** | 100 | None | 0.852 |
| **17** | 100 | sqrt | 0.87106 |
| **18** | 100 | log2 | 0.87106 |
| **19** | friedman\_mse | 10 | None | 0.833 |
| **20** | 10 | sqrt | 0.85 |
| **21** | 10 | log2 | 0.85 |
| **22** | 50 | None | 0.85 |
| **23** | 50 | sqrt | 0.87 |
| **24** | 50 | log2 | 0.87 |
| **25** | 100 | None | 0.852 |
| **26** | 100 | sqrt | 0.87105 |
| **27** | 100 | log2 | 0.87105 |
| **28** | poisson | 10 | None | 0.831 |
| **29** | 10 | sqrt | 0.854 |
| **30** | 10 | log2 | 0.854 |
| **31** | 50 | None | 0.849 |
| **32** | 50 | sqrt | 0.863 |
| **33** | 50 | log2 | 0.863 |
| **34** | 100 | None | 0.852 |
| **35** | 100 | sqrt | 0.868 |
| **36** | 100 | log2 | 0.868 |

In Random Forest, best r2 value is **0.871**(criterion=’absolute\_error’, n\_estimators=100, max\_features= sqrt and log2)

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

Based on R2 values for 4 models, I chose Support Vector Regression Algorithm which gives best r2 value for this dataset as **0.874**((Kernel=’rbf’,c=5000)