**Regression Assignment**

**Problem Statement:**

1. The given dataset contains 1338 rows and 6 columns. The column captions are age, sex, bmi, children,smoker, charges. In this **charges is our output** and the remaining columns are our input parameters.
2. The dataset contains alphabetical values that Machine language does not accept, so we have to convert the nominal data into numerical values by using **One Hot encoding**
3. We have a clear requirement, the client shared the proper dataset which contains multiple input values as column and one output. So we cant use Simple Linear Regression

* The dataset has numerical values, hence we are dealing with numerical values this can come under **Machine Learning.**
* The dataset has proper input and output so it comes under **Supervised Learning**
* The end goal is to predict the charges which is numerical output hence it comes under **Regression**

**R2 Score:**

1. **Multi Linear Regression**

R2 score = 0.78

1. **Support Vector Machine**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Kernel** | **R2 Score** |
| 1 | linear | -0.00967358 |
| 2 | poly | -0.07564862 |
| 3 | rbf | -0.08347749 |
| 4 | sigmoid | -0.07536003 |
| 5 | precomputed | NA |

1. **Decision Tree**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl. No.** | **criterion** | **splitter** | **max\_features** | **R2\_Score** |
| 1 | *squared\_error* | *best* | *sqrt* | 0.714534506 |
| 2 | *friedman\_mse* | *best* | *sqrt* | 0.69721735 |
| 3 | *absolute\_error* | *best* | *sqrt* | 0.743012313 |
| 4 | *poisson* | *best* | *sqrt* | 0.7099544 |
| 5 | *squared\_error* | *best* | *log2* | 0.719080373 |
| 6 | *friedman\_mse* | *best* | *log2* | 0.619377637 |
| 7 | *absolute\_error* | *best* | *log2* | 0.49497597 |
| 8 | *poisson* | *best* | *log2* | 0.731257944 |
| 9 | *squared\_error* | *best* | *None* | 0.690955914 |
| 10 | *friedman\_mse* | *best* | *None* | 0.69204734 |
| 11 | *absolute\_error* | *best* | *None* | 0.676091468 |
| 12 | *poisson* | *best* | *None* | 0.716942048 |
| 13 | *squared\_error* | *random* | *sqrt* | 0.666921259 |
| 14 | *friedman\_mse* | *random* | *sqrt* | 0.636849143 |
| 15 | *absolute\_error* | *random* | *sqrt* | 0.674345537 |
| 16 | *poisson* | *random* | *sqrt* | 0.682472822 |
| 17 | *squared\_error* | *random* | *log2* | 0.6390651 |
| 18 | *friedman\_mse* | *random* | *log2* | 0.636494458 |
| 19 | *absolute\_error* | *random* | *log2* | 0.728501483 |
| 20 | *poisson* | *random* | *log2* | 0.71359545 |
| 21 | *squared\_error* | *random* | *None* | 0.694404794 |
| 22 | *friedman\_mse* | *random* | *None* | 0.636741108 |
| 23 | *absolute\_error* | *random* | *None* | 0.726125491 |
| 24 | *poisson* | *random* | *None* | 0.653306708 |

1. **Random Forest**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl. No.** | **criterion** | **max\_features** | **n\_estimators** | **R2\_Score** |
| 1 | *squared\_error* | *sqrt* | *50* | 0.87036837 |
| 2 | *friedman\_mse* | *sqrt* | *50* | 0.87234024 |
| 3 | *absolute\_error* | *sqrt* | *100* | 0.87273511 |
| 4 | *poisson* | *sqrt* | *100* | 0.87130758 |
| 5 | *squared\_error* | *log2* | *100* | 0.87304159 |
| 6 | *friedman\_mse* | *log2* | *100* | 0.87157384 |
| 7 | *absolute\_error* | *log2* | *50* | 0.87191888 |
| 8 | *poisson* | *log2* | *100* | 0.87031301 |
| 9 | *squared\_error* | *None* | *100* | 0.85769684 |
| 10 | *friedman\_mse* | *None* | *50* | 0.85125364 |
| 11 | *absolute\_error* | *None* | *100* | 0.85127313 |
| 12 | *poisson* | *None* | *100* | 0.85302983 |

1. Based on the below R2 Score we can finalize the **Random Forest Algorithm** with the parameter of n\_estimators=100,criterion='squared\_error',max\_features='log2' is the best one and this **model is** saved with the name of **AssRFR.sav**