**Regression Assignment**

1. Problem statement:

* Machine learning
* Supervised
* Regression

1. About dataset:

There are 1338 rows and 6 columns. We have 5 input columns, age, sex bmi, children and smoker. The output is insurance charge.

1. Pre-processing:

The nominal data text columns sex and smoker were pre-processed using one hot encoding method. By this, we got 8 columns.

1. Best model:

Multiple linear regression R2 score = 0.78913454847886

**Support Vector Machine:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.No** | **Penality** | **Linear** | **Poly** | **rbf** | **sigmoid** |
| 1 | C10 | 0.566512707 | 0.159392113 | -0.018374382 | 0.073055564 |
| 2 | C100 | 0.635950358 | 0.750819225 | 0.390136163 | 0.527560378 |
| 3 | C1000 | 0.744090887 | 0.860584677 | 0.828720635 | 0.143775586 |
| 4 | C2000 | 0.741423887 | 0.860181265 | 0.859691365 | -2.584034602 |
| 5 | C3000 | 0.741423169 | 0.860011623 | 0.8682787351 | -6.826188554 |

SVM regression R2 score = 0.8682787351

**Decision Tree:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No** | **Criterion** | **Splitter** | **Max\_features** | **r\_score** |
| 1 | squared\_error | random | sqrt | 0.661619922 |
| 2 | squared\_error | random | log2 | 0.660992915 |
| 3 | squared\_error | best | sqrt | 0.68484473 |
| 4 | squared\_error | best | log2 | 0.784919417 |
| 5 | friedman\_mse | random | sqrt | 0.67985235 |
| 6 | friedman\_mse | random | log2 | 0.681573675 |
| 7 | friedman\_mse | best | sqrt | 0.761935029 |
| 8 | friedman\_mse | best | log2 | 0.733654818 |
| 9 | absolute\_error | random | sqrt | 0.660931849 |
| 10 | absolute\_error | random | log2 | 0.647477574 |
| 11 | absolute\_error | best | sqrt | 0.726554555 |
| 12 | absolute\_error | best | log2 | 0.679438124 |
| 13 | poisson | random | sqrt | 0.689290497 |
| 14 | poisson | random | log2 | 0.655475049 |
| 15 | poisson | best | sqrt | 0.77505249 |
| 16 | poisson | best | log2 | 0.727416687 |

Decision tree R2 score = 0.784919417

**Random Forest:**

|  |  |  |
| --- | --- | --- |
| **Criterion** | **Max\_features** | **R\_score** |
| squared\_error | sqrt | 0.856585487 |
| squared\_error | log2 | 0.856585487 |
| absolute\_error | sqrt | 0.864012095 |
| absolute\_error | log2 | 0.864012095 |
| friedman\_mse | sqrt | 0.856585487 |
| friedman\_mse | log2 | 0.856585487 |
| poisson | sqrt | 0.857564396 |
| poisson | log2 | 0.857564396 |
| **Default features** | | |
| squared\_error | None | 0.851709179 |

Random Forest R2 score = 0.864012095

Best model is Random Forest

|  |  |
| --- | --- |
| **Criterion** | **R\_score** |
| absolute\_error | 0.864012095 |

1. Justification:

I choose Random Forest as the best model though the SVM regression value is little higher. This is because in SVM regression the penalty is higher and for any input we get the intercept value as output.