**Regression Assignment**

**Problem Statement:**

Wants to predict the insurance charge based on several parameters provided the dataset of the same.

**About Dataset:**

Dataset contains 1338 rows with 6 columns among which two columns (Sex and Smoker) are categorical data of Nominal type.

**Data Preprocessing:**

* The nominal data are converted to numeric data using one hot encoding.
* IN SVM the data is standardized before making the model for more accuracy and to avoid the model getting biased to the large values

**Domain Selection:**

* Machine Learning

**LEARNING SELECTION:**

* Supervised Learning

* Regression

**ALGORITHMS**

* **Multiple Linear Regression - R Square Valure:** 0.789479
* **Support Vector Machine:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.I.NO | KERNEL | C10 | C100 | C1000 |
| 1 | linear | 0.46246 | 0.62887 | 0.76493 |
| 2 | Poly | 0.03871 | 0.61795 | 0.8566 |
| 3 | rbf | -0.03227 | 0.32003 | 0.81020 |
| 4 | sigmoid | 0.03930 | 0.52761 | 0.28747 |

**R Square Value:** 0.8566**(with Kernel = Poly & C = 1000)**

* **Decision Tree:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.I.NO | CRITERION | SPLITTER | MAX\_FEATURES | R\_VALUE |
| 1 | squared\_error | best | sqrt | 0.77630 |
| 2 | squared\_error | best | log2 | 0.58934 |
| 3 | squared\_error | best | None | 0.70127 |
| 4 | squared\_error | random | sqrt | 0.68037 |
| 5 | squared\_error | random | log2 | 0.65285 |
| 6 | squared\_error | random | None | 0.71342 |
| 7 | friedman\_msc | best | sqrt | 0.60492 |
| 8 | friedman\_msc | best | log2 | 0.70603 |
| 9 | friedman\_msc | best | None | 0.70098 |
| 10 | friedman\_msc | random | sqrt | 0.59869 |
| 11 | friedman\_msc | random | log2 | 0.67131 |
| 12 | friedman\_msc | random | None | 0.71472 |
| 13 | absolute\_error | best | sqrt | 0.74588 |
| 14 | absolute\_error | best | log2 | 0.64896 |
| 15 | absolute\_error | best | None | 0.66014 |
| 16 | absolute\_error | random | sqrt | 0.66707 |
| 17 | absolute\_error | random | log2 | 0.64816 |
| 18 | absolute\_error | random | None | 0.72028 |
| 19 | poisson | best | sqrt | 0.71275 |
| 20 | poisson | best | log2 | 0.75088 |
| 21 | poisson | best | None | 0.72278 |
| 22 | poisson | random | sqrt | 0.71997 |
| 23 | poisson | random | log2 | 0.72025 |
| 24 | poisson | random | None | 0.72827 |

**R Square Value:** 0.77630**(Criterion = squared\_error, Splitter = best &**

**MAX\_FEATURES = sqrt)**

* **Random Forest:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.I.NO | CRITERION | MAX\_FEATURES | N\_ESTIMATORS | R\_VALUE |
| 1 | Squared\_error | sqrt | 50 | 0.86899 |
| 2 | Squared\_error | log2 | 50 | 0.86822 |
| 3 | Squared\_error | None | 50 | 0.84952 |
| 4 | Squared\_error | sqrt | 100 | 0.87245 |
| 5 | Squared\_error | log2 | 100 | 0.87193 |
| 6 | Squared\_error | None | 100 | 0.85475 |
| 7 | absolute\_error | sqrt | 50 | 0.87263 |
| 8 | absolute\_error | log2 | 50 | 0.86865 |
| 9 | absolute\_error | None | 50 | 0.85642 |
| 10 | absolute\_error | sqrt | 100 | 0.87228 |
| 11 | absolute\_error | log2 | 100 | 0.87228 |
| 12 | absolute\_error | None | 100 | 0.85477 |
| 13 | friedsman\_mse | sqrt | 50 | 0.87206 |
| 14 | friedsman\_mse | log2 | 50 | 0.86660 |
| 15 | friedsman\_mse | None | 50 | 0.85496 |
| 16 | friedsman\_mse | sqrt | 100 | 0.86795 |
| 17 | friedsman\_mse | log2 | 100 | 0.86971 |
| 18 | friedsman\_mse | None | 100 | 0.85358 |
| 19 | poisson | sqrt | 50 | 0.87207 |
| 20 | poisson | log2 | 50 | 0.87150 |
| 21 | poisson | None | 50 | 0.85805 |
| 22 | poisson | sqrt | 100 | 0.86736 |
| 23 | poisson | log2 | 100 | 0.87056 |
| 24 | poisson | None | 100 | 0.85563 |

**R Square Value:** 0.87245**(Criterion = squared\_error, Estimators = 100 &**

**MAX\_FEATURES = sqrt)**

**Finalized Model:**

**Random Forest ((Criterion = squared\_error, Estimators = 100 & MAX\_FEATURES = sqrt) as it performs better than any other model created with R squared value of** 0.87245