**Assignment – Regression Algorithm**

**1. Problem Statement:**

Todevelop a model which will predict the insurance charges is the Problem Statement.

**2. Info about the dataset:**

Total number of rows - 1338

Total number of columns – 6

**3. Pre-processing method:**

By using One Hot Encoding the categorical column (sex and smoker) is converted into nominal data as 0s and 1s.

**4. Good Model:**

After creating many models, Random Forest is the best model among them. Which gives the highest R2 value that uses criterion(Friedman mse), Max Features(log2) and n\_estimators(100) = 0.873736069

**5. To find the best model by using R2 value:**

**i. Multiple Linear Regression:**

R2 value - 0.7894790349867009

**ii. Support Vector Machine(SVM):**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.NO** | **HYPER PARAMETER** | **LINEAR** | **RBF** | **POLY** | **SIGMOID** |
| 1 | C0.01 | -0.088831334 | -0.089645537 | -0.089568285 | -0.089565016 |
| 2 | C10 | 0.46246841423396834 | -0.032273294 | 0.038716223 | 0.039307144 |
| 3 | C100 | 0.6288792857320369 | 0.3200317832050831 | 0.617956962 | 0.527610355 |
| 4 | C500 | 0.763105805 | 0.664298465 | 0.8263683541268882 | 0.44460610338694795 |
| 5 | C1000 | 0.764931174 | 0.8102064851758545 | 0.856648768 | 0.287470695 |
| 6 | C2000 | 0.7440418308108339 | 0.8547766433500156 | 0.860557928 | -0.593950973 |
| 7 | C3000 | 0.7414236599247678 | 0.8663393965062866 | 0.859893008 | -2.124419479 |

The SVM Regression uses criterion (rbf) and Hyper Parameter(C3000)

= 0.8663393965062866

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.NO** | **CRITERION** | **SPLITTER** | **MAX FEATURES** | **R2 VALUE** |
| 1 | Squared\_error | best | sqrt | 0.7380275138429635 |
| 2 | Squared\_error | random | sqrt | 0.675767737 |
| 3 | Squared\_error | best | log2 | 0.634146098 |
| 4 | Squared\_error | random | log2 | 0.710932355 |
| 5 | Squared\_error | best | none | 0.680721449 |
| 6 | Squared\_error | random | none | 0.667059543 |
| 7 | Friedman\_mse | best | sqrt | 0.725175222 |
| 8 | Friedman\_mse | random | sqrt | 0.70330151 |
| 9 | Friedman\_mse | best | log2 | 0.755616392 |
| 10 | Friedman\_mse | random | log2 | 0.766492961 |
| 11 | Friedman\_mse | best | none | 0.703400927 |
| 12 | Friedman\_mse | random | none | 0.732124852 |
| 13 | Absolute\_error | best | sqrt | 0.66417557 |
| 14 | Absolute\_error | random | sqrt | 0.769800334 |
| 15 | Absolute\_error | best | log2 | 0.662644181 |
| 16 | Absolute\_error | random | log2 | 0.695186144 |
| 17 | Absolute\_error | best | none | 0.667885053 |
| 18 | Absolute\_error | random | none | 0.730315323 |
| 19 | Poisson | best | sqrt | 0.717822807 |
| 20 | Poisson | random | sqrt | 0.712919245 |
| 21 | Poisson | best | log2 | 0.72656031 |
| 22 | Poisson | random | log2 | 0.693120132 |
| 23 | Poisson | best | none | 0.734818476 |
| 24 | Poisson | random | none | 0.7052511 |

**iii. Decision Tree:**

The Decision Tree uses criterion(Absolute error), splitter(random), Max Features(sqrt)= 0.769800334

**iv. Random Forest:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.NO** | **N\_ESTIMATORS** | **CRITERION** | **MAX FEATURES** | **R2 VALUE** |
| 1 | 50 | Squared error | sqrt | 0.869834762 |
| 2 | 100 | Squared error | sqrt | 0.873273486 |
| 3 | 50 | Squared error | log2 | 0.8704457623136355 |
| 4 | 100 | Squared error | log2 | 0.868525199 |
| 5 | 50 | Squared error | none | 0.851892838 |
| 6 | 100 | Squared error | none | 0.851509664 |
| 7 | 50 | Friedman mse | sqrt | 0.864156498 |
| 8 | 100 | Friedman mse | sqrt | 0.871049004 |
| 9 | 50 | Friedman mse | log2 | 0.86625764 |
| 10 | 100 | Friedman mse | log2 | 0.873736069 |
| 11 | 50 | Friedman mse | none | 0.857586168 |
| 12 | 100 | Friedman mse | none | 0.858457834 |
| 13 | 50 | Absolute error | sqrt | 0.870066487 |
| 14 | 100 | Absolute error | sqrt | 0.87196547 |
| 15 | 50 | Absolute error | log2 | 0.868405445 |
| 16 | 100 | Absolute error | log2 | 0.872257395 |
| 17 | 50 | Absolute error | none | 0.850109347 |
| 18 | 100 | Absolute error | none | 0.853181607 |
| 19 | 50 | Poisson | sqrt | 0.867468293 |
| 20 | 100 | Poisson | sqrt | 0.870037705 |
| 21 | 50 | Poisson | log2 | 0.863398648 |
| 22 | 100 | Poisson | log2 | 0.868121134 |
| 23 | 50 | Poisson | none | 0.85192413 |
| 24 | 100 | Poisson | none | 0.851671291 |

The Random Forest uses criterion(Friedman mse), Max Features(log2) and n\_estimators(100) = 0.873736069

**6. Final Model:**

The Random Forest is the final model, when compared with other models which is predicted by the R2 value that uses criterion(Friedman mse), Max Features(log2) and n\_estimators(100) = 0.873736069