REGRESSION ASSIGNMENT

**Problem statement:**

To predict the insurance charges based on the parameters in the dataset given by the client.

**Dataset:**

The dataset contains 1338 rows and 6 columns.

There are 5 input columns – Age, Sex, BMI, Children and Smoker.

The output column is ‘Charges’.

Since the inputs are numerical, the problem comes under Machine learning domain and the ouput is also numerical, it falls under supervised learning- regression.

Two nominal columns Sex and Smoker are converted to numerical columns using ‘One Hot Encoding’.

The various r2 scores by different models are tabulated below.

Multiple linear regression:

R2 score: 0.7894

Support vector machine – Regression:

|  |  |  |  |
| --- | --- | --- | --- |
| S.NO | KERNEL | GAMMA | R2 SCORE |
| 1 | rbf | Scale | -0.0833 |
| 2 | Rbf | auto | -0.0833 |
| 3 | Linear | Scale | -0.0101 |
| 4 | Linear | auto | -0.0101 |
| 5 | Poly | Scale | -0.0756 |
| 6 | Poly | auto | -0.0756 |
| 7 | Sigmoid | Scale | -0.0754 |
| 8 | Sigmoid | auto | -0.0754 |

**Decision Tree:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.NO** | **CRITERION** | **SPLITTER** | **MAX\_FEATURES** | **R2 SCORE** |
| 1 | squared\_error | Best | none | 0.6797 |
| 2 | squared\_error | Best | sqrt | 0.6951 |
| 3 | squared\_error | Best | Log2 | 0.7371 |
| 4 | squared\_error | Random | none | 0.7419 |
| 5 | squared\_error | Random | sqrt | 0.6797 |
| 6 | squared\_error | Random | Log2 | 0.6941 |
| 7 | Friedman\_mse | Best | none | 0.6921 |
| 8 | Friedman\_mse | Best | sqrt | 0.7360 |
| 9 | Friedman\_mse | Best | Log2 | 0.7256 |
| 10 | Friedman\_mse | Random | none | 0.7499 |
| 11 | Friedman\_mse | Random | sqrt | 0.6591 |
| 12 | Friedman\_mse | Random | Log2 | 0.6493 |
| 13 | Absolute\_error | Best | none | 0.6786 |
| 14 | Absolute\_error | Best | sqrt | 0.6777 |
| 15 | Absolute\_error | Best | Log2 | 0.6980 |
| 16 | Absolute\_error | Random | none | 0.7195 |
| 17 | Absolute\_error | Random | sqrt | 0.6804 |
| 18 | Absolute\_error | Random | Log2 | 0.7409 |
| 19 | Poisson | Best | none | 0.7293 |
| 20 | Poisson | Best | sqrt | 0.7196 |
| 21 | Poisson | Best | Log2 | 0.7636 |
| 22 | Poisson | Random | none | 0.6905 |
| 23 | Poisson | Random | sqrt | 0.6543 |
| 24 | Poisson | Random | Log2 | 0.6449 |

**Random Forest Regression:**

|  |  |  |  |
| --- | --- | --- | --- |
| S.NO | CRITERION | MAX FEATURES | R2 SCORE |
| 1 | Squared error | None | 0.8498 |
| 2 | Squared error | Sqrt | 0.8695 |
| 3 | Squared error | Log 2 | 0.8695 |
| 4 | Friedman mse | None | 0.8500 |
| 5 | Friedman mse | Sqrt | 0.8702 |
| 6 | Friedman mse | Log 2 | 0.8702 |
| 7 | Absolute error | None | 0.8526 |
| 8 | Absolute error | Sqrt | 0.8708 |
| 9 | Absolute error | Log 2 | 0.8708 |
| 10 | poisson | None | 0.8491 |
| 11 | poisson | Sqrt | 0.8632 |
| 12 | poisson | Log 2 | 0.8632 |

The models give poor performance, the best score being 0.87.

Under the given circumstance, Random Forest Regression is the best model.