**REGRESSION ASSIGNMENT**

**Problem Statement**

To Predict the insurance charges based on several parameters

**Three Stages**

Stage-1 : Dataset is a csv file with rows and columns so Machine Learning is applicable

Stage 3: Input and output are clear so it is supervised Learning

Stage 3: Output is Numerical Value so regression and multiple inputs and single output so the model can be created with Multiple Linear regression, SVM, Decision Tree, Random Forest Algorithms.

**Dataset Basic information**

No of Columns : 6

No of Rows :1339

**Output/Dependent/Target Variable/Class**

Insurance Charges

**Input/Independent/Features**

Age, Sex, BMI, Children, Smoker

* Age,BMI,Children are Numerical data
* Sex, Smoker are non-numerical/Categorical data

**Pre-Processing**

Sex, Smoker are non-numerical/Categorical data and it is non-ordinal which needs to be converted to numerical data as machine learning algorithms cannot handle non-numerical/Categorical data

**Multiple Linear Regression R2 Score :** 0.7894790349867009

**SVM Evaluation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Regularisation Parameter/Kernel** | **Linear** | **rbf** | **poly** | **Sigmoid** |
| c=10 | 0.4624684 | -0.03227 | 0.038716 | 0.039307 |
| c=100 | 0.6288793 | 0.320032 | 0.617957 | 0.527610 |
| c=1000 | 0.7649312 | 0.810206 | 0.856649 | 0.287471 |
| c=2000 | 0.7440418 | 0.854777 | 0.860558 | -0.593950 |
| c=3000 | 0.7414237 | 0.866339 | 0.859893 | -2.124421 |
| c=4000 | 0.7414199 | 0.871741 | 0.860005 | 0.860005 |

**SVM has R2 Score near to 1 when C=3000 ,kernel=rbf is** 0.866339

**Decision Tree Evaluation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No** | **Criterion** | **Splitter** | **Max\_Features** | **Rvalue** |
| 1 | Squared\_Error | Best | Sqrt | 0.695508 |
| 2 | Squared\_Error | random | Sqrt | 0.664045 |
| 3 | Squared\_Error | Best | log2 | 0.729366 |
| 4 | Squared\_Error | random | log2 | 0.731787 |
| 5 | Squared\_Error | Best | None | 0.691813 |
| 6 | Squared\_Error | random | None | 0.745395 |
| 7 | friedman\_mse | Best | None | 0.719651 |
| 8 | friedman\_mse | random | None | 0.684381 |
| 9 | friedman\_mse | Best | Sqrt | 0.693414 |
| 10 | friedman\_mse | random | Sqrt | 0.662073 |
| 11 | friedman\_mse | Best | log2 | 0.759201 |
| 12 | friedman\_mse | random | log2 | 0.517504 |
| 13 | absolute\_error | Best | log2 | 0.747465 |
| 14 | absolute\_error | random | log2 | 0.71795 |
| 15 | absolute\_error | Best | Sqrt | 0.705268 |
| 16 | absolute\_error | random | Sqrt | 0.680827 |
| 17 | absolute\_error | Best | None | 0.673937 |
| 18 | absolute\_error | random | None | 0.757475 |
| 19 | poisson | Best | None | 0.729177 |
| **S.No** | **Criterion** | **Splitter** | **Max\_Features** | **Rvalue** |
| 20 | poisson | random | None | 0.667646 |
| 21 | poisson | Best | Sqrt | 0.673144 |
| 22 | poisson | random | Sqrt | 0.635689 |
| 23 | poisson | Best | log2 | 0.627296 |
| 24 | poisson | random | log2 | 0.656989 |

**Decision tree has R2****Score near to 1 when criterion=”friedman\_mse”, splitter=”best”, Max\_features=log2 that is R2-score=**0.759201

**Random Forest Evaluation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No** | **Criterion** | **N\_Estimators** | **Max\_Features** | **Rvalue** |
| 1 | Squared\_Error | 50 | Sqrt | 0.8694982 |
| 2 | Squared\_Error | 100 | Sqrt | 0.8709835 |
| 3 | Squared\_Error | 50 | log2 | 0.8694982 |
| 4 | Squared\_Error | 100 | log2 | 0.8709835 |
| 5 | Squared\_Error | 50 | None | 0.8496064 |
| 6 | Squared\_Error | 100 | None | 0.8535522 |
| 7 | friedman\_mse | 50 | None | 0.8497042 |
| 8 | friedman\_mse | 100 | None | 0.8537519 |
| 9 | friedman\_mse | 50 | Sqrt | 0.8704947 |
| 10 | friedman\_mse | 100 | Sqrt | 0.8712499 |
| 11 | friedman\_mse | 50 | log2 | 0.8704947 |
| 12 | friedman\_mse | 100 | log2 | 0.8712499 |
| 13 | absolute\_error | 50 | log2 | 0.8715368 |
| 14 | absolute\_error | 100 | log2 | 0.8713498 |
| 15 | absolute\_error | 50 | Sqrt | 0.8715368 |
| 16 | absolute\_error | 100 | Sqrt | 0.8713498 |
| 17 | absolute\_error | 50 | None | 0.8536495 |
| 18 | absolute\_error | 100 | None | 0.8526642 |
| 19 | poisson | 50 | None | 0.8493333 |
| **S.No** | **Criterion** | **N\_Estimators** | **Max\_Features** | **Rvalue** |
| 20 | poisson | 100 | None | 0.8527751 |
| 21 | poisson | 50 | Sqrt | 0.8632444 |
| 22 | poisson | 100 | Sqrt | 0.8680235 |
| 23 | poisson | 50 | log2 | 0.8632444 |
| 24 | poisson | 100 | log2 | 0.8680235 |

**Random Forest has R2 Score near to 1 when criterion=”absolute\_error”, N\_Estimators=50, Max\_features=log2 that is R2-score=**0.8715368

**Summary of Evaluation:**

* Multiple Linear Regression R2 Score : 0.7894790349867009
* SVM has R2 Score near to 1 when C=3000 ,kernel=rbf is 0.866339
* Decision tree has R2 Score near to 1 when criterion=”friedman\_mse”, splitter=”best”, Max\_features=log2 that is R2-score=0.759201
* Random Forest has R2-Score near to 1 when criterion=”absolute\_error”, N\_Estimators=50, Max\_features=log2 that is R2-score=0.8715368

**Best Model:**

Random Forest with R2-score=0.8715368 for the parameters criterion=”absolute\_error”, N\_Estimators=50, Max\_features=log2.

**Result:**

The best model has been saved and deployed.