**Assignment-Regression Assignment**

**1.Identify the problem statement.**

Looking at the dataset,most of the feature values are in numbers,so machine learning can be chosen.

Every input as corresponding output as it comes under supervised Learning

The output to be predicted is insurance charges(in numbers) is clearly stated in the requirement.so it is regression problem.

**2.Tell basic information about the dataset.**

The dataset consists of 5 input features(age,sex,bmi,children,smoker) and 1 output feature(insurance charges)

The dimension of the dataset is 1338\*6

Except sex and smoker all other feature values are in numbers

Sex and smoker consists of nomial values ,so one hot encoding is to be done.

**3.Mention the preprocessing method if you’re doing any (like converting string to numbers- nominal data)**

The feature values for the input features Sex and smoker are in strings(nominal –Column cannot be compared) ,so it should be converted to numbers by one hot encoding.

**4.Develope a good model with r2 score .you can use any machine learning algorithm ,you can create many models .Finally you have to come up with final model.**

The given dataset is tried with machine learning models like multiple linear regression ,Support vector machine,Decision Tree and Random forest

The results are tabulated .

Support vector machine is seemed to perform low ,so Standardization technique is applied, and the changes in the result is found significant.

On the whole Randomforest algorithm produced r2 score of 0.87730

**5.All the research values should be documented**

**1.MULTIPLE LINEAR REGRESSION**

R2 Score-0.78947

**2. SUPPORT VECTORE MACHINE**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Kernel | Linear | RBF | Poly | Sigmoid |
| Cost parameter C1.0 | -0.11166 | -0.08842 | 0.03871 | -0.08994 |
| C10 | 0.46246 | -0.08196 | -0.09311 | -0.09078 |
| C100 | 0.54328 | -0.12480 | -0.09976 | -0.11814 |
| C500 | 0.62704 | -0.12464 | -0.08202 | -0.45629 |
| C1000 | 0.63403 | -0.11749 | -0.05550 | -1.66590 |
| C1500 | 0.63942 | -0.11238 | -0.02873 | -3.31637 |
| C2000 | 0.68932 | 0.85477() | 0.86055 | -5.61643 |
| C3000 | 0.71353 | -0.10242 | 0.02344 | -12.01904 |

SVM Accuracy for various Kernels(linear,rbf.poly,sigmoid)

Achieved 0.86055 by applying standardization method.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | CRITERION | SPLITTER | MAX FEATURES | R2 value |
| 1 | Squared\_error | best | Sqrt | 0.72684 |
| 2 | Squared\_error | random | Sqrt | 0.73273 |
| 3 | Squared\_error | best | Log2 | 0.74238 |
| 4 | Squared\_error | random | Log2 | 0.65246 |
| 5 | Squared\_error | best | none | 0.67952 |
| 6 | Squared\_error | random | none | 0.74138 |
| 7 | Friedman\_mse | best | Sqrt | 0.69813 |
| 8 | Friedman\_mse | random | Sqrt | 0.66317 |
| 9 | Friedman\_mse | best | Log2 | 0.69000 |
| 10 | Friedman\_mse | random | Log2 | 0.66770 |
| 11 | Friedman\_mse | best | none | 0.69841 |
| 12 | Friedman\_mse | random | none | 0.63324 |
| 13 | Absolute\_error | best | Sqrt | 0.55257 |
| 14 | Absolute\_error | random | Sqrt | 0.64926 |
| 15 | Absolute\_error | best | Log2 | 0.71502 |
| 16 | Absolute\_error | random | Log2 | 0.66873 |
| 17 | Absolute\_error | best | none | 0.66852 |
| 18 | Absolute\_error | random | none | 0.65994 |
| 19 | Poisson | best | Sqrt | 0.70963 |
| 20 | Poisson | random | Sqrt | 0.60963 |
| 21 | Poisson | best | Log2 | 0.72431 |
| 22 | Poisson | random | Log2 | 0.65681 |
| 23 | Poisson | best | none | 0.71701 |
| 24 | Poisson | random | none | 0.77754 |

**3.Decision tree parameters criterion ,splitter,max\_features**

**4.Random Forest Hyperparameters:-** n\_estimators(50,100),Criterion(squared\_error,friedman\_mse,absolute\_error,poisson),Max\_features(sqrt,log2,none)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | n\_estimators | Criterion | Max\_features | R2 Value |
| 1 | 50 | Squared\_error | sqrt | 0.86852 |
| 2 | 100 | Squared\_error | sqrt | 0.87125 |
| 3 | 50 | Squared\_error | Log2 | 0.86585 |
| 4 | 100 | Squared\_error | Log2 | 0.87199 |
| 5 | 50 | Squared\_error | None | 0.85314 |
| 6 | 100 | Squared\_error | None | 0.85176 |
| 7 | 50 | Squared\_error | sqrt | 0.87074 |
| 8 | 100 | Friedman\_mse | sqrt | 0.87190 |
| 9 | 50 | Friedman\_mse | Log2 | 0.86625 |
| 10 | 100 | Friedman\_mse | Log2 | 0.87101 |
| 11 | 50 | Friedman\_mse | None | 0.85293 |
| 12 | 100 | Friedman\_mse | None | 0.85329 |
| 13 | 50 | Absolute\_error | sqrt | 0.87730 |
| 14 | 100 | Absolute\_error | sqrt | 0.87109 |
| 15 | 50 | Absolute\_error | Log2 | 0.86808 |
| 16 | 100 | Absolute\_error | Log2 | 0.86768 |
| 17 | 50 | Absolute\_error | None | 0.85167 |
| 18 | 100 | Absolute\_error | None | 0.85509 |
| 19 | 50 | Poisson | sqrt | 0.87142 |
| 20 | 100 | Poisson | sqrt | 0.86948 |
| 21 | 50 | Poisson | Log2 | 0.87108 |
| 22 | 100 | Poisson | Log2 | 0.87295 |
| 23 | 50 | Poisson | None | 0.84778 |
| 24 | 100 | Poisson | None | 0.85490 |

6.Mention your final model,justify why you have choosen the same.

On the whole Randomforest algorithm produced r2 score of 0.87730

Justification : From the tabulations Randomforest Algorithm performed well with different parameters and highest r2 value is noted .