**Assignment Regression Algorithm**

1. Identify your problem statement.

Problem statement is to predict the insurance charges based on the given parameter.

1. Tell basic info about the dataset(Total number of rows,columns)

The dataset is about insurance company details. There are totally 6 columns and 1338 rows. Header details of each columns are 'age', 'sex', 'bmi', 'children', 'smoker', 'charges'

1. Mention the pre-processing method if you’r doing any

Converting the string type data to numerial data for the 'sex' and 'smoker' columns for easy analysis

1. Develop 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.

Different Machine learning Regression methods to find r2 values

1. Multiple Linear Regression:

R2 Value: 0.789479

1. Support Vector Machine:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.NO | Hyper Parameter | R2 Value | | | |
| Linear | RBF | Poly | Sigmoid |
| 1 | C10 | 0.462468 | -0.03227 | 0.038712 | 0.039307 |
| 2 | C100 | 0.628879 | 0.320031 | 0.617956 | 0.527610 |
| 3 | C500 | 0.763105 | 0.664298 | 0.826368 | 0.444606 |
| 4 | C1000 | 0.764931 | 0.81020 | 0.856648 | 0.287470 |
| 5 | C2000 | 0.744041 | 0.854776 | 0.860557 | -0.593950 |
| 6 | C3000 | 0.741423 | 0.866339 | 0.85989 | -2.124419 |

In SVM regression R2 value is best for kernel rbf= (C=3000): 0.866339

1. Decision Tree:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***SL.NO*** | ***CRITERION*** | ***MAX\_FEATURES*** | ***SPLITTER*** | ***R2 VALUE*** |
| 1 | squared\_error | sqrt | best | 0.69863 |
| 2 | squared\_error | log2 | best | 0.67876 |
| 3 | squared\_error | auto | best | 0.711146 |
| 4 | squared\_error | sqrt | random | 0.686511 |
| 5 | squared\_error | log2 | random | 0.636321 |
| 6 | squared\_error | auto | random | 0.731595 |
| 7 | friedman\_mse | sqrt | best | 0.72192 |
| 8 | friedman\_mse | log2 | best | 0.689316 |
| 9 | friedman\_mse | auto | best | 0.701356 |
| 10 | friedman\_mse | sqrt | random | 0.612828 |
| 11 | friedman\_mse | log2 | random | 0.712738 |
| 12 | friedman\_mse | auto | random | 0.63708 |
| 13 | absolute\_error | sqrt | best | 0.613460 |
| 14 | absolute\_error | log2 | best | 0.706617 |
| 15 | absolute\_error | auto | best | 0.68993 |
| 16 | absolute\_error | sqrt | random | 0.74532 |
| 17 | absolute\_error | log2 | random | 0.620209 |
| 18 | absolute\_error | auto | random | 0.73288 |
| 19 | poisson | sqrt | best | 0.738786 |
| 20 | poisson | log2 | best | 0.740729 |
| 21 | poisson | auto | best | 0.728731 |
| 22 | poisson | sqrt | random | 0.637090 |
| 23 | poisson | log2 | random | 0.705015 |
| 24 | poisson | auto | random | 0.730891 |

In Decision tree regression R2 value is best for CRITERION= absolute\_error, MAX\_FEATURES=sqrt, SPLITTER= random: 0.74532

1. RandomForestRegressor

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***SL.NO*** | ***CRITERION*** | ***MAX\_FEATURES*** | ***n\_estimators*** | ***R2 VALUE*** |
| 1 | squared\_error | sqrt | 50 | 0.86808 |
| 2 | squared\_error | log2 | 50 | 0.86807 |
| 3 | squared\_error | auto | 50 | 0.85671 |
| 4 | squared\_error | sqrt | 100 | 0.872270 |
| 5 | squared\_error | log2 | 100 | 0.873169 |
| 6 | squared\_error | auto | 100 | 0.858568 |
| 7 | friedman\_mse | sqrt | 50 | 0.868586 |
| 8 | friedman\_mse | log2 | 50 | 0.869634 |
| 9 | friedman\_mse | auto | 50 | 0.858075 |
| 10 | friedman\_mse | sqrt | 100 | 0.866332 |
| 11 | friedman\_mse | log2 | 100 | 0.874293 |
| 12 | friedman\_mse | auto | 100 | 0.851027 |
| 13 | absolute\_error | sqrt | 50 | 0.87181 |
| 14 | absolute\_error | log2 | 50 | 0.877052 |
| 15 | absolute\_error | auto | 50 | 0.856974 |
| 16 | absolute\_error | sqrt | 100 | 0.874715 |
| 17 | absolute\_error | log2 | 100 | 0.874195 |
| 18 | absolute\_error | auto | 100 | 0.857207 |
| 19 | poisson | sqrt | 50 | 0.870596 |
| 20 | poisson | log2 | 50 | 0.87148 |
| 21 | poisson | auto | 50 | 0.856465 |
| 22 | poisson | sqrt | 100 | 0.871586 |
| 23 | poisson | log2 | 100 | 0.868924 |
| 24 | poisson | auto | 100 | 0.851993 |

In RandomForest Regressor, R2 value is best for CRITERION= absolute\_error, MAX\_FEATURES=log2, n\_estimators= 50: 0.877052

Best R2 value comes from Random forest Regressor and te final model created for the same.

1. All the research values (R2 scores of the models) should be documented.

It is documented in a tabluer format.

1. Mention your final model, justify why you have chosen the same.

Final model is RandomForest Regressor model because for this model only we get high R2 value.

Best R2 value comes from RandomForest Regressor, R2 value is best for CRITERION= absolute\_error, MAX\_FEATURES=log2, n\_estimators= 50: 0.877052