**1.Identifying the problem statement**

AI used to predict the medical insurance charges based on BMI, gender, number of children and whether the person is smoker or non-smoker.

Stage 1- Machine learning

Stage-2 supervised learning

Stage-3 Regression

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

Dataset contains 1338 rows and 6 columns

5 input columns and 1 output column

**3.) Mention the pre-processing method if you’re doing any (like converting string to number – nominal data)**

Categorical data- 2 columns – nominal data

**4.) 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.**

**5.) All the research values (r2\_score of the models) should be documented. (You can make tabulation or screenshot of the results.)**

**1.MULTIPLE LINEAR REGRESSION –** r2 value is 0.789479

# 2.SUPPORT VECTOR MACHINE-

# 

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **S.NO** | **HYPER PARAMETER** | | | **LINEAR (r value)** | | | **RBF (Non-Linear) (r value)** | **POLY**  **(r value)** | **SIGMOID**  **(r value)** |
| 1 | C=10 | | | 0.46246 | | | -0.03227 | 0.03871 | 0.03930 |
| 2 | C=100 | | | 0.62887 | | | 0.32003 | 0.61795 | 0.52761 |
| 3 | C=500 | | | 0.76310 | | | 0.66429 | 0.82636 | 0.44460 |
| 4 | C=1000 | | | 0.76493 | | | 0.81020 | 0.85664 | 0.28747 |
| 5 | C=2000 | | | 0.74404 | | | 0.85477 | 0.86055 | -0.59395 |
| 6 |  | C=3000 |  |  | 0.74142 |  | 0.86633 | 0.85989 | -2.12441 |

SVM Regression use R2 value (rbf and hyperparameter-C=3000) is 0.86633

# 3.DECISION TREE-

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **S.NO** | **CRITERION** | |  | **MAX**  **FEATURES** | | **SPLITTERS** | | **R VALUE** |
| 1 | Squared error | |  | Auto | | Best | | 0.69879 |
| 2 | Squared error | |  | Auto | | Random | | 0.71977 |
| 3 | Squared error | |  | Sqrt | | Best | | 0.71028 |
| 4 | Squared error | |  | Sqrt | | Random | | 0.66820 |
| 5 | Squared error | |  | Log2 | | Best | | 0.72608 |
| 6 | Squared error | |  | Log2 | | Random | | 0.66872 |
| 7 |  | Absoluteerror |  |  | Auto |  | Best | 0.67072 |
| 8 | Absolute error | |  | Auto | | Random | | 0.73970 |
| 9 | Absolute error | |  | Sqrt | | Best | | 0.65330 |
| 10 | Absolute error | |  | Sqrt | | Random | | 0.72923 |
| 11 | Absolute error | |  | Log2 | | Best | | 0.70927 |
| 12 | Absolute error | |  | Log2 | | Random | | 0.62551 |
| 13 | Friedman\_mse | |  | Auto | | Best | | 0.72082 |
| 14 | Friedman\_mse | |  | Auto | | Random | | 0.70707 |
| 15 | Friedman\_mse | |  | Sqrt | | Best | | 0.74964 |
| 16 | Friedman\_mse | |  | Sqrt | | Random | | 0.61508 |
| 17 | Friedman\_mse | |  | Log2 | | Best | | 0.72424 |
| 18 | Friedman\_mse | |  | Log2 | | Random | | 0.69975 |

**Hypertuning parameters –** criterion= Friedman\_mse, max\_features= Sqrt, splitter= Best has the highest r score – **0.74964**

**RANDOM FOREST:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.NO** | **criterion** | **n\_estimators** | **Random state** | **R value** |
| 1 | Mse | 100 | 0 | 0.85392 |
| 2 | Mse | 100 | 10 | 0.85104 |
| 3 | Mse | 50 | 0 | 0.84988 |
| 4 | Mse | 50 | 10 | 0.85108 |
| 5 | mae | 100 | 0 | 0.85214 |
| 6 | mae | 100 | 10 | 0.85628 |
| 7 | mae | 50 | 0 | 0.85290 |
| 8 | mae | 50 | 10 | 0.85560 |
| 9 | friedman\_mse | 100 | 0 | 0.85400 |
| 10 | friedman\_mse | 100 | 10 | 0.84923 |
| 11 | friedman\_mse | 50 | 0 | 0.84999 |
| 12 | friedman\_mse | 50 | 10 | 0.84982 |

**Hypertuning parameters –** criterion=mae , n\_estimators =100, Random state =10 has the highest r score **– 0.85628**

**6.) Mention your final model, justify why u have chosen the same.**

Final model is SVM regression with r2\_score 0.86633 is better than other algorithms.