**1.Identify your problem statement:**

Client requirement is he want to predict the insurance charge, **insurance charge** is the output, so have clear requirement it’s belongs to **supervised machine learning**

**Three Stages**

* Machine Learning- numbers
* Supervised Machine Learning – requirement is very clear
* Regression – output is continuous values

**2.Basic info about the data set:**

Total 6 columns and 1338 rows.

Input: age, sex, bmi, children, smoker

Output: charge

**3.** **Pre-processing method:**

Pre-processing method is Nominal data (one hot encoding) because input contains text, we are using get\_dummies method and the parameters we are passing dataset, dtype=int, drop\_first=true

**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.**

**Random Forest**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sno | N\_ESTIMATORS | CRITERION | MAXFEATURES | R2 VALUE |
| 1 | 50 | ***squared\_error*** | ***Sqrt*** | 0.8701028385180851 |
| 2 | 100 | ***squared\_error*** | ***Sqrt*** | 0.8718001587457894 |
| 3 | 50 | ***squared\_error*** | ***log2*** | 0.8681996057614894 |
| 4 | 100 | ***squared\_error*** | ***log2*** | 0.8695327331318028 |
| 5 | 50 | ***squared\_error*** | ***None*** | 0.852554550341361 |
| 6 | 100 | ***squared\_error*** | ***None*** | 0.8554279128535067 |
| 7 | 50 | ***absolute\_error*** | ***Sqrt*** | 0.8658715988246394 |
| 8 | 100 | ***absolute\_error*** | ***Sqrt*** | 0.8716449460920973 |
| 9 | 50 | ***absolute\_error*** | ***log2*** | 0.8697756502668389 |
| 10 | 100 | ***absolute\_error*** | ***log2*** | 0.8719872939568064 |
| 11 | 50 | ***absolute\_error*** | ***None*** | 0.849152753559274 |
| 12 | 100 | ***absolute\_error*** | ***None*** | 0.8545582387239524 |
| 13 | 50 | ***friedman\_mse*** | ***Sqrt*** | 0.8688820744852159 |
| 14 | 100 | ***friedman\_mse*** | ***Sqrt*** | 0.8713481734076576 |
| 15 | 50 | ***friedman\_mse*** | ***log2*** | 0.8622413870992105 |
| 16 | 100 | ***friedman\_mse*** | ***log2*** | 0.8700619856800212 |
| 17 | 50 | ***friedman\_mse*** | ***None*** | 0.8465958994853551 |
| 18 | 100 | ***friedman\_mse*** | ***None*** | 0.8592750560471338 |
| 19 | 50 | ***Poisson*** | ***Sqrt*** | 0.8705070945780351 |
| 20 | 100 | ***Poisson*** | ***Sqrt*** | 0.8711288410480799 |
| 21 | 50 | ***Poisson*** | ***log2*** | 0.8638015283716634 |
| 22 | 100 | ***Poisson*** | ***log2*** | 0.869178851414411 |
| 23 | 50 | ***Poisson*** | ***None*** | 0.8522570024085658 |
| 24 | 100 | ***Poisson*** | ***None*** | 0.848362395429203 |

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

**Multiple Linear Regression:**

R2 value: 0.7894790349867009

**Support Vector Machine:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Hyper  Parameter | Linear  (r value) | RBF (NON  LINEAR)  (r value) | POLY  (r value) | SIGMOID  (r value) |
| C10 | 0.462468414233968 | -0.03227329390671052 | 0.038716222760231456 | 0.03930714378274347 |
| C100 | 0.6288792857320346 | 0.3200317832050832 | 0.6179569624059795 | 0.5276103546510407 |
| C500 | 0.7631057975975393 | 0.6642984645143137 | 0.8263683541268934 | 0.44460610338694795 |
| C1000 | 0.7649311738649672 | 0.8102064851758545 | 0.8566487675946524 | 0.28747069486978516 |
| C1500 | 0.7440487080982905 | 0.8427494701978235 | 0.8580889211562686 | -0.06744112500716382 |
| C2000 | 0.7440418308107866 | 0.854776642539298 | 0.860557928865969 | -0.5939509731283508 |
| C3000 | 0.7414236599248631 | 0.8663393953081686 | 0.8598930084494358 | -2.1244194786689863 |

**Decision Tree**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | CRITERION | SPLITTER | MAX FEATURES | R VALUE |
| 1 | ***squared\_error*** | Best | Sqrt | 0.7304275213183836 |
| 2 | ***squared\_error*** | Random | Sqrt | 0.7045493394573541 |
| 3 | ***squared\_error*** | Best | Log2 | 0.6960452317124439 |
| 4 | ***squared\_error*** | Random | Log2 | 0.7287209056740023 |
| 5 | ***squared\_error*** | Best | None | 0.7300613818699182 |
| 6 | ***squared\_error*** | random | None | 0.7209408801405768 |
| 7 | ***friedman\_mse*** | Best | Sqrt | 0.7002606314979025 |
| 8 | ***friedman\_mse*** | Random | Sqrt | 0.6651239025650819 |
| 9 | ***friedman\_mse*** | Best | Log2 | 0.704450536546119 |
| 10 | ***friedman\_mse*** | Random | Log2 | 0.7216578813305338 |
| 11 | ***friedman\_mse*** | Best | None | 0.7306632458144676 |
| 12 | ***friedman\_mse*** | Random | None | 0.7226076851495364 |
| 13 | ***absolute\_error*** | Best | Sqrt | 0.6996688282460393 |
| 14 | ***absolute\_error*** | Random | Sqrt | 0.6726554796535373 |
| 15 | ***absolute\_error*** | Best | Log2 | 0.7205577192254702 |
| 16 | ***absolute\_error*** | Random | Log2 | 0.6920500725802841 |
| 17 | ***absolute\_error*** | Best | None | 0.7139233210656045 |
| 18 | ***absolute\_error*** | Random | None | 0.7374643598233536 |
| 19 | ***poisson*** | Best | Sqrt | 0.4958275379209969 |
| 20 | ***poisson*** | Random | Sqrt | 0.6737218629045889 |
| 21 | ***poisson*** | Best | Log2 | 0.6794969091733786 |
| 22 | ***poisson*** | Random | Log2 | 0.7144451697711103 |
| 23 | ***poisson*** | Best | None | 0.747166745095021 |
| 24 | ***poisson*** | Random | None | 0.7066731783283752 |

**Random Forest**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sno | N\_ESTIMATORS | CRITERION | MAXFEATURES | R2 VALUE |
| 1 | 50 | ***squared\_error*** | ***Sqrt*** | 0.8701028385180851 |
| 2 | 100 | ***squared\_error*** | ***Sqrt*** | 0.8718001587457894 |
| 3 | 50 | ***squared\_error*** | ***log2*** | 0.8681996057614894 |
| 4 | 100 | ***squared\_error*** | ***log2*** | 0.8695327331318028 |
| 5 | 50 | ***squared\_error*** | ***None*** | 0.852554550341361 |
| 6 | 100 | ***squared\_error*** | ***None*** | 0.8554279128535067 |
| 7 | 50 | ***absolute\_error*** | ***Sqrt*** | 0.8658715988246394 |
| 8 | 100 | ***absolute\_error*** | ***Sqrt*** | 0.8716449460920973 |
| 9 | 50 | ***absolute\_error*** | ***log2*** | 0.8697756502668389 |
| 10 | 100 | ***absolute\_error*** | ***log2*** | 0.8719872939568064 |
| 11 | 50 | ***absolute\_error*** | ***None*** | 0.849152753559274 |
| 12 | 100 | ***absolute\_error*** | ***None*** | 0.8545582387239524 |
| 13 | 50 | ***friedman\_mse*** | ***Sqrt*** | 0.8688820744852159 |
| 14 | 100 | ***friedman\_mse*** | ***Sqrt*** | 0.8713481734076576 |
| 15 | 50 | ***friedman\_mse*** | ***log2*** | 0.8622413870992105 |
| 16 | 100 | ***friedman\_mse*** | ***log2*** | 0.8700619856800212 |
| 17 | 50 | ***friedman\_mse*** | ***None*** | 0.8465958994853551 |
| 18 | 100 | ***friedman\_mse*** | ***None*** | 0.8592750560471338 |
| 19 | 50 | ***Poisson*** | ***Sqrt*** | 0.8705070945780351 |
| 20 | 100 | ***Poisson*** | ***Sqrt*** | 0.8711288410480799 |
| 21 | 50 | ***Poisson*** | ***log2*** | 0.8638015283716634 |
| 22 | 100 | ***Poisson*** | ***log2*** | 0.869178851414411 |
| 23 | 50 | ***Poisson*** | ***None*** | 0.8522570024085658 |
| 24 | 100 | ***Poisson*** | ***None*** | 0.848362395429203 |

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

Best model is Random Forest because compare to other algorithm r2\_value’s, Random Forest Algorithm r2\_value is highest (0.8718001587457894) and parameters are given below.

|  |  |  |  |
| --- | --- | --- | --- |
| 100 | ***squared\_error*** | ***Sqrt*** | 0.8718001587457894 |

**7. Kindly create Repository in the name Regression Assignment.**

https://github.com/SENTHILKUMARTH/Week-3-Machine-Learning-Regression/tree/main/Assignment