**Machine Learning Regression Assignment**

**Problem Statement or Requirement:**

A client’s requirement is, he wants to predict the insurance charges based on the several parameters. The Client has provided the dataset of the same

**Identifying problem statement**

Stage -1 : AI Domain selected here is – Machine Learning

Stage- 2 : Learning selection is – Supervised Learning

Stage -3 : Need to predict insurance charges which will be a number, Hence Supervised - Regression

**Basic information about the dataset**

Total number of rows – 1338 rows

Total number of inputs – 5

Input columns are – Age, Sex, BMI, Children, Smoker

**Final output that needs to be predicted here is – Charges**

**Dataset** **Pre**-**processing** **method**

Pre-processing method is – One hot encoding.

Inputs that needs to be converted from string to number are - 1. Sex and 2. Smoker (Nominal data)

**Research Values (R2\_score values of Machine Learning Regression Models).**

**Multiple Linear Regression R2 Value is : 0.7894790349867009**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Support Vector Machine Regression : High Accuracy R2 value is 🡪 0.8663393953081687** | | | | |
| Type of Kernels and its model accuracy values (r2\_scores - model accuracy) | | | | |
| Hyper Parameter Values | Linear | Non-Linear  (rbf - Radial Base Function) | Poly | Sigmoid |
| C=10 | 0.46246841423396834 | -0.03227329390671052 | 0.038716222760231456 | 0.03930714378274347 |
| C=100 | 0.6288792857320369 | 0.3200317832050831 | 0.6179569624059795 | 0.5276103546510407 |
| C=500 | 0.763105805389725 | 0.6642984645143138 | 0.8263683541269009 | 0.44460610338694795 |
| C=1000 | 0.7649311738597411 | 0.8102064851758545 | 0.8566487675946572 | 0.28747069486976173 |
| C=2000 | 0.7440418308107846 | 0.8547766425392979 | 0.8605579258597704 | -0.5939509731283505 |
| C=3000 | 0.7414236599249833 | **0.8663393953081687** | 0.8598930084494356 | -2.1244194786689854 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Decision Tree Regression : High Accuracy R2 value is 🡪 0.7517231237633607** | | | |
| criterion | splitter | max\_features | R\_VALUE |
| squared\_error | best | sqrt | 0.6902731620706302 |
| squared\_error | random | sqrt | 0.635040304315895 |
| squared\_error | best | log2 | 0.6674195671763193 |
| squared\_error | random | log2 | 0.7055881131041393 |
| squared\_error | best | None | 0.6908191115689286 |
| squared\_error | random | None | 0.7169280396242984 |
| friedman\_mse | best | sqrt | 0.7027422692367433 |
| friedman\_mse | random | sqrt | 0.6626837953070767 |
| friedman\_mse | best | log2 | 0.6671288533455892 |
| friedman\_mse | random | log2 | 0.6388281848522321 |
| friedman\_mse | best | None | 0.6945507658669092 |
| friedman\_mse | random | None | 0.7183077593914249 |
| absolute\_error | best | sqrt | 0.68980546932207 |
| absolute\_error | random | sqrt | 0.6646358645351194 |
| absolute\_error | best | log2 | **0.7517231237633607** |
| absolute\_error | random | log2 | 0.7147359189904974 |
| absolute\_error | best | None | 0.6728708895309433 |
| absolute\_error | random | None | 0.6997889557702834 |
| poisson | best | sqrt | 0.541903093996877 |
| poisson | random | sqrt | 0.6869413847509601 |
| poisson | best | log2 | 0.7166222879762963 |
| poisson | random | log2 | 0.6823969839446229 |
| poisson | best | None | 0.716669814490247 |
| poisson | random | None | 0.6961264017361194 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Random Forest Regression : High Accuracy R2 value is 🡪 0.8745913710987558** | | | |
| n\_estimators | criterion | max\_features | R\_VALUE |
| 50 | squared\_error | sqrt | 0.869822743124614 |
| 100 | squared\_error | sqrt | 0.8678922017626693 |
| 50 | squared\_error | log2 | 0.8702602091149049 |
| 100 | squared\_error | log2 | 0.867476502892116 |
| 50 | squared\_error | none | 0.8515482781542119 |
| 100 | squared\_error | none | 0.8536369676876538 |
| 50 | friedman\_mse | sqrt | 0.8623120861061284 |
| 100 | friedman\_mse | sqrt | 0.8711182672255372 |
| 50 | friedman\_mse | log2 | 0.8728112377949552 |
| 100 | friedman\_mse | log2 | 0.8719473984172234 |
| 50 | friedman\_mse | none | 0.8586075841853436 |
| 100 | friedman\_mse | none | 0.8515568403018346 |
| 50 | absolute\_error | sqrt | 0.8662677342178613 |
| 100 | absolute\_error | sqrt | 0.8728433292626412 |
| 50 | absolute\_error | log2 | 0.8699548607247947 |
| 100 | absolute\_error | log2 | 0.8696896714248589 |
| 50 | absolute\_error | none | 0.8545797351410502 |
| 100 | absolute\_error | none | 0.8542037761663623 |
| 50 | poisson | sqrt | 0.8667090151948807 |
| 100 | poisson | sqrt | 0.8697829281096329 |
| 50 | poisson | log2 | **0.8745913710987558** |
| 100 | poisson | log2 | 0.86967928946899 |
| 50 | poisson | none | 0.8536792681298714 |
| 100 | poisson | none | 0.8540299599849885 |

**Conclusion :**

The model we created using Random Forest Regression algorithm is the final machine Learning regression model that has been selected and saved as the best model in order to predict insurance charges amount for the customer inputs.

**Justification why I have chosen the same is :**

Among other research values of various models created with hyper parameters, the highest accuracy R2 value we obtained which is 🡪 **0.8745913710987558,** by using the algorithm “Random Forest Regression” with hyper tuning parameters applied below,

Criterion – poisson

n\_estimators – 50

max\_features – log2