**This is for Data Set🡪 insurance pre.csv**

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.

Form this requirement we clearly know that we need to predict the Insurance Charges.

1. Identify your problem statement:

While looking into the data set, we know that the Input and Output are well defined, Requirements are clear.

So it is Supervised Learning and it is given in Numerical values so we should use Regression.

1. Basic info about the dataset:

Input: Age, Sex, Bmi, Children, Smoker

Output: Charges

1. So for **Multiple Linear Regression** r2 value (r\_score)= 0.7894

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Support Vector Machine** | | | | | |
| S.NO | HYPER PARAMETER | LINEAR  (r value) | RBF  (NON LINEAR)  (r value) | POLY  (r value) | SIGMOID  (r value) |
| 1 | C10 | 0.4624 | -0.0322 | 0.0387 | 0.0393 |
| 2 | C100 | 0.6288 | 0.3200 | 0.6179 | 0.5276 |
| 3 | C500 | 0.7631 | 0.6642 | 0.8263 | 0.4446 |
| 4 | C1000 | 0.7649 | 0.8102 | 0.8566 | 0.2874 |
| 5 | C2000 | 0.7440 | 0.8547 | 0.8605 | -0.5939 |
| 6 | C3000 | 0.7414 | 0.8663 | 0.8598 | -2.1244 |

The **SVM Regression** use **R2**value (RBF) and hyper parameter(c=3000) = 0.8663

|  |  |  |  |
| --- | --- | --- | --- |
| **DECISION TREE** | | | |
| S.NO | **Criterion** | **Splitter** | r value |
| 1 | ***squared\_error*** | ***best*** | 0.6978 |
| 2 | ***squared\_error*** | ***random*** | 0.7316 |
| 3 | ***friedman\_mse*** | ***best*** | 0.6810 |
| 4 | ***friedman\_mse*** | ***random*** | 0.7125 |
| 5 | ***absolute\_error*** | ***best*** | 0.6824 |
| 6 | ***absolute\_error*** | ***random*** | 0.7203 |
| 7 | ***poisson*** | ***best*** | 0.7246 |
| 8 | ***poisson*** | ***random*** | 0.6847 |

The **Decision Tree Regression** use **R2**value **(squared\_error, random)** = 0.7316

|  |  |  |  |
| --- | --- | --- | --- |
| RANDOM FOREST | | | |
| **S.NO** | **HYPER PARAMETER** | **CRITIERION** | **R2** |
| 1 | n\_estimators=10 | Absolute error | 0.8474 |
| 2 | n\_estimators=100 | Absolute error | 0.8554 |
| 3 | n\_estimators=10 | Friedman mse | 0.8467 |
| 4 | n\_estimators=100 | Friedman mse | 0.8513 |
| 5 | n\_estimators=10 | Poisson | 0.8350 |
| 6 | n\_estimators=100 | Poisson | 0.8526 |
| 7 | max-features | Sqrt | 0.8696 |
| 8 | max-features | Log2 | 0.8704 |

The **Random Forest Regression** use **R2**value **(max-features, Log2)** = 0.8704

**So, for this insurance pre data set the best mehod of Regression:**

**Random forest Regression R2**value = 0.8704