1. Problem Statement:

**\*Since it is number format, so this is Machine Learning**

**\*Input and output are clear, so this is Supervised Learning**

**\*Since the output is numerical data so this falls under regression**

**2. Basic info about dataset**

**There are six columns have in the dataset**

**3. Pre-processing Method:**

**In this dataset the smoker was given as categorical data (yes or no). I converted it to numerical data using the get dummies function.**

**4. Find the good model using below machine learning algorithm**

1. **Multiple Linear Regression (R2 value )=0.7894**
2. **Support Vector Machine Regression (R2 Value)**:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.NO | HYPER PARAMETER | LINEAR  (R VALUE) | RBF(NON LINEAR)  (R VALUE) | POLY  (R VALUE)  NON LINEAR | SIGMOID  (R VALUE)  NON LINEAR |
| 1 | C10 | -0.0016 | -0.08196 | -0.09311 | -0.0907 |
| 2 | C100 | 0.5432 | -0.1248 | -0.09976 | -0.1181 |
| 3 | C500 | 0.6270 | -0.1246 | -0.08202 | -0.4562 |
| 4 | C1000 | 0.6340 | -0.1174 | -0.0555 | -1.6659 |
| 5 | C2000 | 0.6893 | -0.1077 | -0.0027 | -5.6164 |
| 6 | C3000 | 0.7590 | -0.0962 | 0.04892 | -12.01904 |

The Linear Regression use R2 value linear and hyper parameter (C3000) = **0.7590**

1. **Decision Tree Regressor(R2 Value):**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.NO | CRITERION | MAX FEATURES | SPLITTER | R2 VALUE |
| 1 | *Squared\_error* | 5 | Best | 0.6984 |
| 2 | *friedman\_mse* | 1000 | Best | 0.7012 |
| 3 | *absolute\_error* | 10 | Best | 0.6789 |
| 4 | *poisson* | 5 | Besst | 0.6872 |
| 5 | *Squared\_error* | 5 | Random | 0.6882 |
| 6 | *friedman\_mse* | 1000 | Random | 0.7095 |
| 7 | *absolute\_error* | 5 | Random | 0.6886 |
| 8 | *poisson* | 10 | Random | 0.7590 |
| 9 | *friedman\_mse* | 5 | Random | 0.7031 |
| 10 | *poisson* | 5 | Random | 0.7188 |

Decision Tree Regression using Poisson criterion. The Best R2 value = **0.7590**

1. **Random Forest Regression**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.NO | CRITERION | n\_estimators | random\_state | R2 VALUE |
| 1 | *Squared\_error* | 50 | 0 | 0.9446 |
| 2 | *friedman\_mse* | 10 | 100 | 0.9433 |
| 3 | *absolute\_error* | 5 | 50 | 0.8825 |
| 4 | *poisson* | 20 | 25 | 0.8947 |

Random Forest Regression using *Squared\_error, friedman\_mse,* *absolute\_error* criterion. The Best R2 value = 0.9446

1. **Adaboost Regression:**

An AdaBoost regressor is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction. As such, subsequent regressors focus more on difficult cases.

|  |  |  |  |
| --- | --- | --- | --- |
| S.NO | n\_estimators | random\_state | R2 VALUE |
| 1 | 100 | 0 | 0.8447 |
| 2 | 50 | 100 | 0.8662 |
| 3 | 20 | 20 | 0.8612 |
| 4 | 100 | 100 | 0.8490 |

Adaboost Regression using n-estimator and random\_state. The Best R2 value = 0.8662

**XG Boosting:**

Overall, XGBoost is a powerful and widely-used tool for regression tasks, and it has been applied successfully to a variety of real-world problems such as predictive modeling, time series forecasting, and customer churn prediction. Advantages: Effective with large data sets.

|  |  |  |
| --- | --- | --- |
| S.NO | Objective | R2 VALUE |
| 1 | Reg: squarederror | 0.8213 |

XG Boost Regression using Objective. The Best R2 value = 0.8213

**LG Boosting**

The light gradient boosting machine regressor (LightGBM) is a breakthrough tree-based ensemble learning approach developed by researchers at Microsoft and Peking University to overcome the efficiency and scalability difficulties of XGBoost in high-dimensional input feature and massive dataset contexts.

|  |  |  |
| --- | --- | --- |
| S.NO | n\_estimators | R2 VALUE |
| 1 | 100 | 0.8660 |

LG Boost Regression using n\_estimators. The Best R2 value = 0.8660