Problem Statement:

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. As a data scientist, you must develop a model which will predict the insurance charges.

1. **Identify your problem statement**

*Domain: Machine Learning*

*Learning: Supervised Learning*

*Type: Regression*

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

*No of rows: 1338*

*No of columns: 6*

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

*Here Dataset having ordinal vales in two columns namely Sex and Smoker. For converting it into numerical value , I have used OrdinalEncoder class from sklearn-preprocessing module*.

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.**

*I have chosen the Random forest model with some specified hypertuning paramters.*

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\_score=0.78947

* Support Vector Machine:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.No | Criterion | Linear | RBF | Sigmoid | Poly |
| 1 | C=1 | -0.01010 | -0.08338 | -0.07542 | -0.07569 |
| 2 | C=10 | 0.46246 | -0.03227 | 0.03930 | 0.03871 |
| 3 | C=100 | 0.62887 | 0.32003 | 0.52761 | 0.61795 |
| 4 | C=500 | 0.76310 | 0.66429 | 0.44460 | 0.82636 |
| 5 | C=1000 | 0.76493 | 0.81020 | 0.28747 | 0.85664 |
| 6 | C=1500 | 0.74404 | 0.84274 | -0.06744 | 0.85808 |
| 7 | C=2000 | 0.74404 | 0.85477 | -0.59395 | 0.86055 |
| 8 | C=3000 | 0.74142 | 0.86633 | -2.12441 | 0.85989 |

SVM uses r2\_score(linear and hypertuning C=3000)=0.86633

* Decision Tree:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | Criterion | Max Features | Splitter | R2\_score |
| 1 | Squared error | sqrt | best | 0.62738 |
| 2 | Squared error | sqrt | random | 0.72020 |
| 3 | Squared error | log | best | 0.65864 |
| 4 | Squared error | log | random | 0.68649 |
| 5 | Friedman mse | sqrt | best | 0.65519 |
| 6 | Friedman mse | sqrt | random | 0.72247 |
| 7 | Friedman mse | log | best | 0.70955 |
| 8 | Friedman mse | log | random | 0.66716 |
| 9 | Absolute error | sqrt | best | 0.76313 |
| 10 | Absolute error | sqrt | random | 0.70024 |
| 11 | Absolute error | log | best | 0.70488 |
| 12 | Absolute error | log | random | 0.69998 |
| 13 | Poisson | sqrt | best | 0.68256 |
| 14 | Poisson | sqrt | random | 0.65257 |
| 15 | Poisson | log | best | 0.67405 |
| 16 | Poisson | log | random | 0.64299 |

Decision tree uses r2\_score (absolute error - criterion , sqrt -max feature ,best splitter )= 0.76313

* Random Forest

|  |  |  |  |
| --- | --- | --- | --- |
| S.No | N\_estimators | Max\_features | R2\_score |
| 1 | 50 | sqrt | 0.86513 |
| 2 | 50 | log2 | 0.86609 |
| 3 | 100 | sqrt | 0.86692 |
| 4 | 100 | log2 | 0.86934 |
| 5 | 200 | sqrt | 0.87166 |
| 6 | 200 | log2 | 0.87439 |
| 7 | 250 | sqrt | 0.87028 |
| 8 | 250 | log2 | 0.87280 |

Random Forest uses r2\_score(n\_estimator=200, max\_feature=log2)= 0.87439

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

*I have chosen Random Forest model with n\_estimator=200 and max\_feature=log2. Since it has r2\_score of 0.87439 , which comparatively higher than other models.*