1. Identify your problem statement :

Here they wanted to predict the insurance charges based on many integer input.

Stage1: Machin Learning

Stage2: Supervised Learning

Stage3: Regression

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

There is #6 column in that one dependand variable i.e: **charges,** others #5 are independent variables.

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

Male & smoker column have only two option like Yes or No/male or female, so I used to covert this string value to integer.

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.

As far analysis says Random forest Alogrithom gives me good score which is 87%. When I am using follow hyper turning parameter:

**Criterion: log2**

n\_estimators: 1000

Max\_features: squared\_error

Multi-Linear regression:

R2\_Score: 0.7894

Simple Vector Machine:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sl.No | Hyper Parameter | Linear | Poly | Rbf | Singmoid |
| 1. | Default(C=1.0) | -0.11166 | 0.06429 | -0.08842 | -0.0899 |
| 2. | C=10 | -0.001617 | -0.09311 | -0.081969 | -0.09078 |
| 3. | C-0.1 | -0.12207 | -0.08625 | -0.0895762 | -0.08974 |
| 4. | C=100 | 0.54328 | -0.09976 | -0.12480367 | -0.118145 |
| 5. | C=1000 | 0.6340369 | -0.055505 | -0.11749 | -1.66590 |
| 6. | C=10000 | 0.744482 |  |  |  |
| 7. | C=1000000 | 0.786068 |  |  |  |

Radom Forest Regression:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sl.No | **criterion** | n\_estimators | **max\_features** | R2\_score |
| 1. | squared\_error | 10 | 1.0(default) | 0.83303 |
| 2. | squared\_error | 10 | Sqrt | 0.852000 |
| 3. | squared\_error | 100 | Sqrt | 0.871027 |
| 4. | squared\_error | 1000 | Sqrt | 0.87192 |
| 5. | squared\_error | 10 | Log2 | 0.852000 |
| 6. | squared\_error | 1000 | Log2 | 0.8719226 |
| 7. | squared\_error | 10 | None | 0.83303 |
| 8. | absolute\_error | 10 | 1.0(default) | 0.83506 |
| 9. | absolute\_error | 100 | 1.0(default) | 0.852009 |
| 10 | friedman\_mse | 10 | 1.0(default) | 0.833166 |
| 11 | friedman\_mse | 100 | 1.0(default) | 0.8540518935 |
| 12 | friedman\_mse | 100 | Sqrt | 0.871054401 |
| 13 | friedman\_mse | 10 | Sqrt | 0.85027 |
| 14 | friedman\_mse | 10 | Log2 | 0.85027 |
| 15 | poisson | 10 | Log2 | 0.854495 |
| 16 | poisson | 100 | Log2 | 0.86801 |
| 17. | poisson | 100 | Sqrt | 0.8680156 |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Decision Tree Regression:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl.No** | **Crieterion** | **Max\_features** | **Spliter** | **E2\_score** |
| 1 | squared\_error | None | Best | 0.69158 |
| 2 | squared\_error | None | Random | 0.692133 |
| 3 | squared\_error | Sqrt | Random | 0.64995 |
| 4 | squared\_error | Log2 | Random | 0.671321 |
| 5 | friedman\_mse | None | Best | 0.7124979 |
| 6 | friedman\_mse | None | Random | 0.6936323 |
| 7 | friedman\_mse | sqrt | Random | 0.609113 |
| 8 | friedman\_mse | Log2 | Random | 0.70905668 |
| 9 | absolute\_error | None | Best | 0.65877 |
| 10 | absolute\_error | sqrt | Best | 0.76348 |
| 11 | absolute\_error | Log2 | Best | 0.7398803064 |
| 12 | absolute\_error | None | Random | 0.72477 |
| 13 | absolute\_error | sqrt | Random | 0.759169 |
| 14 | absolute\_error | Log2 | Random | 0.74747 |
| 15 | poisson | None | Random | 0.741717 |
| 16 | poisson | sqrt | Random | 0.54537 |
| 17 | Poisson | Log2 | Random | 0.73113 |
| 18 | poisson | None | Best | 0.65556 |
| 19 | poisson | sqrt | Best | 0.74170 |
| 20 | Poisson | Log2 | Best | 0.73263 |
| 21 |  |  |  |  |

**Conclusion:**

**When I checked same dataset with all the model, random forest gives me the maximum results of 87%, this is best algorithm for this problem statement. (highlighted in yellow in above table)**