1. Problem statement – To predict the outcome of insurance charges based on input parameters of age, sex, bmi, children, smoker, charges
2. Data set has a total of 1338 rows × 6 columns
3. Pre-processing steps

Out of 6 parameters, age, sex, bmi, children, smoker, charges – the sex column and smoker column is categorical so I will be converting it into numerical value / nominal data

1. Trying out different models:

A - Random Forest

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | criterion | Max features | N\_estimators | R value |
| 1. | Squared\_error | None | 50 | 0.85 |
| 2. | Squared\_error | None | 100 | 0.85 |
| 3. | Squared\_error | Sqrt | 50 | 0.87 |
| 4. | Squared\_error | sqrt | 100 | 0.87 |
| 5. | Squared\_error | Log2 | 50 | 0.86 |
| 6. | Squared\_error | Log2 | 100 | 0.87 |
| 7 | Friedman\_mse | None | 50 | 0.85 |
| 8 | Friedman\_mse | None | 100 | 0.85 |
| 9 | Friedman\_mse | Sqrt | 50 | 0.87 |
| 10 | Friedman\_mse | sqrt | 100 | 0.87 |
| 11 | Friedman\_mse | Log2 | 50 | 0.86 |
| 12 | Friedman\_mse | Log2 | 100 | 0.87 |
| 13 | Absolute\_error | None | 50 | 0.85 |
| 14 | Absolute\_error | None | 100 | 0.85 |
| 15 | Absolute\_error | Sqrt | 50 | 0.86 |
| 16 | Absolute\_error | sqrt | 100 | 0.87 |
| 17 | Absolute\_error | Log2 | 50 | 0.87 |
| 18 | Absolute\_error | Log2 | 100 | 0.87 |
| 19 | poisson | None | 50 | 0.85 |
| 20 | poisson | None | 100 | 0.85 |
| 21 | poisson | Sqrt | 50 | 0.86 |
| 22 | poisson | sqrt | 100 | 0.87 |
| 23 | poisson | Log2 | 50 | 0.86 |
| 24 | poisson | Log2 | 100 | 0.86 |

B - Decision Tree

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | Criterion | Splitter | Max\_features | R value |
| 1 | friedman\_mse | best | None | 0.70 |
| 2 | friedman\_mse | random | None | 0.73 |
| 3 | squared\_error | best | None | 0.6843657056447416 |
| 4 | squared\_error | random | None | 0.7537133297740599 |
| 5 | absolute\_error | best | None | 0.6749612314344458 |
| 6 | absolute\_error | random | None | 0.7460729701361937 |
| 7 | poisson | best | None | 0.7284245497681144 |
| 8 | poisson | random | None | 0.749804147521383 |
| 9 | friedman\_mse | best | sqrt | 0.59 |
| 10 | friedman\_mse | random | sqrt | 0.75 |
| 11 | squared\_error | best | sqrt | 0.68 |
| 12 | squared\_error | random | sqrt | 0.61 |
| 13 | absolute\_error | best | sqrt | 0.79 |
| 14 | absolute\_error | random | sqrt | 0.73 |
| 15 | poisson | best | sqrt | 0.52 |
| 16 | poisson | random | sqrt | 0.55 |
| 17 | friedman\_mse | best | Log2 | 0.69 |
| 18 | friedman\_mse | random | Log2 | 0.58 |
| 19 | squared\_error | best | Log2 | 0.74 |
| 20 | squared\_error | random | Log2 | 0.72 |
| 21 | absolute\_error | best | Log2 | 0.68 |
| 22 | absolute\_error | random | Log2 | 0.76 |
| 23 | poisson | best | Log2 | 0.70 |
| 24 | poisson | random | Log2 | 0.62 |

C – Multiple linear regression R score is 0.7894790349867009

D – Support vector machine

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.No | Hyperparameter | Linear (r value) | RBF (non linear) | Poly (r value) | Sigmoid (r value) |
| 1. | C10 | -0.0016 | -0.08 | -0.09 | -0.09 |
| 2. | C100 | 0.54 | -0.12 | -0.099 | -0.11 |
| 3. | C1000 | 0.63 | -0.11 | -0.05 | -1.6 |
| 4. | C2000 | 0.68 | -0.10 | -0.02 | -5.6 |
| 5. | C3000 | 0.75 | -0.096 | 0.04 | -12.01 |

1. I have chosen Random forest (poisson – sqrt- 100) parameter combo as it gives the highest R score of 0.87