## Requirement:Predict insurance charges

## Goal:

Build a model to estimate insurance costs given customer details like **age, sex, bmi, No. of children, smoking status**

### Dataset:

Row: 1339, Column: 6

**Independent columns:** 5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| age | sex | bmi | children | smoker |

**Dependent column:** 1

|  |
| --- |
| Charges |

#### Stage 1: Domain Selection -> **Machine Learning**

Reason: Dataset is structured numeric data

#### Stage 2: Learning Selection -> **Supervised Learning**

Reason: Dataset includes both inputs and outputs, so this would be Supervised Learning

#### Stage 3: **Regression**

Reason: Output value is continuous numeric values so this would be Regression

Output is numerical

### Pre-Processing method

Features like sex and smoker are ordinal data -> I am going to use “one hot encoding” to keep it simple. Though it is used for nominal data, I see that we can still use it for ordinal as well.

### R\_Score

**Multiple Linear Regression:** R\_score: 0.79

**SVM:** -0.08

**Decision Tree**:

|  |  |  |  |
| --- | --- | --- | --- |
| Criterion | Splitter | Max\_features | R\_score |
| Squared\_error | Random | Log2 | 0.66 |
| ***friedman\_mse*** | Random | Log2 | 0.70 |
| ***absolute\_error*** | Random | Log2 | 0.68 |
| ***poisson*** | Random | Log2 | 0.74 |
| Squared\_error | best | Log2 | 0.66 |

**Random Forest**:

|  |  |  |  |
| --- | --- | --- | --- |
| Criterion | N\_estimators | Random\_state | R\_score |
| Squared\_error | 50 | 0 | 0.85 |
| Squared\_error | 100 | 0 | 0.85 |
| ***absolute\_error*** | 50 | 0 | 0.85 |
| ***friedman\_mse*** | 50 | 0 | 0.85 |

### Conclusion:

**Random forest** is the best model for this dataset