**REGRESSION ALGORITHM ASSIGNMENT**

1. **PROBLEM STATEMENT IDENTIFICATION :**

* The client has provided the required dataset , which consist of both input and output variables
* The requirement is also very clear as mentioned in the question to predict the “**Insurance changes**”

1. **DOMAIN SELECTION :**
   * The given dataset consist of numerical data so , it comes under “**Machine learning**”
2. **LEARNING SELECTION :**

* The input and output data are clearly given in the dataset in the form of numbers so it comes under **“Supervised Learning”**

1. **REGRESSION OR CLASSIFICATION :**
   * In output variable (Target) the data are in the form of Numerical so , it satisfies **“Regression”**
2. **ABOUT DATASET :**

* In “**Insurance\_pre.csv**” dataset , there are 1338(rows) and 6(columns) present .

**3 .PREPROCESSING TECHNIQUES :**

* The given dataset consist of categorical

(Nominal) variables , for converting it into Numerical I used “**one hot encoding**” method , In python , it can be achieved using “**pandas.get\_dummies()**” module

1. **ML – ALGORITHMS :**

* Here , I am going to use Multiple Linear Regression , Support Vector Machine , Decision Tree , Random Forest .

1. **RESEARCH VALUES (r2\_score) :**

**MULTIPLE LINEAR REGRESSION :**

* **The r2\_score of MLR is : 0.7978**

**SUPPORT VECTOR MACHINE :**

|  |  |  |
| --- | --- | --- |
| **Hyper parameter** | **C** | **r2\_score** |
| **linear** | **1** | **-0.012** |
| **10** | **0.5002** |
| **100** | **0.6423** |
| **1000** | **0.7501** |
| **rbf** | **1** | **-0.09** |
| **10** | **-0.04** |
| **100** | **0.35** |
| **1000** | **0.8283** |
| **poly** | **1** | **-0.89** |
| **10** | **0.04** |
| **100** | **0.65** |
| **1000** | **0.8631** |
| **sigmoid** | **1** | **-0.08** |
| **10** | **0.04** |
| **100** | **0.53** |
| **1000** | **0.1720** |

**DECISION TREE :**

|  |  |  |
| --- | --- | --- |
| **criterion** | **splitter** | **r2\_score** |
| **Squared\_error** | **best** | **0.7377** |
| **random** | **0.7169** |
| **Friedman\_mse** | **best** | **0.7267** |
| **random** | **0.7263** |
| **Absolute\_error** | **best** | **0.6762** |
| **random** | **0.7362** |
| **Poisson** | **best** | **0.7517** |
| **random** | **0.7794** |

**RANDOM FOREST :**

|  |  |  |  |
| --- | --- | --- | --- |
| **criterion** | **n\_estimators** | **max\_features** | **r2\_score** |
| **Squared\_error** | **50** | **Sqrt** | **0.8881** |
| **Log2** | **0.8946** |
| **100** | **Sqrt** | **0.8922** |
| **Log2** | **0.8920** |
| **200** | **Sqrt** | **0.8953** |
| **Log2** | **0.8935** |
| **Friedman\_mse** | **50** | **Sqrt** | **0.8902** |
| **Log2** | **0.8930** |
| **100** | **Sqrt** | **0.8911** |
| **Log2** | **0.8918** |
| **200** | **Sqrt** | **0.8937** |
| **Log2** | **0.8951** |
| **absolute\_error** | **50** | **Sqrt** | **0.8853** |
| **Log2** | **0.8897** |
| **100** | **Sqrt** | **0.8947** |
| **Log2** | **0.8917** |
| **200** | **Sqrt** | **0.8950** |
| **Log2** | **0.8956** |
| **poisson** | **50** | **Sqrt** | **0.8947** |
| **Log2** | **0.8927** |
| **100** | **Sqrt** | **0.8942** |
| **Log2** | **0.8944** |
| **200** | **Sqrt** | **0.8935** |
| **Log2** | **0.8930** |

**6.FINAL MODEL :**

* The final model for the insurance prediction dataset is from : “**RandomForestRegressor”** algorithm which is highly accurate then all the other algorithm
* This algorithm gives the r2\_score of 0.8956 with the hyper tuning parameter of **criterion=’absolute\_error’ , n\_estimators=200 ,**

**max\_features=’log2’.**

* *This model has been saved and deployed for further process*