**Assignment-Regression Algorithm**

1. **Identify your problem statement:**

First we have to collect data of information, In that we have to predict insurance price for the give data,

Problem statement is under - Machine learning

- Supervised

- Regression

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

* + - Independent Variable or input are given (Age, Sex, BMI, Smoke, Children)
    - Dependent variable or output are given (charges)
    - Total no of rows = 1338
    - Total no of columns = 6

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

Sex variable and Smoke variable are categorical – ordinal data converted into numbers.

1. **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.**
2. **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 value**: 0.78947
* **Support vector machine**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Hyper parameter** | **linear**  **( r2 value)** | **poly**  **( r2 value)** | **rbf**  **( r2 value)** | **sigmoid**  **( r2 value)** |
| C=10 | 0.42 | 0.03871 | -0.03227 | 0.03930 |
| C=100 | 0.62887 | 0.61795 | -0.32003 | 0.52761 |
| C=1000 | 0.76493 | 0.85664 | 0.81006 | 0.28747 |
| C=3000 | 0.74142 | 0.85989 | 0.86633 | -2.1247 |

The SVM regression use r2 value (rbf and hyper parameter C=3000) r2=0.86633

* **Decision tree:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.no** | **Criterion** | **splitter** | **R2 value** |
| 1. | Squared error | Best | 0.714913 |
| 2. | Squared error | random | 0.71742 |
| 3. | Friedman\_mse | Best | 0.71969 |
| 4. | Friedman\_mse | random | 0.68354 |
| 5. | Absolute error | Best | 0.67114 |
| 6. | Absolute error | random | 0.76107 |
| 7. | Poisson | Best | 0.7256 |
| 8. | Poisson | random | 0.64902 |

The decision tree use r2 value (absolute error, random) r2=0.76107

* **Random forest:**

|  |  |
| --- | --- |
| **n estimators:** | **R2 value** |
| 10 | 0.83315 |
| 50 | 0.84988 |
| 100 | 0.85392 |

The random forest use r2 value (100) r2=0.85392

1. **Mention your final model, justify why u have chosen the same:**

My final model for deployment is in SVM rbf hyper parameter (C=3000) 0.86633 because in this model r2 value is near to 1.