# Assignment-Regression Algorithm

**Date:** 19.09.2024

1. Problem Statement: Prediction of Insurance Charges based on several input parameters

2. About the Dataset:
   **No of Columns**: 6
   **Input** – Age, Sex, BMI, No of Children, Smoker
   **Output** -Insurance Charges
   **No of rows:** 1338

3. Preprocessing Method:
   Categorical data (Nominal) to Numerical data for the inputs
   **One hot Encoding** for column expansion
   1. **Sex** – Male/Female
   2. **Smoker** (Yes/No)

## 4. Sample Data : Insurance_Premium :

1. **Multiple Linear Regression** - $R^2$ Value = 0.7894
2. **Support Vector Machine**:

| S.No | Regularization Parameter 'C' | $R^2$ Value for different Kernel | | | |
|---|---|---|---|---|---|
| | | Linear | Poly | rbf | sigmoid |
| 1 | 1 | -0.010 | -0.075 | -0.083 | -0.075 |
| 2 | 10 | 0.462 | 0.038 | -0.0320 | 0.039 |
| 3 | 100 | 0.628 | 0.617 | 0.320 | 0.527 |
| 4 | 1000 | 0.764 | 0.856 | 0.810 | 0.287 |
| 5 | 3000 | 0.741 | 0.859 | 0.866 | -2.724 |

For the given dataset, c=3000; Kernel = "rbf" fits data better; $r^2 = 0.866$

3. **Decision Tree:**

| S.No | Criteria | Splitter | R_Value |
|---|---|---|---|
| 1 | ***squared_error*** | best | 0.692 |
| 2 | ***squared_error*** | random | 0.712 |
| 3 | ***friedman_mse*** | best | 0.691 |
| 4 | ***friedman_mse*** | random | 0.718 |
| 5 | ***absolute_error*** | best | 0.663 |
| 6 | ***absolute_error*** | random | 0.739 |
| 8 | ***poisson*** | best | 0.662 |
| 9 | ***poisson*** | random | 0.728 |

For the given dataset, criteria = ***absolute_error***; splitter = random fits data better; $r^2$=0.739

4. **Random Forest**

| S.No | Criteria | n_esimators | R_Value |
|---|---|---|---|
| 1 | ***squared_error*** | 10 | 0.843 |
| 2 | ***squared_error*** | 100 | 0.855 |
| 3 | ***friedman_mse*** | 10 | 0.837 |
| 4 | ***friedman_mse*** | 100 | 0.855 |
| 5 | ***absolute_error*** | 10 | 0.840 |
| 6 | ***absolute_error*** | 100 | 0.854 |
| 8 | ***poisson*** | 10 | 0.823 |
| 9 | ***poisson*** | 100 | 0.852 |

For the given dataset, criteria = ***Friedman_mse***; n_estimators = 100 data better; $r^2$=0.855

5. **SVR** seems to be the better model since it gives the highest $R^2$ value among all the algorithms – $R^2$ Value = 0.866