# Assignment-Regression Algorithm

## Problem Statement or Requirement:

A client's requirement is, he wants to predict the insurance charges based on the several parameters. The Client has provided the dataset of the same.

## Question & Answer:

1.) **Identify your problem statement**

- Client wants to predict the insurance charges based on the several parameters.

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

- Client has given the data about the age and sex of the insurer.
- Also provides the BMI value and their family details and the charges collected from the insurer.
- Total No. of Rows: 1338
- Total No. of Columns: 6

3.) **Mention the pre-processing method if you're doing any (like converting string to number – nominal data)**

- Yes, we are doing the pre-processing method for sex and smoker column.
- We are going to convert the both columns from string to numbers – Nominal data.

4.) **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.**

- R2_score value is developed for all the Regression Model.
- Final results for all the Models are as below.

5.) **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 = 0.789134**

- **Support Vector Machine (R2 value)** :

| S.No | Hyper Parameter | Linear | RBF(Non-Linear | Poly | Sigmoid |
|---|---|---|---|---|---|
| | C=0 | -0.01 | -0.083 | -0.075 | -0.075 |
| 1 | C=10 | 0.462 | -0.032 | 0.038 | 0.039 |
| 2 | C=100 | 0.628 | 0.319 | 0.616 | 0.526 |
| 3 | C=500 | 0.763 | 0.661 | 0.828 | 0.442 |
| 4 | C=1000 | 0.764 | 0.81 | 0.854 | 0.212 |
| 5 | C=2000 | 0.743 | 0.854 | 0.8583 | -0.621 |
| 6 | C=3000 | 0.741 | 0.864 | 0.858 | -2.143 |
| 7 | C=4000 | 0.741 | 0.87 | 0.8587 | -5.466 |

- The SVM Regression using R2 value (Linear, C=4000) = **0.87**

➢ **Decision Tree**:

| S.No | Criterion | Max Features | Splitter | R value |
|---|---|---|---|---|
| 1 | squared error | auto | best | 0.686 |
| 2 | squared error | auto | random | 0.756 |
| 3 | squared error | sqrt | best | 0.726 |
| 4 | squared error | sqrt | random | 0.718 |
| 5 | squared error | log2 | best | 0.745 |
| 6 | squared error | log2 | random | 0.68 |
| 7 | friedman_mse | auto | best | 0.694 |
| 8 | friedman_mse | auto | random | 0.683 |
| 9 | friedman_mse | sqrt | best | 0.671 |
| 10 | friedman_mse | sqrt | random | 0.668 |
| 11 | friedman_mse | log2 | best | 0.709 |
| 12 | friedman_mse | log2 | random | 0.68 |
| 13 | absolute_error | auto | best | 0.735 |
| 14 | absolute_error | auto | random | 0.688 |
| 15 | absolute_error | sqrt | best | 0.754 |
| 16 | absolute_error | sqrt | random | 0.712 |
| 17 | absolute_error | log2 | best | 0.764 |
| 18 | absolute_error | log2 | random | 0.757 |
| 19 | poisson | auto | best | 0.67 |
| 20 | poisson | auto | random | 0.75 |
| 21 | poisson | sqrt | best | 0.72 |
| 22 | poisson | sqrt | random | 0.682 |
| 23 | poisson | log2 | best | 0.753 |
| 24 | poisson | log2 | random | 0.642 |

- The Decision Tree Regression use R2 value (absolute_error, log2, best) = **0.764**

- **Random Forest:**

| S.No | Criterion | Max Features | R value | *n_estimators* |
|---|---|---|---|---|
| 1 | squared error | none | 0.855 | 100 |
| 2 | squared error | sqrt | 0.864 | |
| 3 | squared error | log2 | 0.863 | |
| 4 | friedman_mse | none | 0.851 | |
| 5 | friedman_mse | sqrt | 0.864 | |
| 6 | friedman_mse | log2 | 0.865 | |
| 7 | absolute_error | none | 0.857 | |
| 8 | absolute_error | sqrt | 0.869 | |
| 9 | absolute_error | log2 | 0.866 | |
| 10 | poisson | none | 0.851 | |
| 11 | poisson | sqrt | 0.86 | |
| 12 | poisson | log2 | 0.862 | |

The Random Forest Regression use R2 value (absolute_error, sqrt) = **0.869**

6.) **Mention your final model, justify why u have chosen the same.**

**Final Model**:

- R-Squared Value:

R-squared ($R^2$) is defined as a number that tells you how well the independent variable(s) in a statistical model explains the variation in the dependent variable.

It ranges from 0 to 1, where 1 indicates a perfect fit of the model to the data.

- From the definition, R-Squared value lies between 0 to 1 suits the best model

- Hence 1 indicates the perfect model.

- So, here I have chosen the **Support Vector Machine** as a Best model. Because the **R2 value** for the model is **0.87** which is comparatively higher than the other model.

- **SVM supports the best than the other models.**