# REGRESSION ALGORTHIM-ASSIGNMENT

Problem Statement:

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.

As a data scientist, you must develop a model which will predict the insurance charges.

1.) **Identify your problem statement:**

Stage 1: Machine Learning(Dataset in csv file)
Stage 2: Supervised Learning (Requirement is clear that clients want's to predict the insurance charges and while checking input and output, they clearly provide the information).
Stage 3: Regression (because output of this dataset is numerical form).

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

Total number of rows : 1338 rows.
Total number of columns : 6 columns .
It has 5 input as (age, sex, bmi, children, smoker) and 1 output(charges).

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

- Yes we have to do pre-processing for this dataset because we have two categorical column (sex and smoker)
- In sex column we have the (male and female ) in this we have to change as '0' and '1'.
- In smoker column we have the (yes or no) for this also we have to change as '0' and '1'.

As it is nominal data we have to split as per data.

Here is the example of dataset.

| S.no | Age | Sex_male | Sex_Female | BMI | Children | Smoker_Yes | Smoker_No | Charges |
|---|---|---|---|---|---|---|---|---|
| 1 | 19 | 0 | 1 | 27.9 | 0 | 1 | 0 | 16884.92 |
| 2 | 18 | 1 | 0 | 33.7 | 1 | 0 | 1 | 1725.552 |
| 3 | 28 | 1 | 0 | 33 | 3 | 0 | 1 | 4449.462 |
| 4 | 33 | 1 | 0 | 22.705 | 0 | 0 | 1 | 21984.47 |
| 5 | 32 | 1 | 0 | 28.88 | 0 | 0 | 1 | 3866.855 |

**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.**

| S.No | Model | $R^2$ Score (Best Model Score) |
|---|---|---|
| 1 | Multiple Linear Regression | 0.7894790 |
| 2 | Support Vector Machine(SVM) | 0.82636835 |
| 3 | Decision Tree | 0.717000508 |
| 4 | Random Forest | 0.874147851 |

For give dataset Random Forest algorithm is good model.

**5.) All the research values (r2_score of the models) should be documented. (You can make tabulation or screenshot of the results.)**

- **MULTIPLE LINERAR REGRESSION** :

**Model :** Multiple Linear Regression<br>
**$R^2$ score :** 0.7894790

- **SUPPORT VECTOR MACHINE(SVM):**

| S.No | REGULARIZATION PARAMETER | LINEAR $R^2$ Score | RBF(DEFAULT) $R^2$ Score | SIGMOID $R^2$ Score | POLY $R^2$ Score |
|---|---|---|---|---|---|
| **1** | C=1.0(Default) | -0.08095 | -0.089074 | -0.088269 | -0.088302 |
| **2** | C=10 | 0.4624684 | -0.032273 | 0.0393071 | 0.03871622 |
| **3** | C=100 | 0.6288792 | 0.3200317 | 0.52761035 | 0.61795696 |
| **4** | C=500 | 0.76310580 | 0.664298464 | 0.444606103 | 0.82636835 |

**Model:** Support Vector Machine.

**Kernal:** Poly

**Regularization Parameter:** C=500

**$R^2$ Score:** 0.82636835

## DECISION TREE:

| S.No | CRITERION | SPLITTER | $R^2$ SCORE |
|---|---|---|---|
| 1 | squared_error (default) | Best(default) | 0.68862487 |
| 2 | friedman_mse | Best(default) | 0.67937938 |
| 3 | Poisson | Best(default) | 0.717000508 |
| 4 | absolute_error | Best(default) | 0.6981776 |
| 5 | squared_error (default) | Random | 0.6941143 |
| 6 | Friedman_mse | Random | 0.70757116 |
| 7 | Poisson | Random | 0.71563569 |
| 8 | absolute_error | Random | 0.69888384 |

**Model:** Decision Tree**.**

**Criterion:** poisson

**Splitter:** Best(default)

**$R^2$ Score:** 0.717000508

- **RANDOM FOREST:**

| S.No | N_ESTIMATOR | CRITERION | MAX_FEATURE | RANDOM_SATATE | R$^2$ SCORE |
|---|---|---|---|---|---|
| 1 | **50** | *squared_error (default)* | 1.0(default) | None (default) | 0.85405105 |
| 2 | | *absolute_error* | 1.0(default) | None (default) | 0.855780392 |
| 3 | | *friedman_mse* | 1.0(default) | None (default) | 0.848051093 |
| 4 | | *Poisson* | 1.0(default) | None (default) | 0.854245404 |
| 5 | | *squared_error (default)* | sqrt | None (default) | 0.87390831 |
| 6 | | *absolute_error* | Sqrt | None (default) | 0.87005244 |
| 7 | | *friedman_mse* | Sqrt | None (default) | 0.86723481 |
| 8 | | *Poisson* | Sqrt | None (default) | 0.86935213 |
| 9 | | *squared_error (default)* | log2 | None (default) | 0.874147851 |
| 10 | | *absolute_error* | log2 | None (default) | 0.867238039 |
| 11 | | *friedman_mse* | log2 | None (default) | 0.869152362 |
| 12 | | *Poisson* | log2 | None (default) | 0.864473188 |
| 13 | | *squared_error (default)* | None | None (default) | 0.855725680 |
| 14 | | *absolute_error* | None | None (default) | 0.85571374 |
| 15 | | *friedman_mse* | None | None (default) | 0.84951382 |
| 16 | | *Poisson* | None | None (default) | 0.855362569 |
| 17 | | *squared_error (default)* | 1.0(default) | 0 | 0.849832931 |
| 18 | | *absolute_error* | 1.0(default) | 0 | 0.8526655 |

| | | | | | |
|---|---|---|---|---|---|
| 19 | **50** | *friedman_mse* | 1.0(default) | 0 | 0.85007161 |
| 20 | | *Poisson* | 1.0(default) | 0 | 0.84910759 |
| 21 | | *squared_error (default)* | Sqrt | 0 | 0.86958367 |
| 22 | | *absolute_error* | Sqrt | 0 | 0.870814425 |
| 23 | | *friedman_mse* | Sqrt | 0 | 0.87024175 |
| 24 | | *Poisson* | Sqrt | 0 | 0.86323913 |
| 25 | | *squared_error (default)* | log2 | 0 | 0.86958367 |
| 26 | | *absolute_error* | log2 | 0 | 0.870814425 |
| 27 | | *friedman_mse* | log2 | 0 | 0.870241751 |
| 28 | | *Poisson* | log2 | 0 | 0.86323913 |
| 29 | | *squared_error (default)* | None | 0 | 0.84983293 |
| 30 | | *absolute_error* | None | 0 | 0.85266559 |
| 31 | | *friedman_mse* | None | 0 | 0.850071613 |
| 32 | | *Poisson* | None | 0 | 0.8491075 |

**Model:** Random Forest**.**

**N_ESTIMATOR:** 50

**CRITERION:** *squared_error (default)*

**MAX_FEATURE:** log2

**RANDOM_SATATE**: None(default)

**$R^2$ Score:** 0.874147851

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

- As per our research *Random forest algorithm* is my final model.
- My justification is comparatively, this model has given good $R^2$ score.

Finally we found the good model for this dataset is ***Random forest*** and $R^2$ Score is 87%.