# Regression Algorithm-Assignment

## 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.

## Model to be developed:

- Develop a model that predict the insurance premium amount.

Details in dataset:

- Dataset has 1338 rows of records with 1 output and 5 inputs columns.
- Followings are the criteria based on which insurance premium amount has been determined.
- They are, age, smoking status, status of marital & children, body mass index.

Pre-processing requirement:

- Dataset has categorical value which can be converted into meaningful numerical value.
- Categorical type is **nominal** so we need to follow **one-hot-encoding method** to make it meaningful numerical value.

**3 Stages of Model Selection**:

Stage:1 – **Domain Selection**:

- Dataset has majorly numerical value. Model can be created using **Machine Learning domain**.

- Stage:2 – **Model Learning Selection**:
  - Dataset has clear requirements to be build.
  - It has both input and output without missing any records.
  - Hence, Model can be created with **Supervised Learning** method.

Stage:3 – **Regression or Categorical Regression**:

- Since Learning selection is Supervised, we need to identify next whether the output value of model is going to predict will be numerical or categorical value.
- With no doubt, the insurance premium amount is going to be numerical value.
- Hence, the model will be created using **regression type** of algorithms.

**Model Creation**:

- Since the dataset has more than 1 input column, **Simple Linear Algorithm can't be used**.
- Next, Try with all other regression type of algorithms such as **Multiple Linear Algorithm**, **Support Vector Machine (SVM)**, **Decision Tree** and **Random Forest** etc.,

**7 stages of model creation**:

Step-1: Data collection

Step-2: Data Clean-up

Step-3: Data Pre-processing

Step-4: Identify Input/Output values

Step-5: Train data / Test data Split

Step-6: Model Creation

Step-7: Model Evaluation

Step-8: Save the model

**Multiple Linear Algorithm**:

| Sl.No. | Fit_Intercept | r2_score |
|---|---|---|
| 1 | - | 0.789 |
| 2 | TRUE | 0.789 |

Multiple Linear Algorithm has given r2 value same with and/or without hyper tuning parameter.

R2_score is 0.789

**Support Vector Machine (SVM)**:

- R2_score was -0.067 without normalization.
- So, Standardization of data is required.

| Sl.No. | Kernel | gamma | C Value | r2_score |
|---|---|---|---|---|
| 1 | - | - | - | 0.9957 |
| 2 | linear | scale | 1 | 0.9981 |
| 3 | linear | scale | 10 | 0.9981 |
| 4 | linear | scale | 100 | 0.9981 |
| 5 | linear | auto | 1 | 0.9981 |
| 6 | linear | auto | 10 | 0.9981 |
| 7 | linear | auto | 100 | 0.9981 |
| 8 | rbf | scale | 1 | 0.9957 |
| 9 | rbf | scale | 10 | 0.9962 |
| 10 | rbf | scale | 100 | 0.9962 |
| 11 | rbf | auto | 1 | 0.9957 |
| 12 | rbf | auto | 10 | 0.9962 |
| 13 | rbf | auto | 100 | 0.9962 |
| 14 | poly | scale | 1 | 0.9944 |
| 15 | poly | scale | 10 | 0.995 |
| 16 | poly | scale | 100 | 0.9953 |
| 17 | poly | auto | 1 | 0.9944 |
| 18 | poly | auto | 10 | 0.995 |
| 19 | poly | auto | 100 | 0.9953 |
| 20 | sigmoid | scale | 1 | -111.34 |
| 21 | sigmoid | scale | 10 | -8178.42 |
| 22 | sigmoid | scale | 100 | -1138617.8 |
| 23 | sigmoid | auto | 1 | -111.34 |
| 24 | sigmoid | auto | 10 | -8178.42 |

| | | | | |
|---|---|---|---|---|
| 25 | sigmoid | auto | 100 | -1138617.8 |

**Support Vector Machine** with hyper tuning parameter of **kernel ='Linear'** model has given very good r2_score value of 99.81 % .

R2_score value is **0.9981**

**Decision Tree**:

| Sl.No. | criterion | splitter | max_features | r2_score |
|---|---|---|---|---|
| 1 | - | - | - | 0.701 |
| 2 | squared_error | best | sqrt | 0.5593 |
| 3 | squared_error | best | log2 | 0.7215 |
| 4 | squared_error | random | sqrt | 0.6555 |
| 5 | squared_error | random | log2 | 0.6529 |
| 6 | friedman_mse | best | sqrt | 0.701 |
| 7 | friedman_mse | best | log2 | 0.6323 |
| 8 | friedman_mse | random | sqrt | 0.6077 |
| 9 | friedman_mse | random | log2 | 0.6737 |
| 10 | absolute_error | best | sqrt | 0.6654 |
| 11 | absolute_error | best | log2 | 0.7372 |
| 12 | absolute_error | random | sqrt | 0.6847 |
| 13 | absolute_error | random | log2 | 0.6973 |
| 14 | poisson | best | sqrt | 0.6103 |
| 15 | poisson | best | log2 | 0.6217 |
| 16 | poisson | random | sqrt | 0.6247 |
| 17 | poisson | random | log2 | 0.6862 |

Decision Tree with hyper tuning parameter of **criterion = absolute_error, splitter = best and max_features = log2** has given highest r2_score value of 0.**7372**

**Random Forest**:

| Si.No. | n_estimators | criterion | max_features | r2_score |
|---|---|---|---|---|
| 1 | - | - | - | 0.8576 |
| 2 | 100 | squared_error | sqrt | 0.8694 |
| 3 | 100 | squared_error | log2 | 0.8705 |
| 4 | 100 | friedman_mse | sqrt | 0.8724 |
| 5 | 100 | friedman_mse | log2 | 0.8688 |
| 6 | 100 | absolute_error | sqrt | 0.8704 |
| 7 | 100 | absolute_error | log2 | 0.8662 |
| 8 | 100 | poisson | sqrt | 0.8302 |
| 9 | 100 | Poisson | log2 | 0.8281 |

**Random Forest algorithm** with hyper tuning parameter of **n_estimators =100, criterion = friendman_mse, max_features=sqrt** has given hightest r2_score value of **0.8724**

**Model Created to predict the insurance premium:**

**Among all of the regression algorithm tried, Support Vector Machine(SVM)-regression algorithm with hyper tuning parameter of Kernel=Linear, gamma=scale/auto has given extraordinary r2_score value, which is 0.9981, very close to 1.0**

**Since it has given very highest r2_score value, this algorithm will predict the very accurate premium than other models.**

**Hence, I am going to deploy this model to client for the their requirement.**

==========END==========