**Machine Learning Regression**

**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.**

1. Identify the problem Statement?

* As the input data is in a number format given in excel sheet I have chosen the domain as Machine Learning.
* As the requirement is very clear ( i.e. Insurance charges prediction) and there is both input and output data present in the excel the problem statement is Supervised Learning.
* Since prediction output is in numerical value ( i.e. Insurance charges) this falls under the Regression.

1. Basic info about dataset?

* **Age:** This is a numerical data representing the person's age in years.
* **Sex:** This is a categorical data that can be either "male" or "female".
* **BMI:** This is a numerical data representing the person's body mass index.
* **Children:** This is a numerical data representing the number of children the given person has.
* **Smoker:** This is a categorical data that can be either "yes" or "no", indicating whether the person smokes.
* **Charges:** This is the target variable, a numerical value representing the person's insurance charges.

1. Mention the preprocessing method (converting string to data)?

Sex and Smoker status are categorical value that comes under nominal data so we have to preprocess the data by one hot encoding. (i.e dataset=pd.get\_dummies(dataset,drop\_first=True)).

1. r 2 Score for various algorithms?

**Multiple Linear Regression**: (R2 Value) = 0.7894

**SVM:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.NO** | **Hyper parameter** | **Linear**  **(r\_Value)** | **RBF(Non Linear)**  **r\_value** | **Poly** | **Sigmoid** |
| 1 | C10 | 0.4624 | -0.0322 | 0.0387 | 0.0393 |
| 2 | C100 | 0.6288 | 0.3200 | 0.6179 | 0.5276 |
| 3 | C500 | 0.763 | 0.6642 | 0.8263 | 0.4446 |
| 4 | C1000 | 0.764 | 0.8102 | 0.8566 | 0.2874 |
| 5 | C2000 | 0.7440 | 0.8547 | 0.8605 | -0.5939 |
| 6 | C3000 | 0.7414 | 0.8663 | 0.8598 | -2.1244 |

The SVM regression use R2 Value= 0.8663 (Non linear, rbf and hyper parameter C3000)

**Decision Tree:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.NO** | **Criterion** | **Max Feature** | **Splitter** | **R value** |
| 1 | friedman\_mse | auto | random | 0.6663 |
| 2 | friedman\_mse | sqrt | random | 0.6888 |
| 3 | friedman\_mse | auto | best | 0.6902 |
| 4 | friedman\_mse | sqrt | best | 0.6504 |
| 5 | friedman\_mse | log2 | best | 0.6430 |
| 6 | friedman\_mse | log2 | random | 0.7165 |
| 7 | Mae | sqrt | random | 0.7466 |
| 8 | Mae | auto | random | 0.7114 |
| 9 | Mae | sqrt | best | 0.7749 |
| 10 | Mse | sqrt | best | 0.7525 |
| 11 | Mse | sqrt | random | 0.6721 |
| 12 | Mse | auto | random | 0.6763 |

The Decision tree regression use R2 Value= 0.7749 (Mae, sqrt and best)

**Random Forest:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.NO** | **Criterion** | **Max depth** | **N estimator** | **R value** |
| 1 | Mse | 10 | 100 | 0.8585 |
| 2 | Mse | 20 | 100 | 0.8540 |
| **S.NO** | **Criterion** | **Max feature** | **N estimator** | **R value** |
| 1 | Mse | Sqrt | 100 | 0.8733 |
| 2 | Mse | Sqrt | 10 | 0.8616 |
| 3 | Mse | auto | 10 | 0.8334 |
| 4 | Mse | auto | 100 | 0.8535 |
| 5 | Mse | log2 | 10 | 0.8475 |
| 6 | Mse | log2 | 100 | 0.8746 |
| 7 | Mae | Sqrt | 10 | 0.8490 |
| 8 | Mae | Sqrt | 100 | 0.8716 |
| 9 | Mae | auto | 10 | 0.8392 |
| 10 | Mae | auto | 100 | 0.8581 |
| 11 | Mae | log2 | 10 | 0.8528 |
| 12 | Mae | log2 | 100 | 0.8731 |

The Random Forest regression use R2 Value= 0.8746 (Mse, Log2,100)

**The Final Machine Learning Best Method of Regression**

1. The SVM regression use R2 Value= 0.8663 (Non linear, rbf and hyper parameter C3000)
2. The Random Forest regression use R2 Value= 0.8746 (Mse, Log2,100)