Assignment – Regression Algorithm

1.Problem Statement or Requirement:

Multiple input parameters are provided to predict insurance charges. As the input and output are numeric and the requirement deals with numeric values, selecting the **Machine Learning domain.**

Looking at the dataset, the following are observed:

- Requirement is clear with input(s) and Output
- Independent and Dependent variable are present.

Based on the above observation, **Supervised Learning** is adopted.

The output values are numeric hence proceeding with **Regression.**

2. Basic Info about dataset:

The dataset has multiple inputs such as age, gender, children, smoker, bmi and ouput charges (insurance charges)…total 1338 rows x 6 columns

3. The inputs such as gender and smoker columns holds the categorical (nominal data) hence these values are converted in to numeric using the get_dummies.

4. Using Machine Learning Algorithm, created the following:

- Multiple Linear Regression
- Support Vector Machine
- Decision Tree
- Random Forest

5. Accuracy of the Research Models:

- Multiple Linear Regression: r2_score = 1.0
- Support Vector Machine(without standardization) : r2_score = 0.9999999999605653
- Decision Tree (with default parameters): r2_score = 0.9982364920332555
- Random Forest: r2_score = 0.999679694289922

6. The final model selected was Multiple Linear regression with 100% Accuracy when compared to the other models.

SVM(after Standardization)

Hyper Parameter	Linear (r Value)	Rbf(Non Linear)(r Value)	Poly(r Value)	Sigmoid(r Value)
C=1	0.10087994572886794	-0.0808256460270218	-0.04129800453972243	-0.06366537537679862
C=10	0.8529720560922177	-0.008006791942225089	0.3355578386035505	0.14880349061986875
C=100	0.9999999999619384	0.4627044357461626	0.9086496487205619	0.79583334869026
C=500	0.9999999999802569	0.9072503636707163	0.942388513246426	0.6082403582150155
C=1000	0.9999999999814986	0.9635426982159025	0.9541513072972456	-0.08627664821740333
C=2000	0.9999999999825486	0.9858899155822013	0.971028075819288	-2.6673099964651152
C=3000	0.9999999999857418	0.9933146265127804	0.9822555307105733	-6.45827101741801

Decision Tree:

Criterion	Splitter	max_depth	max_leaf_nodes	min_samples_split	min_samples_leaf	R score
Friedman_mse	best	int	int	min	min	0.9989409919072885
Absolute_error	Best	Int	int	min	min	0.9996490574735933
poisson	best	int	int	min	min	0.999683523863742
Friedman_mse	random	int	int	min	min	0.9994120464222314
Absolute_error	random	int	int	min	min	0.9980370731835224
poisson	random	int	int	min	min	0.9985792862032016

Random Forest

N_estimators	Criterion	Max_features	R_score
50	Absolute error	Sqrt	0.9966520724500993
50	Friedman_mse	sqrt	0.9957778930918343
50	poisson	sqrt	0.996078065629805
50	Absolute error	Log2	0.9966520724500993
50	Friedman_mse	Log2	0.9957778930918343
50	poisson	Log2	0.996078065629805