Assignment on Regression

1. Problem Statement

Stage 1- Machine Learning

Dataset has numeric values, so its comes under Machine Learning algorithm

Stage 2-Supervised Learning

Since the output and input is clear, it is fall under the supervised learning

Stage 3- Regression

Output (Insurance charges) values will be in numeric, so it is Regression

2. Dataset has 1338 rows × 6 columns

3. Data pre- processing

Dataset has two categorical columns such as ‘sex’,’smoker’. Since it is Nominal data first we need to covert the categorical values into numeric values by using **one hot encoding.**

4. R Squared Value Analysis

Using the same dataset, created the models using multiple algorithmsin ML. Here the results are:

Linear Regression: R Squared Value: 0.78

Support Vector Machine:

Kernel	C	R_Squared Value
Linear	100	0.62
Linear	3000	0.74
Linear	30000	0.74
Linear	300000	0.74
Rbf	300	0.55
Rbf	3000	0.86
Rbf	30000	0.87
Rbf	300000	0.87
Poly	300000	0.85
Poly	30000	0.85
Poly	3000	0.85
sigmoid	3000	-2.12
sigmoid	30000	-255.44

In SVM , Kernel= rbf and c=30000 gives the best result of R Squared Value as 0.87

Decision Tree:

Criterien	Splitter	Max_features	R_Squared Value
squared_error	best	auto	0.68
squared_error	best	sqrt	0.71
squared_error	best	log2	0.65
squared_error	random	auto	0.68
squared_error	random	sqrt	0.59
squared_error	random	log2	0.66
friedman_mse	best	auto	0.7
friedman_mse	best	sqrt	0.71
friedman_mse	best	log2	0.75
friedman_mse	random	auto	0.67
friedman_mse	random	sqrt	0.67
friedman_mse	random	log2	0.67
absolute_error	best	auto	0.68
absolute_error	best	sqrt	0.76
absolute_error	best	log2	0.71
absolute_error	random	auto	0.76
absolute_error	random	sqrt	0.69
absolute_error	random	log2	0.7
poisson	best	auto	0.72
poisson	best	sqrt	0.71
poisson	best	log2	0.7
poisson	random	auto	0.64
poisson	random	sqrt	0.61
poisson	random	log2	0.57

In Decicion Tree, Criterion= friedman_mse ,splitter=random, max_features=auto gives the best R Squared Value is 0.75 .

Random Forest

n_estimators	criterion	max_features	R_Squared value
50	squared_error	auto	0.84
100	squared_error	auto	0.85
50	squared_error	sqrt	0.86
100	squared_error	sqrt	0.87
50	squared_error	log2	0.86
100	squared_error	log2	0.87
50	friedman_mse	auto	0.85
100	friedman_mse	auto	0.85
50	friedman_mse	sqrt	0.87
100	friedman_mse	sqrt	0.87
50	friedman_mse	log2	0.87
100	friedman_mse	log2	0.87
50	absolute_error	auto	0.85
100	absolute_error	auto	0.85
50	absolute_error	sqrt	0.87
100	absolute_error	sqrt	0.87
50	absolute_error	log2	0.87
100	absolute_error	log2	0.87
50	poisson	auto	0.84
100	poisson	auto	0.85
50	poisson	sqrt	0.86
100	poisson	sqrt	0.86
50	poisson	log2	0.86
100	poisson	log2	0.86

Criterion (friedman_mse and absolute_error) ,Max_features (sqrt and log2) ,n_estimators= 100 gives the best results. R squared value : 0.87

5.The Final Model

SVM and Random Forest algorithms gives best R squared values as 0.87.