PROJECT : INSURANCE CHARGES PREDICTION

3 stages of selection

1) Stage1 - Machine Learning Domain
2) Stage2 - Supervised Learning
3) Stage3 - Regression

Project details

1) Problem statement : Predicting insurance charges based on provided input.

2) Basic information about dataset

- No of rows in datasat - 6 (Age, Sex, BMI, Children, Smoker, Charges)
- No of columns in dataset - 1338
- Input / Independent data - Age, Sex, BMI, Children, Smoker (5No's)
- Output / Dependent data - Charges

3) Preprocessing method - Label encoder to convert ordinal categorical data (sex,smoker) into numerical data

4) Machine learning algorithms used

A. Multiple Linear Regression
B. Support Vector Machine
C. Decition Tree
D. Random Forest

A) Algorithm : Multiple Linear Regression

R^2 value : 0.78947

B) Algorithm : Support Vector Machine

R^2 value : 0.87799

S.No	Parameters		R^2 VALUE	Remarks
	Kernel	Penalty C		
1	Linear	1000	0.76493	
2		5000	0.74141	
3		10000	0.74142	
4		50000	0.74141	
5	Poly	1000	0.85664	
6		5000	0.85956	
7		10000	0.85917	
8		50000	0.85758	
9	rbf	1000	0.81020	
10		5000	0.87477	
11		10000	0.87799	Best Value
12		50000	0.87477	
13	sigmoid	100	0.52761	
14		1000	0.28747	
15		5000	-7.5300	
16	Precomputed			Not applicable for this dataset

C) Algorithm : Decision Tree

R^2 value : 0.74088

S.No	Parameters			R^2 VALUE	Remarks
	Criterion	Splitter	Max features		
1	squared_ error	best	None	0.70739	
2			sqrt	0.74088	Best Value
3			log2	0.73095	
4		random	None	0.71587	
5			sqrt	0.71638	
6			log2	0.5854	
7	friedman_mse	best	None	0.68711	
8			sqrt	0.71901	
9			log2	0.73386	
10		random	None	0.71288	
11			sqrt	0.67338	
12			log2	0.65476	
13	absolute_ error	best	None	0.69672	
14			sqrt	0.73739	
15			log2	0.72877	
16		random	None	0.72269	
17			sqrt	0.63633	
18			log2	0.66665	
19	poisson	best	None	0.66494	
20			sqrt	0.63372	
21			log2	0.66118	
22		random	None	0.62612	
23			sqrt	0.65671	
24			log2	0.68008	

D) Algorithm : Random forest

R^2 value   : 0.872158

S.No	Parameters			R^2 VALUE	Remarks
	Criterion	n_estimators	Max_ features		
1	squared_ error	50	None	0.84988	
2			sqrt	0.86949	
3			log2	0.86950	
4		100	None	0.85392	
5			sqrt	0.87099	
6			log2	0.87099	
7	friedman_mse	50	None	0.84999	
8			sqrt	0.87004	
9			log2	0.87004	
10		100	None	0.85400	
11			sqrt	0.87120	
12			log2	0.87094	
13	absolute_ error	50	None	0.8529	
14			sqrt	0.87215	
15			log2	**0.872158**	Best Value
16		100	None	0.85214	
17			sqrt	0.87172	
18			log2	0.87172	
19	poisson	50	None	0.83321	
20			sqrt	0.82878	
21			log2	0.82878	
22		100	None	0.83321	
23			sqrt	0.82932	
24			log2	0.82932	

5) Conclusion:

Comparison of results

S.No	Algorithm	Parameters	R_Value
1	Multiple Linear Regression		0.78947

2	Support Vector machine	Kernel=rbf,
		C=10000
			0.87799

3	Decision Tree	Criterion= squared_ error,
		Splitter=best,
		Max features = sqrt
			0.74088

4	Random Forest	Criterion=absolute_ error,
		n_estimators=50,
		Max features =log2
			0.87214

Finalized Result:

Finalized Algorithm : Support Vector Machine

Parameters : Kernel=rbf , C=10000

R_value :0.87799