

Insurance Charges Predictor (Hope AI)

Problem Statement :

To Predict the **Insurance Charges**, based on various factors including gender, age, smoker and no. Of children, using the given dataset.

Datset Analysis:

We are given a dataset (**Insurance_pre.csv**) with the following information Age, Sex, Bmi, Children, Smoker, Charges.

From this dataset, we can conclude,

Input : Age, Sex, BMI, Children, Smoker

Output : Charges.

Domain Selection:

- From this dataset, we can get to know that the input and output is clearly given and the output is continuous numerical value.
- Since our Input is Numeric value, we can select **Machine Learning** as our Domain, and the input and output is clearly given so, we can solve this using **Supervised Learning** and the output is continuous numeric value, we can use **Regression** algorithm.

Data Preprocessing:

From the dataset we can see there are some nominal columns were present (**Sex, Smoker**), we convert that nominal columns to number using pandas **get_dummies** method.

Regression Algorithms:

Since our dataset contains multiple inputs, we directly proceed with multiple linear regression and further algorithms.

1.) MULTIPLE LINEAR REGRESSION

Using multiple linear regression algorithm, we achieved the R2 score value as **R2 Score : 0.79**

2.) SUPPORT VECTOR MACHINE

Using support vector machine algorithm, we achieved the R2 score value as **R2 Score : 0.87**

SNO	HYPER PARAMETER	LINEAR (R VALUE)	RBF (R VALUE)	POLY (R VALUE)	SIGMOID (R VALUE)
-----	--------------------	---------------------	------------------	-------------------	----------------------

1	C10	0.46	-0.03	0.04	0.04
2	C100	0.63	0.32	0.62	0.53
3	C500	0.76	0.66	0.83	0.44
4	C1000	0.76	0.81	0.86	0.29
5	C2000	0.74	0.85	0.86	-0.59
6	C3000	0.74	0.87	0.86	-2.12

3.)DECISION TREE:

Using Decision Tree algorithm, we achieved the R2 score value as **R2 Score : 0.79**

SL_NO	CRITERION	MAX FEATURES	SPLITTER	R VALUE
1	SQUARED_ERROR	NONE	BEST	0.7
2	SQUARED_ERROR	NONE	RANDOM	0.71
3	SQUARED_ERROR	SQRT	BEST	0.77
4	SQUARED_ERROR	SQRT	RANDOM	0.6
5	SQUARED_ERROR	LOG2	BEST	0.64
6	SQUARED_ERROR	LOG2	RANDOM	0.69
7	ABSOLUTE_ERROR	NONE	BEST	0.68
8	ABSOLUTE_ERROR	NONE	RANDOM	0.79
9	ABSOLUTE_ERROR	SQRT	BEST	0.68
10	ABSOLUTE_ERROR	SQRT	RANDOM	0.67
11	ABSOLUTE_ERROR	LOG2	BEST	0.73
12	ABSOLUTE_ERROR	LOG2	RANDOM	0.77
13	FRIEDMAN_MSE	NONE	BEST	0.69
14	FRIEDMAN_MSE	NONE	RANDOM	0.76
15	FRIEDMAN_MSE	SQRT	BEST	0.76
16	FRIEDMAN_MSE	SQRT	RANDOM	0.69
17	FRIEDMAN_MSE	LOG2	BEST	0.68
18	FRIEDMAN_MSE	LOG2	RANDOM	0.65
19	POISSON	NONE	BEST	0.73
20	POISSON	NONE	RANDOM	0.69
21	POISSON	SQRT	BEST	0.74
22	POISSON	SQRT	RANDOM	0.72
23	POISSON	LOG2	BEST	0.76
24	POISSON	LGO2	RANDOM	0.73

4. RANDOM FOREST:

Using Random Forest algorithm, we achieved the R2 score value as **0.87**

SL_NO	CRITERION	MAX FEATURES	N_ESTIMATORS	R VALUE
1	SQUARED_ERROR	NONE	10	0.83
2	SQUARED_ERROR	NONE	100	0.85
3	SQUARED_ERROR	SQRT	10	0.85
4	SQUARED_ERROR	SQRT	100	0.87
5	SQUARED_ERROR	LOG2	10	0.85
6	SQUARED_ERROR	LOG2	100	0.87
7	ABSOLUTE_ERROR	NONE	10	0.84
8	ABSOLUTE_ERROR	NONE	100	0.85
9	ABSOLUTE_ERROR	SQRT	10	0.86
10	ABSOLUTE_ERROR	SQRT	100	0.87
11	ABSOLUTE_ERROR	LOG2	10	0.86
12	ABSOLUTE_ERROR	LOG2	100	0.87
13	FRIEDMAN_MSE	NONE	10	0.83
14	FRIEDMAN_MSE	NONE	100	0.85
15	FRIEDMAN_MSE	SQRT	10	0.85
16	FRIEDMAN_MSE	SQRT	100	0.87
17	FRIEDMAN_MSE	LOG2	10	0.85
18	FRIEDMAN_MSE	LOG2	100	0.87
19	POISSON	NONE	10	0.83
20	POISSON	NONE	100	0.85
21	POISSON	SQRT	10	0.85
22	POISSON	SQRT	100	0.87
23	POISSON	LOG2	10	0.85
24	POISSON	LOG2	100	0.87

Model Creation:

From comparing various regression algorithms with their r^2 score values, we conclude that

Support vector Machine (Regression) algorithm with Hyper Tuning Parameter (**Kernal = RBF**, **C = 3000**) gives the better model with efficiency of **0.87**

Created the Maching Learning Model using this algorithm and saved the model with the name **InsuranceChargePredictor.sav**

We can using the saved model in the deployment to predict the insurance charges by passing the required inputs.

-