Insurance Charges Prediction Using Machine Learning Regression Models

Dataset:

This dataset contains information about individuals and their medical insurance charges. It is used to predict insurance charges based on features like age, sex, BMI, children, smoking status, etc.,

Dataset dimensions:

Total Records (Rows): 1338

Total Features (Columns): 6

	age	sex	bmi	children	smoker	charges
0	19	female	27.900	0	yes	16884.92400
1	18	male	33.770	1	no	1725.55230
2	28	male	33.000	3	no	4449.46200
3	33	male	22.705	0	no	21984.47061
4	32	male	28.880	0	no	3866.85520
...	...	...	...	...	...	...
1333	50	male	30.970	3	no	10600.54830
1334	18	female	31.920	0	no	2205.98080
1335	18	female	36.850	0	no	1629.83350
1336	21	female	25.800	0	no	2007.94500
1337	61	female	29.070	0	yes	29141.36030

1338 rows × 6 columns

Pre-Processing methods:

To prepare the dataset for machine learning regression models, the following pre-processing steps were performed.

1. Handling Categorical (Nominal) Data:

Two columns in the dataset were categorical and needed to be converted to numeric values:

Column	Type	Encoding Applied	Details
sex	Nominal	One-Hot Encoding (drop_first = True)	Converted to sex_male (1 = male, 0 = female)
smoker	Nominal	One-Hot Encoding (drop_first = True)	Converted to smoker_yes (1 = smoker, 0 = non-smoker)

2. No Encoding Needed for Numeric Columns:

Columns such as age, bmi, children, charges are numerical. No conversion required.

Column	Type	Action
age	Numeric (int)	Used as-is
bmi	Numeric (float)	Used as-is
children	Numeric (int)	Used as-is
charges	Numeric (float, output)	Used as-is

	age	bmi	children	charges	sex_male	smoker_yes
0	19	27.900	0	16884.92400	0	1
1	18	33.770	1	1725.55230	1	0
2	28	33.000	3	4449.46200	1	0
3	33	22.705	0	21984.47061	1	0
4	32	28.880	0	3866.85520	1	0
...	...	...	...	...	...	...
1333	50	30.970	3	10600.54830	1	0
1334	18	31.920	0	2205.98080	0	0
1335	18	36.850	0	1629.83350	0	0
1336	21	25.800	0	2007.94500	0	0
1337	61	29.070	0	29141.36030	0	1

1338 rows × 6 columns

Model Development & Final Model Selection:

1. **Multiple Linear Regression: R^2 Score: 0.7894**

2. **Support Vector Machine:**

Sl. No.	HYPER PARAMETER	LINEAR (R^2 Score)	RBF (NON-LINEAR) (R^2 Score)	POLY (R^2 Score)	SIGMOID (R^2 Score)
1	Default (C1.0)	-0.1116	-0.0884	-0.0642	-0.0899
2	C10	-0.0016	-0.0819	-0.0931	-0.0907
3	C100	0.5432	-0.1248	-0.0997	-0.1181
4	C500	0.6270	-0.1246	-0.0820	-0.4562
5	C1000	0.6340	-0.1174	-0.0555	-1.6659
6	C2000	0.6893	-0.1077	-0.0027	-5.6164
7	C3000	0.7590	-0.0962	0.0489	-12.0190

Kernel: Linear
Hyperparameter (C): 3000
R^2 Score: 0.7590

3. **Decision Tree:**

Sl. No.	CRITERION	MAX FEATURES	SPLITTER	R^2 Score
1	squared_error	None	best	0.6976
2	squared_error	sqrt	best	0.6966
3	squared_error	log2	best	0.6692
4	squared_error	int (max_features = 2)	best	0.6697
5	squared_error	float (max_features = 0.5)	best	0.7277
6	squared_error	None	random	0.6827
7	squared_error	sqrt	random	0.7215
8	squared_error	log2	random	0.6556
9	squared_error	int (max_features = 2)	random	0.6685
10	squared_error	float (max_features = 0.5)	random	0.6975
11	friedman_mse	None	best	0.6796
12	friedman_mse	sqrt	best	0.6882
13	friedman_mse	log2	best	0.6970
14	friedman_mse	int (max_features = 2)	best	0.7463
15	friedman_mse	float (max_features = 0.5)	best	0.4546
16	friedman_mse	None	random	0.7439
17	friedman_mse	sqrt	random	0.6170
18	friedman_mse	log2	random	0.6696
19	friedman_mse	int (max_features = 2)	random	0.6766
20	friedman_mse	float (max_features = 0.5)	random	0.6207
21	absolute_error	None	best	0.6875
22	absolute_error	sqrt	best	0.7033
23	absolute_error	log2	best	0.7156
24	absolute_error	int (max_features = 2)	best	0.6642
25	absolute_error	float (max_features = 0.5)	best	0.7003

26	absolute_error	None	random	0.6917
27	absolute_error	sqrt	random	0.6347
28	absolute_error	log2	random	0.7486
29	absolute_error	int (max_features = 2)	random	0.7147
30	absolute_error	float (max_features = 0.5)	random	0.6715

Criterion: absolute_error
Splitter: random
max_features: log2
R^2 Score: 0.7486

4. Random Forest:

Sl. No.	CRITERION	MAX FEATURES	N_ESTIMATORS	R^2 Score
1	squared_error	None	100	0.8560
2	squared_error	sqrt	100	0.8719
3	squared_error	log2	100	0.8715
4	squared_error	int (max_features = 2)	100	0.8678
5	squared_error	float (max_features = 0.5)	100	0.8722
6	squared_error	None	10	0.8417
7	squared_error	sqrt	10	0.8637
8	squared_error	log2	10	0.8557
9	squared_error	int (max_features = 2)	10	0.8479
10	squared_error	float (max_features = 0.5)	10	0.8526
11	friedman_mse	None	100	0.8545
12	friedman_mse	sqrt	100	0.8710
13	friedman_mse	log2	100	0.8691
14	friedman_mse	int (max_features = 2)	100	0.8688
15	friedman_mse	float (max_features = 0.5)	100	0.8690
16	friedman_mse	None	10	0.8526
17	friedman_mse	sqrt	10	0.8557
18	friedman_mse	log2	10	0.8479
19	friedman_mse	int (max_features = 2)	10	0.8569
20	friedman_mse	float (max_features = 0.5)	10	0.8592
21	absolute_error	None	100	0.8565
22	absolute_error	sqrt	100	0.8710
23	absolute_error	log2	100	0.8693
24	absolute_error	int (max_features = 2)	100	0.8705
25	absolute_error	float (max_features = 0.5)	100	0.8741
26	absolute_error	None	10	0.8386
27	absolute_error	sqrt	10	0.8569
28	absolute_error	log2	10	0.8583
29	absolute_error	int (max_features = 2)	10	0.8547
30	absolute_error	float (max_features = 0.5)	10	0.8415

Criterion: absolute_error
n_estimators: 100

max_features: float (max_features)
R^2 Score: 0.8741

Final Model Selection:

Among all the regression models developed, including Multiple Linear Regression, Support Vector Machine (SVM), and Decision Tree Regressor – the Random Forest Regressor consistently outperformed the others in terms of R^2 Score, which measures the proportion of variance in the target variable(charges).

Reason for selecting Random Forest:

- Highest R^2 Score (0.8741) – this indicates the best predictive performance among all tested models.
- It aggregates multiple decision trees, reducing variance and overfitting, especially effective on datasets with nonlinear relationships.
- Performs well even with noisy or moderately imbalanced data.
- Hyper tunning parameters (n_estimators, criterion, max_features) allowed the model to generalize well on unseen data.

Final Model: Random Forest Regressor (R^2 Score: 0.8741)

The Random Forest Regressor with absolute error loss and optimized parameters chosen as the final model due to its high predictive accuracy and robust performance in handling the insurance dataset.