Prediction of Medical Insurance Charges using Machine Learning

1.Problem statement

The goal is to create the best model that can predict the medical insurance cost (charges) of a person based on features such as age, gender, BMI, number of children, smoking habits.

2.Dataset Characteristic:

Total Rows - 1,338
Total Columns - 6
Target Column - charges

3.Data Pre-processing

Convert categorical (string) data to numerical:
sex - 0 (female), 1 (male)
Smoker - 0 (no), 1 (yes)

4.Model Development and Evaluation

Experimenting with the following models.

Multiple Linear Regression:

R^2 Score=0.789

Support Vector Machine:

S.No	Hyper Parameter(C)	linear	rbf	poly	sigmoid
1	1.0	-0.11	-0.08	-0.06	-0.08
2	10	-0.001	-0.081	-0.093	-0.090
3	100	0.543	-0.124	-0.099	-0.118
4	500	0.627	-0.124	-0.082	-0.456
5	1000	0.634	-0.117	-0.055	-1.665
6	2000	0.689	-0.107	-0.002	-5.616
7	3000	0.759	-0.096	0.048	-12.010

R^2 Score=0.759 (Hyper parameter C=3000 and kernel=linear)

Decision Tree:

S.No	Criterion	Splitter	Max Features	R2 Value
1	squared_error (Default)	best	None	0.691
2	squared_error	best	sqrt	0.714
3	squared_error	random	sqrt	0.716
4	squared_error	best	log2	0.644
5	squared_error	random	log2	0.543
6	friedman_mse	best	sqrt	0.682
7	friedman_mse	random	sqrt	0.728
8	friedman_mse	best	log2	0.726
9	friedman_mse	random	log2	0.677
10	absolute_error	best	sqrt	0.714
11	absolute_error	random	sqrt	0.711
12	absolute_error	best	log2	0.778
13	absolute_error	random	log2	0.765
14	poisson	best	sqrt	0.731
15	poisson	random	sqrt	0.663
16	poisson	best	log2	0.753
17	poisson	random	log2	0.666

R^2 Score=0.778

Random Forest:

S.No	Criterion	Max Features	n_estimators	R^2 Score
1	squared_error	sqrt	100	0.870
2	squared_error	sqrt	200	0.872
3	squared_error	log2	100	0.867
4	squared_error	log2	200	0.870
5	absolute_error	sqrt	100	0.874
6	absolute_error	sqrt	200	0.871
7	absolute_error	log2	100	0.872
8	absolute_error	log2	200	0.875
9	friedman_mse	sqrt	100	0.872
10	friedman_mse	sqrt	200	0.870
11	friedman_mse	log2	100	0.868
12	friedman_mse	log2	200	0.868
13	poisson	sqrt	100	0.872
14	poisson	sqrt	200	0.870
15	poisson	log2	100	0.871
16	poisson	log2	200	0.869

R^2 Score=0.875

Final Model Selection:

Model	R^2 Score
Multiple Linear Regression	0.789
Support Vector Machine	0.759
Decision Tree	0.778
Random Forest	0.875

The Random Forest Regressor achieved the highest R^2 score of approximately 0.875, indicating strong predictive performance. It was therefore chosen as the final model.

Conclusion:

- This project successfully demonstrated the process of building a regression model to predict insurance charges based on demographic and lifestyle features.

- The Random Forest model performed best, explaining around 87% of the variance in the data. The most influential factors were age, BMI, and smoker status. The model was improved by hyperparameter tuning or using more advanced ensemble techniques.