1. Problem statement

The problem is to predict insurance charges based on customer details such as age, sex, bmi, number of children, and smoking status.

Machine Learning – Supervised Learning – Regression

2. Basic information about the dataset

The dataset contains 1339 rows and 6 columns. The independent variables are age, sex, bmi, children, and smoker, and the dependent variable is insurance charges. It includes both numerical variables (age, bmi, children) and categorical variables (sex, smoker).

3. Data Pre-processing

Since the dataset contains categorical features like sex and smoker**,** they should be converted into numerical values. The categorical features in the dataset are nominal. Hence, one-hot encoding is used for converting the categorical data to numerical values. This step is necessary because machine learning algorithms work only with numerical data.

4. Multiple regression models were developed and implemented to predict insurance charges, including Multiple Linear Regression, Support Vector Machine Regression, Decision Tree Regression, and Random Forest Regression.

5. Model performance comparison and analysis:

In this project, four regression models were implemented to predict insurance charges:

- Multiple Linear Regression (MLR)
- Support Vector Machine (SVM) Regression
- Decision Tree Regression
- Random Forest Regression

The performance of each model was evaluated using the R^2 score, and different parameter combinations were tested.

5.1 Multiple Linear Regression

The Multiple Linear Regression model achieved an R^2 score of 0.789.

5.2 Support Vector Machine

From the SVM results table, it is observed that model performance improved as the regularization parameter C increased. Best performance was obtained with C = 3000 and the RBF kernel achieved the highest R^2 score of 0.866

S.No	Regularization parameter, *C*	Linear *r²_score*	Poly *r²_score*	Rbf *r²_score*	Sigmoid *r²_score*	Gamma
1.	1.0	-0.010	-0.075	-0.083	-0.075	Auto
2.	1.0	-0.010	-0.075	-0.083	-0.075	Scale
3.	100	0.628	0.617	0.320	0.527	Auto
4.	100	0.628	0.617	0.320	0.527	Scale
5.	1000	0.764	0.856	0.810	0.287	Auto
6.	1000	0.764	0.856	0.810	0.287	Scale
7.	2000	0.744	0.860	0.854	-0.593	Auto
8.	2000	0.744	0.860	0.854	-0.593	Scale
9.	3000	0.741	0.859	0.866	-2.12	Auto
10.	3000	0.741	0.859	0.866	-2.12	Scale

5.3 Decision Tree

From the Decision Tree results, it is observed that performance varies significantly based on criterion, splitter, and max_features. The best R^2 score obtained was 0.747 using Squared error criterion, best splitter, and sqrt max_feature.

S.No	Criterion	Splitter	Max_features	r^2_score
1.	Squared error	best	none	0.699
2.	Squared error	random	none	0.694
3.	Squared error	best	sqrt	0.747
4.	Squared error	random	sqrt	0.740
5.	Squared error	best	log2	0.665
6.	Squared error	random	log2	0.701
7.	Absolute error	best	none	0.687
8.	Absolute error	random	none	0.744
9.	Absolute error	best	sqrt	0.736
10.	Absolute error	random	sqrt	0.679
11.	Absolute error	best	log2	0.707
12.	Absolute error	random	log2	0.646
13.	Friedman mse	best	none	0.691
14.	Friedman mse	random	none	0.658
15.	Friedman mse	best	sqrt	0.691
16.	Friedman mse	random	sqrt	0.663

17.	Friedman mse	best	log2	0.716
18.	Friedman mse	random	log2	0.706

5.4 Random Forest

Random Forest Regression produced the best overall performance among all models. Highest R^2 score obtained was 0.871 using Absolute error criterion, 100 estimators, sqrt or log2 max_features. Performance remained consistently high across different criteria.

S.No	**Criterion**	**N_estimators**	**Max_features**	***r^2_score***
1.	Squared error	100	none	0.853
2.	Squared error	100	sqrt	0.870
3.	Squared error	100	log2	0.870
4.	Absolute error	100	none	0.852
5.	Absolute error	100	sqrt	0.871
6.	Absolute error	100	log2	0.871
7.	Friedman mse	100	none	0.854
8.	Friedman mse	100	sqrt	0.870
9.	Friedman mse	100	log2	0.870

Final Model Selection

Based on the comparison of all models,

Model	Best r^2_score
Multiple Linear Regression	0.789
Support Vector Machine	0.866
Decision Tree	0.747
Random Forest	0.871

Random Forest is selected as the final model. Because it achieved the highest r^2 score, handled non-linear relationships effectively, provided stable and reliable predictions across parameter combinations.

Hence, Random Forest Regression is the most suitable model for predicting insurance charges for the given dataset.