ASSIGNMENT DOCUMENT

Name: A. BARKATHULLA

1 Identify your problem statement

Based on the accurate data provided, this falls under the category of machine learning. If the input and output data are clearly defined, we will use the supervised learning method. Since the output data consists of continuous numerical values, we will apply regression techniques.

Machine Learning -> Supervised learning -> Regression

2 Tell basic info about the dataset (Total number of rows, columns)

The given data set contains 1,338 rows and 6 columns. It includes five input columns and one output column.

3 Mention the pre-processing method if you're doing any (like converting string to number – nominal data)

The input columns include age, gender, BMI, number of children, and smoking status.

Since gender and smoking status are categorical values, they will be converted into numerical values using the one-hot encoding algorithm.

4 Develop a good model with r2_score. You can use any machine learning algorithm; you can create many models. Finally, you have to come up with final model.

I have used

1. Multiple Linear Regression
2. Support Vector Machine Regression
3. Decision Tree Regression
4. Random Forest Regression

5 All the research values (r2_score of the models) should be documented. (You can make tabulation or screenshot of the results.)

Assignment: To find the following the machine learning regression method using R^2 value

1. Multiple Linear Regression(R^2 value)= 0.7894790349867009

2. Support Vector Machine

Sl.No.	Hyper parmeter	linear	rbf	poly	sigmoid
1	C=1	-0.010102665	-0.08338239	-0.075699656	-0.075429243
2	C=10	0.462468414	-0.03227329	0.038716223	0.039307144
3	C=100	0.628879286	0.320031783	0.617956962	0.527610355
4	C=500	0.763105798	0.664298465	0.826368354	0.444606103
5	C=1000	0.764931174	0.810206485	0.856648768	0.287470695
6	C=2000	0.744041831	0.854776643	0.860557926	-0.593950973
7	C=3000	0.74142366	0.866339397	0.859893008	-2.124419479

The SVM Regression use R^2 value (rbf, C=3000) = 0.866339397

1. **Decision Tree**

Parameter Value		CRTERIAN			
Spliter	Max Features	**squared_error**	**friedman_mse**	**absolute_error**	**poisson**
best	None	0.69570782	0.711540346	0.668067214	0.72881771
random	None	0.721011855	0.714074918	0.724010131	0.70194947
best	sqrt	0.675469383	0.672665405	0.696076233	0.75913609
random	sqrt	0.696929769	0.696929769	0.69039057	0.626100992
best	log2	0.675469383	0.672665405	0.696076233	0.75913609
random	log2	0.696929769	0.696929769	0.69039057	0.69039057

R^2 value (Crterian= 'poisson', Spiliter='best', Max Features = 'sqrt') = 0.75913609

2. Random Forest Tree

Parameter Value		CRTERIAN			
n_estimators	Max Features	**squared_error**	**friedman_mse**	**absolute_error**	**poisson**
100	None	0.853552161	0.853751864	0.85266421	0.852775
50	None	0.849606351	0.849704226	0.853649521	0.849333
100	sqrt	0.870983465	0.87124993	0.871349834	0.868023
50	sqrt	0.869498178	0.870494655	0.871536806	0.863244
100	log2	0.870983465	0.87124993	0.871349834	0.868023
50	log2	0.869498178	0.870494655	0.871536806	0.863244

R^2 value (Crterian ='absolute_error', n_estimate = 50, Max Features = 'sqrt') =0.871536806

6 Mention your final model, justify why u have chosen the same.

Compare Best model

S.no	Model	R2_value
1	MLR	0.789479035
2	SVM	0.866339397
3	Decision Tree	0.75913609
4	Random Forest Tree	0.871536806

- The Random Forest Tree algorithms were chosen as the best model from the table above because they yield a high R^2 value.
- Best Model is
  Random forest Regressor use R^2 value (absolute_error, None , n_estimate=50) =**0.871536806**