

1.) Identify your problem statement

Ans:

- Machine Learning
- Supervised Learning
- Regression

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

Ans:

- No of rows: 1338
- No of columns: 6

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

Ans:

Splitting the data into multiple columns as one Hot Encoding for the columns sex and smoker.

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.

Ans : Models are attached in the github

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

1. Multiple Linear Algorithm:

S.No	Model Name	R2_Score
1.	Mul_linear_model	0.78

2. SVM

S.No	Model Name	Kernel	R2_Score
1.	SVM_model	rbf	0.81
2.	SVM_model_linear	Linear	0.76
3.	SVM_model_poly	poly	0.85
4.	SVM_model_sigmoid	sigmoid	0.28

3. Decision Tree:

S.No	Model Name	creterion	splitter	R2_Score
1.	d_tree_1	squared_error	best	0.70
2.	d_tree_2	friedman_mse	best	0.69
3.	d_tree_3	absolute_error	best	0.66
4.	d_tree_4	Poisson	best	0.71
5.	d_tree_5	squared_error	random	0.73
6.	d_tree_6	friedman_mse	random	0.67
7.	d_tree_7	absolute_error	random	0.73
8.	d_tree_8	Poisson	random	0.71

4. Random Forest:

S.No	Model Name	creterion	N_estimators	Max_leaf_nodes	Max_samples	R2_Score
1.	r_forest_1	squared_error	100	10	100	0.88
2.	r_forest_2	friedman_mse	100	10	100	0.88
3.	r_forest_3	absolute_error	100	10	100	0.88
4.	r_forest_4	Poisson	100	10	100	0.87

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

- Final Model are **r_forest_1, r_forest_2, r_forest_3**
- These 3 models provided r2_score as 0.88 compare to other models
- Compare to other models Random forest algorithm worked well for this dataset.