Name: Prachi Tiwari Class: 2324-CT5165 Student ID: 23102057

1. First, report what type of algorithm you need to use and why?

**ANSWER**: Given the dataset, which consists of both categorical and numerical values, and the target label "charges" consists of continuous numerical values. The goal is to predict the medical insurance charges. So, given the nature of everything, regression algorithm is best suited as they are designed to predict continuous numerical values. Therefore, I have used GradientBoostingRegressor algorithm.

2. Use all the techniques of pre-processing, cleaning, and normalisation to prepare your data. Report each with justification why you want to use.

**ANSWER**: Following were used to prepare the data:

- The below code snippet is used to determine whether any null values exist in any feature or label column within the train dataset, which is important for ensuring the quality of dataset:

```
training_set.isnull().sum()
```

- Used Label Encoding on categorical data such as “sex”, “smoker” and “region” column to convert into numerical data, because algorithm needs and performs better on numerical data in this case.

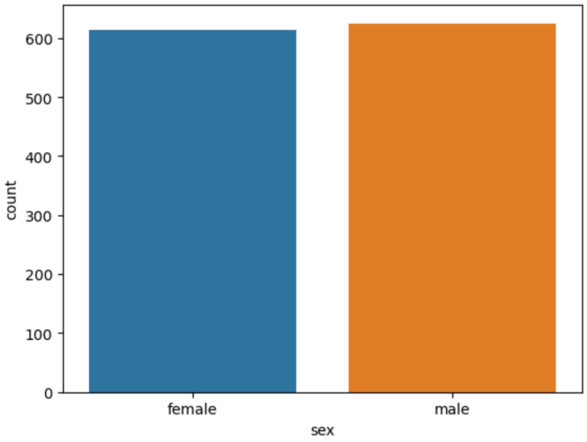
```
label_encoder = LabelEncoder()
training_set['sex_encoded'] = label_encoder.fit_transform(training_set['sex'])
training_set['smoker_encoded'] = label_encoder.fit_transform(training_set['smoker'])
training_set['region_encoded'] = label_encoder.fit_transform(training_set['region'])

label_encoder = LabelEncoder()
testing_set['sex_encoded'] = label_encoder.fit_transform(testing_set['sex'])
testing_set['smoker_encoded'] = label_encoder.fit_transform(testing_set['smoker'])
testing_set['region_encoded'] = label_encoder.fit_transform(testing_set[‘region'])
```

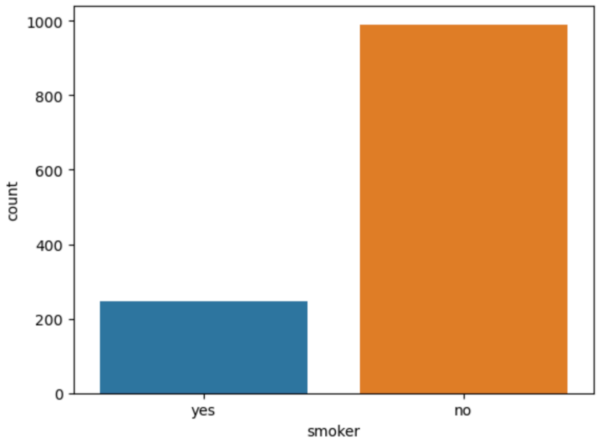
- Normalisation is not performed as the model was performing better without scaling the data.

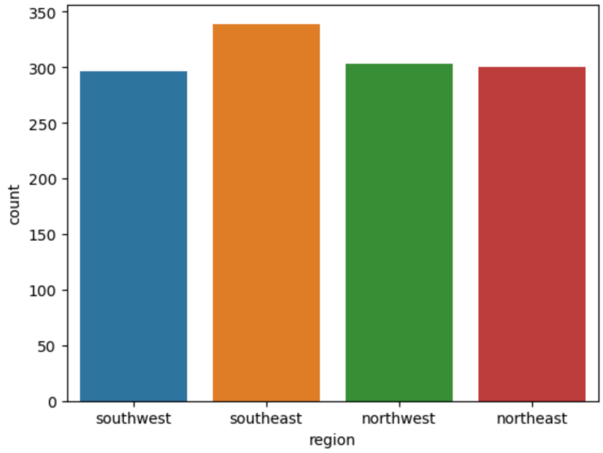
3. Analyse the data by visualising each feature or overall whole dataset with various graphs, bars, etc. to understand the data before applying a model. Report what you learned from visualising the data. Did you find any correlation or discrepancies in the data.

**ANSWER**: Analysed the data by visualising each feature as below:

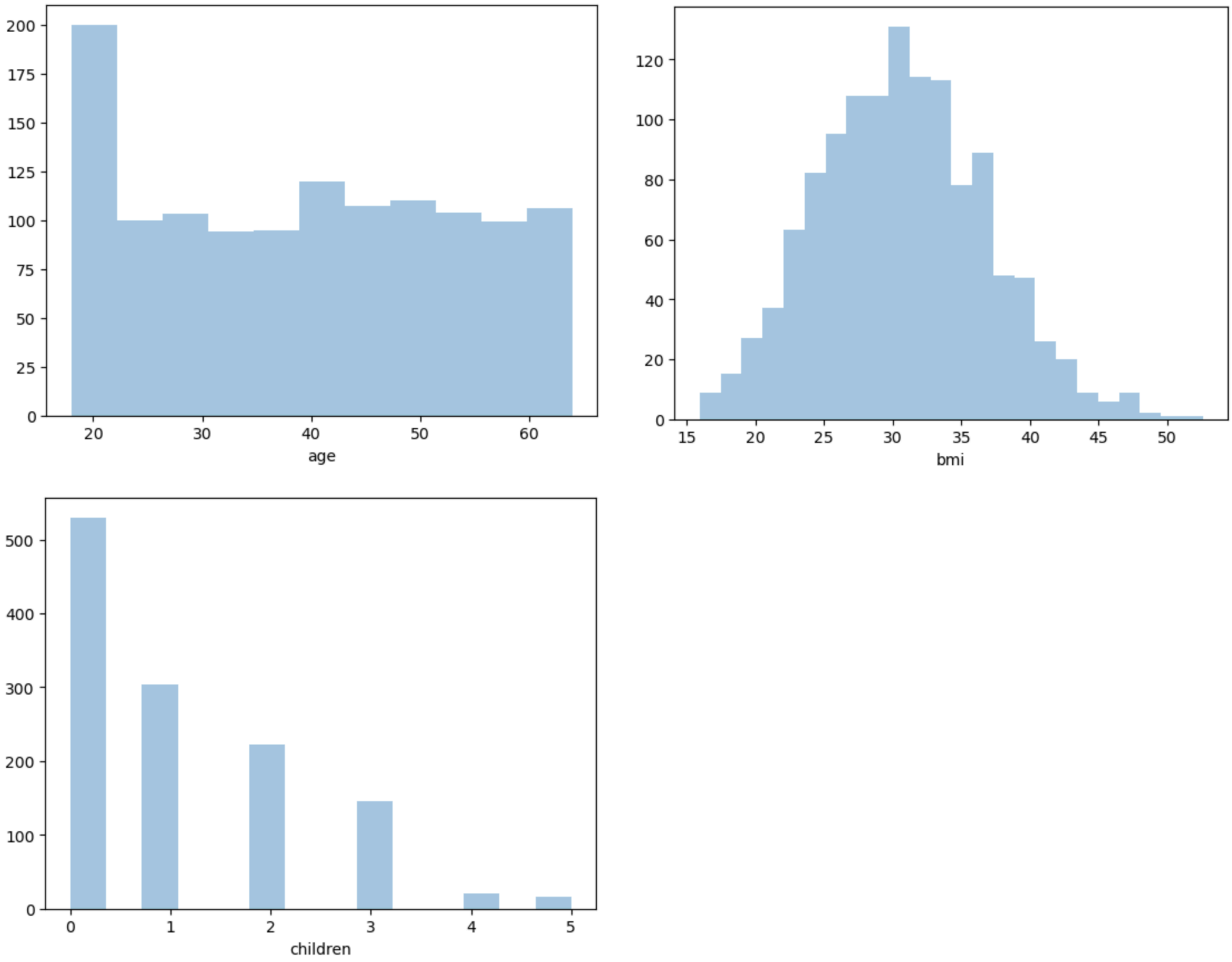
1. <u>Visualised Categorical Data</u>: It can be observed that feature “sex” and “region” are nearly equally distributed, whereas, feature “smoker” has majority of non-smokers.





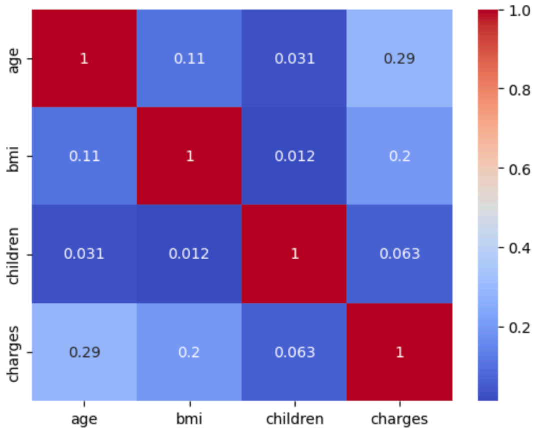


2. <u>Visualised Numerical Data</u>: It can be observed that feature “age” has mostly uniformly distributed data and “bmi” exhibit more normal distribution with majority of data points concentrated in central region, indication most individual have bmi in middle range. While the feature “children” has majority data with no children and low data points with more children.



3. <u>Visualising correlation of variables</u>: The correlation matrix's diagonal elements are marked as '1.0,' indicating perfect correlation between each feature and itself. However, the off-diagonal

values in the matrix imply that there is minimal linear correlation observed between the various features.

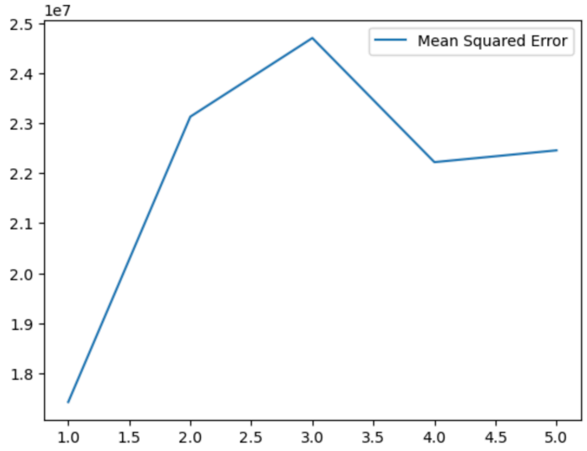


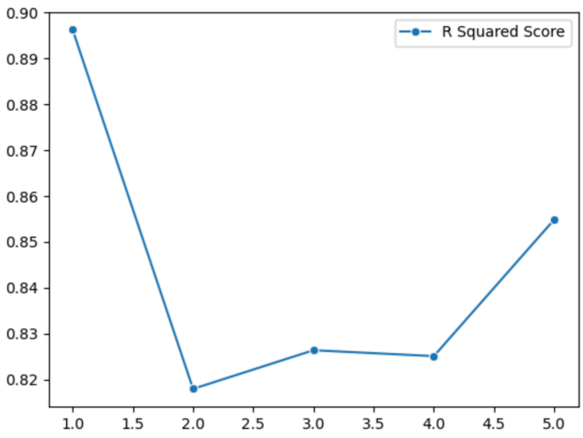
<u>Conclusion</u>: The lack of connection among the numerical features in the training dataset suggests that there is no duplicate information between these features. Each numerical characteristic looks to give different and unique information that can contribute to the prediction model independently.

4. Use the k-fold cross validation approach to find the optimal model that you can apply on the testing data provided. Report the performance e.g loss and loss curve,  validation accuracy and its curve (if applicable) and show when and why you stopped e.g., did you reached convergence or not.

**ANSWER**: On the pre-processed dataset, Kfold cross validation has been applied and number of splits value taken is 5.

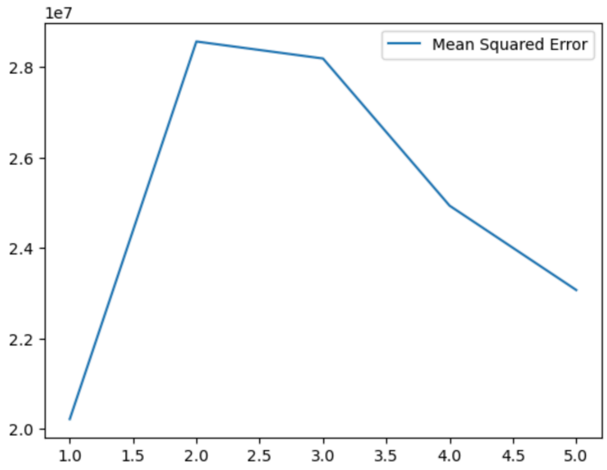
The model used to train the data is GradientBoostingRegressor.
Mean Squared Error (MSE) and R-squared is used to measure the loss.

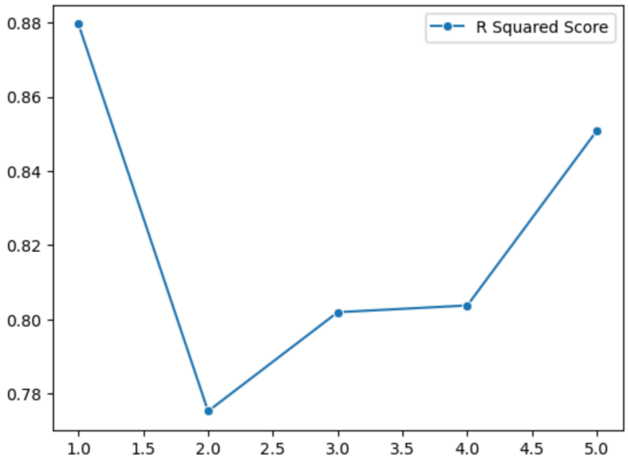




```
Average MSE:  21928446.715915486
Average R-squared:  0.8444870017973811
```

The model used to train the data is RandomForestRegressor.
Mean Squared Error (MSE) and R-squared is used to measure the loss.





```
Average MSE:  24993271.26098957
Average R-squared:  0.8222742996974848
```

Looking at the performance metrics of both the model, GradientBoostingRegressor performs better than RandomForestRegressor, as it has lower MSE and higher R-square. Which means it has better predicted on validation sets.

The purpose for stopping at k-fold value 5 is to avoid model overfitting/underfitting, which would result in bigger swings in validation scores.

5. Report the final total price that you got.

**ANSWER**: The final total price using GradientBoostingRegressor on testing set is:

```
Final total price for new 100 employees is :
1335334.093012557
Final total price for all the employees is :
17745616.793012556
```