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

By analysing the dataset it can be seen that the attributes "age",sex,bmi,children, smoker and region are independent variables whereas the charges is the dependent variable. Here, the dependent variable is continuous in nature so it can be interpreted as a regression problem. Moreover, the comparison between random forest and linear regression clearly shows that the predicted line is much closer to the actual values for random forest**[Fig 1]**. Based on this I have decided to used **Random Forest regressor.** Moreover, the idea behind choosing random forest regressor is that it is less likely to overfit and can capture complex relationships among the variables.In addition to this it can also be seen that the data is not not linear in nature so random forest is more preferred over linear regression.

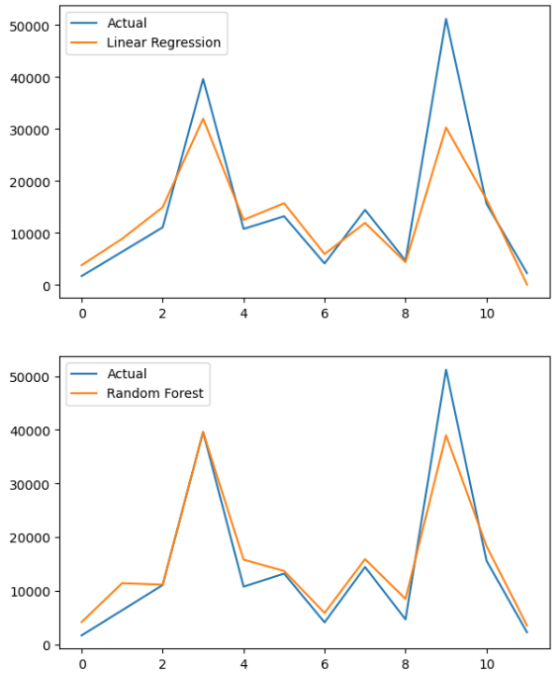

Figure 1: Comparison Between LR and Random Forest

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

The dataset contain values which are in string format as well as numerical format. The data should be first normalized so that the algorithm can interpret the dataset.Below are some steps which I have performed in pre-processing.

- **Scaling:** It is always a good practive to scale the data before providing it to the model. In this the attributes Age and BMI has been scaled using minmax Scalar.
- **Encoding:** The dataset contained few attributes which have categorical values, to deal with this a function called LabelEncoder is used to uniquely map the values. For example, Male=0 and Female=1.**[1]**

0.3 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.

[Fig 1]:- Based on the visualization the dataset contains approximately 64 values of people who are aged 19 and similarly for other age groups.

	age	sex	bmi	children	smoker	region	charges
0	0.021739	0	0.326051	0	1	3	16884.92400
1	0.000000	1	0.486346	1	0	2	1725.55230
2	0.217391	1	0.465319	3	0	2	4449.46200
3	0.326087	1	0.184189	0	0	1	21984.47061
4	0.304348	1	0.352813	0	0	1	3866.85520

Figure 2: Scaling and Encoding

[Fig 2] There are slightly more number of male genders than female.

[Fig 3]:- The graph shows that the BMI is having a normal distribution and people who are aged 30 are having more BMI and the count is more than 120. The normal BMI range for both men and women is 18.5 to 24.9, but it can be seen that more number of people aged 25 to 37 are overweight which will increase the insurance cost that they can get.

[Fig 4] The density of the insurance charges is highest at $10,000 which says that it is a common charge. Moreover, it can be seen that the insurance charges distribution is right skewed which says that most of the insurance policies are affordable to people.

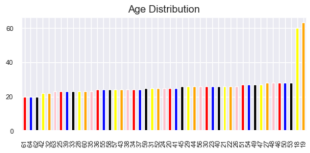

Figure 3: Age Distribution of Dataset

Figure 4: Gender Distribution of Dataset

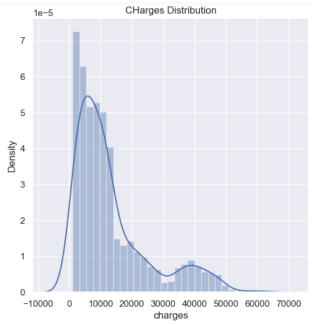

Figure 5: Charges Distribution of Dataset

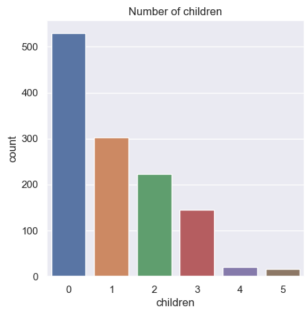

Figure 6: Children Distribution of Dataset

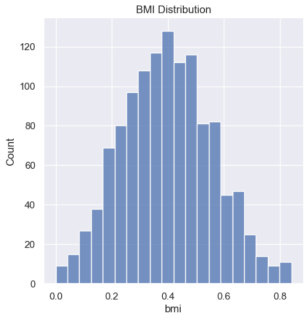

Figure 7: Normal BMI

```
In [1222]: # There are outliers in the dataset which need to be removed to do this I have used IQR method.
           sns.boxplot(Insurance_dataset_Train_df['bmi'])
Out[1222]: <Axes: >
```

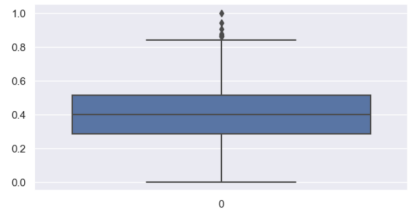

Figure 8: Outlier Detection

[Fig 5]:- The count of people who are not having any children is the highest.

[Fig 6] The BMI seems to be a normal distribution after removing the outliers.

[Fig 7] There are Few outliers which exist in the dataset in the BMI column. I have used the IQR method to detect and remove the outliers**[3]**. Outliers made a significant impact on calculating the final price, after removing these outliers the BMI attribute seems to be more normally distributed.

The main correlation which I found was that the columns smoker,BMI and Age are the important factors that influence the target attribute i.e charges. Moreover, sex,children and region does not seem to affect the target attribute much, so we can say that the coorelation among them is very less.

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

Using the K-Fold cross validation the optimal model which can be used it Random Forest Regressor**[2]**. R-Square score is used to evaluate the performance of a machine learning model, here I have got the R2 as 0.830 and a mean square error of 23706615.52.

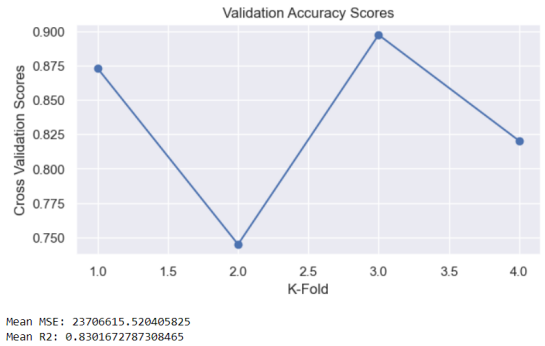

Figure 9: K-Fold Graph

Based on the above graph there is a slight fluctuation between the accuracies from 75% to 90% for kfold=4, if the score is higher we can say that the model is performing correctly.

0.5 Report the final total price that you got.

Total Price of the new employees: 1303485.84
The overall insurance on the organization is:- 17713768.54

1 REFERENCES

1] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html
2] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
3] https://www.analyticsvidhya.com/blog/2022/09/dealing-with-outliers-using-the-iqr-method/