MEDICAL INSURANCE COST PREDICTION

PC 25 KATYAYINI GUPTA
PC 26 SWATI SINGH
PD 19 NIYATI DUBEY
PD 26 ATHARVA LAKKAS
PH 16 AADITYA SINGH

INTRODUCTION

- WE ARE ON A PLANET FULL OF THREATS AND UNCERTAINTY. PEOPLE, HOUSEHOLDS, COMPANIES, PROPERTIES, AND PROPERTY ARE EXPOSED TO DIFFERENT RISK FORMS. AND THE RISK LEVELS CAN VARY. THESE DANGERS CONTAIN THE RISK OF DEATH, HEALTH, AND PROPERTY LOSS OR ASSETS. LIFE AND WELLBEING ARE THE GREATEST PARTS OF PEOPLE'S LIVES. BUT, RISKS CANNOT USUALLY BE AVOIDED, SO THE WORLD OF FINANCE HAS DEVELOPED NUMEROUS PRODUCTS TO SHIELD INDIVIDUALS AND ORGANIZATIONS FROM THESE RISKS BY USING FINANCIAL CAPITAL TO REIMBURSE THEM. INSURANCE IS, THEREFORE,A POLICY THAT DECREASES OR REMOVES LOSS COSTS INCURRED BY VARIOUS RISKS.

BUSINESS PROBLEM

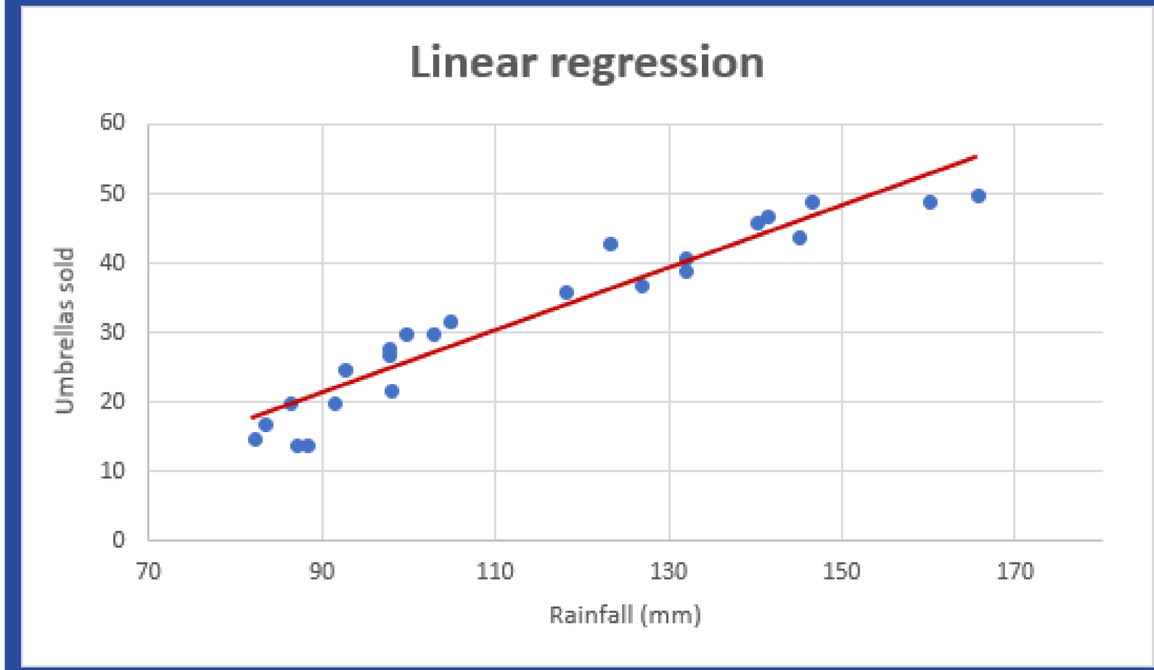
- CONCERNING THE VALUE OF INSURANCE IN THE LIVES OF INDIVIDUALS, IT BECOMES IMPORTANT FOR THE COMPANIES OF INSURANCE TO BE SUFFICIENTLY PRECISE TO MEASURE OR QUANTIFY THE AMOUNT COVERED BY THIS POLICY AND THE INSURANCE CHARGES WHICH MUST BE PAID FOR IT. VARIOUS VARIABLES ESTIMATES THESE CHARGES. EACH FACTOR OF THESE IS IMPORTANT. IF ANY FACTOR IS OMITTED WHEN THE AMOUNTS ARE COMPUTED, THE POLICY CHANGES OVERALL. IT THEREFORE CRITICAL THAT THESE TASKS ARE PERFORMED WITH HIGH ACCURACY. AS HUMAN MISTAKES ARE COULD OCCUR, INSURERS USE PEOPLE WITH EXPERIENCE IN THIS AREA. THEY ALSO USE DIFFERENT TOOLS TO CALCULATE THE INSURANCE PREMIUM.

REGRESSION MODELS

A regression model provides a function that describes the relationship between one or more independent variables and a response, dependent, or target variable.

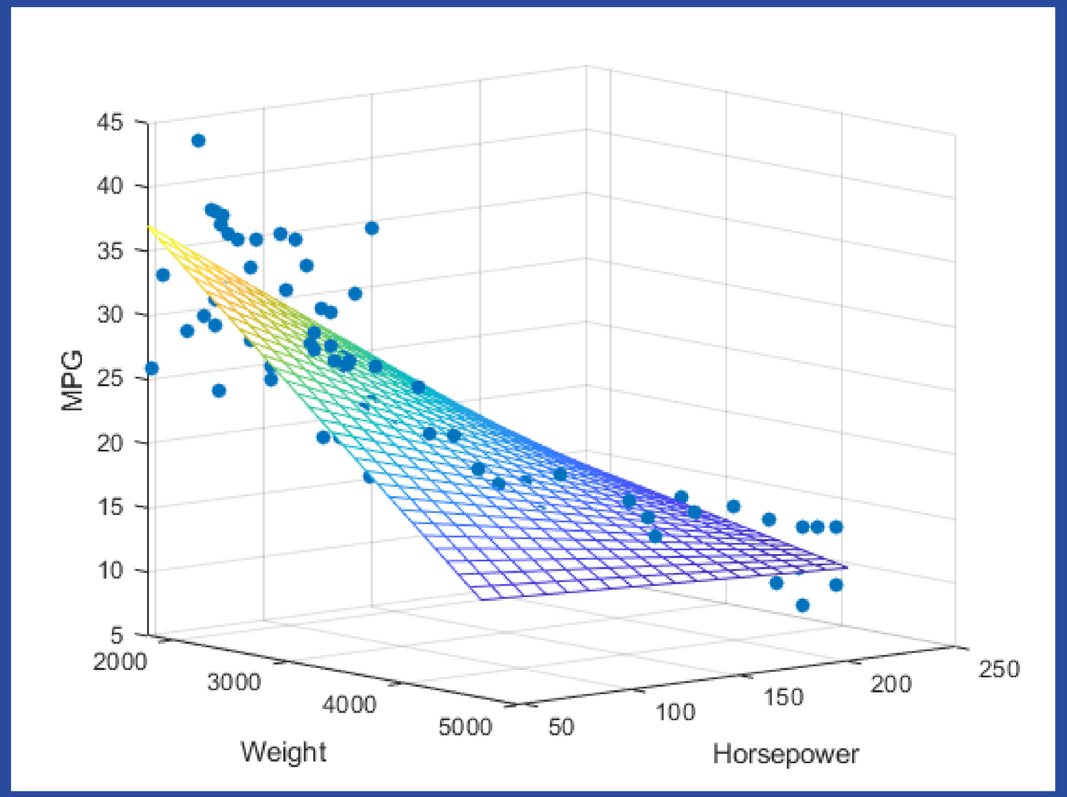
LINEAR REGRESSION

The lower the variability in the data, the stronger the relationship and the tighter the fit to the regression line.



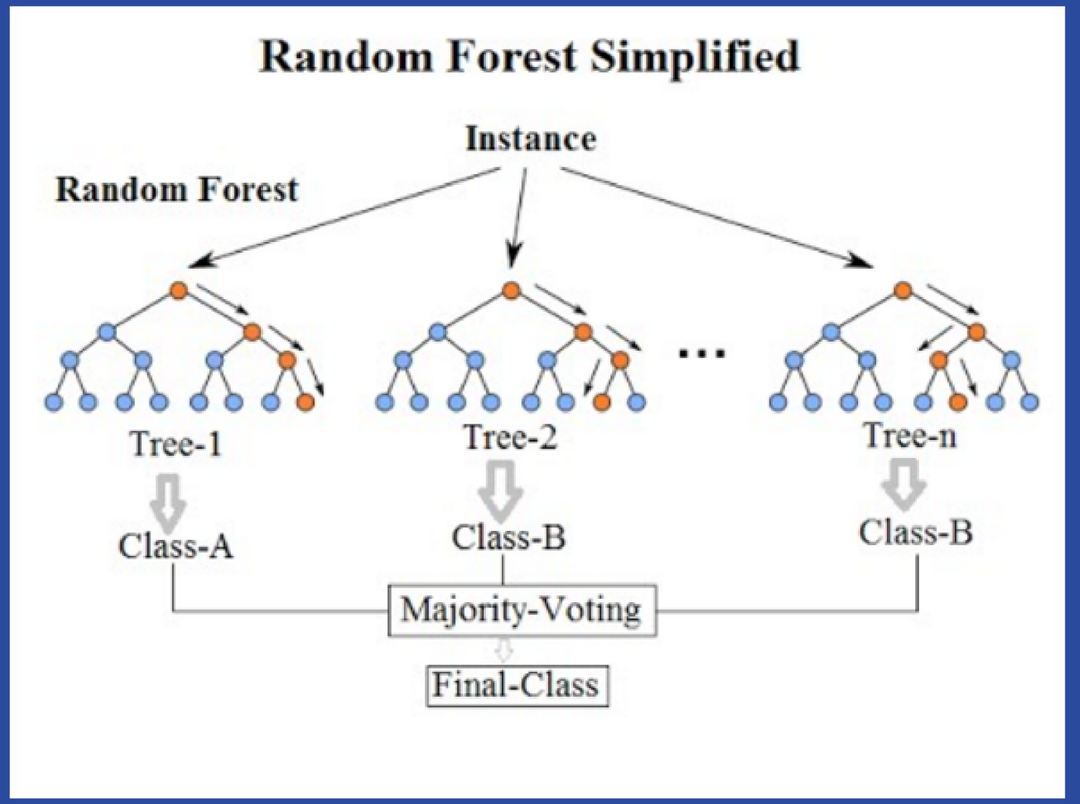
MULTIPLE LINEAR REGRESSION

MULTIPLE LINEAR REGRESSION IS A STATISTICAL ANALYSIS TECHNIQUE USED TO PREDICT A VARIABLE'S OUTCOME BASED ON TWO OR MORE VARIABLES.



RANDOM FOREST REGRESSION

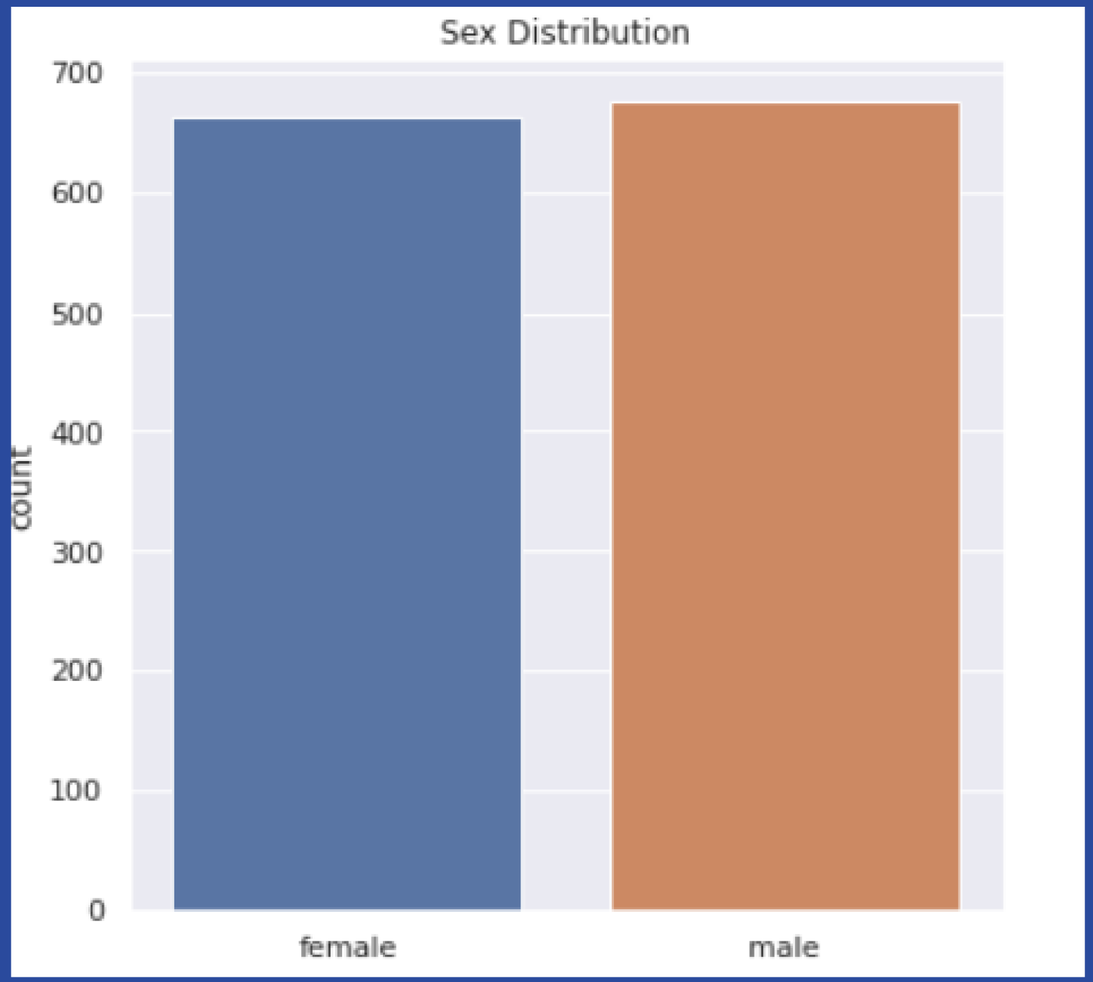
RANDOM FOREST REGRESSION IS A SUPERVISED LEARNING ALGORITHM THAT USES ENSEMBLE LEARNING METHOD FOR REGRESSION.

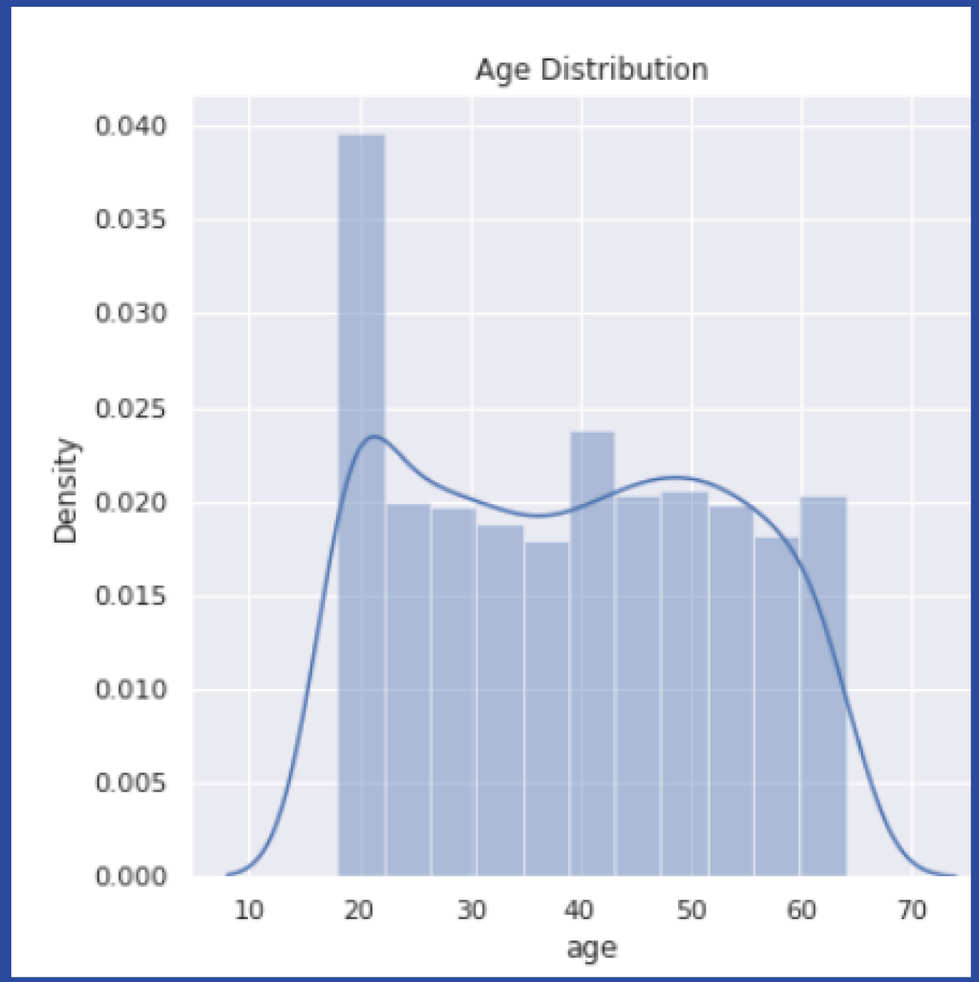


DATA ANALYSIS

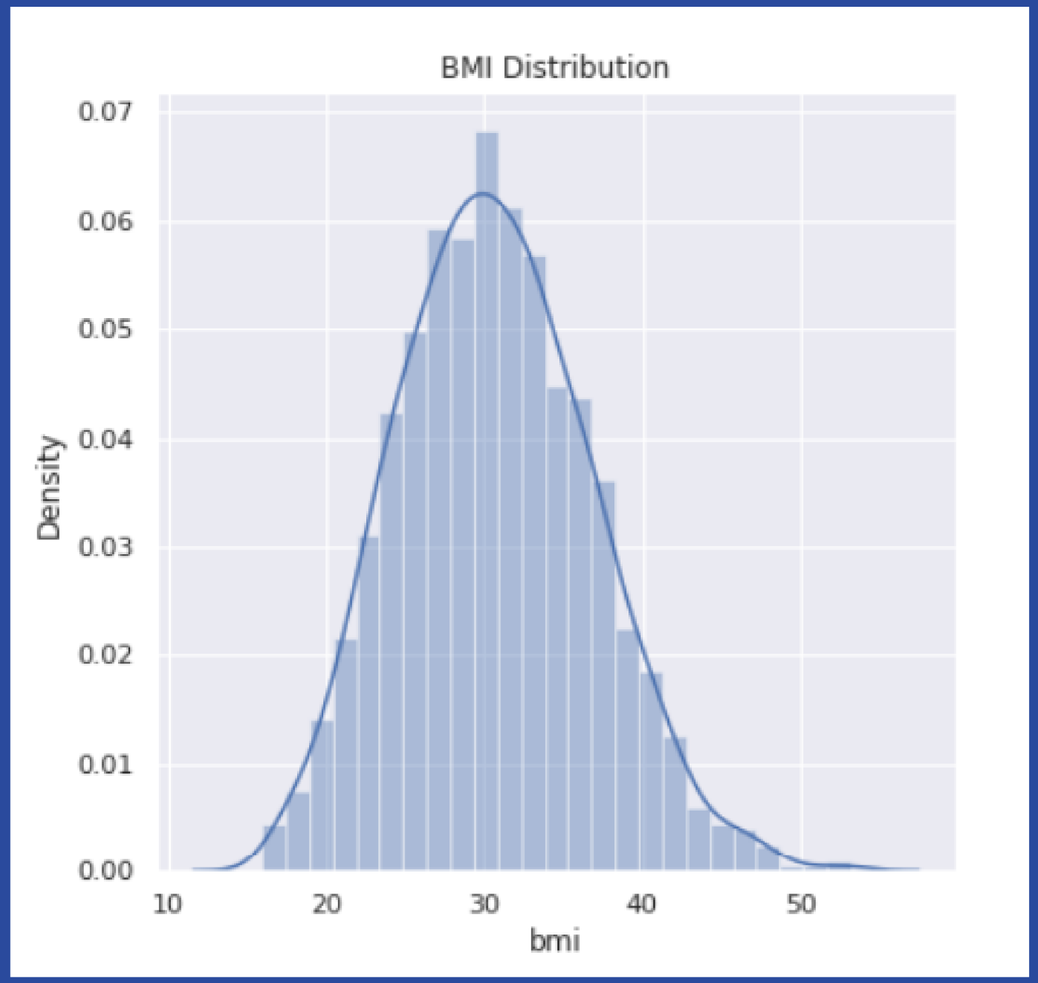
DATA ANALYSIS IS A PROCESS FOR OBTAINING RAW DATA AND CONVERTING IT INTO INFORMATION USEFUL FOR DECISION-MAKING BY USERS.

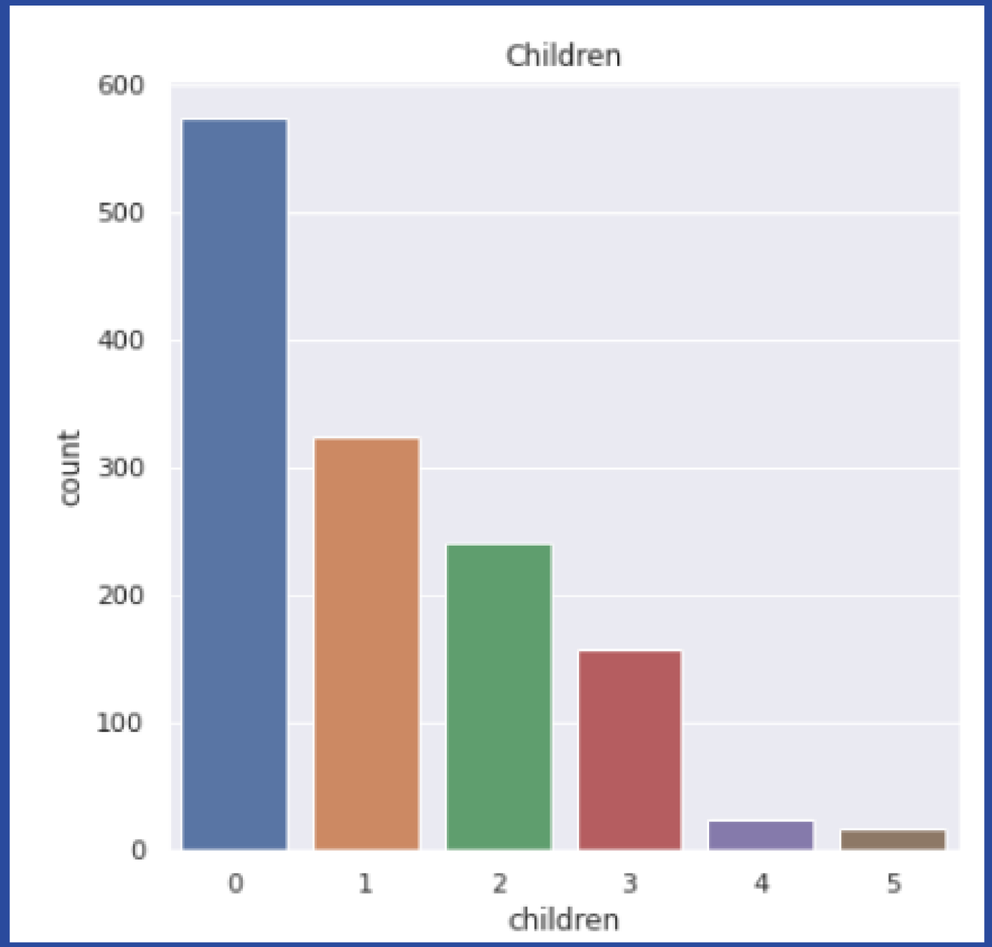
DATA ANALYSIS



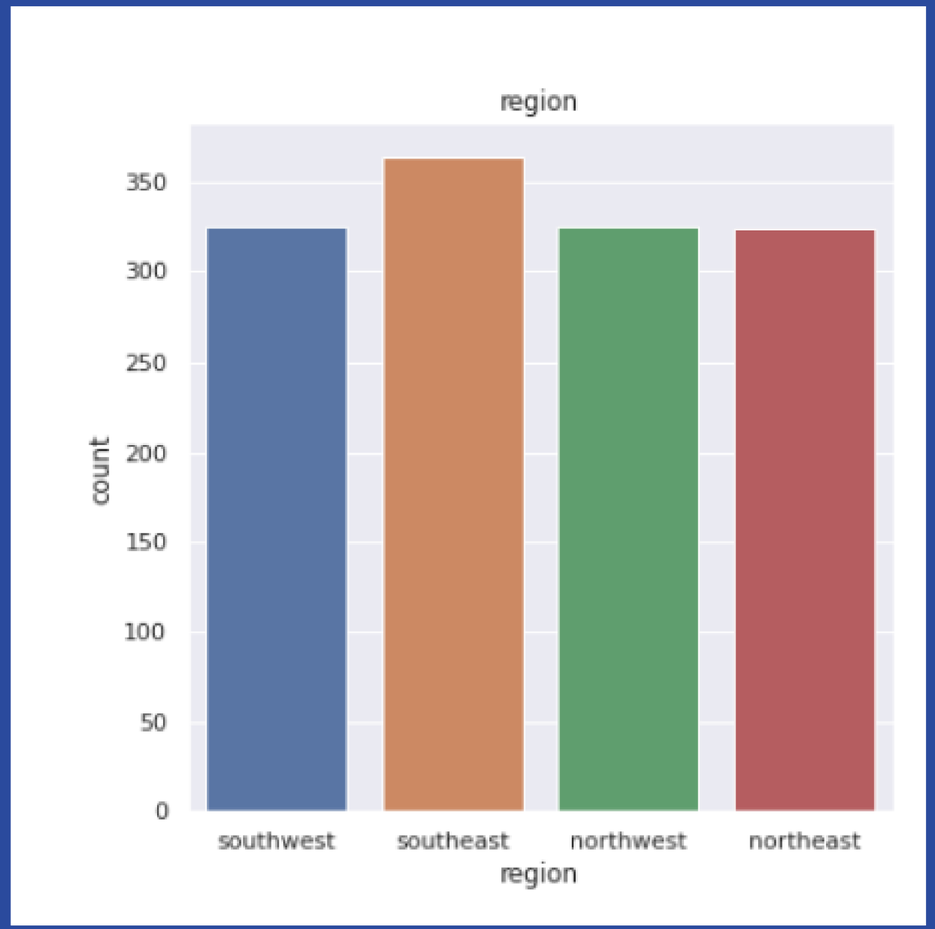


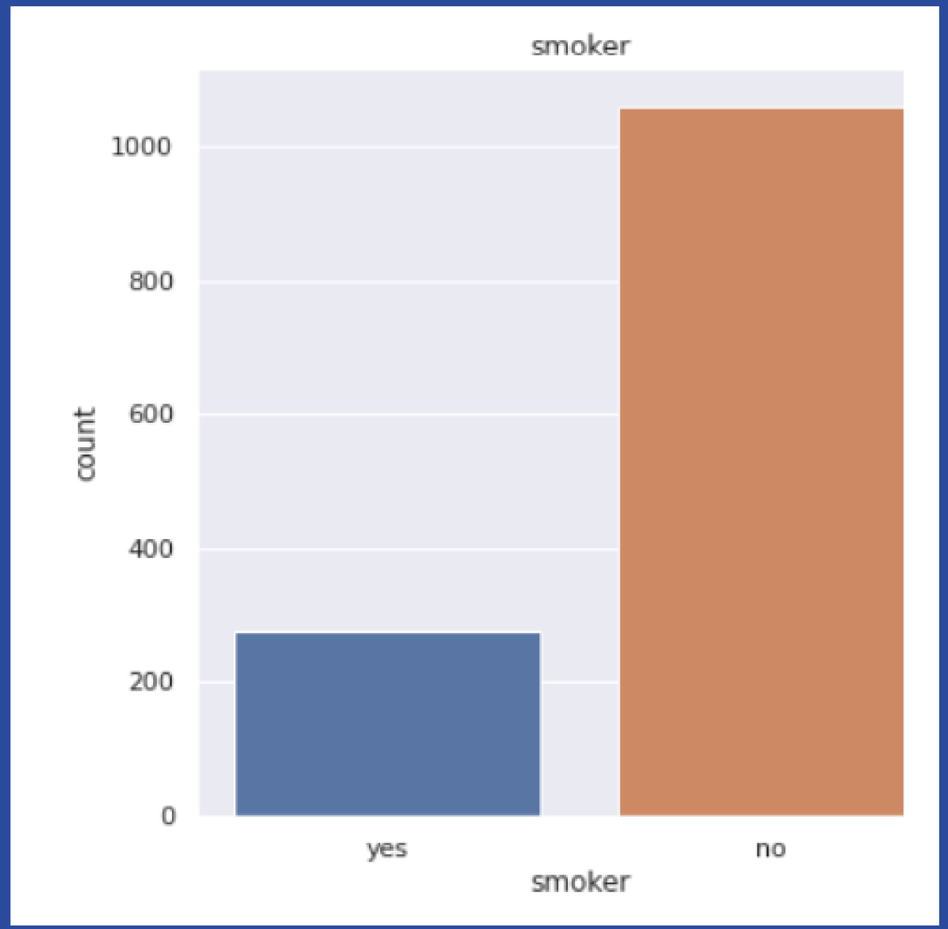
DATA ANALYSIS



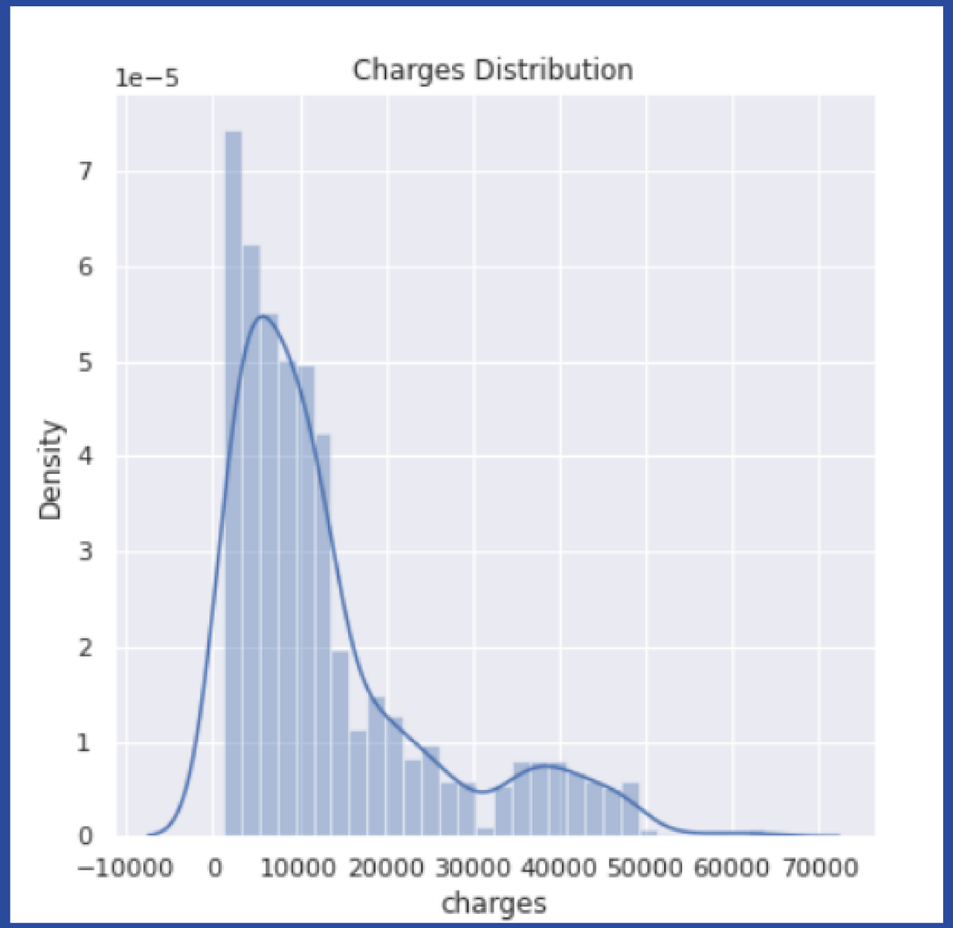


DATA ANALYSIS





DATA ANALYSIS




```python
input_data = (31,1,25.74,0,1,0)

# changing input_data to a numpy array
input_data_as_numpy_array = np.asarray(input_data)

# reshape the array
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = regressor.predict(input_data_reshaped)
print(prediction)

print('The insurance cost is USD ', prediction[0])
```

```
[3760.0805765]
The insurance cost is USD  3760.080576496046
```

DESIGNING & IMPLEMENTATION

1

DATA PREPARATION & CLEANING

The data has been imported from kaggle website. The website provides with a variety of data and the data used for the project is an insurance amount data. The data included various attributes such as age, gender, body mass index, smoker and the charges attribute which will work as the label for the project. The data was in structured format and was stores in a csv file format. The data was imported using pandas library.

TRAINING

Once training data is in a suitable form to feed to the model, the training and testing phase of the model can proceed. During the training phase, the primary concern is the model selection. This involves choosing the best modelling approach for the task, or the best parameter settings for a given model. In fact, the term model selection often refers to both of these processes, as, in many cases, various models were tried first and best performing model (with the best performing parameter settings for each model) was selected.

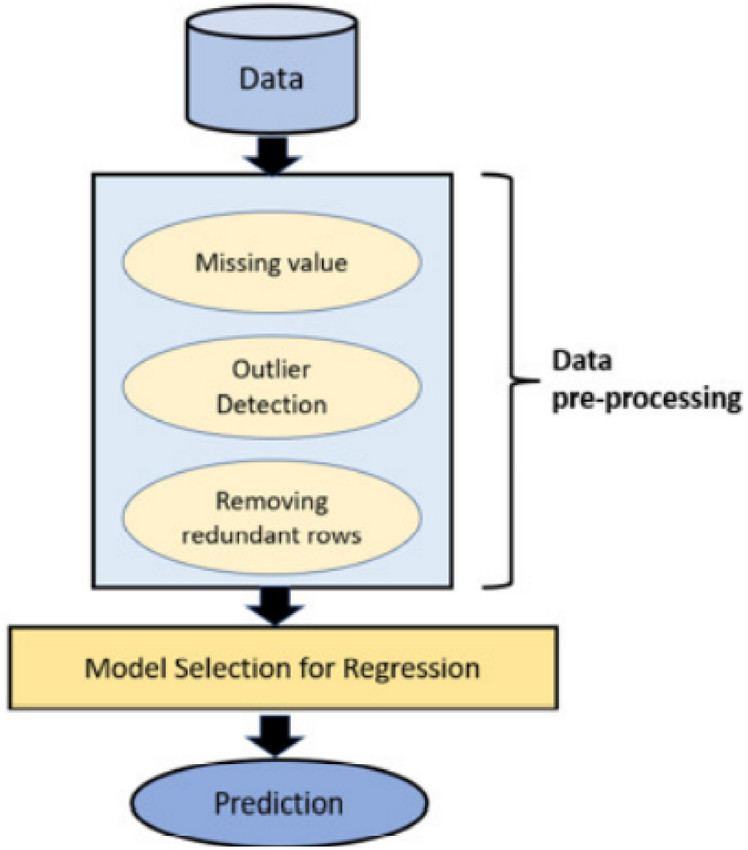
3

PREDICTION

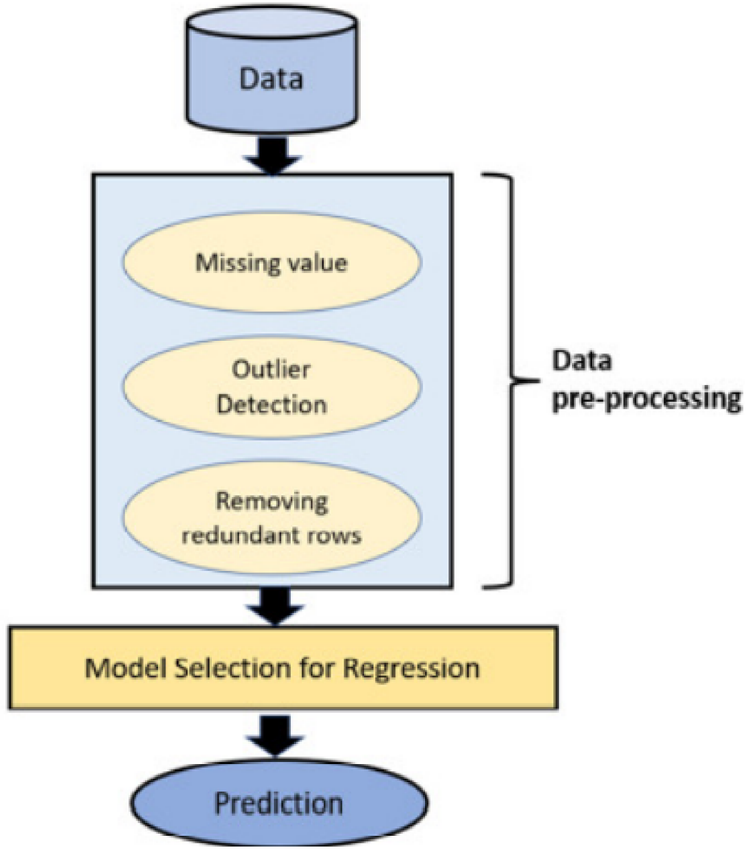
The model was used to predict the insurance amount which would be spent on their health. The model used the relation between the features and the label to predict the amount.
Accuracy defines the degree of correctness of the predicted value of the insurance amount. The model predicted the accuracy of model by using different algorithms, different features and different train test split size. The size of the data used for training of data has a huge impact on the accuracy of data. The larger the train size, the better is the accuracy.
The model predicts the premium amount using multiple algorithms and shows the effect of each attribute on the predicted value.

OBSERVATION

- *From the preceding calculations, it can be seen that Polynomial Regression outperforms other models for the proposed MLHIPS giving an accuracy of 74.97%.*
- *Besides, Polynomial Regression Ridge and Lasso Regression of MLHIPS have achieved an accuracy value of 75.82% and 75.86%, respectively.*
- *The Multiple linear regression model of MLHIPS has achieved an accuracy of 75.86%. whereas the simple linear regression had an accuracy of 62.86% which was the lowest among all.*



- *In the above scenario the dataset is partitioned into 80-20 ratio for training and testing purpose.*
- *In the dataset of 70:30 ratio, the accuracy of the polynomial regression decreases from previous value of 80.97% to 80.54% which is negligible, and similarly the rest of the models have also produced a drop in the accuracy values.*
- *The multiple linear regression, ridge regression and lasso regression accuracy values are reduced by approximately 2% units from their past values.*



FUTURE ROADMAP

- In this paper we discussed some of the traditional regression models for our proposed problem statement, moving forward some of the other techniques like Support Vector Machine (SVM), XGBoost, Decision Tree (CART), Random Forest Classifier and Stochastic Gradient Boosting needs to be addressed as the future work.
- Several optimization techniques such as the Genetic Algorithm or the Gradient Descent Algorithm maybe applied on top of model evaluation. We can also apply some feature selection techniques to our dataset before we train our model to gain a good accuracy value as some of the features may be omitted while predicting the charges.
- Besides a model to perform well a good balanced dataset with a greater number of observations is required which will reduce the variability of the model so in the future if, we get more data than the model can be trained well.

CONCLUSION

we use various machine learning regression models and deep neural networks to forecast charges of health insurance based on specific attributes, on medical cost personal data set from Kaggle.com. The findings are summarized in Table.

Forecasting insurance costs based on certain factors help insurance policy providers to attract consumers and save time in formulating plans for every individual.

THANK YOU!