Medical Insurance Premium Prediction

https://github.com/Shantinii/GA-Capstone-Project

Problem Statement

- find out how the pricing of insurance plans are based on health conditions of individuals

- to study how physical conditions of individual relates to health issues

Data Gathering and Preparation

Added BMI column

Added BMI_Status column

```python
bmi = df['Weight']/((df['Height']/100)**2)
bmi = round(bmi,1)
df['BMI']=bmi
```

```python
: def BMI_Status(bmi):
      if bmi <= 18.4:
          return 'Underweight'
      elif bmi <= 24.9:
          return 'Normal'
      elif bmi <= 29.9:
          return 'Overweight'
      else:
          return 'Obese'
```

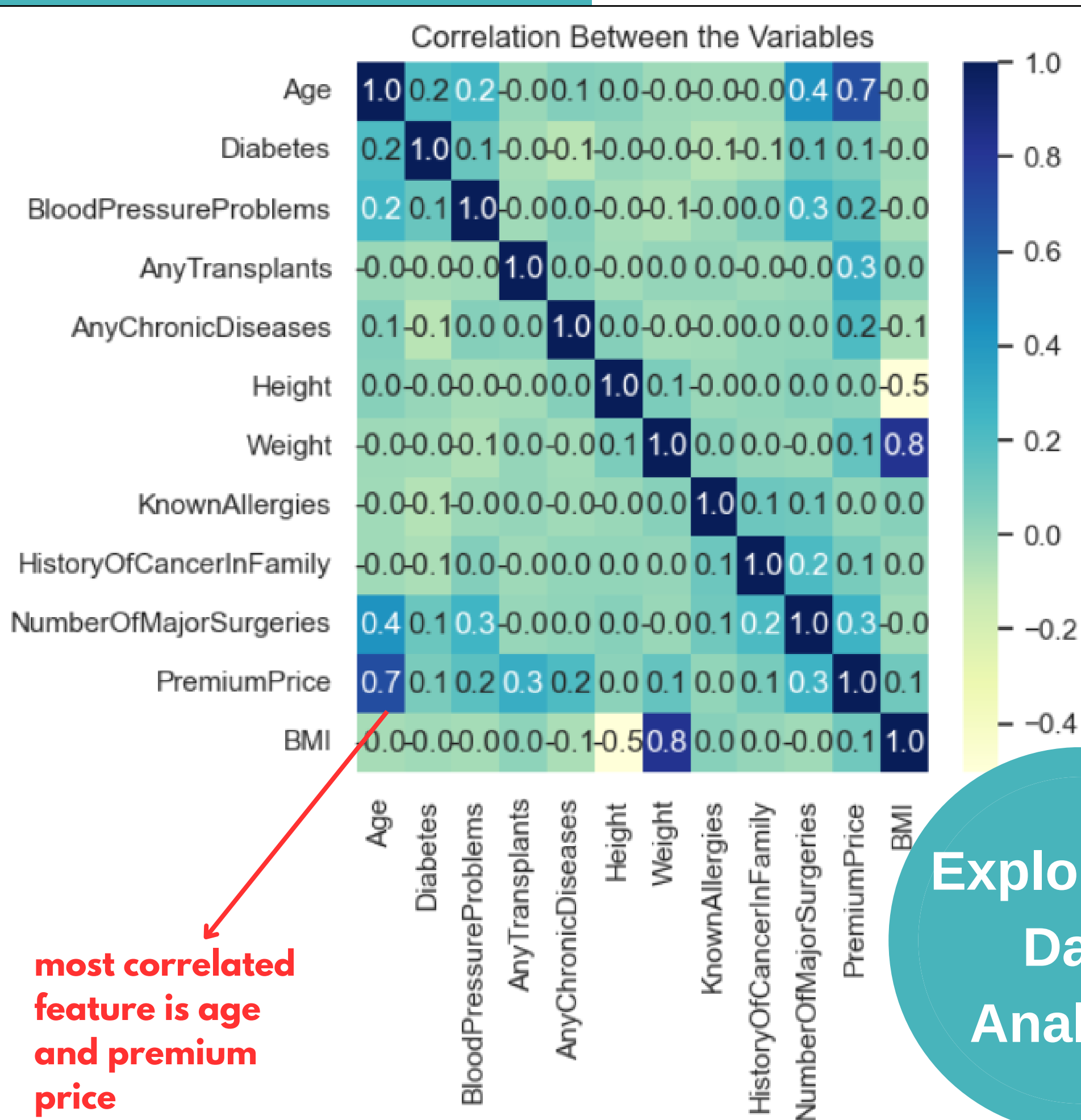| orSurgeries	PremiumPrice	BMI	BMI_Status
0	25000	23.7	Normal
0	29000	22.5	Normal
1	23000	23.6	Normal
2	28000	27.8	Overweight
1	23000	31.9	Obese

Categorize price range

Categorize age range

```python
label=['Basic','Booster','Premier','Supreme']
df['Label']=pr_bins=pd.cut(df['PremiumPrice'],bins=4,labels=label,precision=0)
df.head()
```

```python
agelabel=['Young','Adult','Senior']
df['AgeLabel']=pr_bins=pd.cut(df['Age'],bins=3,labels=agelabel,precision=0)
df.head()
```
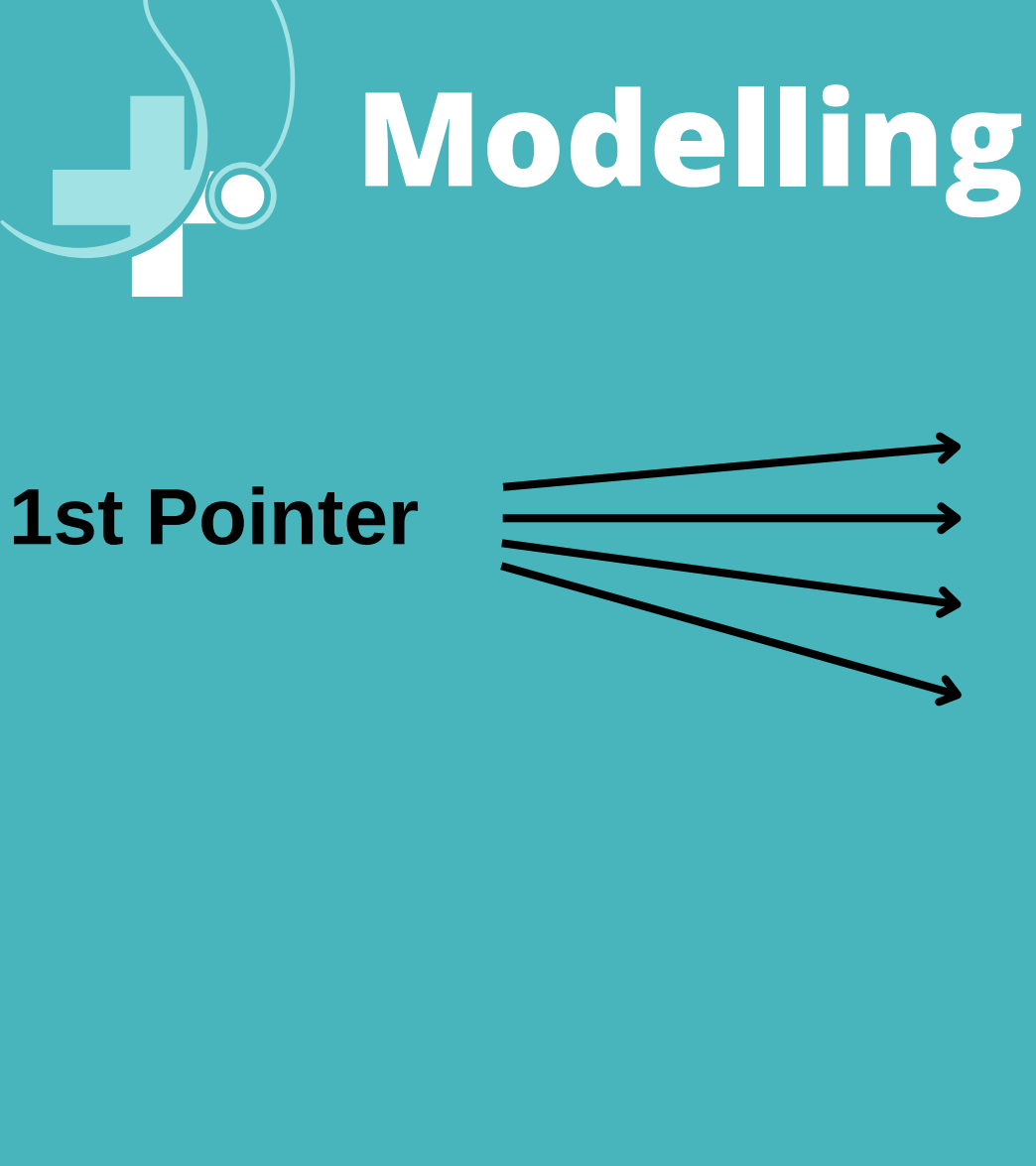
PremiumPrice	Label
25000	Booster
29000	Premier
23000	Booster
28000	Premier
23000	Booster

	Age	AgeLabel	Di
0	45	Adult	
1	60	Senior	
2	36	Adult	
3	52	Senior	
4	38	Adult	

datasets extracted from: https://www.kaggle.com/datasets/tejashvi14/medical-insurance-premium-prediction/code

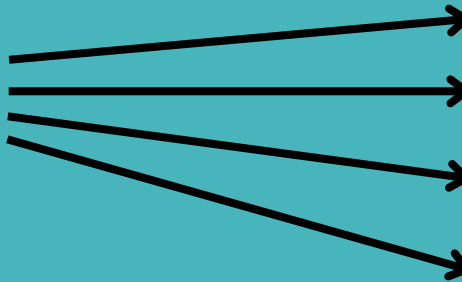Correlation Between the Variables

Strong correlation between weight and BMI.

However, the correlation between age and premium price is more suitable to take into account and do modelling based on it.
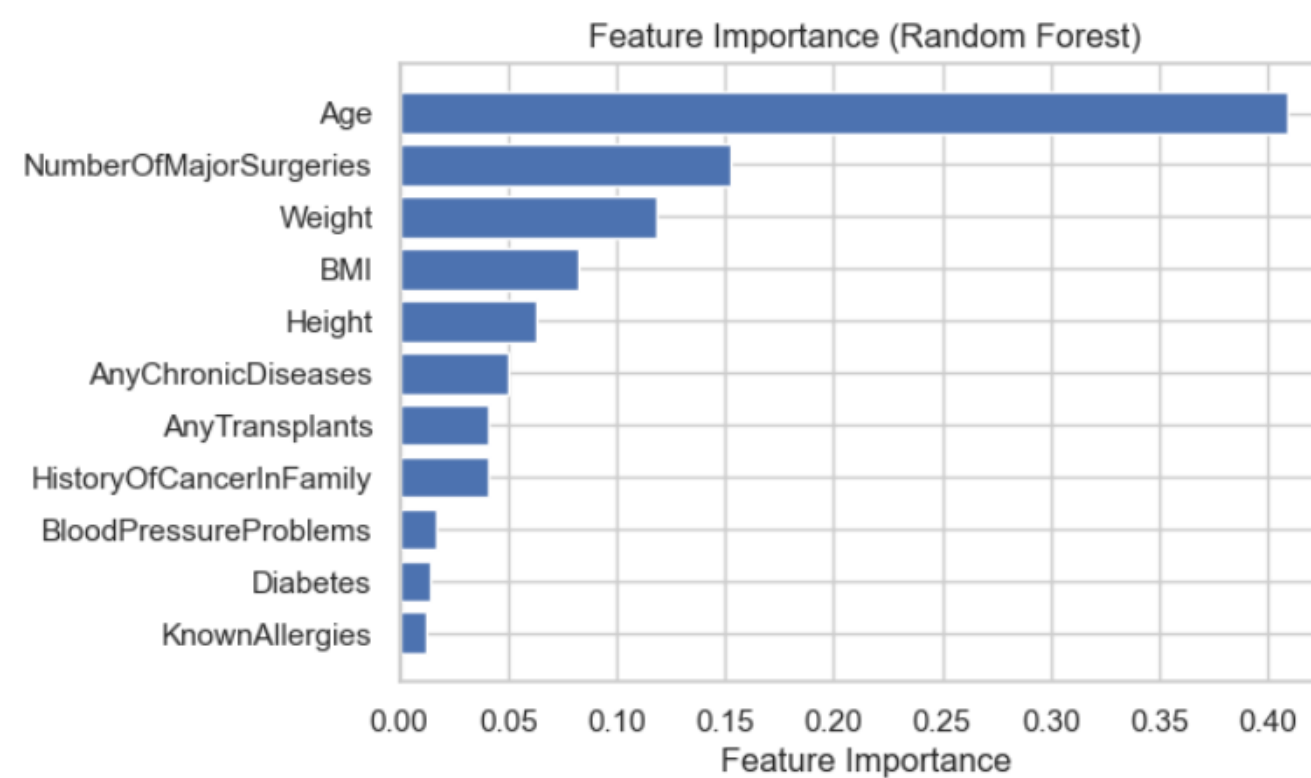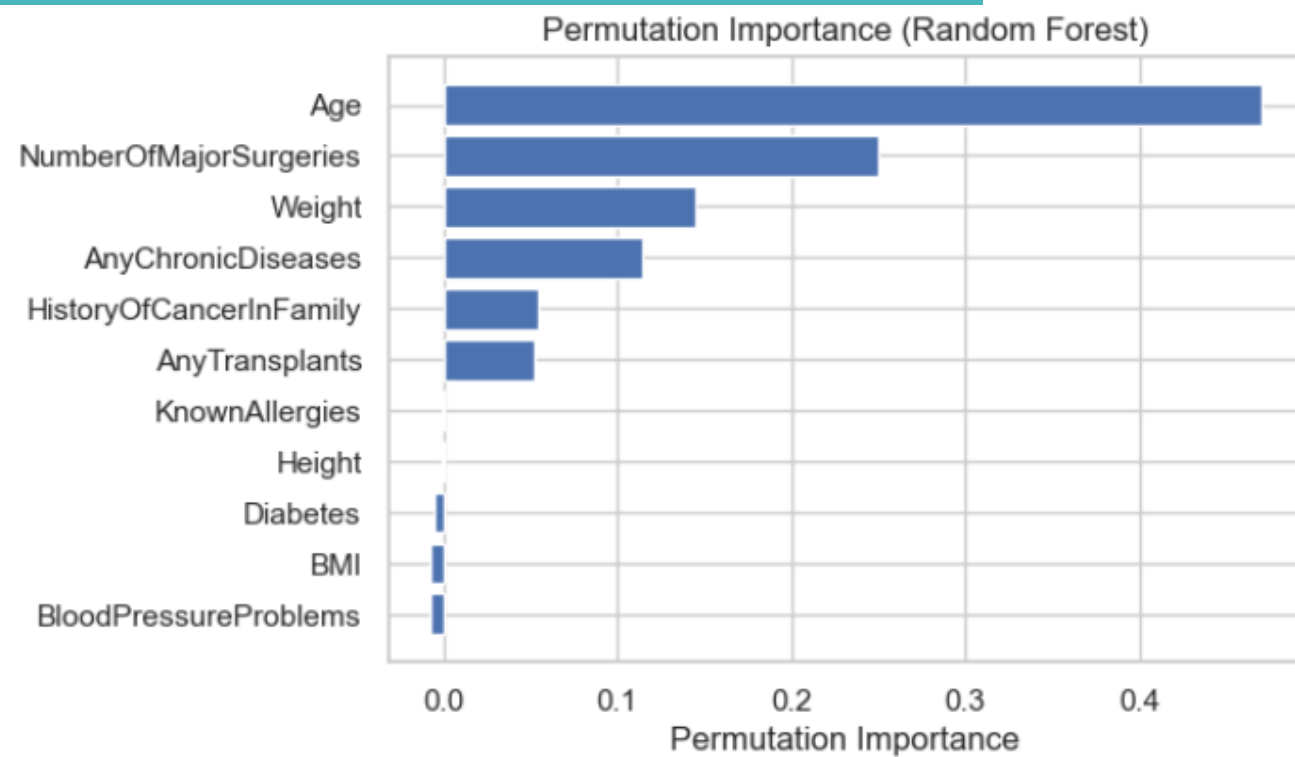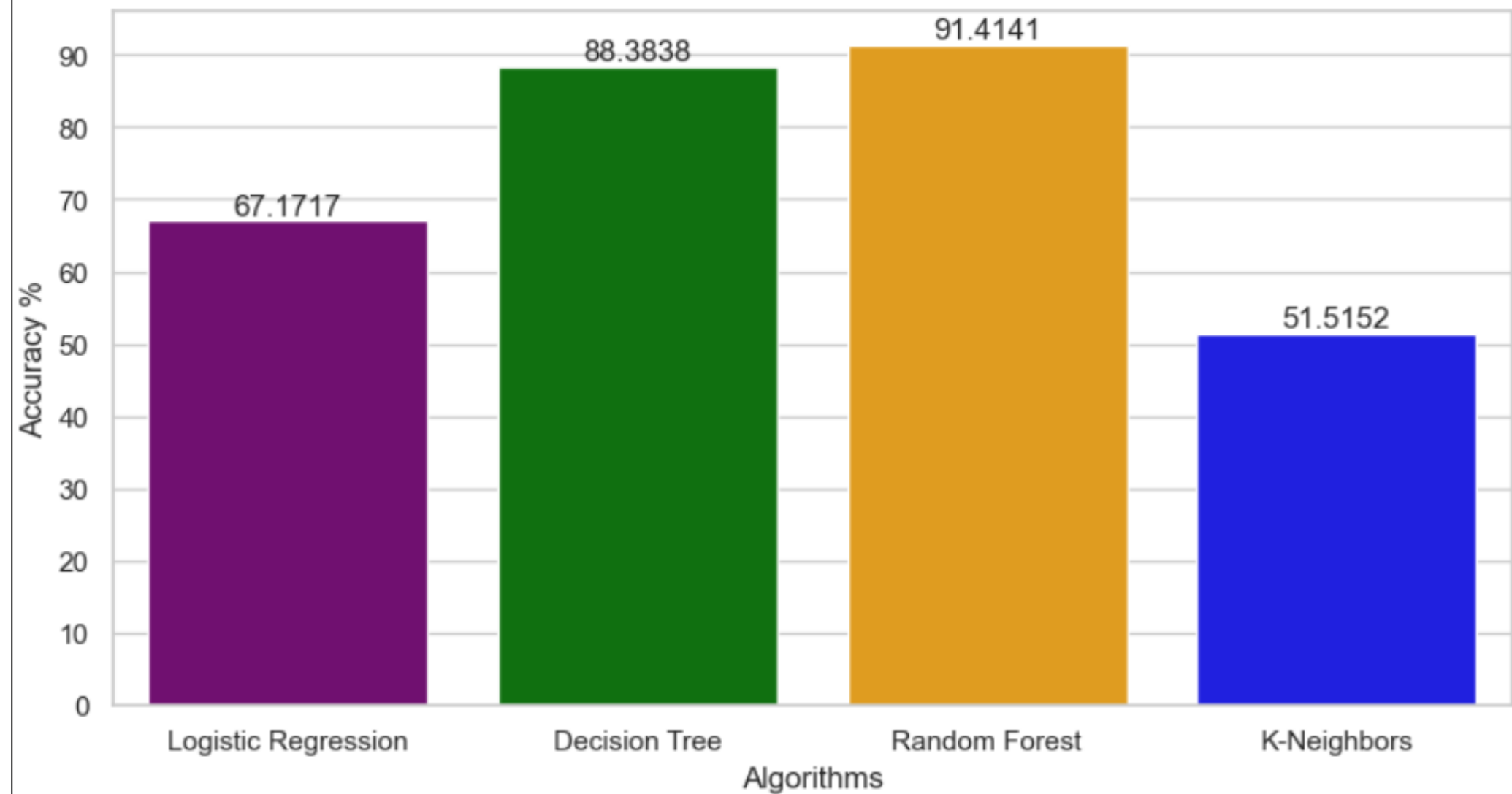
Exploratory Data Analysis

most correlated feature is age and premium price

Modelling

1st Pointer

Logistic Regression
Decision Tree
Random Forest
K-Neighbors

Top three feature i chose was age, number of major surgeries and weight.

1 Recommendation

1. Dataset could add more variables like diet, allergies and physical exercise.
2. Add more descriptive variable like the frequency of surgeries, hospital admission, claiming of insurance and so on.

2 Conclusion

- Highest accuracies score were shown in Random Forest Classifier. Therefore, random forest classifier is best fit model to predict the premium price for health insurance.
- The prediction when using 80% data for training and 20% for testing turnout to be 90% using Random Forest technique.
- The prediction when using 80% data for training and 20% for testing turnout to be 79% using Random Forest technique with the important features were selected
- From the EDA, we can observe that when the age increases, the amount of money one pays for insurance also increases.
- This can be explain with the fact that increasing age influences more illness onto a person.
- And also people those who have done a transplantation pays higher premium than those who have not done any transplantation
- Moreover, those who are obese pays more than others.
- With this, the public should do early health screening and buy health insurance in early age. People should also try to be active by walking or exercising to maintain good health.