Insurance Premium Prediction

PROJECT REPORT

P Maruthi Prasad
pmaruthiprasad22@gmail.com
GITHUB

Table of Contents

Insurance Premium Prediction

1. PROBLEM STATEMENT

Build a Regression model which will predict the Insurance premium depending on multiple factors

2. INTRODUCTION TO PROJECT

A) CASE STUDY EXPLANATION

This report is an analysis on the Insurance Premium. Each year, everybody spent lots of time to identify the best insurance with the best premium. We have Insurance premium data with various different factors which might be affect the premium cost. Here, we use the data and build a regression model which predict the Insurance premium based on other independent variables.

The aim of the project is to reduce the time spent by each individual to identify the best insurance with best premium before purchasing insurance.

B) DOMAIN KNOWLEDGE

Insurance is a way to protect ourself financially against unexpected events that could cause someone significant financial loss.

How it works: Someone purchase an insurance policy from an insurance company. This policy is a contract

that outlines what is covered and how much the insurance company will pay if something bad happens.

Paying Premiums: You pay a regular amount of money, called a premium, to the insurance company. This can be monthly, quarterly, or annually.

Making a Claim: If an event happens that is covered by someone policy (like an accident, illness, or damage to the property), he makes a claim. This means he asks the insurance company to pay for the costs related to that event.

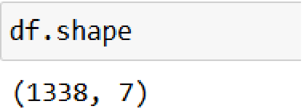
Receiving Payment: If claim is approved, the insurance company will pay the agreed amount to help cover the costs of the loss or damage.

Types of Insurance: Health Insurance, Car Insurance, Home Insurance, Life Insurance

It is very important to purchase the right insurance which covers all the losses which is expected to cover by someone before purchasing, to make this happen each individual should understand the premium required for their insurance based on which each individual can select the right insurance with the right premium cost.

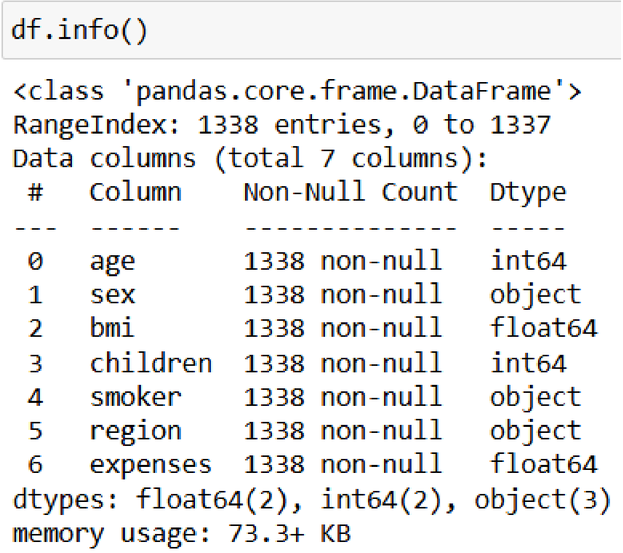
3. EXPLORATORY DATA ANALYSIS

We have insurance dataset which has 1338 observations and 7 variables, variables are shown below



```
df.shape
```

```
(1338, 7)
```



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1338 non-null   int64  
 1   sex       1338 non-null   object 
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64  
 4   smoker    1338 non-null   object 
 5   region    1338 non-null   object 
 6   expenses  1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

As per above dataset has below variables

Input variables:

1- Age

2- Sex

3- bmi

4- children

5- smoker

6- region

Output variables:

7- expenses

We have 4 numeric variables and 3 categorical variables (sex, smoker, region), out of which, 6 are independent variables and 1 is dependent variable (i.e., expenses).

4. UNDERSTANDING ON DOMAIN KNOWLEDGE BASIS:

Will understand each variable here

1. **Age:** Age is the main category which decides the expenses of each individual health, it is natural that as the age increases, the premium amount increases on average by about 8% to 10% for every year of age, according to Ted Bernstein, Director, Life Insurance Concepts Inc.

   Source: How Age Affects Life Insurance Rates (investopedia.com)

2. **Sex:** Gender also important criteria while calculating a insurance premium, it is observed that Typically, health insurance plans tend to impose higher premium costs on women of childbearing age than men. This is because women are statistically more likely to incur medical expenses related to pregnancy, childbirth, and certain gender-specific conditions such as breast and ovarian cancer. Therefore, insurers adjust premiums to account for the higher likelihood of women requiring medical care during their reproductive years.

   Source: How health insurance premium prices vary based on gender? (acko.com)

3. **BMI:** BMI or Body Mass Index is a ratio of weight and height. It can be defined as the weight in kilograms, divided by the square of the height in metres. This ratio is used to classify whether a person is underweight, overweight or obese.

   How do you calculate BMI?

   There are three easy methods to calculate BMI-

   The BMI calculation formula You can easily calculate the BMI manually using the following formula.

   $$\text{The BMI calculation formula} = \frac{\text{Weight of the person (in kg)}}{\text{(Height of the person)2 (in m2)}}$$

   A BMI which is less than 18.5 signifies you are underweight.

   A BMI between 18.5 and 24.9 is normal for a healthy adult.

   A BMI that is between 25 and 29.9 signifies that you are overweight for your height.

   A BMI that is beyond 30 indicates that you are Obese.

   A higher BMI means that the person is more susceptible to coronary heart diseases and other illnesses like diabetes and other weight-related diseases. Concurrently, the medical treatment, along with the cost of medication required for such conditions is high.

   Insurance companies use BMI to determine what your premium amount should be. The reason behind this is pretty straightforward. If they anticipate that your medical spending is bound to be more, it makes a direct impact on your premiums. Higher the expenditure, the higher will be your premium.

   Source: What is BMI, How Does It Affect Your Insurance Premium? | SBI Life

4. **Children:** Even though there is no direct evidence that how a person health will impact by having children, however number of children is one of the category identified to predict the insurance premium

5. **Smoker:** Smoking is a recognised health risk factor that dramatically raises the risk of a number of diseases, including cancer, heart disease, and respiratory conditions. As a result, smokers pay more for

life insurance than nonsmokers do since they are a greater risk to insurers so Smokers obviously pay higher life insurance premiums

Source: Why Are Life Insurance Premiums More Expensive for Smokers | ABSLI (adityabirlacapital.com)

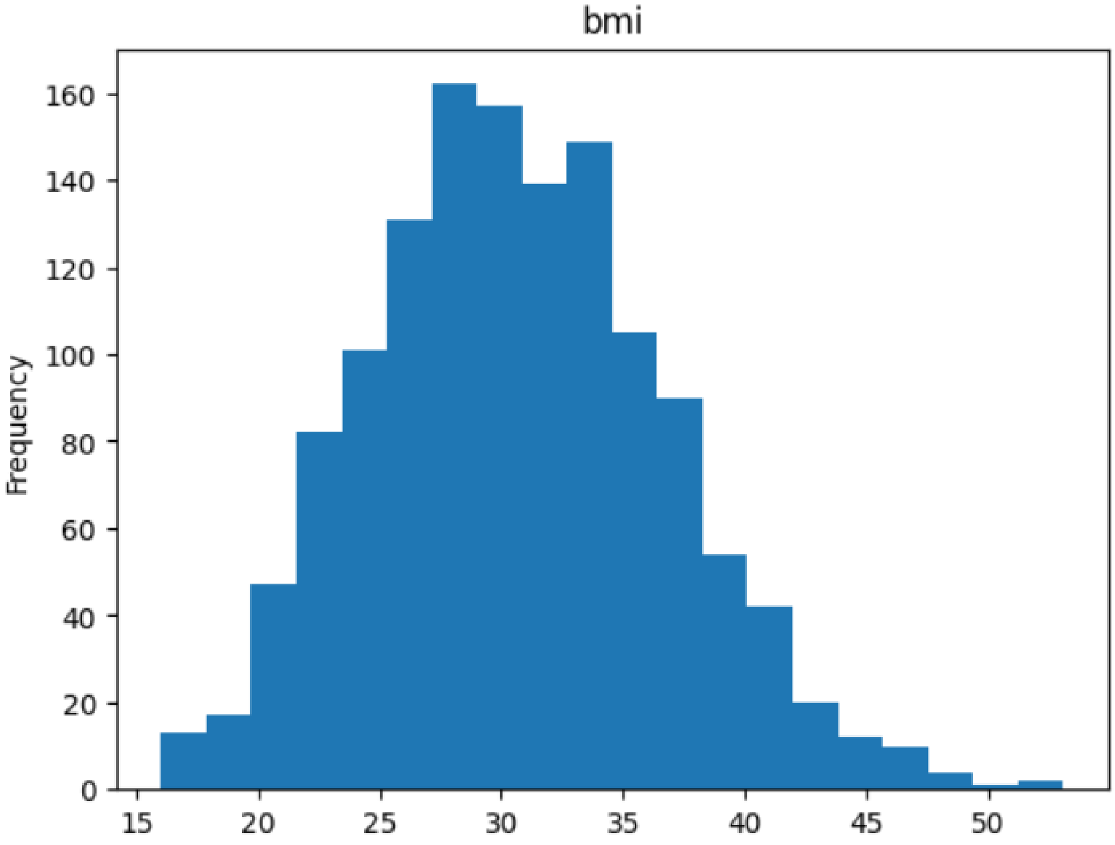
6. **Region:** Some times region also plays a crucial role to decide the insurance premium for instance in India healthcare expenses vary across each city, metropolitan cities incur higher medical charges when compared with towns and non-metro cities. Meical treatment required for 2 days hospitalization is higher is national capital when compared with Bhopal, Madhya Pradesh.

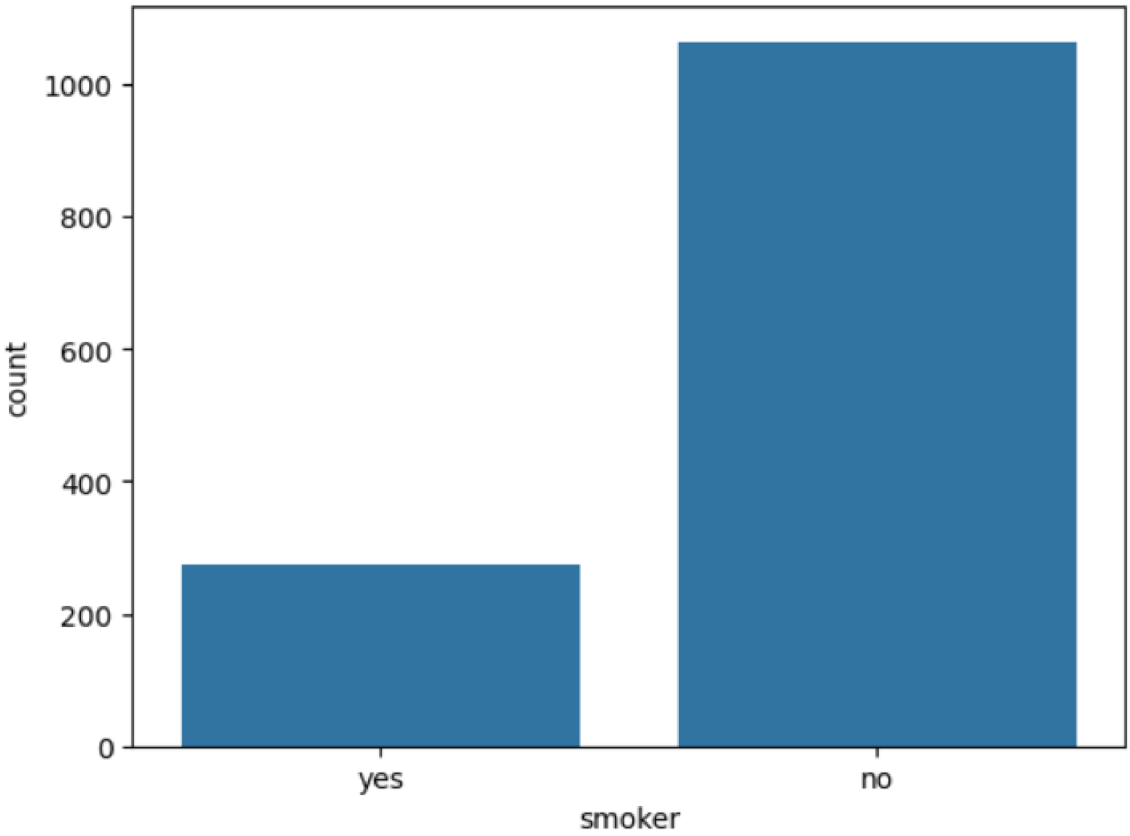
   Source: Why Are Life Insurance Premiums More Expensive for Smokers | ABSLI (adityabirlacapital.com)

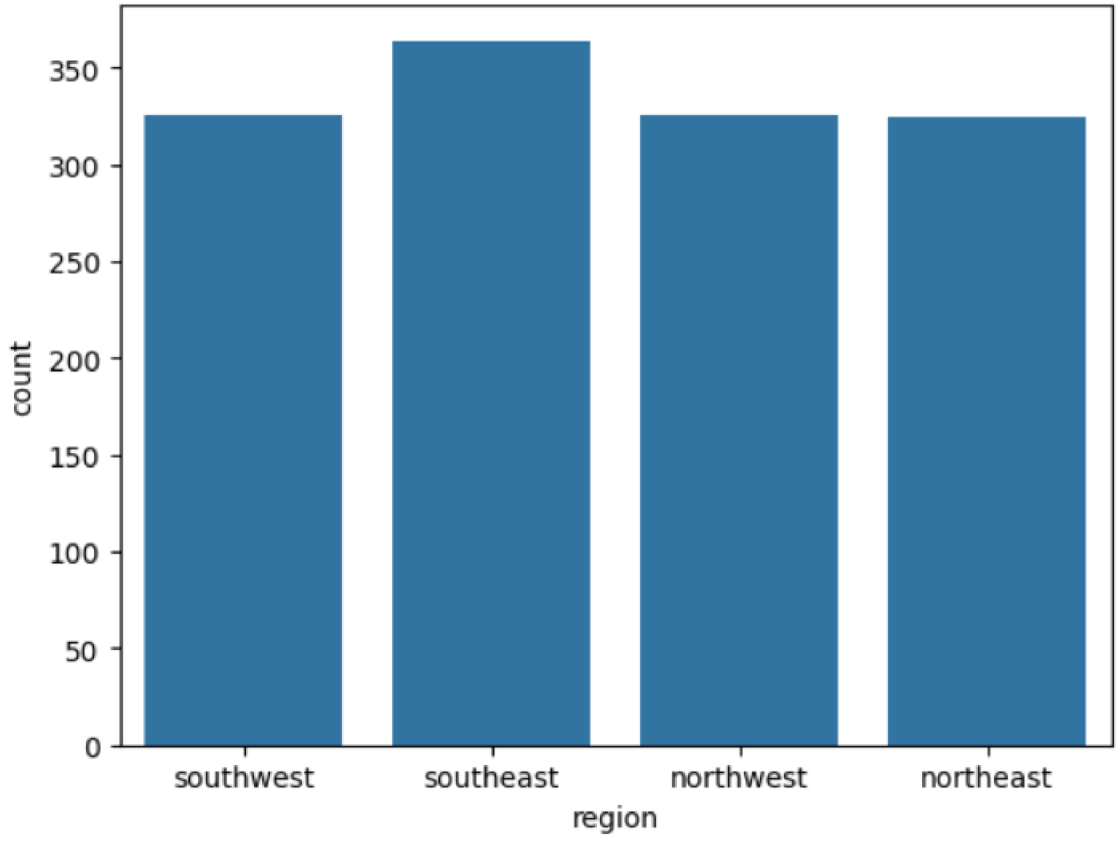
7. **Expenses:** Is the premium amount to be paid by insurer to safeguard him from any unexpected medical issues.

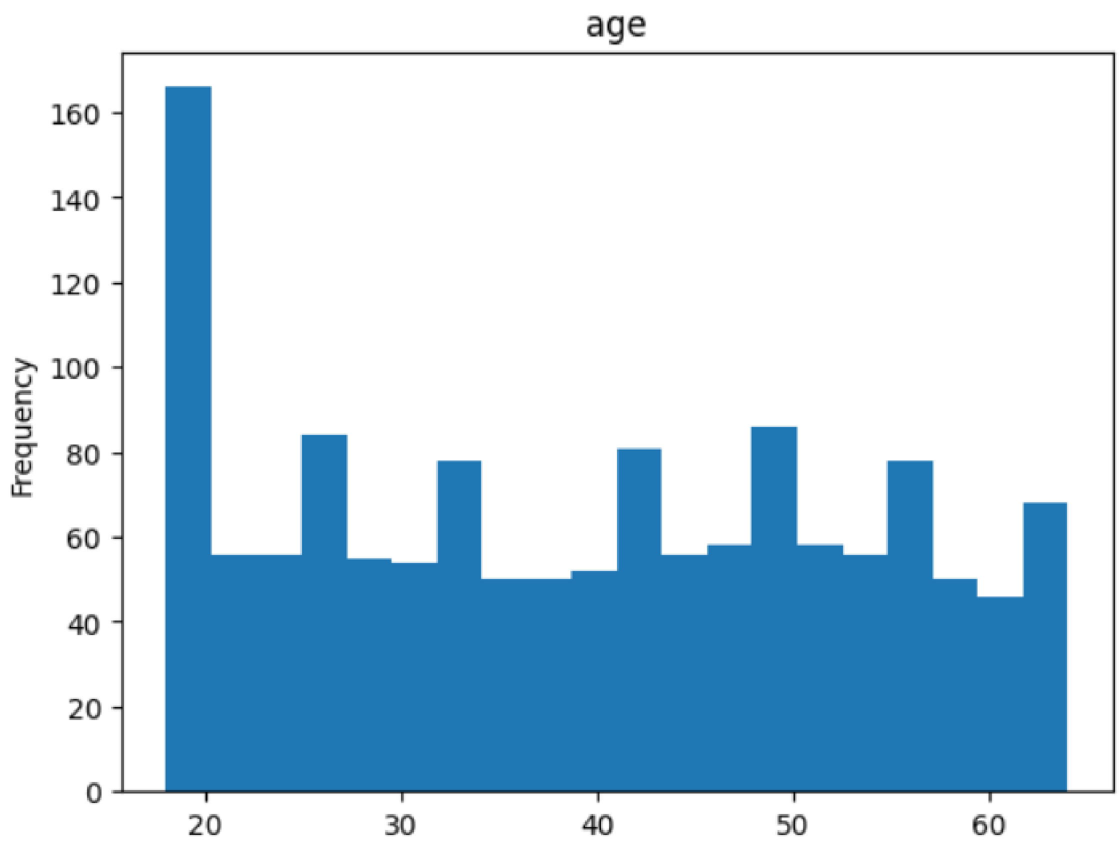
5. UNDERSTAND THE DATASET

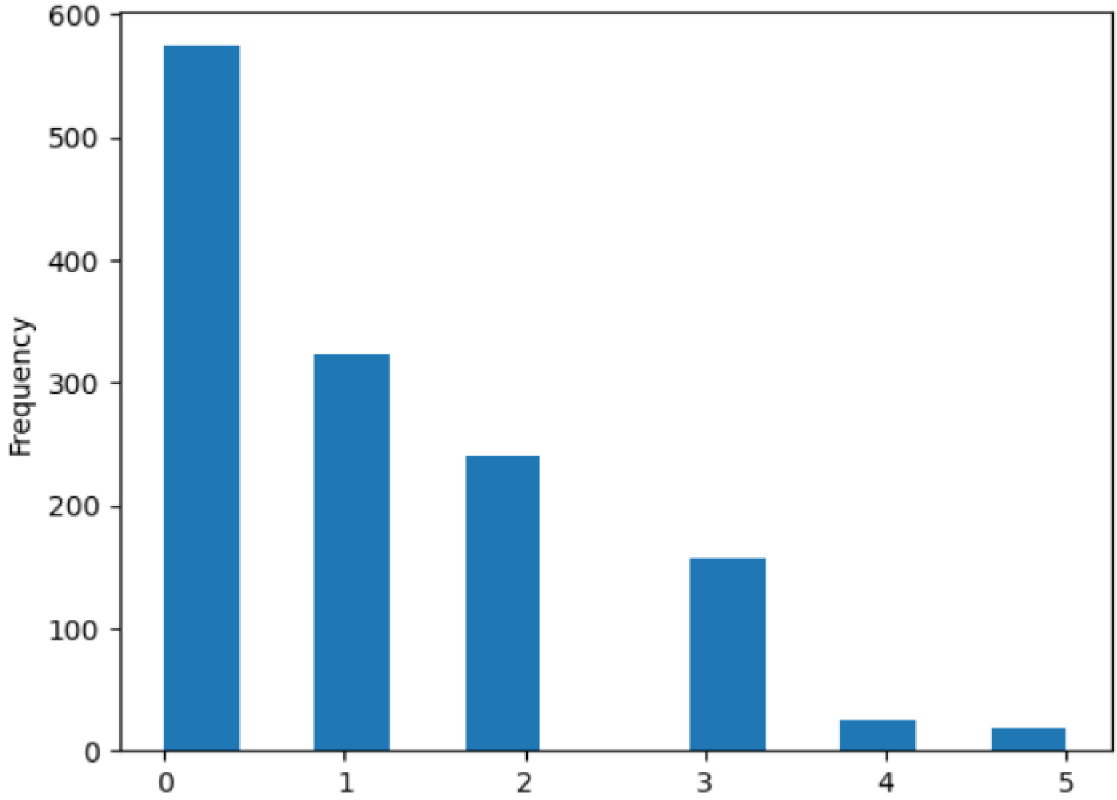
To explain the Distribution of each variable, we are using Histogram and bar graphical representation




count
1000
800
600
400
200
0
yes
no
smoker


count
350
300
250
200
150
100
50
0
southwest
southeast
northwest
northeast
region





KNOWN INSIGHTS OR FIRST LEVEL OF INSIGHTS:

1. BMI is normally distributed

2. Smokers are less in number when compared with non-smokers
3. All four regions have almost same amount of people
4. Young people have more in numbers when compared with all other age group
5. Most of the people have no children since most of the people are unmarried as they are younger and among all parents, having 1 or 2 children are more in number when compared with more than 2 children

6. PREPROCESSING AND CLEANING

A) Missing value analysis

In our data there is no missing values

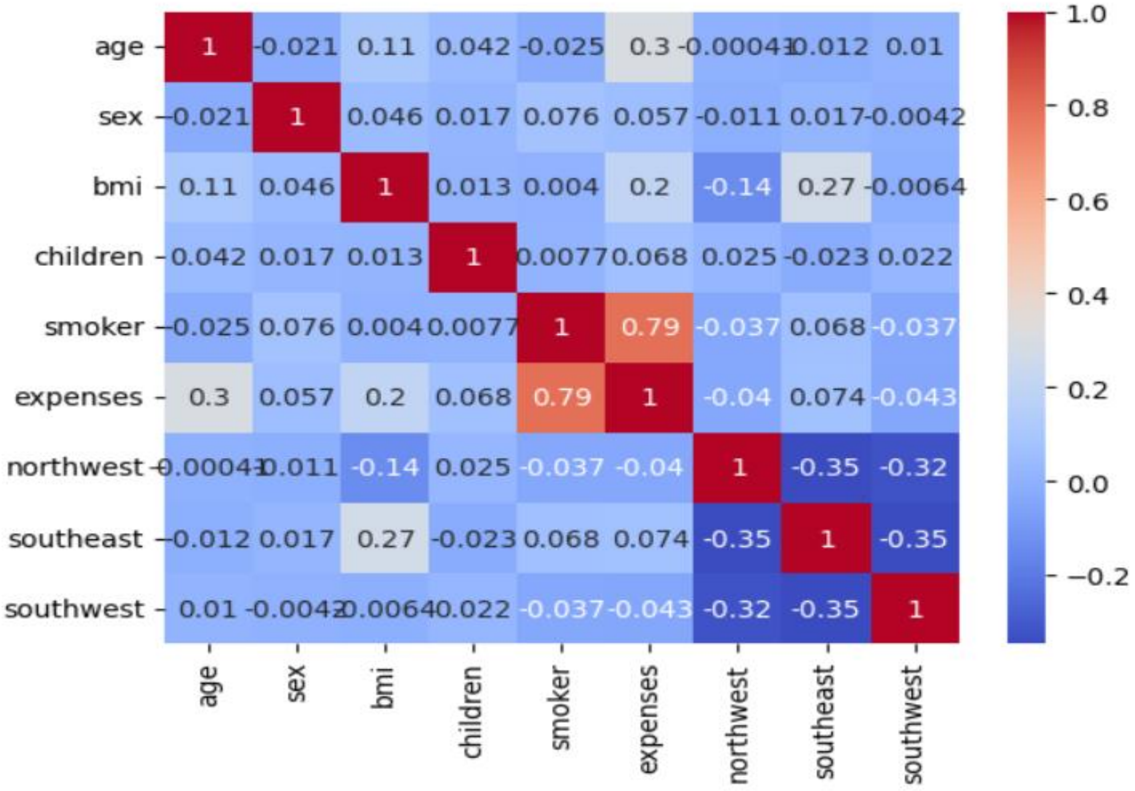
B) Target encoding

Sex and smoke columns are categorical variables and target encoding is done for the same

C) One hot encoding for region column

Region has four categories and one hot encoding is done for the same

D) Correlation

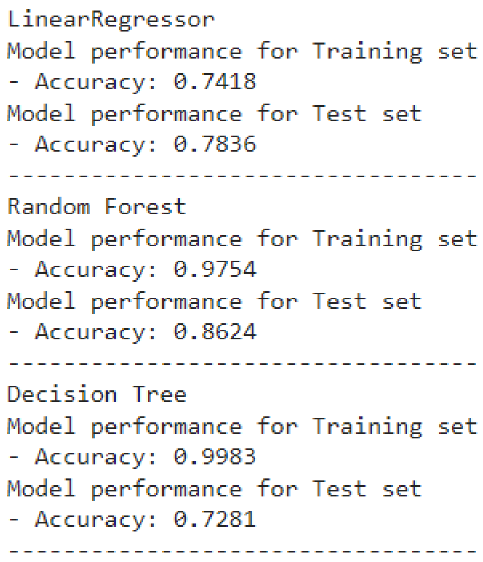


Based on above correlation smoker, BMI and age has highest correlation with expenses column

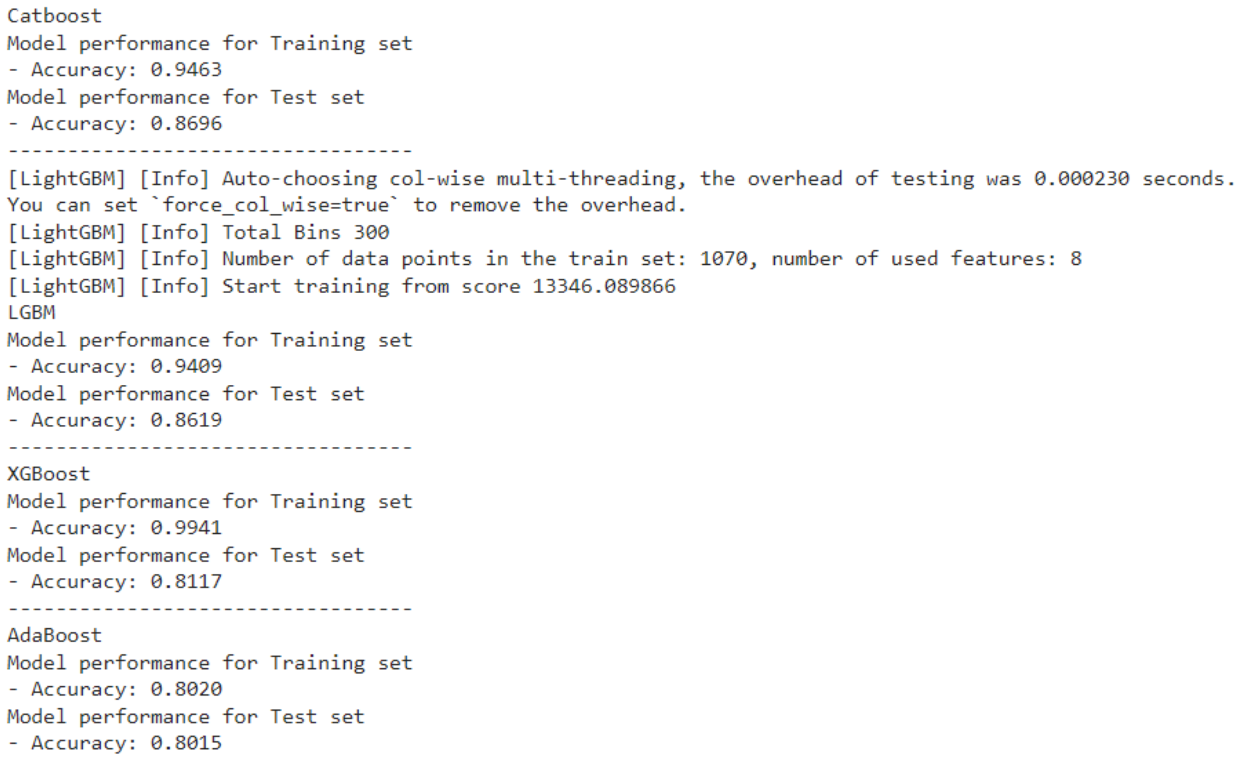
7. MODEL BUILDING:

Tried with different machine learning models such as Linear Regression, Random Forest, Decision Tree, Cat boost, Light GBM, XG boost, Ada boost

Here is the different models and their r2 score here under



```
LinearRegressor
Model performance for Training set
- Accuracy: 0.7418
Model performance for Test set
- Accuracy: 0.7836
----------------------------------
Random Forest
Model performance for Training set
- Accuracy: 0.9754
Model performance for Test set
- Accuracy: 0.8624
----------------------------------
Decision Tree
Model performance for Training set
- Accuracy: 0.9983
Model performance for Test set
- Accuracy: 0.7281
----------------------------------
```



```
Catboost
Model performance for Training set
- Accuracy: 0.9463
Model performance for Test set
- Accuracy: 0.8696
----------------------------------
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing was 0.000230 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 300
[LightGBM] [Info] Number of data points in the train set: 1070, number of used features: 8
[LightGBM] [Info] Start training from score 13346.089866
LGBM
Model performance for Training set
- Accuracy: 0.9409
Model performance for Test set
- Accuracy: 0.8619
----------------------------------
XGBoost
Model performance for Training set
- Accuracy: 0.9941
Model performance for Test set
- Accuracy: 0.8117
----------------------------------
AdaBoost
Model performance for Training set
- Accuracy: 0.8020
Model performance for Test set
- Accuracy: 0.8015
```

Based on all above models, cat boost model giving the best r2 score for both on training dataset i.e. 0.94 and testing dataset i.e. 0.86

8. CONCLUSION:

Insurance premium is a subjected to each insurance firm based on each individual age, BMI, habits (smoking) based on power of Analytics, We have started the journey with 0.86 accuracy to its peak with catboost regression. Yet there are many other influencing factors involved, with the given data, we have built our model.