A Report On

Medical Insurance Premium Prediction

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE

IN THE PARTIAL FULFILLMENT FOR THE AWARD OF THE DEGREE

OF

BACHELOR OF ENGINEERING

IN

COMPUTER ENGINEERING

By

MISS. PRADNYA RAJESH KENJALE

MISS. SNEHAL MAHADEV BHOSALE

MR. ANIKET SURESH KUMBHAR

MR. GAURAV BALKRUSHNA KADAM

UNDER THE GUIDANCE OF

PROF. S. S. Nimbalkar

DEPARTMENT OF COMPUTER ENGINEERING SVPM'S COLLEGE OF ENGINEERING.

MALEGAON(BK)

A/P-MALEGAON(BK)-413115, TAL. BARAMATI, DIST. PUNE (MS) INDIA

2021-2022

Index

Sr.No.	Title	Page No.
1	Introduction	3
2	Problem Statement	4
3	Technology Used	5
4	Output	6
5	References	9
6	Conclusion	10

1. Introduction

Insurance is a policy that eliminates or decreases loss costs occurred by various risks. Various factors influence the cost of insurance. These considerations contribute to the insurance policy formulation. Machine learning (ML) for the insurance industry sector can make the wording of insurance policies more efficient. In this project we demonstrates how Random forest algorithm gives better accuracy than linear regression model.

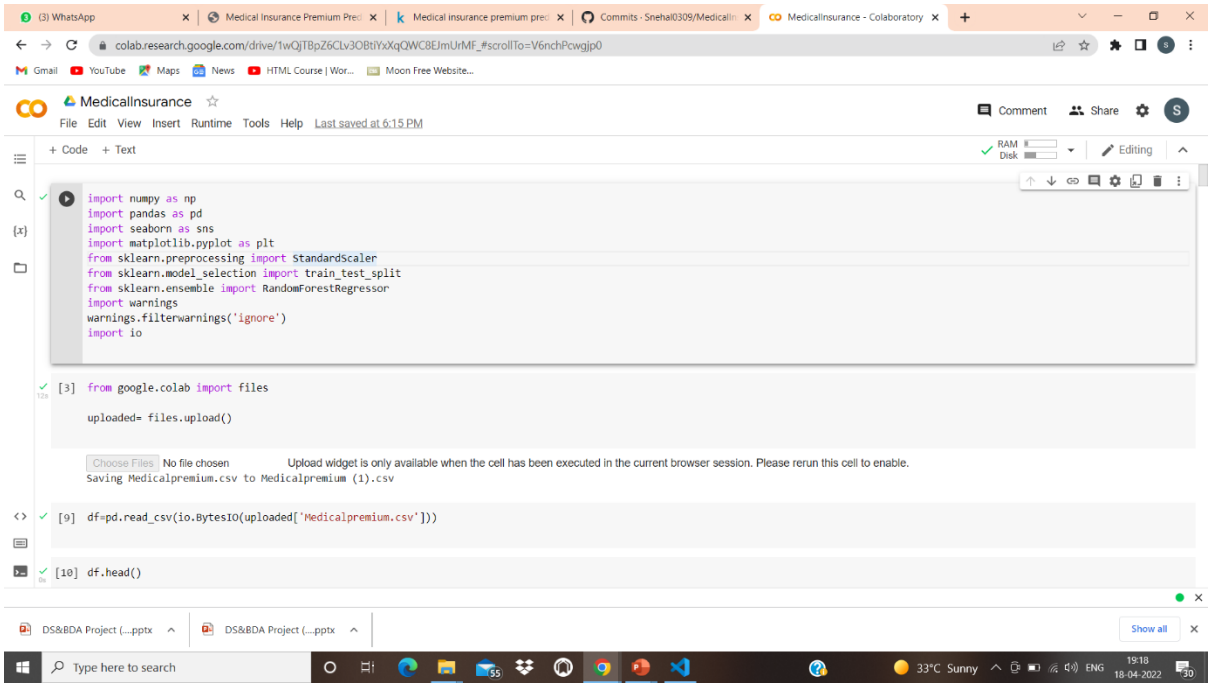
2.Problem Statement

Insurance is a policy that eliminates or decreases loss costs occurred by various risks. Various factors influence the cost of insurance. So, to forecast premium cost using Machine learning algorithm becomes a better way.

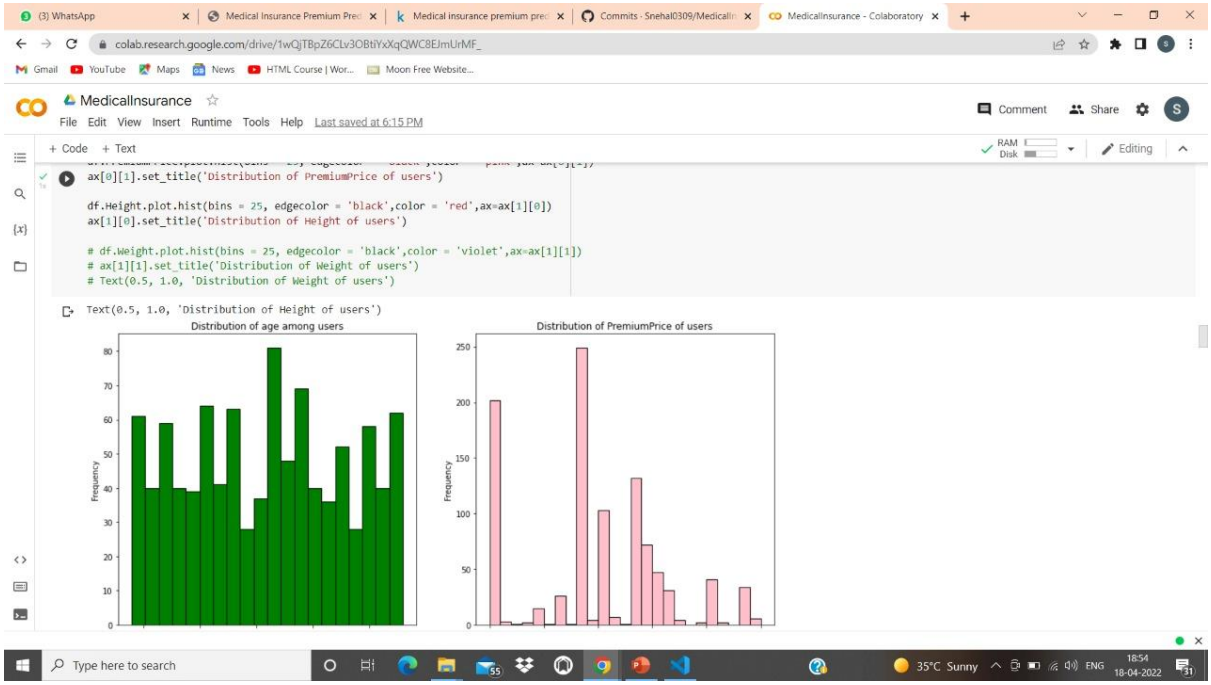
3.Technology

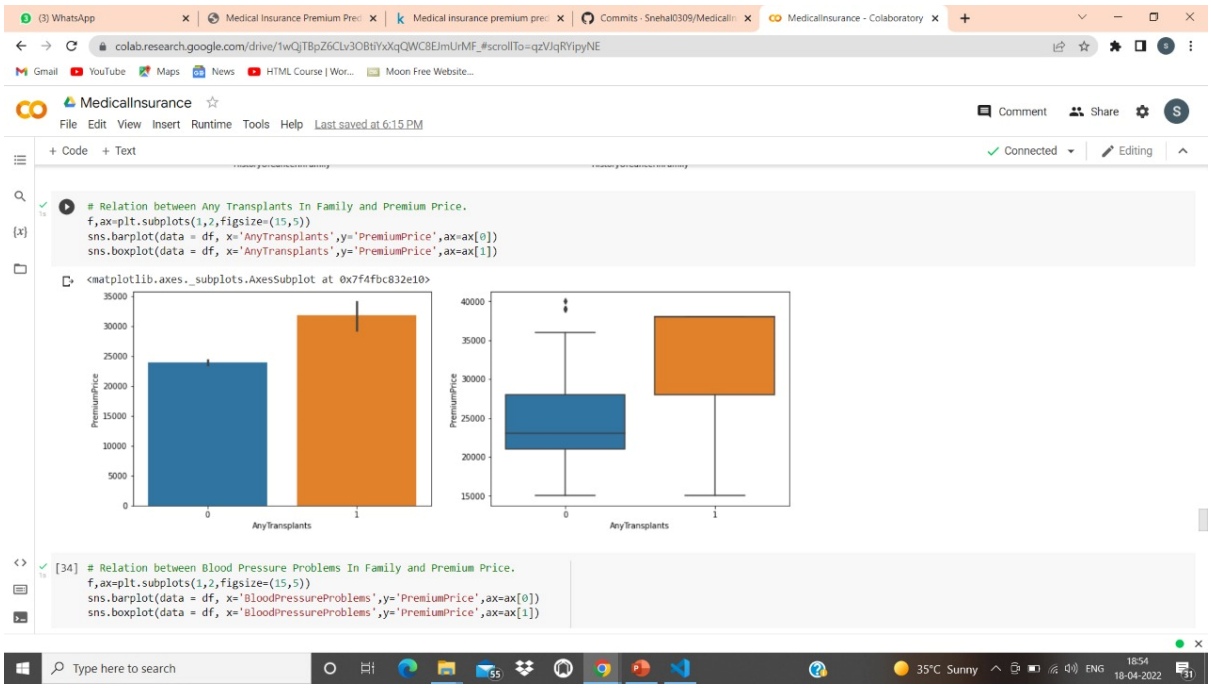
1. Python
2. Algorithm (Random Forest Classifier)

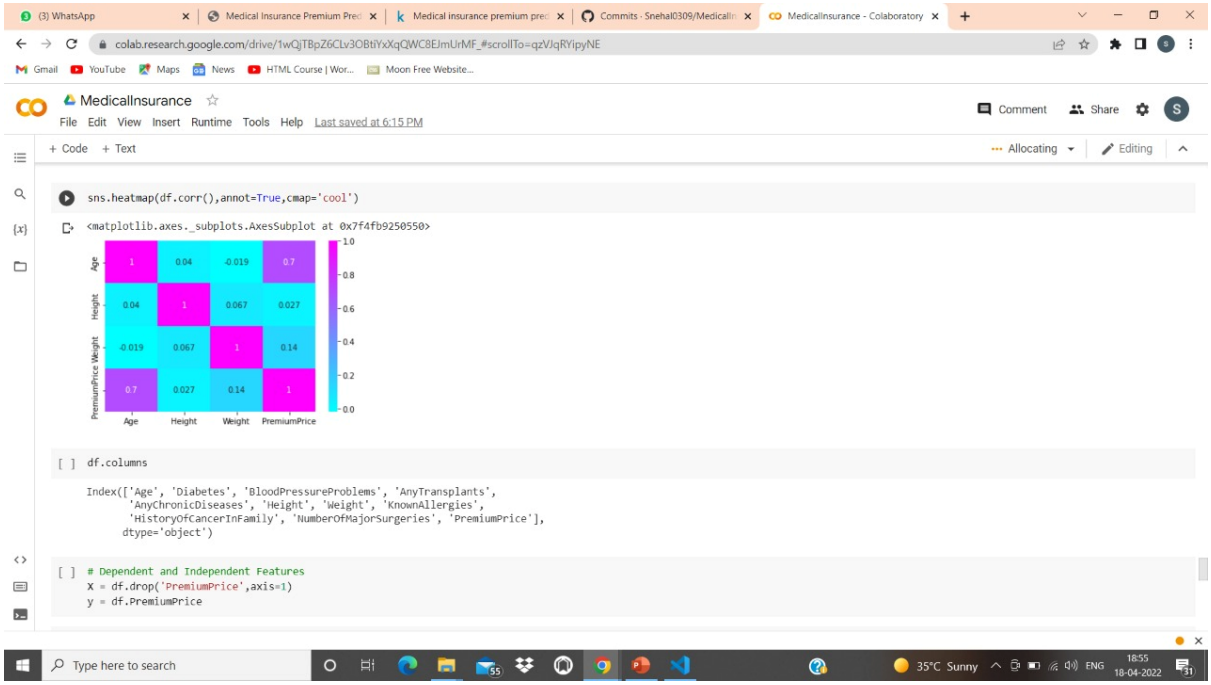
4.Output

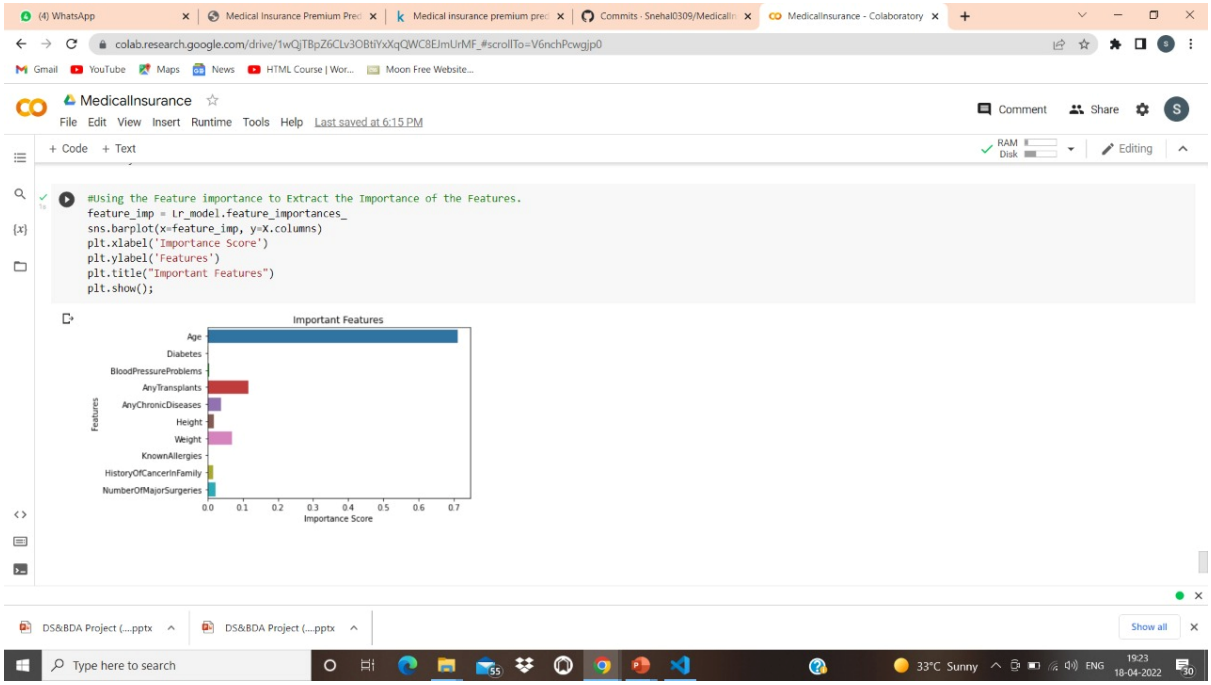

MedicalInsurance
File Edit View Insert Runtime Tools Help Last saved at 6:15 PM
Comment Share
+ Code + Text
RAM Disk Editing
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
import warnings
warnings.filterwarnings('ignore')
import io
[3] from google.colab import files
uploaded= files.upload()
Choose Files No file chosen
Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving Medicalpremium.csv to Medicalpremium (1).csv
[9] df=pd.read_csv(io.BytesIO(uploaded['Medicalpremium.csv']))
[10] df.head()
Type here to search
33°C Sunny
19:18
18-04-2022


MedicalInsurance
File Edit View Insert Runtime Tools Help Last saved at 6:15 PM
Comment Share
+ Code + Text
RAM Disk Editing
Height
Weight
Count
[26] f,ax=plt.subplots(2,2,figsize=(15,15))
df.Age.plot.hist(bins = 20, edgecolor = 'black',color = 'green',ax=ax[0][0])
ax[0][0].set_title('Distribution of age among users')
Type here to search
35°C Sunny
18:53
18-04-2022


MedicalInsurance
File Edit View Insert Runtime Tools Help Last saved at 6:15 PM
Comment Share
+ Code + Text
Editing
ax[0][1].set_title('Distribution of PremiumPrice of users')
df.Height.plot.hist(bins = 25, edgecolor = 'black',color = 'red',ax=ax[1][0])
ax[1][0].set_title('Distribution of Height of users')
# df.Weight.plot.hist(bins = 25, edgecolor = 'black',color = 'violet',ax=ax[1][1])
# ax[1][1].set_title('Distribution of Weight of users')
# Text(0.5, 1.0, 'Distribution of Weight of users')
Text(0.5, 1.0, 'Distribution of Height of users')
Distribution of age among users
Frequency
Distribution of PremiumPrice of users
Frequency
Type here to search
35°C Sunny
ENG
18:54
18-04-2022


MedicalInsurance
File Edit View Insert Runtime Tools Help Last saved at 6:15 PM
Comment Share
+ Code + Text
Connected
Editing
Relation between Any Transplants In Family and Premium Price.
f,ax=plt.subplots(1,2,figsize=(15,5))
sns.barplot(data = df, x='AnyTransplants',y='PremiumPrice',ax=ax[0])
sns.boxplot(data = df, x='AnyTransplants',y='PremiumPrice',ax=ax[1])
<matplotlib.axes._subplots.AxesSubplot at 0x7f4fbc832e10>
PremiumPrice
AnyTransplants
[34] # Relation between Blood Pressure Problems In Family and Premium Price.
f,ax=plt.subplots(1,2,figsize=(15,5))
sns.barplot(data = df, x='BloodPressureProblems',y='PremiumPrice',ax=ax[0])
sns.boxplot(data = df, x='BloodPressureProblems',y='PremiumPrice',ax=ax[1])
Type here to search
35°C Sunny
ENG
18:54
18-04-2022


MedicalInsurance
File Edit View Insert Runtime Tools Help Last saved at 6:15 PM
+ Code + Text
Comment Share
Allocating Editing
sns.heatmap(df.corr(),annot=True,cmap='cool')
<matplotlib.axes._subplots.AxesSubplot at 0x7f4fb9250550>
Age Height Weight PremiumPrice
1 0.04 -0.019 0.7
0.04 1 0.067 0.027
-0.019 0.067 1 0.14
0.7 0.027 0.14 1
1.0 0.8 0.6 0.4 0.2 0.0
df.columns
Index(['Age', 'Diabetes', 'BloodPressureProblems', 'AnyTransplants',
'AnyChronicDiseases', 'Height', 'Weight', 'KnownAllergies',
'HistoryOfCancerInFamily', 'NumberOfMajorSurgeries', 'PremiumPrice'],
dtype='object')
# Dependent and Independent Features
X = df.drop('PremiumPrice',axis=1)
y = df.PremiumPrice
Type here to search
35°C Sunny
ENG
18:55
18-04-2022


MedicalInsurance
File Edit View Insert Runtime Tools Help Last saved at 6:15 PM
+ Code + Text
Comment Share
RAM Disk Editing
#Using the Feature importance to Extract the Importance of the Features.
feature_imp = lr_model.feature_importances_
sns.barplot(x=feature_imp, y=X.columns)
plt.xlabel('Importance Score')
plt.ylabel('Features')
plt.title("Important Features")
plt.show();
Important Features
Age
Diabetes
BloodPressureProblems
AnyTransplants
AnyChronicDiseases
Height
Weight
KnownAllergies
HistoryOfCancerInFamily
NumberOfMajorSurgeries
Features
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
Importance Score
DS&BDA Project (....pptx
DS&BDA Project (....pptx
Show all
Type here to search
33°C Sunny
ENG
19:23
18-04-2022

5.References

1. https://www.kaggle.com/datasets/tejashvi14/medical-insurance-premium-prediction

2. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/#:~:text=Random%20forest%20is%20a%20Supervised,average%20in%20case%20of%20regression.

6.Conclusion

In this project, we have explored the random forest classifier model and applied it to predict premium cost and seen the correlation between predicted and actual results. Random forest classifier algorithm gives more accuracy.