FINAL PROJECT

IST – 687

INTRODUCTION TO DATA SCIENCE

FALL 2022

HEALTH MANAGEMENT ORGANIZATION EXPENSES PREDICTION

PREPARED BY : AYINKYA GADDIME
HENDI KUSHTA
JIANCHEN LI
VATSAL PATEL

Table of Contents

Abstract....3
1. Introduction....4
1.1 A Health Maintenance Organization (HMO) is what, exactly?....4
1.2 The Functions of a Health Maintenance Organization (HMO)....4
1.3 Advantages and Disadvantages of HMOs....4
1.4.1 Advantages of HMOs....5
1.4.2 Disadvantages of HMOs....5
1.5 Goal....5
1.6 Data....6
2. Data Preprocessing and Exploration....6
2.1 Children....6
2.2 Married....7
2.3 Education level....8
2.4 Hypertension....9
2.5 Gender....10
2.6 BMI....11
2.7 Smoker....12
2.8 Location....13
2.9 Location type....14
2.10 Yearly physical....14
2.11 Bivariate Analysis....15
3. Data Pre-processing....16
4. Model Building....17
4.1 Logistic Regression....17
4.2 Decision Tree....18
4.3 Support Vector Machine....19
5. Conclusions and Recommendations....21
Appendixes....21

Abstract

Everybody's life is centered on their health. A healthy body is necessary for every part of our existence. The ability of a person to adapt to their physical, emotional, mental, and social environments is referred to as health. Our lives are moving so quickly that we are forming many bad habits that are bad for our health. Spending a lot of money on physical activity or routine checkups helps one stay healthy, prevent becoming out of shape, and treat illnesses. When we get sick, we frequently overspend, which results in high medical costs.

So, if someone has such factors, an application that can recognize and estimate medical expenses can be created to help them understand the factors that are making them unfit and leading to high medical costs.

We did a data exploration for all the dataset attributes. Based on the researches and results we derived from our tests, we chose to continue with only 8 variables from 13 in total. After data exploration, we did data cleaning. We removed null values, removed cost column from testing data, and created another column named expensive. We divided the dataset into training and testing data. 70% training data and 30% testing. After that we used regression model, decision tree and support vector machine algorithms to create our model. Since we are required to predict a boolean output, we only compared decision tree and support vector machine. From these 2 algorithms, the best performing one was SVM with 87% while decision tree with 86%. In the end we built a shinny app to predict if health cost for a person will be expensive or not.

1. Introduction

1.1 A Health Maintenance Organization (HMO) is what, exactly?

A person who requires health insurance may locate a number of insurance companies with distinctive features. A health maintenance organization (HMO), a sort of insurance plan that offers coverage through a network of doctors, is one insurance provider that is well-liked on the Health Insurance Marketplace.

HMO and preferred provider organization (PPO) plans have a number of significant variances. With an HMO plan, your primary care physician (PCP) will recommend specialists to you, and in order to maintain coverage, you must remain within a network of providers. HMO plans, however, frequently offer lower premiums than PPO policies.

1.2 The Functions of a Health Maintenance Organization (HMO)

The cost of an HMO's health insurance coverage might be either monthly or yearly. An HMO restricts its members' access to healthcare to services delivered by a network of physicians and other healthcare professionals who have agreements with the HMO.
Since the healthcare providers benefit from having people referred to them, these contracts enable rates to be lower than for typical health insurance. They do, however, impose more limitations on HMO members.

Consider the cost of premiums, out-of-pocket expenses, any potential needs you may have for specialist medical care, and how important it is to you to have your own primary care physician when choosing an HMO plan (PCP).

An HMO is a formalized governmental or private organization that offers its members both essential and supplementary healthcare services. The business forms agreements with PCPs, medical facilities, and specialists to secure its network of healthcare professionals. The medical organizations that sign into agreements with the HMO are compensated with an agreed-upon price in exchange for providing a variety of services to the HMO's members. An HMO can offer lower rates than other health insurance plans while still maintaining high standards of care from its network thanks to the negotiated payment.

The Health Maintenance Organization Act of 1973 created the HMO as we know it today. By defining HMOs as "a public or private body structured to offer basic and additional health services to its members," the bill signed by then-President Richard Nixon clarified the term's meaning.
The law also mandates that plans offer basic healthcare to those who are insured in return for regular, set premiums that are determined "using a community rating system."

1.3 Advantages and Disadvantages of HMOs

Before selecting a plan, it's crucial to examine the benefits and drawbacks of HMOs, just as you would with any other alternative. Below, you'll find a list of some of the program's most typical benefits and drawbacks.

1.4.1 Advantages of HMOs

The first and most evident benefit of joining an HMO is the affordable price. Fixed rates that are less expensive than those for conventional health insurance will be paid on a monthly or annual basis. Your co-pays are typically lower than those of other plans, and these policies frequently have low or no deductibles. Your prescription's out-of-pocket expenses will also be reduced.
Additionally, billing is typically less difficult.
You will almost certainly have to deal with the insurer directly as well. That's because you get to pick your PCP, who is in charge of overseeing your care and treatment. Additionally, this expert will promote services on your behalf.
This includes providing you with recommendations for specialized services.
An HMO plan typically offers higher-quality healthcare. This is due to the fact that patients are urged to get yearly checkups and seek treatment as soon as possible.

1.4.2 Disadvantages of HMOs

You have limitations on how you can utilize the plan if you pay for an HMO. You must choose a physician to handle all of your medical requirements, including primary care and referrals. But the network must include this doctor.
This means that even if there are no contracted doctors in your area, you are still liable for any expenses paid if you see a doctor who is not in the network.
If you want your HMO to cover any visits, you'll need references for any specialists. For instance, before you may see a hematologist or dermatologist for the plan to pay for your visit, your PCP must provide a reference. If not, the entire expense is your responsibility.
For some medical claims, such as emergencies, you must fulfill very precise requirements. For instance, the definition of an emergency is typically fairly specific. Your ailment won't be covered by the HMO plan if it doesn't meet the requirements.

1.5 Goal

Our team’s goal was to understand the key drivers for why some people are more expensive (i.e., require more health care), as well as predict which people will be expensive (in terms of health care costs).

Hence, at a high level, you have two goals:

Predict people who will spend a lot of money on health care next year (i.e., which people will have high healthcare costs).
Provide actionable insight to the HMO, in terms of how to lower their total health care costs, by providing a specific recommendation on how to lower health care costs.

1.6 Data

Here are the variables found in data file:

• X: Integer, Unique identified for each person
• age: Integer, The age of the person (at the end of the year).
• location: Categorical, the name of the state (in the United States) where the person lived (at the end of the year)
• location_type: Categorical, a description of the environment where the person lived (urban or country).
• exercise: Categorical, “Not-Active” if the person did not exercise regularly during the year, “Active” if the person did exercise regularly during the year.
• smoker: Categorical, “yes” if the person smoked during the past year, “no” if the person didn’t smoke during the year.
• bmi: Integer, the body mass index of the person. The body mass index (BMI) is a measure that uses your height and weight to work out if your weight is healthy.
• yearly_physical: Categorical, “yes” if the person had a well visit (yearly physical) with their doctor during the year. “no” if the person did not have a well visit with their doctor.
• Hypertension: “0” if the person did not have hypertension.
• gender: Categorical, the gender of the person
• education_level: Categorical, the amount of college education ("No College Degree", "Bachelor", "Master", "PhD")
• married: Categorical, describing if the person is “Married” or “Not_Married”
• children: Integer, Number of children
• cost: Integer, the total cost of health care for that person, during the past year.
Data: The data file is located at: https://intro-datascience.s3.us-east-2.amazonaws.com/HMO_data.csv.

2. Data Preprocessing and Exploration

2.1 Children

Number of children is an important variable to be explored in our project. Aged 45 to 54, those who live with children are healthier than those who do not. Parents between the ages of 18 and 24 smoke more often, engage in less physical activity, and are more likely to be overweight and experience back discomfort than women without children. Living with children and age interact significantly in multi-variable analysis, which has implications for all outcomes (except depression and back pain in men). Being a parent is a significant social factor of health that is strongly influenced by age.

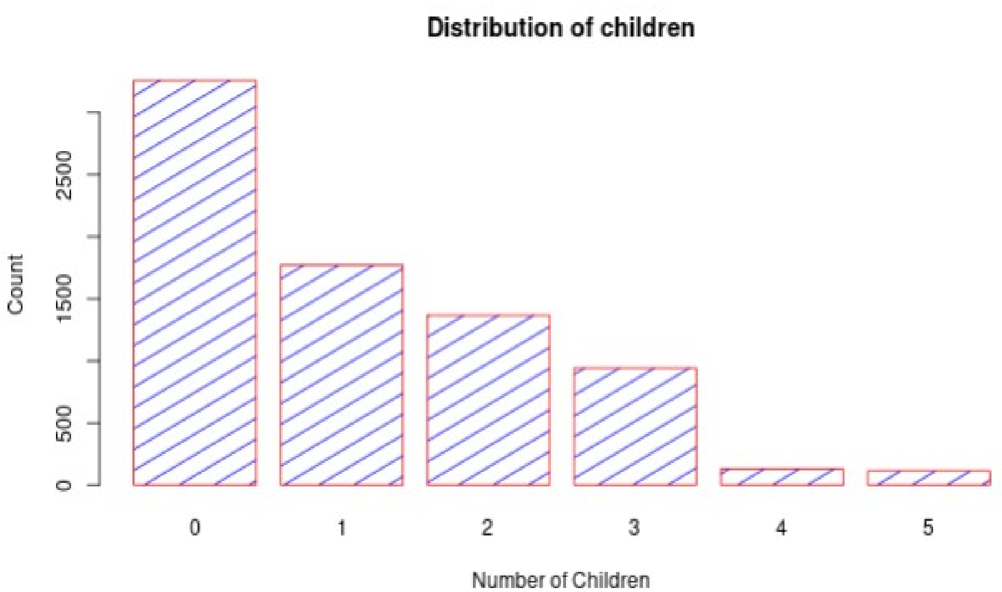
By typing a simple command in R, we can find the the distinct values in children column and the frequency of each number in our data set.

```
table(HMO_df$children)
```

```
0    1    2    3   4   5
3259 1772 1367  942  130  112
```

Number of Subjects having no children - 3259
Number of Subjects having one child – 1772
Number of Subjects having two children – 1367
Number of Subjects having three children – 942
Number of Subjects having four children – 130
Number of Subjects having five children – 112

Our next step is to build a bar plot with children distributions



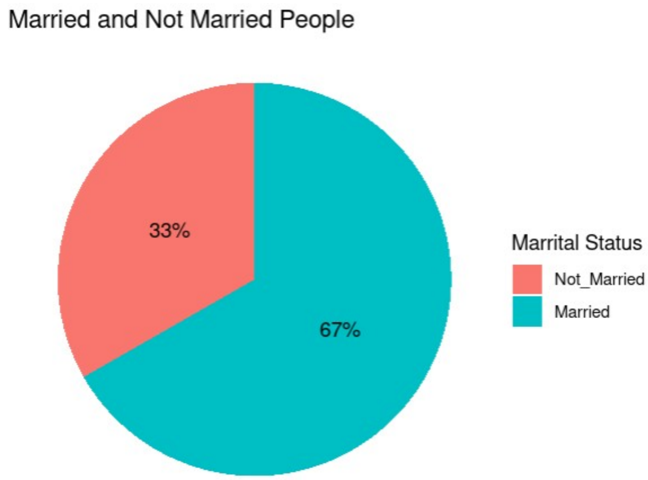
2.2 Married

Married is a categorical variable, which according to our researches is important to take into considerations.
There were different associations between marital status and health outcomes. Never getting married was a major risk factor for developing hypertension and tended to be a major risk factor for male mortality. Being a widowed woman, however, was linked to a decreased incidence of T2D.

```
table(HMO_df$married)
```

```
Married Not_Married
  5060      2522
```

Number of Subjects being married - 5060
Number of Subjects not being married - 2522



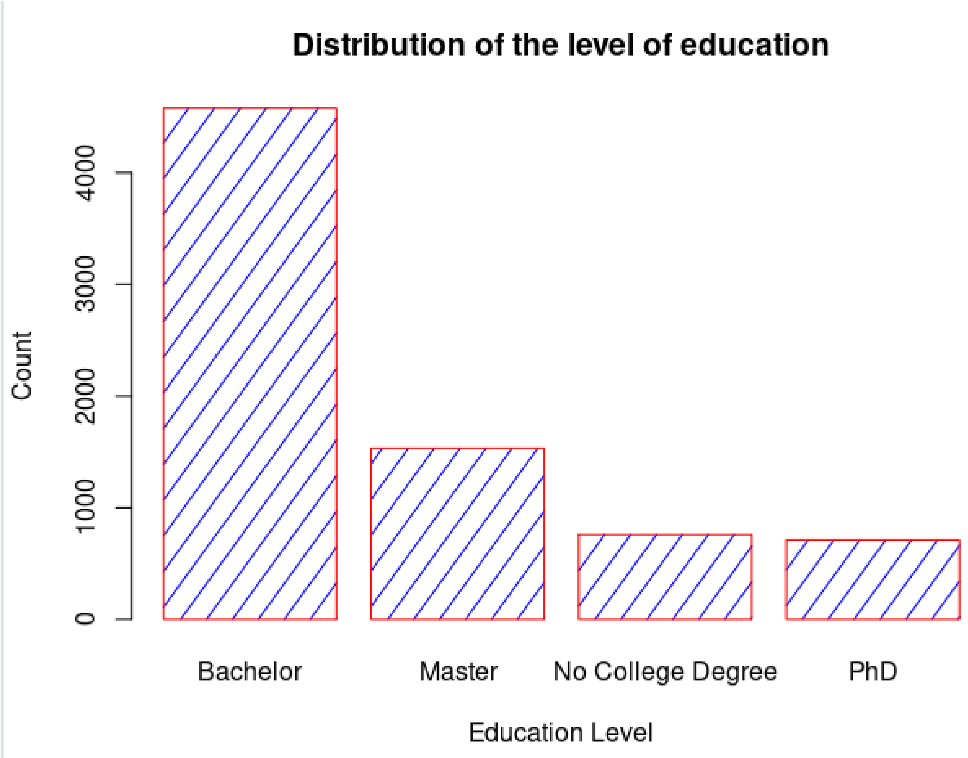
2.3 Education level

Compared to their colleagues who have less education, adults with higher educational attainment tend to live longer and in better health. We emphasize that higher education plays a significant role in determining enrollment rates, life expectancy, infant immunization rates, and infant death. As an additional indicator of health quality, an economy must take premature death into account.
There are 4 categories in education_level variable "No College Degree", "Bachelor", "Master", "PhD".

```
table(HMO_df$education_level)
```

Bachelor	Master	No College Degree	PhD
4578	1533	759	712

Number of Subjects with No College Degree - 4578
Number of Subjects with Bachelor – 1533
Number of Subjects with Master - 759
Number of Subjects with PhD – 712



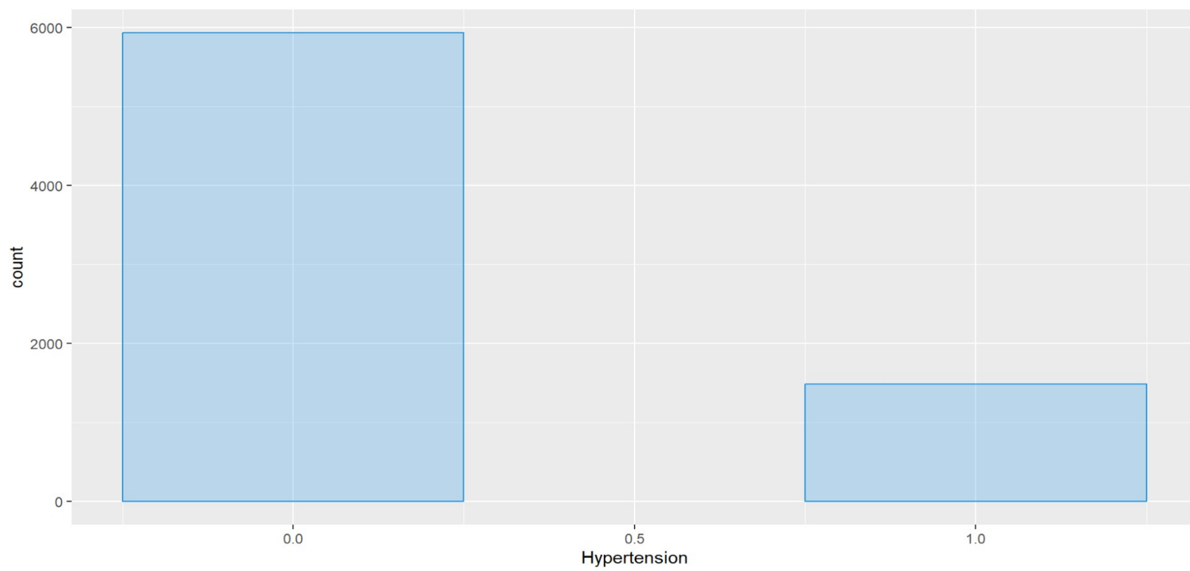
2.4 Hypertension

Hypertension is a world wide disorder that contributes to morbidity, mortality and health cost. There is a prevalence of hypertension among huge amount of adults around the world. In our project we do not consider hypertension is the key variable because it only contains two parameters. This bar chart shows that the Hypertension as the selected variable, as we can see from this chart, the 0.0 means the individual in this dataset has no hypertension, and 1 represents the person who has hypertension, and there are some NA values in this column, our team decided to delete all the NA values since we have more than 7000 rows, these NA values wont be a problem for accuracy.

```
table(df$hypertension)

   0    1
5998 1504
```

Number of has hypertension: 1504
Number of no hypertension: 5998

Next step is building a bar plot of hypertension:



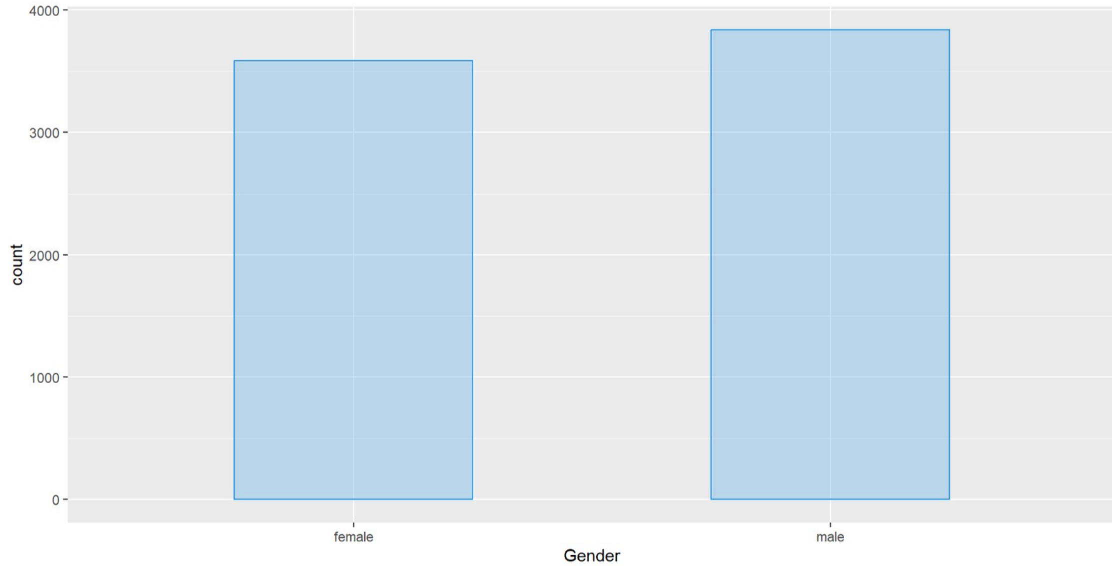
2.5 Gender

In our project the column of gender is a categorical variable which has significant relation to health cost, like disparities between male and female who living in the low income community in access to health care. We can tell that the number of female and male are generally equal to each other, though male is slightly more than the female.

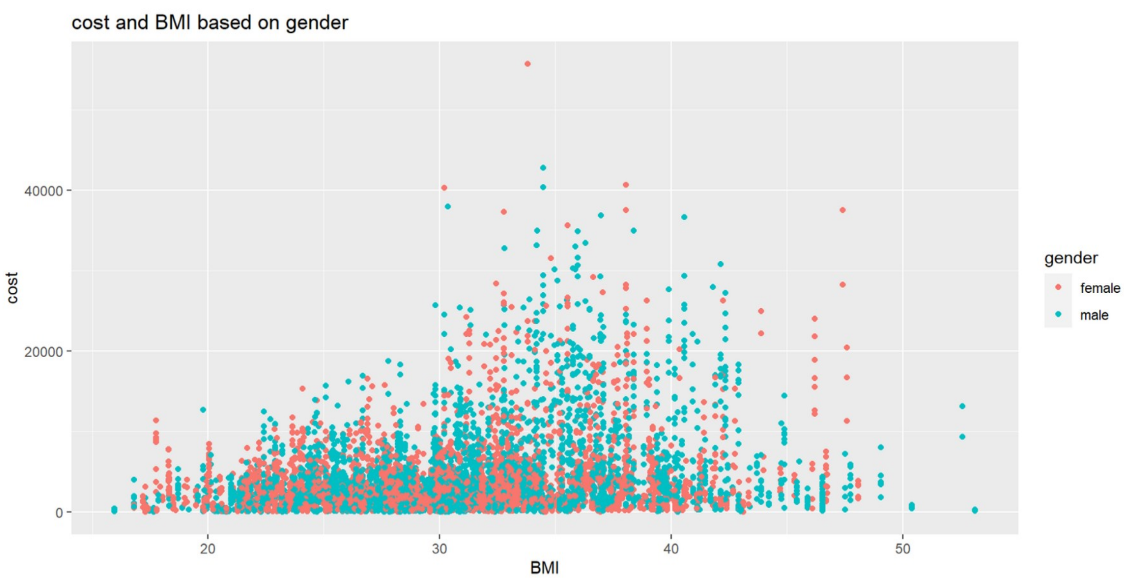
table(df$gender)

female   male
  3662   3920
 The number of male: 3920
The number of female: 3662

Bar plot of gender:



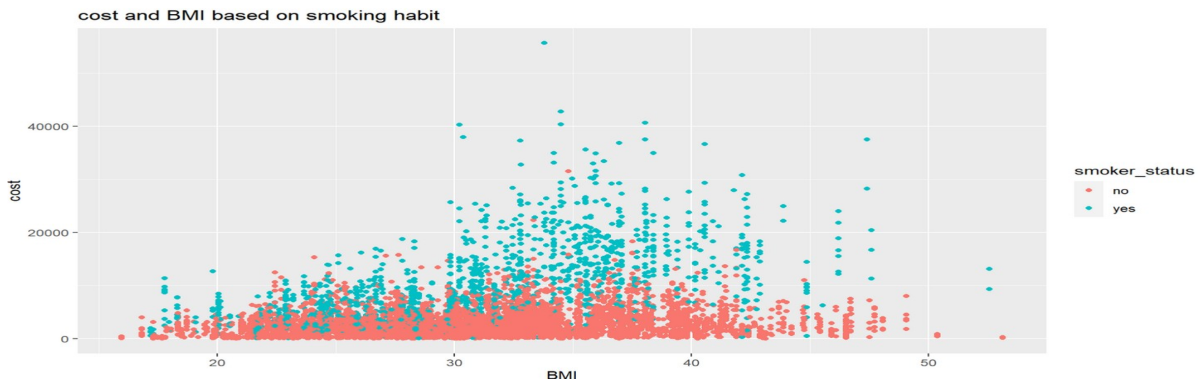
Analysis of gender:



As we can see from the scatter plot of relation between gender and cost, as the BMI rises, there are more males than females who has the higher cost.

2.6 BMI

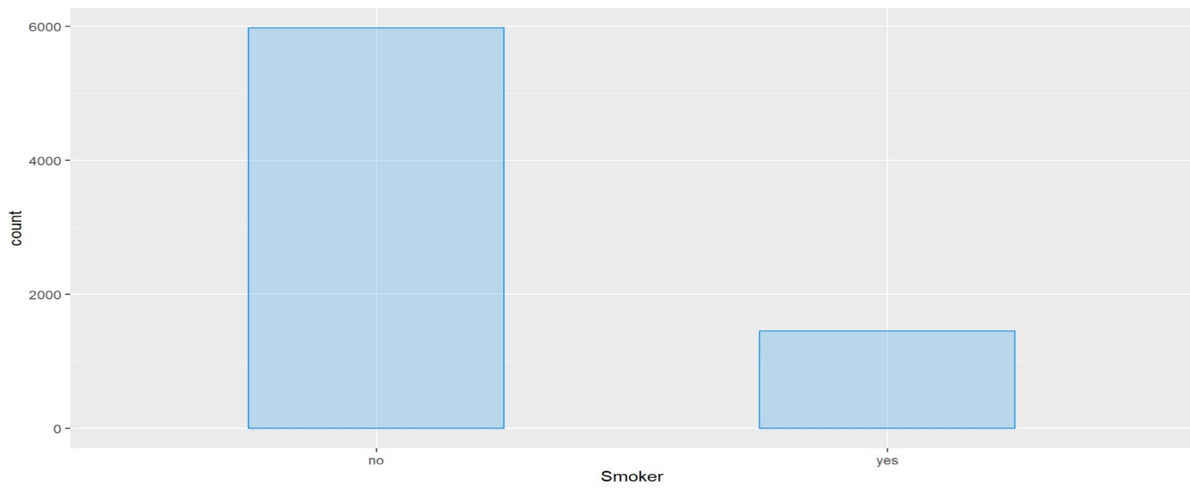
BMI is an assessment of body fat and an excellent predictor of your risk for diseases associated with excess body fat. The greater your BMI, the more likely you are to have certain diseases such as heart disease, high blood pressure, type 2 diabetes, gallstones, breathing difficulties, and some malignancies.

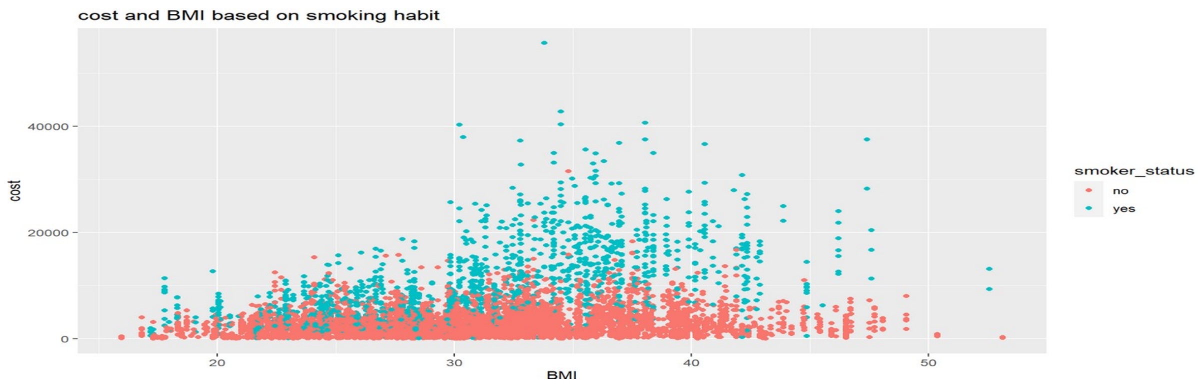


This scatter plot shows the mean distribution of cost of smokers is higher than non-smokers, and the cost of BMI higher than 30 is also higher (hint: technically BMI over 30 is obesity)

2.7 Smoker

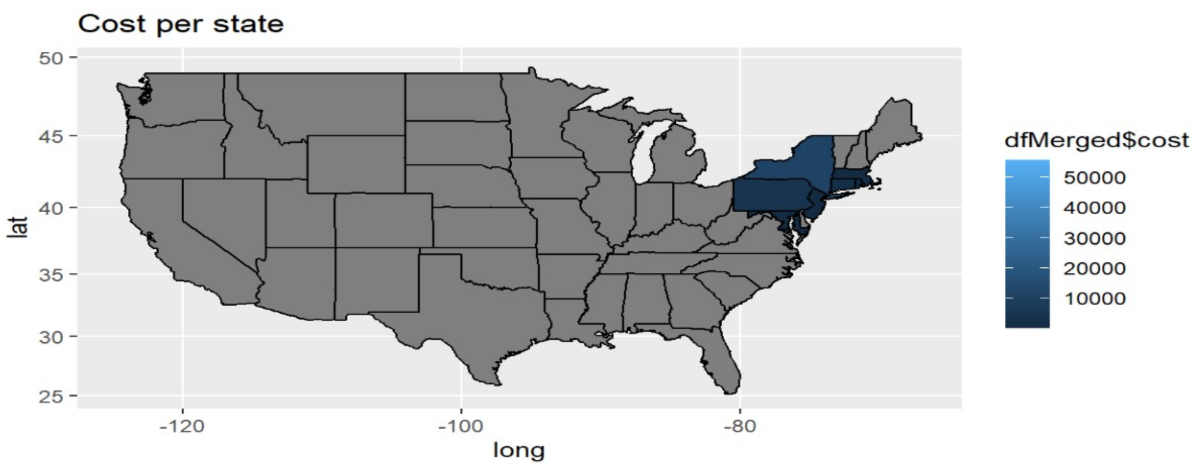
Cigarette smoking kills around 480,000 people in the United States each year. This represents approximately one in every five deaths. Tobacco smoking is linked to cancer, heart disease, stroke, lung disease, diabetes, and chronic obstructive pulmonary disease (COPD), which includes emphysema and chronic bronchitis. Smoking also raises the risk of TB, some eye illnesses, and immune system problems such as rheumatoid arthritis.
Due to these factors, smoking was a critical variable in our model. As shown in the graph in our database, the number of non-smokers outnumbered the number of smokers. Also, if you smoke, your health insurance will cost more than if you do not.

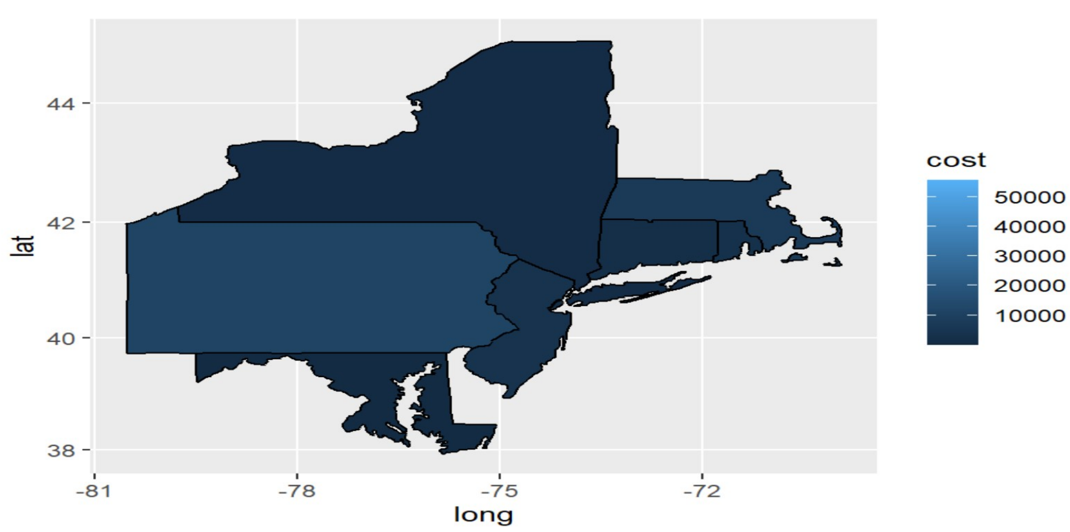




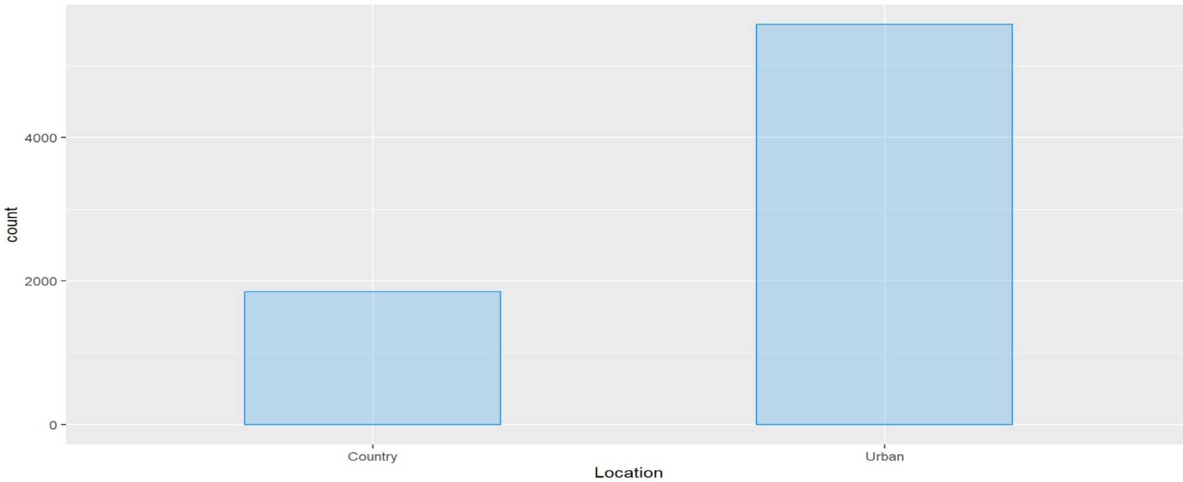
2.8 Location

We plotted our dataset and what we saw was that most of the data provided were in the northeastern part of the USA. As we see, that is why we have plotted twice the map, in the second map it is zoomed. What we can see, New York is the state with the highest cost of living.





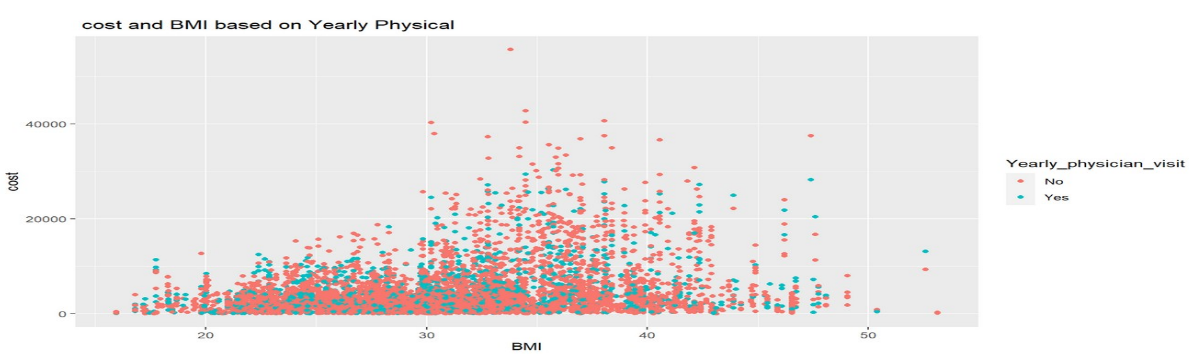
2.9 Location type



The database that was provided to us includes information from cities rather than rural areas. Furthermore, we believed that the location variable was more relevant than the location type variable because knowing the location already tells you whether it is urban or rural. Furthermore, we believed that other criteria, such as geography, had a greater impact on the cost of your insurance than location, but this is not always the case.

2.10 Yearly physical

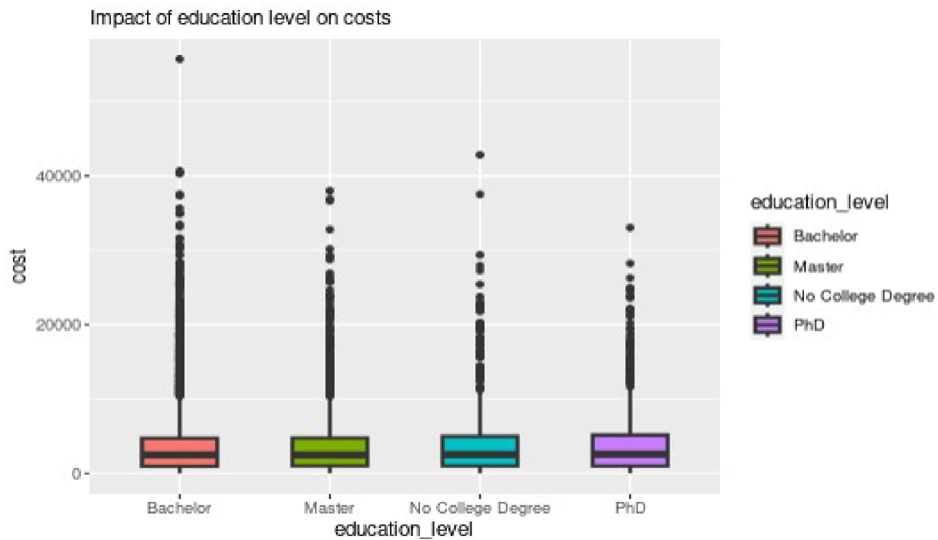
We thought that yearly physical visit was a very important variable and as you can see in the graph below, it has a direct relationship with the cost. People who visit a physician at least once a year and also have a stable bmi will have a less cost than people who don't.
The doctor (or other health expert) may do tests to ensure that the heart, lungs, and other body systems are functioning appropriately. The doctor will most likely ask you questions about your daily routine, medical history, memory, and take routine measurements such as height, weight, and blood pressure.



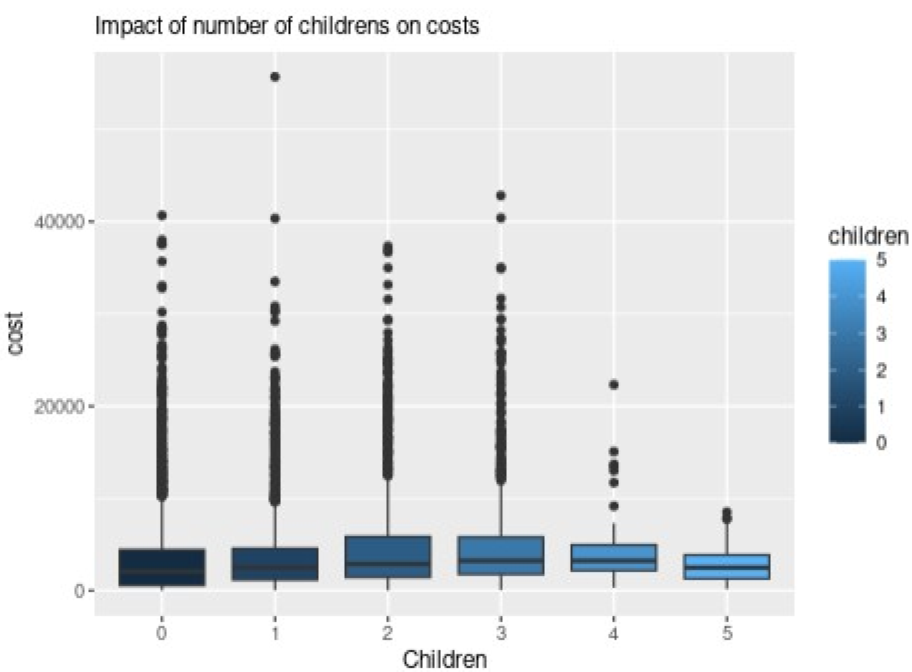
As we can see from the chart, the mean value of the distribution of 'no' status is higher than the 'yes'.

2.11 Bivariate Analysis

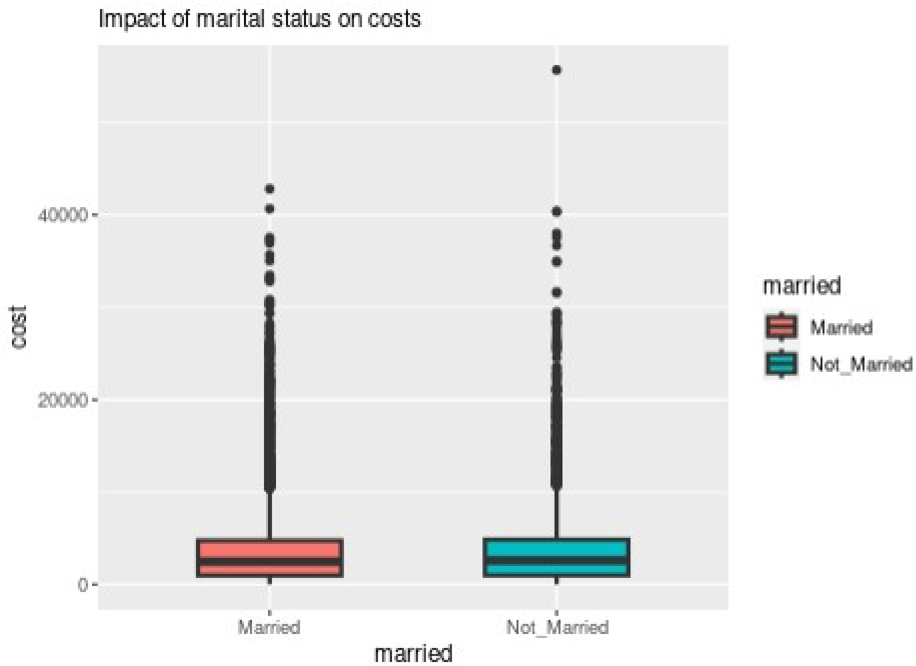
We plotted the bar plot between the education level and costs and it has been noticed people with bachelor education level have the highest costs followed by those that have masters. What we have read, people without any education level, are supported by the government mostly, this is what also is appeared in the box plots. People with Phds, have the least costs.



We compared costs per child to the number of children, and it was found that costs are higher for families with more kids because parents must care for the health of every child, as opposed to families with just one or no kids. However, because families with more than three kids are extremely rare, we set a cap on their costs. Among them, those with three children had the highest expenses.



From the box plot shown below, we see that median cost for both categories married and not married are almost the same. But in some points, being married, seems to have higher costs.



3. Data Pre-processing

Categorical Columns Encoding

There are many categorical variables in our dataset.
children, smoker, location_type, education_level, yearly_physical, # exercise, married, hypertension, gender. We have encoded them mostly with numbers. For example gender instead of male and female we have used 0 and 1 and so on for the other variables.

Splitting Data

We divided the entire dataset into Train and Test with an 70:30 split. We first used the model to the test dataset to evaluate its performance after training it with trained data.

Scaling of Feature

A technique for normalizing the variety of independent variables or features in data is called feature scaling. In essence, feature scaling aids in normalizing the data within the same scale. In order to apply feature scaling to the testing data, we utilized the transform function instead of the fit transform function since the fit transform function first learns all the patterns in the data before applying feature scaling to it. However, since it is impossible to identify patterns from test data, we are unable to apply the fit transform method.

4. Model Building

Here, we'll go over how to divide a dataset into a Train and a Test set using the R programming language. Machine learning algorithms that are applicable for prediction-based algorithms and applications are evaluated using the train-test split. We can compare the output of our own machine learning model to that of other machines using this quick and simple process. By default, the training set contains 70% of the real data, whereas the test set contains 30% of the real data.

To assess how well our machine learning model works, we must divide a dataset into train and test sets. The statistics of the train set are known and are utilized to fit the model.

The test data set, which is the second set, is utilized just for predictions.
We have chosen to compare 3 different machine learning algorithms.

4.1 Logistic Regression

When a dependent variable is dichotomous, the proper regression analysis to use is logistic regression (binary). The logistic regression is a predictive analysis, just like all regression analyses. To describe data and explain the relationship between one dependent binary variable and one or more independent nominal, ordinal, interval, or ratio-level variables, we employ logistic regression.

The Statistics application makes it simple to run the analysis and then interprets the results in plain English. Logistic regressions might be challenging to interpret at times.

After running our regression model we got the results like this.

```
Call:
lm(formula = cost ~ age + children + bmi + smoker + exercise +
    Location_type, data = S1)

Residuals:
   Min     1Q Median     3Q    Max
-12294  -1484   -366   1013  42026

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -9006.381    236.253  -38.12  < 2e-16 ***
age                  102.015      2.666   38.26  < 2e-16 ***
children             237.543     30.930    7.68 1.79e-14 ***
bmi                  182.595      6.289   29.04  < 2e-16 ***
smokeryes           7700.292     94.746   81.27  < 2e-16 ***
exerciseNot-Active  2274.883     86.807   26.21  < 2e-16 ***
Location_typeUrban   -13.886     86.729   -0.16    0.873
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3234 on 7417 degrees of freedom
Multiple R-squared:  0.5729,    Adjusted R-squared:  0.5725
F-statistic:  1658 on 6 and 7417 DF,  p-value: < 2.2e-16
```

Our model was significant as we also see from our p-value, but the R-adjusted was only 57% which show the correlation of the variables in the model. We tried to change the variables many times, but the results were almost the same.

4.2 Decision Tree

A decision tree is a supervised machine learning tool that may be used to classify or forecast data based on how queries from the past have been answered. The model is supervised learning in nature, which means that it is trained and tested using data sets that contain the required categorization.

The decision tree might not always offer a simple solution or choice. Instead, it might give the data scientist choices so they can choose wisely on their own. Decision trees mimic human thought processes, making it generally simple for data scientists to comprehend and evaluate the findings.

A decision tree looks like a tree, naturally. The root node is where the tree starts. A string of decision nodes that represent decisions to be made flow from the root node. Leaf nodes that indicate the decisions' implications emerge from the decision nodes. The leaf nodes that emanate from a decision node indicate the potential replies, and each decision node acts as a question or split point. Like a leaf sprouting on a tree branch, leaf nodes grow from decision nodes. For this reason, we refer to each decision tree branch as a "branch." Below will find our tree model.

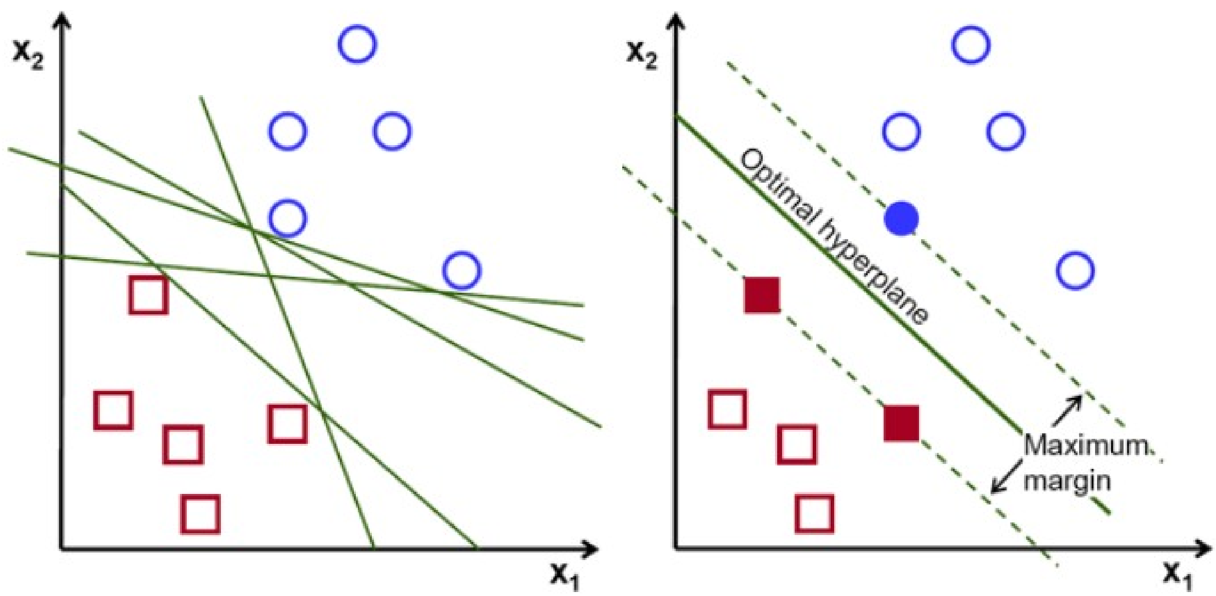
```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE  1632  240
     TRUE     67  286

               Accuracy : 0.862
                 95% CI : (0.847, 0.8761)
    No Information Rate : 0.7636
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5689

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9606
            Specificity : 0.5437
         Pos Pred Value : 0.8718
         Neg Pred Value : 0.8102
             Prevalence : 0.7636
         Detection Rate : 0.7335
   Detection Prevalence : 0.8413
      Balanced Accuracy : 0.7521

       'Positive' Class : FALSE
```

The model performed good, bur as we will also see below, support vector machine performed better in almost all the statistical measures.

4.3 Support Vector Machine

Another straightforward algorithm that every machine learning expert should know how to use is the support vector machine. Many people favor the support vector machine because it offers notable accuracy while using less processing power. SVM, or Support Vector Machine, is a tool that can be used for both classification and regression tasks. However, it is frequently employed in classification goals.

Finding a hyperplane in an N-dimensional space (N is the number of features) that categorizes the data points clearly is the goal of the support vector machine algorithm.



After we built our model, we got the results as below:

```
               Accuracy : 0.8742
                 95% CI : (0.8597, 0.8877)
    No Information Rate : 0.7636
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5939

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9812
            Specificity : 0.5285
         Pos Pred Value : 0.8705
         Neg Pred Value : 0.8968
             Prevalence : 0.7636
         Detection Rate : 0.7492
   Detection Prevalence : 0.8607
      Balanced Accuracy : 0.7548

       'Positive' Class : FALSE
```

In our model we predicted that health costs will be expensive for a person that smokes, does not do physical activities, getting old and is overweight. So if we predict that this persons health care will not be expensive, the cost of the company will be high. When we use only 20 rows from our data, or the test data, there is 86% that our model will predict right, and the sensitivity is almost 96%.
Accuracy although is important, we cannot say that it is good measure to validate our model, because it depends on the data and its balance of classes.
In our model, accuracy is 87% which is mostly by predicting the expensive class.
%95 CI or confidence interval, shows that that are 95% chances that our data will be in the range of 85% to 858. There is only 5% chance that the mean of each sample goes out of our range.
No-Information rate is 76% which shows that only by predicting the majority class, our accuracy drops.
And p-value which shows the correlation of our variables is very low.

5. Conclusions and Recommendations

To conclude, from all 3 models that we chose to compare, support vector machine(SVM) was the best performing one. Compared to decision tree whose accuracy was 86%, SVM accuracy was 87%. SVM sensitivity was 98% while the tree model sensitivity was 96%. The no-information rate was the same and the p-value in both of them showed that there was a high correlation within the variables.

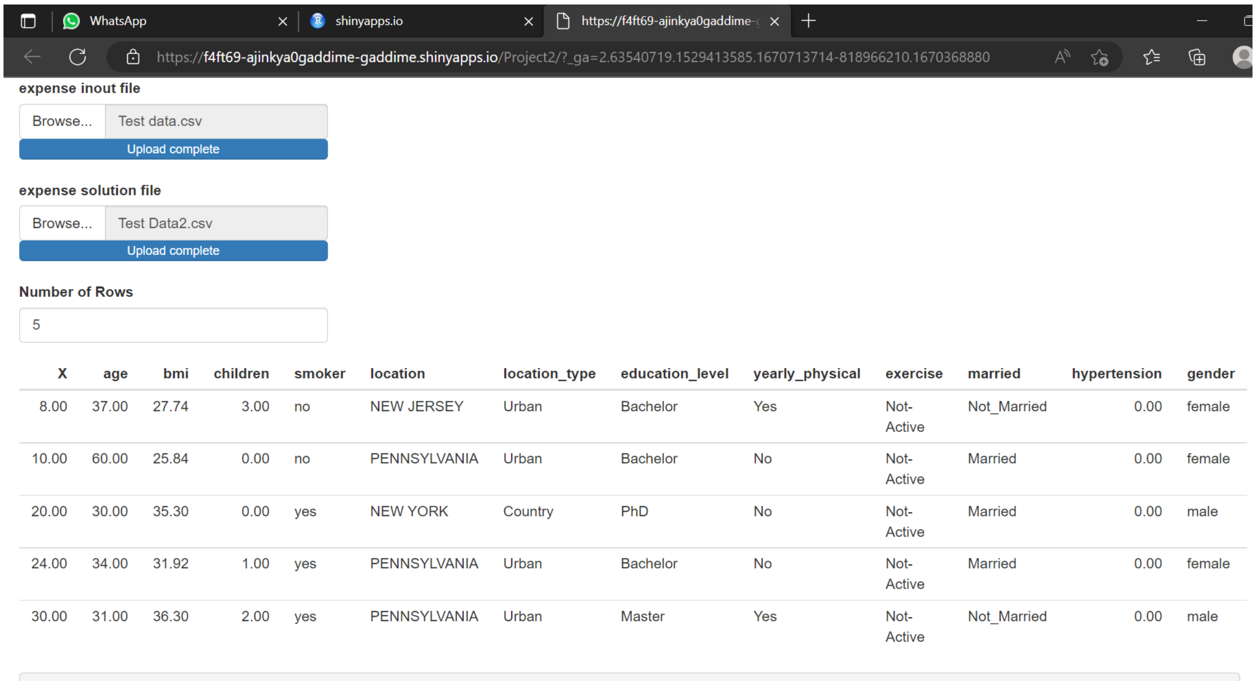
After working on the data, we can conclude that the person who have habit of smoking and not active in exercise tends to spend more than the person who did not smoke and do exercise on regular basis. As a consultant firm we will advise our client that they should motivate their customer to exercise daily and motivate them to quit smoking.

Programs to help you quit smoking may now be covered by your health insurance. Depending on your strategy, that might involve advice on how to stop smoking or using other tobacco products.
Stop-smoking medications varenicline (Chantix) and bupropion (Zyban), as well as nicotine replacement products such gum, lozenges, patches, inhalers, and nasal spray.

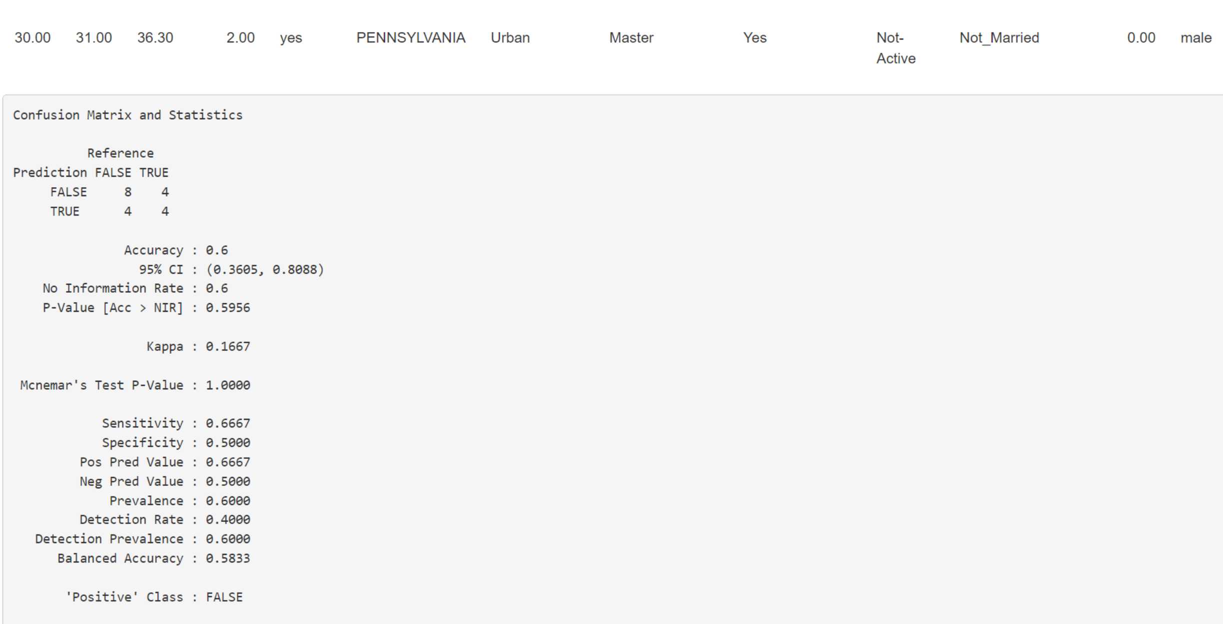
If they choose to participate in the new program extension, policyholders who have chosen Vitality insurance products will receive a loyalty card that will track their grocery purchases at Walmart and other supermarkets. Then, they will receive special offers and cash back on "healthy" purchases, saving them up to $50 each month or $600 per year!
Policyholders can also accrue extra points depending on an item's "healthiness," earning benefits comparable to those offered by credit cards or frequent flyer programs. These benefits might add up to annual premium cuts of up to 15%.

Appendixes

1. Shiny App 1



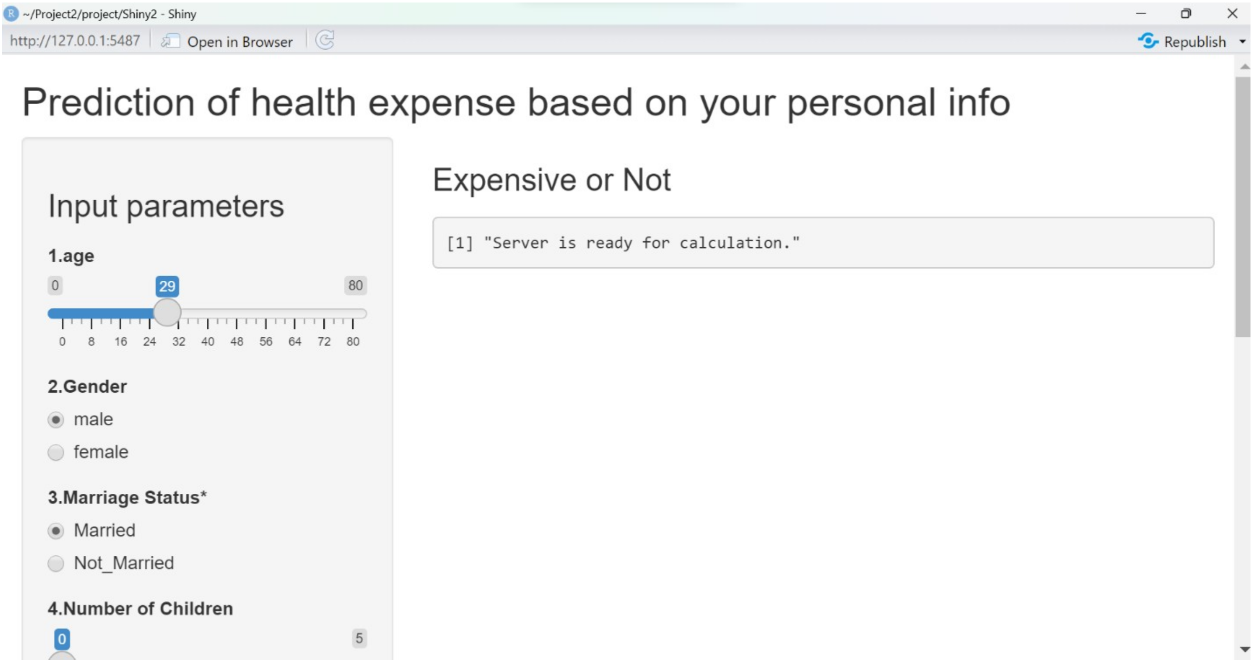
X	age	bmi	children	smoker	location	location_type	education_level	yearly_physical	exercise	married	hypertension	gender
8.00	37.00	27.74	3.00	no	NEW JERSEY	Urban	Bachelor	Yes	Not-Active	Not_Married	0.00	female
10.00	60.00	25.84	0.00	no	PENNSYLVANIA	Urban	Bachelor	No	Not-Active	Married	0.00	female
20.00	30.00	35.30	0.00	yes	NEW YORK	Country	PhD	No	Not-Active	Married	0.00	male
24.00	34.00	31.92	1.00	yes	PENNSYLVANIA	Urban	Bachelor	No	Not-Active	Married	0.00	female
30.00	31.00	36.30	2.00	yes	PENNSYLVANIA	Urban	Master	Yes	Not-Active	Not_Married	0.00	male

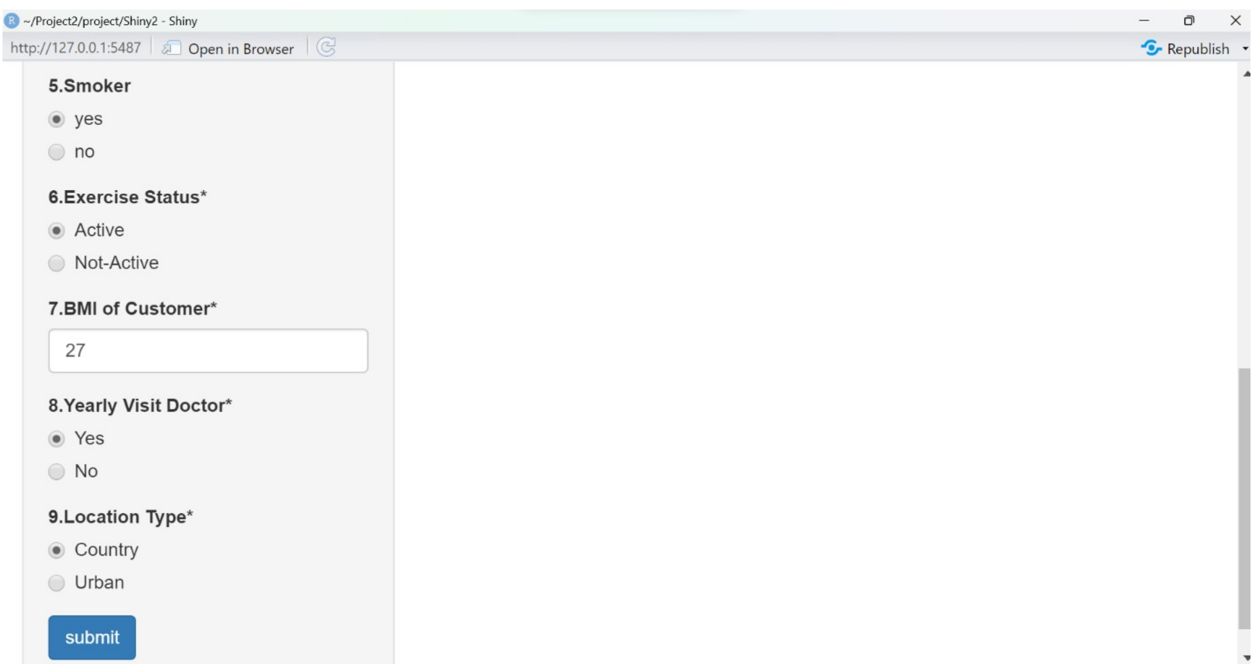


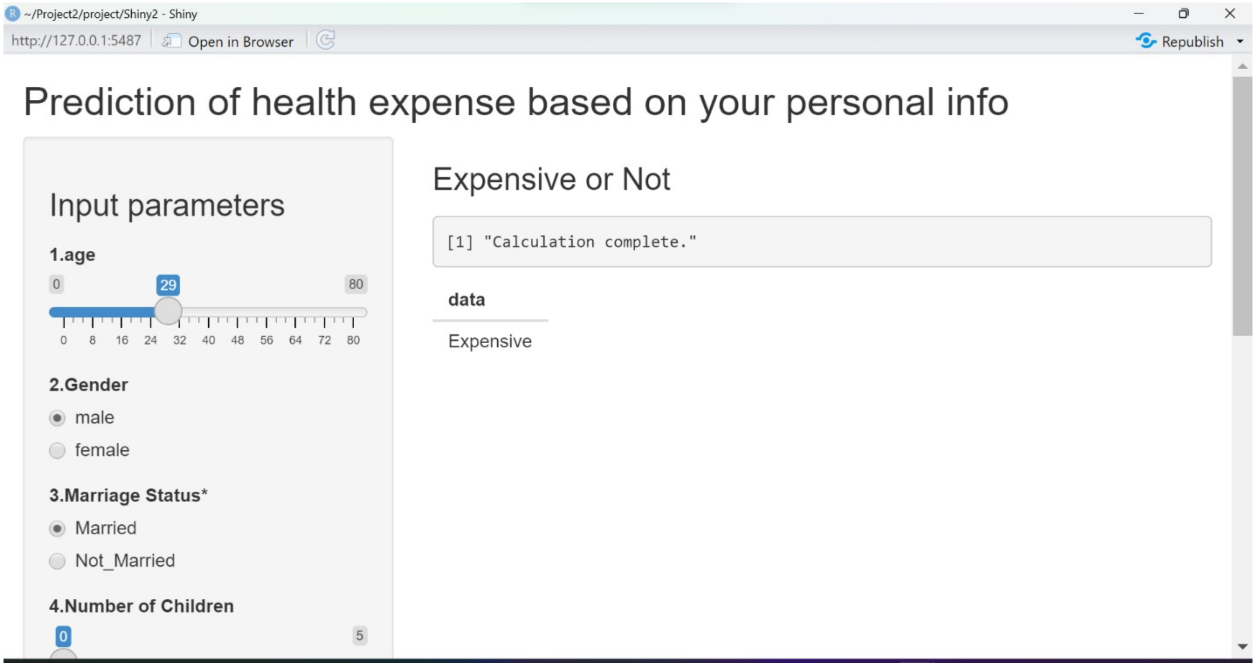
30.00	31.00	36.30	2.00	yes	PENNSYLVANIA	Urban	Master	Yes	Not-Active	Not_Married	0.00	male

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE     8    4
     TRUE      4    4

               Accuracy : 0.6
                 95% CI : (0.3605, 0.8088)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : 0.5956

                  Kappa : 0.1667

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.6667
            Specificity : 0.5000
         Pos Pred Value : 0.6667
         Neg Pred Value : 0.5000
             Prevalence : 0.6000
         Detection Rate : 0.4000
   Detection Prevalence : 0.6000
      Balanced Accuracy : 0.5833

       'Positive' Class : FALSE
```

2. Shiny App 2


~/Project2/project/Shiny2 - Shiny
http://127.0.0.1:5487
Open in Browser
Republish
Prediction of health expense based on your personal info
Input parameters
1.age
0
29
80
0 8 16 24 32 40 48 56 64 72 80
2.Gender
male
female
3.Marriage Status*
Married
Not_Married
4.Number of Children
0
5
Expensive or Not
[1] "Server is ready for calculation."


~/Project2/project/Shiny2 - Shiny
http://127.0.0.1:5487
Open in Browser
Republish
5.Smoker
yes
no
6.Exercise Status*
Active
Not-Active
7.BMI of Customer*
27
8.Yearly Visit Doctor*
Yes
No
9.Location Type*
Country
Urban
submit


~/Project2/project/Shiny2 - Shiny
http://127.0.0.1:5487
Open in Browser
Republish
Prediction of health expense based on your personal info
Input parameters
1.age
0
29
80
0 8 16 24 32 40 48 56 64 72 80
2.Gender
male
female
3.Marriage Status*
Married
Not_Married
4.Number of Children
0
5
Expensive or Not
[1] "Calculation complete."
data
Expensive