**IST-687**

**Introduction to Data Science**

**Analysis of Health Cost Information**

**Project Report**



**Group No. 5**

**Group Member:**

**Tejas Gawade**

**Abhishek Prasun Kumar Singh**

**Malvika Diwan**

**Chintan Patel**

**About the dataset:**

This dataset with the health and medical costs for people between the age groups of 18 and goes above 60. The dataset represents 7 American states. This dataset contains information related to physical exercise, smoking habits, marital status, location, yearly physical, hypertension, gender, and education level of a person. We will factors such as age, Body Mass Index, Smoking status and, exercise to determine how they impact a person’s medical expenses and demonstrate this by analyzing and visualizing this data.

**Objective:**

The main goal of this project is :

* Is to estimate whether or not each person in the dataset will be expensive for the upcoming year. Simply put, will that person spend more or less on medical costs the upcoming year?
* To give the Health Management Organizations practical advice on what initiatives to take to lower the costs of their client's medical care. Additionally, suggest certain business policies to the HMO

**Steps involved in the beginning of the project:**

1. In the first step of this process, we install and load the packages necessary to visualise and analyse the data. The packages used are tidyverse, imputeTS, ggplot, rio, kernlab, caret, e1071, rpart, and rpart.plot.

Graphical user interface, text, application

Description automatically generated

1. In the second step we read the csv file and store it into a data frame ‘df’. Then summarise it using summary(df). There are 8 characters: smoker, location, location\_type, education level yearly physical, exercise, married, X, age bmi, children, hypertension, cost.

Text

Description automatically generated

1. After getting the list of the data in the 3rd step, we check for null values in our dataset only for ‘BMI’. Using the function is.na() we have checked whether are there any null values in the dataset. We have checked the null values for all the columns in the dataset and we found out that there are null values only in 2 columns BMI and hypertension. To remove the null values we created a new filtered data frame and filtered the rows with null values in BMI and for hypertension, we have used na\_interpolation(), these functions fill the null value of the row on the basis of the value in the previous and next row.

A picture containing graphical user interface

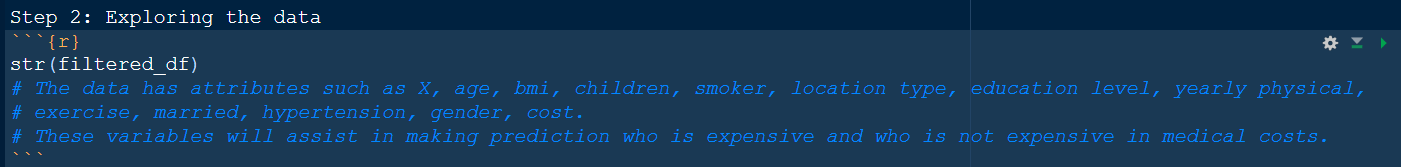
Description automatically generated

Graphical user interface

Description automatically generated

**2. Data Exploration**

2.1 We have explored the attributes in the dataset and their datatypes using different functions such as str() and glimpse(). We got to know that filtered data frame consist of 7504 rows and 14 columns



Text

Description automatically generated

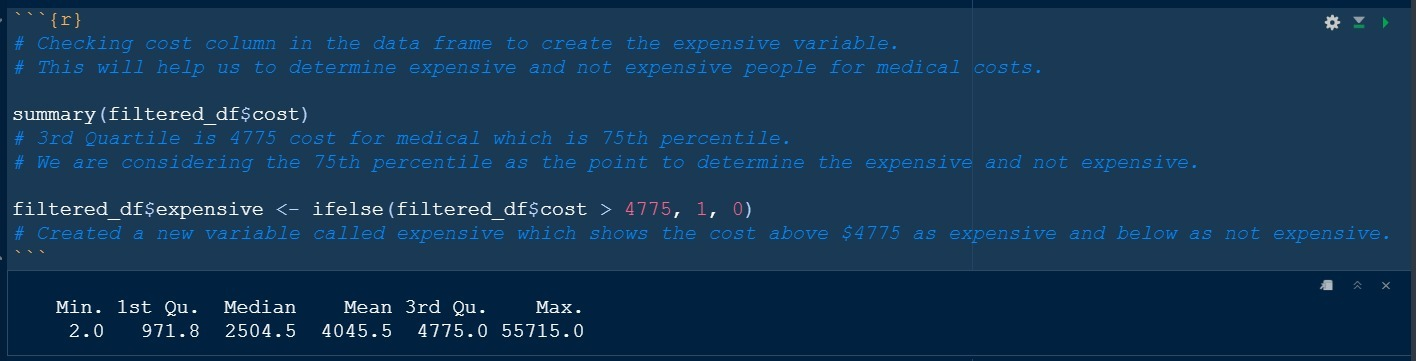
Background pattern

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

2.2 After this step we applied the summary() function on the cost attribute and created a new variable expensive which has the value TRUE if the cost is above $4775 and has the value FALSE if the cost is below $4775. First, we had considered the median of the cost attribute to decide expensive or not, but then we got to know that maximum rows would be expensive therefore we then decided to consider the 3rd quartile value



**3. Data Visualization**

3.1 The first scatter plot consists of BMI on the x-axis and cost on the y-axis. We can see that a person’s education has no impact on the medical costs and we can also infer that as the BMI increase the medical cost also increases. We get to know that BMI has a positive correlation with cost and education type has no relation with the medical cost of a person as People with no college degree also bear the same cost as all the others.

Scatter chart

Description automatically generated with medium confidence

3.2 The second plot also has the same variables on the same axis. But here the determining factor is smoking. Smoking is harmful and a filthy habit and it is common knowledge. This plot shows that a smoker will always have a higher medical expense than a non-smoking person. We can clearly see that smokers have to pay higher medical bill as compared to others. The body mass index is also impacted by smoking and is undesirable for smokers.

Chart, scatter chart

Description automatically generated

3.3 The third plot has cost on the y-axis and cost on the x-axis. The determining factor is exercise or workout. It deals with all sorts of physical activities which a person does to maintain a healthy lifestyle. We can see that people who are physically active and exercise on a frequent basis tend to have lower medical bills and the people who do not work out are prone to have some health disorders and end up paying higher medical bills.

Chart, scatter chart

Description automatically generated

3.4 The fourth plot also has the same variables on the same axis as above. The determining factor is yearly physical. This tells that if a person goes for an annual medical check-up. The plot depicts that people who follow the healthy practice of visiting clinics annually and getting a full body check-up done pay a lesser medical bill than those who don’t. It is advisable for people above 35 years should get a yearly physical because the costs might go above $30000 in many cases. People above 50 years should definitely get a yearly physical otherwise they may end up paying medical bills more than $40000 as we can see in some cases.

Chart, scatter chart

Description automatically generated

3.5 The fifth plot has the same variables on the same axis as above. The determining factor is marital status. This shows whether a person is married or unmarried. According to our plot marital status of a person does have a huge impact on a person’s medical costs. But there are some cases of unmarried people above 50 who end up paying bills more than $30000. This may not directly relate, and other factors may be responsible for this.

Chart, scatter chart

Description automatically generated

3.6 The sixth plot has the same variables on the same axis as above. The determining factor is gender. According to the plot gender of a person is partially unrelated to the medical costs a person bears over a period. But we can see some males above 42 pay higher bills than females of the same age. This may also depend on a person’s lifestyle.

Chart, scatter chart

Description automatically generated

3.7 The seventh plot has the same variables on the same axis as above. The determining factor is hypertension. It’s a condition in which force of the blood against the artery walls is too high. A blood pressure above 140/90 is considered severe. According to our plot people with high blood pressure incur higher medical costs.

Chart, scatter chart

Description automatically generated

3.8 In the next step we an see the bar graphs.

3.9 The first graph shows smokers in our dataset. We can see there are 60000 non- smokers and round 15000 smokers.

Graphical user interface

Description automatically generated

3.10 The next graph represents the 7 states covered in our dataset and shows the people from each one of them. Pennsylvania state is the state from which the most data is collected.

A picture containing graphical user interface

Description automatically generated

3.11 The 3rd graph shows the population represented from urban and rural areas. We have approx. 19000 from the countryside and 58000 from urban cities.

Graphical user interface

Description automatically generated

3.12 The 4th graph shows the age of the population represented in our dataset. We have created a column of age group from the age column and plotted the graph of the age\_group Colum, We can see that most of the data is of the people with age 41-50 and then are the people with age 25-40. We can also infer from the graph that we have fewer data on the people with age 60 and above

Graphical user interface, chart

Description automatically generated

3.12 In the next step we plotted a map showing the most and least expensive states. The most expensive turned to be Pennsylvania followed by New Jersey.

Graphical user interface, text

Description automatically generated

A picture containing chart

Description automatically generated

3.13 The second map shows Body Mass index. Pennsylvania state has the least bmi in our dataset. New York has the highest.

Graphical user interface, text

Description automatically generated

A picture containing graphical user interface

Description automatically generated

3.14 The third map shows medical costs. New jersey has the highest medical expenses in our dataset.

Graphical user interface, text

Description automatically generated

Graphical user interface

Description automatically generated

**Step 4: Preparing the Model**

* 1. In the first step we create a linear model with expensive as as outcome variable and the rest all of the attributes as predictor variables.

Graphical user interface, text

Description automatically generated

Graphical user interface, text

Description automatically generated

* + 1. The outcome from this model, suggests the below attributes to be most significant with a p-value lesser than 0.05.
* age
* BMI
* children
* smokeryes
* locationNew York
* yearly\_physicalYes
* excerciseNot-Active
* hypertension

The model has a predictive power of 42.52% and overall p-value lesser than 0.05.

* 1. We then, created another linear model-this time using all of the significant attributes as predictor variables.

A picture containing graphical user interface

Description automatically generated

Text

Description automatically generated

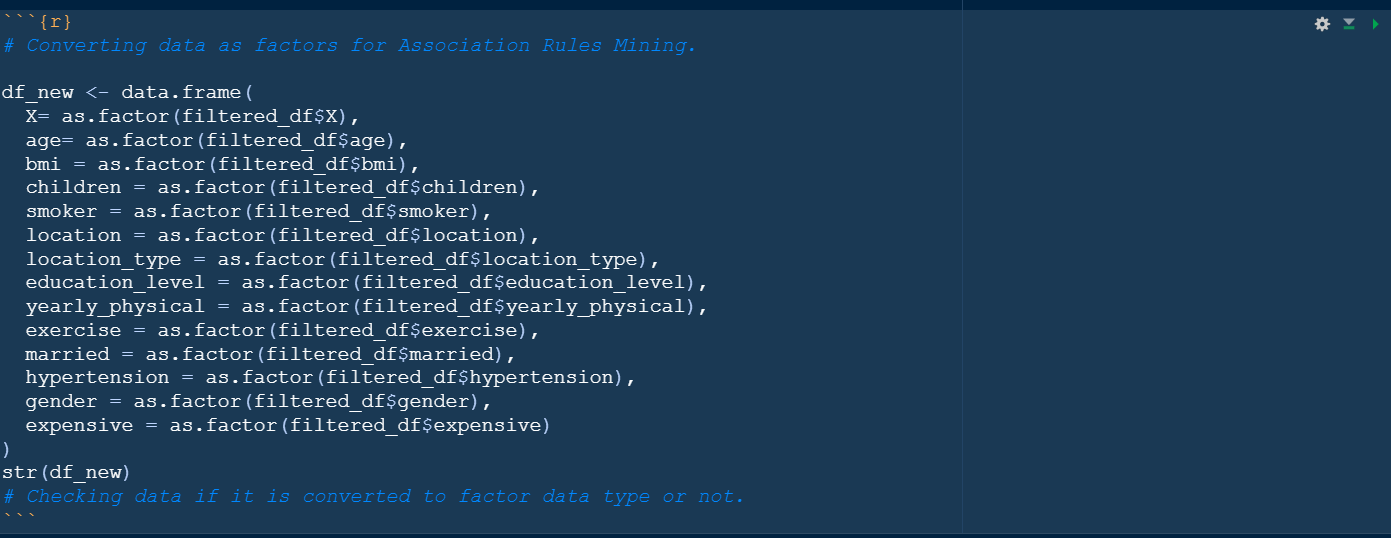
The model has a predictive power 42.43% which is lesser than the linear model created above.

* 1. In order to improve the accuracy, we added 1 more variable which we think can be important as a predictor variable - location.

Text

Description automatically generated

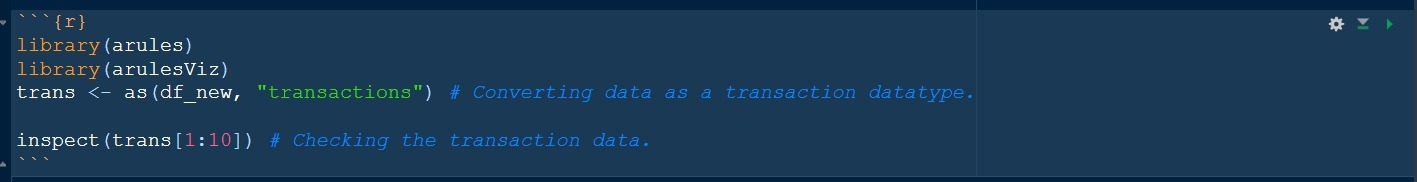
This time the predictive power increased to 42.54%.

* 1. We have used the appropri alogirthm to find out some Assocation rules in, that could help us to understand our data more and we can get some more insights about our attributes and their co-relation. We have created a new data frame and converted all the attributes to factor using as.factor() function as 

Background pattern

Description automatically generated

To find association rules we need to convert the data frame into transactions.



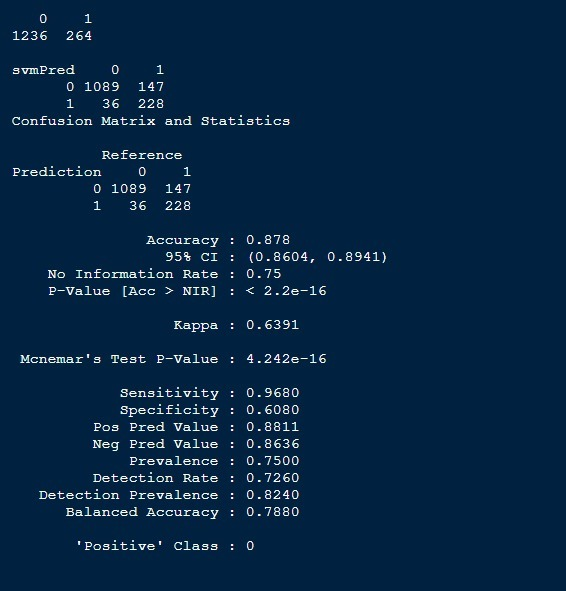
Here we have converted each row into a transaction and we find some association rules related to the person being expensive.

There are some rules that we found out:

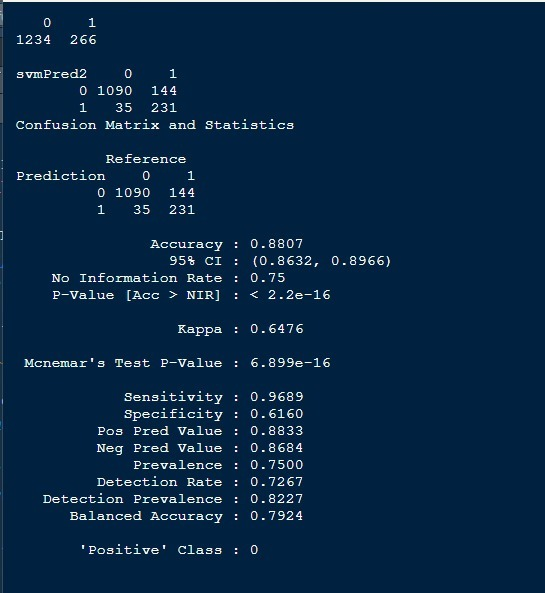
* If a person is older, he/she does not exercise regularly then they tend to have high medical costs.
* If a person is a smoker and he/she is of young age then also the person tends to have high medical expenses.
  1. In order to find the “best” margin and reduce the risk of error on the data, an SVM model was created for predicting the expensive customers.
  2. Splitting the data: We want to use the SVM model to predict whether the customer is expensive or not. Therefore we divided our dataset into trainsets which will be used for training the data and a test set which will be used to test our model prediction.

Our trainset comprised 80% of the data set and the test set comprised 20% of the data set.

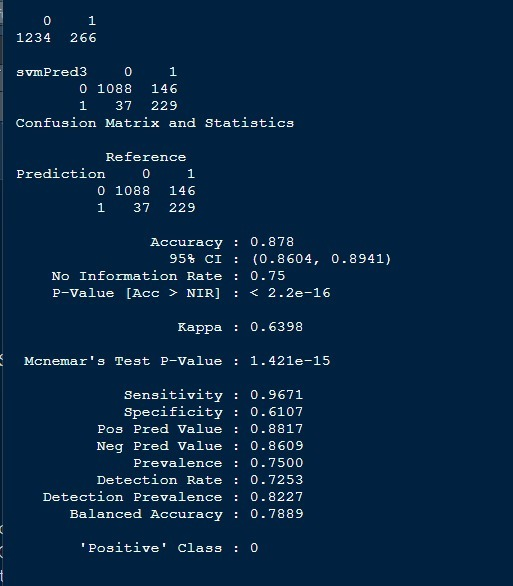
* + 1. In the first model, attribute expensive was taken as the dependent variable and all the other attributes from the data were taken as the predicting variables. The accuracy in predicting who is expensive and who is not with a sensitivity of 96% was observed to be 87.8% for this model.



* + 1. In the second model attribute expensive was taken as the dependent variable and age, bmi, children, smoker, exercise and hypertension data were taken as the predicting variables as these were also significant variables in the linear model. The accuracy in predicting who is expensive and who is not with a sensitivity of 96% was observed to be 88.07% for this model.



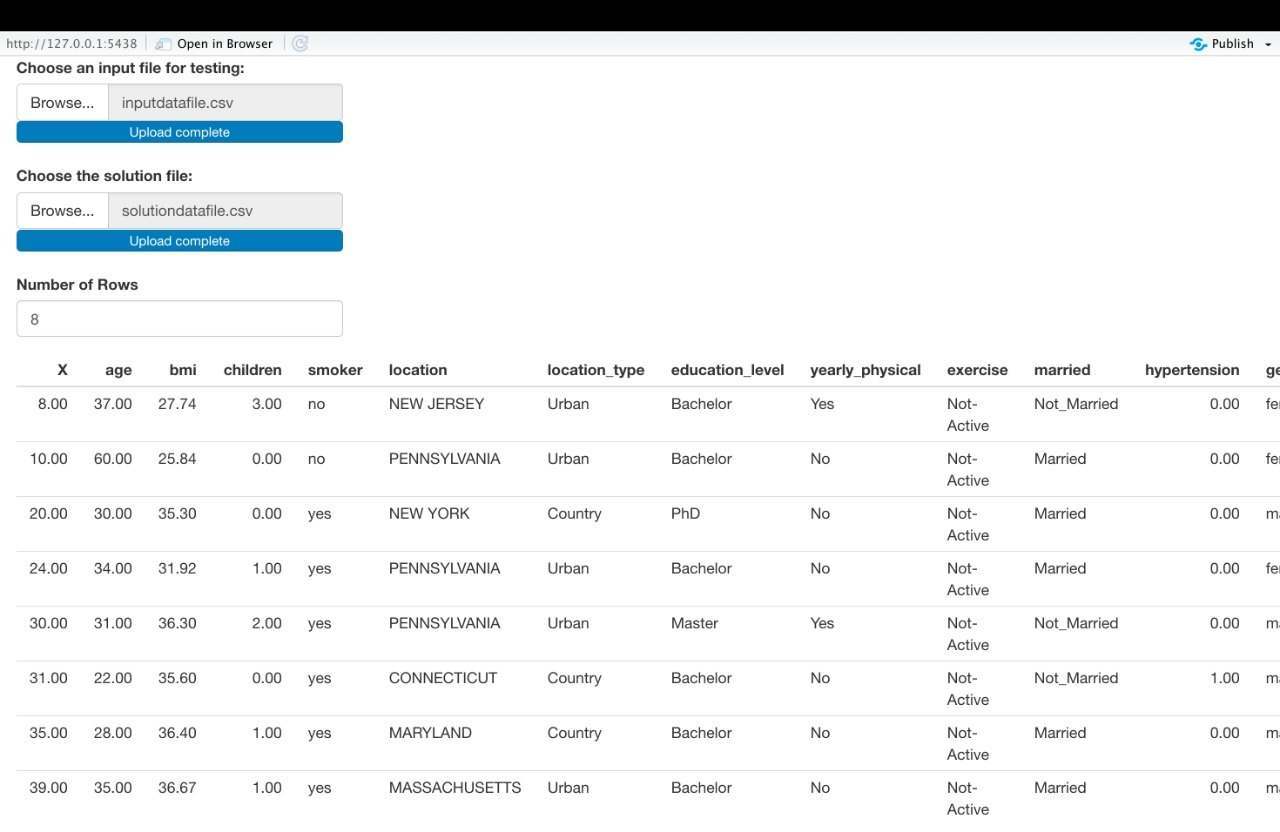
* + 1. In the third svm model- the variable location was added as a predicting variable in addition to the rest in from second model with expensive as the dependent variable. The accuracy in predicting who is expensive and who is not with a sensitivity of 96% was observed to be 87.78% for this model.

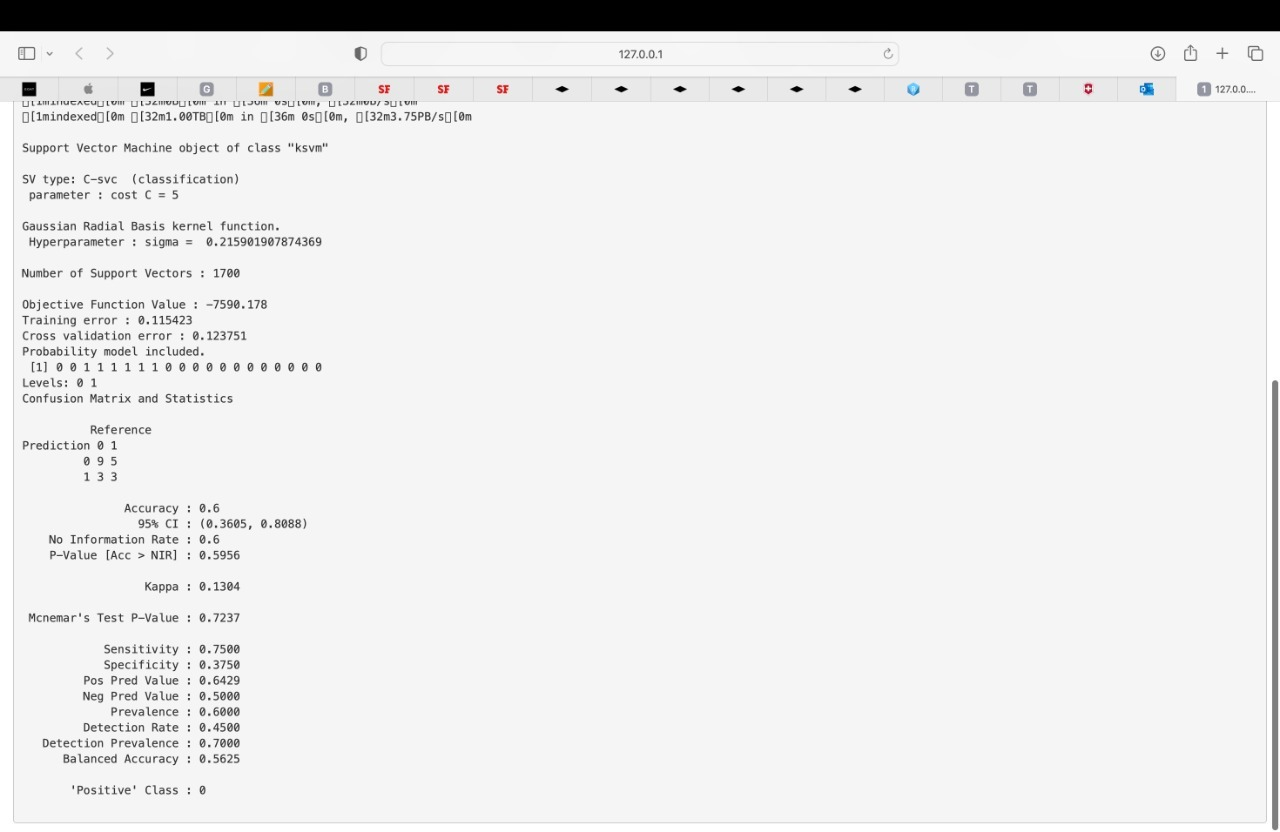


Hence, we conclude that the second SVM model has the highest accuracy in predicting expensive customers. The predicting variable in the second SVM model were: age, bmi, children, smoker, exercise, and hypertension.

**Testing the data set provided and showing the result in Shinny App:**

We have cleaned the test data set provided by the professor and predicted whether the person will be expensive or not. We have shown the results of the prediction and accuracy of the prediction on the Shinny app:





We got 60% accuracy and a sensitivity of 75% on the test data set.

**Actionable Insights and Business Suggestions:**

**Q1 Sort the consumers into categories and discuss underlying values based on the distribution of the American states.**

Ans- Seven states, including Connecticut, Maryland, Massachusetts, New Jersey, New York, Pennsylvania, and Rhode Island, were used to collect the data. It was noted that Pennsylvanians make up the majority of insurance subscribers. Maryland comes next with about 650 individuals. With about 490 people, Massachusetts has the fewest number of people in our dataset. New Jersey, with a slight difference and 500 persons, is second-to-last.

The HMO must broaden its reach and get residents of other states to enroll in their health insurance program. This would allow us to see a wider picture, and with more information and data, we would be able to offer better solutions.

**Q2 Sort the data into the two geographical groups of urban and rural areas and then use those classifications to generate useful insights.**

Ans- Due to the higher population density in metropolitan areas compared to rural areas, it has been found that these locations have the highest amount of health issues. The graph shows that about 1950 citizens live in rural areas and about 5800 live in metropolitan areas. The following factors may contribute to urban health problems:

* Compared to urban cities, rural areas have advantages in terms of lifestyle, such as less noise and air pollution. Because they consume healthier foods and get enough exercise, people who live in rural areas are less stressed and have better mental health.
* Additionally, we recommend that the HMO promote its offerings and enhance accessibility for rural inhabitants.

**Q3 Classify the information based on the dataset's age demographic, then list the observations.**

Ans- The minimum age range included in the data collection was 18, while the remainder is older. The age group is divided into four brackets: 18–25, 26–40, 41–60, and the last category, for people over 60. The youngest individuals in the first group (18 to 25) total around 1800. There are 2100 persons in the 26–40 age range. The largest age range for this graph is represented by the third category, which has 3000 people in it. and the least number of people are found in the category with those over 60. However, it was found that, compared to other age groups, the information for individuals over 60 was fewer.

**Q4 Draw conclusions based on what you notice about a person's exercise habit.**

Ans- Regularly physically active people are found to have lower medical costs than those who are not. Exercise and physical activity of any kind are known to be beneficial to health over the long term. The HMO can collaborate with neighborhood gyms in various communities and perhaps offer a discount or other incentives to its clients. People will be inspired to join health clubs and live longer, healthier lives as a result of this!

**Q5 Describe how smoking affects a person's long-term medical expenses.**

Ans- Our data shows that smoking has a significant impact on medical costs. Areas where it is thought that a higher proportion of people smoke can host awareness initiatives. To reduce the risk of health problems linked to smoking, regular health exams for adults over 40 ought to be encouraged.