

## Health Insurance Cost Predictability

**Author Name:** Sanjir Inam Salsabil

*Department of Computer Science, Kulliyah of Information and Communication Technology, International Islamic University Malaysia, Kuala Lumpur, Malaysia.*

[sanjirsalsabil1@gmail.com](mailto:sanjirsalsabil1@gmail.com)

### **Author Note**

*This paper holds all the information & research about the project of Data Science*

*Matric no: 1431937*

*Section: 01*

*Date: 16/12/2018*

*Submitted to: Dr. Raini Binti Hassan*

## Table of Contents

| Section  | Page    |
|--|---------|
| Abstract   | 3       |
| Research Question  | 3       |
| Hypothesis   | 3       |
| Introduction   | 3       |
| Objectives   | 4       |
| Expected Outcomes  | 4       |
| Literature Review  | 4 - 6   |
| Machine Learning Model <ul style="list-style-type: none"> <li>• Calculation</li> <li>• Entropy</li> <li>• Information Gain</li> <li>• Workflow</li> </ul>  | 7 - 9   |
| Experimental Setup (Data Science Process) <ul style="list-style-type: none"> <li>• Data Collection</li> <li>• Dataset</li> <li>• Data Processing</li> <li>• Data Cleaning</li> <li>• Exploratory Data Analysis (EDA)</li> <li>• Machine Learning Algorithm</li> <li>• Data Visualization</li> <li>• Data Product</li> <li>• Feedback Loop</li> </ul> | 10 - 19 |
| Model Testing  | 19 - 22 |
| Hypothesis Testing   | 23      |
| Data Analytics Tools   | 23      |
| Expected Challenges or Limitations   | 24      |
| Appendix   | 24 - 25 |
| References   | 26      |

**Abstract:** According to US research statistics, uninsured Americans have a higher mortality rate, as well as a higher chance of going into debt (Erica Block, 2017). They are more likely to die suddenly from a curable disease due to insufficient money<sup>[1]</sup>. Furthermore, in logical terms those who are older, overweight/underweight, have smoking habits or many children have higher chances of health risks and being hospitalized. This paper illustrates the research work of our data science project and analyze our hypothesis to be correct or not by predicting health insurance cost using a machine learning model.

**Keywords:** Health Insurance Cost, Machine Learning Model, Smoking, Prediction etc.

**Research Question:** Is there any correlation for a rise of health insurance cost due to smoking habits, BMI, age and the no. of children?

**Hypothesis:** It is assumed that people who smoke, are overweight and have many children, tend to have more health risks, thus the risk of getting hospitalized is higher compared to others.

**Introduction:** Planning for our financial expenses is crucial if we want to have a stable lifestyle. This issue becomes more significant for those who have families. One of the major concerns, is our health, and this is where health insurance comes into the picture. Although many people don't like the idea of having to pay for insurance, but it can be a life saver. Our data science project aims to help both insurance companies as well as normal citizens. Insurance companies would find it easier to figure out how much they should charge customers that possesses certain criteria such as Smoking Habits, BMI, Age, No. of Children etc. Apart from that, people would also have more knowledge on the different cost that they would have to incur if they fall into any of those criteria. Thus, they can use this information to have a much better financial plan, as well as perhaps to change their habits or lifestyle, for example, stop smoking, try to lose weight, etc. To predict the outcomes of our research and analyze the hypothesis, we train a model using a machine learning algorithm called two class boosted decision tree.

## **Health Insurance Cost Predictability**

### ***Objectives***

- Learn predictive analysis through data science processes.
- Analyze and visualize dataset to predict the health insurance cost accurately.
- Develop a machine learning model using Two Class Boosted Decision Tree and implement in real life scenarios.
- Measure the influences of certain criteria on health insurance cost.
- Raise awareness to help people know how health insurance costs can be altered.
- Aid people to have a better financial plan for their future.
- Conclude the result based on the outcome of our constructed algorithm to except or reject the hypothesis.

### ***Expected Outcomes***

Successfully accomplish predictive analysis of hypothesis through data science processes and build a machine learning model that can predict health insurance cost labeled high or low.

### ***Literature Review***

Enquiring about health insurance options, would result in you receiving a range of various prices from each insurance company even though the policies are very similar. This is due to them assigning different values to components that belong to your risk profile. The risk profile comprises of a person's physical and medical risk factors as well as lifestyle and personal risk factors (K. Botkin, 2018).

The insurance companies tend to use historic data and research to deduce the factors and a benchmark for the risk profiles. Once they review a customer's risk profile, it is compared to the standard benchmark that they have created and would lead to the company deciding whether they should provide you with insurance or not.

If the company decides to provide you with health insurance, the premium charges would be calculated based on the risk factors and your individual application. The methods and standards for calculating the premiums vary from one company to another, thus resulting in different premium costs.<sup>1</sup>

**Physical and medical risk factors.** Those who have a high body mass index (BMI), would normally be charged significantly higher premiums compared to a person who has normal weight. This is due to them having a high risk of acquiring diseases including diabetes, sleep apnea, and heart and joint problems. Tobacco or cigarette users are also penalized for having a higher risk of developing cancer and other health issues. Ex-smokers who quit recently are also charged extra, as it takes time for the body to recover from the damages done by smoking. Furthermore, other factors include gender and age, whereas women and older people tend to be charged higher premiums. Pre-existing medical conditional as well as family history are also factors that may result in a person's premium cost to be higher (K. Botkin, 2018).

**Lifestyle and personal risk factors.** People who work in hazardous environments or locations that have high injury possibilities would have to pay higher premiums compared to those that have safer jobs. This also applies to those who live and reside in these types of locations where climate changes could be a factor. Being married actually helps to reduce premium costs as it is believed that married couples tend to be healthier and live longer than single people. Those who apply for health insurance for the first time, might be charged a higher premium as companies may suspect that you are taking up the insurance finally because of planned medical check-ups or may have developed a certain health problem. Insurance companies considers many of these factors takes them into account when deciding the type of premium to charge on a certain individual (K. Botkin, 2018).

---

<sup>1</sup> Botkin, K. (n.d.). 10 Factors That Affect Your Health Insurance Premium Costs. Retrieved from <https://www.moneycrashers.com/factors-health-insurance-premium-costs/>

On average, a person who smokes may have to pay an extra 15-20% charge on premiums compared to non-smokers.<sup>2</sup> If the normal premium payment was around \$500 for an individual, a smoker would have to pay \$600 or more in order to be insured by the specified company (K. Mercadante, 2013).

According to HealthMarkets (an insurance company), **at least 15.5% of US adults are smokers** which comprises of 17.5% for men and 13.5% for women. About one quarter of all the smokers come from poverty while 14.3% are just at the border line of poverty. Furthermore, 24.1 percent of all smokers have no high school diploma while 40.6 percent are GED recipients. Only 19.7 percent have completed their high school diploma, 7.7 percent have received an undergraduate degree and as little as 4.5 percent have achieved a post-graduate degree.<sup>3</sup>

So how to quit smoking? This comes to mind for many, but only a few manage to overcome the issue at hand. This is due to the addiction and mindset of the person. Steps you may take includes, setting a quit date and following it, keep yourself engaged with other beneficial activities while having the urge to light a cigarette. Try not to replace the habit with something else like gum, as it would make you feel like it's a sacrifice. Let people around you know that you have decided to quit smoking and ask to help you with this new challenge. Changing daily habits can have a major improvement on your overall health and may also assist you with getting rid of your smoking habits.

In the United States, employers are permitted by the Affordable Care Act to charge overweight (above 30 BMI) patients an extra 30-50% higher premium cost for health insurance. These extra charges would help employers recover some of the increased health-care costs incurred from being overweight. Patients are sometimes provided with incentives to encourage them to seek medical advice for losing weight or certain programs that would help to guide and motivate them to try and lose that extra unhealthy weight (J. Crowley, 2016).

---

<sup>2</sup> Mercadante, K. (2013, March 15). How Smoking Affects Your Health Insurance Premiums -. Retrieved from <https://www.mcmha.org/smoking-affects-health-insurance-premiums/>

<sup>3</sup> HealthMarkets. (2016). What You Need to Know About Smoking and Health Insurance. Retrieved from <https://www.healthmarkets.com/content/what-you-need-know-about-smoking-and-health-insurance>

### Machine Learning Model

In our project we apply Two Class Boosted Decision Tree algorithm to achieve our goal from the selected dataset. The Decision Tree algorithm builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf node. The core algorithm for building decision trees called \*ID3\* by J. R. Quinlan which employs a top-down, greedy search through the space of possible branches with no backtracking. ID3 uses Entropy and Information Gain to construct a decision tree.

- Calculate the entropy using frequency tables.
- Information Gain: The information gain is based on the decrease in entropy after a dataset is split on an attribute.
- Calculate entropy of the target.
- The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain or decrease in entropy.
- Choose attribute with the largest information gain as the decision node, divide the dataset by its branches and repeat the same process on every branch.
- A branch with entropy of 0 is a leaf node.
- A branch with entropy more than 0 needs further splitting.
- The ID3 algorithm is run recursively on the non-leaf branches, until all data is classified.

#### ***Calculation:***

Gradient Boosted Decision Tree is an optimized algorithm which uses multiple decision trees. A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous).

### Entropy

ID3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one. To build a decision tree, we need to calculate two types of entropy using frequency tables as follows:

a) Entropy using the frequency table of one attribute:

$Entropy (Set) =$

$$\sum_{i=1}^c -p_i \log_2 p_i$$

b) Entropy using the frequency table of two attributes:

$E (T, X) =$

$$\sum_{c \in x} P(c)E(c)$$

### Information Gain

The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain.

**Step 1:** Calculate entropy of the target.

**Step 2:** The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain or decrease in entropy.

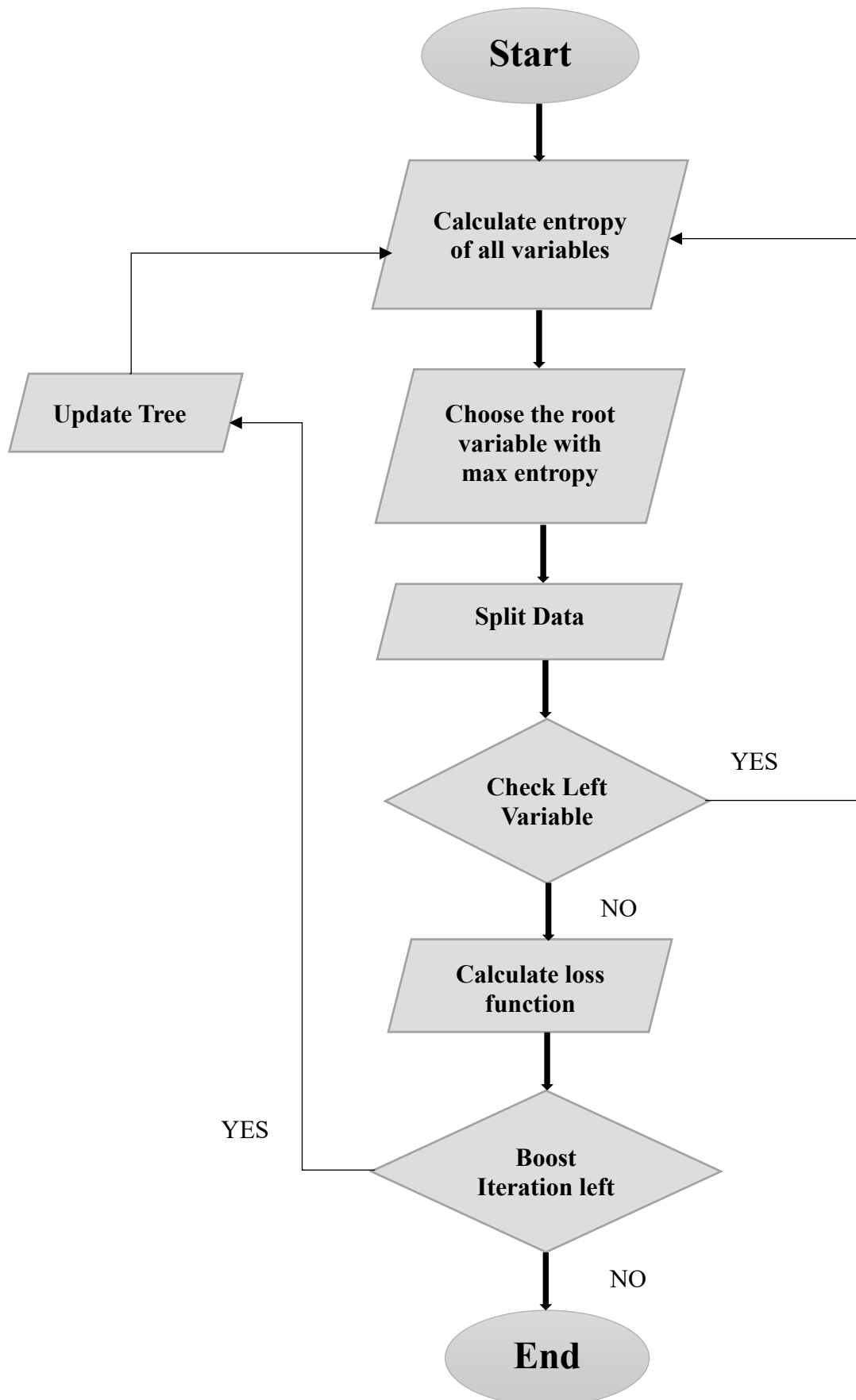
**Step 3:** Choose attribute with the largest information gain as the decision node, divide the dataset by its branches and repeat the same process on every branch.

**Step 4a:** A branch with entropy of 0 is a leaf node.

**Step 4b:** A branch with entropy more than 0 needs further splitting.

**Step 5:** The ID3 algorithm is run recursively on the non-leaf branches, until all data is classified.



**Workflow**

### Experimental Setup

We follow the data science process to analyze the problem and to complete our project successfully.

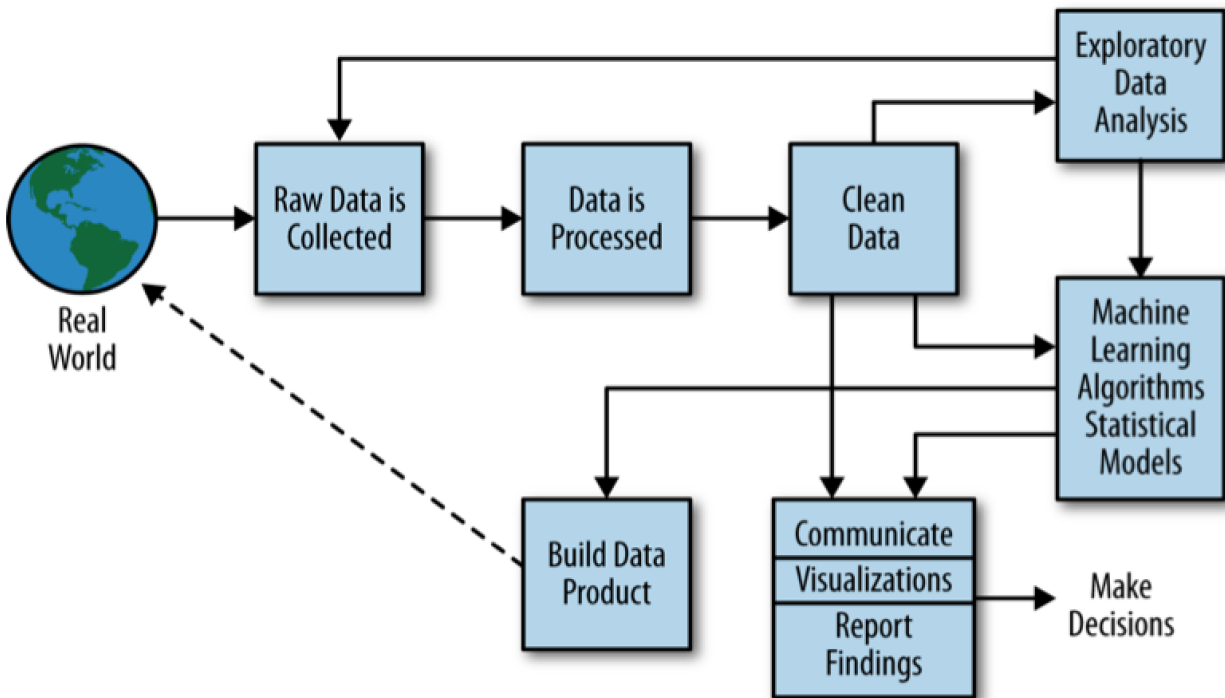


Figure 1: Data Science Process

### Data Collection

Raw data collection from the real world is the first step of the data science process. We retrieved an online dataset for records on Americans which is available on Kaggle. The dataset contains records on age, BMI, smoker and the no. of children as well as the charges for health insurance.

### Dataset

The dataset consists of 8 different features and 1338 observations. Those 8 features are age, sex, BMI, children, smoker, region, amount and charge. Charge feature is considered as the dependent variable. It has two values which are high and low. Rest of others are independent variable. So, the scenario of the dataset is a supervised classification problem.

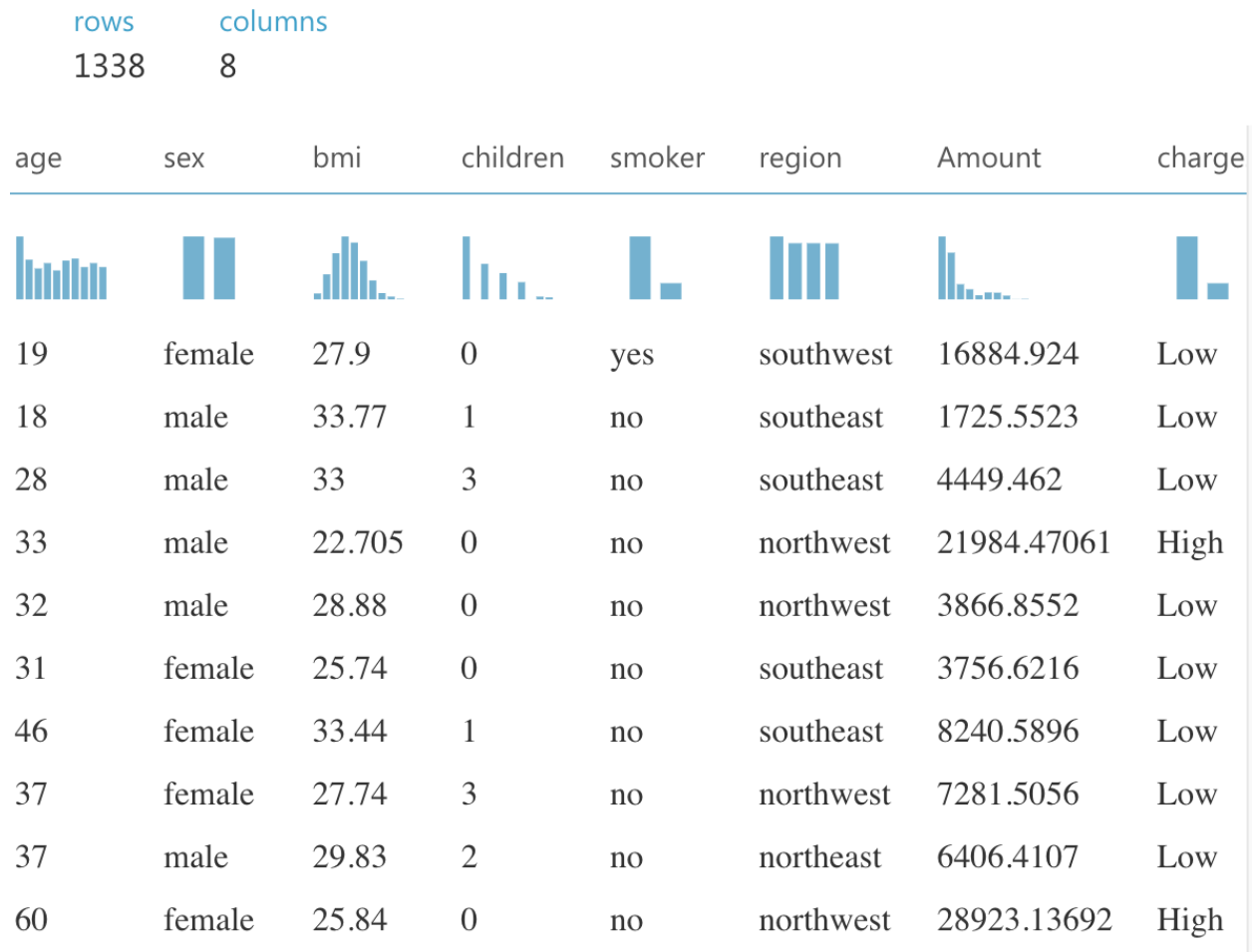


Figure 2: Dataset of American's Health Insurance Cost

### Data Processing

We started to explore the dataset further with a goal of finding any issue and then apply data engineering method to prepare the dataset for our model. So, we removed the undesired columns that might not be an important feature considering our prediction. We removed region and amount.

Amount column is a redundant column as it is a subset of charges column. We considered high charges when the amount is higher than 20000 to make this dataset easier to interpret.

After that, we converted the feature type to our desired feature types. In this dataset ‘sex’ was a string feature but through ‘Edit Metadata’ option we changed into categorical feature. We did the same for the children and smoker feature as those features had only 3-4 unique values from 1338 observations.

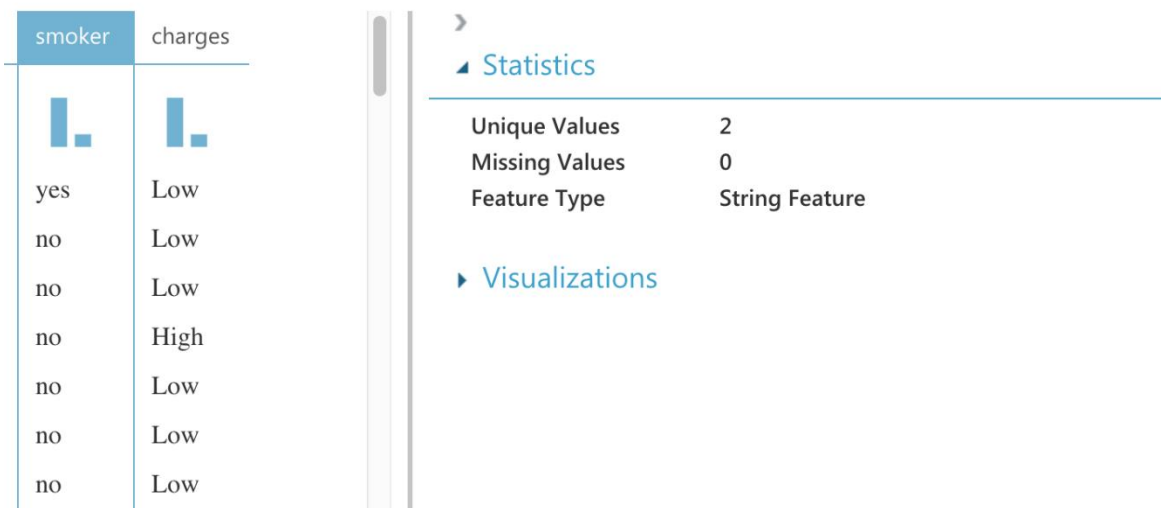


Figure 3a: Before Processed

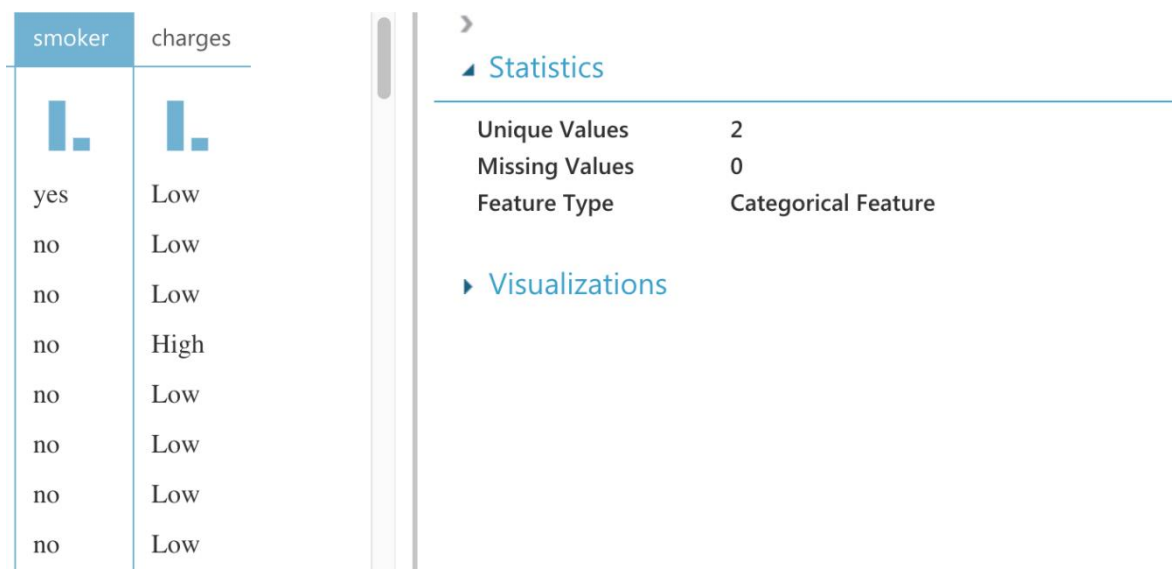


Figure 3b: After Processed

## Data Cleaning

Once dataset is processed, we checked for missing values. If we would any missing values, we would change it with mean of that column. But we did not find any missing values as the dataset is pretty much cleaned.

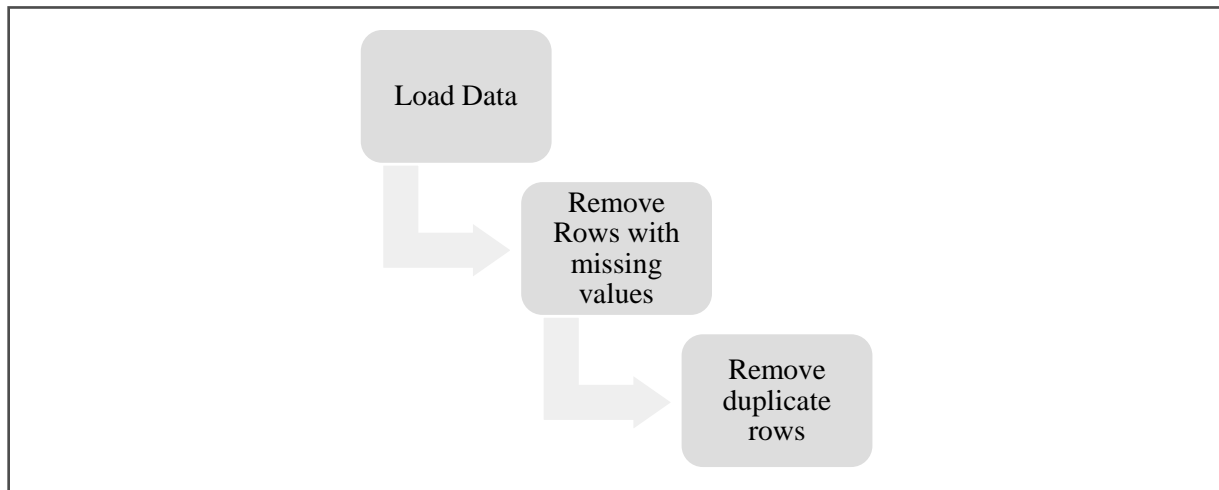


Figure 4a: Data Cleaning Process

The screenshot shows a data cleaning workflow in a software interface. The workflow consists of four steps: 'insurance.csv', 'Select Columns in Dataset', 'Edit Metadata', and 'Clean Missing Data'. The 'Clean Missing Data' step is highlighted with a blue border and numbered '1' and '2'. The interface shows 'Finished running' with a green checkmark and 'Draft saved at 19:02:12'. On the right, the 'Clean Missing Data' configuration panel is visible, showing 'Columns to be cleaned' as 'age, children', 'Minimum missing value...' as '0', 'Maximum missing value...' as '1', 'Cleaning mode' as 'Replace with mean', and 'Cols with all missing val...' as 'Remove'.

Finished running ✓

Draft saved at 19:02:12

insurance.csv

Select Columns in Dataset ✓

Edit Metadata ✓

Clean Missing Data ✓

1 2

Clean Missing Data

Columns to be cleaned

**Selected columns:**  
Column names:  
age, children

Launch column selector

Minimum missing value...  
0

Maximum missing value...  
1

Cleaning mode  
Replace with mean

Cols with all missing val...  
Remove

Figure 4b: Filling Missing Data

### Exploratory Data Analysis (EDA)

EDA is one of the most important part in data science process to understand the data and make them clear to compute. In this section we plotted histogram, boxplots to check the outliers of the data, where we found only the BMI feature has outliers. So, further clean it and check again. We also summarized the data where we found mean, median, max values and so on for numeric features.



Figure 5a: Finding Outliers from Boxplots














| Feature   | Count   | Unique Value Count  | Missing Value Count   | Min   | Max   | Mean  | Mean Deviation  | 1st Quartile  | Median  | 3rd Quartile  | Mode  | Range   |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| age   | 1338  | 47  | 0   |   |   |   |   |   |   |   |   |   |
| sex   | 1338  | 2   | 0   |   |   |   |   |   |   |   |   |   |
| bmi   | 1338  | 548   | 0   | 15.96   | 53.13   | 30.663397   | 4.897871  | 26.29625  | 30.4  | 34.69375  | 32.3  | 37.17   |
| children  | 1338  | 6   | 0   |   |   |   |   |   |   |   |   |   |
| smoker  | 1338  | 2   | 0   |   |   |   |   |   |   |   |   |   |
| charges   | 1338  | 2   | 0   |   |   |   |   |   |   |   |   |   |

Figure 5b: Summary of Data (Numeric Features)











| Sample Variance   | Sample Standard Deviation   | Sample Skewness   | Sample Kurtosis   | P0.5  | P1  | P5  | P95   | P99   | P99.5   |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
| 37.187884   | 6.098187  | 0.284047  | -0.050732   | 17.355075   | 17.89515  | 21.256  | 41.106  | 46.4079   | 47.5452   |

Figure 5c: Summary of Data (BMI)

We also calculated the correlation of variables to understand the relationship between them.

#### PEARSON CORRELATION TEST RESULT

| Attributes | Age   | BMI   | Children |
|------------|-------|-------|----------|
| Age        | 1.00  | 0.109 | 0.042    |
| BMI        | 0.109 | 1.00  | 0.013    |
| Children   | 0.042 | 0.013 | 1.00     |

#### SPEARMAN RANK CORRELATION TEST RESULT

| Attributes | Age   | BMI   | Children |
|------------|-------|-------|----------|
| Age        | 1.00  | 0.108 | 0.057    |
| BMI        | 0.108 | 1.00  | 0.016    |
| Children   | 0.057 | 0.016 | 1.00     |

### Machine Learning Algorithm

At this part we do the main business where we use Two Way Boosted Decision Tree algorithm to predict the health insurance cost. We split the dataset and 80% of the data we feed in the training set and 20% for the test set. So, the ratio of split data is 8:2. After feeding data from the dataset, we got an accuracy of 91.4% from our trained model. It means our model can predict possibilities more than 91 percent accurately.

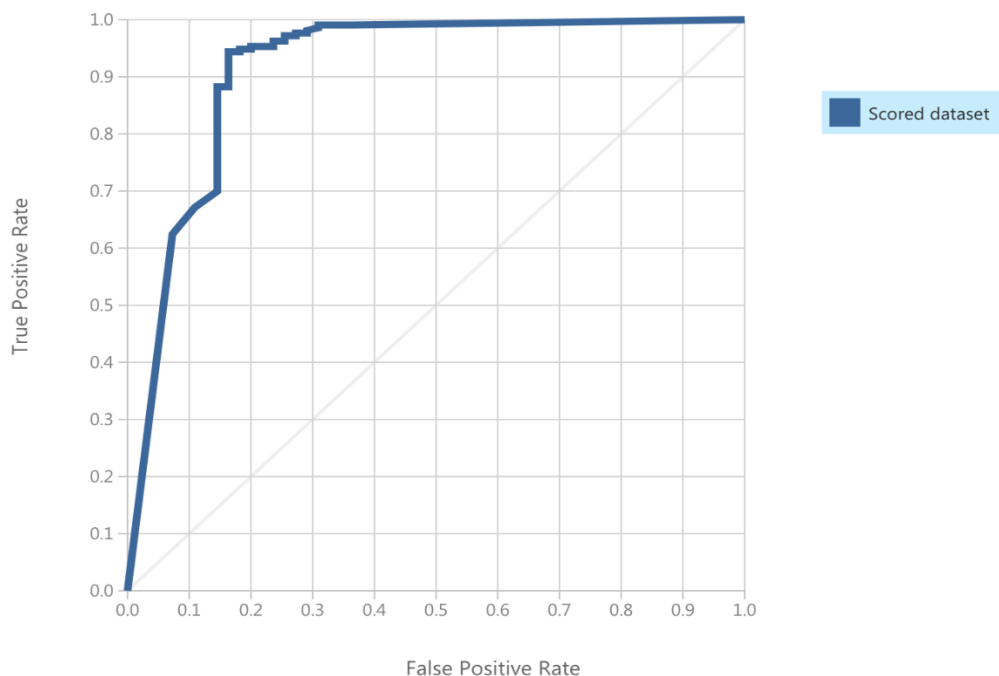


Figure 6a:  
ROC curve



|                |                |              |              |            |              |
|----------------|----------------|--------------|--------------|------------|--------------|
| True Positive  | False Negative | Accuracy     | Precision    | Threshold  | AUC          |
| <b>203</b>     | <b>10</b>      | <b>0.914</b> | <b>0.940</b> | <b>0.5</b> | <b>0.891</b> |
| False Positive | True Negative  | Recall       | F1 Score     |            |              |
| <b>13</b>      | <b>42</b>      | <b>0.953</b> | <b>0.946</b> |            |              |
| Positive Label | Negative Label |              |              |            |              |
| <b>Low</b>     | <b>High</b>    |              |              |            |              |

Figure 6b: Accuracy of trained model

### Data Visualization

Understanding data is a very important thing for any data scientist to solve problems. And data visualization is the medium to understand the data. After finishing all pre-processing, we visualized our dataset. We compare features with labeled data and did crosstab analysis, multi-box plotting.

Figure 7a: Crosstab Analysis (Smoker vs Scored Labels)

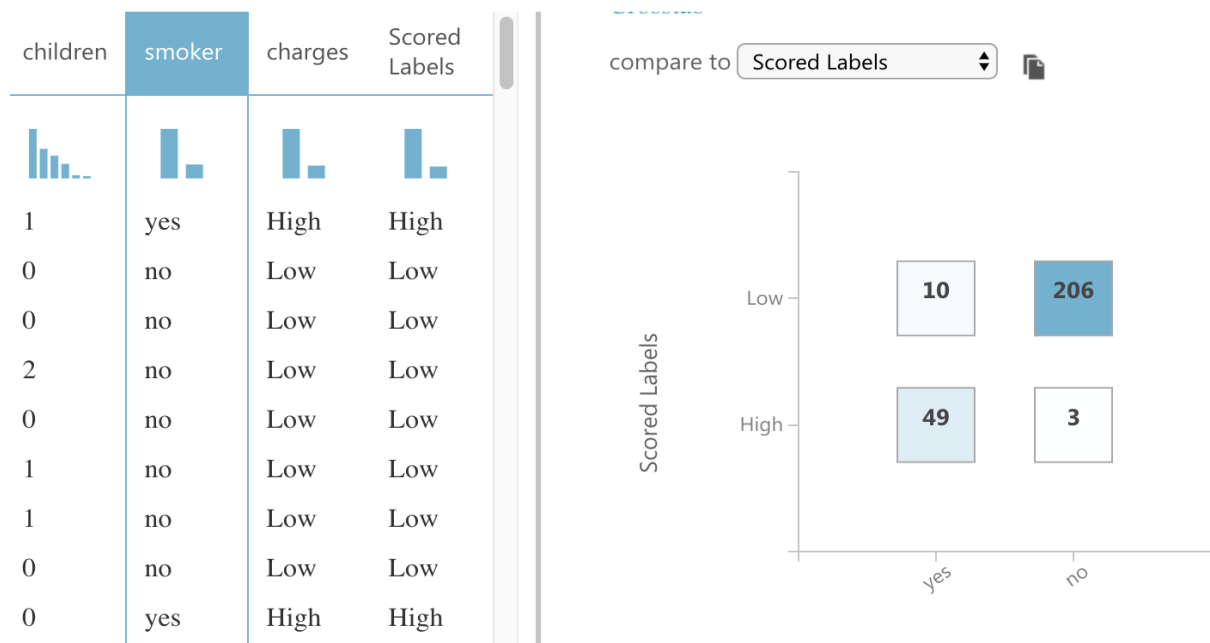
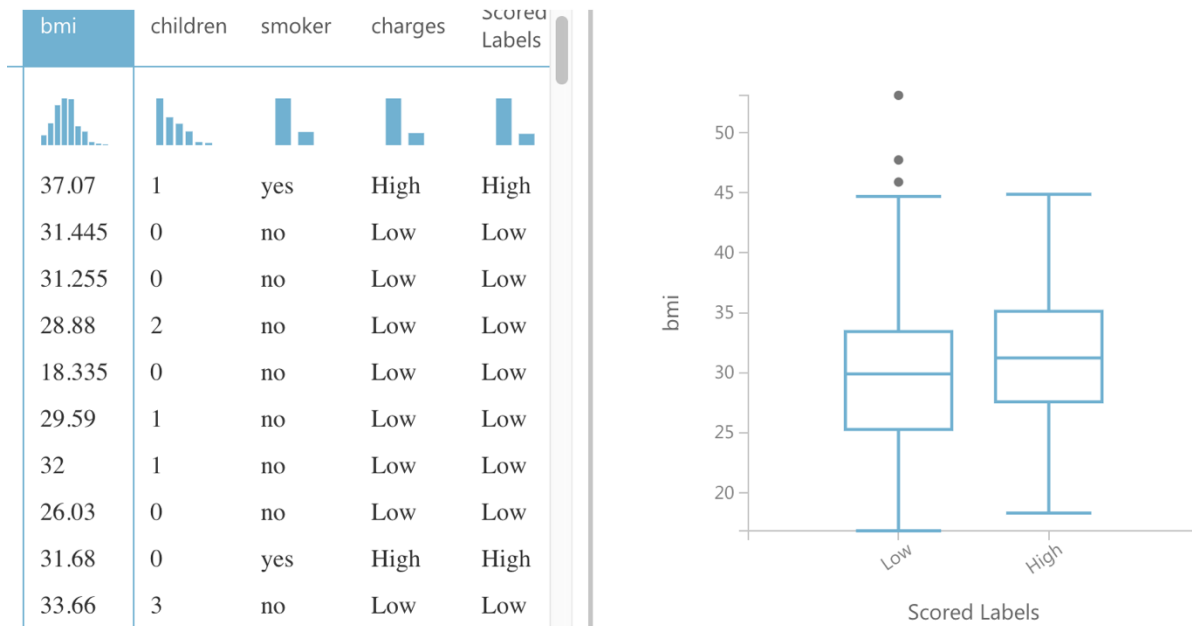


Figure 7b: Crosstab Analysis (Children vs Scored Labels)

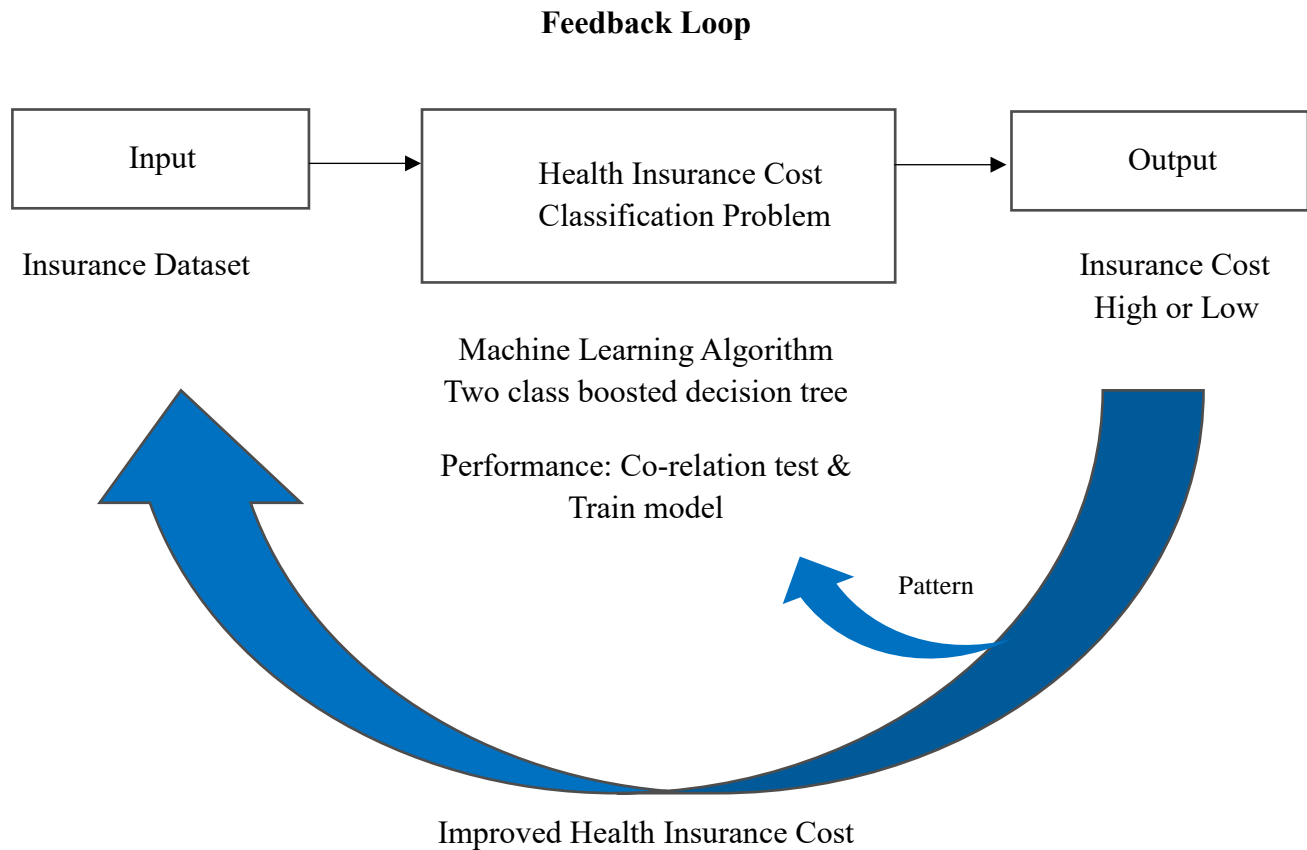
Figure 7c: Multi-box plotting (BMI vs Scored labels)



### Data Product

| Independent Variable | Dependent Variable |
|----------------------|--------------------|
| Age                  | Charges            |
| Sex                  |                    |
| BMI                  |                    |
| No. of Children      |                    |
| Smoker               |                    |

From this table we can easily understand that our data product will predict charges in terms of age, sex, bmi, number of children and smoker.



Generating more data that can be fed into the model to further refine it and helps to make it more accurate.

### Model Testing

To test the model, first we made a simple decision tree for our dataset to observe how it looks like.

#### R code for the tree:

```

Insurance <- read.csv("insurance.csv")
set.seed(1)
Insurance
install.packages("tree")
library(tree)
library(rpart)
library(rattle)
  
```

```

library(rpart.plot)
library(RColorBrewer)

head(Insurance)
tail(Insurance)
str(Insurance)
tree <- rpart(charges ~ age+sex+bmi+children+smoker, data = Insurance)
fancyRpartPlot(tree)

```

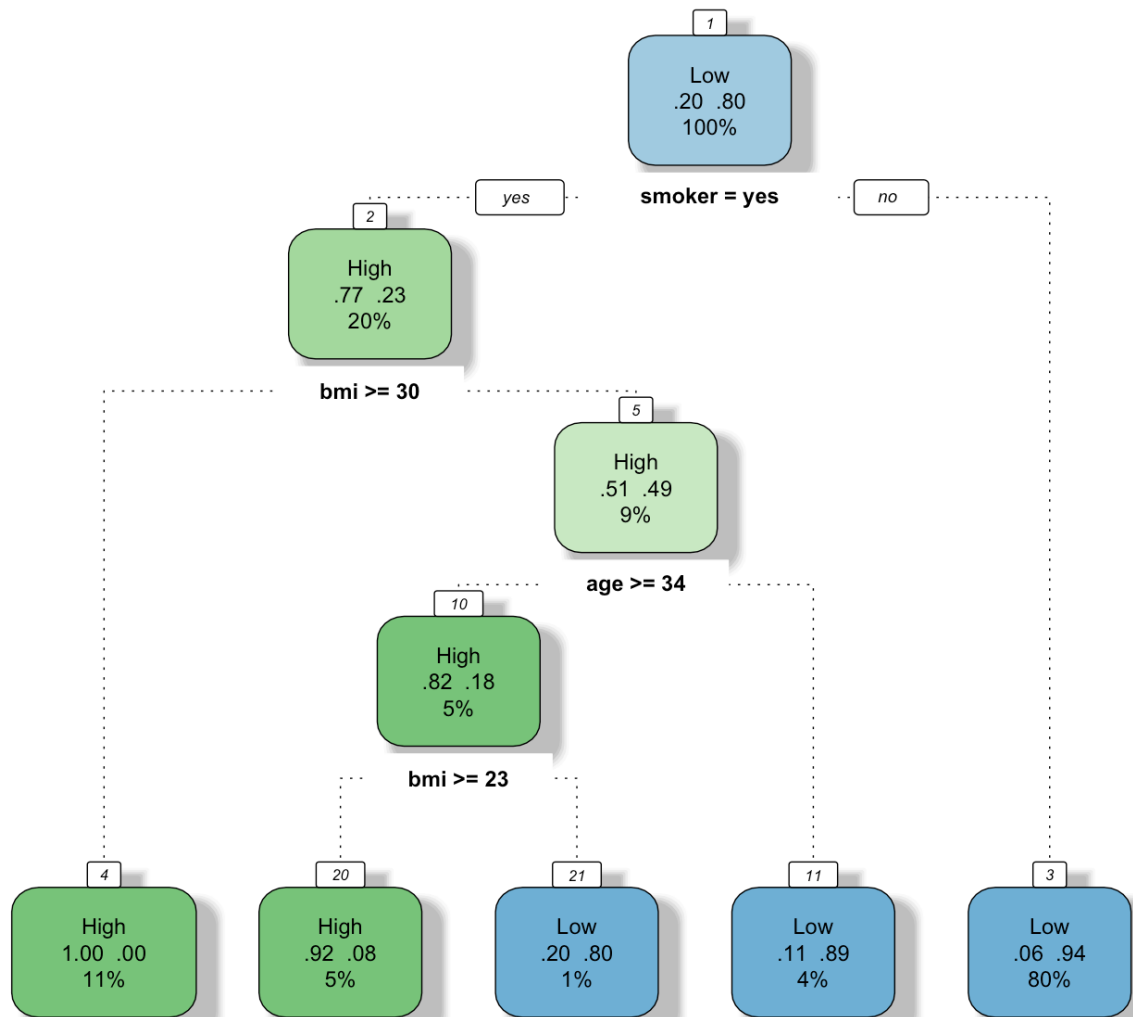


Figure 8a: Decision Tree

After that we tested our train model how accurate it is.



Test Data Science Project [Predictive Exp.] Service

## Enter data to predict

AGE

19

SEX

female

BMI

28

CHILDREN

1

SMOKER

no



✓ 'Data Science Project [Predictive Exp.]' test returned ["19","female","28","1","no","Low","0.999776542186737"]...

Figure 8b: Model Testing



Test Data Science Project [Predictive Exp.] Service

## Enter data to predict

AGE

SEX

BMI

CHILDREN

SMOKER



✓ 'Data Science Project [Predictive Exp.]' test returned ["31","male","32","3","yes","High","1.11993622340378E-05"]...

Figure 8c: Model Testing

### ***Hypothesis Testing***

From the decision tree we can see that, if a person is not smoker there are 80% of chance that he his medical insurance cost will be low. If a person is smoker but his bmi is less than 30 and age is less than 34, his medical insurance cost will be low as well.

In the second part if we look at our trained model, it gives us almost same result that our first decision tree had given. If a person 19 years of old having 1 child and his bmi is 28 with no smoking habits would get a minimal health insurance cost plan. In contrast, a 31 year aged person with three child and having 32 bmi with smoking habits would have a health insurance plan which will cost him very much.

In the crosstab analysis where we compare smoker with the scored label it showed those who has smoking habits tends to have costly insurance plan than those who does not have this bad habit.

In the Pearson and Spearman correlation test we got the relationships between variables but not the expected variable which is “charges” as this feature is not numeric. So, we cannot put the results of this tests in our hypothesis testing.

From this analysis we could easily say that there is correlation for a rise of health insurance cost due to smoking habits, BMI, age and the no. of children. There is no way to reject null hypothesis. So, people who smoke, are overweight and have many children, tend to have more health risks, thus the risk of getting hospitalized is higher compared to others and that would cost a higher health insurance plan.

### ***Data Analytics Tools***

- **R-Programming language**

R is a programming language and free software environment for statistical computing and graphics. The purpose of using this tool is to compare values.

- **Microsoft Azure**

Azure is a cloud-based platform for building, deploying, and managing services and applications, anywhere. The purpose of using this tool is to clean data, replace missing data, process data, implement machine learning algorithms and predict result.

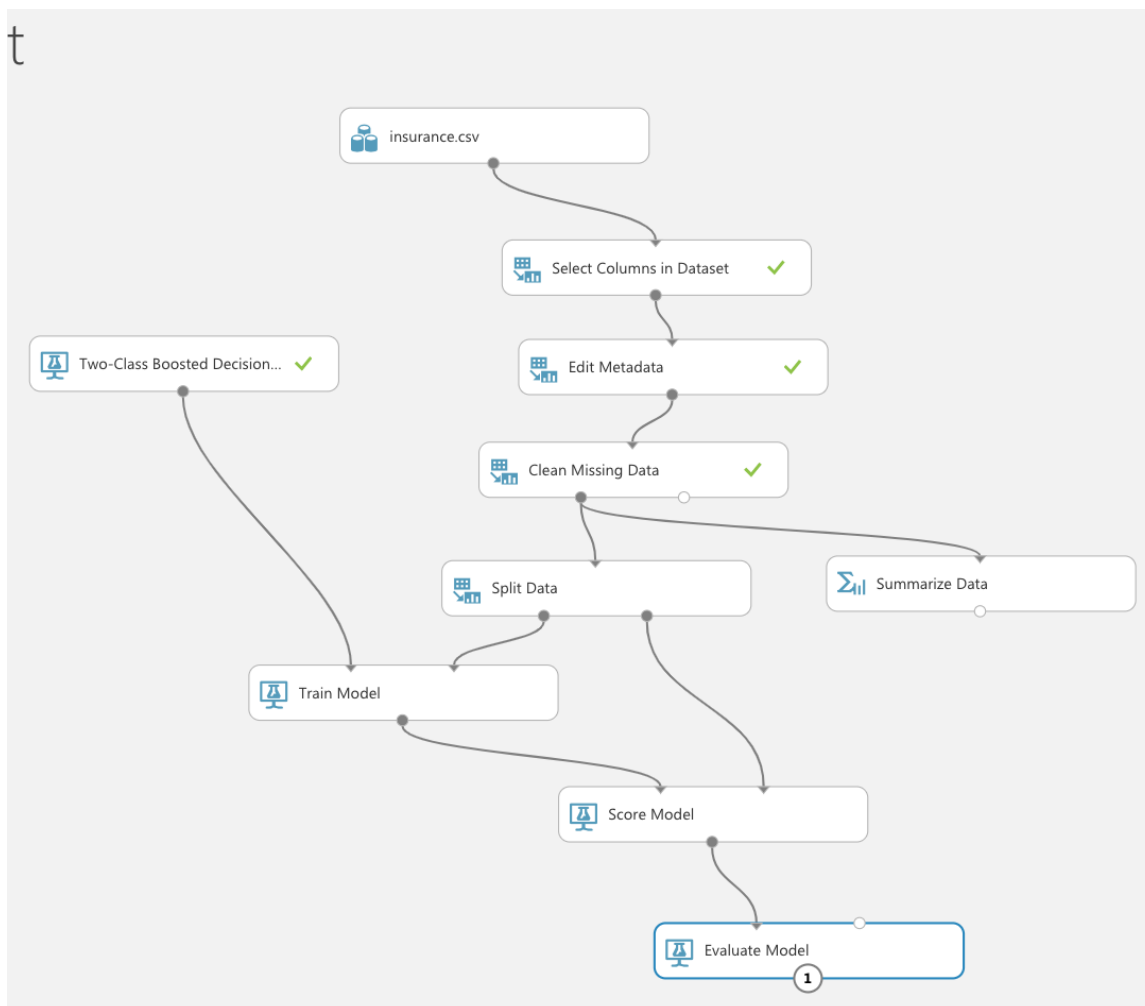
## Expected Challenges or Limitations

- To use data analytic tools accordingly.
- Lack of machine learning knowledge and implementation.
- Limited access to online dataset.

## Appendix

### ➤ Training Experiment

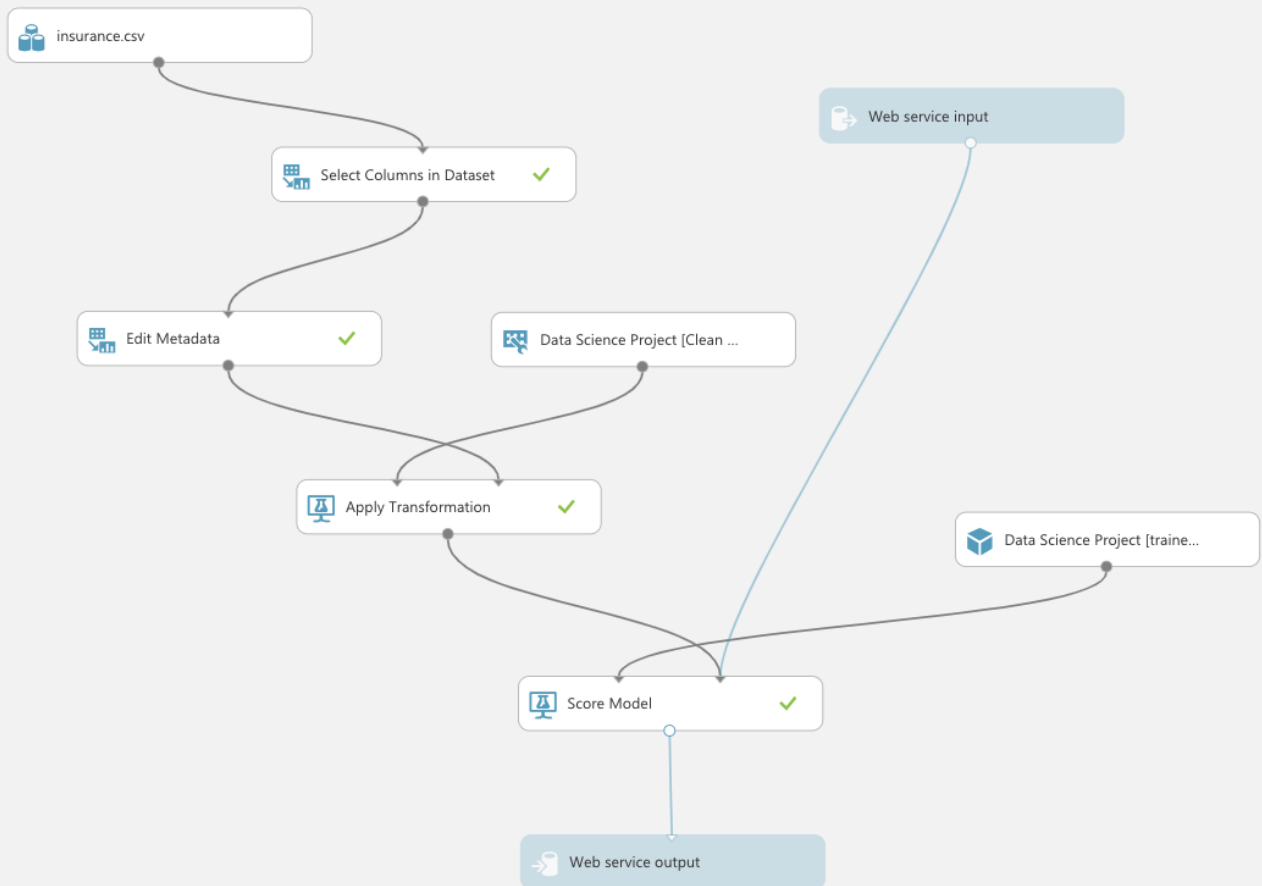
t





➤ **Web Deployed Predictive Experiment**

## Predictive Exp.]



## References

1. Botkin, K. (n.d.). 10 Factors That Affect Your Health Insurance Premium Costs. Retrieved from <https://www.moneycrashers.com/factors-health-insurance-premium-costs/>
2. HealthMarkets. (2016). What You Need to Know About Smoking and Health Insurance. Retrieved from <https://www.healthmarkets.com/content/what-you-need-know-about-smoking-and-health-insurance>
3. Mercadante, K. (2013, March 15). How Smoking Affects Your Health Insurance Premiums. Retrieved from <https://www.mcmha.org/smoking-affects-health-insurance-premiums/>