MEDICAL INSURANCE PREDICTION

INTRODUCTION

OVERVIEW

The net protections premiums over the world are expanding day by day. But, most of these costs can be anticipated fair by dispensing with smoking and bringing down BMI (Body Mass List) by few focuses. By utilizing AI and ML, we offer assistance clients to get it how much smoking increments their premium by anticipating how much they will got to pay inside seconds. Thus, the clients can see the make exceptional way of life choices make on their protections charges.

PURPOSE

The purpose of this project is to predict the insurance premium of a person by analyzing his lifestyle choices and making him aware of the impact of more smoking ,Unbalanced BMI in his life. It consider the impacts of smoking, BMI, sexual orientation and locale to decide how much these variables can account for our increase/decrease in protections premium.It consider the impacts of smoking, BMI, sexual orientation and locale to decide how much these variables can account for our increase/decrease in protections premium.

LITERATURE SURVEY

EXISTING PROBLEM

The problem of this project is to consider the effects of smoking,BMI,gender and region to determine how much these factors can account for our increase/decrease in insurance premium.

PROPOSED SOLUTION

Here, in this project I used LINEAR REGRESSION technique which is one among the regression techniques which come under the supervised learning where the model is getting trained on a labeled dataset. In classification, learning algorithms takes the input data and map the output to a discrete output like True or False In regression, learning algorithms maps the input data to continuous output like weight, cost, etc.

THEORITICAL ANALYSIS

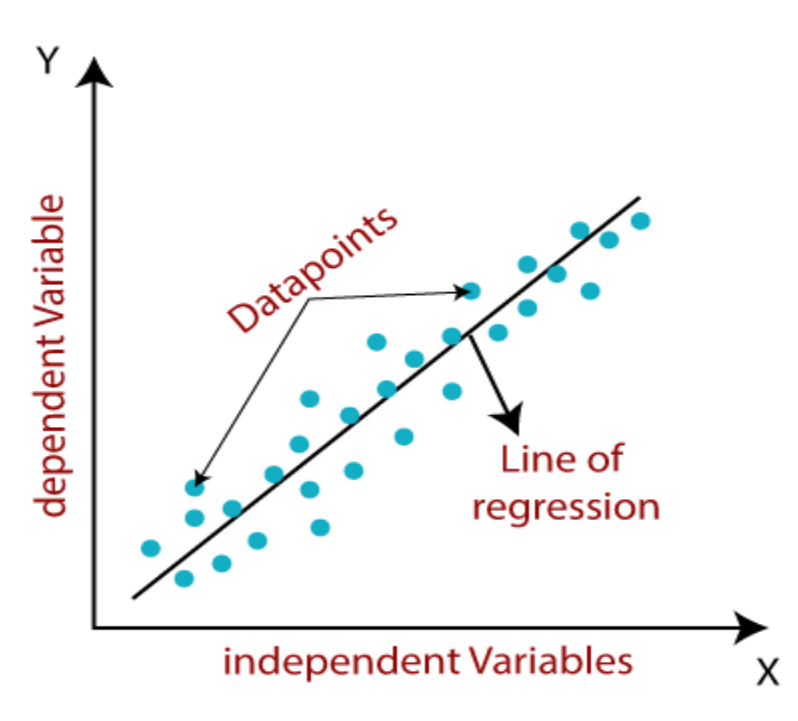
Training has to be done first with the data associated. By filtering and various machine

learning models accuracy can be improved.

Types of Machine Learning

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.
Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.
The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



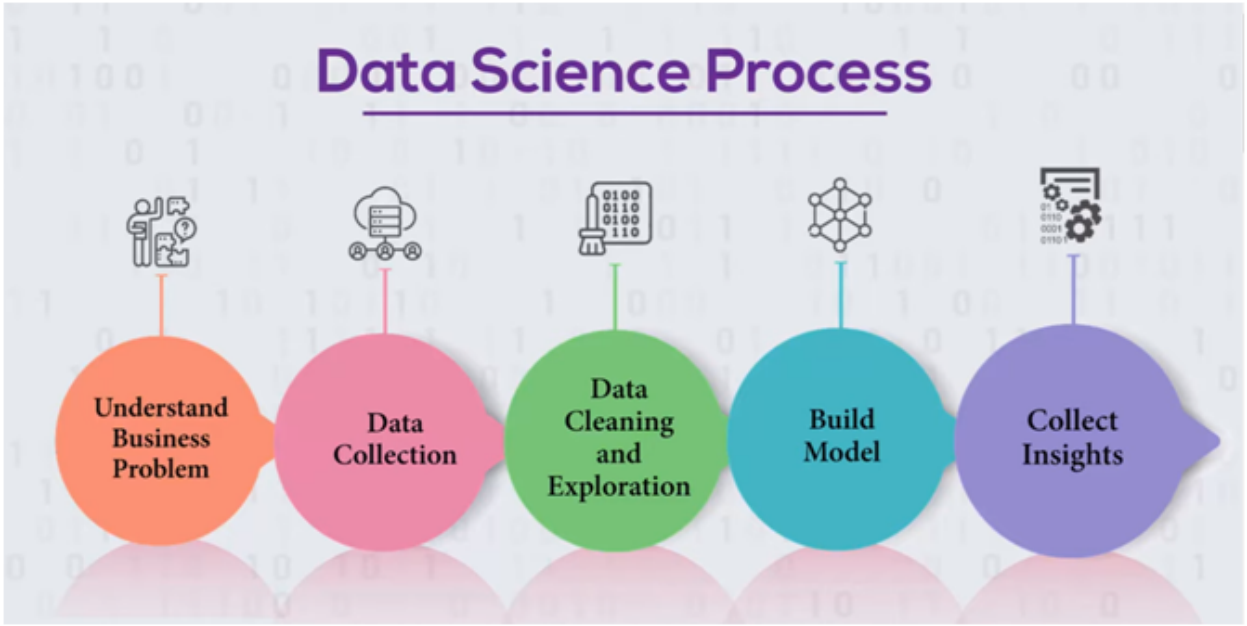
Mathematically, we can represent a linear regression as:

$$y= a_0+a_1x+ \varepsilon$$

- Y= Dependent Variable (Target Variable)
- X= Independent Variable (predictor Variable)
- a0= intercept of the line (Gives an additional degree of freedom)
- a1 = Linear regression coefficient (scale factor to each input value).
- ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

BLOCK DIAGRAM



HARDWARE/SOFTWARE DESIGNING:

Hardware requirements :

> Processor: INTEL core i5 Processor – 4 core processor, 4.20 GHz Turbofrequency, 6 MB intel smart cache.

> 8 GB memory; 1 TB hard disk drive

> 12 GB DDR4-2933 SDRAM (1 x 4 GB, 1 x 8 GB)

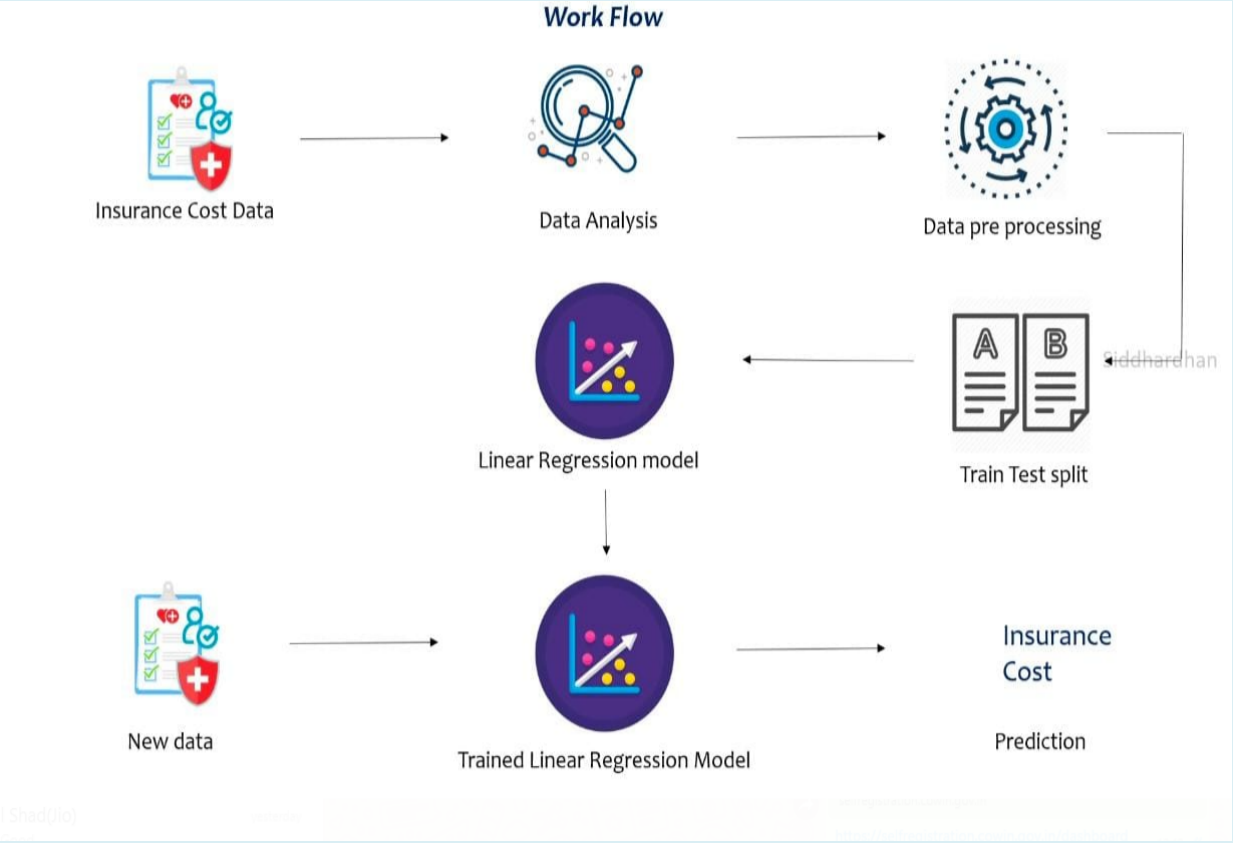
> Intel Heatsink to keep temperature under control.

Software requirements:

Eclipse, JDK, Weka.

- JDK – Java Development Kit
- Eclipse – It is Integrated Development Environment to run java programs .
  The java code in eclipse is compared with the weka software and results obtained are compared and checked.
- Weka - data mining Tool(build using java).
  open Source GUI
  The dataset can be compiled without writing any java code as it contains inbuilt functions to perform all the activities.
- Microsoft Excel - Using this we can organise,format and calculate data with formuls and is useful for data analysis for machine learning.

FLOWCHART:



Step 1: Collection of Data

- Resource for Dataset: https://www.kaggle.com/mirichoi0218/insurance.

Dataset consists of 1338 records. Each record contains the below data for specific person. The data was in structured format and was stores in a csv file. Dataset is not suited for the regression to take place directly. So cleaning of dataset becomes important for using the data under various regression algorithms.

Columns

- **Age:** age of primary beneficiary
- **Sex:** insurance contractor gender, female, male
- **BMI:** Body mass index, providing an understanding of body, weights that are relatively high or low relative to height,
  objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- **children:** Number of children covered by health insurance / Number of dependents
- **smoker:** Smoking
- **Region:** the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- **charges:** Individual medical costs billed by health insurance

Data Wrangling:

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues.

Stages of Data Preprocessing are:

- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation

Analyze Data:

Data visualization is a technique that uses an array of static and interactive visuals within a specific context to help people understand and make sense of large amounts of data. The data is often displayed in a story format that visualizes patterns, trends and correlations that may otherwise go unnoticed.

Data Visualization:

DATA VISUALIZATION Can be done by HISTOGRAMS , SCATTER PLOT , BOX TRACE etc..These histograms, boxplot are represented in both eclipse and weka. In weka the graphs are ploed for each and evary individual aribute.The dataset has 7 ATTRIBUTES/VARIABLES and 1338 INSTANCES .

We have two types of variables available in this dataset.

They are : Continuous variable

Categorical variable.

They are also refered as Numeric or Nominal Type.

RESULT:

Fig 1: This figure shows aver all over view of the data .

---gives no.of rows and columns present.

---displays first 7 rows of dataset

---displays last 7 rows of dataset

---structure or type of variable

---all mathemativcal calculations

```
<terminated> DataAnalysis (3) [Java Application] C:\Program Files\Java\jdk-12.0.2\bin\javaw.exe  (May 5, 2021, 10:18:32 AM – 10:19:02 AM)
Data Analysis
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
1338 rows X 7 cols
                                insurance.csv
 age  |   sex    |   bmi    | children  | smoker  |   region    |    charges    |
-------------------------------------------------------------------------------
  19  |  female  |    27.9  |         0 |     yes |  southwest  |    16884.924  |
  18  |    male  |   33.77  |         1 |      no |  southeast  |    1725.5523  |
  28  |    male  |      33  |         3 |      no |  southeast  |     4449.462  |
  33  |    male  |  22.705  |         0 |      no |  northwest  |  21984.47061  |
  32  |    male  |   28.88  |         0 |      no |  northwest  |    3866.8552  |
  31  |  female  |   25.74  |         0 |      no |  southeast  |    3756.6216  |
  46  |  female  |   33.44  |         1 |      no |  southeast  |    8240.5896  |
                               insurance.csv
 age  |   sex    |   bmi    | children  | smoker  |   region    |    charges    |
-------------------------------------------------------------------------------
  23  |  female  |   33.4   |         0 |      no |  southwest  |  10795.93733  |
  52  |  female  |   44.7   |         3 |      no |  southwest  |    11411.685  |
  50  |    male  |  30.97   |         3 |      no |  northwest  |  10600.5483   |
  18  |  female  |  31.92   |         0 |      no |  northeast  |    2205.9808  |
  18  |  female  |  36.85   |         0 |      no |  southeast  |    1629.8335  |
  21  |  female  |   25.8   |         0 |      no |  southwest  |     2007.945  |
  61  |  female  |  29.07   |         0 |     yes |  northwest  |   29141.3603  |
      Structure of insurance.csv
 Index  |  Column Name  |  Column Type  |
----------------------------------------
    0   |          age  |       INTEGER |
    1   |          sex  |        STRING |
    2   |          bmi  |        DOUBLE |
    3   |     children  |       INTEGER |
    4   |       smoker  |        STRING |
    5   |       region  |        STRING |
    6   |      charges  |        DOUBLE |
                                                    insurance.csv
 Summary  |         age          |  sex  |         bmi          |       children       |  smoker  |  region  |       charges
-------------------------------------------------------------------------------------------------------------------------------------
   Count  |              1338    |  1338 |              1338    |              1338    |    1338  |    1338  |              1338
     sum  |             52459    |       |  41027.624999999985  |              1465    |          |          |    17755824.990759
    Mean  |  39.20702541106125   |       |  30.663396860986524  |  1.0949177877429015  |          |          |  13270.422265141255
     Min  |                18    |       |             15.96    |                 0    |          |          |           1121.8739
     Max  |                64    |       |             53.13    |                 5    |          |          |          63770.42801
   Range  |                46    |       |             37.17    |                 5    |          |          |  62648.554110000005
Variance  |  197.40138665754355  |       |  37.18788360977321   |  1.4532127456669055  |          |          |  146652372.15285477
Std. Dev  |  14.049960379216147  |       |  6.098186911679012   |  1.205492739781914   |          |          |  12110.011236693992
  Unique  |                      |     2 |                      |                      |       2  |       4  |
     Top  |                      |  male |                      |                      |      no  |  southeast  |
Top Freq.  |                      |   676 |                      |                      |    1064  |     364  |
```

BOXPLOT :

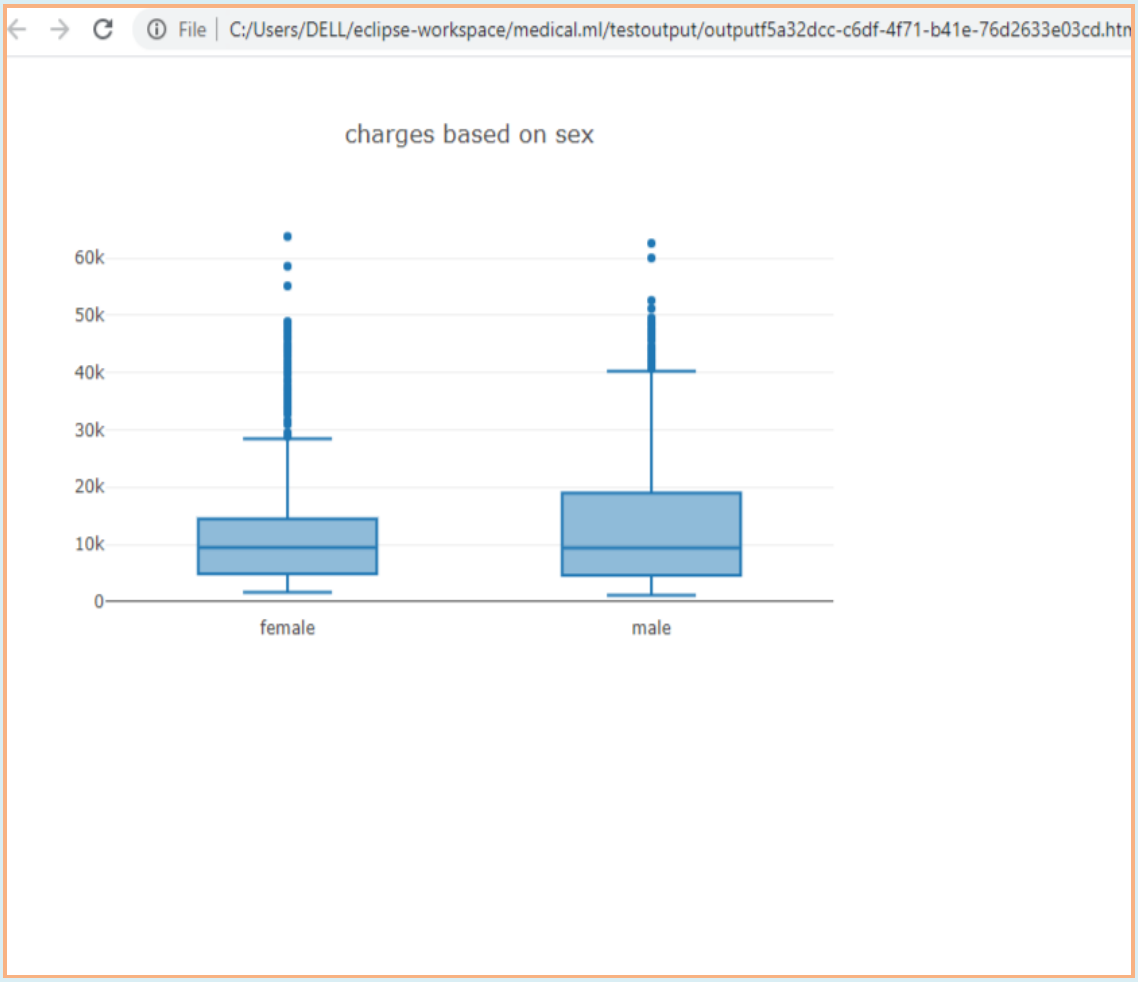
Boxplot is used only for catogorical columns.

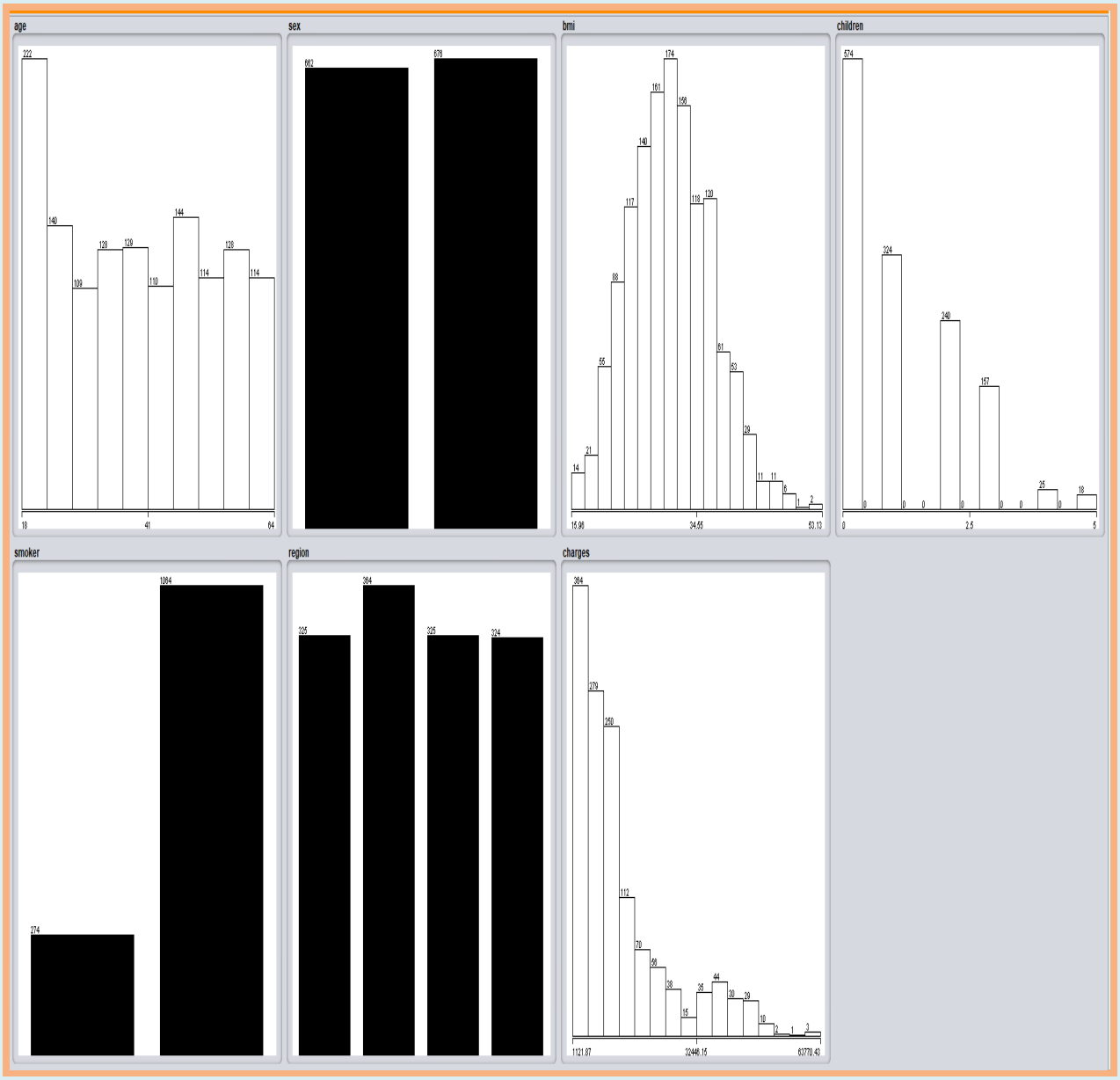
Box plots (also called **box-and-whisker plots** or **box-whisker plots**) give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data. A box plot is constructed from five values: the minimum value, the first quartile, the median, the third quartile, and the maximum value. We use these values to compare how close other data values are to them.

Box plot for categorical column name smoking is as follow:

- which is used to describe about the smoking activity of people.

Box plot for categorical column sex is as follow:

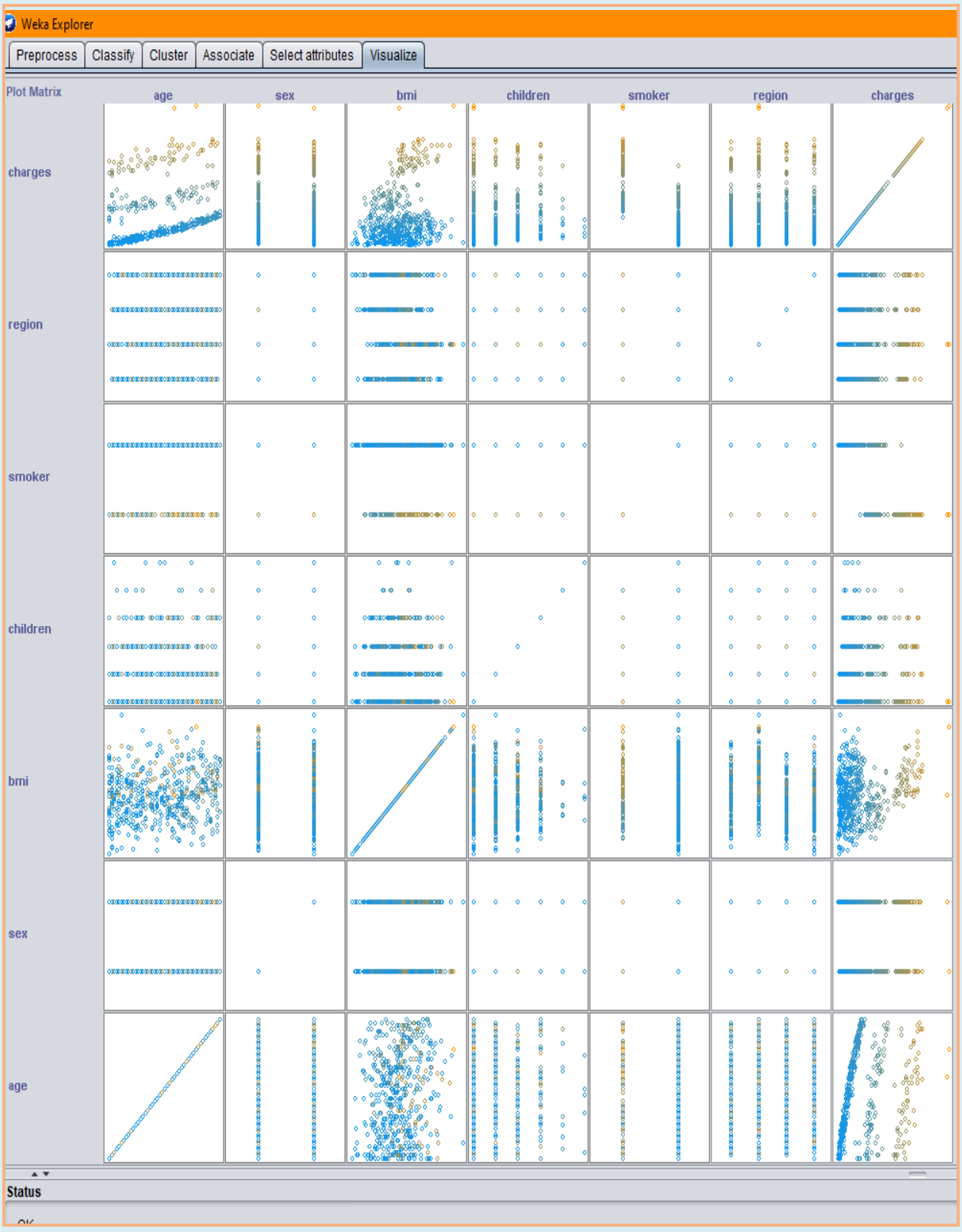




Scatter plots :

Scatter plots primary uses are to observe and show relationships between two numeric variables. The dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole. Identification of correlational relationships are common with scatter plotS

> The following shows the scatter ploting among various attributes present in dataset.


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
age
sex
bmi
children
smoker
region
charges
Status

- The Dataset has 7 attributes with 1388 instances. As mentioned earlier each attribute has its own Data visualization based on its data.
- By using the training set, we predicts the charges.
- As in linear regression we express everything in the form

$$Y=mX+c$$

The predicted charges would also be in the Y=mX+c form and they are as follows :
The equation if of the form: y=mx+c i.e ,

charges = (257.0064 * age) +
(338.6413 * bmi) +
(471.5441 * children)+
(23843.8749 *smoker=yes) +
(782.7452 * region=northwest,northeast,southeast) +
(-858.4696 * region=southeast) +
(-12948.1277)

- For each and every row , we obtained an actual and predicted value along with the error value in weka

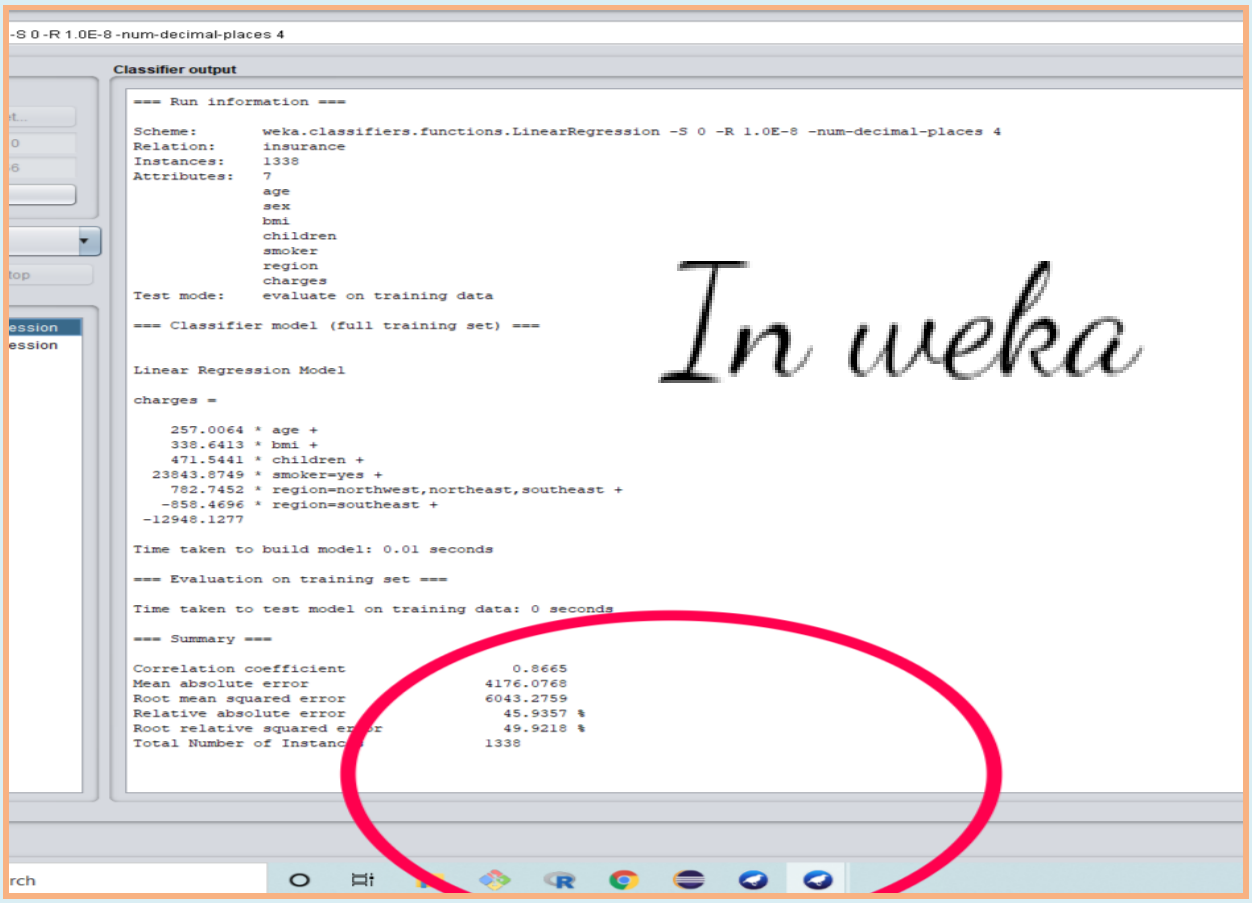
```
Time taken to build model: 0.01 seconds

=== Predictions on training set ===

    inst#     actual  predicted      error
        1  16884.924  25226.962   8342.038
        2   1725.552   3509.725   1784.173
        3   4449.462   6762.123   2312.661
        4  21984.471   4004.68  -17979.791
        5   3866.855   5838.784   1971.929
        6   3756.622   3659.974    -96.648
        7   8240.59   10594.152   2353.563
        8   7281.506   8152.397    870.891
        9   6406.411   8388.613   1982.203
       10  28923.137  12005.493 -16917.644
       11   2721.321   3138.953    417.632
       12  27808.725  35657.3     7848.575
       13   1826.843   4612.281   2785.438
       14  11090.718  14853.204   3762.486
       15  39611.758  32026.155  -7585.603
       16   1837.237    737.115  -1100.122
       17  10797.336  12093.874   1296.538
       18   2395.172   1820.667   -574.504
       19  10602.385  15091.476   4489.091
       20  36837.467  30559.978  -6277.489
       21  13228.847  15447.782   2218.935
       22   4149.736   6205.587   2055.851
       23   1137.011   3149.932   2012.921
       24  37701.877  31697.685  -6004.191
```

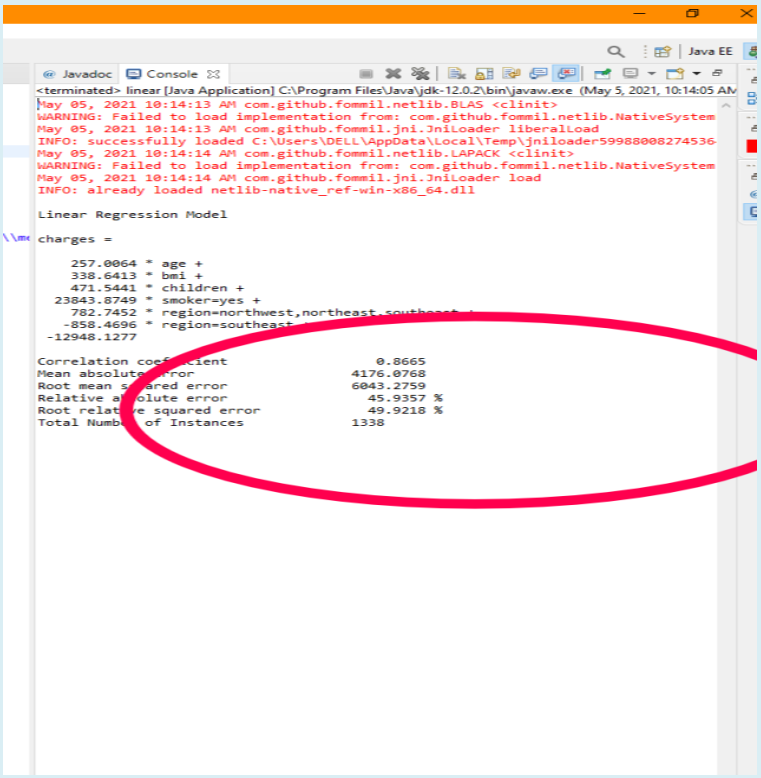
- Since there are a number of instances available , we can obtain a single instance in Eclipse by the following code:

```
System.out.println(lreval.predictions().get(125));
```

The above code returns the Actual and Predicted value of instance 125 in Dataset. As indexing starts from 0 in java , get(125) returns the 126 th instance values….

- The overall Summary of the test data is also obtained both in eclipse and Weka. The final corelation coefficient , root mean square value etc.. all are obtained as follows:





ADVANTAGES & DISADVANTAGES

Early health insurance amount prediction can help in better contemplation of the amount needed. Where a person can ensure that the amount he/she is going to opt is justified. Also it can provide an idea about gaining extra benefits from the health insurance.

Our project does not give the exact amount required for any health insurance company but gives enough idea about the amount associated with an individual for his/her own health insurance.

APPLICATIONS

Project allows a person to get an idea about the necessary amount required according to their own health status. Later they can comply with any health insurance company and their schemes & benefits keeping in mind the predicted amount from our project. This can help a person in focusing more on the health aspect of an insurance rather than the futile part.

Health insurance is a necessity nowadays, and almost every individual is linked with a government or private health insurance company. Factors determining the amount of insurance vary from company to company. Also people in rural areas are unaware of the fact that the government of India provides free health insurance to those below poverty line. It is very

complex method and some rural people either buy some private health insurance or do not invest money in health insurance at all. Apart from this people can be fooled easily about the amount of the insurance and may unnecessarily buy some expensive health insurance.

CONCLUSION

Various factors were used and their effect on predicted amount was examined. It was observed that a person's age and smoking status affects the prediction most in every algorithm applied. Attributes which had no effect on the prediction were removed from the features.
The effect of various independent variables on the premium amount was also checked. The attributes also in combination were checked for better accuracy results.

FUTURE SCOPE

Premium amount prediction focuses on persons own health rather than other companies insurance terms and conditions. The models can be applied to the data collected in coming years to predict the premium. This can help not only people but also insurance companies to work in tandem for better and more health centric insurance amount.

BIBILOGRAPHY

- https://www.kaggle.com/mirichoi0218/insurance
- https://waikato.github.io/weka-wiki/

APPENDIX

SOURCE CODE:

☞ https://github.com/eegabhavani/Medical_Insurance_prediction