

MEDICAL INSURANCE PREDICTION

INTRODUCTION :

OVERVIEW :

The gross insurance premiums across the world are increasing day by day. But, most of these costs can be prevented just by eliminating smoking and lowering BMI (Body Mass Index). In this project we study and perform analysis on age, smoker, gender, BMI, region and how much difference they make on a person Medical Insurance Premium. Hence, the customers can see the make drastic lifestyle choices make on their insurance 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. By this Project we make a Person understand that by how much smoking he make can increase/decrease his Medical insurance Premium by predicting the data.

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 :

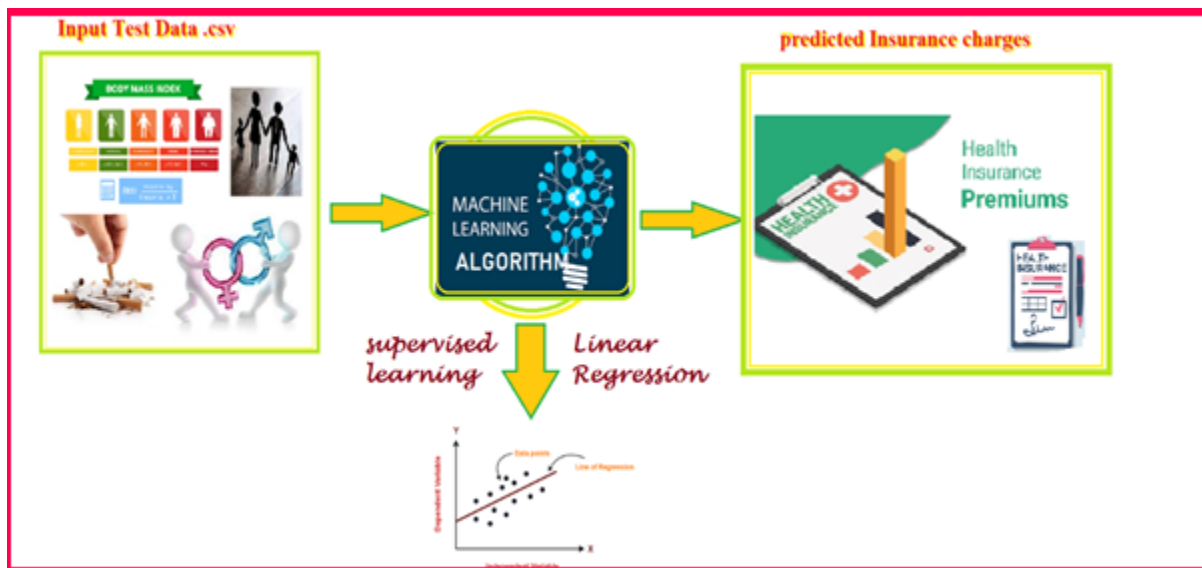
The solution is by implementing the model using LINEAR REGRESSION. In Linear regression algorithm, it shows a linear relationship between a dependent/predicted value (y) and one or more independent/training data (X) variables.

- ▲ The dependent variable in this project refers to "charges "
- ▲ The independent variables refers to age, Bmi, smoker, children, region.

Hence, by determining the PREDICTED Value of the data, we can determine the insurance that needs to be paid by a customer.

THEORITICAL ANALYSIS :

Block diagram :



BLOCK DIAGRAM DESCRIPTION:

Step1: The Machine model takes the input a dataset/csv file where the dataset/csv file contains different columns like

- ⇒ 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,
- ⇒ 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.

Step2: MachineLearning Algorithm

An appropriate algorithm is applied to the Dataset and necessary outputs are obtained. In prediction of Medical insurance premium we use: LINEAR REGRESSION.

- ⇒ In linear regression we calculate data and obtain results in the form of $y=mx+c$
- ⇒ The equation for Medical insurance premium can be referred as:

$$\text{CHARGES} = \text{BASECOST} + m * \text{ATTRIBUTE}$$

By applying the above algorithm the actual and predicted values are calculated along with the error.

Step3 : RESULT

By comparing actual and predicted values, we can obtain the detailed information about the project and the difference/errors in the charges in Medical Insurance Premium.

HARDWARE OR SOFTWARE REQUIREMENTS :

HARDWARE REQUIREMENTS :

- Processor: INTEL core i5 Processor – 4 core processor, 4.20 GHz, Turbo frequency, 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 Heat sink to keep temperature under control

SOFTWARE REQUIREMENTS :

1. WEKA :

weka stands for **WAIKATO ENVIRONMENT FOR KNOWLEDGEANALYSIS**.

- a. It is a machine learning software that is written in Java.
- b. It consists of several machine learning algorithms and it is an open source GUI.
- c. The dataset can be compiled without writing any java code as it contains inbuilt functions to perform all the activities.

2.Eclipse :

- Eclipse is an integrated development environment (IDE) used in computer programming.
- The java code in eclipse is compared with the weka software and results obtained are compared and checked.

3.Microsoft Excel :

- Through Microsoft excel we can organize, format and calculate data with formulas and is useful for data analysis for machine learning.

EXPERIMENTAL INVESTIGATIONS :

STEP1 : EXPLORATORY DATA ANALYSIS :

1. In Exploratory data analysis, we will learn about the number of different techniques used to understand the dataset(" insurance.csv") which is being used.
2. Resource for Dataset: <https://www.kaggle.com/mirichoi0218/insurance>
3. In EDA , we will :
 - a. 1.understand the types of variables in dataset.
 - b. 2. perform Data cleaning
4. The type of variable can be understood by performing different operations in eclipse.
The head, shape, summary about the data set can be obtained by the following code.

```
public class DataAnalysis {  
    public static void main(String args[]){  
        System.out.println("data Analysis");  
        try {  
            Table insurance_data = Table.read().csv("C:\\Users\\srivatsav\\eclipse-workspace\\org.ml\\src\\main\\java\\org\\ml\\insurance.csv");  
  
            System.out.println(insurance_data.shape());           //displays number of rows and columns  
  
            System.out.println(insurance_data.first(7));          // displays first 7 rows of dataset  
  
            System.out.println(insurance_data.last(7));           // displays last 7 rows of dataset  
  
            System.out.println(insurance_data.structure());        //structure or type of variable  
  
            System.out.println(insurance_data.summary());          //all the mathematical calculations like range, min etc.  
  
        } catch (IOException e){    e.printStackTrace();    }  
    }  
}
```

output :

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

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	

STEP2 : DATA PREPROESSING :

In the insurance.csv dataset, since there is no missing data available, there is no need to perform any data cleaning or integration or etc. Hence, the dataset is ready to perform data visualization for all the available attributes in data set.

STEP3 : DATA VISUALIZATION :

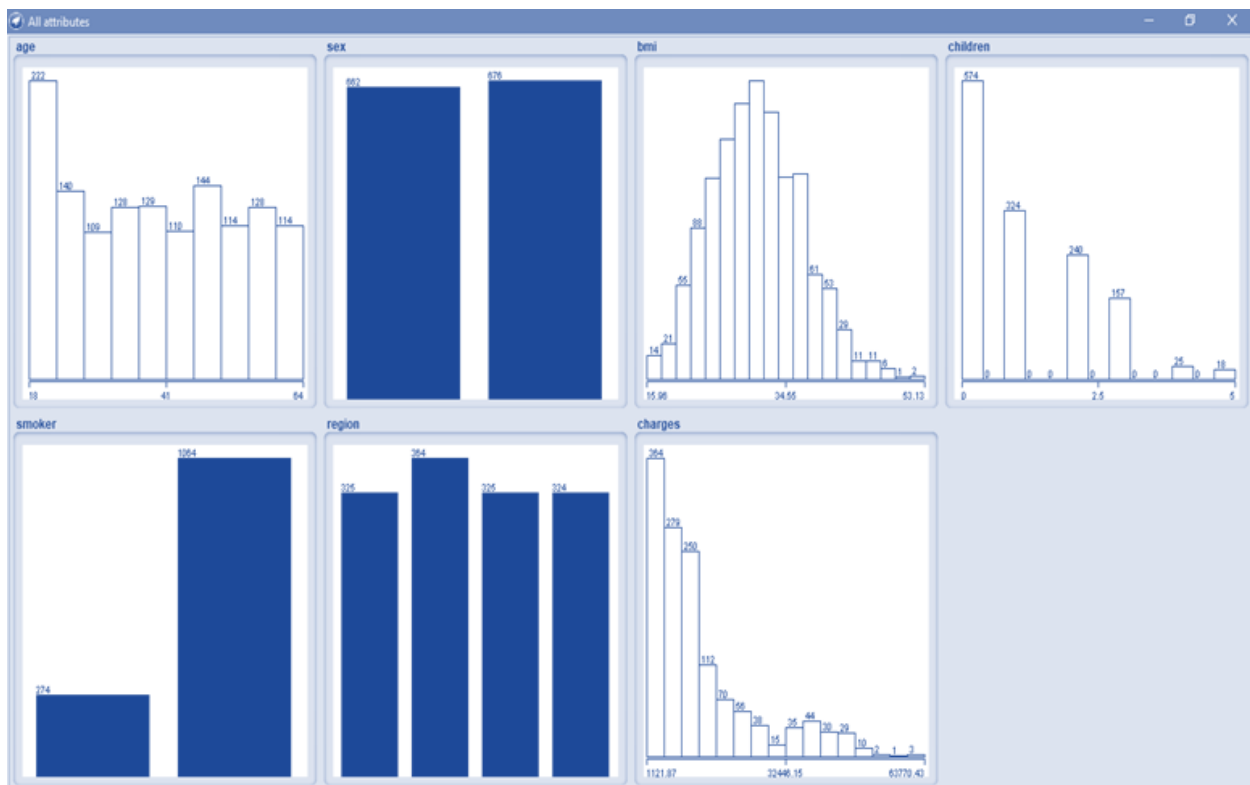
DATA VISUALIZATION Can be done by **HISTOGRAMS , SCATTER PLOT , BOX TRACE etc..** These histograms, boxplot are represented in both eclipse and weka. In weka the graphs are plotted for each and every individual attribute. The dataset has 7 ATTRIBUTES/VARIABLES and 1338 INSTANCES. We have two types of variables available in this dataset. they are :

- Continuous variable
- Categorical variable.

➤ They are also referred as Numeric or Nominal Type.

Viewer							
Relation: insurance							
No.	1: age	2: sex	3: bmi	4: children	5: smoker	6: region	7: charges
	Numeric	Nominal	Numeric	Numeric	Nominal	Nominal	Numeric
1	19.0	fem...	27.9	0.0	yes	south...	16884.9...
2	18.0	male	33.77	1.0	no	south	1725.55

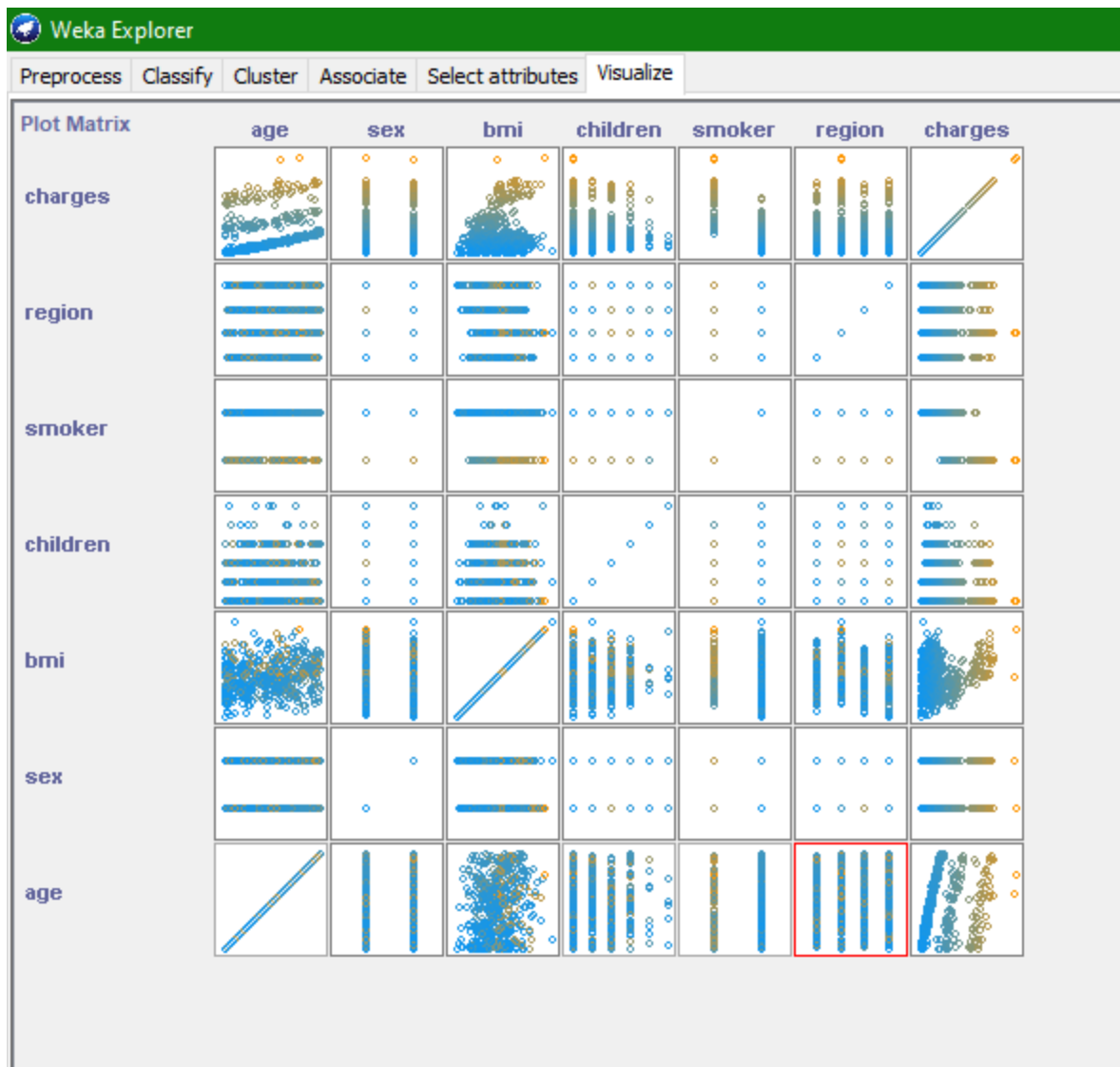
- Based on type of variable :
mean , mode , min ,max , unique , top ,top frequency , range ,variance, are calculated for each and every attribute by weka.
- Every attribute has its own mean, range etc. as mentioned in above figure.
- The data visualization based on HISTOGRAM of the attributes are shown below.
- Every attribute is measured against the charges and the histograms are represented as shown below..



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 plotting among various attributes present in dataset.

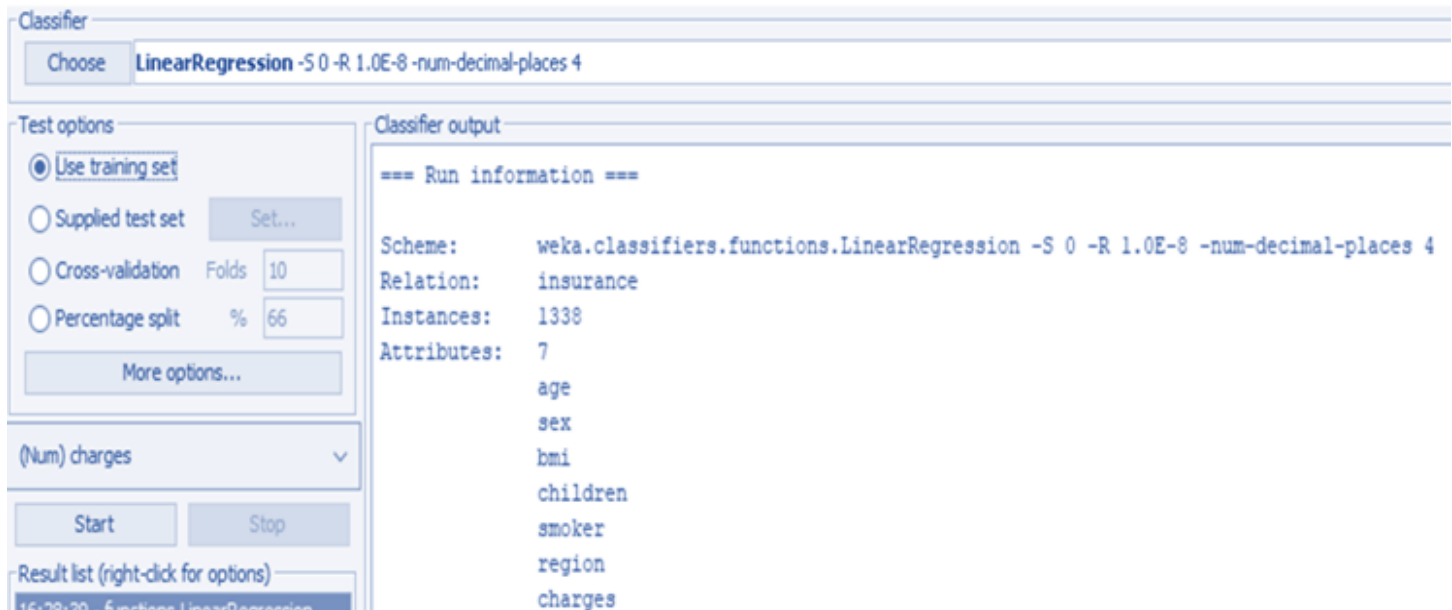


4. BUILDING AND TESTING MACHINE LEARNING MODEL :

AS we have the whole data which is cleaned and no missing values, we will build a model which can predict the charges for medical insurance premium.

- Hence a linear regression model can be used to predict the charges. Generally it is used when we want to predict the value of a variable based on the value of another variable.
- The linear regression can be done in both weka and Eclipse . where after compiling both the software produces the Predicted charges with respect to actual charges.

- The error is also obtained along with actual and predicted values. Hence for every actual value, we can obtain a predicted value which is charges.
- Since we are using linear Regression, the output equation will be of the form $y = mX + C$, where y, x are dependent and independent variables.
- In our project, we use the training data set to obtain the values.
- ◆ The process of linear Regression in weka is depicted as follows :



- 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 predict 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 is of the form: $y = mx + c$ i.e.,

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

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

Classifier output

```
=== 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
25	6203.902	7777.366	1573.464
26	14001.134	12941.295	-1059.839
27	14451.835	11843.555	-2608.28
28	12268.632	14012.027	1743.395
29	2775.192	104.588	-2670.604

- ★ 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(12));
```

OUTPUT :

```
May 07, 2021 5:34:20 PM com.github.fommil.jni.JniLoader l
INFO: already loaded netlib-native_ref-win-x86_64.dll
NUM: 1826.843 4612.281233075533 1.0
```

The above code returns the Actual and Predicted value of instance 12 in Dataset. As indexing starts from 0 in java , get(12) returns the 13th 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 :

```
May 07, 2021 5:41:28 PM com.github.fommil.jni.JniLoader load
INFO: already loaded netlib-native_ref-win-x86_64.dll

Correlation coefficient          0.8665
Mean absolute error             4176.0768
Root mean squared error         6043.2759
Relative absolute error         45.9357 %
Root relative squared error     49.9218 %
Total Number of Instances      1338
```

In eclipse

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.63 seconds

=== Summary ===

Correlation coefficient          0.8665
Mean absolute error             4176.0768
Root mean squared error         6043.2759
Relative absolute error         45.9357 %
Root relative squared error     49.9218 %
Total Number of Instances      1338
```

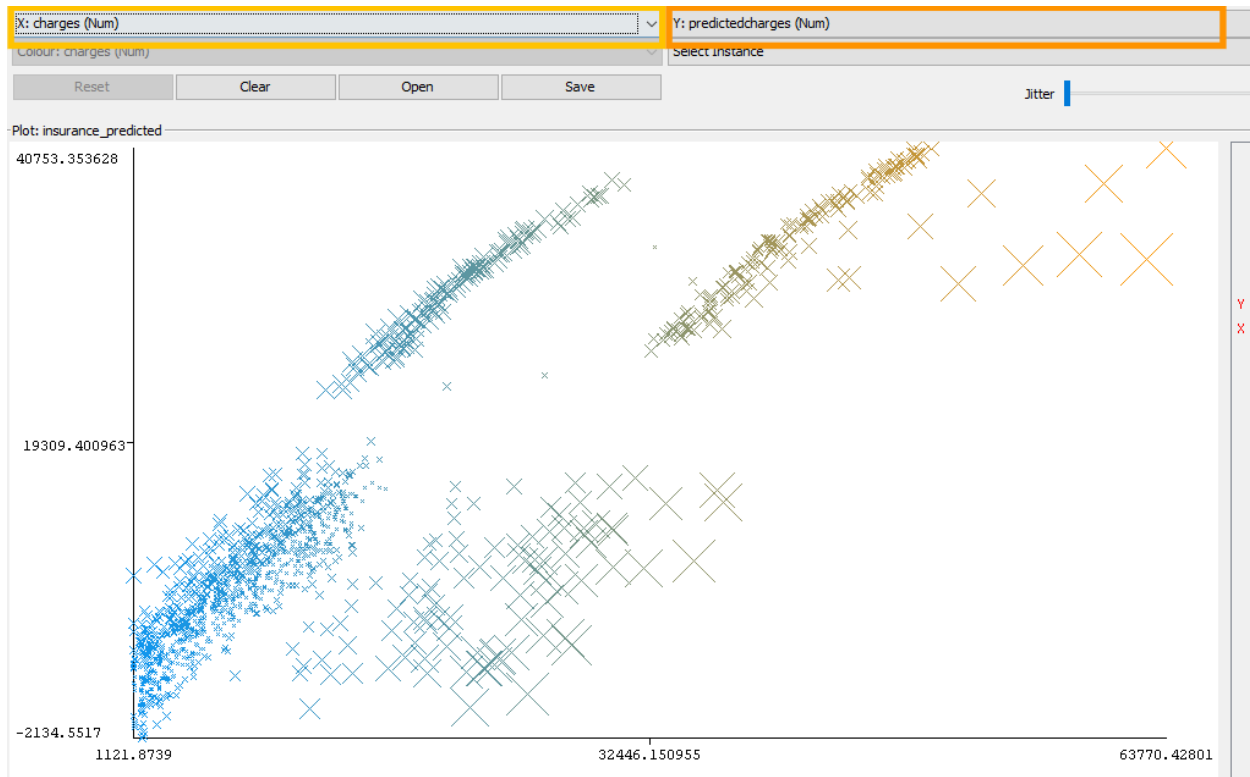
In Weka.

FLOWCHART :



RESULT :

The output of the Medical insurance prediction model is shown in the below figure . Finally we obtained actual and predicted values and obtained an equation in the form $y=mx+c$ which proves the dataset had turned it into LINEAR REGRESSION model and we obtained the predicted charges for every current charges in an diagrammatically model.



Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **LinearRegression** -S 0 -R 1.0E-8 -num-decimal-places 4

Test options

☒ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Num) charges

Start Stop

Result list (right-click for options)

21:50:26 - functions.LinearRegression

Classifier output

Test mode: evaluate on training data

=== Classifier model (full training set) ===

Linear Regression Model

charges =

```

257.0064 * age +
338.6413 * hmi +
471.5441 * children +
23843.8749 * smoker=yes +
782.7452 * region=northwest,northeast,southeast +
-858.4696 * region=southeast +
-12948.1277
    
```

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correlation coefficient	0.8665
Mean absolute error	4176.0768
Root mean squared error	6043.2759
Relative absolute error	45.9357 %
Root relative squared error	49.9218 %
Total Number of Instances	1338

Status OK

Log x 0

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 health insurance.

There are no disadvantages in checking into a medical insurance prediction. It is the basic responsibility of every customer/Person to check his Insurance and Predict the charges based on his lifestyle choices.

APPLICATIONS

There are many applications of linear regression in our day to day life. It is probably one of the most popular and well inferential algorithms in statistics.

1. Marks scored by students based on number of hours studied (ideally)- Here marks scored in exams are independent and the number of hours studied is independent.

2. Predicting crop yields based on the amount of rainfall- Yield is a dependent variable while the measure of precipitation is an independent variable.

3. Predicting the Salary of a person based on years of experience- Therefore, Experience becomes the independent while Salary turns into the dependent variable.

CONCLUSION

Various factors were used and their effect on the 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 and the charges are predicted for the

insurance premium.

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.

BIBLIOGRAPHY

- 📄 <https://www.kaggle.com/mirichoi0218/insurance>
- 📄 <https://waikato.github.io/weka-wiki/>
- 📄 https://en.wikipedia.org/wiki/Linear_regression

APPENDIX

- 📄 [Source Code- Github](#)