Medical Insurance Prediction

Introduction

Overview:

In this project for the given dataset which consists of information about clients which have taken earlier for medical insurance prediction. The task is to collect insights of data, plot the attributes and finally develop a machine learnig model which predict the charges for the given input data. Data consists of several attributes like age, sex, region, bmi etc. The charges are the class for which training is done. As this is a regression problem, here I've been using java for developing with the help of weka api. In this I have used a machine learning algorithmfor supervised learning. For this the data is split into training data and test data sets with ratio 7:3, and models are trained on the training data. Test data is used for the compute the performance of the model

Purpose:

The purpose of this project is to help the people and the insurance companies to know how the various conditions like age, bmi,smoking or not etc.,
affect the charges of the premium and also help to predict the charges for the given conditions.

Literature Survey

Existing Methods:

Charges vary from company to company and they have their parameters, charges which are made up of collected data. Here they use clients 'Demographics, daily habits any othermedical records to calculate charges. butmany don't have any solutionor integrated software which can predict the charges.

Proposed Solution:

Here we develop a machine learning model which predict the charges for the given input parameters.It also help the clients to determine the charges for existing parameters.

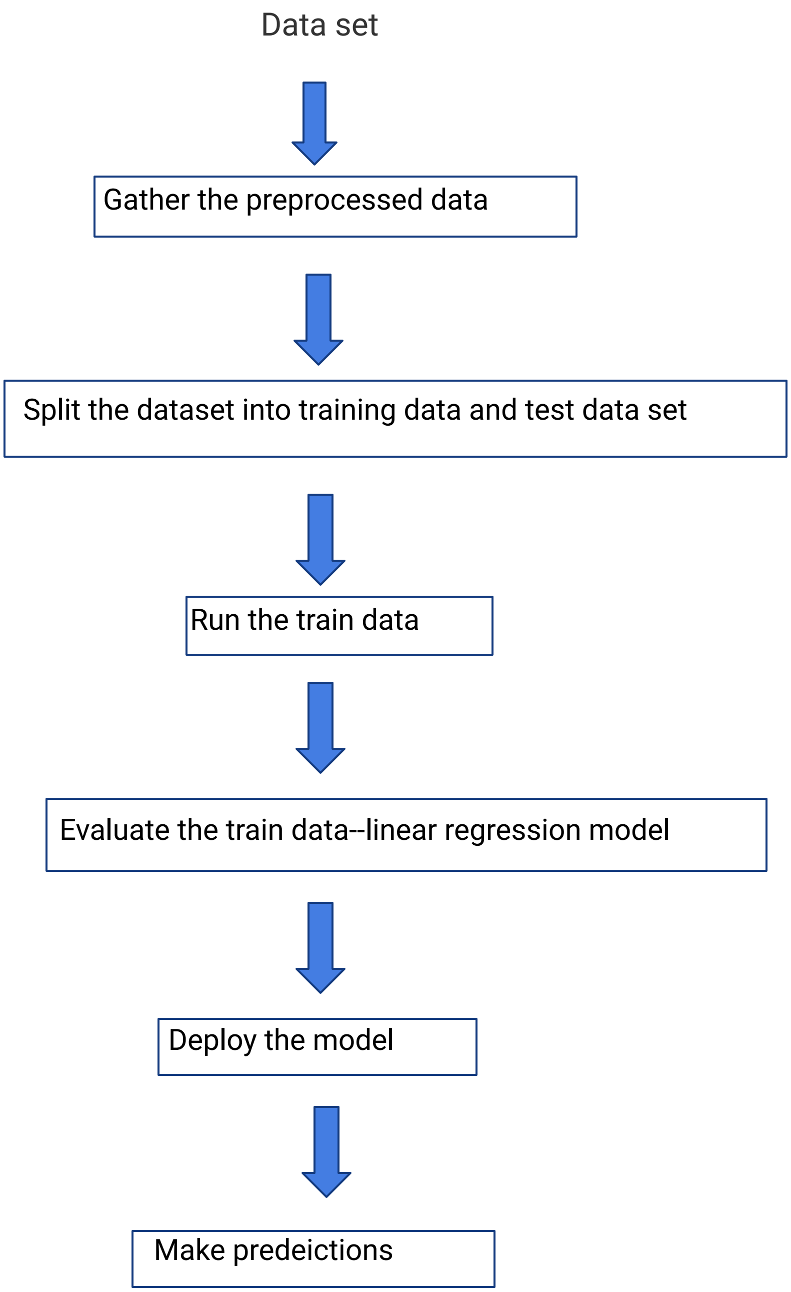
Dataset Description:

- 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

Task:

Effect of various parameters like age,bmi,smoking or not etc., to determine how much these factors can account for our increase/decrease in in surance premium.

Theoritical Analysis:


Data set
Gather the preprocessed data
Split the dataset into training data and test data set
Run the train data
Evaluate the train data--linear regression model
Deploy the model
Make predeictions

For developing the model ,the hardware requried is PC working on 64bit processor witn internet connection. Software requrirements are Java, Eclipse IDE for developing the project and Weka gui.

Experimental Investigations:

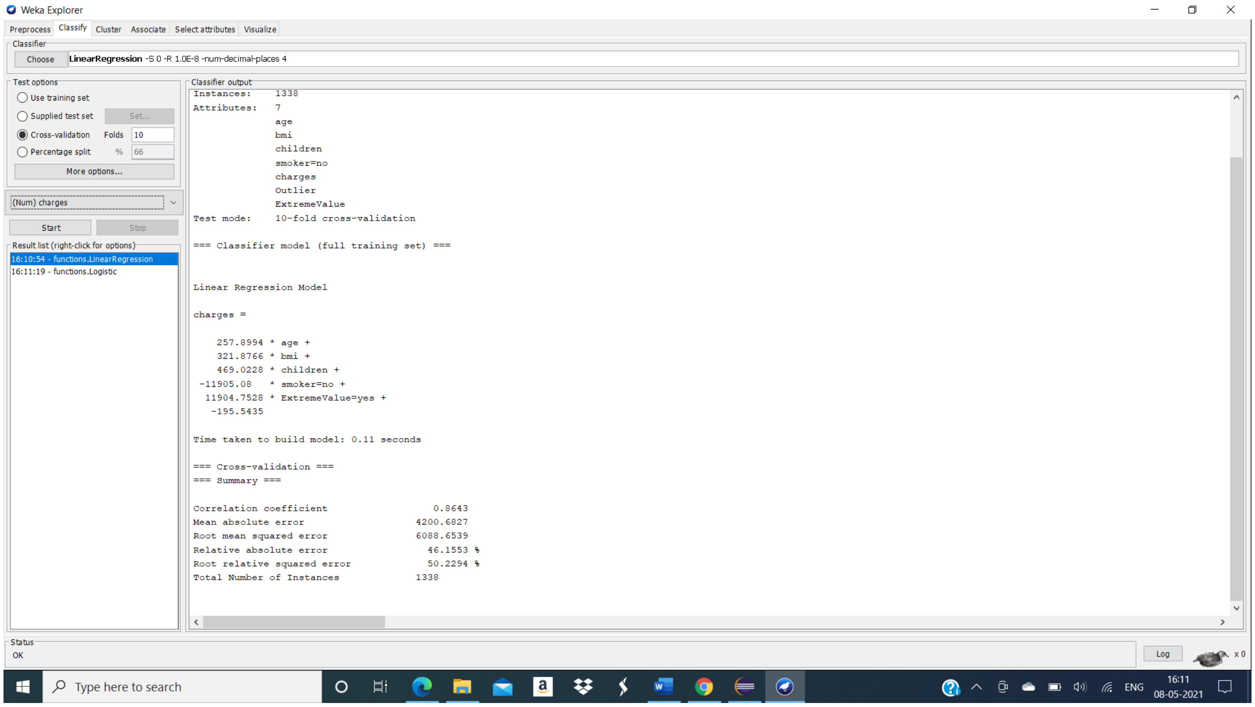
First the dataset hsa been imported from kaggle website. Dataset consists of serval attributes like sex,age, bmi etc., which in the form of csv file format.Next step is data processing which includes data cleaning, data integration and reduction and also data transformation.

The presence of missing, incomplete or corrupted data leads to wrong result while performing any functions like average,mean etc. These inconsistencies must be removed before analysing the data. This can be done by data processing.

Weka gui software, it is an open source code/platform for training various Machine learning algorithms. it is used to process and visualize the data. It also help in slipting the data . Now the data splited into train and test dataset.

Next we use eclipse IDE to build the linear regression algorithm for obtained data. Finally training and testing are done through the developed machine learing model.

Result:



Classifier: LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4

Test options:
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

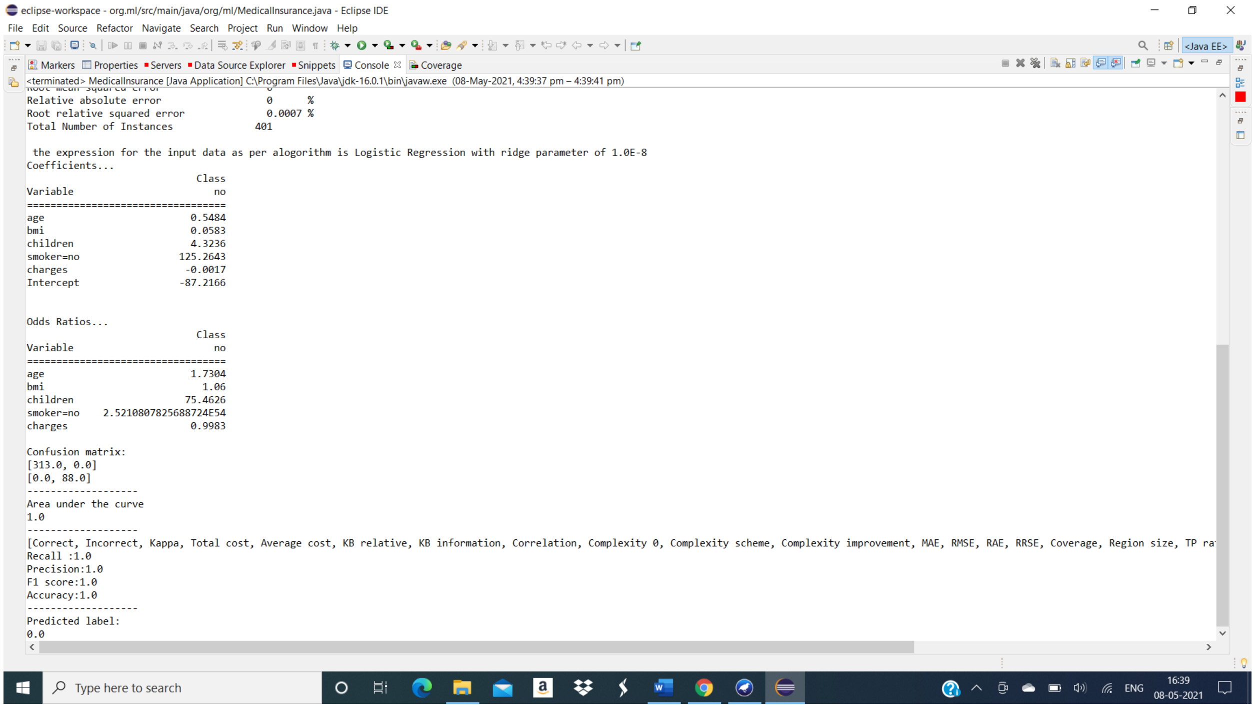
(Num) charges

Result list (right-click for options)
16:10:54 - functions.LinearRegression
16:11:19 - functions.Logistic

Classifier output

```
Instances:    1338
Attributes:   7
              age
              bmi
              children
              smoker=no
              charges
              Outlier
              ExtremeValue
Test mode:    10-fold cross-validation

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


Linear Regression Model

charges =

    257.8994 * age +
    321.8766 * bmi +
    469.0228 * children +
 -11905.08   * smoker=no +
  11904.7528 * ExtremeValue=yes +
   -195.5435

Time taken to build model: 0.11 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8643
Mean absolute error                   4200.6827
Root mean squared error               6088.6539
Relative absolute error                 46.1553 %
Root relative squared error             50.2294 %
Total Number of Instances             1338
```

Status
OK



eclipse-workspace - org.ml/src/main/java/org/ml/MedicalInsurance.java - Eclipse IDE
File Edit Source Refactor Navigate Search Project Run Window Help
Markers Properties Servers Data Source Explorer Snippets Console Coverage
Type here to search
16:39
08-05-2021

```
<terminated> MedicalInsurance [Java Application] C:\Program Files\Java\jdk-16.0.1\bin\javaw.exe  (08-May-2021, 4:39:37 pm – 4:39:41 pm)
Relative absolute error                  0      %
Root relative squared error              0.0007 %
Total Number of Instances              401

 the expression for the input data as per alogorithm is Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
                              Class
Variable                         no
==================================
age                          0.5484
bmi                          0.0583
children                     4.3236
smoker=no                  125.2643
charges                     -0.0017
Intercept                  -87.2166


Odds Ratios...
                              Class
Variable                         no
==================================
age                          1.7304
bmi                            1.06
children                    75.4626
smoker=no    2.5210807825688724E54
charges                      0.9983

Confusion matrix:
[313.0, 0.0]
[0.0, 88.0]
-------------------
Area under the curve
1.0
-------------------
[Correct, Incorrect, Kappa, Total cost, Average cost, KB relative, KB information, Correlation, Complexity 0, Complexity scheme, Complexity improvement, MAE, RMSE, RAE, RRSE, Coverage, Region size, TP ra
Recall :1.0
Precision:1.0
F1 score:1.0
Accuracy:1.0
-------------------
Predicted label:
0.0
```

Conclusion:

As we can see from the output the most factors affecting the charges of a person is the age, bmi and most significantly the habit of smoking. So, in order to reduce the charges, one must quit smoking and work on body to improve the BMI. Machine learning can be helpful for calculating the Insurance premium charges to agreat accuracy.

Future Scope:

Premium amount prediction focuses on persons own health but not on the terms and conditions of a particular company. This model is used to predict the premium. This will not only help the companies but also the clients to predict the charges for the existing parameters.

References:

https://www.kaggle.com/mirichoi0218/insurance

economictimes.indiatimes.com

saedsayad.com/decision_tree_reg.htm