Medical Insurance Prediction

Introduction

Overview:

In this project for the given data set containing information about the clients which have earlier taken the medical Insurance premium, the task is to collect insights of data, plot the attributes and finally develop a machine learning model which can predict the charges for given input .Data consists of several attributes like age, sex, bmi etc. The charges is 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 three machine learning algorithms for supervised learning. For this the data is split into test and training sets, and models are trained on the training data.Test data is used for the comparison of performance of the models.

Purpose:

This will help people and insurance companies to know how the various conditions like age, bmi, smoking or not etc affect the charges of the premium and help to predict the charges for given conditions.This could also be used to know which factors one can work on to reduce the charges.

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 other medical records to calculate charges . but many don't have any solution or integrated software which can predict the charges.

Proposed Solution:

Here we develop a Machine learning model which can predict the charges for given input parameters. This will not only help the companies but also the clients to determine the charges for existing parameters.

Dataset description:

Attributes

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of the 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 ,Yes ,No
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

Task:

The effects of various parameters like age, sex etc (from given data) to determine how much these factors can account for our increase/decrease in insurance premium.

Theoretical Analysis:

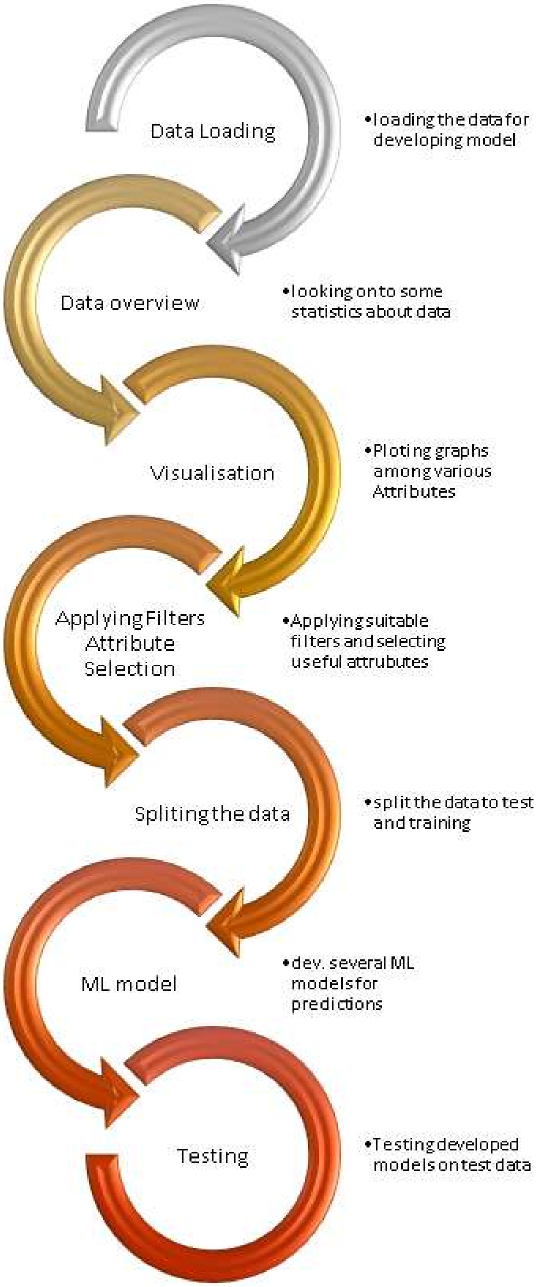
Input the data to ML model
Running the ML Model
output is the predicted charges

For developing the model, the hardware required is a PC working on 64bit processor with internet connection.Software requirements are Java, Eclipse for developing projects in java external Weka api.weka is an open source code/platform for training various Machine learning algorithms ,it comes along with visualization tools.

For the project the use algorithms are Linear regression , Multilayer Perceptron model,Random Forest . For the testing the data is split after applying the nominal to binary filter, and attribute selection . testing is done using the 10 fold cross validation using random shuffle.

Evaluation metrics for the model- Correlation coefficient, Mean absolute error, Root mean squared error, Relative absolute error, Root relative squared error.

Flow chart


Data Loading
•loading the data for developing model
Data overview
•looking onto some statistics about data
Visualisation
•Ploting graphs among various Attributes
Applying Filters Attribute Selection
•Applying suitable filters and selecting useful attrubutes
Spliting the data
•split the data to test and training
ML model
•dev. several ML models for predictions
Testing
•Testing developed models on test data

Results

Following are the model evaluation metrics

1. Linear Regression

```
charges =

253.9781 * age +
 316.4362 * bmi +
577.3602 * children +
 24043.5075 * smoker=yes +
-11869.439

prediction performance

Correlation coefficient                0.8589
Mean absolute error                 4083.0794
Root mean squared error               5957.9169
Relative absolute error               47.7659 %
Root relative squared error             51.1475 %
Total Number of Instances              534
```

2. Multi layer Perceptron Model

```
Linear Node 0
        Inputs    Weights
        Threshold    0.6506451642695577
        Node 1    -0.6756934197170261
        Node 2    -0.9861615658536667
        Node 3    -0.3881271133391064
Sigmoid Node 1
         Inputs    Weights
        Threshold    -0.8960110518023242
        Attribute age    -1.4730549666875787
        Attribute bmi    -0.13862918688389828
        Attribute children    -0.4525545758491905
        Attribute smoker=yes    -1.3973148128431674
         Attribute region=southwest    0.07348344529302792
Sigmoid Node 2
           Inputs    Weights
           Threshold    3.925306896724941
           Attribute age    -0.42863808882442284
           Attribute bmi    -7.787327461564774
           Attribute children    0.12538007176268712
           Attribute smoker=yes    -5.792378289356256
           Attribute region=southwest    -0.005221560160357173
Sigmoid Node 3
             Inputs    Weights
             Threshold    -1.8769489234639098
             Attribute age    -0.28854466008584645
             Attribute bmi    1.230187764444332
             Attribute children    0.2109904911700036
             Attribute smoker=yes    0.873963970283533
             Attribute region=southwest    0.019225142023235317

        prediction performance

        Correlation coefficient              0.9224
        Mean absolute error                2359.92
        Root mean squared error              4519.453
        Relative absolute error             27.6075 %
        Root relative squared error           38.7986 %
        Total Number of Instances             534
```

3.Random forest

```
        prediction performance

        Correlation coefficient              0.9189
        Mean absolute error                2562.6986
        Root mean squared error              4594.487
        Relative absolute error             29.9797 %
        Root relative squared error           39.4427 %
        Total Number of Instances             534
```

As from the above results we can see that multi layer Perceptron does the job best and comparable next to it is the Random Forest model and finally we have Linear regression Model.

Advantages

- With this people can themselves know what factors are responsible for the insurance charges and can work on that factor to reduce the charges
- Companies can integrate this into applications for there easier to calculate charges.

Disadvantages

- Even with the good accuracy the model might not be accurate for the data containing attributes other than the trained attribute.
- It cannot still predict the charges with 100 percent accuracy.

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 a great accuracy . This opens the door for further extension to other fields.

Future Scope

Using the trained weight and model this can be deployed in to devices for easy use . This training procedure can be extended to many more attributes which affect the health.

References

- Medical Cost Personal Datasets | Kaggle
- weka.classifiers.evaluation.output.prediction (weka-stable 3.8.5 API) (sourceforge.io)
- Use weka in your java code - Weka Wiki (waikato.github.io)