**Medical Insurance Prediction**

**Introduction**

***Overview:***

In this projectforthe given data set with information about clients which have earliertaken the medicalInsurance 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 for given input.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 algorithm for supervised learning. For this the data is split into test and trainingsets with ratio 3:7, and models are trained on the training data. Test data is used for the compute the performance of the model.

***Purpose:***

This will help people and insurance companies to know how the various conditionslike age, bmi, smoking or not etc. affect the chargesof 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 othermedical records to calculate charges. but many don't have any solutionor 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 existingparameters.

***Dataset description:***

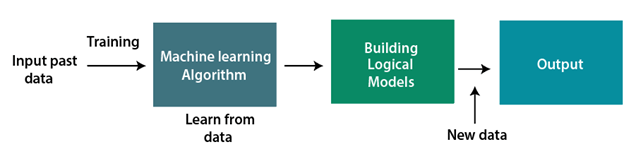
age: age of primary beneficiary  
sex: insurance contractor gender, female, male  
bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height,  
objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9  
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:***

The effects of variousparameters like age, sex etc (fromgiven data) to

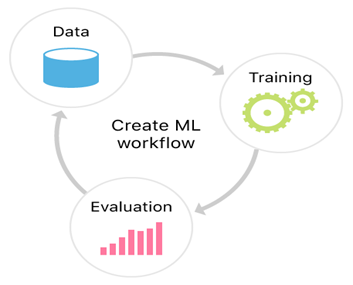
determine how much these factorscan account for our increase/decrease in insurance premium.

**Theoretical Analysis:**

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 algorithm is Linear regression. For the testing the data is split after applying the nominal to binary filter, and attribute selection . testing is done using the 7 fold cross validationusing random shuffle.

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

**Flow chart**

***Results:***

Following are the model evaluation metrics

1. ***Linear Regression***

**charges:**

257.8994\* age +

321.8766 \* bmi +

469.0228 \*children+

-23809.8328 \* smoker=no

**prediction performance**

|  |  |
| --- | --- |
| Correlation coefficient | 0.8643 |
| Mean absoluteerror | 4200.6827 |
| Root mean squared error | 6088.6539 |
| Relative absoluteerror | 46.1553% |
| Root relative squared error | 50.2294% |
| Total Numberof Instances | 1338 |

**Advantages**

With this people can themselves know what factorsare responsible for the insurancecharges and can work on that factor to reduce the charges

* 1. Companies canintegratethis into applications for there easier to calculatecharges.

**Disadvantages**

* 1. Even with the good accuracy the model might not be accurate for the data containing attributes other than thetrained attribute.

It cannot still predict the charges with 100 percentaccuracy.

**Conclusion**

As we can see from the output the most factorsaffecting the chargesof 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 improvethe BMI

Machine learning can be helpful for calculating the Insurance premium charges to agreat accuracy. This opens the door for further extensionto other fields.

**Future Scope**

Using the trainedweight and model this can be deployedin to devices for easy use.

This training procedurecan be extended to many more attributes which affect the health.

**References**

1. Medical Cost Personal Datasets Kaggle
2. weka.classiflers.evaluation.output.prediction(weka-stable 3.8.5 API)(sourceforge.io)