**Insurance Purchase Prediction**

**Abstract:**

Earning money is not at all useful if you don’t have a life. Think about yourself and your family and the importance of life. You know the importance and benefits of having a Health Insurance Policy.

Now-a-days, buying a health insurance policy for yourself and your family is important because medical care is expensive, especially in the private sector. Hospitalization can burn a hole in your pocket and derail your finances. It will become even tough, if the person, who brings in the money, is now in a hospital bed. All this can be avoided by just paying a small annual premium which would lessen your stress in case of medical emergencies.

A good health insurance policy would usually cover expenses made towards doctor consultation fees, costs towards medical tests, ambulance charges, hospitalization costs and even post-hospitalization recovery costs to a certain extent.

The purpose of our project is to explore the use of machine learning algorithms to predict the charges that a common man needs to pay. And it predict how much charges it will cost per annum by basing on the some personal details and their BMI, and we will provide the best companies to choose to be the member of that company.

There are several machine learning algorithms that are used to predict the premium charges are decision tree, Logistical Regression, SVM (support Vector Machine), and random Forest. We are using the decision tree in this project to predict the data.

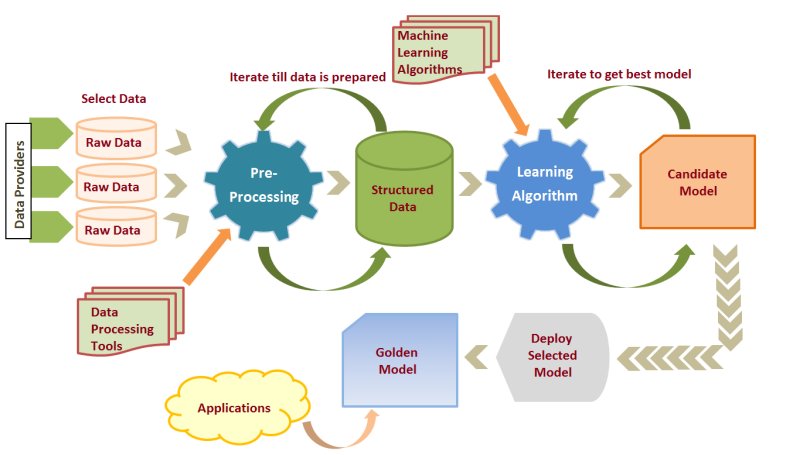
**Introduction:**

**Machine learning:**

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. **Machine learning focuses on the development of computer programs**that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. **The primary aim is to allow the computers learn automatically**without human intervention or assistance and adjust actions accordingly.

Basically Machine Learning is of three types:

* **Supervised Learning**: Learning from characterized data.
* **Unsupervised Learning**: Learning from raw data.
* **Reinforcement Learning**: Learning from self mistakes (or) self learning data.



Steps involved in Machine learning are

* 1.Data collection
* 2.Data wrangling
* 3. Analyze the data
* 4.Train the algorithm
* 5.Test algorithm
* 6. Deployment

**Classification:**

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc.

The types of classification algorithms in Machine Learning:

**Linear Classifiers:** Logistic Regression, Naive Bayes Classifier, Support Vector Machines

* **Decision tree:**

Given a data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data.

**Advantages:** Decision Tree is simple to understand and visualize, requires little data preparation, and can handle both numerical and categorical data.

**Disadvantages:**Decision tree can create complex trees that do not generalize well, and decision trees can be unstable because small variations in the data might result in a completely different tree being generated.

* **KNN model:**

Neighbors’ based classification is a type of lazy learning as it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the k nearest neighbors of each point.

**Advantages:**This algorithm is simple to implement, robust to noisy training data and effective if training data is large.

**Disadvantages:**Need to determine the value of K and the computation cost is high as it needs to computer the distance of each instance to all the training samples.

* **Random forest**:

Random forest classifier is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement.

**Advantages:**Reduction in over-fitting and random forest classifier is more accurate than decision trees in most cases.

**Disadvantages:**Slow real time prediction, difficult to implement, and complex algorithm.

* **Support vector machine (SVM):**

Support vector machine is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

**Advantages:**Effective in high dimensional spaces and uses a subset of training points in the decision function so it is also memory efficient.

**Disadvantages:**The algorithm does not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

**PYTHON:**

Python, as a high-level programming language, allows you to focus on core functionality of the application by taking care of common programming tasks. The simple syntax rules of the programming language further makes it easier for you to keep the code base readable and application maintainable. Main reasons to use python language are:

* **Readable and Maintainable Code**
* **Multiple Programming Paradigms**
* **Compatible with Major Platforms and Systems**
* **Robust Standard Library**
* **Many Open Source Frameworks and Tools**
* **Simplify Complex Software Development**
* **Adopt Test Driven Development.**

Objective of Research:

This project deals with the calculation of charges that has to be paid by the person based of certain parameters like age, gender, BMI, number of children, whether he is a smoker or not and finally based on the region where he lives. We will link up with several companies and provide the assurance to the users by basing on their conditions.

To provide the best and optimum charges that can be provide by a company in some harsh situations to the user. Our government is also providing several health care schemes some of those are:

* **Central Government Health Scheme (CGHS)**
* **Universal Health Insurance Scheme (UHIS)**
* **Employment State Insurance Scheme (ESIS)**

**We make sure that these health insurance policies to reach to normal person and when these are applicable by basing on several circumstances.**

**Problem Statement:**

The insurance companies are under some privileged conditions that may not provide insurance at every condition and also may not be provided with full amount. In those conditions we are providing them to linkup with the companies by choosing the suitable one with the certain parameters.

## How to Select the Right Insurance Policy

It’s difficult to select the best insurance policies as all insurance company provides a similar type of insurance plan. Hence some of the important points that any Person should look before purchasing any plans are:

1. Sum Assured

2. Minimum Entry Age and renewability clause

3. Children

4. Inclusion and Exclusion

5. No Claim Bonus

6. Other Benefits

7. BMI

**Industry Profile:**

The insurance sector plays a critical role in a country’s economic development. It acts as a mobilizer of savings, a financial intermediary, a promoter of investment activities, a stabilizer of financial markets and a risk manager. The life insurance sector plays an important role in providing risk cover, investment and tax planning for individuals; the non-life insurance industry provides a risk cover for assets. Health insurance and pension systems are fundamental to protecting individuals against the hazards of life, and India, as the second-most populous nation in the world, offers significant potential for that type of cover. Furthermore, fire and liability insurance are essential for corporations to safeguard infrastructure projects and investment risks. Private insurance systems complement social security systems and add value by matching risk with price.

**Review of literature:**

The dataset is taken from the following websites to predict the charges that are to be paid by a person those datasets are:

[www.kaggle.com](http://www.kaggle.com)

[www.data.gov.in](http://www.data.gov.in)

**Data collection:**

Data collection is defined as the procedure of collecting, measuring and analyzing accurate insights for research using standard validated techniques. A researcher can evaluate their hypothesis on the basis of collected data. In most cases, data collection is the primary and most important step for research, irrespective of the field of research. The approach of data collection is different for different fields of study, depending on the required information.

**Methodology:**

Here we have used model called decision tree in python using machine learning

* A tree has many analogies in real life, and turns out that it has influenced a wide area of machine learning, covering both classification and regression
* In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making
* Decision Trees are type of Supervised Machine Learning where the data is continuously split according to certain parameter
* The tree can be explained by two entities, namely decision nodes and leaves

**Figures and Table**

1. Statistical Techniques and Visualization
2. Data modeling and visualization: applying data to a model-high accuracy-RUC curve
3. Findings and Suggestions: what are drawbacks that u r observed and how u use this problem statement and provide solution

**Conclusion:**

There are several insurance companies that are maintained public and private sectors. It is difficult to predict a person that which company he/she need to visit. So, by basing on some basic details of the person and acclaim the charges that he/she need to pay. And provide with the best company that is suitable for a person by basing the firms and conditions.