**HEALTHINSURANCEPURCHASEPREDICTION**

**1.1INTRODUCTION**

* Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc).
* Python has a simple syntax similar to the English language.
* Python has syntax that allows developers to write programs with fewer lines than some other programming languages.
* Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be very quick.
* Python can be treated in a procedural way, an object-orientated way or a functional way.

Artificial Intelligence is an approach to make a computer, a robot, or a product to think how smart human think. AI is a study of how human brain think, learn, decide and work, when it tries to solve problems. And finally, this study outputs intelligent software systems. The aim of AI is to improve computer functions which are related to human knowledge, for example, reasoning, learning, and problem-solving.

The intelligence is intangible. It is composed of

* Learning
* Reasoning
* Problem Solving
* Perception
* Linguistic Intelligence

**1.2 OBJECTIVES OF RESEARCH**

The main objective of the project is to predict the premium covered in particular insurance plan so that he needn’t go to the insurance company or insurance agent.

To obtain estimate whether a person gets his charges based on gender, children, bmi and age as inputs.To obtain accuracy based on previous data.

**1.3 PROBLEM STATEMENT**

Health insurance is an insurance that covers the whole or a part of the risk of a person incurring medical expenses, spreading the risk over a large number of persons. By estimating the overall risk of health care and health system expenses over the risk pool, an insurer can develop a routine finance structure, such as a monthly premium or payroll tax, to provide the money to pay for the health care benefits specified in the insurance agreement. The benefit is administered by a central organization such as a government agency, private business, or not-for-profit entity.And here we are predicting health insurance coverage.

**2. REVIEW OF LITERATURE**

Dramatic advances in artificial intelligence and machine-learning technologies have accelerated the ability of insurers to predict risk. Algorithms can find trends and patterns that help forecast the probability of a risk situation occurring again. By utilizing internal and external data sources, algorithms are selected according to how a specific model fits with the insurer’s data. This model is applied to predict or detect the likelihood of an event happening, such as a person needing medical attention abroad for travel insurance or a house flooding for home insurance. Insurance and assistance provider The Collision Group uses a variety of predictive analytical tools to flash through terabytes of data to find variables, some of which it hadn’t considered, to help predict customer risk and purchasing behavior.

CASE STUDY In this case analysis of hospital data was done to optimize and balance human resources, medication and time spent on each patient to improve clinical outcomes. It performs spectral partitioning of the graph that was built using the data from the health-care agency. Understanding the structure of the data and capturing hidden interrelationships helped to improve the existing resource allocation schema. As a result, created a model of resource harness that stopped overspending and improved the quality of patient's care.

In this case analysis of patient’s symptoms was taken to predict the development of the disease. It demonstrates principal component analysis and support vector machine classifier. Healthcare data analytics allows us to find patterns that help to recognize early stages of the disease and predict its development. This predictive model provides the hospital with anopportunity to control the occurrence of epidemics as well as be more accurate in early diagnosis of the disease.

**3. DATA COLLECTION**

A data set (or dataset) is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question.

Before you had to the bank to pick up you requires charges based on bmi, it’s helpful to know all the little details that will make the process easier.

It is hard to know in advance, what kind of data will be helpful in future. We considered dataset which consists of previous years data about bmi charges. Using previous data we can easily predict present situation. A dataset consists of columns as Gender, Bmi, Children, Age, Charges.Here we consider Gender, Bmi, Children, Age as independent variables and Charges as dependent variable. The independent variables are given as inputs which are represented by variable x and dependent variable is given as output which is represented by variable y.

**3.1 AGE**

It shows the age of the customer. The length of time that a person has lived or a thing has existed.

**3.2 BMI**

It shows the bmi of the customer.BMI is a person's weight in kilograms (kg) divided by his or her height in meters squared. The National Institutes of Health (NIH) now defines normal weight, overweight, and obesity according to BMI rather than the traditional height/weight charts. Overweight is aBMI of 27.3 or more for women and 27.8 or more for men.

**3.3 CHILDREN**

It shows no of children the customer have.

**3.4 SMOKER**

It says whether the person is smoker or not.

**3.5 GENDER**

It says the gender of the customer

**3.6 REGION**

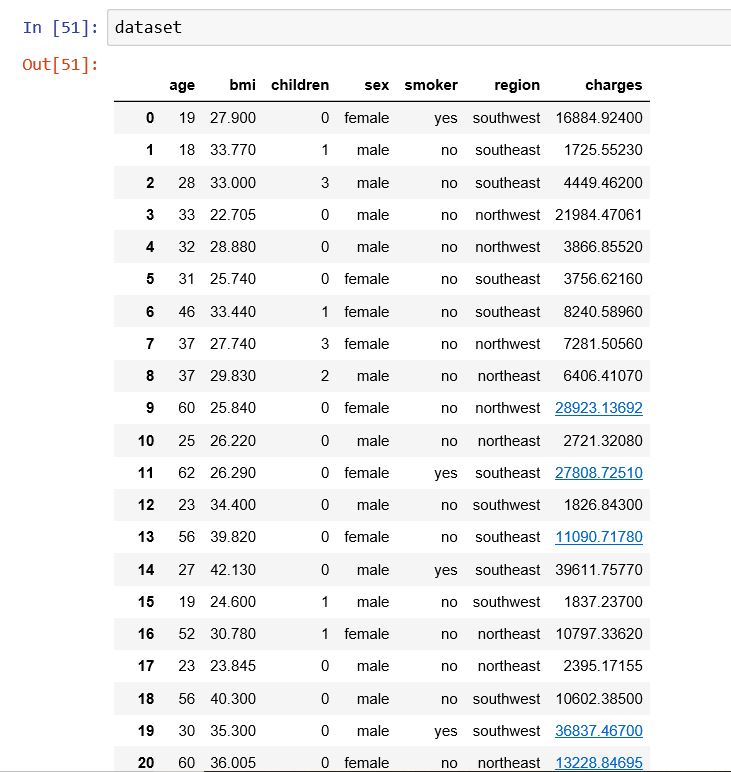
It says where the customer lives.

**3.7 CHARGE**

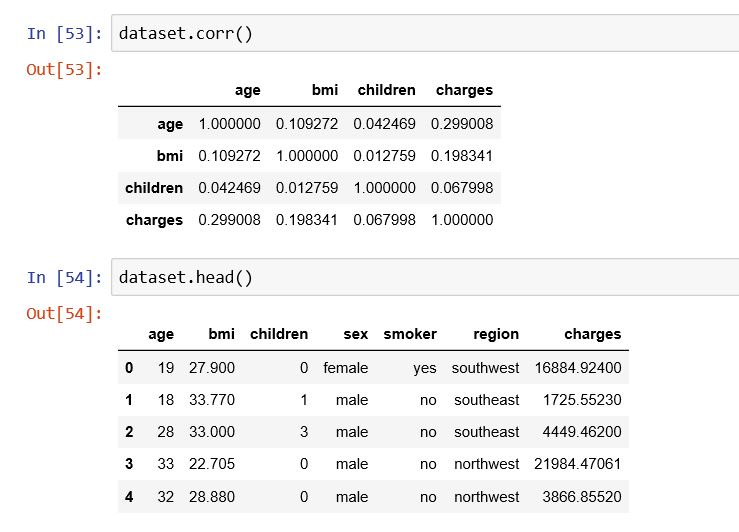
It says what the customer should pay.

**4. METHODOLOGY**

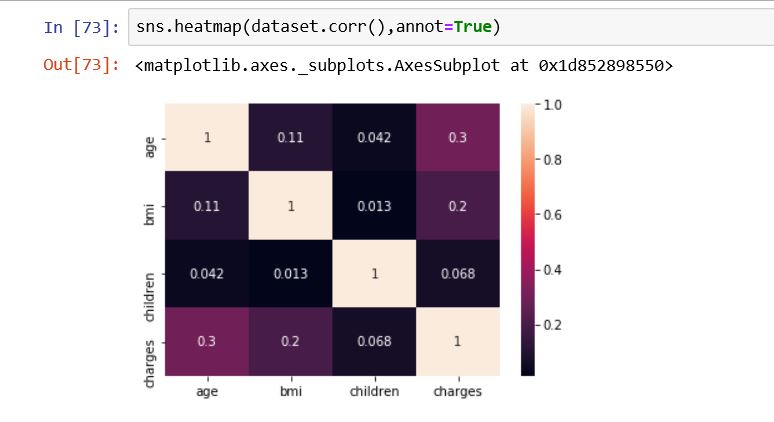
**4.1 EXPLORATORY DATA ANALYSIS**

**4.1.1FIGURESANDTABLES**

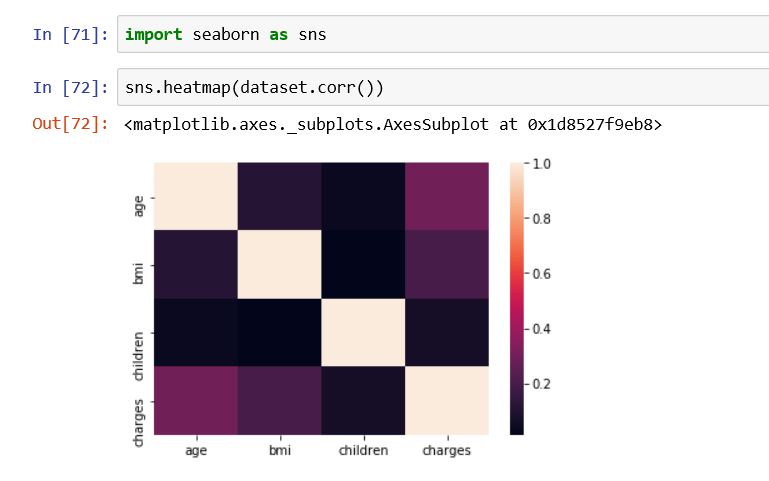
**Table:1.1**

****

**Table:1.2**

****

**Fig 1.1**

****

**Fig 1.2**

**4.2 STATICAL TECHNIQUES AND VISUALIZATION**

**NUMPY:**

NumPy stands for ‘Numerical Python’ or ‘Numeric Python’. It is an open source module of Python which provides fast mathematical computation on arrays and matrices. Since, arrays and matrices are an essential part of the Machine Learning ecosystem, NumPy along with Machine Learning modules like Scikit-learn, Pandas, Matplotlib, TensorFlow, etc. complete the Python Machine Learning Ecosystem.

NumPy provides the essential multi-dimensional array-oriented computing functionalities designed for high-level mathematical functions and scientific computation. Numpy can be imported into the notebook using import numpy as np.

|  |
| --- |
|  |

NumPy’s main object is the homogeneous multidimensional array. It is a table with same type elements, i.e, integers or string or characters (homogeneous), usually integers. In NumPy, dimensions are called axes. The number of axes is called the rank.

**PANDAS:**

Similar to NumPy, Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures and data analysis tools. Unlike NumPy library which provides objects for multi-dimensional arrays, Pandas provides in-memory 2d table object called Data frame. It is like a spreadsheet with column names and row labels.

Hence, with 2d tables, pandas is capable of providing many additional functionalities like creating pivot tables, computing columns based on other columns and plotting graphs. Pandas can be imported into Python using import pandas aspd.

New columns and rows can be easily added to the dataframe. In addition to the basic functionalities, pandas dataframe can be sorted by a particular column.

Dataframes can also be easily exported and imported from CSV, Excel, JSON, HTML and SQL database.

**MATPLOTLIB:**

Matplotlib is a 2d plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments. Matplotlib can be used in Python scripts, Python and IPython shell, Jupyter Notebook, web application servers and GUI toolkits. matplotlib.pyplot is a collection of functions that make matplotlib work like MATLAB. Majority of plotting commands in pyplot have MATLAB analogs with similar arguments.

On the X array below we saying... include all items in the array from 0 to 4.On the y array below we are saying... just use the column in the array mapped to the **row**. The **Charges** column.We are using group by to view the distribution of values in our **Charges** column. Recall that this column is our **target variable**. It's that thing we are trying to predict.

Charge prediction is a regression problem. Accuracy is only really useful when there are an even distribution of values in a data set.This module for Node-RED contains a set of nodes which offer machine learning functionalities. Such nodes have a python core that takes advantage of common ML libraries.These flows create a dataset, train a model and then evaluate it. Models, after training, can be use in real scenarios to make predictions.

Flows and test datasets are available in the 'test' folder. Make sure that the paths specified inside nodes' configurations are correct before trying to execute the program.

**4.3 DATA MODELING**

Here we are using multi regression model. Multiple Regression and Beyond offers a conceptually oriented introduction to multiple regression (MR) analysis and structural equation modeling (SEM), along with analyses that flow naturally from those methods. By focusing on the concepts and purposes of MR and related methods, rather than the derivation and calculation of formulae, this book introduces material to students more clearly, and in a less threatening way. In addition to illuminating content necessary for coursework, the accessibility of this approach means students are more likely to be able to conduct research using MR or SEM--and more likely to use the methods wisely.

* Covers both MR and SEM, while explaining their relevance to one another
* Also includes path analysis, confirmatory factor analysis, and latent growth modeling
* Figures and tables throughout provide examples and illustrate key concepts and techniques

**5. REFERNCES**

**https://kaggle.com/datasets**

**https://archive.ics.uci.edu/ml/datasets.php**

**https://data.gov.in/**

**https://www.kaggle.com/hhs/health-insurance**

**6. CONCLUSION**

The aim of this research is to propose and implement a rule based system to predict the charges from the collection of past data. This has been achieved by multi level regression algorithm on previous dataset.