****

**BENAZIR BHUTTO SHAHEED UNIVERSITY LYARI**

**FINAL PROJECT REPORT**

**(Artificial Intelligence)**

**Group Members:**

* Salman Abdul Rahim (BCS-616)
* Muhammad Saad (BCS-597)
* Zafar Hayat Khan (BCS-635)

**Submitted To:**

Ma’am Nadiya Mirbahar

**INSURANCE CHARGES PREDICTION USING**

**MACHINE LEARNING**

**INTRODUCTION:**

It has become a necessity for almost every person nowadays to sign up for the health insurance facility. These perks are being granted by many of the government as well as private insurance companies all across the globe. Concerning the value of insurance in the lives of individuals, it becomes important for the insurance companies to remain sufficiently precise while quantifying the amount covered by their policy of depicting the insurance charges over the customers.

Our project goal is to provide an idea about the insurance charges that a person require in order to fulfill the terms and conditions of an insurance company. According to our project, the criteria of insurance charges does not rely on any specific insurance company as we have taken out general estimations depending on the health status of an individual. It is purely a machine learning problem justifying multiple linear regression problem. The insurance dataset that we are using in this project is taken from kaggle while it contains seven features i.e. age, gender, bmi, children, smoker, region, and charges. In addition to this, the Machine Learning algorithm that we are up to use for our dataset is LinearSVR (Linear Support Vector Regressor).

**PROBLEM STATEMENT:**

There are various factors that influence the cost of insurance. These considerations contribute in the origination of some insurance policies. Each of the factors carry an important role when the amounts are being calculated. Ignorance of any factor might cause the policy to change from top to bottom. Hereby, it becomes critical to perform this task with high accuracy. Machine Learning plays an essential part to solve this issue. The technique of supervised learning pulls through the goal to automate the insurance price prediction. The linear regression model learns the insurance data from the past and can give out accurate insurance charges for the new set of data. This on the one hand reduces human effort to assume manual calculations, and on the other hand, can improve insurance company’s expediency through automation.

**TOOLS AND LIBRARIES:**

In our project, we have demonstrated the insurance charges prediction through the most robust programming language, Python. It is considered as one of the best and efficient tool for Artificial Intelligence, especially to compute and deal the Machine Learning problems. There are various significant packages and libraries with predefined methods which help in solving any of the Machine Learning query in a faster as well as systematic way. All we need is to know about these packages and libraries in order to deal with our data perfectly. Here in our project, we are using the following python libraries for data analysis, preprocessing and building our Machine learning model:

1. Pandas
2. NumPy
3. Matplotlib
4. Seaborn
5. Sci-kit Learn

Let’s have a brief look on these libraries and what utilities they provide for the Python developers and A.I enthusiasts.

1. **Pandas:**

Pandas is an important data preprocessing library written for the Python language. It is an effectual tool to create and manipulate Data Frame objects. In addition, we can also load the in-memory data files with different formats into the Python development environment. For the preprocessing of data, Pandas is sufficient to carry out data cleaning, data alignment, visualization and even handling of the missing data. It is surely the best data representation tool with the benefit to write less and gain more.

1. **NumPy:**

NumPy, (Abbreviated as Numerical Python), is another fast and robust Python library used to deal with multi-dimensional arrays and numeric data. It possesses large-scale mathematical functions that include in different domains like Linear Algebra, Matrices, Fourier Transform, Fourier Series, Statistics, etc. Furthermore, this Python library also contains such special functions that aid in creating charts and graphs with visualization libraries in Python. Hence, we can consider it the second to none package for array manipulations.

1. **Matplotlib:**

Matplotlib is a data visualization library integrated with the Python language. It contains numerous packages from which the mostly used is the pyplot package. We can produce many graphs according to the nature of our data. These different charts and graphs include line plots, scatter plots, bar graphs, pie charts, etc. It deals with both types of plots, 2D, as well as 3D plots.

1. **Seaborn:**

Another data visualization library is the Seaborn library which is basically the advance version of Matplotlib. If you are working with the statistical methods through Python and you need visuals and graphs, the Seaborn library is the best for that. From line plots to kde-plots and heatmaps, the Seaborn library is the all in one package for extensive and detailed data visualization.

1. **Sci-kit Learn:**

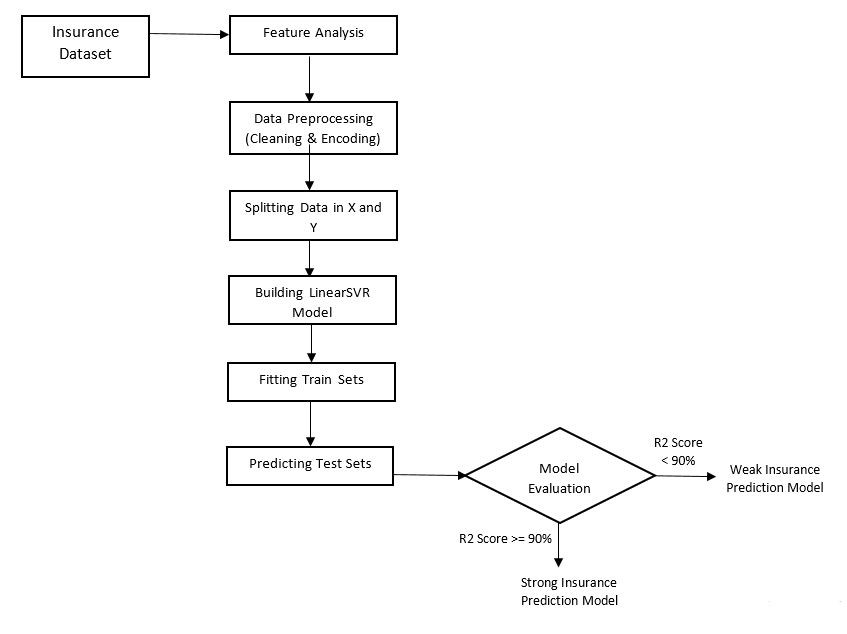
Sci-kit learn (shortly referred as sklearn) is a Python library that supports predictive data analysis tasks. It is a complete package for machine learning algorithms and data mining techniques. All of these methods, techniques and algorithms are based on statistical methods and modeling. Some of the significant techniques that the Sci-kit Learn library affords are Regression, Classification, Clustering, Association, etc. It also features methods and functions that are relevant for finding out model accuracies, losses, deviations, and many other evaluation metrics. Moreover, there is also an essence of useful data preprocessing methods.

**THE LINEAR SUPPORT VECTOR REGRESSOR:**

Support Vector Regressor is the predecessor of the SVC (Support Vector Classifier) associated with the SVM machine learning technique. The SVR works same as the SVC i.e. both of them aim to attain maximal margin. The main focus is on minimizing the regression error of the hyperplane through taking different traces over the dataset.

Now let us take a look on the one of the versions of SVR algorithm that we have used here in this project i.e. the Linear Support Vector Regressor. It is the type of Support Vector Regressor which holds the linear kernel (to approximate target value through an estimating single line). There is a tolerance function of epsilon which aims to minimize the regression error while adjusting the hyperplane. In addition, the C parameter regularizes features and controls the fitting parameters of LinearSVR.

The LinearSVR is one of the popular algorithms for regression problems especially when we have a dataset with more than 1000 values. This is one of the reasons that we are using this exceptional regressor for our Insurance Price Prediction in this project as we initially have 1338 values in our dataset.

**METHODOLOGY:**

The above diagram shows the methodological model and the workflow of our Insurance Price Prediction project. There are a few steps that initiate from the loading of our dataset in the Jupyter IDE to the predictions that we have carried out. Obviously, we have utilized all the necessary Python libraries discussed above for the step by step and clear workflow.

**MODEL EVALUATION AND RESULTS:**

**CONCLUSION:**