



**PROJECT REPORT**

**Medical Insurance cost prediction using multiple linear regression model**

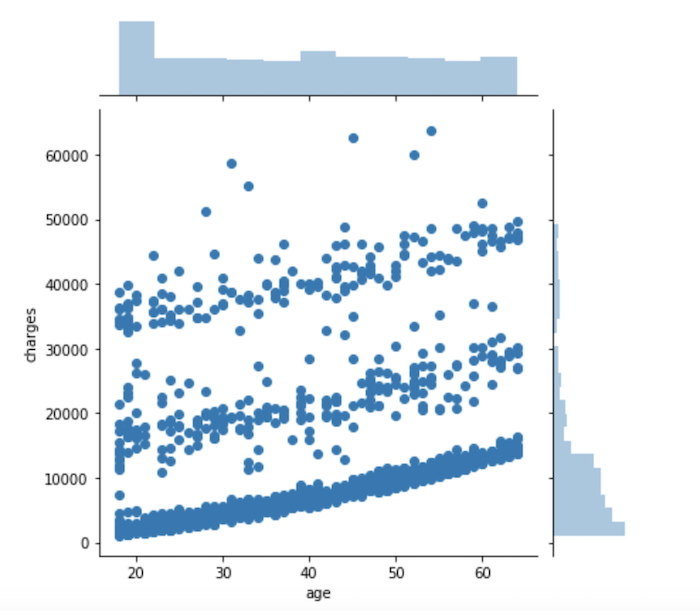
**NAME: Falak Khan**

**BATCH: Machine Learning and AI Batch A3**

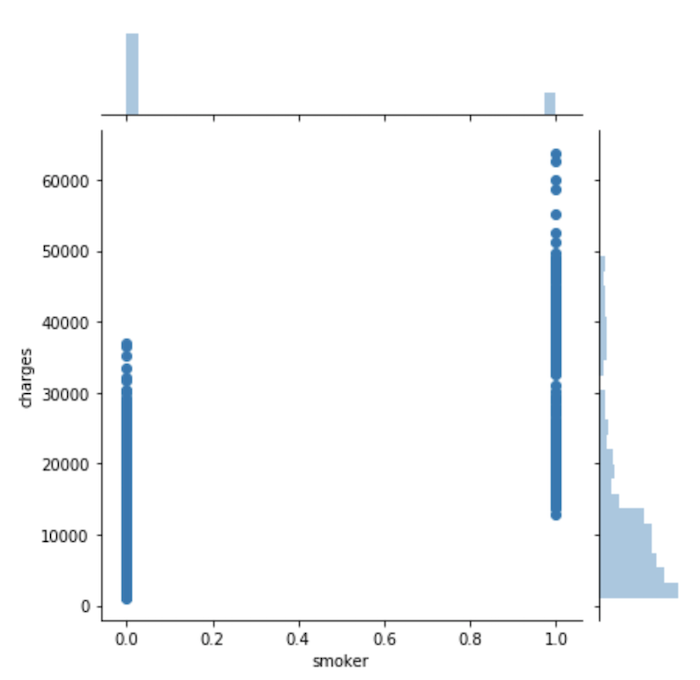
1. **TABLE OF CONTENT:**

|  |  |
| --- | --- |
| 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 (**target – y)** |

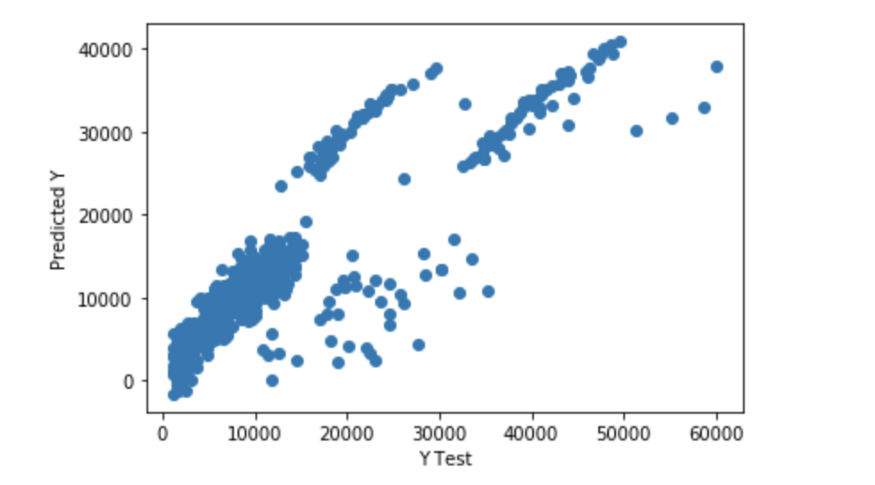
1. **TABLE OF FIGURES/GRAPHS:**
2. Here we see that as Age goes up Charges for health insurance also trends up



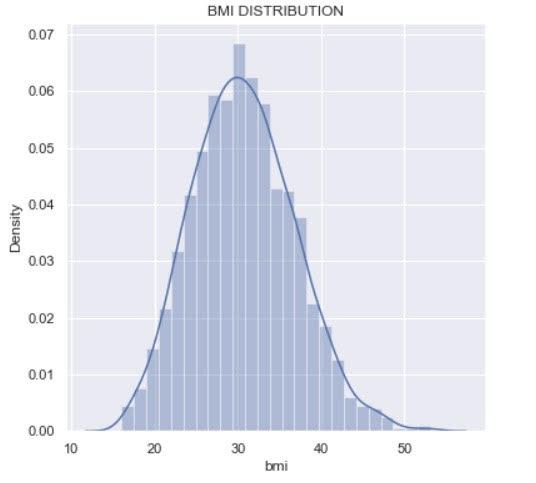
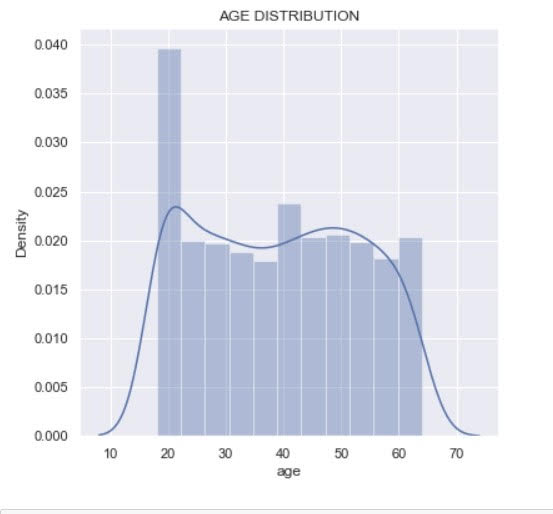
1. Here we see that charges for smokers are higher than non-smokers

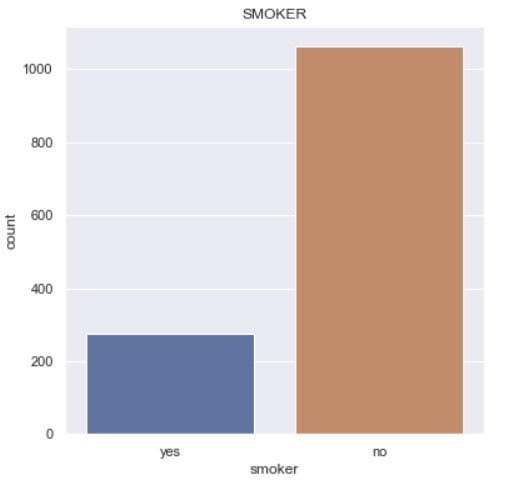
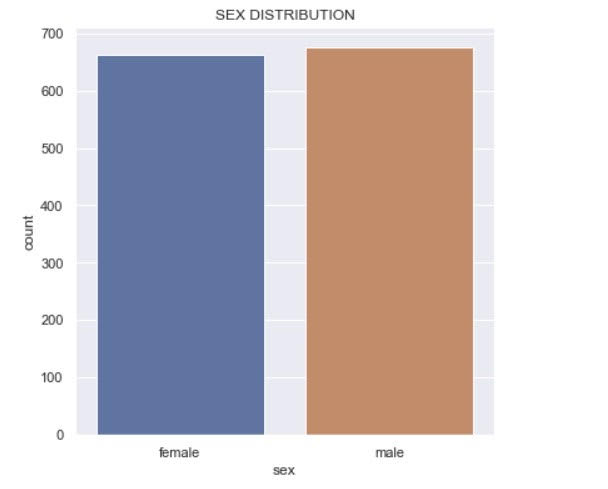


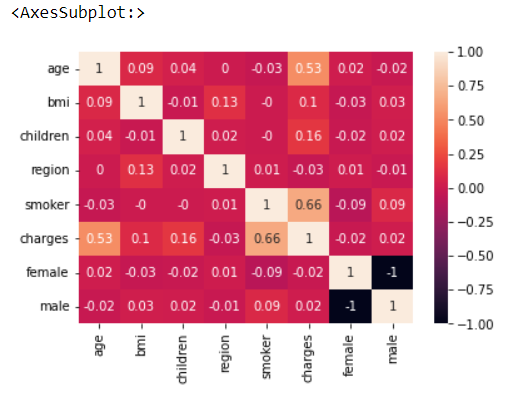
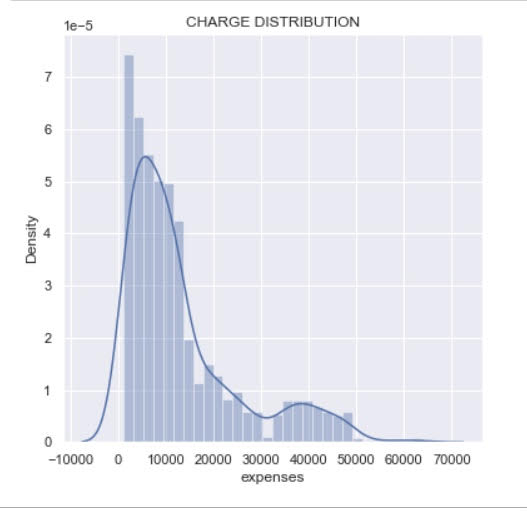
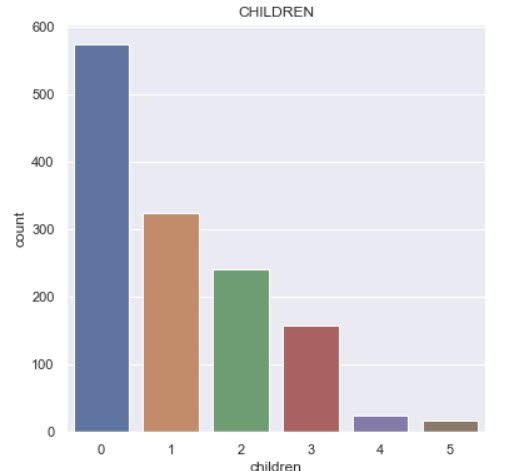
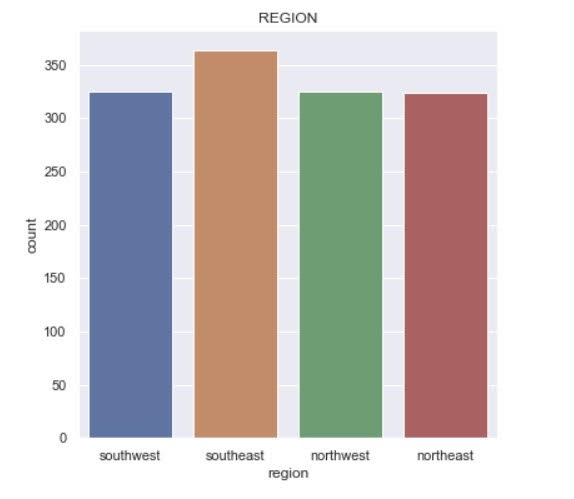
1. Here we can see the correlation between predicted and actual results.



**Data Visualization using Graphs:**







Correlation Matrix using Heatmap

**c. ABSTRACT:**

Insurance is a policy that eliminates or decreases loss costs occurred by various risks. Various factors influence the cost of insurance. These considerations contribute to the insurance policy formulation. Machine learning (ML) for the insurance industry sector can make the working of insurance policies more efficient. This study demonstrates how different models of regression can forecast insurance costs. And we will compare the results of models, for example, Multiple Linear Regression, Random Forest Regressor, CART, XGBoost,

Insurance fees are based on different variables. As a result, insurance fees are continuous values. Regression is the best choice available to fulfil our needs. We use multiple linear regression in this analysis since there are many independent variables used to calculate the dependent(target-y) variable. for this study, the dataset for cost of health insurance is used

d. **INTRODUCTION:**

We are on a planet full of threats and uncertainty. People, households, companies, properties, and property are exposed to different risk forms. And the risk levels can vary, these dangers contain the risk of death, health, and property loss or assets. Life and well-being are the greatest parts of people's lives, but risks cannot usually be avoided. Insurance is therefore a policy that decreases or removes loss costs incurred by various risks concerning the value of insurance in the lives of individuals it becomes important for the companies of insurance to be sufficiently precise to measure or quantify the amount covered by this policy and the insurance charges which must be paid for it. Various variables estimate these charges each factor of this is important, if any factor is omitted when the amounts are computed the policy changes overall.

It is therefore critical that these tasks are performed with high accuracy as human mistakes could occur, insurers use people with experience in this area. They also use different tools to calculate the insurance premium. Machine Learning is beneficial here, ml may generalize the effort or method to formulate the policy.

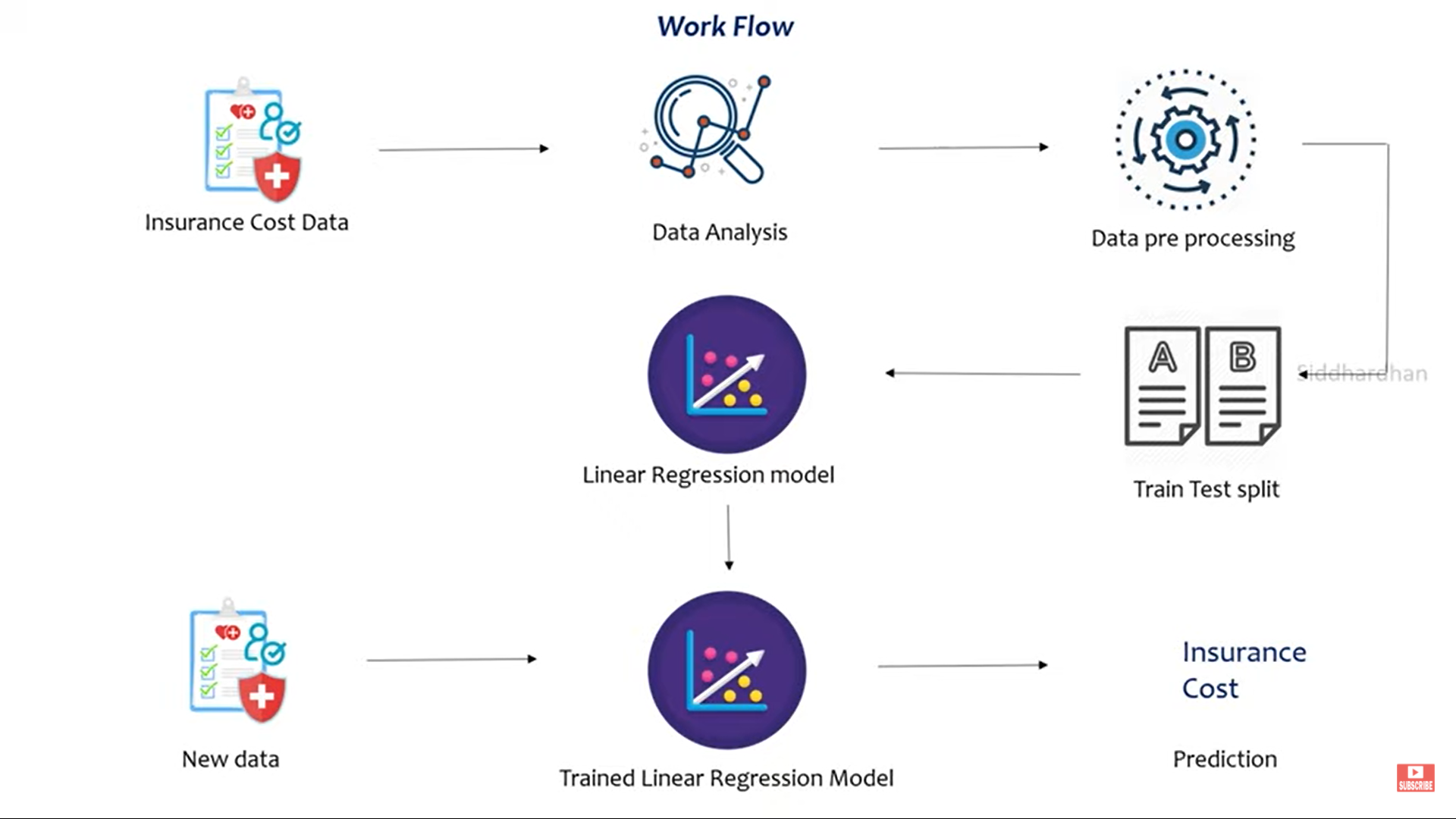
These ml models can be learned by themselves. The model is trained on insurance data from the past, the requisite factors to measure the payments can then be defined as the model inputs, then the model can correctly anticipate insurance policy costs. This decreases human effort and resources and improves the company's profitability, Thus the accuracies can be improved with these Machine Learning models.

e. **LIBRARIES USED:**

* **NumPy** is an open-source numerical Python library. NumPy contains a multi-dimensional cluster and network information structures. It tends to be used to play out various numerical procedures on clusters, for example, trigonometric, measurable and mathematical schedules.

* **Pandas** is a library composed for the Python programming language for information control and examination. Specifically, it offers information structures and activities for controlling numerical tables and time arrangement, which is a Panel Data. In this way, the library is named is Pandas. Pandas' library is based on NumPy bundle.
* **Matplotlib** is a plotting library for the Python programming language and its numerical science expansion of NumPy. It gives an article arranged API to implant plots into applications utilizing broadly useful GUI.
* **Seaborn**- Used for visualization.
* **Scikit-learn** (formerly scikits.learn and also known as sklearn) It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

f. **ALGORITHM/ FLOWCHART**:



g. **Data Preprocessing:**

**Dealing with categorical data:**

In our dataset we have 3 independent variables that are categorical data

1. Sex
2. Smoker
3. Region

**Dealing with Skewed distribution of the ‘Charges’ feature:**

As we can see the graph shown above, the ‘charges’ feature is left skewed and thus causes outliers, to resolve this we need to convert it to a normal distribution using Log Transforms, with this the ‘charges’ feature is now normally distributed and free of outliers.

**Dealing with outliers in the ‘BMI’ feature:**

As in the graph shown above the ‘BMI’ feature has outliers which cause problems for Linear Regression model. Since the outlier values are less in number, we resolve this by dropping the rows with the outlier values.

**Scaling the Dataset:**

We are going to scale our data using StandardScaler from sklearn, since Linear Regression expects that the data is normally distributed.

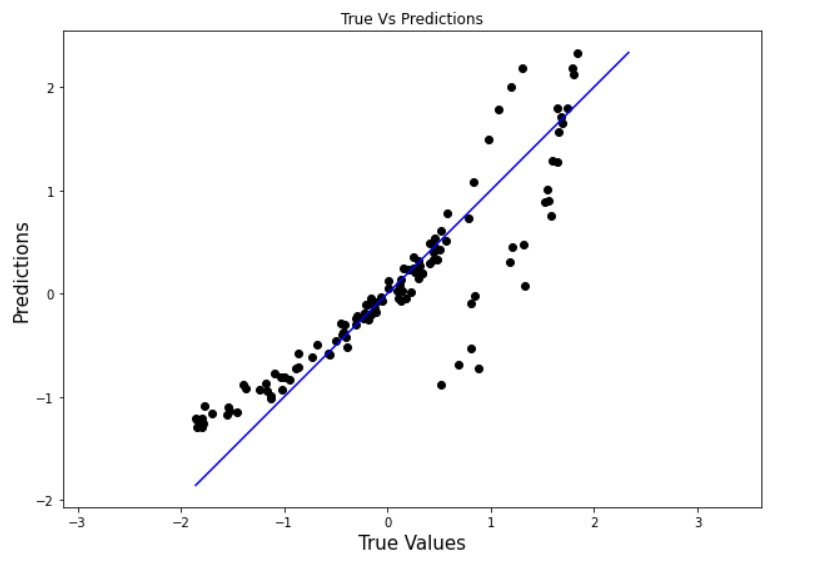
**Algorithmic Models created:**

1. **Linear Regression:**

We are going to use sklearn library in python to implement Linear Regression, we are going to split our dataset into test and training data and fit the Regression model with the training data and evaluate the model based on its RMSE and R-square scores over the test data.

**R squared value: 0.7999053396503137**

It means that our model has can explain 80% of the variance of the target function using independent variables.



**h. CONCLUSION:**

In this model, we have projected strategies for anticipating medical coverage. Medical coverage is a kind of protection item that explicitly ensures the health expenses or care of the protection individuals in the event that they become sick or have a mishap. The degree of treatment in crisis offices shifts altogether relying upon what kind of medical coverage an individual has by this we foresee the protection charges of an individual. Machine learning can significantly minimize these individual efforts in policymaking, as ML models can do cost calculation in a short time, while a human being would be taking a long time to perform the same task. This will help businesses improve their profitability. The ML models can also manage enormous amounts of data.

**THANK YOU.**

Project by -

Falak Khan

Batch A3

**REFERENCES:**

* 1. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., "Scikitlearn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, Oct 2011.
* 2. E. Wang; and G. Gee, "Larger Issuers, Larger Premium Increases: Health insurance issuer competition post-ACA," 2015.
* 3. C. C. a. A. Semanskee, "Analysis of UnitedHealth Group’s Premiums and Participation in ACA Marketplaces," 2016. 4. Ng, "Machine learning for Housing Price Prediction Mobile Application," Masters, Department of Computing, Imperial College London, 2015