# MEDICAL INSURANCE COST PREDICTION

MASABATTULA DURGA PRASADU[1], Smt.DURGA DEVI 2

[1] **Assistant Professor MCA DEPT,** Dantuluri Narayana Raju College **,Bhimavaram, Andharapradesh**

**Email id: -** adurgadevi760@gmail.com

[2]PG Student of MCA, **Dantuluri Narayana Raju College,** Bhimavaram, Andharapradesh

**Email id:-** durgaprasadumasabattula@gmail.com

## ABSTRACT

The health care costs constitute a significant fraction of the U.S. economy. Nearly 20% of the Gross Domestic Product (GDP) is spent on health care. The health spending in the US is the highest among all developed nations in absolute numbers as well as a percentage of the economy. The U.S. government bears a large portion of seniors' health expenditure through itsMedicare program. The growing health related expenses combined with the fact that the baby-boomer generation is retiring, and hence they will be eligible for Medicare, puts a great burdenon the U.S. exchequer. Therefore, it is essential to contain health related payments through all means possible

In this work, we will develop a medical price prediction system using machine learning algorithms which will aid in steering patients to cost effective providers and thereby curb health spending. The policymakers can also use the tool to better understand which providers are relatively expensive and take punitive actions if necessary. The prediction of the medical price will be done using implementing Random Forest Regression algorithm in machine learning.

Additionally, we plan to include the experiments on the same data with other machine learning models such as Gradient Boosted Trees and Linear Regression and compare results. The findings from these experiments will also be included.

*Key terms*- **Health care, GDP, medical price prediction system, Random Forest Regression, machine learning, GradientBoostedTrees, Linear Regression**

## 1 INTRODUCTION

The health care costs in the U.S. account for 17.80% of the national output [2], which is the highest among all developed nations. The U.S. government runs Medicare insurance program for seniors and bears nearly half of seniors' total health care spending [3], [4], [5]. The number of seniors is set to expand dramatically with the ongoing and impending retirement of the baby- boomer generation, which is expected to swell the ranks of Medicare beneficiaries a lot more in the coming years. Consequently, the Medicare outlay for the government has to increase and thatadds a great strain on the budget. Therefore, ways and means to control health care costs and thereby, slow down the rise of Medicare spending have assumed great significance. One of the key components to restrain the rise in health care costs is access to an accurate medical price prediction system. That is, if patients have accurate information on medical pricing such as, a certain medical procedure costs dollar amount X at hospital A, the same procedure comes with aprice tag of Y at hospital B then they have the opportunity to choose the provider that costs themless.

In this project our goal is to predict medical prices based on the data we have in hand. In the firstfew chapters of this report, we will compare the work of various authors in the area of price prediction and we will also provide the information in detail, about some of the techniques used in health care domain to predict the health care prices. Later, we will propose the design of a new system which will use Medicare payment datasets. The proposed system can be called as a medical price prediction system. Such a system will be useful for patients, and government officials alike. Patients can use the price prediction tool to choose the most cost-efficient. providers. It can be used by Medicare administrators to forecast expenditure for future months and years and plan the budget accordingly. Additionally, high cost providers can be identified using the system. Deeper investigations may be subsequently carried out involving high charging providers and punitive measures may be initiated against them when necessary [2]. We will build the proposed system by implementing two machine learning algorithms from the scratch.

The first algorithm is, Regression Tree and the second one is, Random Forest Regression. While implementing the Random Forest regression algorithm, we will make use of Regression Trees algorithm to build base trees. In the end, we will also include the results from other two machine learning algorithms which are, Linear Regression and Gradient Boosted Decision Trees. These two algorithms we will not implement from the scratch but we will use in-built libraries of them from python's scikit-learn tool kit [6] to build our machine learning model based on the dataset we have. Shows the organization of this project. The following are some of the questions which will be answered in this project report: What are the different approaches people have used to predict various types of prices? Which machine learning techniques can be used in this area? What is Classification and Regression in machine learning? How will the new proposed system work? Classification.

## 2.LITERATURE SURVEY AND RELATED WORK

Price prediction is a popular problem. There are several price prediction systems which are used to predict different kinds of prices. Some of them include stock prices [7], [8], [9], [10], [11], home prices [12], [13], electricity prices [14], [15], [16] etc. The techniques used in above papershave been varied such as fuzzy logic, neural networks, genetic algorithms, Naive Bayesian method and others.

In the health domain, there have been numerous works on predicting medical prices in different contexts [17], [18], [19], [20], [21]. For example, Moran et al. [17] use generalized linear regression methods to predict Intensive Care Unit (ICU) costs and use patient demographics, DRG (Diagnostic Related Group), length of stay in the hospital and a few others as features.Sushmita et al. [18] attempt to predict future health-care costs of individuals based on their medical and cost history. The expected costs for the future 3-, 6-, 9-, and 12-months are projected. They apply linear regression and Random Forest regression analysis for cost prediction. Lahiri et al. [19] employ classification algorithms to predict whether an individual's health care costs will increase in the next year given the health care costs for the previous year. Researchers have also used hierarchical regression analysis to tackle price prediction problem. Multilevel linear regression is used to determine effects of patient and physician characteristics on diagnostic testing [22]. Hierarchical decision trees are used for classification tasks where theclass labels are hierarchical in nature [23].

The problem which will be addressed in this project is distinct from the problems addressed by other researchers. We will tackle the problem of predicting costs of treating DRGs at a hospital

located anywhere in the U.S. using Medicare payment datasets and picking only relevantattributes from the dataset which are useful for our problem.In the next chapter, we provide information regarding different machine learning techniques usedin the price prediction systems.

## 3 Implementation Study And PROPOSED WORK AND ALGORITHM

To predict medical insurance costs using python , you can use libraries like scikit-learn and pandas.first ,you need a dataset with information like age,bmi,smoking status etc.then,you can train a model that learns pattern from this data to predict the insurance costs. You can use regression model like linear regression and random forest.these model analyze the data and Make predictions base on the pattern s they find .finally you can evaluate the performance of your model and make predictions on new data.itsreporesent the data in the form of graphs and bars.

### 3.1 PROPOSED SYSTEM:

The purpose of being able to classify what activity a person is undergoing at a given time is to allow computers to provide assistance and guidance to a person prior to or while undertaking a task.
The difficulty lies in how diverse our movements are as we perform our day-to-day tasks.

There have been many attempts to use the various machine learning algorithms to accurately classify a person's health details, so much so that Google have created an Activity Recognition API for developers to embed into their creation of mobile applications

### 3.2 Algorithm

Regression:
The target variable or the label in a regression problem holds continuous values. That means the target variable is continuous e.g. 1240, 1256, 4800.89 etc. These continuous numerical values of this target variable are not finite. These values can range from any real value to another real value. The mapping function given in the previous section then would try to predict the continuous value in the given range of the target variable with the help of input variables.

For example, given a set of input variables, an algorithm would try to predict the house price. In this example the house price is a label and it can have any value within the given range of valuesof training data. There are many algorithms used to solve a regression problem. Some of the algorithms are Decision Trees usually called as Regression Trees, Random Forest Regression, Linear Regression etc.

Tree Approaches and Regression Algorithms:

A price prediction is a regression task because any kind of a price is a continuous value and as mentioned in the previous section, regression problem tries to predict a continuous value [24]. In a regression task the dependent variables may or may not be continuous valued but the final outcome being predicted should be continuous.
A typical price prediction system will take the set of input variables, apply the technique chosen to get the final outcome and based on that predicts the price. Some of these techniques are machine learning algorithms which are related to trees approaches. These algorithms are decision trees and ensemble methods like random forest regression, gradient boosted regression trees etc. These algorithms are popular because of their simplicity and their efficiency over other complicated techniques. We will discuss these algorithms as we are going to use them to solve our medical payment price prediction problem.

a. Decision Trees

Decision trees [25] are a class of popular machine learning models that are used for classification and regression. Fig. 2 shows a simple decision tree used for classification that uses features to

classify a person into two categories fit or unfit. The feature set contains age, whether a personeats a lot of pizzas and whether a person exercise in the morning.

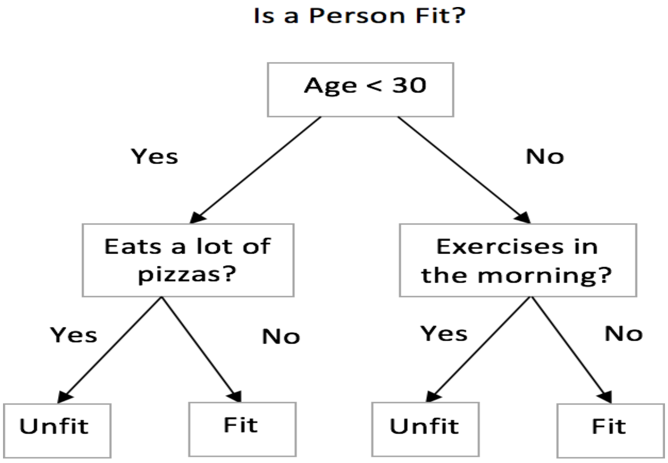


Fig 2. . An example of Decision Tree classification

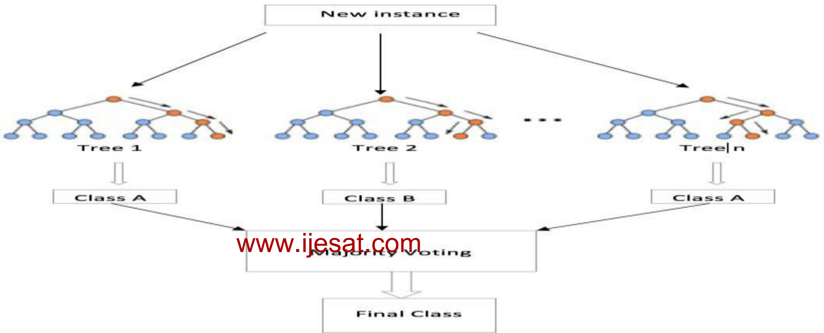
While using the decision trees one of the best practice is to calculate the importance of each feature [28]. The basic idea is to capture the relative importance of features in a particular dataset. Features are the different variables in a dataset whose values differentiate each row in the data. The feature having the highest importance gets the priority and data is split according tothat and so on. The constructed decision tree is subsequently used to classify or predict a new instance. If the decision tree is being used as a classifier, then the new instance is classified into one of the classes provided and if it is being used as a regressor then, it is used to predict the outcome.

A classification tree predicts the class of a new instance by taking the mode of all the class values at leaf nodes. A regression tree predicts the value by taking mean of all the values at leafnodes [28]. There are many algorithms to build regression trees, some of them are AID, CART,M5 and GUIDE [27]. AID and CART construct piecewise constant regression trees. Among these algorithms, CART uses binary split on the node where as others use multiple split on the node.

b. Ensemble Methods

The Random Forests [30], gradient-boosted trees [29] are called ensemble methods, for constructing multiple decision trees. Ensemble methods take multiple weak learners, such asdecision trees, and construct a strong learner from them such as random forest. The RandomForests and gradient-boosted trees both can be used for classification and regression task.

In ensemble methods a classification task for a new instance will be done by taking the majorityof votes from each tree. The class getting the highest votes will be chosen as the final target value for the



new instance. Fig.  shows a simple illustration of random forest classifier.

Fig 3.  An example of Random Forest
classification

On the other hand, in a regression task, the new instance is passed through all the trees and the outcomes from all the individual trees are aggregated to produce an overall outcome. Fig. shows a simple illustration of random forest regressor.

Like decision trees, we can calculate the importance of each feature in the ensemble method also.They are calculated by computing the importance of features of individual trees and then averaging them across the trees. Ensemble methods are more robust because they decrease the tendency of a single decision tree to overfit the training data.

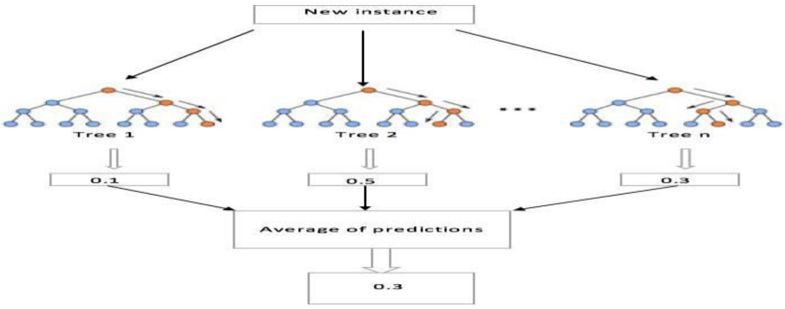


Fig 5.   An example of Random Forest regression

## **3.** .2 Unsupervised Learning:

The second type of machine learning technique is unsupervised learning. Unlike supervised learning, unsupervised learning algorithms don't learn from the input data. That is, the data in case of unsupervised learning don't have labels associated. An algorithm only takes input variables and finds patterns in the given data. It then tries to predict the right answer from thosepatterns. The most common examples of unsupervised learning are clustering algorithms and association rule mining algorithms. For example, in clustering the algorithm tries to find pattern in a given set of input variables. These patterns are called clusters. When the new input data comes for testing, an algorithm tries to put it into the right cluster formed during the training phaseIn the next chapter, we propose a medical price prediction system which will be used to predictmedical prices. Additionally, we will provide the details of the dataset information and feature selection from the dataset

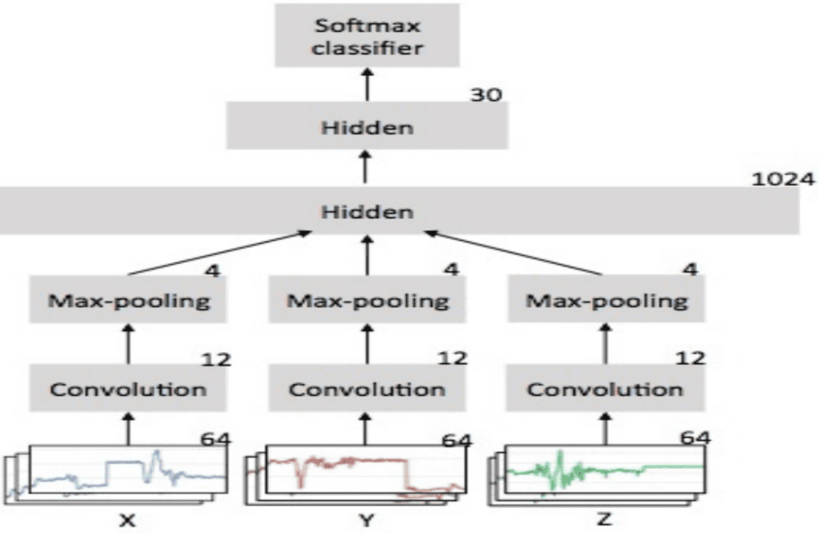
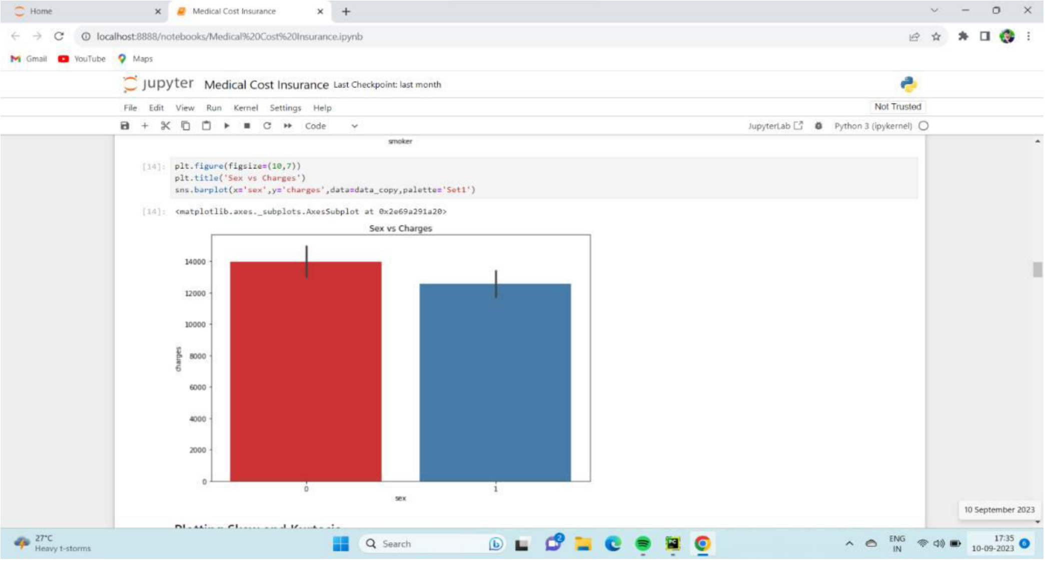


Fig6: - Flow chart showing entire process

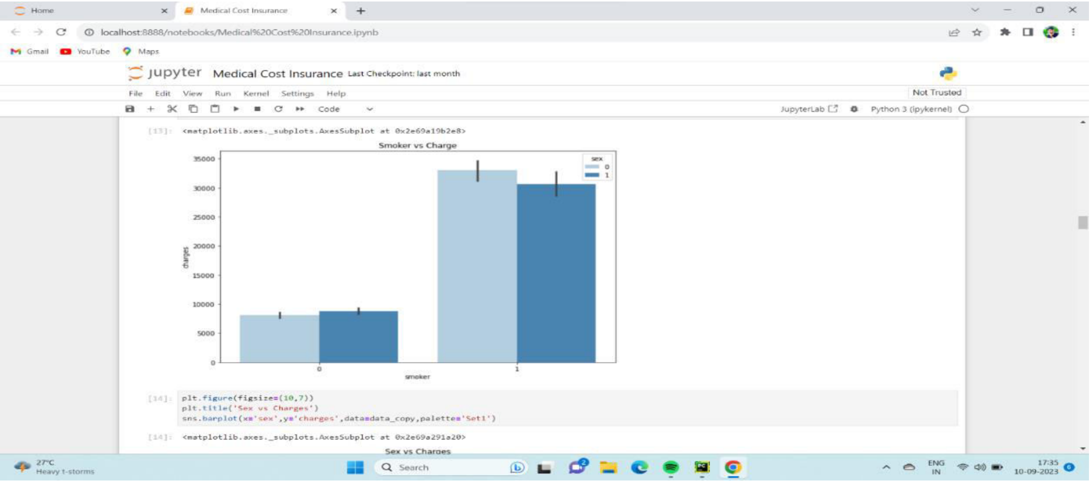
## 4.RESULTSANDDISCUSSION SCREENSHOTS



```
[14]: plt.figure(figsize=(10,7))
      plt.title('Sex vs Charges')
      sns.barplot(x='sex',y='charges',data=data_copy,palette='Set1')

[14]: <matplotlib.axes._subplots.AxesSubplot at 0x2e69a291a20>
```

Fig 7 :- The above image represents the data will show in bar graphs

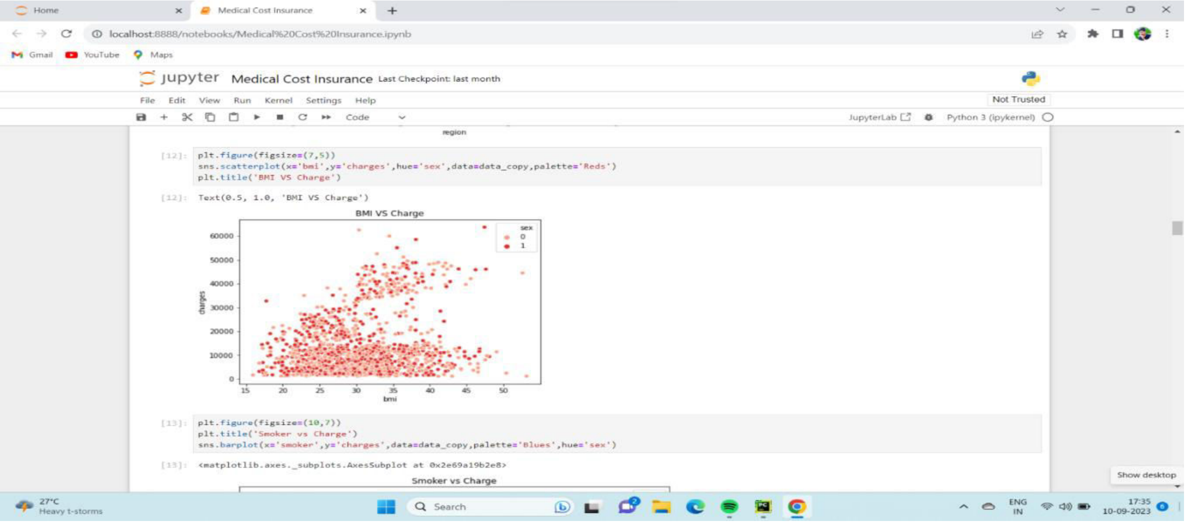


```
[13]: <matplotlib.axes._subplots.AxesSubplot at 0x2e69a19b2e8>

[14]: plt.figure(figsize=(10,7))
      plt.title('Sex vs Charges')
      sns.barplot(x='sex',y='charges',data=data_copy,palette='Set1')

[14]: <matplotlib.axes._subplots.AxesSubplot at 0x2e69a291a20>
```

Fig 8 :- The above image represents the data will show in bar graphs



```
[12]: plt.figure(figsize=(7,5))
      sns.scatterplot(x='bmi',y='charges',hue='sex',data=data_copy,palette='Reds')
      plt.title('BMI VS Charge')

[12]: Text(0.5, 1.0, 'BMI VS Charge')

[13]: plt.figure(figsize=(10,7))
      plt.title('Smoker vs Charge')
      sns.barplot(x='smoker',y='charges',data=data_copy,palette='Blues',hue='sex')

[13]: <matplotlib.axes._subplots.AxesSubplot at 0x2e69a19b2e8>
```

Fig 9:- The above image represents the data will show in plotted line

## 5. CONCLUSION

In this project, we have successfully implemented Regression Trees and Random Forest Regression algorithms from scratch to predict the medical prices from the input dataset. We also compared the results of Regression Trees, Random Forest Regression, Gradient Boosted Regression Trees and Linear Regression for the same dataset. From the results, we cannot conclude which model performed the best because the model performance can vary depending upon the configuration tried while testing. Hence, the model performing best for some configurations can give unsatisfactory results for some other configurations. Overall for the test configuration parameters, the order of performance of each model from the best to worst is Gradient Boosted Regression Trees, Random Forest Regression, Regression Trees and Linear Regression. The average medical payments predicted by Gradient Boosted Regression Trees, Random Forest Regression and Regression Trees are close to the actual values of payments.

### 5.1 FURTHER ENCHANCEMENT:

As possible future work for this project, we can add new features to the dataset we are already using. These newly added features can be totally new or can be derived from the dataset itself. For example, we can calculate distinct DRG counts feature from the DRG Definition column. It will give the count of each unique DRG in the entire dataset. We think it will be useful because, the dataset has the top 100 most frequently billed Diagnostic-Related Groups (DRGs) and a widevariation in the prices for a given DRG among different providers is observed. Hence, it will be important to know if the particular DRG has more impact on the project more features can improve the accuracy of prediction as we can add randomness in the selection of features while building the individual trees for the Random Forest algorithm.

Another addition in the future work can be, making the system even more scalable. Right now we are using few thousands of records to train and test the algorithm. In future, we can try to scale the algorithm for a larger dataset having at least a million records and see the results for it. To make the system scalable we can make use of distributed frameworks like Spark and Hadoop.These frameworks can handle big data efficiently.

### 6. REFERENCES

- ❖ A. Tike and S. Tavarageri. (2017). A Medical Price Prediction System using HierarchicalDecision Trees. In: IEEE Big Data Conference 2017. IEEE.
- ❖ “National Health Expenditures 2015 Highlights,” CMS.gov, 2015. [Online]. Available:https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends- and-Reports/NationalHealthExpendData/downloads/highlights.pdf
- ❖ J. Cubanski, C. Swoop and T. Neuman,“How Much Is Enough? Out-of-Pocket Spending Among Medicare Beneficiaries: A Chartbook,” The Henry J. Kaiser Family Foundation, Menlo Park, CA, 2014. [Online]. Available: http://files.kff.org/attachment/ how-much-is- enough-out-of-pocket-spending-among-medicare-beneficiaries-a-chartbook-
- ❖ J. Cubanski and T. Neuman,“The Facts on Medicare Spending and Financing,” The Henry J. Kaiser Family Foundation, Menlo Park, CA, 2017. [Online]. Available: http://files.kff.org/attachment/Issue-Brief-The-Facts-on-Medicare-Spending-and-Financing. [Accessed Nov. 2, 2017].
- ❖ The Henry J. Kaiser Family Foundation. (2017). Total Number of Medicare Beneficiaries. [Online]. Available: https://www.kff.org/ medicare/state-indicator/total-medicare- beneficiaries/. [Accessed Nov. 2, 2017].
- ❖ Scikit-learn.org. (2018). About us — scikit-learn 0.19.1 documentation. [online] Available at: http://scikit-learn.org/stable/about.html#people [Accessed 23 Apr. 2018].
- ❖ R. Hafezi, J. Shahrabi, and E. Hadavandi, “A bat-neural network multi- agent system (bnnmas) for stock price prediction: Case study of dax stock price,” Applied Soft Computing, vol. 29, pp. 196–210, 2015.
- ❖ H. Lee, M. Surdeanu, B. MacCartney, and D. Jurafsky, “On the importance of text analysisfor stock price prediction.” in LREC, 2014, pp. 1170–1175.