Medical Health Insurance Cost Prediction

by using Machine Learning

**Rakesh Nemu, Ankit Lakra, Manjari Gupta**

**Banaras Hindu University, India**

***Abstract*: Nowadays, the expenses of health care increases. *Insurance is a policy that eliminates or decreases loss costs occurred by various risks. As there are a greater quantity of new viruses stepping into human beings, there's a want to expect health costs.*** ***This form of prediction allows the governments to decide regarding health issues.***

***People are aware of the importance of health care costs. Various factors affect the coverage price. These issues make contributions to the insurance policy method. In this work, we can expand a scientific rate prediction gadget by use of machine learning algorithms as a way to resource in steerage sufferers to value powerful vendors and thereby scale down fitness spending.*** ***We build a model to predict the medical insurance cost of a person based on gender. This dissertation demonstrates how distinctive models can forecast insurance fees. And we will compare the results of models, for example Multiple Linear Regression, Polynomial regressor, Decision tree regressor, Random Forest Regressor, XGBoost, k-Nearest Neighbours, Gradient Boosting, and AdaBoost regressor. This paper offers the best approach to the XGBoost model with an MAE value of 2232.5097, RMSE value of 4073.399 and R -squared value of 0.8908.***

***Keywords*: *regression, machine learning, forecast, insurance ,Kaggle***

# INTRODUCTION

The goal of this project is to lets in someone to get a concept approximately the vital amount required according to their own fitness status. Later they could follow any medical insurance organisation and their schemes & blessings keeping in mind the expected quantity from our assignment. This could assist a person in focusing greater on the fitness component of a coverage rather than the futile part [7].

Health insurance (sometimes called health coverage) pays for some or all of the cost of the health services you receive, like doctors’ visits, hospital stays, and visits to the emergency room. It helps keep your health care costs predictable and affordable.[10] A health insurance agency can best make money if it collects extra than it spends at the hospital treatment of its beneficiaries. However, even though a few conditions are extra typical for certain segments of the population, clinical prices are tough to expect given that most money comes from uncommon situations of the patients [1].

Various factors evaluate insurance charges in lives of individuals. Each factor of those is critical. if any component is not noted whilst the amounts are computed, the policy changes universal. If human calculate he doesn't get proper accuracy. But insurers can do this task properly, who uses several tools to calculate the insurance premium. ML is one of the helping tools that helps to build model to formulate the policy. ML models can be learned by themselves. The model is trained on insurance data from the past [9]. 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 ml. our objective is to forecast insurance charges in this article. the value of insurance fees is based on different variables.

In this paper, we actually want to improve the accuracy of the model. Here, we have used several regressor algorithms which have been already applied in the available paper about this topic and also, we have applied polynomial regression in this model which has been not applied in others papers.

# DATASET

We have taken dataset from **Kaggle** site [2]. The data set includes seven attributes see **Table I** and sample in **Table II**; We have divided the data set in two parts in 3:1 ratio- training data and test data.

**Table I. Dataset overview [9]**

|  |  |
| --- | --- |
| Name | Description |
| age | Age of the customer |
| BMI | Body Mass Index |
| The Number of Kids | Number of children the client has |
| gender | Male / Female |
| smoker | whether a client is a smoker or not |
| region | where the client lives  southwest, southeast, northwest or northeast |
| Charges (target variable) | Medical Cost the client pays |

**Table II**

A screenshot of a computer

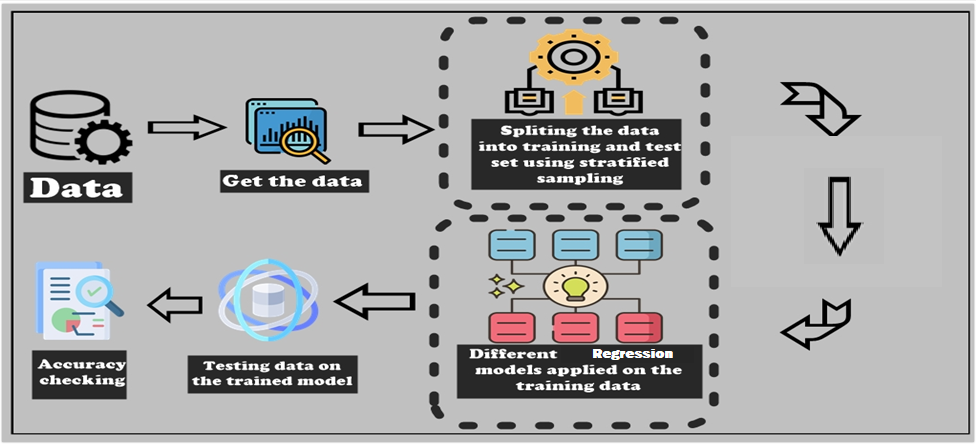
Description automatically generated with medium confidence

# DATA PRE-PROCESSING

***Data pre-processing*** is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms [12].

The major tasks of data pre-processing are given below: 

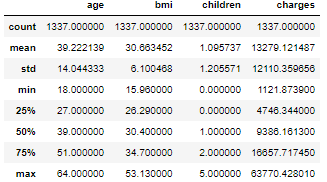
**Figure I (workflow of data pre-processing)**

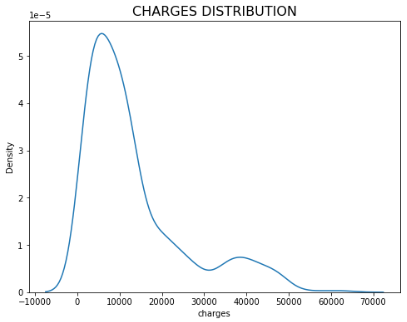
In this experiment, the dataset has been found for our analysis from Kaggle [2] which one of the most used platforms for collecting the data. It contains 1338 rows of data and the following columns: age, gender, BMI, children, smoker, region, insurance charges. In these features insurance charges is a dependent variable and the remaining features are called as independent variables. Since our dependent variable is continuous, so our model is a regression model. For this regression model we have used different kind of regressor machine learning algorithm. Our whole analysis is based on this flow chart given below: 

**Figure II (workflow of supervised regression model)**

Firstly, we have imported the libraries and loaded the dataset. Then, we have checked the missing value and the duplicated values in the dataset. There were no missing values and one duplicated value. Then we have dropped the duplicated value. The numeric variable of the dataset would then be examined.

**Descriptive statistic of numerical variable**

**Table III** From the above table we see that mean value is greater than the median value for the charges. This implies that the distribution of charges is right skewed



**Figure III**

Then we have used **Feature Encoding**. Here, we have changed the categorical features as **sex, smoker** by **label encoder** and also **region** feature has been changed into numeric value by **One-Hot-Encoding**. After this the region column transferred into for new column.

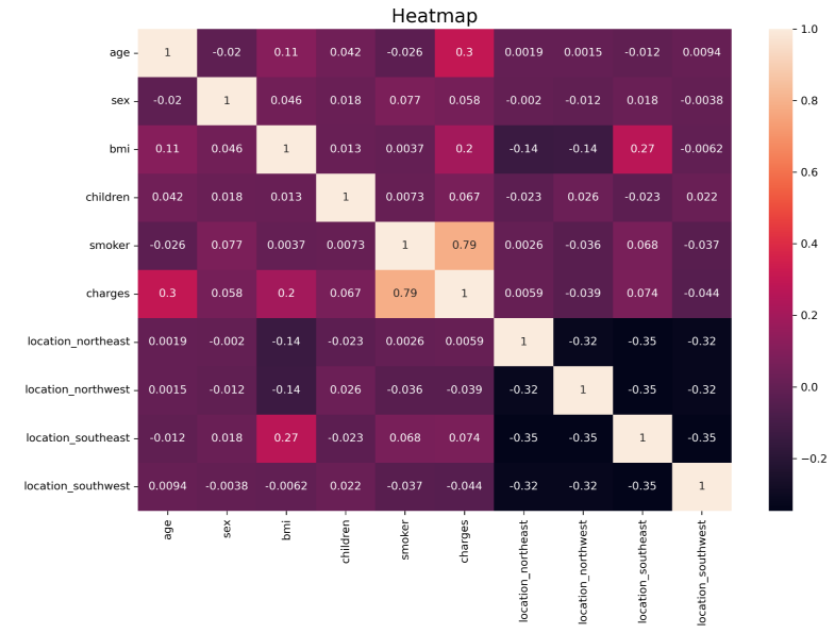
**Categorical variables after translated into numeric or binary values**

**Table IV**

|  |  |
| --- | --- |
| **Name** | **Description** |
| Age | Age of the client |
| BMI | Body mass Index |
| Number of Children | Number of children the client has |
| Sex | Male/Female  1= Male  0=Female |
| Smoker | Whether the client is smoker or not  1= Yes  0= No |
| location\_northeast | Where the client lives  1= Yes  0= No |
| location\_northwest | Where the client lives  1= Yes  0= No |
| location\_southeast | Where the client lives  1= Yes  0= No |
| location\_southwest | Where the client lives  1= Yes  0= No |
| Charges | Insurance chargers of client |

Then we checked the relationship between the numeric variables.

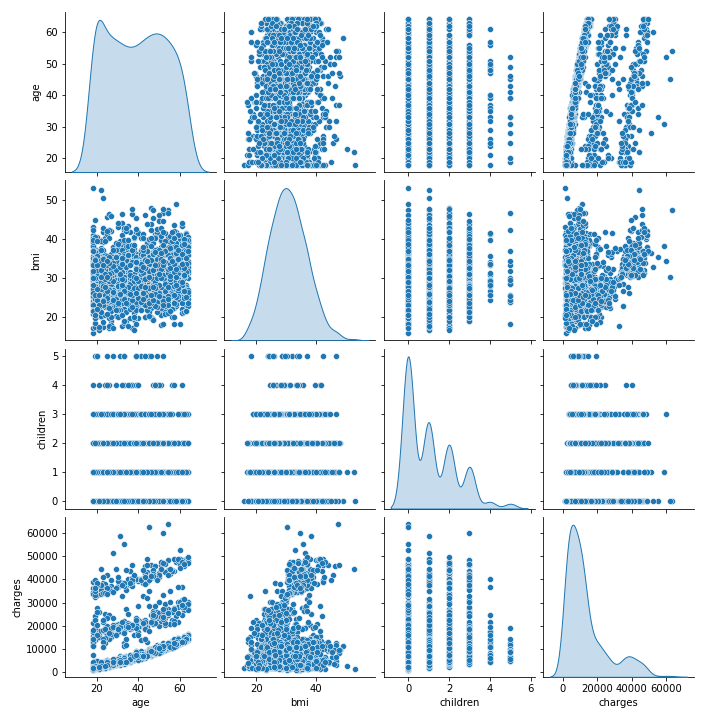
**Correlation between variables**



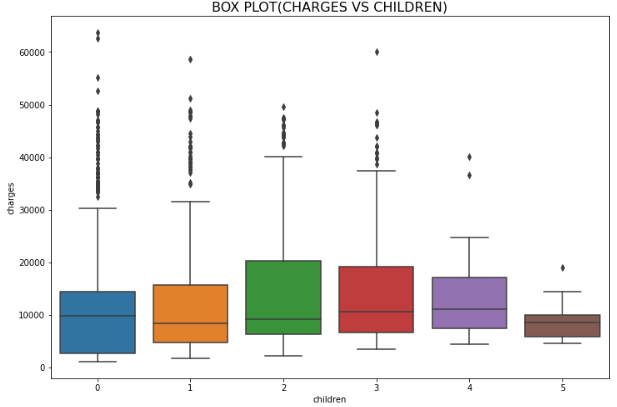
**Figure IV**

The above figure shows that variables which most influenced charges are age, bmi and smoker.

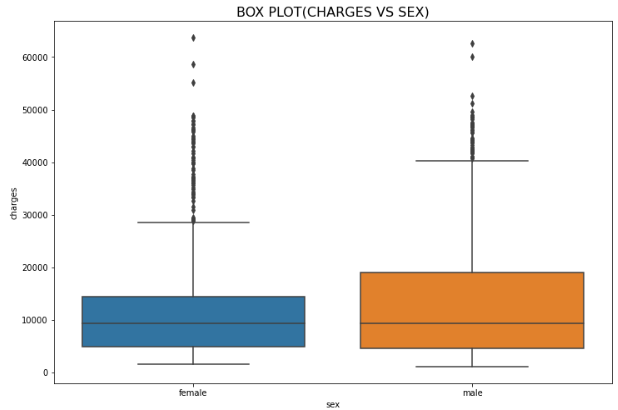
Then we examine the independent variable with dependent variable by scatter plot, hist plot, count plot and distribution plot.



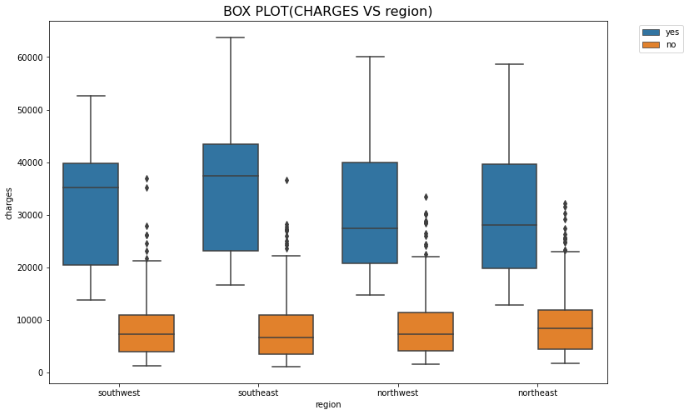
**Figure V**

****

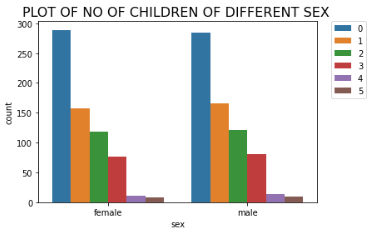
**Figure VI**

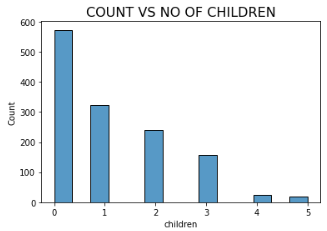
****

**Figure VII**

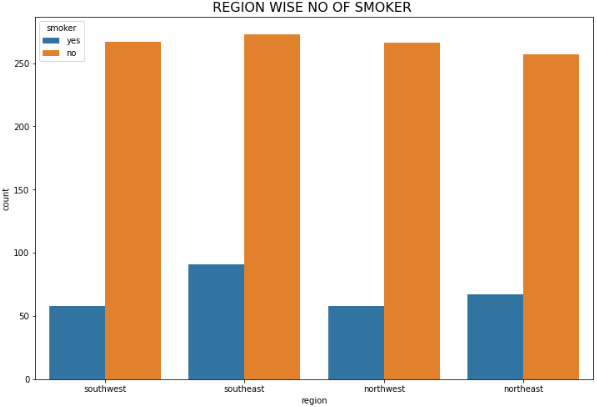


**Figure VIII**

Figure V shows the plot between independent and dependent variables. Figure VI shows the impact of the number of children on charges. Figure VII shows the impact of gender variable on charges Figure VIII shows the impact of the region variable on charges according to the smokers. Then we have plotted the count plot of categorical column. **Figure IX**



**Figure X**



**Figure XI**

Figure IX shows the plot of no. of children of different gender. Figure X shows the number of clients according to their children. Figure XI shows the number of smoker and non-smoker client in the different region.

# RELATED WORK

While we've a huge data set on which we would really like to carry out predictive analysis or pattern recognition, Machine learning is the way to move [8].

“Health Insurance Amount Prediction” by Nidhi Bhardwaj and Rishabh Anand in 2020, it was concluded that they have used three model but they found Gradient Boosting Regression model which is built upon decision tree is the best [7].

Lakshmana Rao, Chandra Sekhar Koppireddy and Vijay Kumar applied MLR with backward elimination technique and they observed that age and bmi are features which decides the dependent variable [20].

Anuja Tike [21] suggested a paper “MEDICAL PRICE PREDICTION SYSTEM" in 2018. In this paper, Regression, Trees, Random Forest Regression, Gradient Boosted Regression Trees and Linear Regression to predict the medical prices from the input dataset have been successfully implemented and also compared the results of those algorithms for the same dataset. Gradient Boosted Regression Trees performed the best (88% accuracy) in this paper.

Albers Uzila [1] suggested a paper"MEDICAL Cost PREDICTION " in 2021. For Linear Regression, feature engineering plays an important role to improve the model. In this paper, we apply this technique by making polynomial combinations of features with degree 2. No multicollinearity found in Linear Regression model, but many found in Polynomial Regression model. Some assumptions on Linear Regression may break down in the process.

Dimitris Bertsimas [20] et.al applied modern data-mining

methods provide quantifiable predictions of medical costs

and represented a powerful tool for the prediction of healthcare costs.

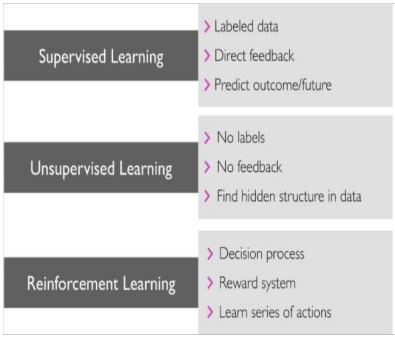
T Takeshima [21] et.al applied a statistical machine learning method to develop a medical cost prediction model. a regression model with least absolute shrinkage and selection operator (LASSO), was developed using the excess logarithmic cost as explained valuables, and dummy valuables of each disease and their interaction terms as explanatory valuables.

A research was done by Mohamed Hanafy [8] uses feature scaling in independent and dependent variable, various machine learning regression models and deep neural networks to forecast charges of health insurance based on specific attributes, on medical cost personal data set.

# REGRESSION MODELS

**Machine learning** can be defined as the process of teaching a computer system which allows it to make accurate predictions after the data is fed. It is the field of the study that gives computer the ability to learn without being explicitly programmed.

The different type machine learning system is shown by given figure: [6]



**Figure XII**

* ***Linear Regression:*** Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous or numeric variables. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression [14]. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.
* ***Decision trees Regression:*** Regression or classification models in decision tree regression builds in the form of a tree structure [7]. It is a tree-structured classifier with three types of nodes namely root node, interior node and leaf node. This algorithm is very useful for solving decision-related problems. Here dataset is segmented into subsets while at the same time an associated decision tree is incrementally developed. Here we have two results as decision node and leaf node. Decision node have two or more branches, each representing values. Leaf node represents Numerical targeted decision. Root node give the proper predicted result.
* ***Polynomial Regression:*** Polynomial Regression is a form of Linear regression known as a special case of Multiple linear regression which estimates the relationship as an nth degree polynomial [15]. The Polynomial Regression equation is given below:

y= …...+

It is a linear model with some modification in order to increase the accuracy. The dataset used in Polynomial regression for training is of non-linear nature. It makes use of a linear regression model to fit the complicated and non-linear functions and datasets. Hence, "In Polynomial regression, the original features are converted into Polynomial features of required degree (2, 3,...,n) and then modelled using a linear model."

* ***Random Forest :***  Random Forest is made by combining different model. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. An important feature of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.
* ***Gradient Boosting Regression:***Gradient boosting [6] is a type of boosting machine learning algorithm. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error.
* ***AdaBoost Regressor:***  An AdaBoost [4] regressor is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction. As such, subsequent regressors focus more on difficult cases. This class implements the algorithm known as AdaBoost.R2 [5]
* ***XGBoost (Extreme Gradient Boosting) :***  XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library[11]. XGBoost builds upon: supervised machine learning, decision trees, ensemble learning, and gradient boosting. XGBoost is a highly accurate implementation of gradient boosting that pushes the limits of computing power for boosted tree algorithms, being built largely for energizing machine learning model performance and computational speed.

K***-Nearest Neighbours:*** K nearest neighbours [19] is a simple algorithm that stores all available cases and predict the numerical target based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970’s as a non-parametric technique. Algorithm A simple implementation of KNN regression is to calculate the average of the numerical target of the K nearest neighbours. Another approach uses an inverse distance weighted average of the K nearest neighbours. KNN regression uses the same distance functions as KNN classification.

# IMPLEMENTATION AND RESULT

The objective of the analysis is to predict the medical insurance cost which depend on the features like age of the client, BMI of the client, number of children of the client, location where the client lived and also the client is smoker or not. These features contribute to our target variable prediction of insurance costs. For the measurement of the target variable, several regression models are applied in this analysis. At first the dataset is loaded. Then the data is pre-processed. Then we converted features of the dataset except target variables into independent variables and target variables into dependent variables. Then dataset is split into two part where the first part is called training dataset and the second one is called test dataset. For this analysis we took 75% of total data for training dataset and rest 25% for test dataset. After this we applied several machine learning algorithm one after another. During applying ML model , each of the model is trained with training dataset and then evaluated with test dataset. In this analysis we used Mean absolute error (MAE), root mean squared error (RMSE) and R-squared As a standard for evaluating and also for the accuracy of these model.

* ***MAE (Mean absolute error):***  MAE[16] is the mean of the absolute difference of the given value and the predicted values of the target variable over the dataset

MAE =

* ***RMSE (Root mean squared error):***  RMSE[16] is the square root of the mean of the squared of the difference of given values and the predicted value of target variable over the dataset.

RMSE =

Where, N= number of data present in the dataset

= predicted value of the target value

Y= given value of the target variable.

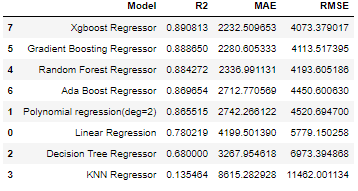
* ***R-Squared (co-efficient of the determination) :*** R-squared[16] represent the co-efficient of how well the values fit compared to the given values. The range of R-squared value is between 0 to 1 where 1 is the best possible score and 0 is worst score. If any model has higher R-squared value then the model is the better model.

For predicting the target variable we used various machine learning algorithm like Linear Regression, Polynomial Regression, Decision Tree Regressor, KNN Regressor, Random Forest Regressor, Support Vector Regressor, Gradient Boosting Regressor, Ad boost Regressor and XGBoost (extreme Gradient Boosting) regressor. All of these models are trained on the basis of training data and evaluated on test data. We compared different model according to the R-squared, MAE and RMSE in the below

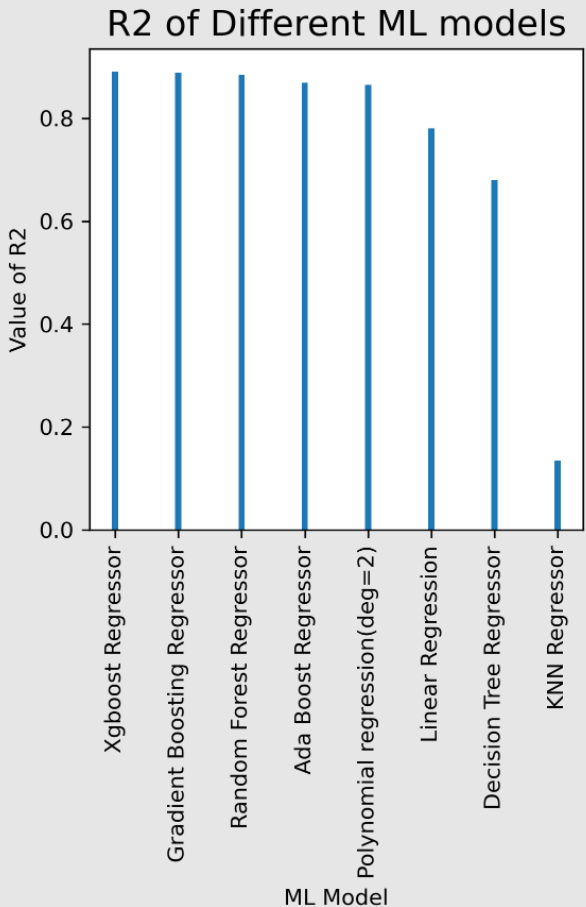
table :

# Model Performance

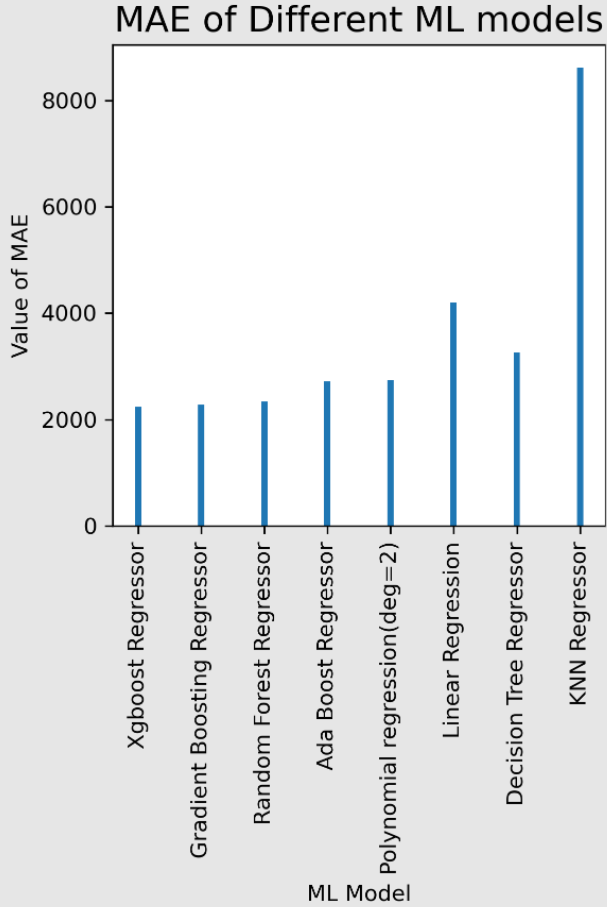
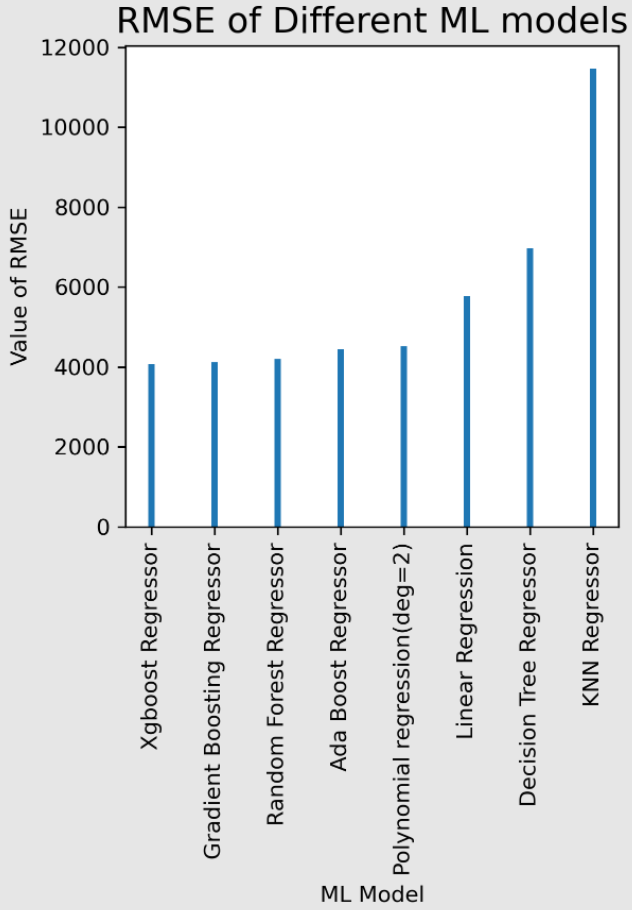
**Table V**



Now we compared the R-squared values of different ML model graphically. R-squared is most widely used measure for analysing. After applying several machine learning model we found that XGBoost regressor performed well on this data set.



# Figure XIII

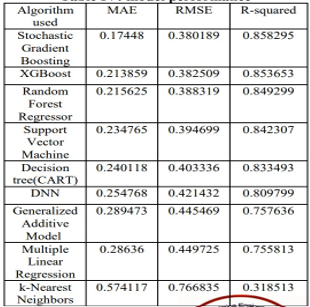
After this we compared the MAE and RMSE of different ML model graphically. MAE and RMSE are also the measures which decide performance of Machine learning models. After applying several machine learning model we found that XGBoost regressor performed well on this data set.  

**Figure XIV**

**DISCUSSION**

In this section, we discuss about result comparison with other available paper on the web about “Medical Health Cost Prediction”. A research done by Mohamed Hanafy [8] got the result is shown below:

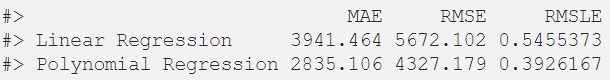
**Table VI**



From table VI we see that he has used several supervised regression ML algorithm and deep learning. Here stochastic Gradient Boosting gives best result(~85% accuracy).

A research done by Albers Uzila[1] got the result is shown below:

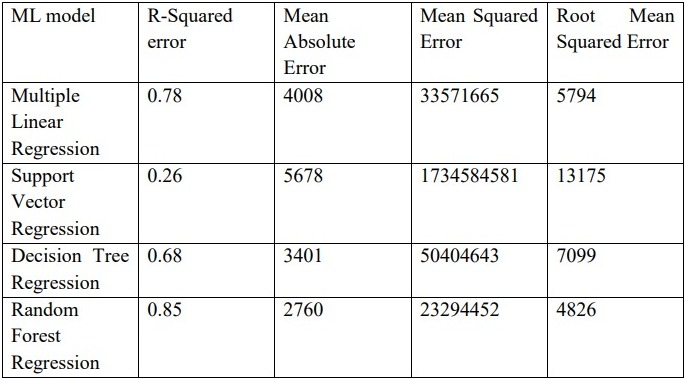
**Table VII**



From the above table VII, we see that he used only Linear Regression and polynomial Regression. He has got accuracy ~73% and ~83% in Linear Regression and polynomial Regression model respectively.

A research done by Lakshmanarao [20] et.al got the result is shown below:

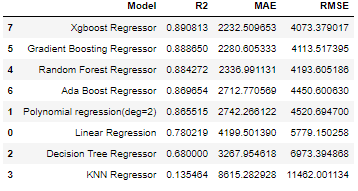
**Table VIII**



We see that he has used four algorithms which present in above table VIII. He has got better accuracy(~85%) in Random Forest Regression.

In this paper, we have used different supervised ML algorithms like Multiple Linear Regression, Polynomial regressor, Decision tree regressor, Random Forest Regressor, XGBoost, k-Nearest Neighbors, Gradient Boosting, and Ada boost regressor. Here we have got best accuracy ~89% in XGBoost model which is better accuracy than the above-mentioned papers about this topic. We have mentioned the performance of the model in the below table IX

**Table IX**



# CONCLUSION

In this paper we have used various machine learning regression models to predict health insurance charge based on specific attributes. We have got 89.08% accuracy by XGBoost (extreme Gradient Boosting) regressor which is no doubt a great accuracy achieved of this kind of problem.

The dataset which we have collected from Kaggle has 1338 data records. To increase the data records of our data set we have also applied data augmentation and got 28098 data records. But this augmentation was not successful because when we applied several ML model over the new dataset which we got by data augmentation, decreases the accuracy. The main reason of decreasing accuracy over the new dataset is the categorical features of the original dataset. After augmentation the values of the categorical features in the dataset is unbiased.

Out of all analysis XGBoost regressor model has given better result. We have done certain settings in parameter of the XGBoost regressor algorithm and achieved this accuracy. Predicting insurance charges based on certain factors help insurance policy providers to attract consumers and save time in formulating plans for every individual. 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. That’s why we used machine learning models.

# FUTURE WORK

However, we have reached a very good accuracy, it’s not the end. We will try our best to improve the model to get 100% accuracy where our model can correctly predict health insurance charged by the health insurance company. More data always increases the accuracy of any algorithm. So, getting more data is an issue. In future, if more data is available from public health insurance companies than we can improve the accuracy.

# REFERENCES

1. Albers Uzila Mar 23,

2021 DATA SCIENCE IN R

Medical Cost Prediction

https://towardsdatascience.com/medical-cost-prediction-4876e3449adf

1. Kaggle Medical Cost Personal Datasets. Kaggle Inc. [https://www.kaggle.com/mirichoi0218/insurance.](https://www.kaggle.com/mirichoi0218/insurance)
2. Y. Freund, R. Schapire, “A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting”, 1995..
3. Drucker, “Improving Regressors using Boosting Techniques”, 1997.
4. https://www.displayr.com/gradient-boosting-the-coolest-kid-on-the-machine-learning-block/.
5. Health Insurance Amount Prediction Nidhi Bhardwaj , Rishabh Anand Delhi, India Dr. Akhilesh Das Gupta Institute of Technology & Management.
6. https://www.semanticscholar.org/paper/Predictive-analytics-in-health-care-using-machine-Nithya-Ilango/248f27f0b53255e8958afb4bf1a5c20c3e236a4b
7. https://www.researchgate.net/publication/348559741\_Predict\_Health\_Insurance\_Cost\_by\_using\_Machine\_Learning\_and\_DNN\_Regression\_Models
8. https://www.medicareinteractive.org/get-answers/medicare-basics/medicare-overview/introduction-health-insurance
9. https://www.nvidia.com/en-us/glossary/data-science/xgboost/#:~:text=XGBoost%2C%20which%20stands%20for%20Extreme,%2C%20classification%2C%20and%20ranking%20problems
10. https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/
11. Algorithmic Prediction of Health Care Costs and Discovery of Medical Knowledge, Dimitris Bertsimas.

https://www.researchgate.net/publication/228605587\_Algorithmic\_Prediction\_of\_Health\_Care\_Costs\_and\_Discovery\_of\_Medical\_Knowledge.

1. JavaTpoint <https://www.javatpoint.com/linear-regression-in-machine-learning>
2. Polynomial regressor <https://www.analyticsvidhya.com/blog/2021/07/all-you-need-to-know-about-polynomial-regression/>
3. Prediction of medical costs using regression algorithms A. Lakshmanarao, Chandra Sekhar Koppireddy, G.Vijay Kumar
4. Random forest https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
5. KNN:- <https://medium.com/analytics-vidhya/k-neighbors-regression-analysis-in-python-61532d56d8e4>
6. Prediction of medical costs using regression algorithms A. Lakshmana Rao, Chandra Sekhar Koppireddy, G.Vijay Kumar
7. Bertsimas, M. V. Bjarnad´ottir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala, and G.Wang, “Algorithmic prediction of health-care costs,” Operations Research, vol. 56, no. 6, pp. 1382–1392, 2008.<https://people.csail.mit.edu/gjw/papers/healthcare.pdf>
8. A MEDICAL PRICE PREDICTION SYSTEM - Anuja Tike, San Jose State University.
9. Development of Medical Cost Prediction Model Based on Statistical Machine Learning Using Health Insurance Claims Data - Takeshima T, Keino S, Aoki R, Matsui T, Iwasaki K https://doi.org/10.1016/j.jval.2018.07.738