ENHANCING THE ACCURACY FOR MEDICAL COST TO PREDICT THE HEALTH INSURANCE USING POLYNOMIAL REGRESSION ALGORITHM OVER RANDOM FOREST REGRESSION ALGORITHM

Vengala Rashmika[1], M. Amanullah[2*]

Article History: Received: 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

Aim: The main objective of the research study is to improve the accuracy for Medical costs for Health Insurance using Polynomial Regression Algorithm compared with Random forest Regression Algorithm.
Materials and Methods: The dataset needed for the Medical cost prediction for health insurance is acquired from Google's Kaggle Website. The data set columns have the columns patient name, age, sex, bmi, smoker, children, region. In these features insurance charges are dependent variables and the remaining features are called independent variables. In regression analysis, predict the values of dependent variables using independent variables. The data sets are imported and Novel Polynomial regression Algorithm and Random forest regression Algorithms are tested. The number of groups are 2 for Two Algorithms with the G-power value of 80%. The sample size is 49 per group.
Results: The results are acquired in the form of accuracy for the inputs provided. The IBM SPSS tool is used in order to obtain the results. From these results the author has obtained, a statistical significance difference was observed between the Novel Polynomial Regression and has an accuracy of 84.12%. Random forest Regression Algorithm 83.17%, which is more accurate than the value. The independent sample T-Test was performed to find the mean, standard deviation, standard error mean significance between the groups.
Conclusion: In this paper, based on the results obtained, the Polynomial Regression Algorithm has more accuracy than the Random forest Regression Algorithm.

Keywords: Medical Cost, Health Insurance, Random forest Regression, GDP, Novel Polynomial Regression, Accountability, Machine learning.

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and technical Science, Saveetha University, Chennai, Tamil Nadu, India. Pincode: 602105.
[2*]Project Guide, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India. Pincode: 602105.

1. Introduction

In this project our goal is to predict medical prices based on the data obtained. In the first few chapters of this report, compare the work of various authors in the area of price prediction and also provide the information in detail (Breiman 2017), about some of the techniques used in the healthcare domain to predict the health care prices (Miner et al. 2014). Later, the design is proposed of a new system which will use Medicare payment datasets. The proposed system can be called a medical price prediction system (Gerds and Kattan 2021). 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 2 providers. It can be used by Medicare administrators to forecast expenditure for future months and years and plan the budget accordingly (Shiny Irene et al. 2021). 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. The proposed system is built by implementing two machine learning algorithms from scratch (Lewis 2007). The first algorithm is a regression tree and the second one is random forest regression. While implementing the Random forest regression algorithm, the Regression tree algorithm is used to build base trees. In the end, the results will be included from two other machine learning algorithms which are Linear Regression and Gradient Boosted Decision Trees (Tike and Tavarageri 2017). These two algorithms will not be implemented from scratch but will use in-built libraries of them from python's scikit-learn tool kit to build our machine learning model based on the dataset. They apply novel polynomial regression and Random Forest regression analysis for cost prediction.

Employ classification algorithms (Kim et al. 2021) 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 (Ericta 2007) to tackle price prediction problems. Multilevel linear regression is used to determine effects of patient and physician characteristics on diagnostic testing (Austin and Sutton 2018). Hierarchical decision trees are used for classification tasks where the class labels are hierarchical in nature.Our team has extensive knowledge and research experience that has translated into high quality publications(Pandiyan et al. 2022; Yaashikaa, Devi, and Kumar 2022; Venu et al. 2022; Kumar et al. 2022; Nagaraju et al. 2022; Karpagam et al. 2022; Baraneedharan et al. 2022; Whangchai et al. 2022; Nagarajan et al. 2022; Deena et al. 2022)

The problem in the existing research of Medical cost prediction for Health Insurance is less accuracy. There are certain algorithms with more accuracy when comparing it with existing ones. The main aim of the study is to improve the accuracy of Health insurance by implementing a Novel Polynomial Algorithm.

2. Materials and Methods

The proposed work is done in the Machine Learning lab, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. The number of groups is 2 for two algorithms. The sample size is 40 per group in total 80.

The dataset named 'Medical Insurance cost' ("Cost-Effectiveness Analysis and Cost-Benefit Analysis" 2013a) is downloaded from Google Kaggle Website. The data in this dataset explains about the Health Insurance provided for each patient varying with the factors. Many factors that affect how much you pay for health insurance are not within your control ("Cost-Effectiveness Analysis and Cost-Benefit Analysis" 2013b). Health insurance is calculated based on the patient's expenses. In this database, the information about the patients and the analytics about the patients with different factors are obtained (Qudsi 2015). Price prediction history shows that authors have used machine learning techniques in this domain extensively (Sohn et al. 2022). Health domain is no exception for this where medical prices are being predicted using health related data. Broadly machine learning techniques are categorized into two types of learning, supervised and unsupervised learning.

The Health Insurance is an important eye-opener for emergency needs during accidents and disease pandemic situations (Bondio, Sporing, and Gordon 2020). Many of the people will lag to hit financially and to bear the operational censoring expenses during treatment. The need for health insurance changes from youth to old age depending on your lifestyle and genetics (Bondio, Sporing, and Gordon 2020; Bergua et al. 2022). Random forest regression (RFR) model is a collection of multiple decision trees. RFR model is an estimator that fits several classifying decisions on the subsamples of the data and uses averaging criteria to improve the accuracy and control overfitting problems (Muremyi et al. 2020). In the case of a classification problem, RFR uses voting criteria. Each tree in the RFR makes its prediction, and at

the end, a class is assigned to a new test point based on the maximum voting (Muremyi et al. 2020; Kim et al. 2021). In the case of regression, it takes an average of all the numeric values predicted by the individual decision trees. In this way, it improves the accuracy and controls the overfitting of a model

Polynomial Regression

Polynomial Regression is a Machine learning based algorithm which is also a form of Linear regression model (van Mens et al. 2022). Our model can be improved by features of the prediction, specifically, by making new features that capture the interactions between existing features (Misra et al. 2021). This is called Polynomial regression. The idea is to generate a new feature matrix consisting of all Novel polynomial combinations of the features with degree less than or equal to the specified degree.

Step 1: Import the Packages required.
Step 2: Import the dataset into the code environment.
Step 3: Assign the features of dataset such as Age, Sex, Bmi, Children, Smoking, Region. The next process is checking the data for correction. After the corrections store the data into dataframes. Since predicting the insurance costs, charges will be our target feature.
Step 4: Once after importing the data, processes such as encoding are to be performed. The dataset should be chosen and start pre-processing the input so that the model can be used.
Step 5: In this Polynomial and Random forest regression algorithms the author uses the Backward Elimination method in order to work our way down.
Step 6: The determination of required parameters are done so that the model is good to fit. The parameters taken are predicted and performed.
Step 7: Further analysis is performed and the measurement of accuracy is done successfully.

Random Forest Regression

Random Forest algorithm can also be used for both regression as well as classification tasks. It is a supervised learning algorithm where the algorithm tests on some data and tests its performance on new data (Jain and Chatterjee 2020; Kim et al. 2021). The problem with decision trees is they can be easily overfit. Random forest tries to avoid this problem of overfitting. As the name suggests Random Forest consists of a number of trees. Decision Trees act as a weak learner for GDP. The strategy used while building the random forest is creating a strong learner from multiple weak learners which will minimize the errors produced by weak learners. Random Forest tries to minimize overfitting by introducing randomness in the creation of trees in the forest. The randomness is introduced in selection of GDP features and selection of subsamples used while building each tree.

Step 1 : Import the packages required.
Step 2 : Import the dataset taken from kaggle into the code environment.
Step 3: Assign the features of dataset such as Age, Sex, Bmi, Children, Smoking, Region. The next process is checking the data for correction.After the corrections store the data into dataframes. Since predicting the insurance costs, charges will be our target feature.
Step 4: Once after importing the data, processes such as encoding are to be performed. The dataset should be chosen and start pre-processing the input so that the model can be used.
Step 5: Import the Random forest regression algorithm dataset from the kaggle and predict the output for the testing GDP datasets.
Step 6: The determination of required parameters are done so that the model is good to fit. The parameters taken are predicted and performed.
Step 7: Further analysis is performed and the measurement of accuracy is done successfully.

For Polynomial Regression Algorithm, the test size is 40%of the total dataset and the remaining of 60% is used for training the datasets. Accuracy of both the algorithms are tested from sample sizes of 40 to 60. The dataset used for this paper on Machine Learning based Algorithms are obtained from Google's official dataset website Kaggle.

Statistical Analysis

The statistical software used for performing analysis is IBM SPSS version 21.0. IBM SPSS is a statistical software tool used for the analysis of data. The datasets are normalized and then the data is converted into arrays. The number of clusters needed are visualized and analyzed and the existing algorithms are obtained. For the Polynomial Regression algorithm, it is observed that if the number of censoring iterations increased in the GDP, then the error rate decreased and accuracy increased. It is declared that the Polynomial Regression Algorithm shows higher value

compared with the Random forest Regression Algorithm.

3. Results

In Table 1 a file collection of people with the charges obtained and the details of the people is given.The dataset is taken from Kaggle and it contains Age, Sex, BMI, Children, Smoking, Region and charges obtained by this bases using Machine Learning. The age column is the respective number of each patient with different ages, bmi of the patients, and the smoking cause for charges. These the bmi and the charges are represented in numerical values which are taken from the dataset collected on the first format. The statistical comparison of the charges with respect to the region and the smoking cause using two sample groups was done through SPSS version 21. Analysis was done for mean, standard deviation, independent T-Test.

The Outcome of the Polynomial Regression and Random forest Regression algorithm. Which predicted values are compared to the values and these outcomes are shown as tables and bar graphs.

Fig 1 shows the violin plots between real estate and medical prices for it.

Fig 2 displaying the distribution of charges with accuracies in medical cost insurance.

Fig 3 graph is shown with feature importances in smoking, bmi, age, children, region, sex categories.

And have noticed that the sex and region dont have noticeble differences for each category terms of charges given. It is observed that there is an increasing trend in the charges as the number of children increases. Lastly, Smoker seems to make a significant change to charges given by Health Insurance .

The bar chart plotted with the accuracies of both the algorithms for different sample sizes is represented. The bar chart is plotted by taking algorithms as x-axis and accuracy as y-axis. The bar chart shows that the Polynomial Regression Algorithm is more accurate than the Linear Regression Algorithm.

The last row shows the average of accountability of accuracy of both the algorithms. At sample size 49, the average accuracy of Polynomial Regression Algorithm is 84.12% and Random forest Regression Algorithm is 83.17%.

4. Discussion

From the results of the study the Polynomial Regression Algorithm has a better performance than the Random forest Regression Algorithm. Polynomial Regression has an accuracy of 84.12% whereas Random forest Regression has an accuracy of 83.17%.

Random forest regression (RFR) model was also applied in the same way as other models(Zhao et al. 2021). The RFR model showed an MSE of 0.76 and an R-square score of 0.987 for the training data. However, it showed an MSE of 5 and an R-square score of 0.92 for the test data(Etzioni, Mandel, and Gulati 2021). These results indicate the superior predictive performance of the RFR method as compared to other models.

The results were compared between Regression models, Random Forest Regression and Novel polynomial Regression for the same dataset. According to the results, it is not concluded that which model performed the best because the model performance can vary depending upon the GDP 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 Random Forest Regression. The average medical payments predicted by Random Forest Regression and Novel Polynomial Regression are close to the actual values of payments.

5. Conclusion

In this paper, the results obtained in executing several Algorithms based on the various data samples using the Novel Polynomial Regression Algorithm (84.12%) and Linear Regression Algorithm (83.17%) are presented. The Polynomial Regression Algorithm was used to test the accuracy of medical charges for Health Insurance and was shown to be more accurate than the Random forest Regression Algorithm.

Declarations

Conflict of Interests

No conflict of interest in this manuscript.

Authors Contributions

Author VR was involved in data collection, data analysis, and manuscript writing. Author SS was involved in conceptualization, data validation, and critical review of the manuscript.

Acknowledgement

We would like to thank our management, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing facilities and opportunities for our research study.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. The Big Event
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences
4. Saveetha School of Engineering

6. References

Austin, Zubin, and Jane Sutton. 2018. *Research Methods in Pharmacy Practice: Methods and Applications Made Easy*. Elsevier Health Sciences.

Baraneedharan, P., Sethumathavan Vadivel, C. A. Anil, S. Beer Mohamed, and Saravanan Rajendran. 2022. "Advances in Preparation, Mechanism and Applications of Various Carbon Materials in Environmental Applications: A Review." *Chemosphere*. https://doi.org/10.1016/j.chemosphere.2022.134596.

Bergua, Valérie, Céline Meillon, Karine Pérès, Jean-François Dartigues, Jean Bouisson, and Hélène Amieva. 2022. "Routinization: Risk Factor or Marker of Adjustment to Negative Health Issues?" *International Journal of Geriatric Psychiatry* 37 (3). https://doi.org/10.1002/gps.5682.

Bondio, Mariacarla Gadebusch, Francesco Sporing, and John-Stewart Gordon. 2020. *Medical Ethics, Prediction, and Prognosis: Interdisciplinary Perspectives*. Routledge.

Breiman, Leo. 2017. *Classification and Regression Trees*. Routledge.

"Cost-Effectiveness Analysis and Cost-Benefit Analysis." 2013a. *Medical Decision Making*. https://doi.org/10.1002/9781118341544.ch10.

———. 2013b. *Medical Decision Making*. https://doi.org/10.1002/9781118341544.ch10.

Deena, Santhana Raj, A. S. Vickram, S. Manikandan, R. Subbaiya, N. Karmegam, Balasubramani Ravindran, Soon Woong Chang, and Mukesh Kumar Awasthi. 2022. "Enhanced Biogas Production from Food Waste and Activated Sludge Using Advanced Techniques – A Review." *Bioresource Technology*. https://doi.org/10.1016/j.biortech.2022.127234.

Ericta, Natasha C. 2007. *Hierarchical Regression Model on Evaluating Effectiveness of Art Intervention Programs in Schools*.

Etzioni, Ruth, Micha Mandel, and Roman Gulati. 2021. *Statistics for Health Data Science: An Organic Approach*. Springer Nature.

Gerds, Thomas A., and Michael W. Kattan. 2021. *Medical Risk Prediction Models: With Ties to Machine Learning*. CRC Press.

Jain, Vishal, and Jyotir Moy Chatterjee. 2020. *Machine Learning with Health Care Perspective: Machine Learning and Healthcare*. Springer Nature.

Karpagam, M., R. Beaulah Jeyavathana, Sathiya Kumar Chinnappan, K. V. Kanimozhi, and M. Sambath. 2022. "A Novel Face Recognition Model for Fighting against Human Trafficking in Surveillance Videos and Rescuing Victims." *Soft Computing*. https://doi.org/10.1007/s00500-022-06931-1.

Kim, Joung Ouk Ryan, Yong-Suk Jeong, Jin Ho Kim, Jong-Weon Lee, Dougho Park, and Hyoung-Seop Kim. 2021. "Machine Learning-Based Cardiovascular Disease Prediction Model: A Cohort Study on the Korean National Health Insurance Service Health Screening Database." *Diagnostics (Basel, Switzerland)* 11 (6). https://doi.org/10.3390/diagnostics11060943.

Kumar, P. Ganesh, P. Ganesh Kumar, Rajendran Prabakaran, D. Sakthivadivel, P. Somasundaram, V. S. Vigneswaran, and Sung Chul Kim. 2022. "Ultrasonication Time Optimization for Multi-Walled Carbon Nanotube Based Therminol-55 Nanofluid: An Experimental Investigation." *Journal of Thermal Analysis and Calorimetry*. https://doi.org/10.1007/s10973-022-11298-4.

Lewis, Mitzi. 2007. *Stepwise Versus Hierarchical Regression: Pros and Cons*.

Mens, Kasper van, Sascha Kwakernaak, Richard Janssen, Wiepke Cahn, Joran Lokkerbol, and Bea Tiemens. 2022. "Predicting Future Service Use in Dutch Mental Healthcare: A Machine Learning Approach." *Administration and Policy in Mental Health* 49 (1): 116–24.

Miner, Linda, Pat Bolding, Joseph Hilbe, Mitchell Goldstein, Thomas Hill, Robert Nisbet, Nephi Walton, and Gary Miner. 2014. *Practical Predictive Analytics and Decisioning Systems for Medicine: Informatics Accuracy and Cost-Effectiveness for Healthcare Administration and Delivery Including Medical Research*. Academic Press.

Misra, Debdipto, Venkatesh Avula, Donna M. Wolk, Hosam A. Farag, Jiang Li, Yatin B. Mehta, Ranjeet Sandhu, et al. 2021. "Early Detection of Septic Shock Onset Using Interpretable Machine Learners." *Journal of*

Clinical Medicine Research 10 (2). https://doi.org/10.3390/jcm10020301.

Muremyi, Roger, Dominique Haughton, Ignace Kabano, and François Niragire. 2020. "Prediction of out-of-Pocket Health Expenditures in Rwanda Using Machine Learning Techniques." *The Pan African Medical Journal* 37 (December): 357.

Nagarajan, Karthik, Arul Rajagopalan, S. Angalaeswari, L. Natrayan, and Wubishet Degife Mammo. 2022. "Combined Economic Emission Dispatch of Microgrid with the Incorporation of Renewable Energy Sources Using Improved Mayfly Optimization Algorithm." *Computational Intelligence and Neuroscience* 2022 (April): 6461690.

Nagaraju, V., B. R. Tapas Bapu, P. Bhuvaneswari, R. Anita, P. G. Kuppusamy, and S. Usha. 2022. "Role of Silicon Carbide Nanoparticle on Electromagnetic Interference Shielding Behavior of Carbon Fibre Epoxy Nanocomposites in 3-18GHz Frequency Bands." *Silicon*. https://doi.org/10.1007/s12633-022-01825-1.

Pandiyan, P., R. Sitharthan, S. Saravanan, Natarajan Prabaharan, M. Ramji Tiwari, T. Chinnadurai, T. Yuvaraj, and K. R. Devabalaji. 2022. "A Comprehensive Review of the Prospects for Rural Electrification Using Stand-Alone and Hybrid Energy Technologies." *Sustainable Energy Technologies and Assessments*. https://doi.org/10.1016/j.seta.2022.102155.

Qudsi, Dini Hidayatul. 2015. *Predictive Data Mining of Chronic Diseases Using Decision Tree: A Case Study of Health Insurance Company in Indonesia*.

Shiny Irene, D., V. Surya, D. Kavitha, R. Shankar, and S. John Justin Thangaraj. 2021. "An Intellectual Methodology for Secure Health Record Mining and Risk Forecasting Using Clustering and Graph-Based Classification." *Journal of Circuits Systems and Computers* 30 (08): 2150135.

Sohn, Minsung, Daseul Moon, Patricia O'Campo, Carles Muntaner, and Haejoo Chung. 2022. "Who Loses More? Identifying the Relationship between Hospitalization and Income Loss: Prediction of Hospitalization Duration and Differences of Gender and Employment Status." *BMC Public Health* 22 (1): 232.

Tike, Anuja, and Sanket Tavarageri. 2017. "A Medical Price Prediction System Using Hierarchical Decision Trees." *2017 IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/bigdata.2017.8258396.

Venu, Harish, Ibham Veza, Lokesh Selvam, Prabhu Appavu, V. Dhana Raju, Lingesan Subramani, and Jayashri N. Nair. 2022. "Analysis of Particle Size Diameter (PSD), Mass Fraction Burnt (MFB) and Particulate Number (PN) Emissions in a Diesel Engine Powered by Diesel/biodiesel/n-Amyl Alcohol Blends." *Energy*. https://doi.org/10.1016/j.energy.2022.123806.

Whangchai, Niwooti, Daovieng Yaibouathong, Pattranan Junluthin, Deepanraj Balakrishnan, Yuwalee Unpaprom, Rameshprabu Ramaraj, and Tipsukhon Pimpimol. 2022. "Effect of Biogas Sludge Meal Supplement in Feed on Growth Performance Molting Period and Production Cost of Giant Freshwater Prawn Culture." *Chemosphere* 301 (August): 134638.

Yaashikaa, P. R., M. Keerthana Devi, and P. Senthil Kumar. 2022. "Advances in the Application of Immobilized Enzyme for the Remediation of Hazardous Pollutant: A Review." *Chemosphere* 299 (July): 134390.

Zhao, Shirong, Jamie Browning, Yan Cui, and Junling Wang. 2021. "Using Machine Learning to Classify Patients on Opioid Use." *Journal of Pharmaceutical Health Services Research: An Official Journal of the Royal Pharmaceutical Society of Great Britain* 12 (4): 502–8.

Tables and Figures

Table 1. Represents the File containing details about the patients with factors as age, sex, bmi, children, smoker, region and the charges based on it.

S.No	Age	Sex	Bmi	Children	Smoker	Region	Charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200

3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Table 2. Statistical results of Polynomial Regression and Random Forest Regression algorithms. Mean accuracy value, standard deviation and standard error mean for PRA and RFRA algorithms are obtained for 10 iterations. It is observed that the PRA (84.12%) algorithm performed better than the RFRA (83.17%) algorithm.

Algorithms (Accuracy)	**Sample (N)**	**Mean**	**Std Deviation**	**Std Error Mean**
Polynomial Regression	10	84.1190	3.11967	.98653
Random forest Regression	10	83.1670	2.96446	.93744

Table 3. Independent sample t-test of the significance level Polynomial Regression Algorithm and Random forest Regression algorithm results with two tailed significant values (p=<.001). Therefore both the PRA an RFRA algorithms have a significance level less than 0.05 with a 95% confidence interval.

Accuracy	**Levene's Test for Equality of Variances**		**T-test of Equality of Means**					**95% of the confidence interval of the Difference**	
	F	**Sig.**	**t**	**df**	**Sig (2-tailed)**	**Mean Difference**	**Std Error Difference**	**Lower**	**Upper**
Equal Variance Assumed	.040	.843	.700	18	<.001	.95200	1.36090	-1.90714	3.81114
Equal Variance Not Assumed			.700	17.953	<.001	.95200	1.36090	-1.90767	3.81167

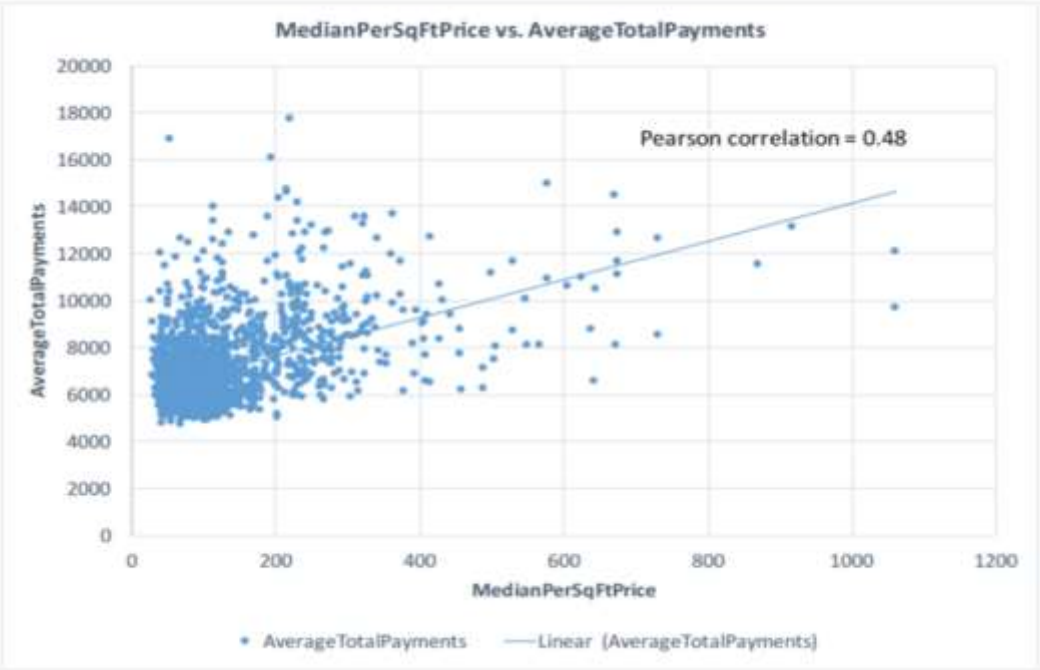

Fig. 1. The relation between real estate and medical prices

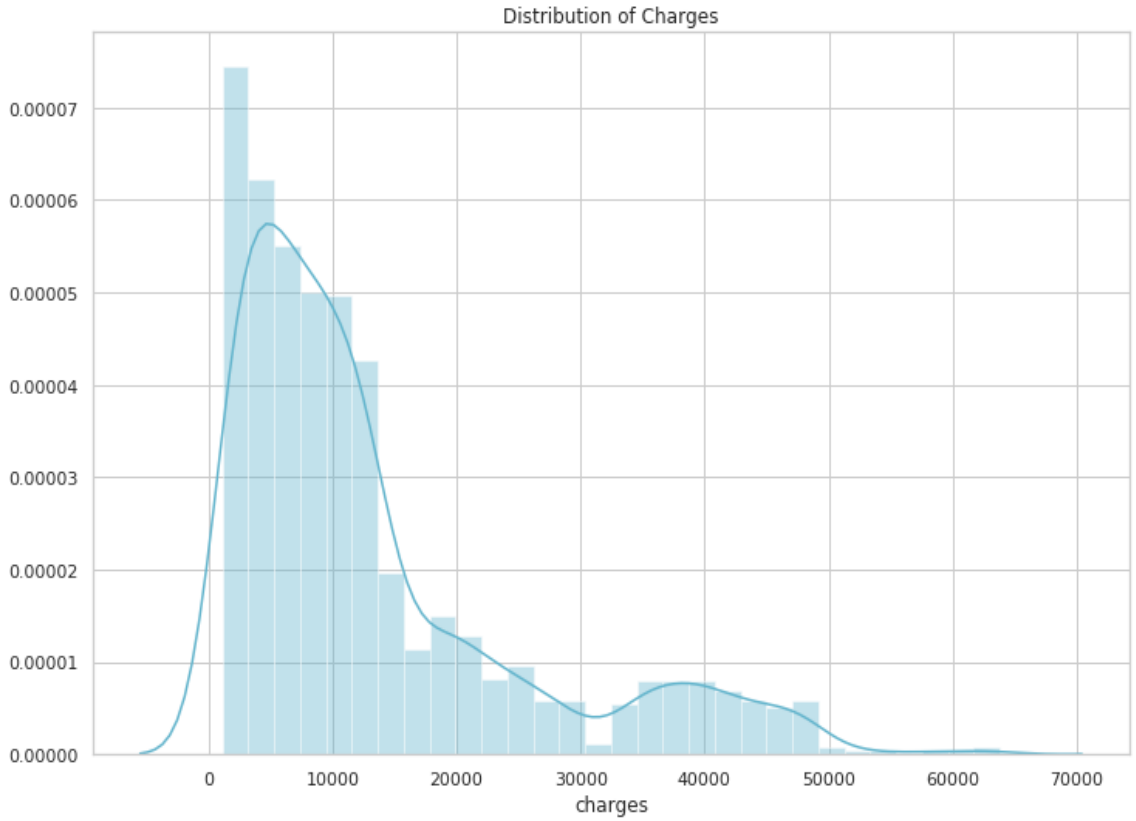

Fig. 2. Displaying the distribution of charges in medical cost insurance.

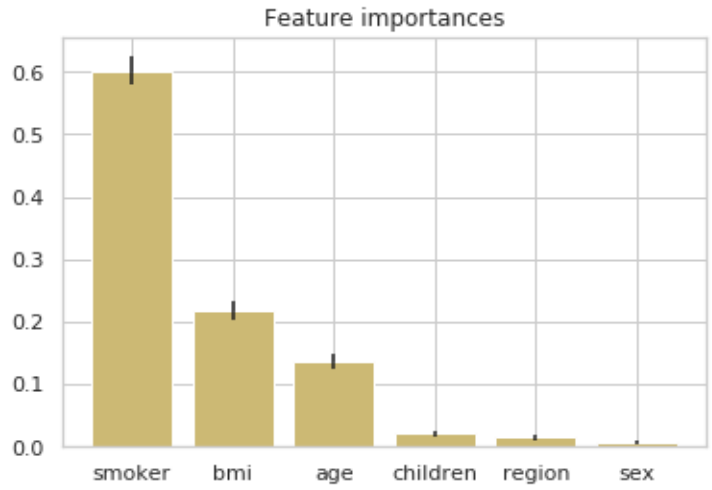

Fig. 3. Displaying the feature importances in smoking, bmi, age, children, region, sex categories.

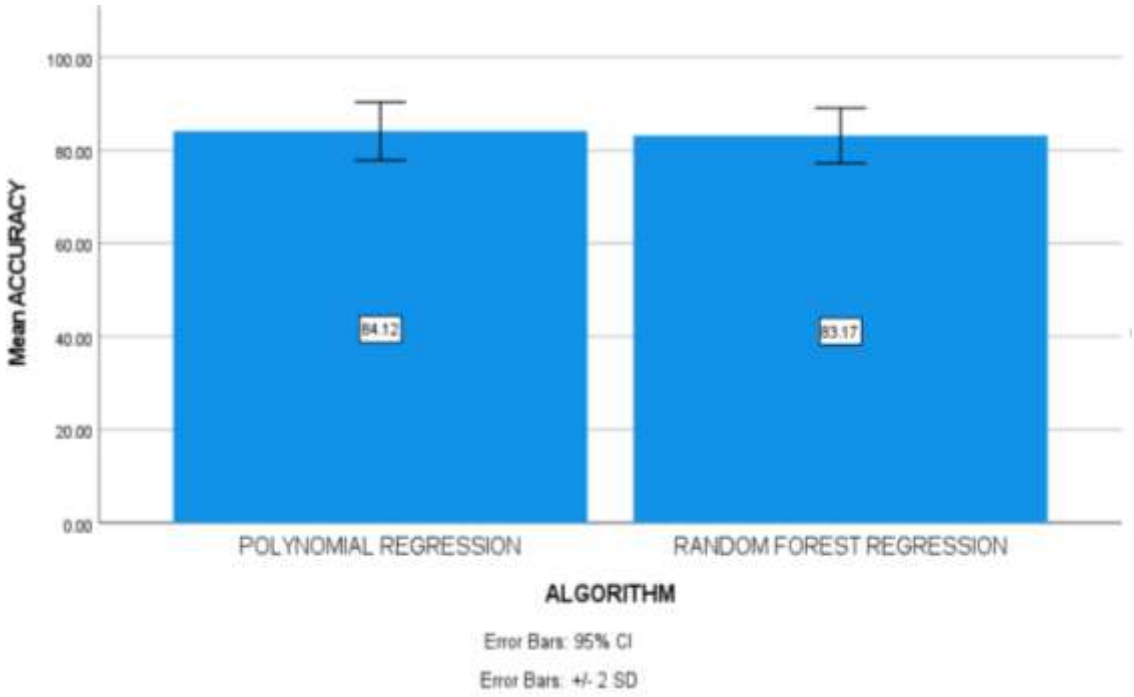

Fig. 4. Comparison of Polynomial Regression algorithm and Random forest Regression Algorithm in terms of mean accuracy. The accuracy of PRA is better than RFRA and the standard deviation of PRA is slightly better than the RFRA algorithm. X-axis: (GROUPS) PRA vs RFRA algorithm and Y- axis: Mean accuracy of prediction ±2 SD