ENHANCING THE ACCURACY FOR MEDICAL COST TO PREDICT THE HEALTH INSURANCE USING POLYNOMIAL REGRESSION ALGORITHM OVER LASSO 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 Lasso 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 Polynomial regression Algorithm and Lasso regression Algorithms are tested. The number of groups are 2 for two Algorithms with the G-power value of 80%. The sample size is 20 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, statistical significance difference was observed between the Polynomial Regression and has an accuracy of 83.94% and Lasso Regression Algorithm 75.06%, 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.The study has a significance value of p=0.001 (p<0.05) two-tailed.

Conclusion: In this paper, based on the results obtained, the Polynomial Regression Algorithm has more accuracy than Lasso Regression Algorithm.

Keywords: Medical Cost, Health Insurance, MLR, Linear Regression, Novel Polynomial Regression, Machine learning, Lasso Regression.

[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

Due to unforeseen climate change, complicated chronic diseases, and mutation of viruses, hospital administration's top challenge is to know about the Length of stay (LOS) of different diseased patients in the hospitals (Pfutzenreuter and de Lima 2021). Hospital management does not exactly know when the existing patient leaves the hospital; this information could be crucial for hospital management (Irmawati, Department of Medical Record and Health Information, and Semarang 2018; "Cases of Management and Assessment of Hospital Culture Construction" 2020; Ramachandra, n.d.). It could allow them to take more patients for admission. As a result, hospitals face many problems managing available resources and new patients in getting entries for their prompt treatment. Therefore, a robust model needs to be designed to help hospital administration predict patients' LOS to resolve these issues (Fu 2021).

For this purpose, a very large-sized data (more than 2.3 million patients' data) related to New-York hospitals patients and containing information about a wide range of diseases including Bone-Marrow, Tuberculosis, Intestinal Transplant, Mental illness, Leukemia, Spinal cord injury, Trauma, Rehabilitation, Kidney and Alcoholic Patients, HIV Patients, Malignant Breast disorder, Asthma, Respiratory distress syndrome, etc. It has been analyzed to predict the LOS (McGuire and Van Kleef 2018). Six machine learning (ML) models taken are: multiple linear regression (MLR), lasso regression (LR). The selected models predictive performance was checked using R square and mean square error (MSE) as the performance evaluation criteria (Kim and Kim 2020). A variety of ML models have been used to predict the LOS of the patients, including unsupervised and supervised ML models. In unsupervised and supervised ML, the model is trained on an unlabeled and labeled dataset, respectively. Some conditions are, however, more prevalent for certain segments of the population (van Egmond et al. 2021). An example of this is throat cancer which is more likely among smokers than non-smokers, and heart diseases such as cardiomyopathy may be more likely among the obese. The difference in performance for the different estimators wasn't much with both linear regression and lasso regression giving the same accuracy rate of 76% and elastic net regression and the orthogonal matching pursuit CV gave different results (Mohanty et al. 2022). The Elastic Net Regression gave the lowest result. It's interesting because tuning the hyper parameters for Elastic Net Regression, if I normalized the dataset; it even reduced the performance more to about 0.02 accuracy rate (Cattaneo and Escanciano 2017). Tuning the Hyperparameters for the linear and lasso regression made little or no difference.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 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 10 per group in total 20.

The dataset named 'Medical Insurance Cost' ("Cost-Effectiveness Analysis and Cost-Benefit Analysis" 2013; Pfutzenreuter and de Lima 2021; Irmawati, Department of Medical Record and Health Information, and Semarang 2018)(Natarajan, Frenzel, and Smaltz 2017)("Cost-Effectiveness Analysis and Cost-Benefit Analysis" 2013; Pfutzenreuter and de Lima 2021; Irmawati, Department of Medical Record and Health Information, and Semarang 2018) is downloaded from Google Kaggle Website. The data in this dataset explains about the Health Insurance provided for each patient varying with the factors (Zhao et al. 2021). Many factors that affect how much you pay for health insurance are not within your control. Health insurance is calculated based on the patient's expenses (Natarajan, Frenzel, and Smaltz 2017). In this dataset information about the patients and the analytics about the patients with different factors was given.

The Health Insurance is an important eye-opener during the emergency need during accidents and disease pandemic situations (Reinhart et al. 2021). Many of the people will lag to hit financially and to bear the operational censoring expe nses during treatment (Kim and Kim 2020). The need for health insurance changes from youth to old age depending on your lifestyle and genetics.

Polynomial Regression

Polynomial Regression is a Machine learning based algorithm (Srinivas, Sucharitha, and Matta 2021) which is also a form of Linear regression model. This model can be improved by features of the prediction, specifically, by making new features that capture the interactions between existing features. This is called Polynomial regression (Cobb et al. 2020). The idea is to generate a new feature matrix consisting of all 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 these Polynomial and Lasso regression algorithms the person uses the Backward Elimination method in order to work his 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.

Lasso Regression

Lasso regression (LR) model was applied in a way very similar to the MLR model. The LR model showed an MSE of 42.58 and an R-square score of 0.31 for the training data (Williams et al. 2022). For the test data, the LR model showed an MSE of 42.19 and an R-square score of 0.310. Thus, for both cases (training and testing), MSE was even higher than MLR, and the R-square score was very low, resulting in low model performance.

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 Lasso regression algorithm dataset from the kaggle and predict the output for the testing 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 the Novel Polynomial Regression Algorithm, the test size is 20% of the total dataset and the remaining of 80% is used for training the datasets. Accuracy of both the algorithms are tested from sample sizes of 20 to 80. 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 novel Polynomial Regression algorithm, it is observed that if the number of censoring iterations increased, then the error rate decreased and accuracy increased. It is declared that the novel Polynomial Regression Algorithm shows higher value compared with the Lasso 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 Novel Polynomial Regression and Lasso Regression algorithm which are predicted values are compared to the values and these outcomes are shown as tables and bar graphs.

In Table 2 Mathematical results of Polynomial Regression and Lasso Regression algorithms. It means accuracy, standard deviation and standard error mean that such LRA and PRA algorithms are received 10 times. It is noted that the PRA algorithm (83.94%) performed better than the LRA algorithm (75.06%).

In Table 3 Independent sample evaluation of the Polynomial Regression Algorithm value. The results of the Lasso Regression algorithm with significant values have two tails (p = <. 001). So both the PRA and LRA algorithms have a value of less than 0.05 with a confidence interval of 95%.

In Figure 1 Displaying Univariate distribution plot of length of stay. They are plotted with the accuracies obtained for duration of stay.

In Figure 2 violin plot is shown with respective medical charges per age. This is plotted by accuracy of the number of ages of the persons and their charges.

In Figure 3 violin plot is shown with respective medical charges per bmi. This is plotted by accuracy of the rate of bmi and their charges.

In Figure 4 violin plot is shown with respective medical charges per child. The accuracy is taken by the number of the children included and not included in smoking.

It has been 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, smokers seem to make a significant change to charges given by Health Insurance.

In Figure 5 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. From the bar chart, it can be seen that the Novel Polynomial Regression Algorithm is more accurate than the Lasso 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 83.94% and Lasso Regression Algorithm is 75.06%.

4. Discussion

The results of the study shows that the Polynomial Regression Algorithm has a better performance than the Lasso Regression Algorithm. Polynomial Regression has obtained an accuracy of 83.94% compared to Lasso Regression which has an accuracy of 75.06%.

MLR incurs a larger penalty for the underestimation of the actual variable than the overestimation. Also, the MLR metric only considers the relative error between the predicted and the actual value, and the scale of the error is not significant (The Law The Law Library 2018). On the other hand, RMSE value increases in magnitude if the scale of error increases. This means RMSLE should be more useful than RMSE when underestimation is undesirable. MAE and MLR are indifferent to the direction of errors. Since the errors are squared before they are averaged, the MLR gives a relatively high weight to large errors. This means the RMSE should be more useful than MLR when large errors are particularly undesirable(Denuit, Hainaut, and Trufin 2019). Knowing the metrics and accountability, it validated the performance of our model simply by applying new data insurance.csv to it and seeing the metrics score. Here k-fold cross validation is not done since the data is small.

No multicollinearity found in the Lasso Regression model, but many found in the Polynomial Regression model. This makes sense since some features in Novel Polynomial Regression were created by multiplying two features from the Lasso Regression model. It is recommended that future work involve more variables in the given dataset to build a more accurate model that could predict hospital cost insurance more accurately.

5. Conclusion

In this paper, the results obtained in executing several Algorithms based on the various data samples using the Polynomial Regression Algorithm (83.94%) and Lasso Regression Algorithm (75.06%) 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 Lasso Regression Algorithm.

Predicting hospital length of stay will help hospitals estimate resources available for the patients and manage the available resources efficiently. MLR with the help of graphs was performed to develop essential insights from the data. By MLR, it concludes that maximum stay was between 0 to 5 days with the meantime of each patient 5.3 days and more than 50 years old patients spent more days in the hospital. Based on the average LOS, it was also observed that the patients with diagnoses related to birth complications spent more days in the hospital than other diseases. Six Machine Learning models were employed and evaluated by using the 10-fold CV approach. Linear multiple regression (LMR), Lasso regression (LR) were the chosen models in this analysis.

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

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.

"Cases of Management and Assessment of Hospital Culture Construction." 2020. *Studies on Hospital Management Transformation*. https://doi.org/10.1142/9789811211645_0016.

Cattaneo, Matias D., and Juan Carlos Escanciano. 2017. *Regression Discontinuity Designs: Theory and Applications*. Emerald Group Publishing.

Cobb, Sharon, Mohsen Bazargan, Jessica Castro Sandoval, Cheryl Wisseh, Meghan C. Evans, and Shervin Assari. 2020. "Depression Treatment Status of Economically Disadvantaged African American Older Adults." *Brain Sciences* 10 (3). https://doi.org/10.3390/brainsci10030154.

"Cost-Effectiveness Analysis and Cost-Benefit Analysis." 2013. *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.

Denuit, Michel, Donatien Hainaut, and Julien Trufin. 2019. *Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions*. Springer Nature.

Egmond, Marie Beth van, Gabriele Spini, Onno van der Galien, Arne IJpma, Thijs Veugen, Wessel Kraaij, Alex Sangers, et al. 2021. "Privacy-Preserving Dataset Combination and Lasso Regression for Healthcare Predictions." *BMC Medical Informatics and Decision Making* 21 (1): 266.

Fu, Xian-Zhi. 2021. "Financial Protection Effects of Private Health Insurance: Experimental Evidence from Chinese Households with Resident Basic Medical Insurance." *International Journal for Equity in Health* 20 (1): 122.

Irmawati, Dr, Department of Medical Record and Health Information, and Poltekkes Kemenkes Semarang. 2018. "The Accuracy of the Indonesian Social Health Insurance Patients Control Code towards the Hospital Cost Services." *Journal of Medical Science And Clinical Research*. https://doi.org/10.18535/jmscr/v6i10.201.

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, Hyunju, and Younkyoung Kim. 2020. "Factors Influencing the Use of Health Services by Trauma Patients according to Insurance Type and Injury Severity Score in South Korea: Based on Andersen's Behavioral Model." *PloS One* 15 (8): e0238258.

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.

McGuire, Thomas G., and Richard C. Van Kleef. 2018. *Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets: Theory and Practice*. Academic Press.

Mohanty, Somya D., Deborah Lekan, Thomas P.

McCoy, Marjorie Jenkins, and Prashanti Manda. 2022. “Machine Learning for Predicting Readmission Risk among the Frail: Explainable AI for Healthcare.” *Patterns (New York, N.Y.)* 3 (1): 100395.

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.

Natarajan, Prashant, John C. Frenzel, and Detlev H. Smaltz. 2017. *Demystifying Big Data and Machine Learning for Healthcare*. CRC Press.

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.

Pfutzenreuter, Thais Carreira, and Edson Pinheiro de Lima. 2021. “MACHINE LEARNING IN HEALTHCARE MANAGEMENT FOR MEDICAL INSURANCE COST PREDICTION.” https://doi.org/10.14488/enegep2021_ti_st_354_1820_42095.

Ramachandra, D. L. n.d. *Essentials of Hospital Management & Administration*. Educreation Publishing.

Reinhart, Alex, Logan Brooks, Maria Jahja, Aaron Rumack, Jingjing Tang, Sumit Agrawal, Wael Al Saeed, et al. 2021. “An Open Repository of Real-Time COVID-19 Indicators.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (51). https://doi.org/10.1073/pnas.2111452118.

Srinivas, Mettu, G. Sucharitha, and Anjanna Matta. 2021. *Machine Learning Algorithms and Applications*. John Wiley & Sons.

The Law The Law Library. 2018. *Health Insurance Issuers Implementing Medical Loss Ratio (Mlr) Requirements Under Patient Protection and Affordable Care ACT (Us Department of Health and Human Services Regulation) (Hhs) (2018 Edition)*. Createspace Independent Publishing Platform.

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.

Williams, Ross D., Aniek F. Markus, Cynthia Yang, Talita Duarte-Salles, Scott L. DuVall, Thomas Falconer, Jitendra Jonnagaddala, et al. 2022. “Seek COVER: Using a Disease Proxy to Rapidly Develop and Validate a Personalized Risk Calculator for COVID-19 Outcomes in an International Network.” *BMC Medical Research Methodology* 22 (1): 35.

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 Lasso Regression algorithms. Mean accuracy value, standard deviation and standard error mean for LRA and PRA algorithms are obtained for 10 iterations. It is observed that the PRA (83.94%) algorithm performed better than the LRA (75.06%) algorithm.

Algorithms (Accuracy)	Sample (N)	Mean	Std Deviation	Std Error Mean
Polynomial Regression	10	83.9400	3.13921	.99270
Lasso Regression	10	75.0680	3.05188	.96509

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

	Levene's Test	T-test of Equality of Means	95% of the confidence

Accuracy	**for Equality of Variances**		**t**	**df**	**Sig (2-tailed)**	**Mean Difference**	**Std Error Difference**	**interval of the Difference**	
	F	**Sig.**						**Lower**	**Upper**
Equal Variance Assumed	.001	.977	6.415	18	<.001	8.88200	1.38451	5.97326	11.79074
Equal Variance Not Assumed			6.415	17.986	<.001	8.88200	1.38451	5.97309	11.79091

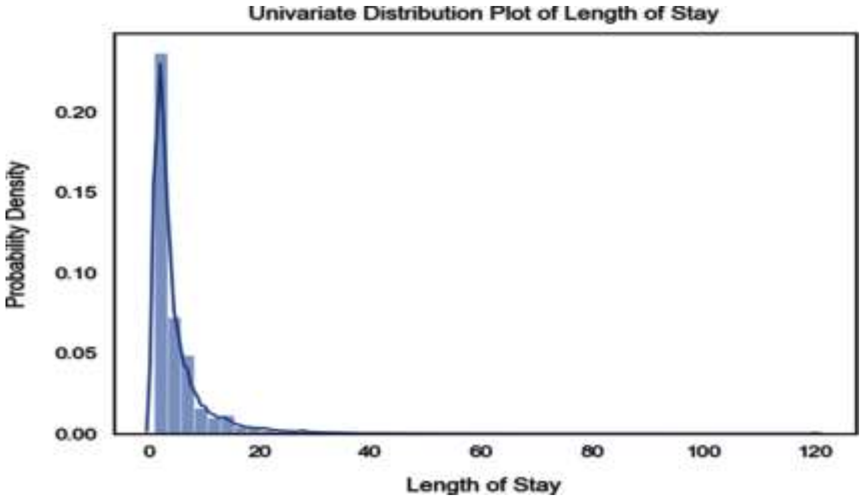

Fig.1. Displaying Univariate distribution plot of length of stay

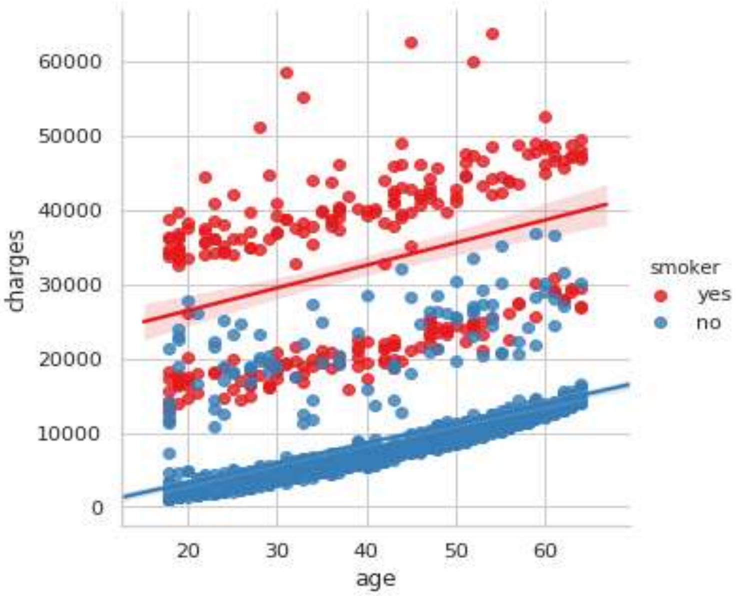

Fig. 2. Displaying the violin plots for medical charges per age in the smoking category.

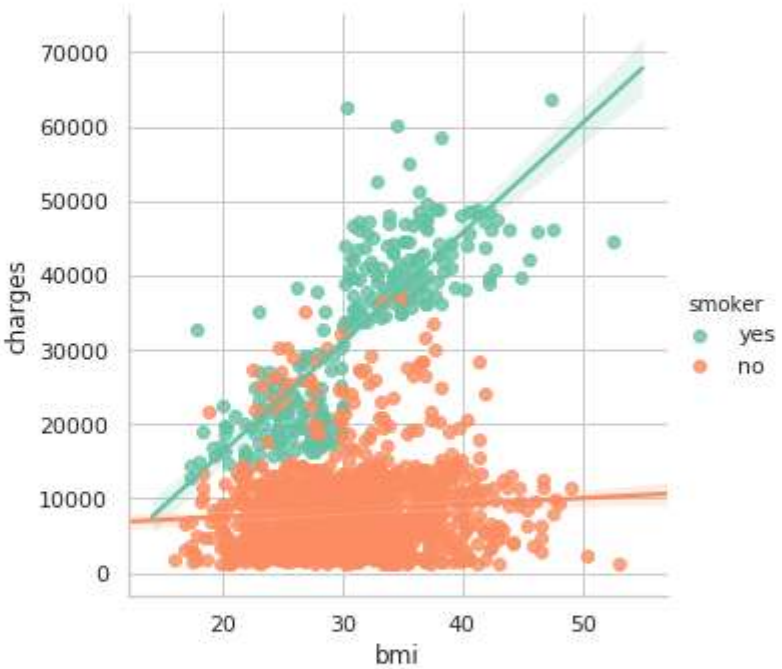

Fig. 3. Displaying the violin plots for medical charges per bmi in smoking category.

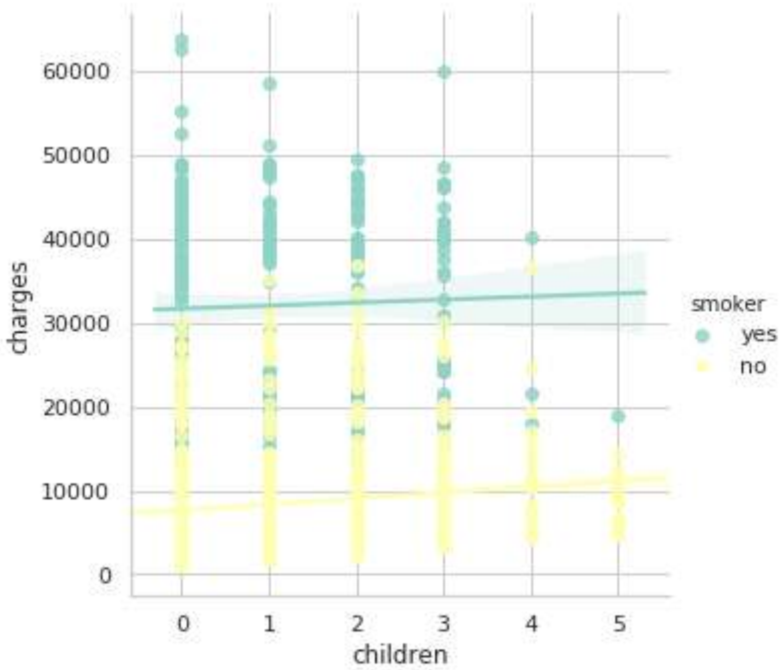

Fig. 4. Displaying the violin plots for medical charges per child in the smoking category.

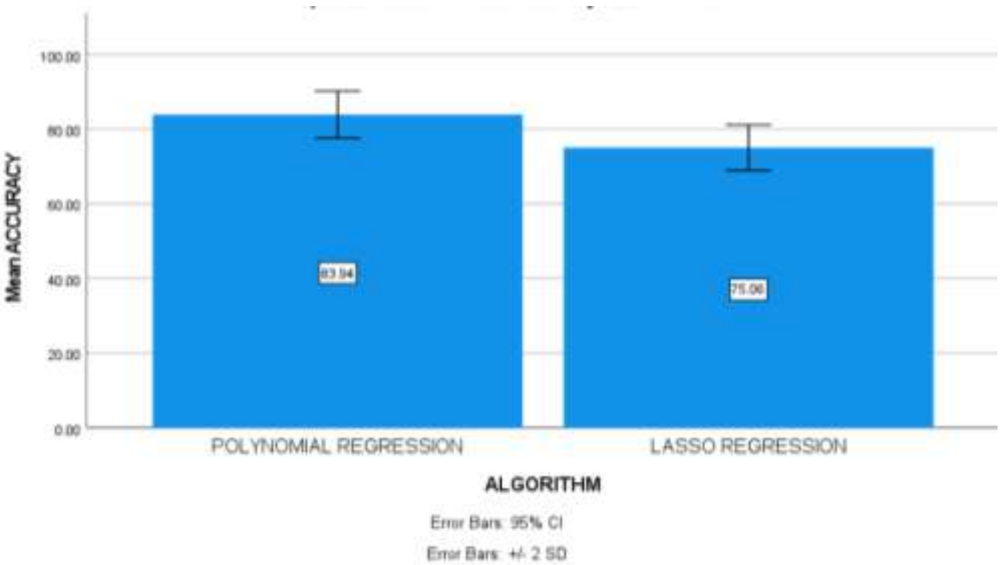

Fig. 5. Comparison of Polynomial Regression algorithm and Lasso Regression Algorithm in terms of mean accuracy. The accuracy of PRA is better than LRA and the standard deviation of PRA is slightly better than the LRA algorithm. X-axis: (GROUPS) PRA vs LRA algorithm and Y- axis: Mean accuracy of prediction ±2 SD