Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 1162553, 13 pages
https://doi.org/10.1155/2021/1162553

Research Article

A Computational Intelligence Approach for Predicting Medical Insurance Cost

Ch. Anwar ul Hassan,[1] Jawaid Iqbal,[1] Saddam Hussain,[2] Hussain AlSalman,[3] Mogeeb A. A. Mosleh,[4] and Syed Sajid Ullah[5]

[1]*Department of Computer Science, Capital University of Science and Technology, Islamabad, Pakistan*
[2]*School of Digital Science, Universiti Brunei Darussalam, Jalan Tungku Link, Gadong BE1410, Brunei Darussalam*
[3]*Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia*
[4]*Faculty of Engineering and Information Technology, Taiz University, Taiz 6803, Yemen*
[5]*Department of Electrical and Computer Engineering, Villanova University, Villanova, PA, USA*

Correspondence should be addressed to Saddam Hussain; saddamicup1993@gmail.com and Mogeeb A. A. Mosleh; mogeebmosleh@taiz.edu.ye

Received 16 October 2021; Accepted 23 November 2021; Published 28 December 2021

Academic Editor: Ewa Rak

In the domains of computational and applied mathematics, soft computing, fuzzy logic, and machine learning (ML) are well-known research areas. ML is one of the computational intelligence aspects that may address diverse difficulties in a wide range of applications and systems when it comes to exploitation of historical data. Predicting medical insurance costs using ML approaches is still a problem in the healthcare industry that requires investigation and improvement. Using a series of machine learning algorithms, this study provides a computational intelligence approach for predicting healthcare insurance costs. The proposed research approach uses Linear Regression, Support Vector Regression, Ridge Regressor, Stochastic Gradient Boosting, XGBoost, Decision Tree, Random Forest Regressor, Multiple Linear Regression, and k-Nearest Neighbors A medical insurance cost dataset is acquired from the KAGGLE repository for this purpose, and machine learning methods are used to show how different regression models can forecast insurance costs and to compare the models' accuracy. The results shows that the Stochastic Gradient Boosting (SGB) model outperforms the others with a cross-validation value of 0.0.858 and RMSE value of 0.340 and gives 86% accuracy.

1. Introduction

People's healthcare cost forecasting is now a valuable tool for improving healthcare accountability. The healthcare sector produces a very large amount of data related to patients, diseases, and diagnosis, but since it has not been analyzed properly, it does not provide the significance which it holds along with the patient healthcare cost [1].

A health insurance policy is a policy that covers or minimises the expenses of losses caused by a variety of hazards. A variety of factors influence the cost of insurance or healthcare [2]. For a variety of stakeholders and health departments, accurately predicting individual healthcare expenses using prediction models is critical [3]. Accurate cost estimates can help health insurers and, increasingly, healthcare delivery organisations to plan for the future and prioritise the allocation of limited care management resources [2]. Furthermore, knowing ahead of time what their probable expenses for the future can assist patients to choose insurance plans with appropriate deductibles and premiums. These elements play a role in the development of insurance policies [4].

In the insurance sector, ML can help enhance the efficiency of policy wording. In healthcare, ML algorithms are

particularly good at predicting high-cost, high-need patient expenditures [5]. ML can be categorized into three different types [6], as shown in Figure 1. These types are supervised machine learning (i.e., a task-driven approach) used for classification/regression and all data labeled; unsupervised machine learning (i.e., a data-driven approach) used for clustering and all data unlabeled; and reinforcement learning (i.e., learning from mistakes) used for decision making.

In this study, we used supervised ML models to demonstrate and compare the accuracy of various regression models, including Linear Regression (LR), Stochastic Gradient Boosting (SGB), XGBoost (XGB), Support Vector Regression (SVR), k-Nearest Neighbors (kNN), Ridge Regressor (RR), Decision Tree (CART), Random Forest Regressor (RFR), and Multiple Linear Regression (MLR). Table 1 describes the notation guide for each algorithm as well as additional abbreviations.

In addition, the main contributions of this works can be summarized as follows:

(i) Investigating the applicability of the machine learning-based computational intelligence approach for predicting healthcare insurance cost in the healthcare industry section.

(ii) Comparing the performance results of the most popular machine learning algorithms for forecasting the costs of healthcare insurance by using a public dataset.

(iii) Providing a guide for developers to choose the appropriate machine learning method when developing an effective healthcare insurance cost prediction system.

The rest of the paper is structured as follows: The related work is discussed in Section 2. Section 3 describes the suggested system. Section 4 contains the experimental outcomes. Finally, Section 5 summarises our findings.

2. Related Work

The research efforts connected to information exploration utilising ML algorithms are addressed in this section. On the subject of claim prediction, a number of publications have been published previously.

Several ML algorithms were used by researchers and practitioners to analyse medical data and estimate health insurance costs [7]. Different ML approaches were utilised for medical data analysis in studies [8–11]. In [12], the authors implement the XGB model for predicting health insurance cost and performed flexible imputation of missing data [13]. In [14], the authors compared the performance of the LR and XGB techniques in predicting the presence of a small number of accident claims, and the results showed that logistic regression is a more effective model than XGB because of its interpretability and strong predictability [14].

Data mining (DM) and machine learning (ML) techniques are widely used for insurance cost prediction and medical fraud detection [15]. Using the Extreme Gradient Boosting algorithm, we improved the accuracy of a decision tree classifier for predicting healthcare insurance fraud [16].

Detection of healthcare fraud using machine learning methods is a significant step for embedding the role of medical providers [17]. On the basis of their personal and financial information, the authors analyse three classifiers that can predict and estimate fraudulent claims as well as the proportion of premiums paid by various clients. The methods Random Forest, J48, and Naive Bayes are employed for classification, and the results are presented in Table 1. Random Forest surpasses the other strategies in terms of financial performance, depending on the synthetic dataset used in the analysis. Hence, they concentrate on bogus claims rather than insurance claim forecasts [18], which is a mistake.

Machine learning methods have been widely used to forecast healthcare costs, although the data used varies, such as the Japanese Public Health Insurance Database [19] and nationwide claims database in France [20] which are used in machine learning applications for predicting individual healthcare costs. Ensemble Regression and LR-based healthcare cost insurance prediction are performed in [21]. Another example for predicting professional costs, pharmacy costs, medication cost, and inpatient and outpatient costs for healthcare is in [22]. In [23], the authors applied M5, RF, CART, LR, GB, and DT for the prediction of medical insurance cost.

In [24, 25], hierarchical Decision Trees and other ML models are used for predictive analytics of healthcare costs. They also suggested that machine learning tools and techniques are critical in the healthcare sector and that they are exclusively used in the diagnosis and prediction of medical insurance costs. Similarly, the underwriting process and medical investigations necessary by the insurance firm to profile the applicants' risks can be difficult and costly [26]. According to [27], the insurance sector collects a lot of information from the applicant, which can take a long time. The insurance agent will normally need applicants to submit a variety of medical tests or documentation. The insurance firm then evaluates the customer's profile and decides whether or not to accept the application. After that, the premiums are determined [28]. On average, it takes at least 30 days to process an application. On the other hand, nowadays many are hesitant to pay for slow services. Because the underwriting procedure is lengthy and time-consuming, customers are more inclined to transfer to a competitor or forgo purchasing life insurance coverage. Poor underwriting methods may cause customers to be unsatisfied, resulting in a reduction in insurance sales. As a result, anticipating the most important aspects that influence the risk assessment process can aid in streamlining and improving insurance procedures [29, 30].

Medical insurance, according to many experts and practitioners, is an absolutely vital component of the medical field's infrastructure. Medical costs, on the other hand, are hard to estimate because the vast majority of the money comes from individuals suffering from unusual diseases. Various machine learning methods are employed in the prediction process. The accuracy of these methodologies' predicted results, on the other hand, is not particularly high.

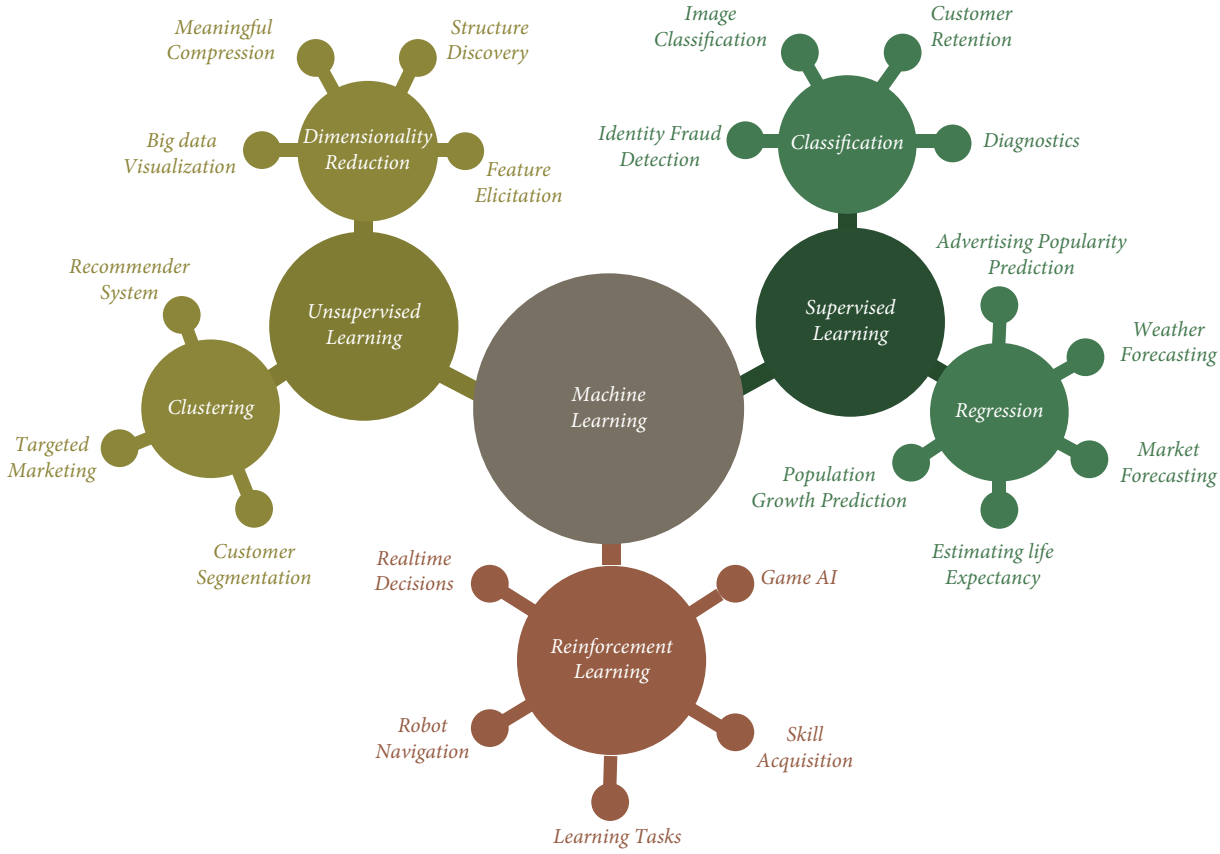

Figure 1: Types of machine learning.

Table 1: Notation guide.

Notations	Description
ML	Machine learning
SVR	Support Vector Regression
RR	Ridge Regressor
kNN	k-Nearest Neighbors
LR	Linear Regression
DT	Decision Tree (CART)
RFR	Random Forest Regressor
SGB	Stochastic Gradient Boosting
XGB	XGBoost
BMI	Body mass index
NN	Neural network
DM	Data mining
RMSE	Root mean squared error
CV	Cross-validation

Although machine learning models are capable of discovering hidden patterns, the training period precludes them from being employed in real time. Because of this, the research tries to develop new ensembles for estimating individual insurance prices in order to attain high forecast accuracy. Several ensemble models, including those based on boosting, bagging, and assembling techniques, were employed to address medical insurance cost prediction problems in this study. The results of the experiments demonstrate that the new assembling model based on machine learning techniques has a higher prediction accuracy for accomplishing the specified job than the previous model.

3. Methodology

We have performed machine learning techniques on medical insurance data. The medical insurance cost dataset is gained from KAGGLE's repository [31], and we performed the data preprocessing. After preprocessing, we select the

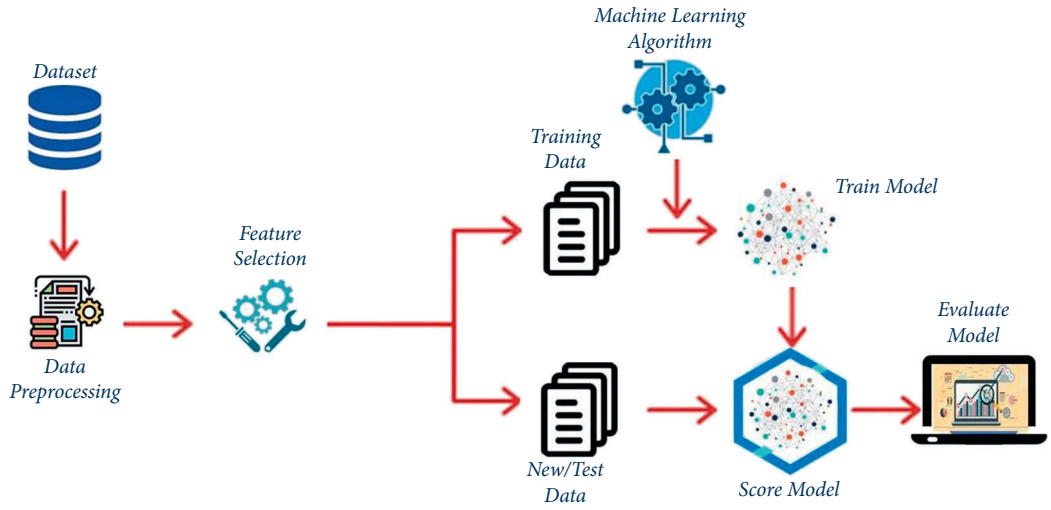

Figure 2: Working methodology.

Table 2: Dataset description.

S/N	Feature name	Description	Value
(1)	Age	One of the most important aspects of health care is age	It has an integer value
(2)	Sex	Gender	(Male = 1, female = 0)
(3)	Body mass index (BMI)	Understanding the human body: weights that are exceptionally high or low in relation to height	An objective body weight index (kg/m^2) based on the height-to-weight ratio, ideally 18.5 – 25
(4)	Children	Number of children/dependents	It has an integer value
(5)	Smoker	Smoking state	(Smoker = 1, nonsmoker = 0)
(6)	Region	Area of residence	(Northeast = 0, northwest = 1, southeast = 2, southwest = 3)
(7)	Charges	Medical costs paid by healthcare insurance	It has an integer value

features by performing feature engineering. Then, the dataset is split into two parts, train and test datasets; about 70% of the total data are used for training, while the rest is for testing. The training dataset is used to create a model that predicts medical insurance costs for the year, while the test dataset is used to evaluate the regression models. For regression exploring the dataset, then categorical values are converted to numerical values. The steps of our working methodology are shown in Figure 2.

3.1. Dataset. The medical cost personal datasets are obtained from the KAGGLE repository. This dataset contains seven attributes, and it was uploaded by Miri Choi in 2018 [31]. The description of the dataset is described in Table 2, and conversion of categorical feature values to numerical values is given in Table 3.

3.2. Feature Engineering and Correlation Matrix. When it comes to machine learning, feature engineering is the process of extracting features from raw data while applying domain expertise in order to improve the performance of ML algorithms. In the medical insurance cost dataset, attributes such as smoker, BMI, and age are the most important factors that determine charges. Also, we see that sex, children, and region do not affect the charges. We might drop these 3 columns as they have less correlation by plotting the heat map graph to see the dependency of dependent value on independent features. The heat map makes it easy to identify which features are most related to the other features or the target variable. Outcomes are shown in Figure 3.

4. Results and Analysis

The results of applied ML models are discussed in this section. Now for this, we can proceed with exploratory data analysis for plotting feature vs. feature (charges) for data visualization.

4.1. Age vs. Charges. We can see in Figure 4 that with the growing age, the insurance charges are going to be increased. For example, when the age touches 64, the insurance charge is 23000, as shown in Figure 4. Age is shown on the *x-axis*, and charges are given on the *y-axis*.

4.2. Region vs. Charges. Insurance charges vary concerning certain regions as shown in Figure 5. The health insurance charges in the southeast are greater than in other regions. The region is displayed on the *x-axis*, and charges are shown on the *y-axis*.

Table 3: Categorical features to numerical values.

Feature name/attribute	Categorical value	Numerical value
Sex	Male	0
	Female	1
Smoker	No	0
	Yes	1
Region	Northwest	0
	Northeast	1
	Southeast	2
	Southwest	3

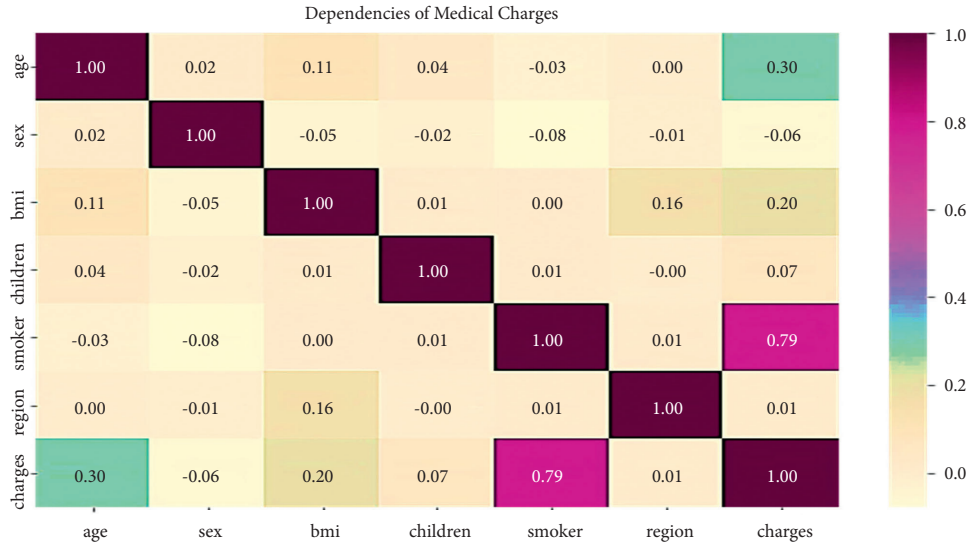

Figure 3: Correlation matrix with a heat map.

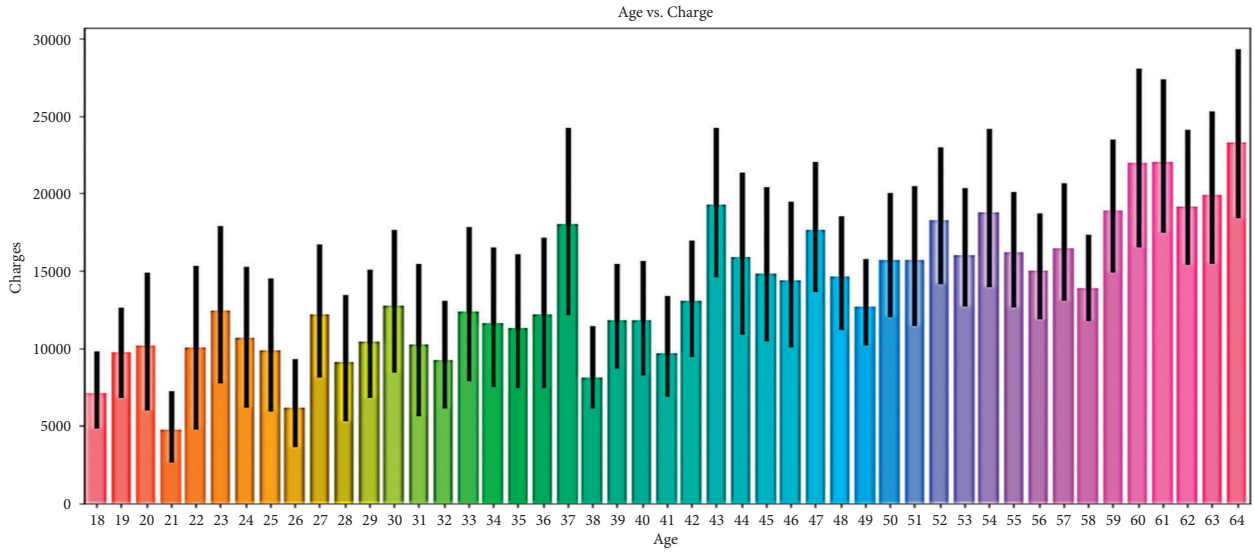

Figure 4: Age vs. insurance charge features.

4.3. BMI vs. Charges. In Figure 6, the zero value is used to represent the females and one value is used for the males. The BMI values of sex or gender types (male and female) are given in the *x-axis*, and the charges are presented in the *y-axis*. It can be clearly seen that when the values of BMI are varied, the insurance charges will vary accordingly as shown in Figure 6.

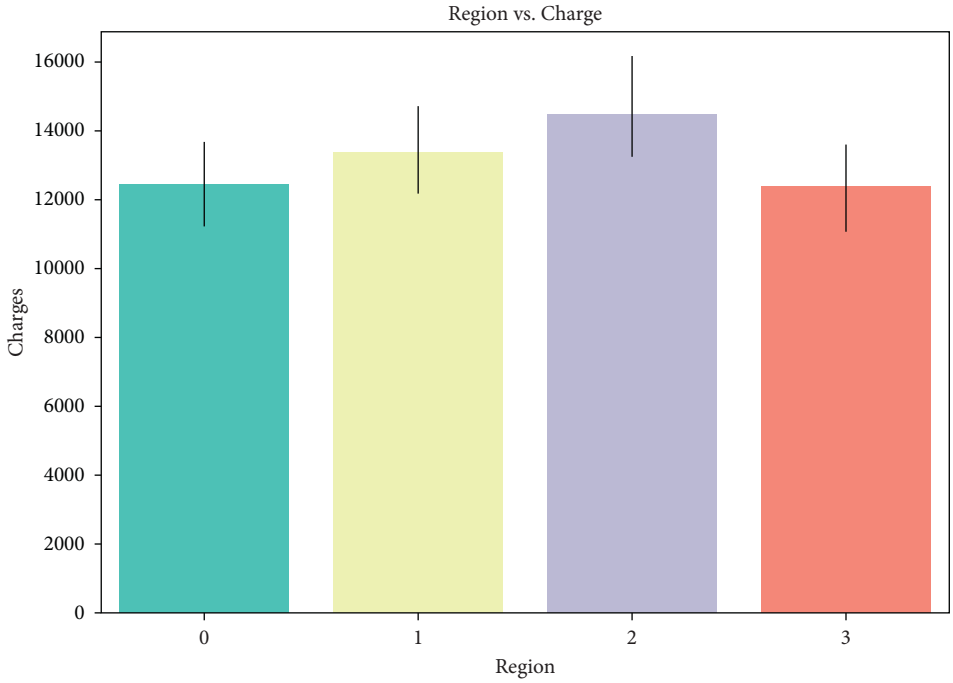

Figure 5: Region vs. insurance charge features.

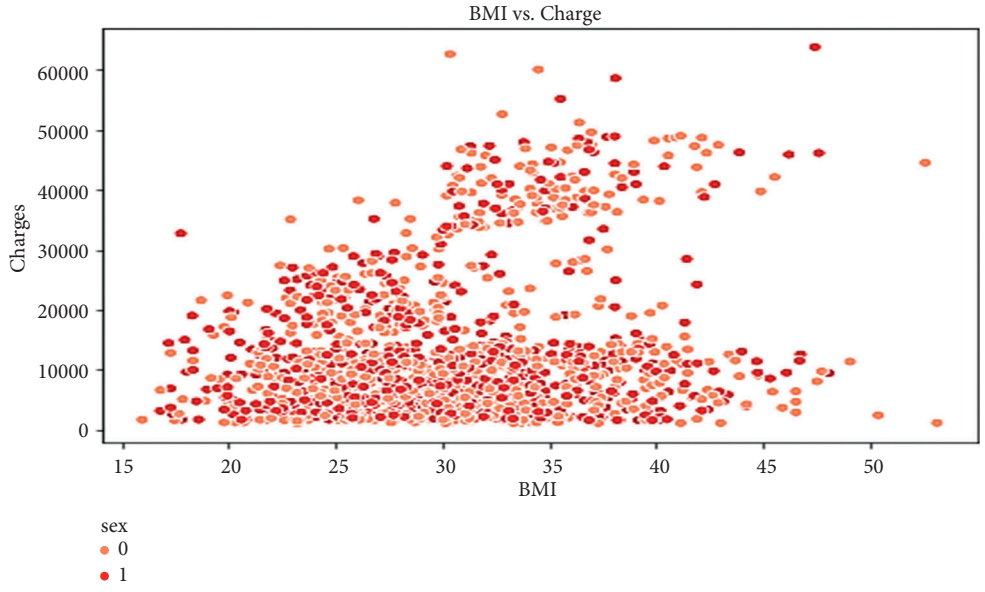

Figure 6: BMI vs. insurance charge features.

4.4. Smoker vs. Charges. Figure 7 illustrates that as a normal smoker, the medical insurance cost varies slightly. However, men are more addicted and passionate to smoking as compared to women so the health insurance cost for females is greater as compared to the males. We can see in Figure 7 that with the increase of smoking habits, the insurance charges are going to be decreased for men and increased for women. Smokers' values are shown on the *x-axis*, and charges are shown on the *y-axis*.

4.5. Sex vs. Charges. The medical insurance charges for the female gender are always greater than for the male as shown in Figure 8. It gives the sex types on the *x-axis* and the charges on the *y-axis*. The figure illustrates that the insurances charges for the female are 14000, and for the male, the charges are around 13000.

4.6. Skew and Kurtosis. Skewness is a metric that quantifies symmetry in a given scenario, or more specifically, the lack of it. If a distribution or data set appears the same on all sides of

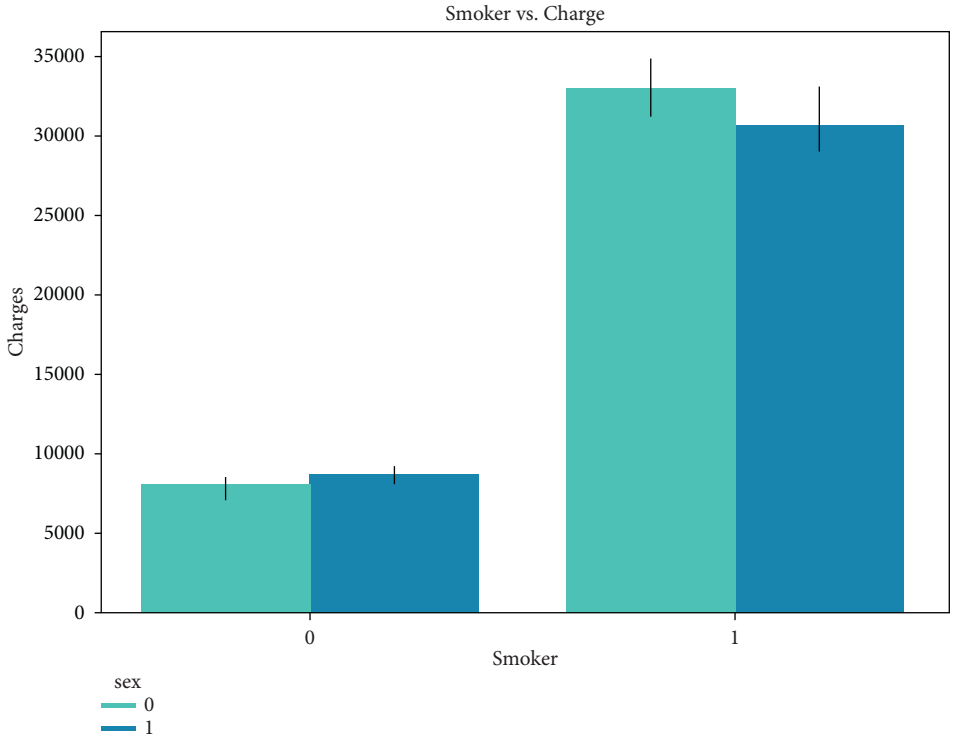

Figure 7: Smoker vs. insurance charge features.

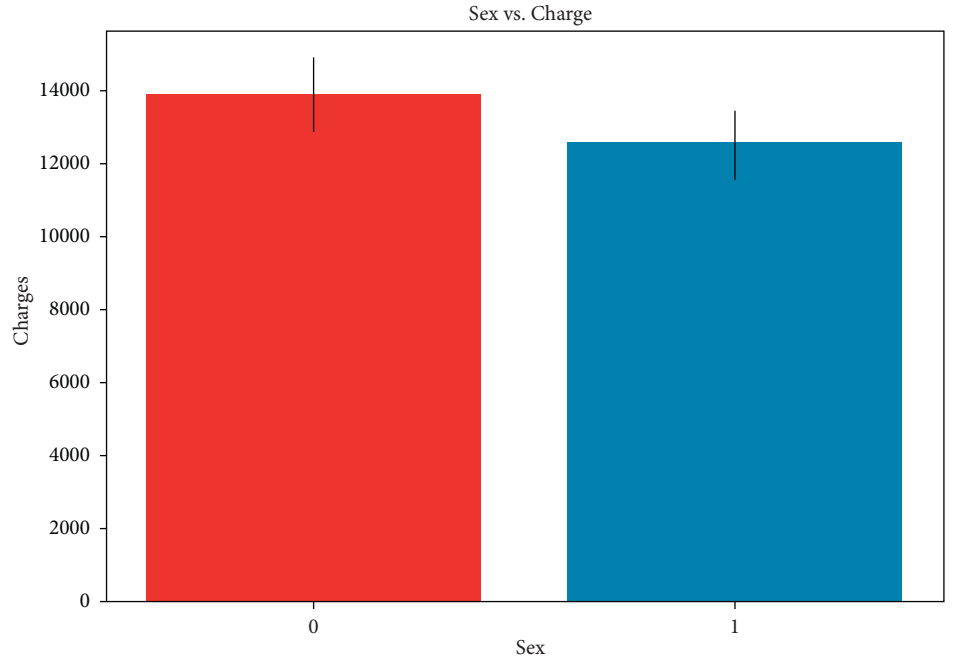

Figure 8: Sex vs. insurance charge features.

the graph to the left and right of the centre point, it is said to be symmetric. Kurtosis is a measure of how heavy-tailed or light-tailed the data are when compared to the normal distribution, according to the normal distribution. Heavy tails or outliers are more probable in data sets with a high kurtosis than data sets with a low kurtosis. When there is a low kurtosis in a data collection, it is more likely that there will be no outliers [32]. The most extreme instance would be if there is a uniform distribution. Table 4 displays the values for the skew and kurtosis of the attributes of a medical dataset.

There might be a few outliers in charges, but we cannot say that the value is an outlier as there might be cases in

Table 4: Attributes skew and kurtosis value.

Attributes	Skew	Kurtosis
Age	0.056	−1.245
Sex	0.021	−2.003
BMI	0.284	−0.051
Smoker	1.465	0.146
Region	0.038	−1.329
Charges	1.516	1.606

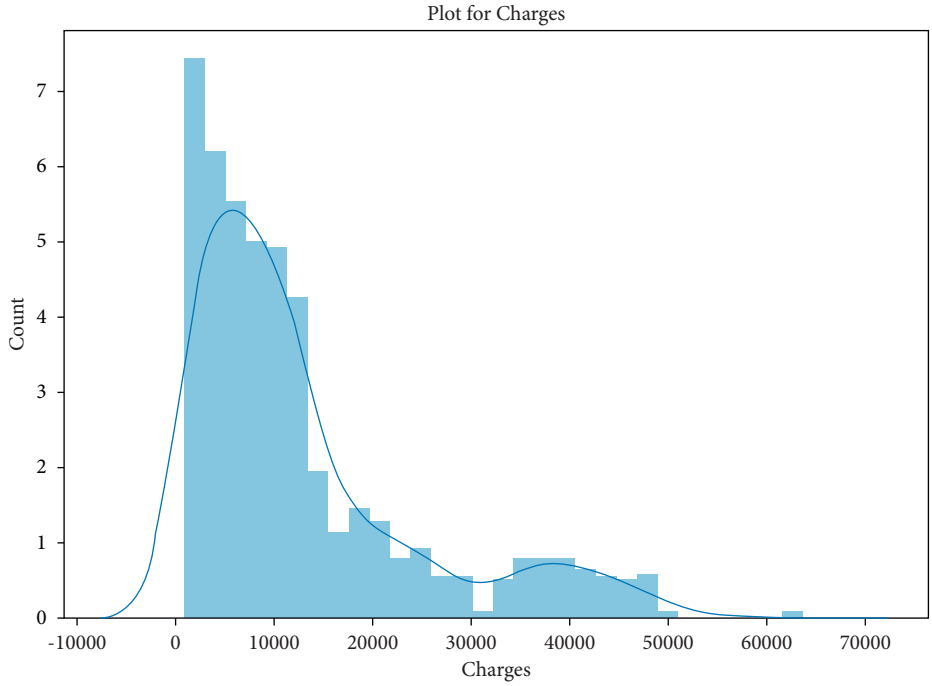

Figure 9: Plot for charges.

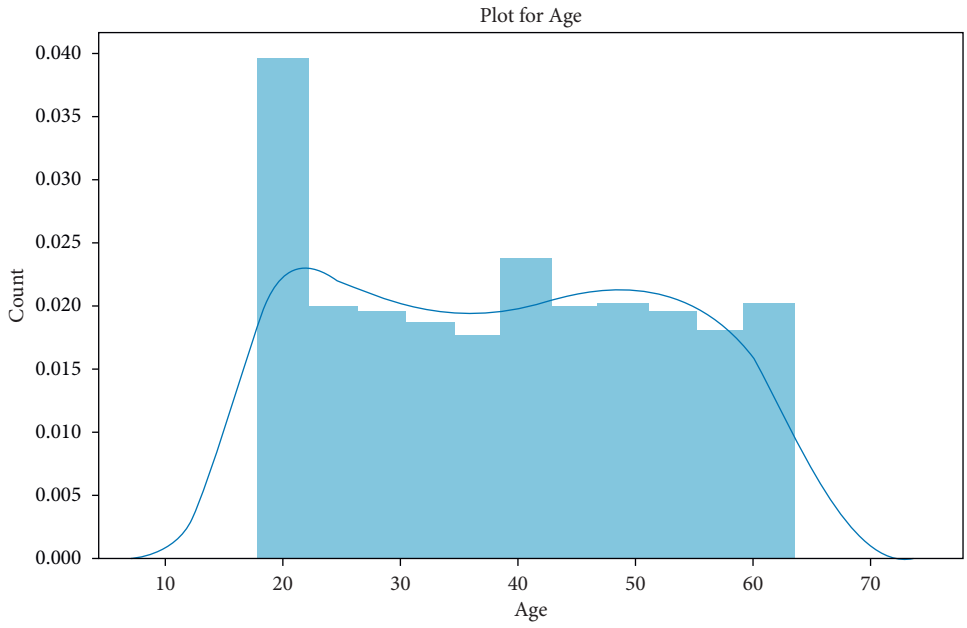

Figure 10: Plot for age.

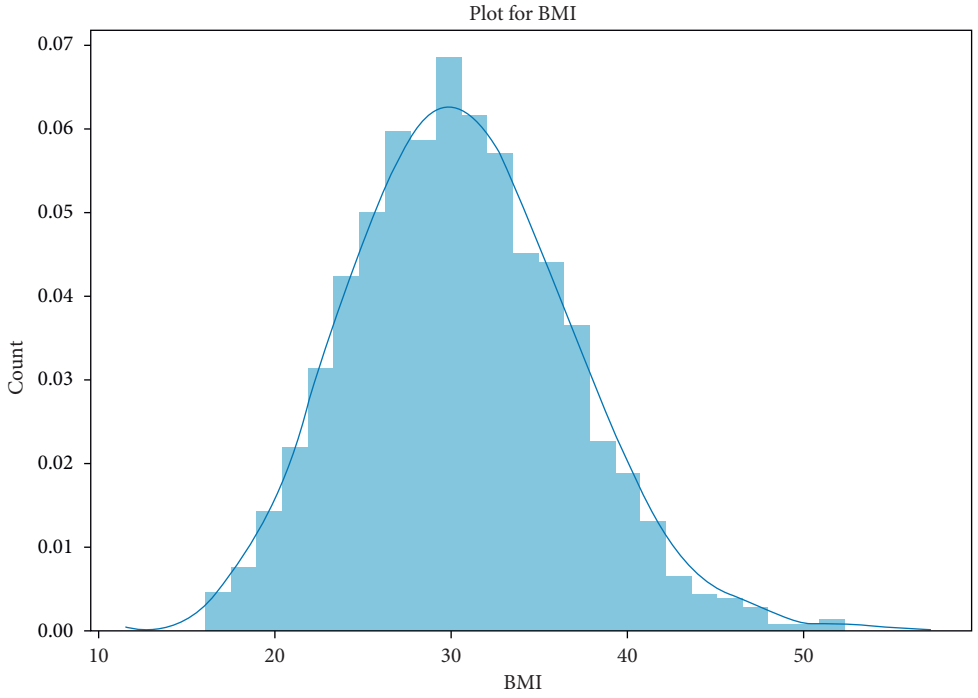

Figure 11: Body mass index plot.

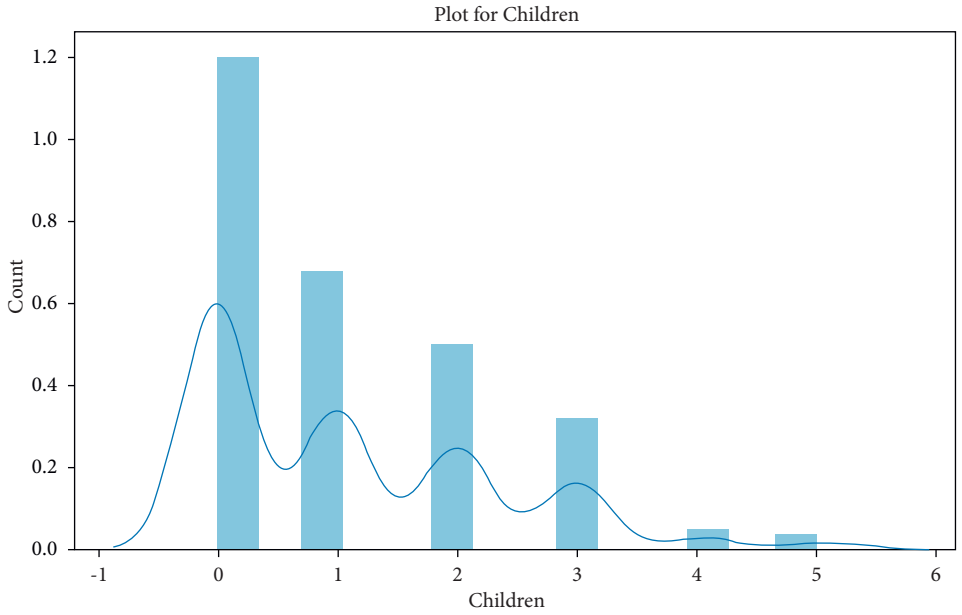

Figure 12: Children skew and kurtosis plot.

which charge for medical care was very less actually. The skew value of charges is 1.516, and the kurtosis value is 1.606 as shown in Figure 9.

The skew value of the age plot is 0.056, and the kurtosis value is −1.245 as shown in Figure 10.

According to BMI, 0.284 and −0.051 are the skew and kurtosis values of BMI, respectively, as shown in Figure 11.

For children, 0.938 and 0.2020 are the skewness and kurtosis values of children, as shown in Figure 12.

In case of smokers, 1.465 and 0.146 are the skewness and kurtosis values of smokers as shown in Figure 13.

Considering region, −0.038 and −1.329 are the skew and kurtosis values of region, respectively, as shown in Figure 14.

4.7. Performance of ML Algorithms. The performance of all the algorithms in terms of RMSE (root mean squared error), training and test scores, and cross-validations is shown in

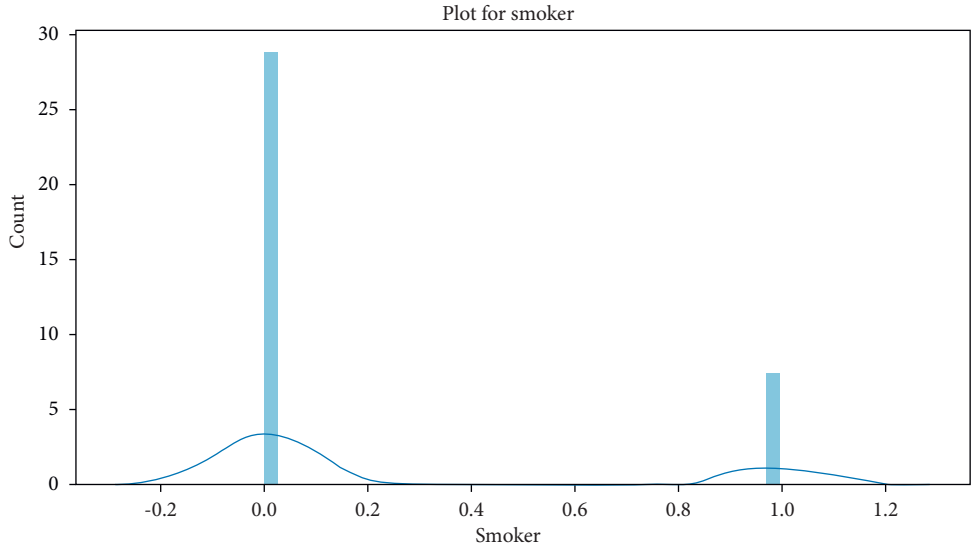

Figure 13: Smoker skew and kurtosis value plot.

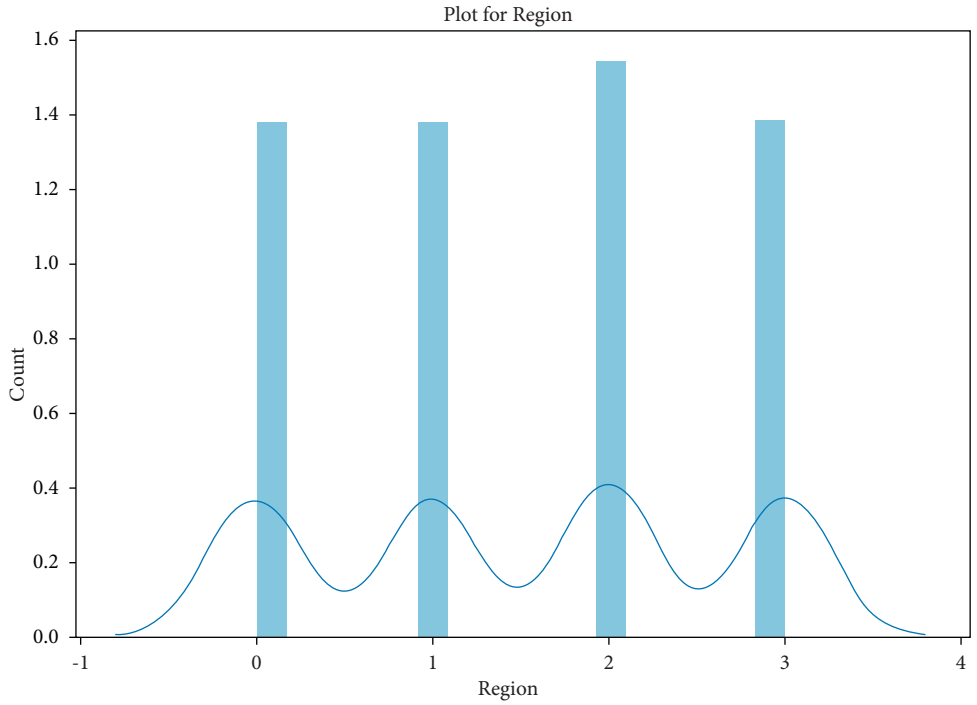

Figure 14: Region skewness and kurtosis plot.

Table 5: Outcomes of machine learning models.

Model	RMSE	R2_score (training)	R2_score (test)	Cross-validation
Linear Regression	0.479808	0.741410	0.782694	0.744528
Ridge Regressor	0.465206	0.741150	0.783800	0.825999
Support Vector Regression	0.358771	0.847234	0.871283	0.842307
Random Forest Regressor	0.347522	0.874422	0.879228	0.849299
Stochastic Gradient Boosting	0.340189	0.17448	0.898595	0.858293
XGBoost	0.342509	0.831859	0.883683	0.853654
Decision Tree (CART)	0.363336	0.820118	0.873213	0.833492
Multiple Linear Regression	0.409725	0.74636	0.794312	0.755814
k-Nearest Neighbors	0.726835	0.274117	0.356719	0.318517

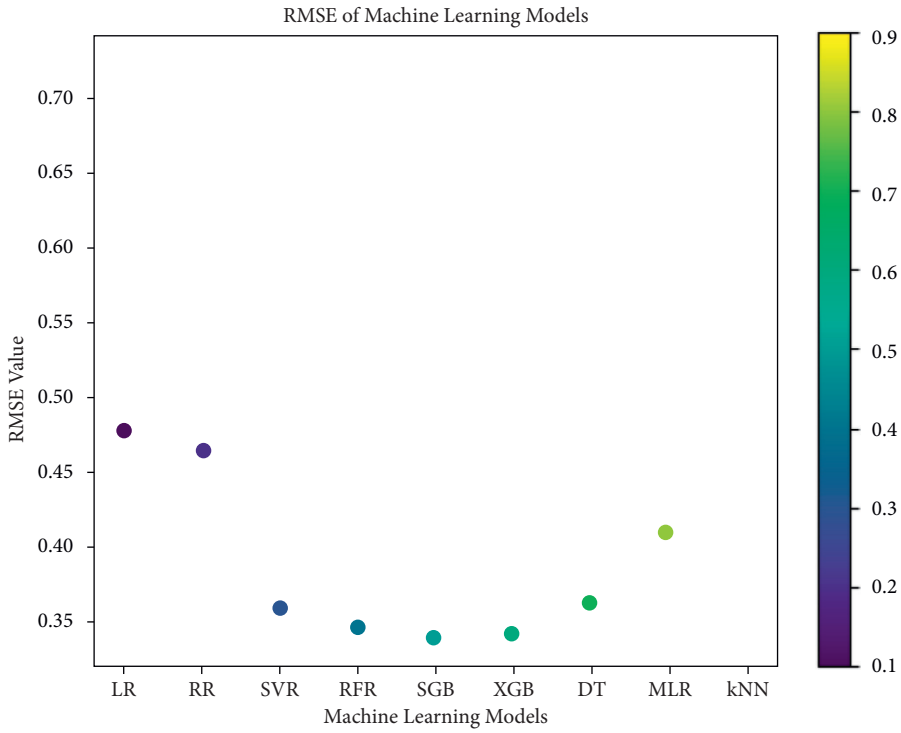

Figure 15: RMSE value graph for machine learning techniques.

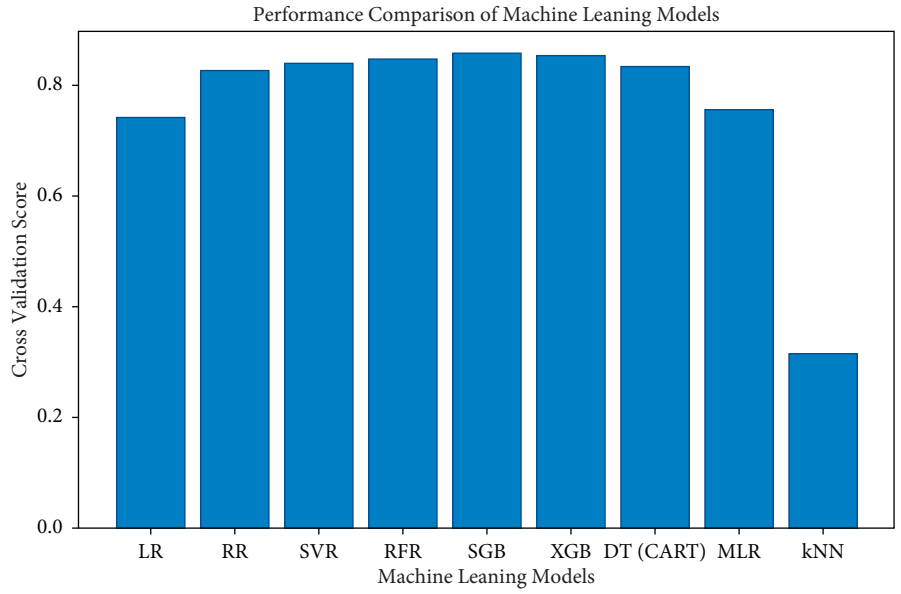

Figure 16: Graph for all models to compare their performance.

Table 5. In Figure 15, the RMSE value of all machine learning (ML) algorithms is visualized for better understanding. By comparing the RMSE value of these ML models, in comparison to the other ML models, k-Nearest Neighbors provides a high RMSE value of 0.726835.

By comparing the performance of all these machine learning algorithms, we conclude that Stochastic Gradient Boosting, XGBoost, and Random Forest Regression performed better as compared to the other ML algorithms and these models achieved almost 86%, 85%, and 85% accuracy, respectively, as shown in Figure 16.

5. Conclusion

Machine learning (ML) is one aspect of computational intelligence that can solve different problems in a wide range of applications and systems when it comes to leveraging historical data. Predicting medical insurance costs is still a

problem in the healthcare industry that needs to be investigated and improved. In this paper, by using a set of ML algorithms, a computational intelligence approach is applied to predict healthcare insurance costs. The medical insurance dataset was obtained from the KAGGLE repository and was utilised for training and testing the Linear Regression, Ridge Regressor, Support Vector Regression, XGBoost, Stochastic Gradient Boosting, Decision Tree, Random Forest Regressor, k-Nearest Neighbors, and Multiple Linear Regression ML algorithms. The regression of this dataset followed the steps of preprocessing, feature engineering, data splitting, regression, and evaluation. The resultant outcome revealed that Stochastic Gradient Boosting (SGB) achieved a high accuracy of 86% with an RMSE of 0.340.

In future work, we will use nature-inspired and metaheuristic algorithms to modify the parameters of machine learning and deep learning approaches on multiple medical health-related datasets.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research was supported by the Researchers Supporting Project number (RSP-2021/244), King Saud University, Riyadh, Saudi Arabia.

References

[1] B. D. Sommers, "Health insurance coverage: what comes after the ACA?" *Health Affairs*, vol. 39, no. 3, pp. 502–508, 2020.

[2] B. Milovic and M. Milovic, "Prediction and decision making in health care using data mining," *Kuwait Chapter of the Arabian Journal of Business and Management Review*, vol. 1, no. 12, 2012.

[3] M. A. Morid, K. Kawamoto, T. Ault, J. Dorius, and S. Abdelrahman, "Supervised learning methods for predicting healthcare costs: systematic literature review and empirical evaluation," in *Proceedings of the AMIA Annual Symposium Proceedings*, vol. 2017, American Medical Informatics Association, Washingdon, DC, USA, November 2017.

[4] M. Kumar, R. Ghani, and Z. S. Mei, "Data mining to predict and prevent errors in health insurance claims processing," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 65–74, Washingdon, DC, USA, July, 2010.

[5] C. Yang, C. Delcher, E. Shenkman, and S. Ranka, "Machine learning approaches for predicting high cost high need patient expenditures in health care," *BioMedical Engineering Online*, vol. 17, no. 1, pp. 131–220, 2018.

[6] M. Iqbal and Z. Yan, "Supervised machine learning approaches: a survey," *ICTACT Journal on Soft Computing*, vol. 5, no. 3, 2015, https://www.potentiaco.com/what-is-machine-learning-definition-types-applications-and-examples/.

[7] B. Panay, N. Baloian, J. A. Pino, S. Peñafiel, H. Sanson, and N. Bersano, "Predicting health care costs using evidence regression," *Multidisciplinary Digital Publishing Institute Proceedings*, vol. 31, no. 1, p. 74, 2019.

[8] M. U. Ghani, T. M. Alam, and F. H. Jaskani, "Comparison of classification models for early prediction of breast cancer," in *Proceedings of the International Conference on Innovative Computing (ICIC)*, Lahore, Pakistan, November.2019.

[9] K. Shaukat, F. Iqbal, T. M. Alam et al., "The impact of artificial intelligence and robotics on the future employment opportunities," *Trends in Computer Science and Information Technology*, vol. 5, no. 1, pp. 50–54, 2020.

[10] X. Yang, M. Khushi, and K. Shaukat, "Biomarker CA125 feature engineering and class imbalance learning improves ovarian cancer prediction," in *Proceedings of the IEEE Asia-Pacific Conf. on Computer Science and Data Engineering (CSDE)*, pp. 1–6, Gold Coast, Australia, December 2020.

[11] T. M. Alam, M. M. A. Khan, M. A. Iqbal, W. Abdul, and M. Mushtaq, "Cervical cancer prediction through different screening methods using data mining," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 2, 2019.

[12] M. A. Fauzan and H. Murfi, "The accuracy of XGBoost for insurance claim prediction," *International Journal of Advanced Software Computer Applications*, vol. 10, no. 2, 2018.

[13] B. S. Van, *Flexible Imputation of Missing Data*, CRC Press, Boca Raton, FL, USA, 2018.

[14] J. Pesantez-Narvaez, M. Guillen, and M. Alcañiz, "Predicting motor insurance claims using telematics data-XGBoost versus logistic regression," *Risks*, vol. 7, no. 2, 2019.

[15] L. S. Chen and J. C. Chen, "Using data mining methods to detect medical fraud," in *Proceedings of the 2020 International Conference on Management of e-Commerce and e-Government*, pp. 89–93, Jeju Island, South Korea, July 2020.

[16] N. A. Akbar, A. Sunyoto, M. R. Arief, and W. Caesarendra, "Improvement of decision tree classifier accuracy for healthcare insurance fraud prediction by using Extreme Gradient Boosting algorithm," in *Proceedings of the International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, pp. 110–114, IEEE, Jakarta, Indonesia, November, 2020.

[17] J. M. Johnson and T. M. Khoshgoftaar, "Medical provider embeddings for healthcare fraud detection," *SN Computer Science*, vol. 2, no. 4, pp. 1–15, 2021.

[18] G. Kowshalya and M. Nandhini, "Predicting fraudulent claims in automobile insurance," in *Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp. 1338–1343, IEEE, Coimbatore, India, April 2018.

[19] Y. Nomura, Y. Ishii, Y. Chiba et al., "Does last year's cost predict the present cost? An application of machine leaning for the Japanese area-basis public health insurance database," *International Journal of Environmental Research and Public Health*, vol. 18, no. 2, 2021.

[20] A. Vimont, H. Leleu, and I. Durand-Zaleski, "Machine learning versus regression modelling in predicting individual healthcare costs from a representative sample of the nationwide claims database in France," *The European Journal of Health Economics*, pp. 1–13, 2021.

[21] D. M. Shyamala, P. Swathi, R. M. Purushotham et al., "Linear and ensembling regression based health cost insurance prediction using machine learning," in *Smart Computing Techniques and Applications*, Springer, New York, NY, USA, 2021.

[22] S. Sushmita, S. Newman, J. Marquardt et al., "Population cost prediction on public healthcare datasets," in *Proceedings of the 5th International Conference on Digital Health*, pp. 87–94, Florence, Italy, May 2015.

[23] I. Duncan, M. Loginov, and M. Ludkovski, "Testing alternative regression frameworks for predictive modeling of health care costs," *North American Actuarial Journal*, vol. 20, no. 1, pp. 65–87, 2016.

[24] B. Nithya and V. Ilango, "Predictive analytics in health care using machine learning tools and techniques," in *Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 492–499, IEEE, Madurai, India, June 2017.

[25] A. Tike and S. Tavarageri, "A medical price prediction system using hierarchical decision trees," in *Proceedings of the IEEE International Conference on Big Data (Big Data)*, pp. 3904–3913, IEEE, Boston, MA, USA, December 2017.

[26] N. Boodhun and M. Jayabalan, "Risk prediction in life insurance industry using supervised learning algorithms," *Complex & Intelligent Systems*, vol. 4, no. 2, pp. 145–154, 2018.

[27] J. M. Carson, C. M. Ellis, R. E. Hoyt, and K. Ostaszewski, "Sunk costs and screening: two-part tariffs in life insurance," *Journal of Risk & Insurance*, vol. 87, no. 3, pp. 689–718, 2020.

[28] A. C. Wuppermann, "Private information in life insurance, annuity, and health insurance markets," *The Scandinavian Journal of Economics*, vol. 119, no. 4, pp. 855–881, 2017.

[29] A. E. Prince, "Tantamount to fraud: exploring non-disclosure of genetic information in life insurance applications as grounds for policy rescission," *Health Matrix*, vol. 26, 2016.

[30] S. Peñafiel, N. Baloian, J. A. Pino et al., "Associating risks of getting strokes with data from health checkup records using Dempster-Shafer Theory," in *Proceedings of the IEEE 20th International Conference on Advanced Communication Technology (ICACT)*, pp. 239–246, Gang'weondo, Korea, February 2018.

[31] M. Choi, "Medical cost personal datasets," 2018, https://www.kaggle.com/mirichoi0218/insurance.

[32] D. B. Madan and K. Wang, "Option implied VIX, skew and kurtosis term structures," *International Journal of Theoretical and Applied Finance*, vol. 24, no. 5, Article ID 2150030, 2021.