Yanbu Journal of Engineering and Science

Volume 22 | Issue 2 Article 2

2025

Prediction, Comparison and Visualization of Premium Pricing using Machine Learning methods

Lal S Jan
Hera British School, Jeddah AND Physics, Government Postgraduate College Mardan, Pakistan

Farzan Tashfeen
NED University of Engineering and Technology

Mohid Hussain
NED University of Engineering and Technology

Nadi M Alresheedi
General Studies, Yanbu University College

Syed Muhammad Hassan Ahmer
General Studies, Yanbu University College

See next page for additional authors

Follow this and additional works at: https://yjes.researchcommons.org/yjes
The journal in which this article appears is hosted on Digital Commons, an Elsevier platform.

Recommended Citation
Jan, Lal S; Tashfeen, Farzan; Hussain, Mohid; Alresheedi, Nadi M; Ahmer, Syed Muhammad Hassan; and Albarkati, Nahla M (2025), Prediction, Comparison and Visualization of Premium Pricing using Machine Learning methods, *Yanbu Journal of Engineering and Science*: Vol. 22: Iss. 2, 23-33.
DOI: https://doi.org/10.53370/1658-5321.1199
Available at: https://yjes.researchcommons.org/yjes/vol22/iss2/2

This Research Article is brought to you for free and open access by Yanbu Journal of Engineering and Science. It has been accepted for inclusion in Yanbu Journal of Engineering and Science by an authorized editor of Yanbu Journal of Engineering and Science.

Prediction, Comparison and Visualization of Premium Pricing using Machine Learning methods

Authors

Lal S Jan, Farzan Tashfeen, Mohid Hussain, Nadi M Alresheedi, Syed Muhammad Hassan Ahmer, and Nahla M Albarkati

This research article is available in Yanbu Journal of Engineering and Science: https://yjes.researchcommons.org/yjes/vol22/iss2/2

YANBU JOURNAL OF ENGINEERING AND SCIENCE 2025;22:23–32

Scan the QR to view the full-text article on the journal website

RESEARCH ARTICLE

Prediction, Comparison and Visualization of Premium Pricing using Machine Learning methods

Lal S Jan [a,b,*], Farzan Tashfeen [c], Mohid Hussain [c], Nadi M Alresheedi [d], Syed Muhammad Hassan Ahmer [d], Nahla M Albarkati [d]

[a] Hera British School, Jeddah, Saudi Arabia
[b] Physics, Government Postgraduate College Mardan, Pakistan
[c] NED University of Engineering and Technology, Karachi, Pakistan
[d] General Studies, Yanbu University College, Saudi Arabia

ABSTRACT

This study investigates the efficacy of machine learning in forecasting and comparing health insurance premiums. The all-inclusive goal line is to improve the accuracy of premium pricing and minimize the manual labor associated with the traditional methods and present data trends in a logical manner. An open-source dataset of US medical premiums has been utilized and a systematic machine learning model in python programming language has been developed. This model took into account factors like age, bmi, smoking habits, and geographical location. The XGBoost model excels other models with an RMSE value of 0.349 and an adjusted R^2 of 90.1% and this is closely tracked down by the authors' Artificial Neural Network model which achieved an accuracy of 93.07%. These numerical research findings accentuate the transformative potential of machine learning in offering precise and accurate pricing that stands to the advantage of both insurer and policyholder.

Keywords: Premium, Prediction, Health insurance, Machine learning, Pricing, Artificial intelligence

1. Introduction

Health insurance is a safe haven for many individuals against the financial burdens of medical bills. Upon payment of a certain premium, an individual can secure coverage under a health insurance policy [1–11]. The computation of this premium is influenced by a number of factors that leads to variability in the cost of health insurance policies. For instance, age significantly influences the cost of a health insurance plan [12–17]. Younger individuals naturally have fewer health issues than older ones and normally require less extensive medical treatment. As such, the cost of treating an older person is higher and consequently they are charged a higher premium. Several factors impact the insurance premium of a health insurance policy that coerces to a unique premium amount for each individual. In recent years, the healthcare industry has witnessed a surge in the application of artificial intelligence *(AI)*, performing medical-related tasks at a much faster rate [18–20]. The ability of *AI* to gather, process and provide appropriate results from data accelerates disease detection and minimizes errors, thereby substantially shortening the diagnosis-treatment-recovery cycle. For instance, healthcare professionals and organizations utilize chatbots to gather basic patient information before a consultation, improving efficiency for both the doctor and the patient [21–25]. In fact, *AI* and machine learning *(ML)* have made significant advances in the health insurance sector. In effect, *AI* facilitates faster claim settlements by detecting fraudulent claims, enabling insurers to process valid claims

Received 28 February 2024; revised 28 December 2024; accepted 26 January 2025.
Available online 16 July 2025

* Corresponding author.
E-mail address: lalsaidjan@gmail.com (L. S. Jan).

https://doi.org/10.53370/001c.129048
1658-5321/© 2025 The Author(s). Yanbu Journal of Engineering and Science. This is an open access article under the CC BY 4.0 Licence (https://creativecommons.org/licenses/by/4.0/).
Published by Yanbu Industrial College, Saudi Arabia. The journal is hosted on Digital Commons, an Elsevier Platform.

more efficiently [26]. Additionally, the potential of *AI* to recommend healthier habits and behaviors to clients offers cost-effectiveness to insurers thereby reducing avoidable healthcare expenditures. The traditionally lengthy and time-consuming procedure of underwriting has been accelerated by *AI* based predictive analysis. Research in the field of *AI* with regard to healthcare can enhance efficiency, improve treatment efficacy and reduce burnout among medical professionals. The *AI* and *ML* technologies have the competency to expedite insurance claim processing and reduce the burden of electronic health annals and, at the same time, provide individual health insurance plans so as to transform the healthcare experience in developing countries [27, 28]. The landscape of premium pricing prediction using *ML* methods has been explored in a number of research studies. A literature appraisal of the research work done in this realm reveals that the scope of utilizing *ML* algorithms for premium pricing prediction is vast and multi-faceted. The research work done by Keshav Kaushik, Akashdeep Bhardwaj, Ashutosh Dhar and Dwivedi Rajani Singh [29] trained and evaluated an artificial intelligence network-based regression model to predict health insurance premiums. The authors predicted the health insurance cost incurred by individuals on the basis of a host of parameters like age, gender, bmi (body-mass-index), number of children, smoking habits, and geographical location with a consequential accuracy of 92.72%. The research team of Ch. Anwar ul Hassan, Jawaid Iqbal, Saddam Hussein, Hussain Al Salman, Mogeeb A. A. Mosleh and Syed Sajid Ullah [30] exploited a variety of machine learning algorithms to predict medical insurance costs. They calculated the skewness and kurtosis for each attribute so as to check the outliers. Their research encompassed data preprocessing and feature engineering in consort with regression and evaluation. In their findings, the Stochastic Gradient Boosting model stood out prominent and they claimed an accuracy of 86%. Takeshima T, Keino S, Aoki R, Matsui T and Iwasaki K, [31] developed a prediction model by using health insurance claims data. They applied the least absolute shrinkage and selection operator (LASSO) regression model and used excess logarithmic medical cost as explained valuables and diseases as explanatory valuables in order to avoid overfitting. In their research work, Sen Hu, Adrian O'Hagan, James Sweeney, and Mohammad Hossein Ghahramani [32] analyzed lapse behavior in life insurance policies. The integration of spatially linked demographic census data provided insights for insurers to assess and prevent lapsing risks, enhancing the calculation and preparation of capital reserves. In the same vein, Angela D. Kafuria [33] developed a predictive model for determining health insurance charges using machine learning algorithms. The model was trained on attributes like age, sex, bmi, number of children and region. They highlighted in their research that Gradient Boosting and Random Forest Regression are the most effective predictive models. Nonetheless, the work of Y. Angeline Christobel and Suresh Subramanian [34] compared four regression analysis algorithms to predict insurance namely Linear Regression, Ridge Regression, Lasso Regression and Polynomial Regression. They concluded in their research findings that among the four models the Polynomial Regression was the most effective model with an accuracy of 88%.

The literature review reveals the fact that ML models hold significant promise in predicting health insurance premiums that can revolutionize the insurance industry in terms of predicting medical costs and analyzing the customer behavior. Nevertheless, predicting health insurance premiums in the healthcare business by incorporating ML models and algorithms is still a subject that has to be investigated wisely, contemplated sensibly and then optimized meticulously. Consequently, in this research endeavor the authors are going to examine the ML models in the programming language of python for an advanced prediction, comparison and visualization of the health insurance premiums. It is anticipated that the model will help reduce manual pricing labor, improve the accuracy and interpretation of the data for decision-making to the advantage of both the agencies and customers.

2. Methodology

The research method for this study is comprised of several steps and has been progressively elucidated in the results and discussion section, viz. Section 3 of the present article. Section 3 has been split into different subheadings in order to logically illuminate the procedure. An open-source dataset of US medical premiums has been utilized and a systematic machine learning model in python programming language has been developed. This model took into account factors like age, bmi, smoking habits, and geographical location. The authors imported data and analyzed it and then ensured the feature engineering in orchestration with data visualization. In inference, the authors constructed an artificial neural network model (ANN model) and tested the model for accurate predictions. The salient topographies of the ANN model are presented in Section 3.6 of the results and discussion section.

3. Results and discussions

The present research study targeted enhancements in the precision of pricing for medical insurance providers and, to achieve that end, it scientifically examined the usefulness of machine learning models in forecasting health insurance premiums. The research focused on addressing the manysided aspects that may affect the insurance premiums such as age, bmi, smoking habits, and geographical location and incorporated a diverse set of machine learning algorithms like Linear Regression, Ridge Regressor, Support Vector Regression, Random Forest and Artificial Neural Networks (*ANN*). The metrics used for comparison were RMSE (Root Mean Square Error) and R^2 (coefficient of determination). A systematic and in-depth description of our research data and the ensuing generation of tabulated results and graphs is given in the following Sections 3.1 to 3.6.

3.1. Data import and preprocess

In this study we utilized the US health insurance dataset which has been used previously in several research studies among which the prominent research studies are undertaken by researchers like Kaushik et al. [29], Hassan et al. [30] and Takeshima et al. [31]. This dataset is widely recognized for its comprehensiveness and the key attributes like age, sex, bmi, the number of children, smoking habits, and regional details are tested and judged against the insurance charges. The choice of the authors to use this dataset in the present examination is based on its proven reliability and relevance to our research objectives. In order to remove duplication, the dataset was preprocessed and cleaned.

3.2. Feature engineering

We transformed the categorical columns into binary formats for model compatibility. Precisely, the sex column was converted to binary, where 0 denotes female and 1 indicates male. In the same vein, the smoker column was binarized with 0 for 'no' and 1 for 'yes'. The region column was feature engineered where the categorical entries were transformed using dummy variables and were then appended to the main dataset. This way, the original 'region' column was removed to avoid redundancy.

3.3. Exploratory analysis

To visualize the distribution of all the attributes in the dataset including age, sex, bmi, children, smoker, charges and regional information (i.e., Northwest, Southeast, Southwest), a comprehensive histogram was generated that can be appreciated very clearly from Fig. 1. This initial graphical examination helped in understanding the data distribution for each attribute. Next, to explore the distribution of three key variables, i.e., age, bmi and charges, a multifaceted visualization was designed that is visibly perceptible from Fig. 2. This was achieved using box plots, violin plots, and histograms for each variable. Subsequently, to display the correlation coefficients between all the numerical attributes, a bivariate map was generated as shown in Fig. 3. This Fig. 3 is, in fact, a digitization of the attributes. In order to gain a clear and deeper understanding of the relationships amongst the different features in the dataset, a Pair-Grid Plot was constructed which is distinctly noticeable from Fig. 4. This visual Pair-Grid provides a matrix scheme where each feature is plotted against every other feature. In conclusion, Fig. 5 was generated in the form of a bar graph for a fair comparison of the performance of different Machine Learning Models. The LASSO Regression shows a minimum performance while the XGBoost Regression outperforms the other models.

3.4. Statistical significance of parameters

With the intention of determining the statistical significance of each feature, an Ordinary Least Squares (OLS) regression model was fitted to the data. The relevant information and the important parameters are listed in Table 1. The skewness and kurtosis of this OLS model match very well with the observations of Ch. Anwar ul Hassan et al. [30]. The *p*-values of the t – statistic were then derived from the summary of the model to assess the significance of each feature in predicting the target variable. Additionally, based on the threshold of $p > 0.05$ for feature significance, likewise we dropped the 'sex' and the region-specific features (northwest, southeast, southwest) as well. The overall model was marked as significant based on the threshold of $p > 0.05$ for F–statistic. The multicollinearity was tested using the variance inflation factor (VIF) and subsequently the model was tested for auto-correlation with JB–test. Values of the variance inflation factor (VIF) of the premium attributes are recorded in Table 2. The values of the coefficient of determination that is to say, the R^2 score, of these machine learning models are enumerated in Table 3.

3.5. Implementation of ML models

The dataset was analyzed using several machine learning models to predict the target variable. Multiple algorithms were selected to compare performance across different modelling techniques. Each model

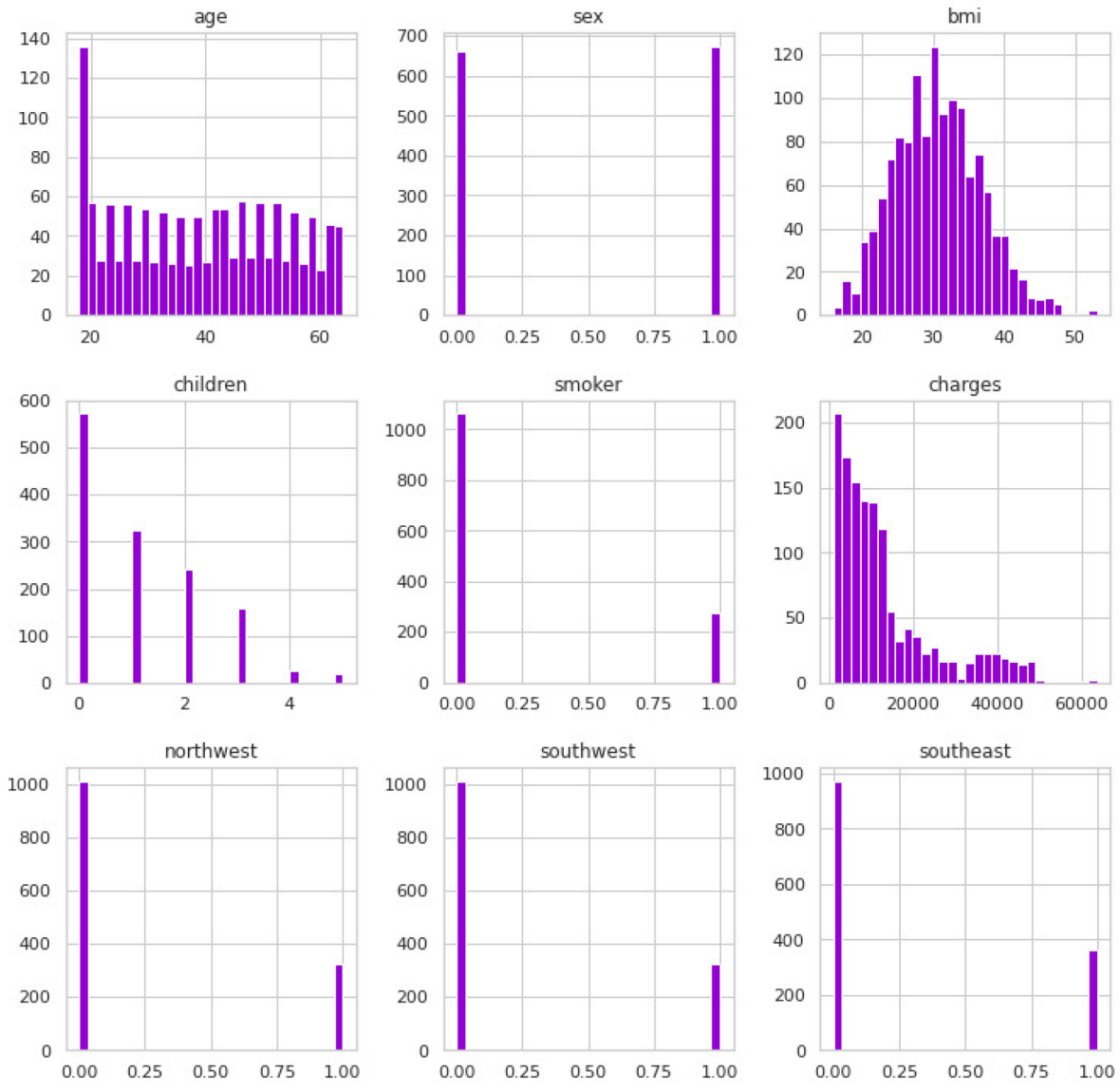

Fig. 1. Histograms for the distribution of attributes. The histograms visualize the distribution of all the attributes in the dataset including age, sex, bmi, children, smoker, charges and regional information.

was evaluated using standard regression metrics – mean absolute error (MAE), root mean squared error (RMSE) and R-squared score (R^2). Although MAE measures the average magnitude of errors in predictions, RMSE amplifies large errors due to squaring, whereas the R^2 calculates the proportion of variance in the target explicated by the predictions. A comparison of the performance of the various machine learning models that have been explored in different research studies has been described in Table 4.

3.6. Artificial neural network regression model

The dataset was split into 20% testing and 80% training data and was further subdivided into training and validation sets. The artificial neural network model, i.e., the ANN model, was constructed by means of the Tensor Flow's Keras API emphasizing five dense layers with Exponential Linear Unit (ELU) activation functions. The drop-out layers were incorporated for regularization. The model was optimized using the Adam optimizer. The model was trained for 100 epochs with a batch size of 8 and a validation split of 20%. The model accomplished a test accuracy of 93.07% and an R^2 score of 92.25%.

By examining the R^2 values given in Tables 3 and 4 and the bar diagram plotted in Fig. 5, it is comprehensible that the performance of the XGBoost is the best with an adjusted R^2 value of 90.1% and this is closely followed by the Artificial Neural Network model

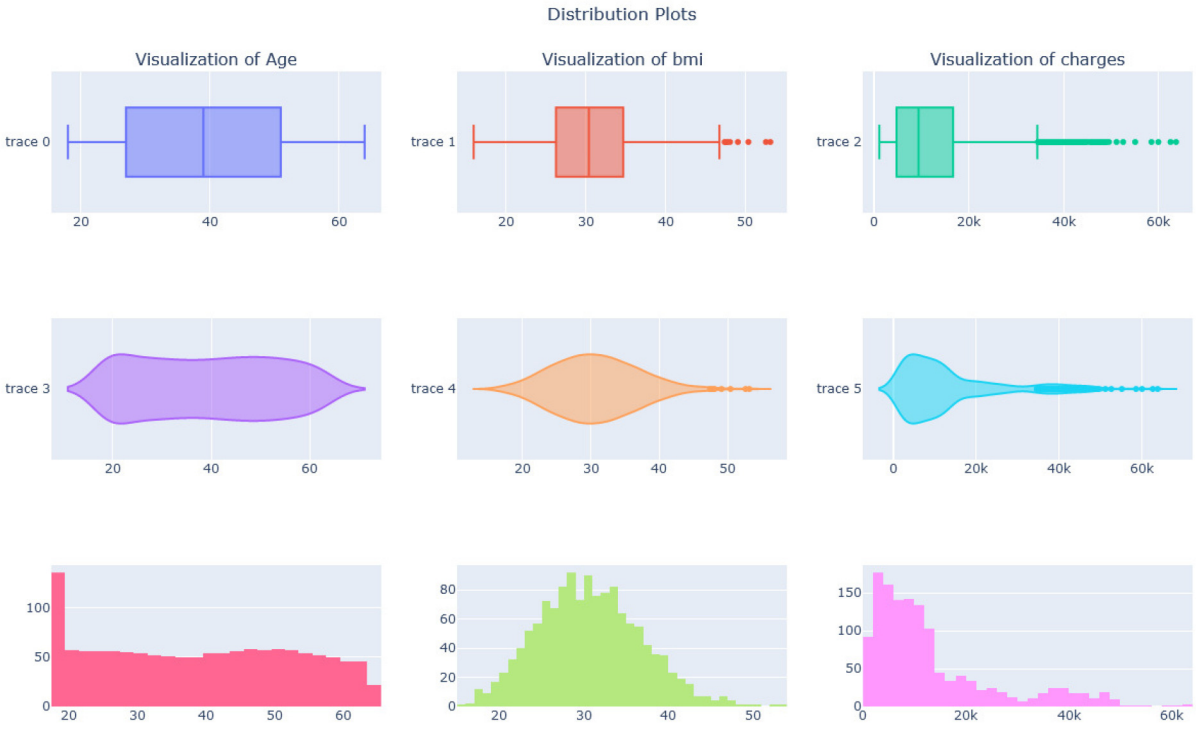

Fig. 2. Distribution plots for the visualization of three key variables, i.e., age, bmi and charges. The visualization is generated by means of box plots, violin plots, and histograms for each variable.

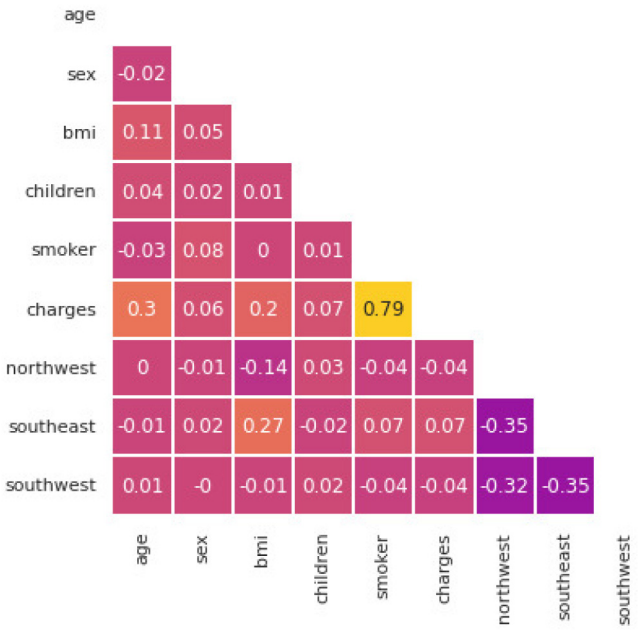

Fig. 3. Digitization of attributes. It is a bivariate map that displays the correlation coefficients between all the numerical attributes.

constructed in the present examination. The ANN model was optimized by taking advantage of the Adam optimizer. The model was trained for 100 epochs with a batch size of 8 and a validation split of 20%. The model accomplished a test accuracy of 93.07% and an R^2 score of 92.25%. These results are decidedly promising in comparison with the research work done by Keshav et al. [29] in which they trained and evaluated an artificial intelligence network-based regression model to predict health insurance premiums. The authors predicted the health insurance cost incurred by individuals on the basis of a host of

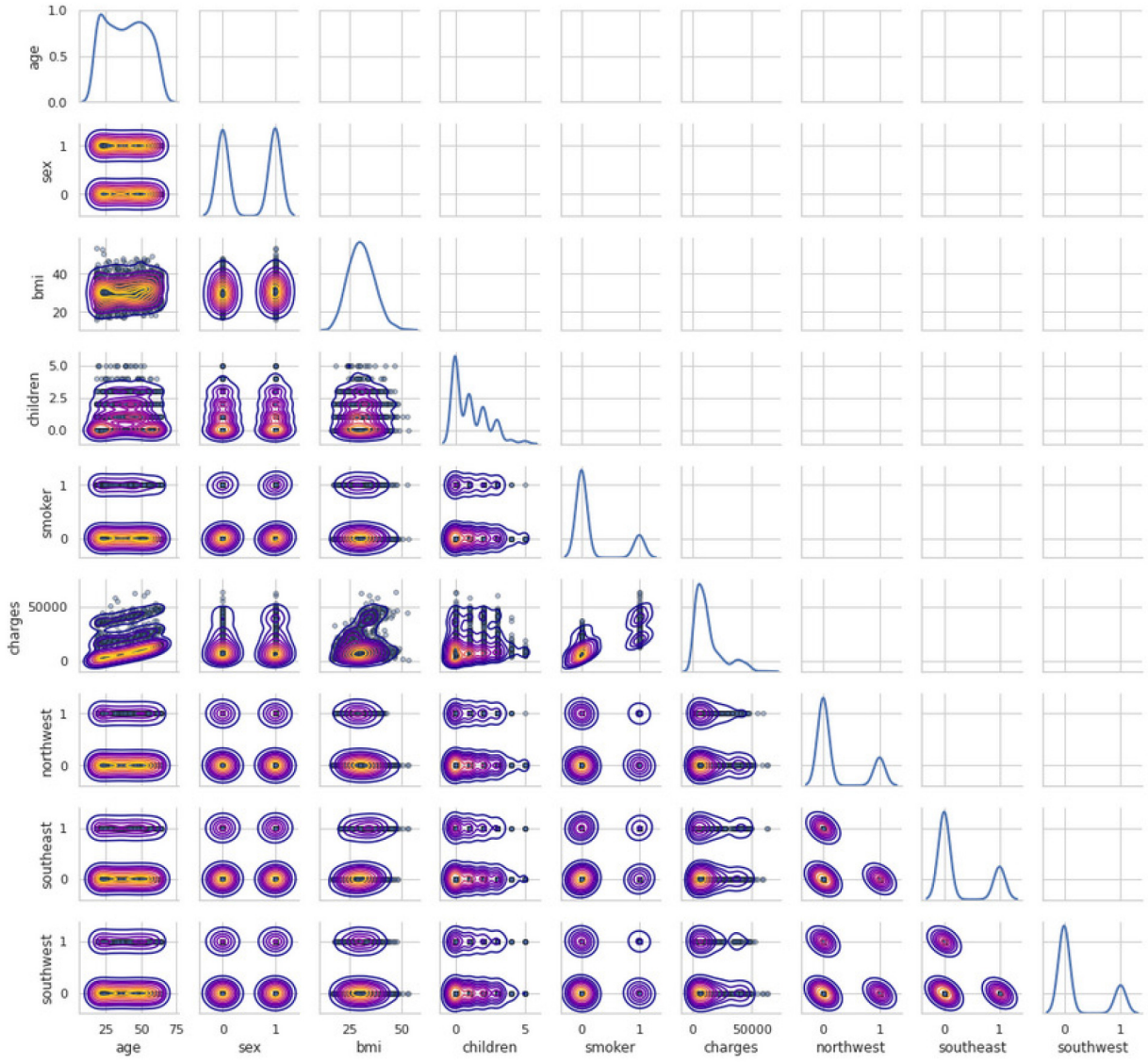

Fig. 4. Pair-Plot; this plot represents the relationship between the different features in the dataset. It is a matrix of plots where each parameter is plotted against every other parameter.

parameters like age, gender, bmi, number of children, smoking habits, and geographical location with a consequential accuracy of 92.72%. Equivalent research verdicts are forecasted by the research team of Ch. Anwar ul Hassan et al. [30] In their findings, the Stochastic Gradient Boosting model stood out prominent and claimed an accuracy of 86%. Respectively, Takeshima et al. [31] developed a prediction model by using the data of health insurance claims. They applied the LASSO regression model and used excess logarithmic medical cost as explained valuables and diseases as explanatory valuables in order to avoid overfitting. In their research work, Sen Hu et al., [32] analyzed lapse behavior in life insurance policies. The incorporation of spatially linked demographic census data provided insights for insurers to assess and prevent lapsing risks, enhancing the calculation and preparation of capital reserves. Likewise employing machine learning algorithms, Angela D. Kafuria [33] developed a predictive model for determining health insurance charges. They highlighted in their research that Gradient Boosting and Random Forest Regression are the most effective predictive models. Whereas Angeline Christobel and Suresh Subramanian [34] compared four regression analysis algorithms to predict insurance and concluded that Polynomial Regression was the most effective model with an accuracy of 88%. This comparison elucidates that the ANN model utilized in the present research work is decidedly effective in prediction, comparison and visualization of

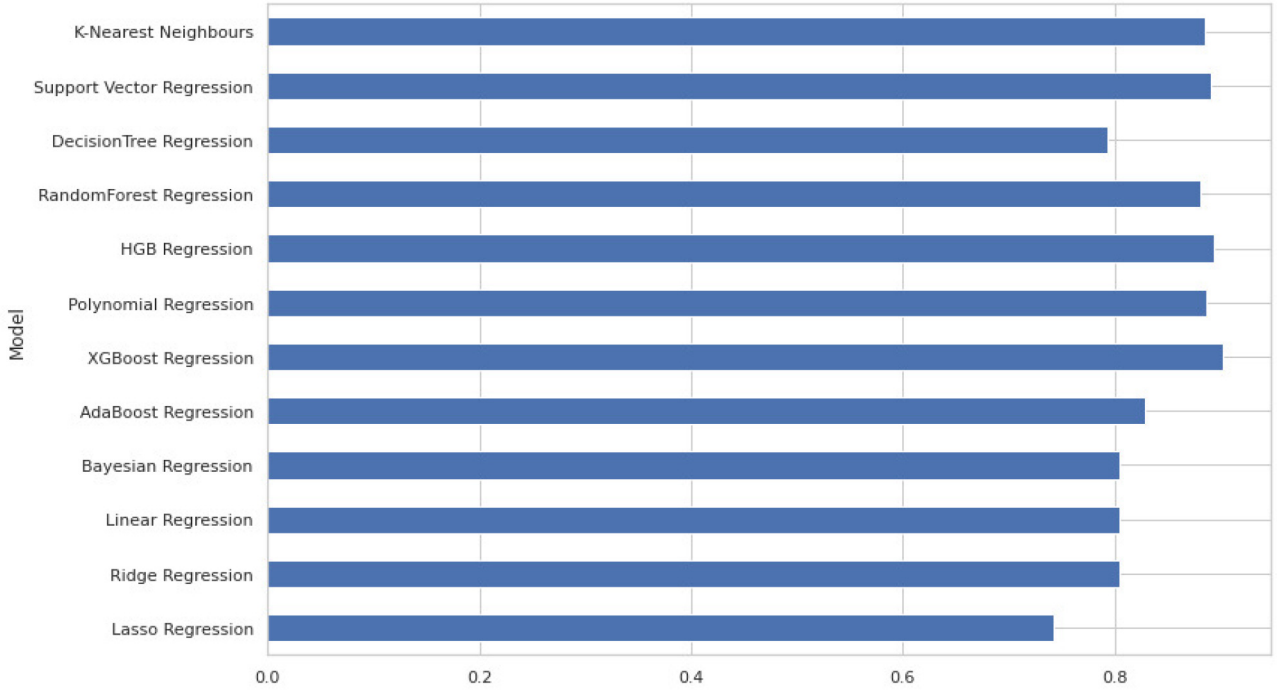

Fig. 5. Bar graph for the comparison of performance of the Machine Learning Models. The LASSO Regression shows a minimum performance while the XGBoost Regression outperforms the other models.

Table 1. List of the important parameters from the ordinary least squares fitting.

Results: Ordinary least squares			
Model:	OLS	Adj. R-squared:	0.749
Dependent Variable:	charges	AIC:	27094.2154
No. Observations:	1337	BIC:	27140.9991
Df Model:	8	Log-likelihood:	–13538
Df Residuals:	1328	F-statistic:	500
R-squared:	0.751	Prob (F-statistic):	0
		Scale:	3.68E + 07

Variable	Coef.	Std. Err.	t	P > |t|	[0.025]	[0.975]
const	–11936.5575	988.2274	–12.0788	0.0000	–13875.2146	–9997.9004
age	256.7646	11.9122	21.5547	0.0000	233.3958	280.1334
sex	–129.4815	333.1952	–0.3886	0.6976	–783.1278	524.1648
bmi	339.2504	28.6113	11.8572	0.0000	283.1222	395.3786
children	474.8205	137.8969	3.4433	0.0006	204.3009	745.3401
smoker	23847.3288	413.3479	57.6931	0.0000	23036.4428	24658.2149
northwest	–349.2265	476.8238	–0.7324	0.4641	–1284.6365	586.1835
southeast	–1035.2656	478.8670	–2.1619	0.0308	–1974.6838	–95.8474
southwest	–960.0814	478.1059	–2.0081	0.0448	–1898.0066	–22.1561

Omnibus:	299.816	DurbinWatson:	2.089
Prob (Omnibus):	0	JarqueBera (JB):	716.552
Skew:	1.211	Prob (JB):	0.000
Kurtosis:	5.646	Condition No.:	311

Table 2. Values of the variance inflation factor (VIF) of the premium attributes.

Feature	VIF
Const	35.504594
Age	1.016794
Bmi	1.106742
Children	1.004017
Smoker	1.012100
Northwest	1.517673
Southeast	1.651779
Southwest	1.529044

Table 3. Comparison of the MAE, RMSE and adjusted R^2 values of the different machine learning models.

Machine Learning Models	MAE	RMSE	Adjusted R^2
Linear Regression	0.346	0.244	0.801
Ridge Regression	0.346	0.495	0.804
Lasso Regression	0.391	0.568	0.738
Bayesian Regression	0.346	0.495	0.801
Polynomial Regression	0.221	0.377	0.885
XGBoost	0.184	0.349	0.901
AdaBoost	0.388	0.446	0.839
Histogram-based Gradient Boosting Regression Tree (HGB)	0.214	0.365	0.892
Random Forest Regressor	0.217	0.388	0.878
Decision Tree	0.253	0.509	0.790
Support Vector Regression	0.175	0.369	0.890
K-Nearest Neighbors	0.223	0.378	0.884

Table 4. Comparative evaluation of machine learning models for predictive accuracy.

	Ch. Anwar ul Hassan et al. [30]		Kaushik et al. [29]		Kafuria et al. [33]		The Current Study	
Model	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
XGBoost	0.343	88.30%	Not used	Not used	Not used	Not used	0.349	90.20%
Random Forest	0.348	87.90%	Not used	Not used	4783.8	85.30%	0.388	88.10%
SVR	0.359	87.10%	Not used	Not used	Not used	Not used	0.369	89.10%
ANN	Not used	Not used	0.27	92.60% (*accuracy* = 92.72%)	Not used	Not used	0.324	92.25% (*accuracy* = 93.07%)
Linear Regression	0.48	78.20%	0.49	74.90%	Not used	Not used	0.244	80.40%
Ridge Regression	0.465	78.30%	Not used	Not used	Not used	Not used	0.495	80.40%
Lasso Regression	Not used	Not used	Not used	Not used	6384.3	73.90%	0.568	74.20%
Polynomial Regression	Not used	Not used	Not used	Not used	Not used	Not used	0.377	88.60%
AdaBoost	Not used	Not used	Not used	Not used	Not used	Not used	0.446	82.90%
Histogram-based Gradient Boosting Regression Tree	Not used	Not used	Not used	Not used	Not used	Not used	0.366	89.30%
K-Nearest Neighbors	0.727	35.60%	Not used	Not used	5236.7	82.50%	0.378	88.50%

premium pricing. The end results uphold the ultimate evidence that machine learning offers a robust and efficient scheme for accurate premium pricing and the advantage of the use of these technological tools goes directly to both the insurers and policyholders.

4. Conclusion

The research work dealt with the effectiveness of machine learning in prediction and comparison of health insurance premiums in the real-world insurance settings. Utilizing an open-source dataset of US medical premiums, the authors have developed a thorough machine learning model in python programming language. Age, bmi, smoking habits, and geographical location were the main factors in the model. The results are encouraging demonstrating the transformative capacity of machine learning techniques in streamlining the prediction and evaluation of health insurance premiums. Curiously, XGBoost outperformed other models with an adjusted R^2 of 90.1% which was closely tracked down by the artificial neural network model of the authors that has

achieved an accuracy of 93.07%. To further refine the machine learning models, future work is needed that should explore the incorporation of nature-inspired and meta-heuristic algorithms. In brief, our research underscores the potential of machine learning to revolutionize Medical Insurance price prediction, offering valuable insights for patients, policymakers, and researchers alike. By continuing to innovate and refine our methodologies, we can contribute to more informed decision-making and sustainable healthcare practices in the face of evolving market dynamics and climate variability.

Acknowledgements

The authors are grateful to the NED University Karachi, Pakistan and the Royal Commission of Yanbu Colleges and Institutes, Saudi Arabia for providing a working environment to complete this research undertaking. This research has not received any funding or financial support from any source or agency.

Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
bmi	Body-Mass-Index
SVR	Support Vector Regression
ANN	Artificial Neural Network

Authors contribution

Mr. Farzan Tashfeen and Mr. Mohid Hussain collected the basic data and graphs. The tables and graphs were configured by Dr Nadi Alresheedi, Dr Nahlah Albarkati and Mr. Muhammad Hassan Ahmer. Dr Lal Said Jan organized and composed the research paper.

Conflict of interest

All authors declare that there is no conflict of interests.

References

1. Badawy M, Ramadan N, Hefny HA. Healthcare predictive analytics using machine learning and deep learning techniques: a survey. J Elec Syst and Inf Technol. 2023;10:40:2–45.
2. Shawni D, Bose P, Bandyopadhyay SK. Forecasting health insurance premium using machine learning approaches. Asia-Pacific Journal of Sci and Tech. 2022;28(6):1–13.
3. Orji Ug, Ukwandu El. Machine learning for an explainable cost prediction of medical insurance. Science Direct. 2024;15:100516. doi: 10.1016/j.mlwa.2023.100516.
4. Rehan H. Artificial intelligence and machine learning: The impact of machine learning on predictive analytics in healthcare. Innovative Computer Science Journal. Published online 2023:1–20. https://www.researchgate.net/publication/383611185.
5. Your total costs for health care: Premium, deductible, and out-of-pocket costs. https://www.healthcare.gov/choose-a-plan/your-total-costs/.
6. Age and Sex | CMS. https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/age-and-gender.
7. Cevolini A, Esposito E. From pool to profile: Social consequences of algorithmic prediction in insurance. Big Data Soc. Published online 2020:7. doi: 10.1177/2053951720939228.
8. van den Broek-Altenburg EM, Atherly AJ. Using social media to identify consumers' sentiments towards attributes of health insurance during enrollment season. Appl Sci. 2019;9:2035.
9. Hanafy M, Mahmoud OMA. Predict health insurance cost by using machine learning and DNN regression models. Int J Innov Technol Explor Eng. 2021;10:137–143.
10. Bhardwaj N, Anand R. Health insurance amount prediction. Int J Eng Res. 2020;9:1008–1011.
11. Boodhun N, Jayabalan M. Risk prediction in life insurance industry using supervised learning algorithms. Complex Intell Syst. 2018;4:145–154.
12. World Health Organization (WHO). WHO issues first global report on Artificial Intelligence (AI) in health and six guiding principles for its design and use who.int? June 28, 2021. https://www.who.int/news/item/28-06-2021-who-issues-first-global-report-on-ai-in-health-and-six-guiding-principles-for-its-design-and-use.
13. Goundar S, Prakash S, Sadal P, Bhardwaj A. Health insurance claim prediction using artificial neural networks. Int J Syst Dyn Appl. 2020;9:40–57.
14. Health Insurance Premium Prediction with Machine Learning. Accessed May 9, 2022. https://thecleverprogrammer.com/2021/10/26/health-insurance-premium-prediction-with-machinelearning/.
15. Ejiyi CJ, Qin Z, Salako AA, et al. Comparative analysis of building insurance prediction using some machine learning algorithms. Int J Interact Multimed Artif Intell. 2022;7:75–85.
16. Fauzan MA, Murfi H. The accuracy of XGBoost for insurance claim prediction. Int J Adv Soft Comput Appl. 2018;10:159–171. Accessed May 9, 2022. https://www.claimsjournal.com/news/national/2013/11/21/240353.htm.
17. Rustam Z, Yaurita F. Insolvency prediction in insurance companies using support vector machines and fuzzy kernel C-Means. J Phys Conf Ser. 2018;1028:012118.
18. Pereira J, Díaz O. Using health chatbots for behavior change: A mapping study. Journal of Medical Systems. 2019;43(5). doi:10.1007/s10916-019-1237-1.
19. Kumar Sharma D, Sharma A. Prediction of health insurance emergency using multiple linear regression technique. Eur J Mol Clin Med. 2020;7:98–105.
20. Rukhsar L, Bangyal WH, Nisar K, Nisar S. Prediction of insurance fraud detection using machine learning algorithms. Mehran Univ Res J Eng Technol. 2022;41:33–40.
21. AI in the insurance sector. Insurance Europe. Accessed August 15, 2023. https://www.insuranceeurope.eu/publications/2608/artificial-intelligence-ai-in-the-insurance-sector/.
22. Azzone M, Barucci E, Giuffra Moncayo G, Marazzina D. A machine learning model for lapse prediction in life insurance

contracts. Expert Syst Appl. 2022;191:116261. doi: 10.1016/j.eswa.2021.116261.
23. Sun JJ. Identification and prediction of factors impact america health insurance premium. Master's Thesis. National College of Ireland; 2020. Accessed May 9, 2022. http://norma.ncirl.ie/4373/.
24. Lui E. Employer health insurance premium prediction. Accessed May 17, 2022. https://cs229.stanford.edu/proj2012/Lui.EmployerHealthInsurancePremiumPrediction.
25. Prediction of Health Expense-Predict Health Expense Data. Accessed May 9, 2022. https://www.analyticsvidhya.com/blog/2021/05/prediction-of-health-expense/.
26. Yang C, Delcher C, Shenkman E, Ranka S. Machine learning approaches for predicting highcost high need patient expenditures in health care. Biomed Eng Online. 2018;17:131. doi: 10.1186/s12938-018-0568-3.
27. Shyamala Devi M, Swathi P, Purushotham Reddy M, et al. Linear and ensembling regression based health cost insurance prediction using machine learning. Smart Innov Syst Technol. 2021;224:495–503. doi: 10.1007/978-981-16-1502-3_49.
28. Transforming healthcare with AI: The impact on the workforce and organizations. McKinsey & Company. March 10, 2020. https://www.mckinsey.com/industries/healthcare/our-insights/transforming-healthcare-with-ai.
29. Kaushik K, Bhardwaj A, Dwivedi AD, Singh R. Machine learning-based regression framework to predict health insurance premiums. International Journal of Environmental Research and Public Health. 2022;19(13):7898. doi: 10.3390/ijerph19137898.
30. Hassan CAU, Iqbal J, Hussain S, AlSalman H, Mosleh MAA, Ullah SS. A computational intelligence approach for predicting medical insurance cost. Mathematical Problems in Engineering. Published online 2021:1–13. doi:10.1155/2021/1162553.
31. Takeshima T, Keino S, Aoki M, Matsui T, Iwasaki K. Development of medical cost prediction model based on statistical machine learning using health insurance claims data. value in health. 2018;21:S97. doi: 10.1016/j.jval.2018.07.738.
32. Hu S, O'Hagan A, Sweeney J, Ghahramani M. A spatial machine learning model for analyzing customer's lapse behavior in life insurance. Annals of Actuarial Science. Published online 2020:1–27. doi: 10.1017/s1748499520000329.
33. Kafuria AD. Predictive model for computing health insurance premium rates using machine learning algorithms. International Journal of Computer. 2022;44(1):21–38.
34. Christobel YA, Subramanian S. An empirical study of machine learning regression models to predict health insurance cost. Webology. 2022;19(2):29–35.