FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 8 No. 6, December, (Special Issue) 2024, pp 354 - 362
DOI: https://doi.org/10.33003/fjs-2024-0806-3120

PREDICTIVE MODEL FOR HEALTH INSURANCE COST USING SELF-ORGANIZING MAPS AND XGBOOST

Ogunsanwo, Gbenga O.

Department of Computer Science, College of Science and Information Technology, Tai Solarin University of Education, Ogun State. Nigeria

*Corresponding authors' email: ogunsanwogo@tasued.edu.ng

ABSTRACT

Machine Learning (ML) techniques are gaining more adoption in every sector in order to improve their services. The healthcare industry is not left behind in this development of adopting ML predictive model to increase their efficiency and productivity.The paper developed a predictive healthcare insurance cost Model using Self-Organizing Maps (SOM) and XGBoost models. In this study, two models, SOM and XGBoost were deployed for medical insurance cost prediction using the dataset from KAGGLE's repository which consists of 1338 instances and 7 predicting parameters. The dataset were preprocessed and thereafter divided into 80% for training and 20% for testing. The comparative result from the prediction showed that the two models achieved impressive outcomes, and whereas the XGBoost model achieved the results of MAE score of 2432.04, MSE of 18030522.49 and RMSE of 4246.24. The SOM model achieved result of MAE score of 3978.29, MSE of 32775593.13 and RMSE of 5724.997216804203. The quantization error of 0.5135462765843376 and topographic error of 0.9730941704035875 generated for SOM model developed. The study concluded that XGBoost outperformed SOM for the insurance predictive model developed and the model is recommended for healthcare sectors to assist in decision making as regard to health insurance cost. More future works can be done using more predicting factors in the dataset and other machine learning algorithms can be employed.

Keywords: Insurance cost, Machine learning, SOM, XGBoost, Predictive model

INTRODUCTION

Health insurance is a critical measure that has been put in place in most countries of the world to provide optimizing provider cost management, prompt and accurate reimbursements. Health insurance is a process of billing, looking over, verifying medical records and payment authorization. The insurance company depends seriously on data analytics to evaluate risk, set premiums and manage claims. Accurate cost estimates will be of great help to health insurers and assist healthcare delivery organizations to prioritize the allocation of limited care management resources while planning for the future. Predicting insurance cost is a major task that assists insurance company to ascertain profitability while providing affordable pricing. In the past traditional method, which uses statistical methods to evaluate historical data and identify variables causing costs, like age, health status, location and the lifestyle have been widely used with some flaws. Machine Learning (ML) is the process of permitting computer to learn from input data and makes prediction (Ravi *et al.*, 2021). ML can help in the insurance sector to improve the effectiveness of policy wording (Hassan et al., 2021). It could also predict high-need and high-cost patient expenditures in healthcare sector (Yang *et al.*, 2018). ML can be classified into three major types. These include supervised machine learning, unsupervised machine learning and reinforcement learning. The features of learning problems involve tasks that must be learnt, experience that will be gained and improved performance measures (Taiwo *et al.*, 2023). However, the complex interplay between these factors often results in non-linear relationships and hidden patterns that are difficult to capture using conventional techniques. Therevarious works done in the area of insurance cost prediction, for example, Sabarinath & Mathew (2024) developed a medical insurance cost prediction system using machine learning techniques. They applied Support Vector Regression (SVR), linear regression (LR) and Ridge regression (RR) algorithms to construct predictive model. So also Mishra et al. (2024) proposed a model to forecast health insurance cost. They employed Gradient Boosted Trees, Support Vector Machine, Random Forest Regression and Linear Regression to build models. Orji & Ukwandu (2023) developed a model to predict medical insurance costs. They utilized Random Forest, Extreme Gradient Boosting and Gradient-boosting Machine to construct model. Kulkarni et al. (2022) developed proposed a model to predict medical insurance cost. They used Decision Tree Regression, Gradient Boosting Regression and Linear Regression to construct models. Hassan et al. (2021) proposed a model to forecast medical insurance cost. They employed Random Forest Regressor, Linear Regression, Ridge Regressor, Support Vector Regression, Decision Tree, Stochastic Gradient Boosting, XGBoost, k-Nearest Neighbors, and Multiple Linear Regression to build models. Hanafy & Mahmoud (2021) developed a model to predict health insurance cost. They applied Support Vector Machine, Multiple Linear Regression, CART, Generalized Additive Model, XGBoost, Random Forest Regressor, Stochastic Gradient Boosting, k-Nearest Neighbors and Deep Neural Network to build models. Rao et al. (2023) developed a model to interpret the premium prediction of health insurance. They applied Random Forest Regression, Linear Regression and Multiple Regression to build models. Rohan et al. (2023) developed a model to forecast Medical Insurance Premium. They used Support Vector Machine, Random Forest Regression Linear and Regression algorithms to build models.. Kaushik et al. (2022) proposed a model to predict health insurance premiums. They employed artificial neural network to construct model. Choi et al. (2022) developed a model to forecast high-cost of National Health Insurance Service of Korea. They used XGBoost, logistic regression and random forest algorithms to build models. Bhardwaj and Anand (2020) developed a model to predict health insurance amount. They utilized Gradient

Boosting Regression, Multiple Linear Regression and Decision tree regression algorithms to build models. Thorat et al. (2023) proposed model to predict medical insurance cost. They applied Decision Tree regression, Gradient boosting regression and multiple linear regression to construct models. Benarji et al. (2023) developed a model to forecast minimum health insurance premium. They employed XGB Regressor, Decision Tree Regressor, Linear Regression, Gradient Boosting Regressor, K Neighbours Regressor, LGBM Regressor and Random forest algorithms to build models. Orji & Ukwandu (2023) developed a model to predict explainable cost of medical insurance. They deployed Gradient-boosting Machine (GBM), Random Forest (RF) and Extreme Gradient Boosting (XGBoost) to build models. Pandey et al. (2021) proposed a model to forecast medical insurance cost. They applied Support Vector Regression, Linear Regression and Random Forest Regressor to build models. In recent years, Self-Organizing Maps (SOM), a type of unsupervised learning algorithm, has gained attention for their ability to visualize and cluster high-dimensional data. SOMs are particularly effective in identifying patterns, detecting anomalies, and segmenting data based on underlying similarities. By using SOMs, the development of predictive model for insurance costs can help in numerous methods such as Clustering and Segmentation: SOMs can group policyholders into clusters with similar risk profiles, aiding in more personalized premium calculations and other benefits such as feature relationship, anomaly detection, and dimensionality reduction as Insurance datasets often involve many interrelated variables. SOMs reduce these into an intuitive two-dimensional map without losing important relationships.. Despite, these numerous benefit, the use of SOMs in insurance cost prediction remains under-explored compared to other machine learning methods. This study aims to bridge that gap by leveraging SOMs to cluster and analyze policy holder data, identify significant factors affecting insurance costs, and ultimately enhance predictive accuracy. In this study, we used unsupervised ML algorithm and supervised to build insurance cost predicitve models and compare the accuracy of the two models. The objective of this study is to make use of SOM and XGBoost models to predict healthcare insurance cost in the health sector, compare the performance results of machine learning algorithms used to forecast cost of healthcare insurance using public dataset from kaggle repository and provide a guide for effective healthcare insurance cost prediction system.

MATERIALS AND METHODS

The Dataset used in this study was retrieved from kaggle.com and the file was saved in csv format, it consists of 1338 instances and 7 predicting parameters which are: age, sex ,bmi , children, smoker, region and charges.as seen in Table 1. The work flow for the insurance cost predictive model as shown in Figure 1.

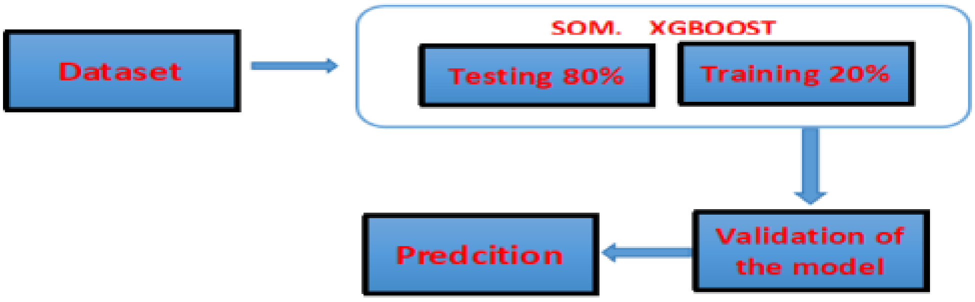

Figure 1: Architectural Model of the study

Table 1: Sample of the dataset

Age	Sex	Bmi	children	Smoker	Region	Charges
19	Female	27.9	0	Yes	southwest	16884.924
18	Male	33.77	1	No	Southeast	1725.5523
28	Male	33	3	No	Southeast	4449.462
33	Male	22.705	0	No	northwest	21984.47061
32	Male	28.88	0	No	northwest	3866.8552
31	Female	25.74	0	No	Southeast	3756.6216
46	Female	33.44	1	No	Southeast	8240.5896

Data preprocessing

The stage of preprocessing is necessary in order to improve the quality of the dataset and to improve the model accuracy. Here data cleaning was done in python environment, the duplicates were removed and One-hot encoding was done on the dataset in which the columns that are not necessary were dropped. The dataset are grouped into two variables dependent and independent variables. The x = ['age', 'bmi', 'children', 'smoker', 'sex', 'region'] and y = ['charges']

Model Development

In order to develop the insurance cost predictive model, the dataset used was imported into google colab environment as seen in Figure 2. The dataset was divided into 80% training and 20% testing as seen in Figure 3. Thereafter, two algorithms were employed which are Self-Organizing Maps (SOM), and XGBoost model.

	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

Figure 2: Sample of processed dataset

SOM model

SOM is an example of artificial neural network used to produce a low dimensional representation of high dimensional input data. SOM has six phases. It is useful for visualization and analysis of complex dataset (Kohonen, 2001).

The mathematical Model of SOM

Stage 1 Initialization: The SOM comprises of networks of neurons which are arranged in 2-dimensional grid. These weight vectors are randomized first.

Stage 2 Competitive Learning: For each input data point, Distance between the input vector and individual neutrons weight is calculated using Euclidean distance. Thereafter neuron with weight closer to the input vector is identified as the Best Matching Unit (BMU). Then the BMU's weight and the weight vectors of its neighboring neurons are updated

Stage 3 Weight Update

The weight updated of the SOM model is showing equation 1

$$w_i(t+1) = w_i(t) + \alpha(t) * h_{ci}(t) * [x(t) - w_i(t)] \quad (1)$$

Where:

$w_i(t)$ is the weight vector of neuron i at time t.

$\alpha(t)$ is the learning rate, which decreases over time.

$h_{ci}(t)$ is the neighborhood function, which defines the influence of the BMU on its neighbors. It decreases with distance from the BMU and over time.

$x(t)$ is the input vector at time t.

Stage 4 Neighborhood Function

The Neighborhood Function make used of the Gaussian function as shown in equation 2

$$h_{ci}(t) = \exp(-d_{ci}^2 / (2\sigma(t)^2)) \quad (2)$$

Where:

d_{ci} is the distance between neuron i and the BMU c.

$\sigma(t)$ is the neighborhood radius, which decreases over time.

Stage 5 Training Iterations: here stage 2 and 3 are repeated many iteration at a predefined epoch is attained or the weight vector converge

Stage 6 : SOM forms a low dimensional representation of the input data after training , where similar input vector are mapped on the grid. This map can be used for visualization

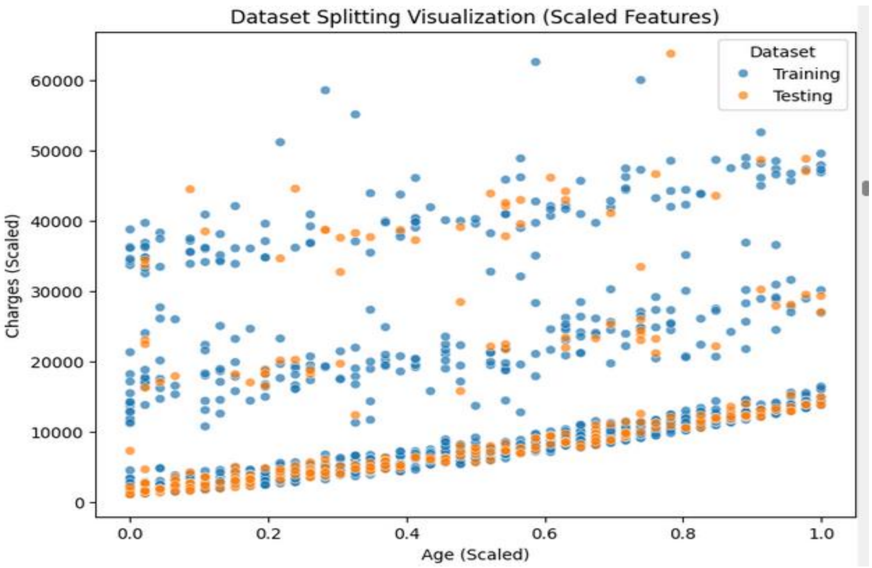

Figure 3: Testing and Training

SOM Model Development

The SOM Distance Map

The SOM distance map is a good instrument used for visualizing and interpreting the relationships between data points in SOM. It provides insights into clusters structure of the data and the boundaries. In general, SOM distance map helps in SOM analysis. It is constructed by calculating the Euclidean distance between the weight vectors of neighboring neurons on the SOM grid. In this study, the distance map reveal cluster of parameter with similar behavior as seen in Figure 4. Regions with high similarity were represented with dark colours. The location of a data point on the map corresponds to the neuron with the closest weight vector. The values on the Map used to represent the similarity between. When the values are small it suggests high similarity and high values suggest low similarity (Kohonen, 2001). The SOM map revealed more of small values among neurons this suggests high similarity among predicting factors such as age sex , bmi charges, children , smokers.

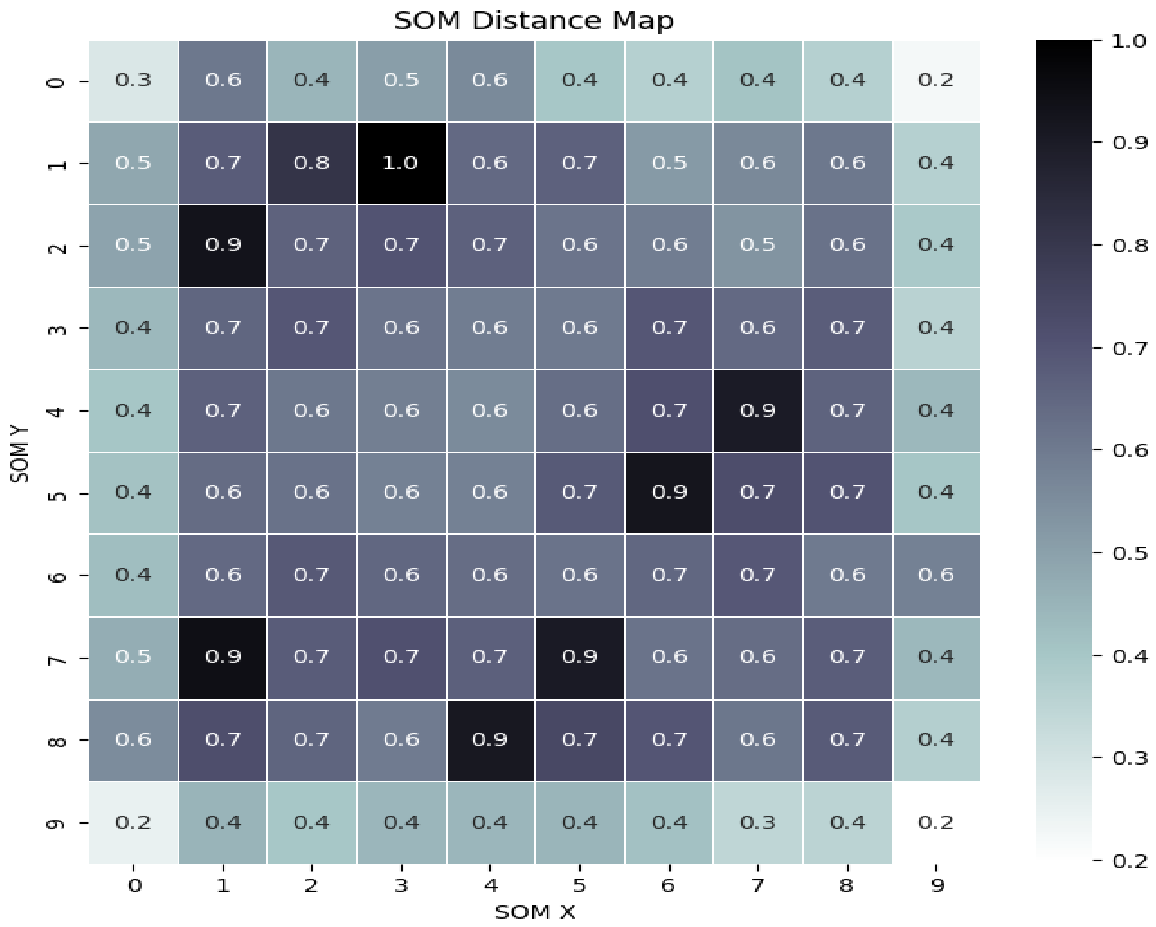

Figure 4: SOM Distance Map

RESULTS AND DISCUSSION

SOM Validation using validating metrics

These metrics provide understandings into how well the model is predicting the target variable insurance charges. In SOMs, Topographic Error estimates the proportion of data points where the first and second best matching units (BMUs) are not adjacent on the map grid. It shows how adequate the SOM preserves the topology of the input data. Table 2 shows the results of Quantization error and Topographic error of the SOM Model. The quantization error of 0.5135462765843376 and topographic error of 0.9730941704035875 generated shows a lower quantization error which signifies that the SOM is representing the data with less distortion (Vesanto & Alhoniemi, 2000). The Mean Absolute Error (MAE) measures the average absolute difference between actual and predicted values: Means Square Error (MSE) is the average of the squared differences between the predicted values and the actual values. It measures the average squared prediction error. Root Means Square Error (RMSE) is the square root of the MSE. It represents the average magnitude of the prediction errors in the same units as the target variable. The values of MAE, MSE and RMSE are shown in Table 3.

Table 2: Quantization error and Topographic error of SOM

S/No	Validating Metrics	Values
1	Quantization error	0.5135462765843376
2	Topographic error	0.9730941704035875

Table 3: SOM Validating Metric

S/No	Validating Metrics	Values
1	MAE	3978.2877792873137
2	MSE	32775593.132415872
3	RMSE	5724.997216804203

The Learning Curve

Learning curves are graphical representations that show the relationship between the performance of a machine learning model and the amount of training data used to train the predictive model. The MSE, RMSE and MAE learning curve for the insurance predictive model developed is seen in Figure 5, 6 and 7 respectively. The Learning curve shows High Training Score and High Validation Score which converge. This indicates a good fit that the model is performing well on both training and validation sets, and the scores are converging as more training data is used. This suggests the model is generalizing well to unseen data

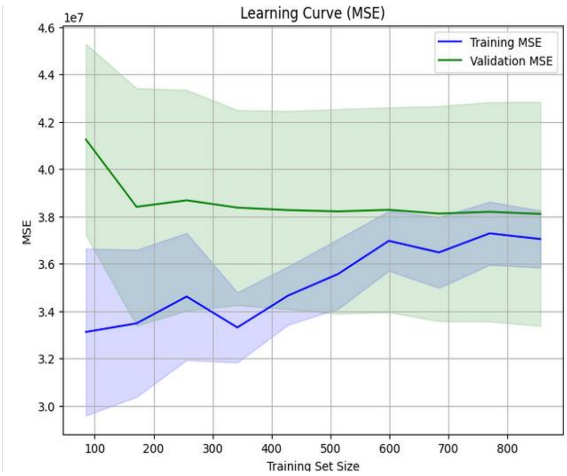

Figure 5: MSE Learning curve

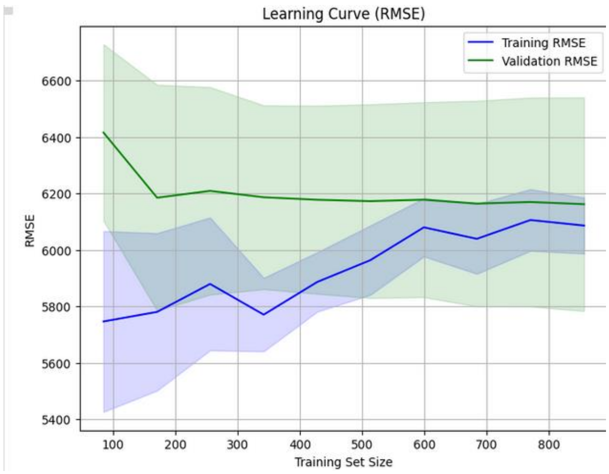

Figure 6: RMSE Learning curve

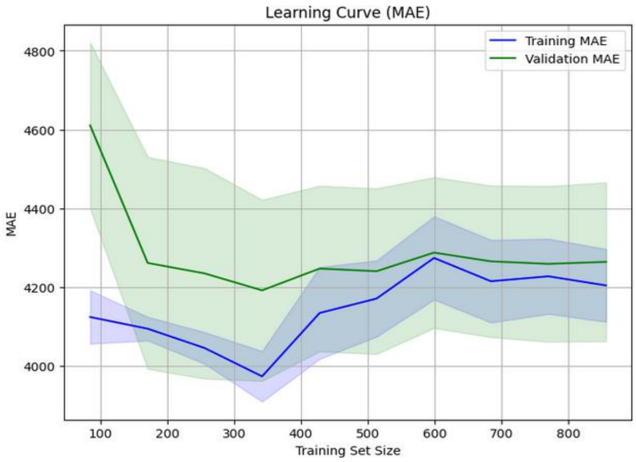

Figure 7: MAE Learning curve

The Scatter plot for Insurance cost predictive Model
This scatter plot explores the relationship between body mass weight (bmi) and insurance charges. Each dot represents a person, with their bmi on the horizontal axis and their charges on the vertical axis. By observing the distribution of the dots, we can determine if there's a trend—whether charges tend to increase, decrease, or stay relatively constant as bmi increases as seen in Figure 8. The scatter plot revealed a relationship between insurance charges and smoker as the dot for smokers overlap each other and increase rapidly.

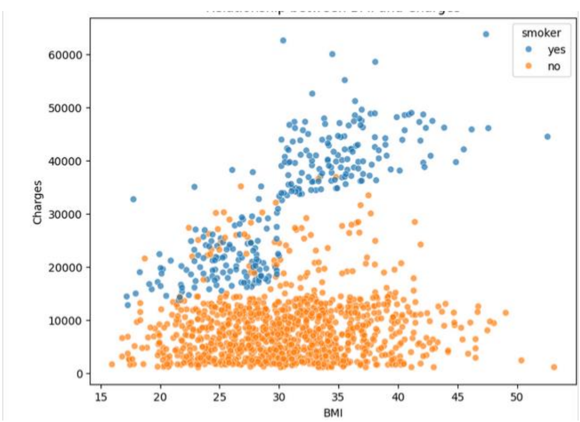

Figure 8: Relation between charges and Bmi

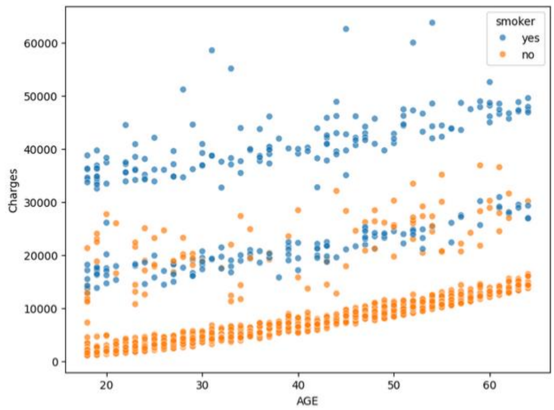

Figure 9: Charges and Age

The Scatter plot for Insurance cost predictive Model
This scatter plot explores the relationship between age and insurance charges as seen in Figure 9. It also explores the relationship between sex and insurance charges as senn in Figure 10 and finally scatter plot explores the relationship between children and insurance charges as seen in Figure 11. Each dot represents a person, with their age on the horizontal axis and their charges on the vertical axis. By observing the distribution of the dots, we can determine if there's a trend—whether charges tend to increase, decrease, or stay relatively constant as age increases.

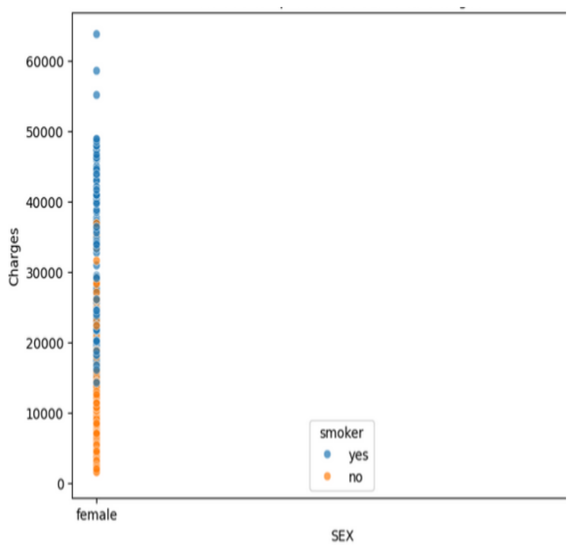

Figure 10: Scatter plot for sex and charges

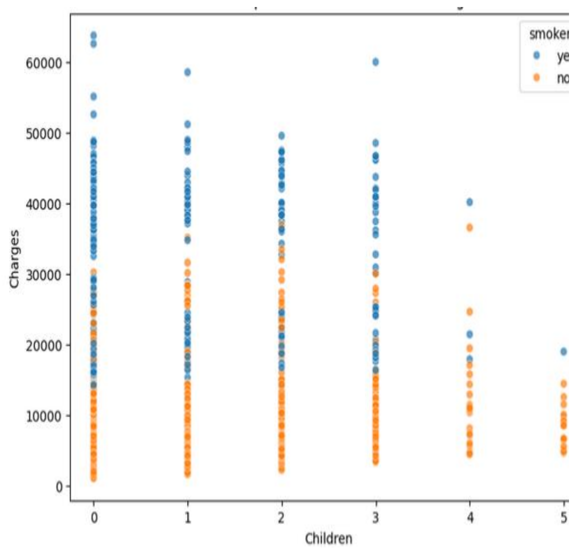

Figure 11: Scatter plot for children and charges

The box plot for charges against children
The interquartile range (IQR), containing the middle 50% of the data for each sex. This shows where the bulk of the charges fall for each group. The size of the shape of the boxes (IQR) is larger and the size of whisker is longer in the case of the male which suggests that there is variability in charges among the male while smaller box (IQR) in female indicates that charges are more tightly clustered around the median. As seen in Figure 13. The box also reveal that charges varies with the number of children as seen in Figure 12

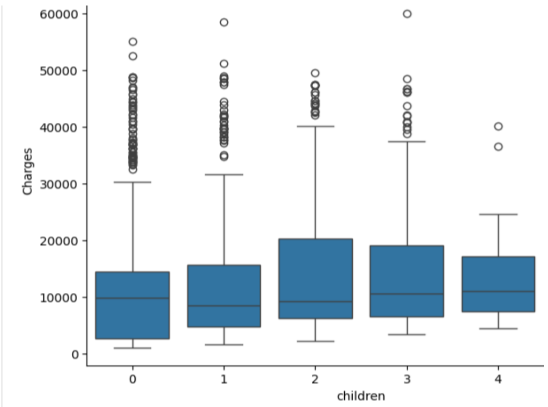

Figure 12: Box plot for charges and children

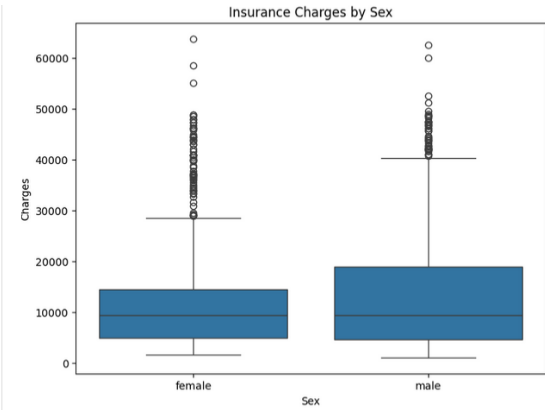

Figure 13: box plot for charges and sex

XGBoost Model
The understanding of the scatter plot helps to gain insights into XGBoost model's strengths and weaknesses. The points cluster around the ideal line shows tighter cluster indicates higher accuracy as seen in figure 14. The plot indicates consistency in the model prediction which suggests that the issues of bias have been taken off. The points spread out which also reveals high variance this propose that the model is sensitive to small changes as seen in figure 15.. The box plot also reveals that there are longer whiskers for smoker which shows that charges increase in the case of smoker. as seen in figure 16.

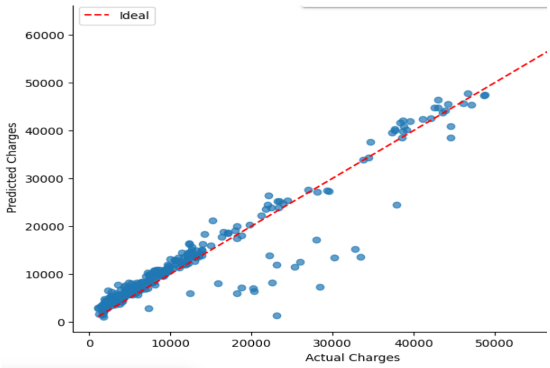

Figure 14: XGBoost model

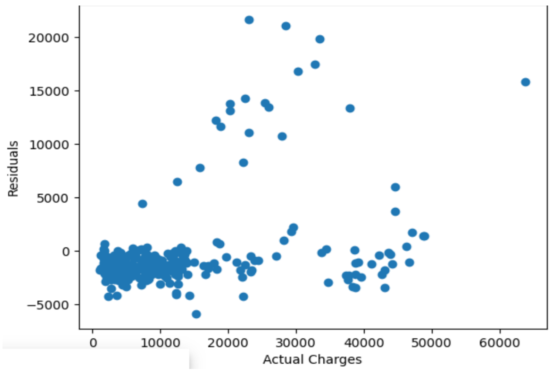

Figure 15: Residuals against charges

Table 4: XGBoost Validating Metrics

S/No	Validating Metrics	Values
1	MAE	2432.049324739739
2	MSE	18030522.491324932
3	RMSE	4246.236273610423

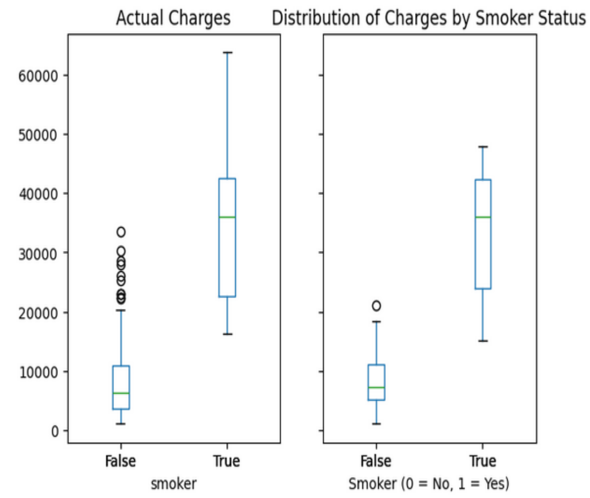

Figure 16: box plot of charges against charges

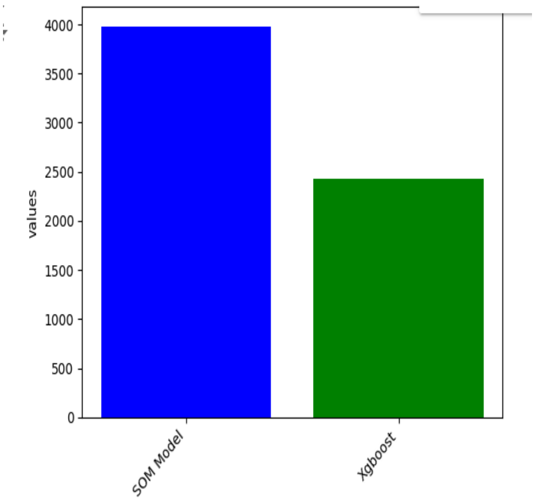

Figure 17: MAE Comparison

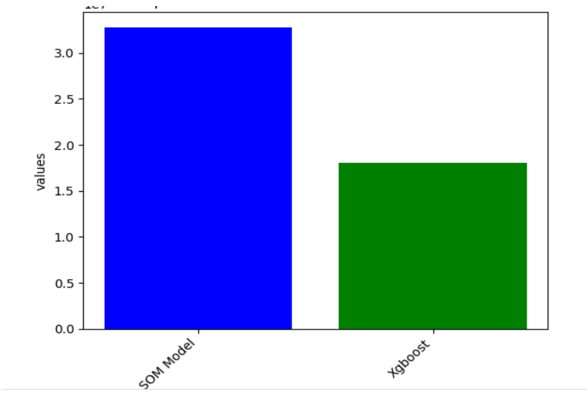

Figure 18: MSE Comparison

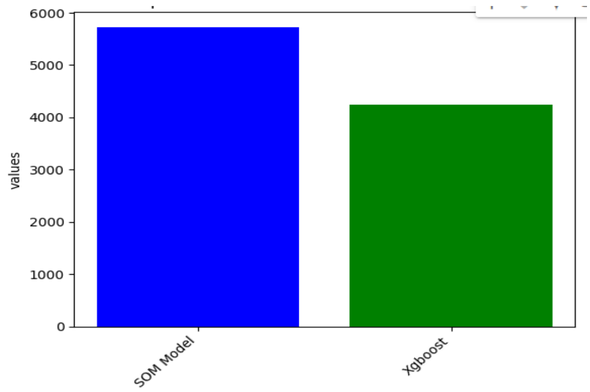

Figure 19: RMSE Comparison

Table 5: Performance evaluation of SOM and XGBoost

S/No	Validating Metrics	MAE Values	MSE values	RMSE Values
1	SOM	3978.29	32775593.13	5724.99
2	XGBoost	2432.05	18030522.49	4246.24

Discussion of the findings

The developed models were each evaluated based on their performance on validating metrics. The result of the model for insurance cost predictive model developed shows that XGBoost model has significantly lower values for MAE, MSE and RMSE compared to the SOM model as seen in Table 5. This indicates that XGBoost model outperform the SOM model in terms of prediction accuracy. It has lower average errors and is good in reducing both small and large errors. It demonstrate better predictive performance compared to the SOM model. The SOM distance map a good instrument used for visualizing and interpreting the relationships between data points in SOM was developed. The SOM map revealed more of small values among neurons this suggests high similarity among neuron with the closest weight vector in line with the study of (Kohonen, 2001). The quantization error of 0.5135462765843376 and topographic error of 0.9730941704035875 generated shows a lower quantization error which signifies that the SOM is representing the data with less distortion in line with the findings of (Vesanto & Alhoniem 2000). The Learning curve shows high training Score and high validation Score which converge. This indicates a good fit that the model is performing well on both training and validation sets. The study developed a scatter plot to explore the relationship between predicting parameters e.g. bmi, sex, children and insurance charges. Each dot represents a parameter on the horizontal axis and their charges on the vertical axis. The box plots also reveal that charges varies with the number of children and suggests that there is variability in charges among the male When compared with the works of Chandrashekha et al (2024) which Forecast Medical Insurance Costs with RMSE of 4684320.39 the models in the study are better with lower RMSE 4246.24. This suggests the models are more useful than that of their work.

CONCLUSION

In this paper, two models, SOM and XGBoost were deployed for medical insurance cost prediction using the dataset from

KAGGLE's repository. The comparative result from the prediction showed that the two models achieved impressive outcomes, and whereas the XGBoost model achieved better result of MAE score of 2432.04, MSE of 18030522.49 and RMSE of 4246.24 while the SOM model achieved result of MAE score of 3978.29, MSE of 32775593.13 and RMSE of 5724.997216804203. The SOM map developed revealed small values among neurons this suggests high similarity among neuron with the closest weight vector. The quantization error of 0.5135462765843376 and topographic error of 0.9730941704035875 generated shows a lower quantization error which signifies that the SOM is representing the data with less distortion. The Learning curve shows high training Score and high validation Score which converge. This indicates a good fit that the model is performing well on both training and validation sets. The study developed a scatter plot to explore the relationship between predicting parameters e.g. bmi, sex, children and insurance charges. The box plots also revealed that charges varies with the number of children and suggests that there is variability in charges among the male The model is recommended for healthcare sector to assist in decision making as regards to health insurance cost. More future works can be done using more predicting factors in the dataset and other machine learning algorithms can be employed.

REFERENCES

Benarji, P. S., Kumar, K. P., Sushanth, M., Gogi, M. & Neeraja, D. (2023). Minimum health insurance premium prediction using health parameters. *International Research Journal of Engineering and Technology (IRJET)*, 10(9), 585-589. e-ISSN: 2395-0056. p-ISSN: 2395-0072.

Bhardwaj, N. & Anand, R. (2020). Health insurance amount prediction. *International Journal of Engineering Research & Technology (IJERT),* 9(5), 1008-1011. ISSN: 2278-0181. Available at http://www.ijert.org

Chandrashekhar H. Patil, Ashish Garud, Ajinkya Sonawane, Akash Gavhane and Mayur Jadhav (2024) Forecasting Medical Insurance Costs: A Data-Driven Approach Grenze International Journal of Engineering and Technology, 01.GIJET.10.2.559_1

Choi, Y., An, J., Ryu, S. & Kim, J. (2022). Development and evaluation of machine learning-based high-cost prediction model using health check-up data by the national health insurance service of Korea. *International Journal of Environmental Research and Public Health*,19(13672), 1-16. https://doi.org/10.3390/ijerph192013672

Hanafy, M. & Mahmoud, O. M. A. (2021). Predict health insurance cost by using machine learning and DNN regression models. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 10(3), 137-143. ISSN: 2278-3075 (Online). https://doi.org/10.35940/ijitee.C8364.0110321

Hassan, C. A., Iqbal, J., Hussain, S., AlSalman, H., Mosleh, M. A. A. & Ullah, S. S. (2021). A computational intelligence approach for predicting medical Insurance cost. *Mathematical Problems in Engineering*, Volume 2021, pp. 1-13. https://doi.org/10.1155/2021/1162553

Kaushik, K., Bhardwaj, A., Dwivedi, A. D. & Singh, R. (2022). Machine learning-based regression framework to predict health insurance premiums. *International Journal of Environmental Research and Public Health,* 19(7898), 1-15. https://doi.org/10.3390/ijerph19137898

Kohonen, T. (2001). Self-Organizing Maps. Springer Series in Information Sciences (3rd ed.). Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-642-56927-2

Kulkarni, M., Meshram, D. D., Patil, B., More, R., Sharma, M. & Patange, P. (2022). Medical insurance cost prediction using machine learning. *International Journal for Research in Applied Science & Engineering Technology (IJRASET),* 10(XII), 449-456. ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538. Available at www.ijraset.com

Mishra, S., Kapoor, R., Yukti & Mahesh, G. (2024). Health insurance cost prediction using machine learning. *International Research Journal of Modernization in Engineering Technology and Science*, 6(2), 438-442. e-ISSN: 2582-5208. Available at www.irjmets.com DOI: https://www.doi.org/10.56726/IRJMETS49200

Orji, U. & Ukwandu, E. (2023). Machine learning for an explainable cost prediction of medical insurance. *Machine Learning with Applications*, 15 (2024) 100516. Available at www.elsevier.com/locate/mlwa. https://doi.org/10.1016/j.mlwa.2023.100516

Pandey, R. K., Chaitanya, K. K., Menta, M. & Gujarathi, Y. (2021). Implementation of predicting medical insurance cost with machine learning techniques. *International Journal of Food and Nutritional Sciences (IJFANS)*, 10(11), 162-165. ISSN PRINT 2319 1775 Online 2320 7876.

Rao, V. S., Iswarya, M., Hamza, S. A. & Satish (2023). Interpreting the premium prediction of health insurance through random forest algorithm using supervised machine learning technology. *International Journal of Innovative Science and Research Technology,* 8(5), 726-731. ISSN No: 2456-2165.

Ravi, A., Nandhini, R., Bhuvaneshwari, K., Divya, J. & Janani, K. (2021). Traffic management system using machine learning algorithm. *International Journal of Innovative Research in Technology (IJIRT),* 7(11), 303-308. ISSN: 2349-6002.

Rohan, D., Rohan, P., Rohith, B., Rohith, A., Rohith, T., Sri, K. R. & Sagar, K. M. (2023). Medical insurance premium prediction using machine learning. *International Journal of Research Publication and Reviews,* 4(12), 1465-1470. ISSN 2582-7421. Available at www.ijrpr.com

Sabarinath, U. S. & Mathew, A. (2024). Medical Insurance Cost Prediction. *Indian Journal of Data Communication and Networking (IJDCN),* 4(4), 1-4.

Taiwo, E. O., Ogunsanwo, G.O. & Alaba, O. B. (2023). Predictive modelling of business success using random forest, jrip and naïve Bayes algorithms. *FNAS Journal of Scientific Innovations,* 5(2), 30-37.

Thorat, M., Patel, M., Kute, Y., Sharma, M. & Bhosale, S. (2023). Medical insurance cost prediction using machine learning. *International Journal of Science, Engineering and Technology,* 11(3), 1-4. ISSN (Online): 2348-4098 ISSN (Print): 2395-4752.

Ugochukwu Orji & Elochukwu Ukwandu (2024) Machine learning for an explainable cost prediction of medical insurance Machine Learning with Applications 15 (2024) 100516

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. IEEE Transactions on neural networks, 11(3), 586-600.

Yang, C., Delcher, C., Shenkman, E. & Ranka, S. (2018). Machine learning approaches for predicting high cost high need patient expenditures in health care. *BioMedical Engineering Online*, 17(1), 131-220.