JOIV : Int. J. Inform. Visualization, 8(1) - March 2024 481-491

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Comparative Analysis of Machine Learning Algorithms for Health Insurance Pricing

Yoon-Teck Bau [a], Shuhail Azri Md Hanif [a,*]

[a] *Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, Cyberjaya, 63100, Malaysia*

Corresponding author: [*]shuhail.azri@gmail.com

***Abstract*— Insurance is an effective way to guard against potential loss. Risk management is primarily employed to protect against the risk of a financial loss. Risk and uncertainty are inevitable parts of life, and the pace of life has led to a rise in these risks and uncertainties. Health insurance pricing has emerged as one of the essential fields of this study following the coronavirus pandemic. The anticipated outcomes from this study will be applied to guarantee that an insurance company's goal for its health insurance packages is within the range of profitability so that the insurance company will also choose the most price-effective course of action. The US Health Insurance dataset was utilized for this study. This health insurance pricing prediction aims to examine four different types of regression-based machine learning algorithms: multiple linear regression, ridge regression, XGBoost regression, and random forest regression. The implemented model's performance is assessed using four evaluation metrics: MAE, MSE, RMSE, and R2 score. Random forest regression outperforms all other algorithms in terms of all four evaluation metrics. The best machine learning algorithm, random forest, is further enhanced with hyperparameter tuning. Random forest with hyperparameter tuning performs better for three evaluation metrics except for MAE. To gain further insights, data visualizations are also implemented to showcase the importance of features and the differences between actual and predicted prices for all the data points.**

***Keywords*— Health insurance pricing, machine learning algorithms, regression, multiple linear regression, ridge regression.**

Manuscript received 5 Dec. 2022; revised 21 Jul. 2023; accepted 28 Aug. 2023. Date of publication 31 Mar. 2024.

I. INTRODUCTION

We live in a world that is full of uncertainty. People are all exposed to numerous types of risk, and there are different levels of danger and uncertainty. Risk is the possibility that something harmful or unexpected may occur. However, risks and uncertainties may not always be averted in our daily lives, which has resulted in the financial industry developing a variety of services to shield people from them by using their own money as compensation. Insurance is introduced as an effective way to guard against financial loss.

One of the multiple forms of insurance is health insurance, which pays for an individual's medical costs. The sum of money that a customer continually pays to an insurance company monthly or yearly in return for this guarantee is called a premium. An individual who has purchased a health insurance policy receives coverage by paying a specific premium.

Several factors influence the pricing of health insurance. The insurance business must precisely know the price of the health insurance package to sell health insurance, but guessing is not an appropriate strategy. Insurance companies are now utilizing machine learning algorithms to make a predictive model for health insurance pricing as a better approach to this research problem. The algorithms used to predict the prices should be accurately chosen to ensure that their aim is within the threshold of profitable pricing and, therefore, the insurance company will be able to make the cost-effective decision.

In the past, insurance companies have relied on insurance consultants to mediate with potential customers to develop insurance premium packages. To expedite the process, machine learning can be utilized to predict the health insurance pricing of the insurance premiums packages with the help of historical customer data. In-depth research is necessary to fully comprehend how machine learning can contribute to health insurance pricing and why it should be. Multiple machine learning algorithms will be researched in this study for exposure. It is essential to comprehend their machine learning algorithm variations and how they will affect their accuracy in predicting health insurance pricing. By concentrating on the correct customer data, the accuracy

in predicting health insurance pricing helps insurance companies manage their cost-effective decisions. The ideal machine learning algorithm will be selected by evaluating its performance after testing each machine learning algorithm. Throughout this research, the efficacy of various machine learning algorithms will determine the most critical contributing algorithm.

This research aims to obtain a health insurance pricing dataset, analyze the dataset, preprocess data for machine learning algorithms, implement four machine learning algorithms, evaluate and compare the performance of machine learning algorithms used, and visualize a predictive model for health insurance pricing.

II. Material and Method

A. Overview of Machine Learning for Health Insurance Pricing

The primary factor influencing customer choice is price, as insurance companies frequently undercharge to broaden customer numbers and generate more business. The cornerstone of insurance companies is undoubtedly competitive pricing, as multiple insurance companies compete to offer the best prices and services. As a result, customers may compare the prices of several insurance providers, thanks to the rising price comparison websites in the insurance sector. Without a doubt, the customers will pick the most inexpensive plan. Accuracy in predicting the competitive price of an insurance plan is a vital part of attracting a customer. However, human error may occur during the insurance underwriting process, and for insurance companies, inaccurate pricing is a considerable risk since it can result in operational and bankruptcy risks. In a recent study of auto insurance premium leakage, missing or incorrect underwriting data has jeopardized insurers' rating strategies, costing them at least $29 billion annually [1]. Machine learning is used to achieve better results, avoid this issue, and make insurance pricing more accurate.

Machine learning has become more prevalent in commercial applications due to technological advancements and the reality of big data in various sectors. Health insurance firms have been using artificial intelligence and machine learning to enhance business operations and better serve customers. Artificial intelligence can accomplish multiple tasks at a much faster rate as it may collect data, process it, and present the user with the best outcome [2], [3], [4], [5], [6], [7], [8], [9]. In this case, it may be particularly suited to jobs that insurance consultants frequently carry out at a slower pace because, previously, generalized linear models were used to establish most insurance rates [10].

Various datasets have been used in related studies which heavily rely on machine learning algorithms to estimate healthcare expenses. Most studies have improved performance by including cost-related inputs such as prior total expenses and prescription prices. In contrast, some studies only depend on demographic and clinical information, such as diagnostic groups and medical tests, to create predictions [11], [12], [13]. Among all the machine learning algorithms used in healthcare price prediction, gradient boosting has consistently emerged as a top performer in accurately predicting healthcare costs, which is an ensemble learning approach that sequentially integrates weak regression tree models, using iterative optimization to minimize loss and use the minor absolute deviation.

Machine learning in health insurance pricing will undoubtedly help insurance companies as it may assist them in maximizing the use of the data, they have access to and improving their operations in various ways. Machine learning may assist insurance pricing by predicting the premiums resulting from machine learning algorithms trained on past data and applied to current data. Applying these algorithms may make evaluating risk, claims, and consumer behavior more transparent as they have more significant and quicker prediction accuracy than humans [14]. As a result, insurance companies may grow their business with more personalized plans for their customers and precisely identify potential new customers.

B. Review of Machine Learning for Health Insurance Pricing

Multiple machine learning techniques for resolving issues with insurance pricing prediction or healthcare have been published over the years. In several areas, machine learning has enhanced the diagnosis of illnesses and healthcare interventions, assisting in developing complex predictive analytics models to predict chronic disease [15]. It is also widely used by car insurance companies to determine the decision to renew the insurance based on data on business channel features, no-claim discounts, car age, and new car purchase prices [16]. These studies show that machine learning improves illness diagnosis and a wide range of areas, including the insurance sector.

In the insurance sector, machine learning algorithms can also be implemented using historical data from past customers to predict the premium prices for new customers. The authors of this research did a predictive analysis of the cost of medical insurance based on customer information, including gender, age, smoking status, body mass index (BMI), number of children, location, and premium costs [17]. This kind of research occurs because the calculations used to determine health insurance premiums are complicated, and actuaries that insurance companies often employ must consider the prices that are suitable to insurance companies as they must make profits by amassing more money than they need to cover their customer's medical expenses to stay in business [18]. In addition, this research is also conducted due to common mistakes made by humans. The old method of calculating health insurance costs is complex for insurance firms because human involvement in this process might occasionally result in incorrect or flawed outcomes [19].

Machine learning has several benefits for health insurance pricing, such as large-scale data analysis. Machine learning algorithms can analyze massive volumes of data and see patterns and trends humans would miss. As a result, it may be able to assess the risk profile of their policyholders and determine the insurance premium prices more accurately. Other than that, machine learning algorithms may be trained on past data to forecast a customer risk profile or likelihood of filing a claim, and this information may be utilized to create customized insurance prices.

In this age of modern technology, data has paved the way for machine learning to become a strategic ally for calculating health insurance pricing. Previously, actuaries employed

statistical techniques to evaluate risk and pricing. Before the 1980s, they used linear regression, but this has changed with the development of the generalized linear model (GLM).

However, because GLM is neither real-time nor dynamic, it is less accurate today. This is the critical factor driving machine learning's rise in popularity because machine learning algorithms can assist actuaries in moving toward dynamic and data-driven pricing from data preparation to real-time pricing. The use of machine learning algorithms generates precise and fair analyses of the cost of health insurance packages while proving how actuarial models can be outperformed by machine learning models [20], [21].

Machine learning can be used in health insurance pricing, such as risk assessment. Based on variables like age, gender, pre-existing conditions, and medical history, machine learning algorithms may be trained on historical data to estimate a policyholder's risk profile. This information may be utilized to create more personalized insurance premium prices. Other than that, it can also be helpful in fraud detection, where patterns and abnormalities in health insurance claim data pointing to fraud can be found using machine learning algorithms. This can aid in reducing costs and enhancing the accuracy of risk assessment by assisting insurers in identifying and preventing false claims. Overall, using machine learning in health insurance pricing can assist insurers in personalizing their goods and services, better understanding of managing risk, and increasing productivity, which may increase client satisfaction.

With machine learning, a form of artificial intelligence, insurance companies may anticipate outcomes more accurately without being explicitly instructed. Regression algorithms are used to predict health insurance premiums using machine learning. Regression belongs to one of the subcategories of machine learning and artificial intelligence, which is supervised learning. Regression helps determine how independent traits or variables relate to a dependent feature. It is also a machine learning predictive modeling technique that is frequently used to forecast continuous outcomes. In some cases, classification and ensemble methods are also utilized by researchers to predict insurance premiums.

In this study, it is expected that this research will comprehend each proposed model's idea and operation. Past research did not examine multiple high-performance machine learning algorithms in a study. This study will use four high-performance machine learning algorithms as a comparative analysis. Additionally, four performance score evaluations will be found to gauge their effectiveness. Their final scores will be compared to identify the best machine learning algorithm for health insurance pricing from the US Health Insurance dataset. The same dataset will be used to test each strategy to ensure the research is not biased toward machine learning algorithms. Data visualizations are also implemented to showcase the importance of features and the differences between actual and predicted prices for all the data points.

C. *Machine Learning Algorithms*

In this research problem domain, machine learning is often utilized to solve the problem of forecasting insurance premiums. Machine learning algorithms discover the data's underlying patterns in a way controlled by a particular set of hyperparameters. Trial and error are used to find the ideal set of hyperparameters that results in the model that provides the more accurate prediction. Multiple machine learning algorithms may be used to examine and predict outcomes based on input data in the context of regression algorithms.

1) *Regression Algorithms:*

Regression is the most widely used machine learning algorithm in predictive modeling due to its usefulness in forecasting, time series modeling, and determining the causal connection between variables. Regression is a statistical technique used in machine learning to forecast a continuous outcome variable (the dependent variable) by one or more predictor variables (the independent variables).

To create a regression model, a machine learning algorithm is trained on a dataset containing the predictor and outcome variables. The dataset is often used to create a training set and a test set, with the training set used to develop the model and the test set used to assess its effectiveness. The method calculates the coefficients of the predictor variables to create a model that can generate predictions based on an equation.

For example, a regression model might forecast a house's price based on size and location. [22] conducted a house price prediction using a machine learning model that utilized locational, structural, and neighborhood attributes. Then, the dataset on the size, location, and cost of houses would be used to train the model to create a model that can forecast the price of a property based on its estimation of the coefficients of the size and location variables.

Regression methods come in various forms, including multiple linear regression, ridge regression, and more. The exact problem being addressed, and the data type being used will determine the algorithm to use. Despite the different types of regression in supervised learning, the researchers' steps to conduct predictive modeling are usually similar. Figure 1 shows the usual steps of conducting predictive models.

Fig. 1 Steps for Predictive Modeling

The regression method is the most commonly employed in predicting insurance premium prices. Ridge and most minor absolute shrinkage and selection operators (LASSO) are used to fit their machine learning model for lapse prediction in life insurance contracts [23]. Additionally, in a recent study of health insurance cost prediction using regression models, multiple regression models, such as ridge regression, LASSO regression, linear regression, multiple linear regression, and polynomial regression, are used to create a method for predicting insurance cost and pricing in real time [10]. The

study explains why machine learning algorithms, or particularly regression, are being utilized in health insurance pricing, and it will help insurance businesses in the market by making quick and straightforward premium value determinations. Moreover, supervised learning in machine learning is also more accurate than generalized linear models [18].

A study predicted the cost of health insurance premiums based on an individual's age, sex, BMI, number of children, smoking habits, location of residence, and individual insurance premiums billed by health insurance by using regression algorithms, which are multiple linear regression (MLR), LASSO, and random forest regression [18]. The scores of this study demonstrated that random forests had the highest level of effectiveness compared to the others while suggesting insurance companies develop the algorithm.

Predicting the insurance premium charge may help governments forecast the price to help them decide on health-related issues [24], [25]. The study compared the performance of four regression models: multiple linear regression, decision tree regression, support vector regression, and random forest regression. The study's findings showed that random forests outperform the other three models by their evaluation metrics. Additionally, it was discovered that age and BMI are features that decide the dependent variable by using multiple linear regression with backward elimination.

A study by [26] used a diverse method of regression algorithms. This paper offers a computational intelligence technique for forecasting healthcare insurance costs. The suggested study methodology uses random forest regressor, multiple linear regression, ridge regression, and linear regression. This study aims to compare the forecasting accuracy of several regression models and demonstrate how they might estimate insurance costs. However, the results of this study indicate that an ensemble method model performs better than the other algorithms with an accuracy of 86%. Regression is a powerful technique for analyzing variables' relationships and predicting continuous outcome variables. It is frequently employed in various fields, including economics, finance, and health care.

2) Ensemble Algorithms:

Other than straightforward regression algorithms used in predicting insurance premiums, ensemble algorithms are also occasionally employed. Machine learning approaches known as ensemble algorithms combine the predictions of several models to get more accurate predictions. Ensemble approaches come in various forms, such as boosting, bagging, and stacking. The accuracy of forecasts may be increased, and the variance of the model can be decreased by using ensemble approaches. By combining the predictions of many models, ensemble techniques may be utilized to increase the accuracy of forecasts in the context of forecasting insurance premiums.

Boosting algorithms successively train several weak models, combining their predictions to get a final prediction. Overall, a more robust model results from training each weak model to fix the error produced by the prior model. Bagging algorithms simultaneously train several different models and then integrate the results to provide a single forecast, which can lower the model's variance and increase the predictability of results. The last ensemble algorithm is the stacking algorithm. Multiple base models are trained using stacking techniques, and their predictions are then used as features in a model that provides the final prediction. As a result, the model may capture complicated relationships between the basic models.

In a study by [27], the authors focused on ensemble algorithms and developed three new algorithms for managing medical insurance costs using supervised learning models. The authors built boosting models based on stochastic gradient descent and regression trees. The algorithms bagged classification, regression tree (CART), and random forest were also proposed. Accuracy was higher with the boosting and stacking ensembles than bagging, k-nearest neighbor, support vector machine, regression tree, linear regression, and stochastic gradient boosting used to construct the stacking. The random forest method is employed to merge the forecasts. The examination of the cited studies demonstrated that ensembles are more successful than a single machine-learning algorithm.

Leverage machine learning algorithms are used to predict medical costs, aiming to guide customers towards more affordable healthcare options [28]. The authors utilized two ensemble algorithms, random forest, and gradient boosting. The study revealed that the gradient boosting produced the highest accuracy in predicting the price of medical costs. Ensemble algorithms can help estimate insurance premiums because they may increase the accuracy of the forecasts by combining the results of many models. They can also make the model more stable and lower the variance of the predictions. However, they can be more challenging to build and require more processing power to train and test. In general, ensemble approaches can help anticipate insurance rates and increase the precision of such projections.

D. Theoretical Framework

1) Machine Learning Algorithm – Multiple Linear Regression:

When attempting to determine the relationship between variables, the term regression is used. That relationship is employed in machine learning to forecast how future events will turn out. Linear regression is one of the regression techniques that has been widely used for predicting the value of a variable based on the value of another variable. A regression model known as linear regression uses a straight line to estimate the relationship between a dependent variable and one independent variable. The simple linear model in the statistical world can be a single variable linear regression, as in Eq. 1.

$$Y = \beta 0 + \beta 1 + \varepsilon \quad (1)$$

There will always be some disparity between the values predicted by linear regression models and the actual values. That is when an error term is included that considers the discrepancy and aids in prediction. However, multiple predictors are frequently used to predict the outcome due to the dataset. For instance, using the dataset in this study, knowing whether the dependent variable and the six independent variables are connected in any way is a crucial step. Hence, the multiple linear regression model is a better algorithm for this scenario than simple linear regression. Like simple linear regression, multiple linear regression is a statistical method that assesses the strength of the connection between several independent variables and a dependent

variable using a straight line, much like simple linear regression. In a simple linear regression, there is only one independent variable and one dependent variable. In contrast, in a multiple linear regression, there are many predictor variables, and the value or outcome of the dependent variable is now calculated based on the values of the predictor variables. The equation is now modified from simple linear regression to Eq. 2.

$$Y = \beta 0 + \beta 1x11 + \beta 2x2 + \ldots + \beta mxm + \varepsilon \qquad (2)$$

In this study, Y is the insurance premium price predicted using several predictors or independent variables, which is x in the equation. It offers a straightforward method for predicting dependent variables by determining the Y corresponding to an x collection. In multiple linear regression, some of the independent variables may be correlated, so checking these before developing the regression model is essential. This is because of the foundation of multiple linear regression, which is the presumption that the connection between the dependent and independent variables is linear. Additionally, it is assumed that the independent variables have little to no association with each other, so if two independent variables are highly correlated, then only one of them should be used in the regression model.

The benefit of multiple regression is that it presents the prediction from various decision-making perspectives. It enables more complex models where numerous variables could be in play for a given outcome. However, any drawbacks of employing a multiple regression model are often related to the data employed using inadequate data and incorrectly assuming a correlation to be a cause of misleading results.

2) Machine Learning Algorithm: Ridge Regression:

Another approach for predicting health insurance pricing using machine learning is ridge regression. Ridge regression is one of the subcategories of regression algorithms in machine learning for analyzing multicollinear multiple regression data. Multicollinearity is a situation where more than two independent variables have strong correlations. Statistical conclusions will be less accurate if independent variables are multicollinear. When multicollinear independent variables are present in linear regression models, a ridge regression estimator was created as a potential substitute for the imprecision of the algorithm. The fundamental goal of ridge regression is to take the dataset and fit a new line into it without overfitting the model.

Although the variations in linear regression estimates are relatively wide and may be fairly distant from the real value, they tend to be unbiased. Ridge regression lowers the standard errors by adding some bias to the regression estimates. In essence, it seeks to obtain more accurate estimations. The ridge regression estimator is as in Eq. 3.

$$\beta = (X^{T}X + \lambda I)^{-1}X^{T}Y \qquad (3)$$

Linear regression is similar to ridge regression, but the penalty term ensures that the coefficients are not too significant. The penalty term is lambda, which is a positive value. By increasing the value of lambda, the penalty term becomes more robust and less likely to overfit. This might be helpful whenever there is a lot of noise in the data since it keeps the model from being overly sensitive to specific data points.

The advantage of ridge regression is that it prevents the model from overfitting and does not need unbiased estimators. When there is a considerable amount of multivariate data and more predictor variables than outcome variables, ridge regression performs better. When there is multicollinearity, the ridge estimator seems to be quite helpful in enhancing the least-squares estimate as the adverse effects of correlated features on regression models are often severe. Still, these effects are significantly reduced when the penalty term is applied.

Ridge regression has the drawback of being computationally demanding, as it requires more data to produce accurate results. Since it can be applied even when the data contains outliers, it can result in unpredictable outcomes, which makes it challenging to interpret the model's findings.

3) Machine Learning Algorithm: XGBoost Regression:

XGBoost is one of the new, well-known, and practical implementations of the gradient-boosted trees algorithm, which is based on function approximation by optimizing certain loss functions and using a variety of regularization approaches. XGBoost is a supervised learning technique that is very beneficial for predicting problems with data sets and missing values. It has been developed and known to be very effective, fast, adaptable, and portable in many cases. The main goal of the development of XGBoost was to enhance the performance and computational speed of machine learning models. Figure 2 shows an example of XGBoost implementation.

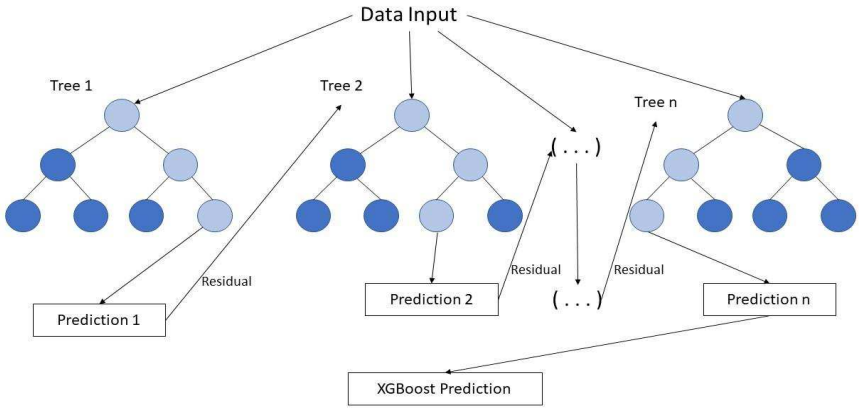


Fig. 2 XGBoost Implementation

XGBoost is a boosting method that combines classifiers with weights adaptively altered at each step to give more weight to the examples incorrectly categorized or known as weak learners in the earlier phase. Boosting uses weighted copies of the learning sample as the basis for classifiers. Each independent variable is given a significant weight in XGBoost before being put into the decision tree that predicts outcomes. Before being put into the following decision tree, previously mispredicted variables are given additional weight. Then, a robust and precise model is created by combining these various predictors. To make a high level of accurate predictors, XGBoost adds predictors sequentially to the model and corrects earlier predictors.

Similar to random forest, XGBoost is capable of handling massive data sets. Tree algorithms like XGBoost and random forest can be used with nonlinear or clustered data without normalizing features. The main difference between random forests and XGBoost is that random forests are built in parallel while XGBoost is built sequentially. In comparison to XGBoost, random forests are more straightforward to tune. Some disadvantages of this method include that this XGBoost algorithm may overfit the data when dealing with noisy data.

4) Machine Learning Algorithm: Random Forest Regression:

Random forest regression is an ensemble machine learning method commonly used for regression and other training-required tasks. It builds multiple decision trees to increase prediction effectiveness further. Bagged decision trees in random forest regression produce many trees. This means a random forest comprises several trees built in a specific random manner. In a random forest, the trees run in parallel. Therefore, they do not interact as they run [29]. Figure 2 shows an example of random forest prediction implementation.

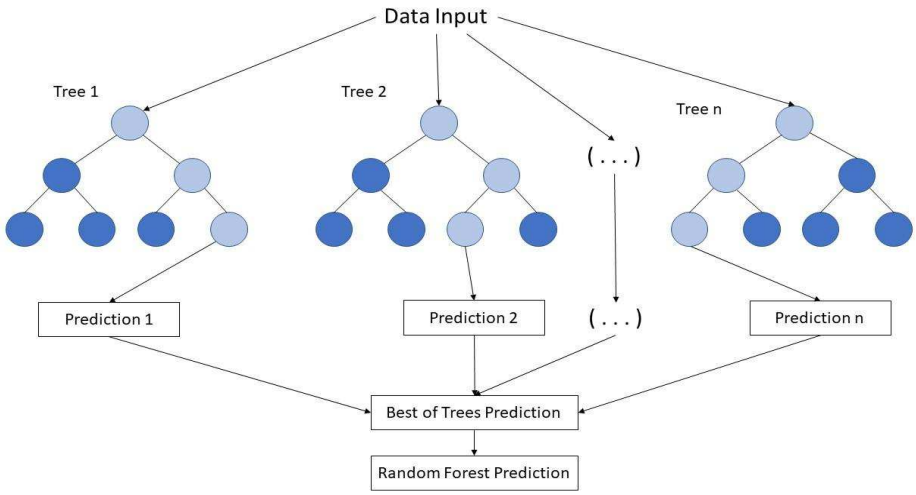


Fig. 3 Random Forest Implementation

In the case of regression, they begin at the tree's root and proceed in splits depending on possible outcomes until they reach a leaf node, at which point the conclusion is revealed. As in Figure 2, a separate tree sample of rows is used to build each tree, and a different tree sample of characteristics is chosen for splitting at each node. The next step is that each tree will make a unique prediction on its own, and then a single outcome will be produced by averaging these predictions, depending on how many trees are created. The best of the projections made by the forest's trees represent a prediction from random forest regression. As a result, the averaging improves the random forest's accuracy and overfitting over a single decision tree. The sum of all models for the random forest regressors can also be represented in the following Eq. 4.

$$g(x) = f_0(x) + f_1(x) + f_2(x) + f_3(x) + \ldots + f_n(x) \quad (4)$$

Regression requires Predictions over the individual decision tree's mean or average. Random forest regression compensates for the inclination of decision trees to overfit their training set. Random forest typically outperforms single-decision trees. However, the data quality might also impact how well they function [30]. A random forest model is a robust and precise regression algorithm. It often delivers excellent results on various issues involving non-linear connections. In cases where random forest regression delivers a poor score result, it may be due to inadequate data wrangling and the utilization of unnecessary variables. In addition, it may not escape from some disadvantages, such as the potential for overfitting and the need to select how many trees to include in the model on our own.

5) Evaluation Metrics:

Model evaluation is a significant element of creating a robust and accurate machine learning model. Machine learning algorithms use evaluation metrics to evaluate a model's performance and compare other models' performance. It may assess how effectively the model can predict the dependent variable from the independent variables by providing a numerical assessment of its accuracy and goodness of fit. For regression tasks, the four most popular metrics are mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared (R2) score. MAE measures the average absolute difference between predicted and actual values, while MSE measures the average squared difference. RMSE is the square root of MSE, which is easier to grasp and comprehend than the MSE since it is given in the same units as the original data. R-squared

measures how well the model fits the data in a regression problem. It indicates how much of the variance in the dependent variable is explained by the independent variables in the model. As for MAE, MSE, and RMSE, a lower value indicates a lower error in the algorithms, while in R-squared, higher numbers suggest a better fit of the model to the data and can vary from 0 to 1.

E. Methodology

In this research, an implementation will be made for development. It is used to show the functionality of the finished system, which is a crucial phase in this research process. In this study, the implementation will be expected to load the dataset first before performing exploratory data analysis (EDA) then by data pre-processing, and lastly by the algorithm implementations. Four high-performing machine learning algorithms will be implemented: multiple linear regression, ridge regression, XGBoost regression, and random forest regression.

1) Dataset:

The implementation will start with loading the dataset. The dataset used for model training in the implementation is the US Health Insurance dataset. It is widely used in machine learning related to research and work. The dataset contains information on the health insurance coverage of customers, including their age, sex, body mass index (BMI), number of children, smoking status, residential area, and charges of each customer.

The dataset obtained is in a comma-separated values (CSV) format, which requires a Python library, Pandas, to load it. The Pandas library is a well-known library for handling data in the Python programming language, as it offers a variety of features and resources for reading, modifying, and analyzing data. Thus, the read CSV function from Pandas is used to read data into a Pandas data frame from a CSV file, as it is easier to store and work with data in data frames, which are two-dimensional data structures similar to tables.

2) Exploratory Data Analysis:

Upon successfully loading the dataset, gaining insight from the data by performing EDA is important, as it allows one to comprehend the data better and spot any problems or potential areas of interest that need more analysis. Similar to loading the dataset, the Pandas library would also be mainly utilized in EDA due to its usefulness in performing essential steps.

In this section, the distribution of premium prices will be visualized first. This dataset clearly shows that the distribution of premium prices is right-skewed, as in Figure 4. This shows that premium prices are more frequent at the lower end of the pricing range and less frequent at the upper back, as customers take them.

The next step would be investigating the relationship between a dependent variable and its independent variables to comprehend their relationship. In this case, the dependent variable is 'charges,', and the independent variables are ‘age,’ ‘sex,’ ‘body mass index,’ ‘children,’ ‘smoker’, and ‘region.’ It is essential as it may help recognize the data's underlying relationships and identify significant variables that may affect the result. In Figure 5, a higher age would have higher charges. This indicates the greater risk of filing a claim or having more risk in life is frequently linked with older customers, which might be contributing reasons.

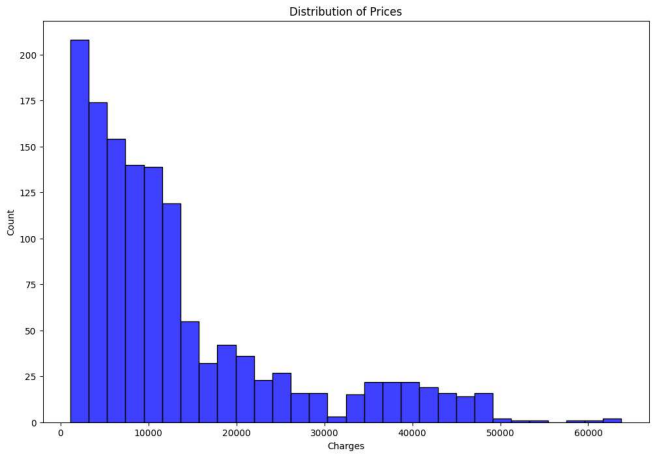


Fig. 4 Distribution of Premium Prices

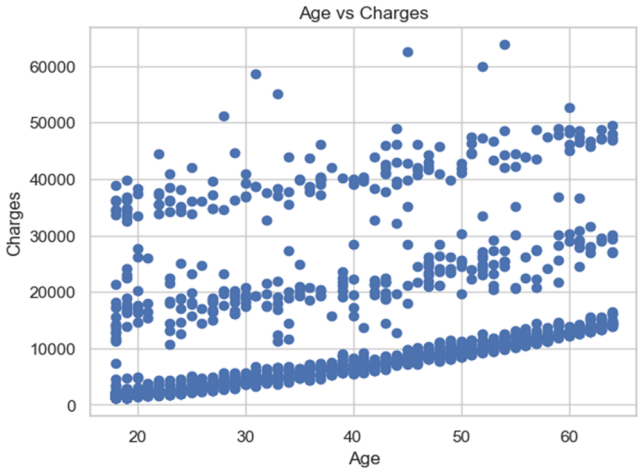


Fig. 5 Scatter plot of Age vs Charges

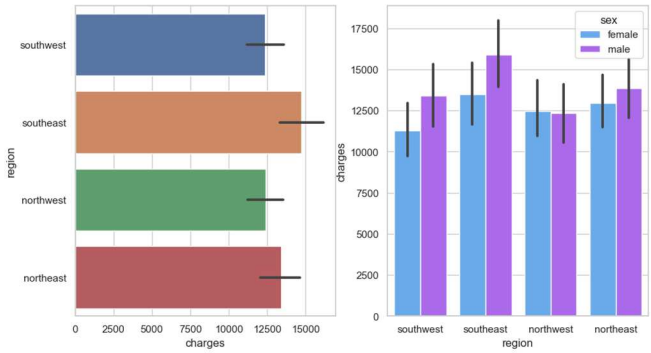


Fig. 6 Bar Plot of Premium Prices against Region and Sex

Then, the following predictor variables would be region and gender. Figure 6 on the left shows that the southeast has the highest prices while the southwest and northwest have the lowest. This could be due to some possible factors: the cost of healthcare in the area is more significant to compensate for those expenditures, a region with a higher risk profile, or a different region with different insurance regulations. The following plot at the right in Figure 6 shows that in every area, male customers tend to pay higher premium prices except for the northwest region, which is almost the same as the female charges. It can be said that males can be subject to higher premium prices. Due to their perceived increased risk of filing

a claim, males may be subject to higher rates. They may result from a greater accident rate, or a higher percentage of dangerous activities compared to females.

Next, it is essential to look at the smoking status of a customer. In Figure 7, it is demonstrated that a smoker has a way higher insurance premium price than a non-smoker on average. Smokers may pay higher insurance rates since they are thought to be more likely to file a claim. This might be a result of the increased risk of smoking-related illnesses such as lung cancer and stroke. Insurance firms may increase their rates to offset the higher risk of filing a claim and the medical expenditures.

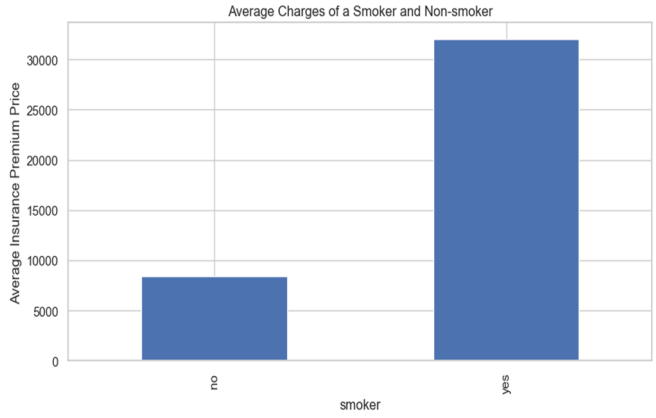


Fig. 7 Bar Plot of Average Premium Prices of a Smoker and Non-smoker

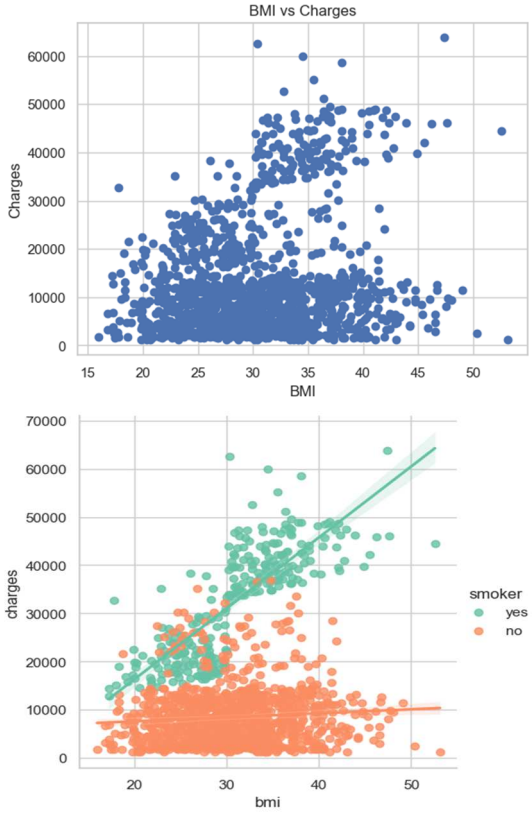


Fig. 8 Scatter plot of BMI vs Charges

On the upper side of Figure 8, it is unclear that a higher BMI would have higher charges. No particular trend can be established between BMI and the premium price to a customer. However, on the lower side of Figure 8, the charges of the premium insurance paid by a customer who smokes significantly grow for people with a higher BMI. This may be because a higher BMI is frequently seen as a risk factor in addition to smoking and is associated with a higher chance of developing specific health issues, including diabetes, heart disease, and stroke, which result in higher insurance premium prices. Because of their BMI, customers at a higher risk of filing a claim may be subject to increased rates from insurance providers.

The last independent variable would be the number of children. As seen in Figure 9, there is no significant trend in the average charges by the number of children. This may be because customers with more children, for instance, may occasionally pay lower prices since it is thought that there are fewer things like fewer accidents or dangerous conduct. However, people with more children may pay more significant premiums under other circumstances because they require more insurance coverage. For instance, a family with additional children may need more excellent health insurance protection because they may incur more medical costs.

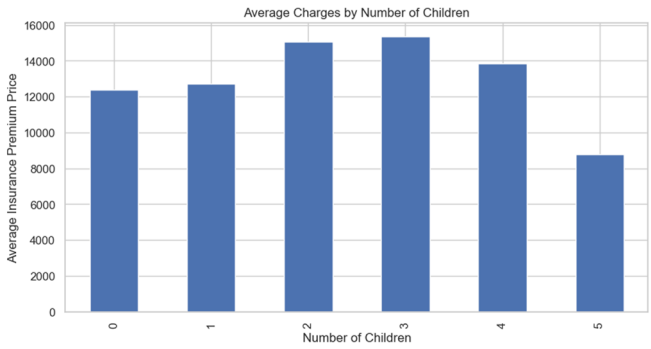


Fig. 9 Bar Plot of Average Premium Prices of a Smoker and Non-smoker

Figure 9 above shows that the customer with the highest number of children has the lowest premium price. This could be supported by Figure 10 below, which shows that customers with more children smoke less.

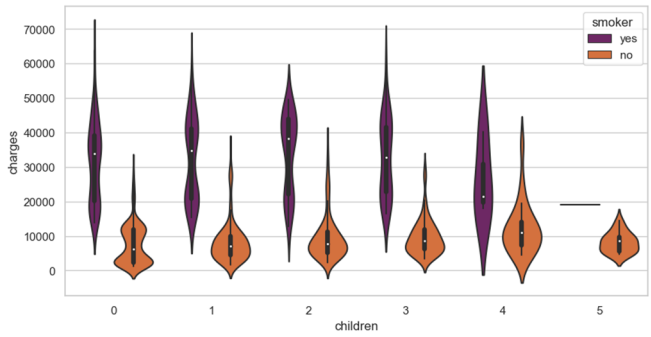


Fig. 10 Violin plot of Charges by Number of Children and Smokers

With the information gained above, the last step would be to check the correlation between each variable. Figure 11 shows the correlation between each variable. There is no significant correlation between variables except for smoking and the charges. It is expected to be highly correlated with the findings and plot.

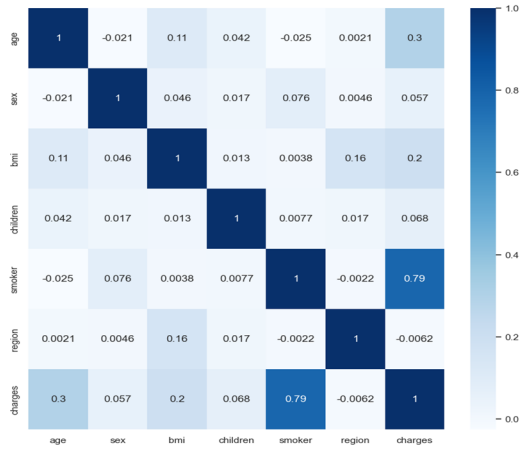


Fig. 11 Heatmap of the variables

3) Data Pre-processing:

Upon getting insight into the data during EDA, is it expected to process the data before executing the simulation to make it more suitable for analysis and modeling? In this dataset, there are no missing values, which means there is no need to replace the missing values or drop the missing rows or columns.

However, in the dataset, three columns are categorical variables: "sex," "smoker," and "region." Categorical variables must be encoded to be utilized in machine learning algorithms and made readable by machines. Thus, LabelEncoder converts the category data type to numerical data type for these three columns.

4) Regression Implementations:

After data pre-processing, the data is suitable for machine learning algorithms. Multiple linear regression, ridge regression, XGBoost, and random forest were implemented to address this problem. To complete this phase, it is necessary to import sklearn, pandas, matplotlib, and other Python modules that were needed during the implementations. Before fitting the data into the regression algorithms, the dataset will be split using a train-test split with a 30 percent test size.

After the train-test split phase, all four algorithms will then be implemented. The selected algorithm can then be further evaluated using evaluation metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared score. Among all the four algorithms, the best-performing algorithms are then fine-tuned by performing hyperparameter tuning using GridSearch cross-validation (GridSearchCV). It will find the optimal set of hyperparameters for a specific algorithm by searching through a given parameter grid. It uses a set of parameters to search for and gauge the model's effectiveness. It performs a cross-validation technique for each combination of parameters. GridSearchCV then returns the parameters that performed the best on the validation set. A new set of hyperparameters was built for the best-performing algorithm based on the outcomes of the hyperparameter tuning.

5) Data Visualization Implementations

After getting the results of each algorithm, data visualizations of the differences between actual and predicted will be implemented for all the algorithms. These visualizations clearly understand the data and help interpret the results. The pseudocode of how the algorithms do the visualizations is shown below.

Actual vs Predicted Data Visualization of Algorithms

---

1. algorithm_names = [Multiple Linear Regression, Ridge Regression, XGBoost Regression, Random Forest, Random Forest with GridSearchCV]
2. for each algorithm in algorithm_names:
3. Plot title of the algorithm
4. Plot y-axis range limits between -15,000 and 25,000
5. Plot y-axis label is 'Price Difference'
6. Plot x-axis label is 'Customer Number.'
7. Construct a scatter plot of differences between actual and predicted prices for the algorithm
8. Construct the horizontal line of y=0 for the scatter plot

III. Results and Discussion

All the machine learning algorithms implemented are evaluated using the evaluation metrics: MAE, MSE, RMSE, and R2 scores. The results of the algorithms are recorded in Table I. Table I records the findings of the algorithms with the default parameters for all algorithms and shows that the random forest method fared best compared to the other algorithms. It has the lowest MAE, MSE, RMSE and the highest R2 score among all the models. This suggests that the random forest model had the highest predictive power among all the algorithms utilized in this study. Notably, the XGBoost method outperformed the linear and ridge regression models with better MAE, MSE, and RMSE values and a higher R2 score. The linear and ridge models, on the other hand, had higher error values and lower R2 scores. Based on these results, it can be said that tree-based models, which are the ensemble methods, perform better than linear and ridge models in forecasting the cost of health insurance, with random forest being the most accurate of the models tested. The Random Forest algorithm of building an ensemble of independent decision trees can lead to better overall performance than XGBoost, which relies on sequential boosting and may incorrectly boost the wrong earlier predictors.

TABLE I
EVALUATION METRICS SCORE

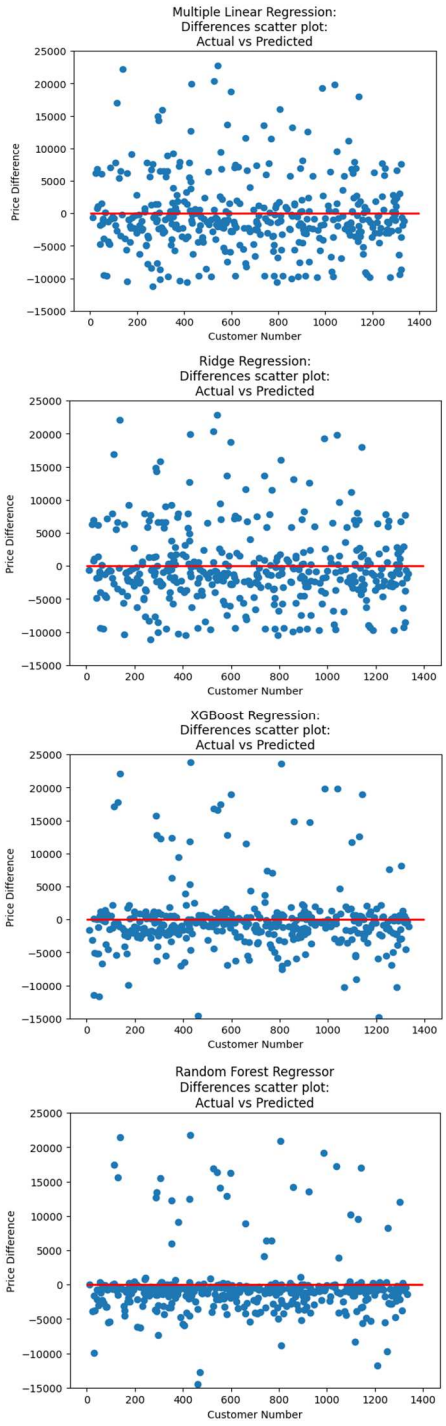
Model	MAE	MSE	RMSE	R2 Score
Linear	4155.24	33805466.9	5814.25	0.77
Ridge	4167.79	33839690.23	5817.19	0.77
XGBoost	2873.59	26515250.37	5149.29	0.82
Random Forest	**2572.27**	**21614066.78**	**4649.09**	**0.85**

As random forest performs the best in Table I, the algorithm is fine-tuned by performing hyperparameter tuning using GridSearchCV. The algorithm's final results after applying hyperparameter tuning are shown in Table II. The MAE of the two models is also relatively close, indicating similar performance in this metric. Except for the MAE, all three-evaluation metrics, MSE, RMSE, and R2 Score, have improved after applying GridSearchCV to random forest. This implies that the random forest with the GridSearchCV model outperforms the default parameters random forest model in predicting health insurance pricing.

TABLE II
EVALUATION METRICS SCORE

Model	MAE	MSE	RMSE	R2 Score
Linear	**2332.32**	21614066.78	4649.09	0.85
Random ForestGS	2617.76	**19698670.62**	**4438.32**	**0.86**

The visualization of the differences between the actual and predicted prices for all data points is shown in Figure 12. As visualized, both random forest plots are scattered closely to the zero horizontal line compared to the other algorithms, which indicates there are slight differences in the predicted data from the original data with random forest, with GridSearchCV implemented being the most of the expected data points are closest to 0.



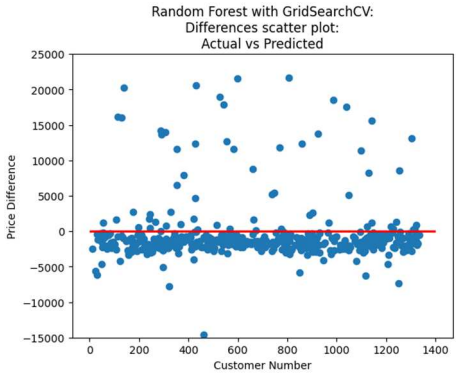


Fig. 12 Actual vs Predicted Data Visualization of Algorithms

IV. CONCLUSION

The price of health insurance is a challenging issue that calls for the application of modern technologies like machine learning. This study investigates the suitability of artificial intelligence methods based on machine learning for estimating health insurance prices. The high-performing machine learning algorithms for predicting healthcare insurance prices have had their performance outcomes compared, such as multiple linear regression, ridge regression, random forest regression, and XGBoost regression, to address this problem. The algorithms show excellent results in the evaluation metrics tested, and the random forest algorithm produced the best results. After applying the hyperparameters tuning on the random forest algorithm using the GridSearchCV, the algorithm showed three more improvements in the evaluation metrics assessed. Additionally, visualizations of the differences between the predicted and actual price for all data points for these models demonstrated their ability to predict health insurance pricing accurately. Healthcare providers and insurance companies can use these findings to make more informed decisions about insurance pricing and coverage.

Additional adjustments can be made to enhance the effectiveness of the algorithms used. In future work, this research can be extended by incorporating more advanced machine learning algorithms, such as deep learning methods like convolutional neural networks and recurrent neural networks, to increase the predictions for health insurance pricing. Another direction for future research is the application of ensemble approaches, which hybridize many algorithms that may be investigated to raise overall accuracy.

REFERENCES

[1] L. Zhou, Q. Chen, Z. Luo, H. Zhu, and C. Chen, "Speed-Based Location Tracking in Usage-Based Automotive Insurance," in *Proceedings - International Conference on Distributed Computing Systems*, Institute of Electrical and Electronics Engineers Inc., Jul. 2017, pp. 2252–2257. doi: 10.1109/ICDCS.2017.278.

[2] K. Kaushik, A. Bhardwaj, A. D. Dwivedi, and R. Singh, "Article Machine Learning-Based Regression Framework to Predict Health Insurance Premiums," *Int J Environ Res Public Health*, vol. 19, no. 13, Jul. 2022, doi: 10.3390/ijerph19137898.

[3] V. Kuleto *et al.*, "Exploring opportunities and challenges of artificial intelligence and machine learning in higher education institutions," *Sustainability (Switzerland)*, vol. 13, no. 18, Sep. 2021, doi:10.3390/su131810424.

[4] S. A. Kalogirou, "Artificial intelligence for the modeling and control of combustion processes: A review," *Progress in Energy and Combustion Science*, vol. 29, no. 6. pp. 515–566, 2003. doi:10.1016/S0360-1285(03)00058-3.

[5] R. H. Hariri, E. M. Fredericks, and K. M. Bowers, "Uncertainty in big data analytics: survey, opportunities, and challenges," *J Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0206-3.

[6] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Computer Science*, vol. 2, no. 3. Springer, May 01, 2021. doi: 10.1007/s42979-021-00592-x.

[7] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *Eurasip Journal on Advances in Signal Processing*, vol. 2016, no. 1. Springer International Publishing, Dec. 01, 2016. doi: 10.1186/s13634-016-0355-x.

[8] J. H. Thrall *et al.*, "Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success," *Journal of the American College of Radiology*, vol. 15, no. 3, pp. 504–508, Mar. 2018, doi: 10.1016/j.jacr.2017.12.026.

[9] Erik Brynjolfsson and Tom Mitchell, "What can machine learning do? Workforce implications".

[10] Embrechts P. Actuarial versus financial pricing of insurance. The Journal of Risk Finance. Mar 1;1(4):17-26, 2000.

[11] B. Panay, N. Baloian, J. Pino, S. Peñafiel, H. Sanson, and N. Bersano, "Predicting Health Care Costs Using Evidence Regression," MDPI AG, Nov. 2019, p. 74. doi: 10.3390/proceedings2019031074.

[12] M. Amin Morid, K. Kawamoto, T. Ault, J. Dorius, and S. Abdelrahman, "Utah Health Plans for," 2013.

[13] P. Yang, H. Qiu, L. Wang, and L. Zhou, "Early prediction of high-cost inpatients with ischemic heart disease using network analytics and machine learning," *Expert Syst Appl*, vol. 210, Dec. 2022, doi: 10.1016/j.eswa.2022.118541.

[14] M. Eling, D. Nuessle, and J. Staubli, "The impact of artificial intelligence along the insurance value chain and on the insurability of risks," *Geneva Papers on Risk and Insurance: Issues and Practice*, vol. 47, no. 2, pp. 205–241, Apr. 2022, doi: 10.1057/s41288-020-00201-7.

[15] T. Pfutzenreuter and E. de Lima, "Machine Learning in Healthcare Management for Medical Insurance Cost Prediction," 2022, pp. 1323–1334. doi: 10.37885/220207863.

[16] H. D. Wang, "Research on the features of car insurance data based on machine learning," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 582–587. doi: 10.1016/j.procs.2020.02.016.

[17] M. hanafy and O. M. A. Mahmoud, "Predict Health Insurance Cost by using Machine Learning and DNN Regression Models," *International Journal of Innovative Technology and Exploring Engineering*, vol. 10, no. 3, pp. 137–143, Jan. 2021, doi: 10.35940/ijitee.C8364.0110321.

[18] A. D. Kafuria, "Predictive Model for Computing Health Insurance Premium Rates Using Machine Learning Algorithms," *International Journal of Computer*, [Online]. Available: http://ijcjournal.org/

[19] S. Panda, B. Purkayastha, D. Das, M. Chakraborty, and S. K. Biswas, "Health Insurance Cost Prediction Using Regression Models," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, COM-IT-CON 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 168–173. doi: 10.1109/COM-IT-CON54601.2022.9850653.

[20] R. Kshirsagar *et al.*, "Accurate and Interpretable Machine Learning for Transparent Pricing of Health Insurance Plans," 2021. [Online]. Available: www.aaai.org

[21] S. Badillo *et al.*, "An Introduction to Machine Learning," *Clin Pharmacol Ther*, vol. 107, no. 4, pp. 871–885, Apr. 2020, doi:10.1002/cpt.1796.

[22] N. H. Zulkifley, S. A. Rahman, N. H. Ubaidullah, and I. Ibrahim, "House price prediction using a machine learning model: A survey of literature," *International Journal of Modern Education and Computer Science*, vol. 12, no. 6, pp. 46–54, 2020, doi:10.5815/ijmecs.2020.06.04.

[23] M. Azzone, E. Barucci, G. Giuffra Moncayo, and D. Marazzina, "A machine learning model for lapse prediction in life insurance contracts," *Expert Syst Appl*, vol. 191, Apr. 2022, doi:10.1016/j.eswa.2021.116261.

[24] A. Lakshmanarao, C. S. Koppireddy, and G. V. Kumar, "Prediction of medical costs using regression algorithms." [Online]. Available: www.joics.org

[25] N. K. Yego, J. Kasozi, and J. Nkurunziza, "A comparative analysis of machine learning models for the prediction of insurance uptake in kenya," *Data (Basel)*, vol. 6, no. 11, Nov. 2021, doi:10.3390/data6110116.

[26] C. A. ul Hassan, J. Iqbal, S. Hussain, H. AlSalman, M. A. A. Mosleh, and S. Sajid Ullah, "A Computational Intelligence Approach for Predicting Medical Insurance Cost," *Math Probl Eng*, vol. 2021, 2021, doi: 10.1155/2021/1162553.

[27] N. Shakhovska, N. Melnykova, V. Chopiyak, and M. Gregus Ml, "An ensemble methods for medical insurance costs prediction task," *Computers, Materials and Continua*, vol. 70, no. 2, pp. 3969–3984, 2022, doi: 10.32604/cmc.2022.019882.

[28] A. Kumar Sahu, G. Sharma, J. Kaushik, K. Agrawal, and D. Singh, "Health Insurance Cost Prediction by Using Machine Learning." [Online]. Available: https://ssrn.com/abstract=4366801

[29] R. Samala, H.-P. Chan, L. Hadjiiski, and S. Koneru, "Hazards of data leakage in machine learning: a study on classification of breast cancer using deep neural networks," Feb. 2020, p. 39. doi:10.1117/12.2549313.

[30] B. Janet, A. Ghosh, and J. A. Kumar R, "End-to-End Encryption and Prediction of Medical Insurance Cost," in *2022 6th International Conference on Trends in Electronics and Informatics, ICOEI 2022 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 846–850. doi: 10.1109/ICOEI53556.2022.9777238.