Analyzing Insurance Cost Estimation: A Supervised Regression Approach

Tonni Das Jui
Department of Computer Science
Baylor University
Waco, TX
tonni_jui1@baylor.edu

Pablo Rivas, *Senior, IEEE*
Department of Computer Science
Baylor University
Waco, TX
pablo_rivas@baylor.edu

***Abstract*—While health insurance systems have gained traction in both developed and developing nations, underdeveloped countries continue to strive for cost-effective healthcare solutions. Designing effective insurance predictive models is critical, relying on variables such as age and average monthly medical costs. Researchers are exploring supervised and unsupervised models to accommodate diverse data and prevent financial strain on all stakeholders due to high medical expenses. This study contributes to this effort by evaluating the efficiency of various supervised models in predicting insurance costs. We assess their model performance and provide insights for developing robust insurance systems.**

***Index Terms*—Insurance cost estimation, EDA, Linear Regression, Ridge regression, Neural Network, Support Vector regression, Random forest, Polynomial feature transformation, Hyperparameter Tuning, cross-validation.**

Type of submission: Short Research Paper

I. Introduction

Health insurance systems are crucial for ensuring equitable access to healthcare and avoiding bankruptcies, yet their design and implementation remain challenging, particularly in underdeveloped countries [1]. Despite significant advancements in developed and developing nations, underdeveloped regions continue to seek cost-effective healthcare solutions [2], [3]. A well-designed health insurance system necessitates accurate predictive models that assess risks and costs based on diverse variables [4]–[6].

Health data's increasing complexity and diversity have driven researchers to explore various regressive machine learning (ML) models to enhance the accuracy of insurance cost predictions [7], [8]. Supervised and unsupervised learning approaches are being investigated to develop robust models to mitigate the financial risks associated with high medical costs [9], [10]. These models aim to prevent the potential bankruptcy of insurance stakeholders by accurately predicting insurance costs and managing data diversity [11], [12].

In this study, we contribute to this ongoing research by evaluating the efficiency of various supervised models in predicting insurance costs. We utilize a dataset from Brett Lantz's book, Machine Learning with R, which is simulated based on demographic statistics from the United States [13]. By comparing the performance of different models, we aim to provide insights that can aid in developing more effective and sustainable health insurance systems, particularly in underdeveloped countries.

The remainder of the paper is organized into several sections. Section II reviews related work and background studies. Section III outlines the experimental methodology, including a discussion of the dataset, feature analysis, and the models employed. The experimental results are presented in Section IV, and the paper concludes with Section V.

II. Background Study

One of the world's most pressing issues is the escalating cost of healthcare. Consequently, many stakeholders have invested in forecasting health expenses to mitigate the risk of financial insolvency due to medical costs, where machine learning models have gained significant traction [14]–[16]. While machine learning encompasses numerous potential assumptions, its efficacy hinges on selecting an appropriately precise algorithm for the given problem domain and adhering to the correct model construction, training, and deployment procedures [17]. In this regard, Various supervised techniques were discovered to analyze past medical history as a predictor and forecast potential medical costs by adopting regression and classification strategies [1]. Among the regressive models, simple and multiple linear regression [18], polynomial regression [19], [20], ridge regression [21], [22], and lasso regression [23] are noteworthy. For example, Sushmita et al. have utilized a random forest and linear regressive model to predict a person's quarterly future medical costs from the past medical expense history [24]. Another model employed an extensive linear regression approach to estimate the intensive care unit (ICU) stay cost. This model utilized features such as patient profile data, diagnosis-related groups (DRGs), duration of hospital stay, and additional characteristics [14]. In addition to these regression models, Lahiri et al. implemented a classification algorithm to forecast whether an individual's medical expenses would increase in the following year by considering the medical expenses from the previous year [16].

However, the influence of neural network architecture has emerged as another research branch for predicting insurance cost, and it has become difficult to ensure a perfect model to achieve an ideal estimate for varieties in a country's financial condition [25]. Researchers examined patient-level health care

costs in both the short and long term and discovered a clear temporal link over various time spans [26]. The supervised learning methods for predicting healthcare costs using Artificial Neural Network (ANN) and the Ridge regression model using Empirical Evaluation produce approximations but struggle for time-series implementation [27]. Feature engineering is crucial in capturing temporal patterns in the data and reducing the number of features with minimal loss [28]. An analysis regarding the implementation of these various strategies can aid in the research of developing an insurance cost estimation system.

III. Methodology

This section discusses the data we utilized and the models we explored for predicting insurance costs.

A. Dataset

This dataset, derived from the repository of Lantz's book, was simulated using demographic statistics from the US Census Bureau, thus approximating real-world conditions [13]. The dataset includes the age of the primary beneficiary (excluding those over 64 years, as government programs typically cover them), the policy holder's gender (male or female), body mass index (BMI, calculated as weight in kilograms divided by height in meters squared), the number of children or dependents covered by the insurance plan, smoking status (yes or no), and the beneficiary's residence in the US. It encompasses a total of 1338 patients, capturing features related to patient demographics and total medical costs incurred by the insurance plan over the calendar year. Table I represents the summary of the dataset.

TABLE I: Description of the dataset features and feature values.

Features	Value type	Value description	Detailed info.
Age	An integer	A number indicating beneficiaries age	range = 18 to 64 mean = 39.2
Sex	A string	male or female	# female = 662 # male = 676
BMI	A float	weight (kg) divided by height (m) squared	min value = 15.96 mean value = 30.66 max value = 53.13
Children	An integer	# children covered by the insurance plan	Range = 0 - 5 years mean value = 1.059
Smoker	A string	yes or no	smokers = 1064 non-smokers = 274
Region	A string	four geographic regions (categorical)	southeast = 364 people northeast = 324 people southwest = 325 people northeast = 325 people

Among these features, sex, smoker, and region are categorical features, and their categories are detailed in the table I. Aside from these features, the dataset contains another attribute that represents medical charges as the marker to train the dataset, indicating suitability for a supervised model. Also, to mitigate a moderate noise level within our dataset, we calculated the interquartile range (IQR) and excluded data points falling outside this range. Outliers, defined as those data points exceeding the IQR (e.g., charges above $30,000$), were considered noise within the dataset. Table II illustrates the IQR differences for each feature.

TABLE II: Mean of the Inter Quartile range ($(Q_3 - Q_1)/2$) for all features of dataset.

Features	IQR	Features	IQR
Age	24.00	sex	1.00
BMI	8.40	children	2.00
smoker	0.00	region	1.00
charges	11899.62		

The IQR filter eliminated 283 outlier data, resulting in a final dataset size 1055. We followed standard segmentation for splitting training and testing data, with 15% and 5% allocated for testing and validation, and the remaining 80% for training. The developed models were applied to the testing dataset following model training on the training set.

B. Feature analysis

We also conduct further feature analysis to investigate the correlation between the number of children, age, and BMI as illustrated in Figure 1 of the exploratory data analysis. By examining their correlations with health insurance charges, we found that charges increase with age and the number of children but decrease with BMI. Based on these findings, we synthesized a new feature called the "stress level," which integrates the impacts of age, number of children, and BMI, to replace the original three features following equation (1).

$$data[stress\_level] = \frac{data[children] * data[age]}{data[BMI]} \quad (1)$$

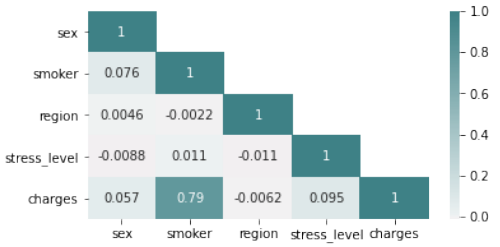


Fig. 1: Correlation between charges with other features including stress_level and excluding age, children, BMI

We also plotted the value distribution to observe the distribution of the targeted attribute, charges/costs. We noticed in Figure 2 that costs were most dense initially and gradually fell afterward. Figure 3 indicates that most peoples' insurance costs (target attribute) were between 5000 to $15,000$.

Lastly, as the 'smoker' attribute is highly correlated with the targeted attribute 'charges/costs', we analyze them further. Figure 4 illustrates the distribution of charges for smokers versus non-smokers, revealing that smokers face a significantly wider range of charges. The boxplot in Figure 5 displays

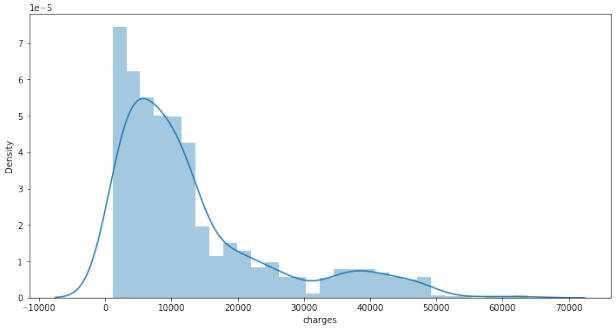


Fig. 2: Charges distribution per number of people.

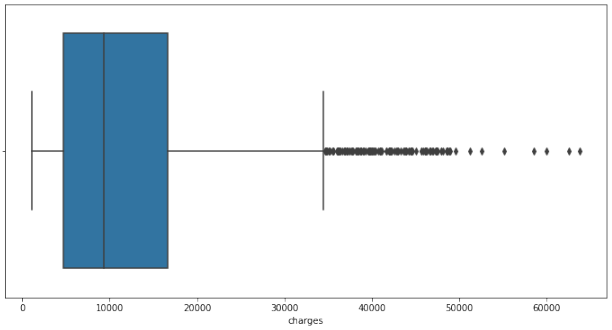


Fig. 3: Costly charge outliers.

charges for both groups, excluding the outliers based on the IQR, indicating our min-max normalization to scale these features in the range $(0, 1)$.

C. Models

We aim to predict health insurance charges using supervised learning methods [29]. The testing dataset includes output labels, enabling the comparison between predicted and actual values. We employ linear regression, ridge regression, SVM, random forest, and NN algorithms to enhance prediction accuracy, utilizing k-fold cross-validation for performance optimization. Linear regression models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data [30]. Ridge regression is also a type of linear regression that includes a regularization term to penalize large coefficients and avoid overfitting [22]. We explored support-vector machines (SVM) following linear kernel from linear and rbf kernel as linear SVMs (or logistic regression) [31]. Then, we experimented with another method, random forest, which is a method for classification, regression, and other tasks that works by constructing a large number of decision trees during training and then predicting the class that is the mode of the classes or the average prediction of the individual trees [32]. While increasing the trees, Random Forest adds more randomness to the model. When splitting a node, it looks for the best function among a random subset of features rather than the most appropriate feature. As a result, a wide range of variety leads to a stronger model in general. We also upgraded the random forest model with polynomial feature transformation, bias-variance analysis, hyperparameter tuning, and k-fold cross-validation and conducted our experiment for this upgraded model.

Our last experimental model is the most influential one. We implemented a neural network model to predict the targeted categorical attribute in our datasets. Neural network approaches are better for fault tolerance and predictive analysis as the hidden layers of neural networks are used to improve prediction accuracy. We implemented our data set on a neural network and got the result with five layers where in four of them, we used the RELU activation function, and in one, we used linear activation. We used six input parameters. There were 128 neurons in the first hidden layer, 256 neurons in the second, third, and fourth hidden layers, and 1 in the last layer. We mention all our results for all these models in the section IV-B.

IV. Experiment

In this section, we discuss our evaluation function and experimental results.

A. Cost function

We utilized root mean square error (RMSE) as the cost function over mean absolute error (MAE). The equation 2 overviews our cost function.

$$RMSE = \sqrt{\sum_{i=1}^{n} (y_i - y_i^P)^2 / n} \tag{2}$$

The RMSE is a quadratic scoring rule that calculates the average magnitude of the error. Since errors are squared before being averaged, the RMSE gives large errors with a relatively high weight, and RMSE appears to be greater than MAE. As a result, the RMSE is most useful when significant errors are undesirable. When dealing with large error values, RMSE performs better in representing efficiency.

B. Experimental results

Table III presents the summarized results, including the RMSE scores for each model. The table highlights that the tuned neural network model achieves the lowest error among all models. Aside from the neural network, the ridge regression, random forest, and upgraded random forest models display lower testing errors, with the latter two models showing nearly identical results. Notably, the upgraded random forest model exhibits the lowest training error among all the models despite its testing error being comparable to that of the ridge and un-tuned random forest models. Based on the higher testing error of the upgraded random forest model, it can be inferred that this model is likely overfitted. Lastly, the support vector machine (SVM) and linear regression models demonstrate the poorest performance, yielding the highest training and testing errors. From a different perspective, while the linear regression and SVM models do not achieve the expected performance, they do not exhibit signs of overfitting,

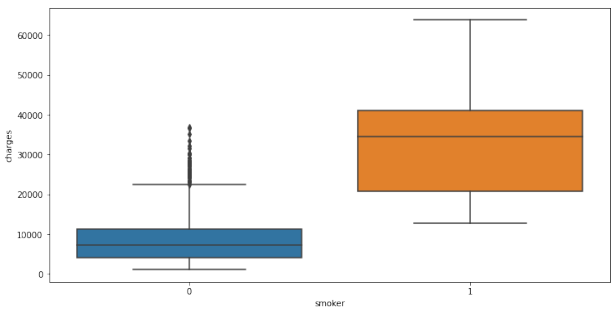


Fig. 4: Charges distribution for smoker and non-smoker.

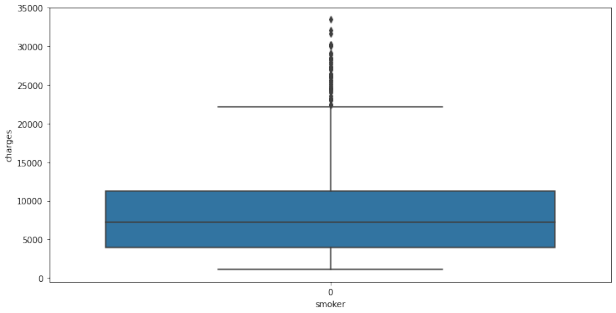


Fig. 5: Charges distribution for smoker and non-smoker excluding the outliers based on the IQR.

as their training and testing errors are relatively similar. Likewise, the ridge regression, random forest, and neural network models display only slight discrepancies between their training and testing errors, further indicating an absence of significant overfitting in these models.

TABLE III: Training and testing error for experimental regression models in insurance cost prediction Lantz's dataset [13].

Approach	**Training sample error**	**Testing sample error**
Linear	15645	15746
Ridge	6128	5559
SVM	12898	12795
Random Forest	6128	5558
Upgraded Random Forest	1860	4674
Neural Network	2739	3098

We also plot the testing sample performance in the figures 6 to visualize our experimental models' performances. Figure 6 represents the plot for illustrating how similar insurance costs (output) are compared to predicted insurance costs or how prediction and actual output differ.

For the linear and ridge regression models, Figures 6a, 6b, and 6c illustrate the discrepancy between the predicted values (green) and the actual values (blue), as the two do not align closely. In contrast, Figure 6f demonstrates that the predicted insurance costs (y) nearly coincide with the actual insurance costs for the improved random forest model. Furthermore, Figure 6d highlights the closest alignment between the predicted and actual insurance costs for the tuned neural network model. An important observation in Figures 6d and 6f is the close alignment between the actual and predicted insurance costs. This indicates that the tuned neural network and the tuned random forest models exhibit lower prediction errors than other approaches.

V. Conclusion

In conclusion, this study explored various supervised regression models to predict insurance costs based on a dataset reflecting real-world demographic statistics. The experimental results demonstrate that while traditional models like linear regression and SVM did not yield high prediction accuracy, they maintained consistency between training and testing errors, indicating minimal overfitting. More advanced models, including ridge regression, random forest, and especially the tuned neural network, significantly improved prediction accuracy. The tuned neural network model exhibited the best performance, closely aligning predicted and actual values, particularly in comparison to other models. The results also highlight the importance of hyperparameter tuning and feature engineering, as seen in the improved random forest model, which delivered substantial gains in training accuracy, albeit with potential signs of overfitting. These findings underscore the efficacy of advanced regression techniques in insurance cost estimation and provide a foundation for future studies to build on, especially by refining feature synthesis and model optimization strategies.

References

[1] D. U. Himmelstein, D. Thorne, E. Warren, and S. Woolhandler, "Medical Bankruptcy in the United States, 2007: Results of a National Study," *The American Journal of Medicine*, vol. 122, no. 8, pp. 741–746, Aug. 2009. [Online]. Available: https://www.amjmed.com/article/S0002-9343(09)00404-5/abstract

[2] K. Bhatia, S. S. Gill, N. Kamboj, M. Kumar, and R. K. Bhatia, "Health insurance cost prediction using machine learning," in *2022 3rd International Conference for Emerging Technology (INCET)*, 2022, pp. 1–5.

[3] OECD, *Health at a Glance 2023*, 2023. [Online]. Available: https://www.oecd-ilibrary.org/content/publication/7a7afb35-en

[4] W. Yip, H. Fu, A. T. Chen, T. Zhai, W. Jian, R. Xu, J. Pan, M. Hu, Z. Zhou, Q. Chen *et al.*, "10 years of health-care reform in china: progress and gaps in universal health coverage," *The Lancet*, vol. 394, no. 10204, pp. 1192–1204, 2019.

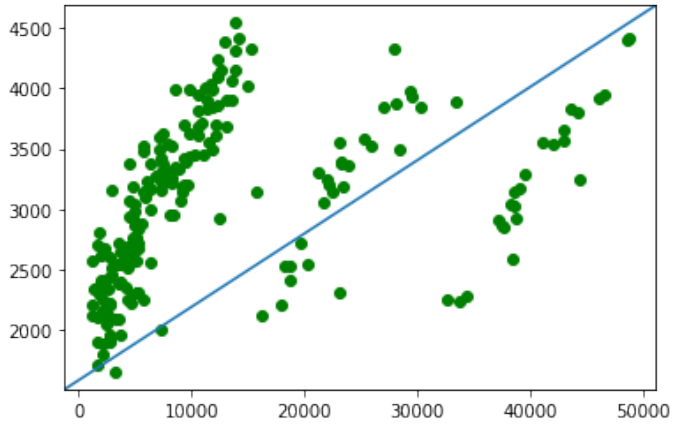
[5] E. Siegel, *Predictive analytics: The power to predict who will click, buy, lie, or die*. John Wiley & Sons, 2013.

[6] A. Taha, B. Cosgrave, W. Rashwan, and S. McKeever, "Insurance reserve prediction: Opportunities and challenges," in *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2021, pp. 290–295.

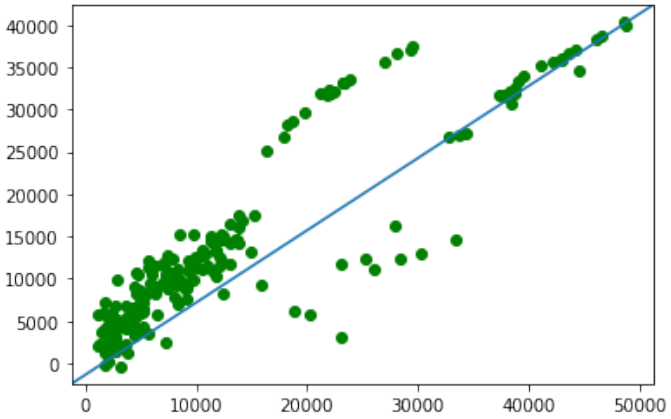
[7] J. H. Chen and S. M. Asch, "Machine learning and prediction in medicine—beyond the peak of inflated expectations," *The New England journal of medicine*, vol. 376, no. 26, p. 2507, 2017.

[8] A. Alghamdi, T. Alsubait, A. Baz, and H. Alhakami, "Healthcare analytics: A comprehensive review," *Engineering, Technology & Applied Science Research*, vol. 11, no. 1, pp. 6650–6655, 2021.

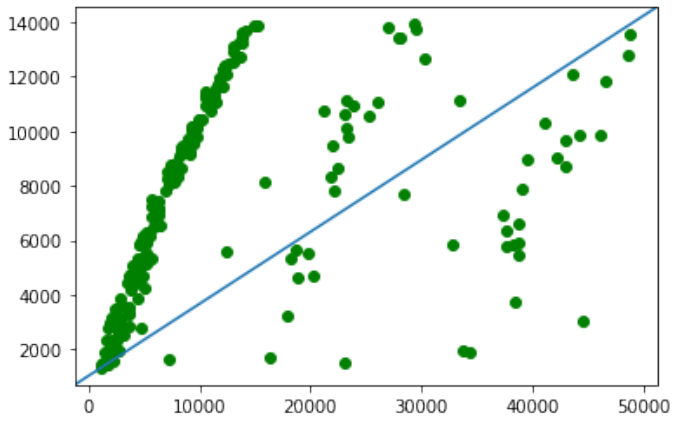
[9] P. Singh, S. Singh, and D. Singh, "An introduction and review on machine learning applications in medicine and healthcare," in *2019 IEEE conference on information and communication technology*. IEEE, 2019, pp. 1–6.



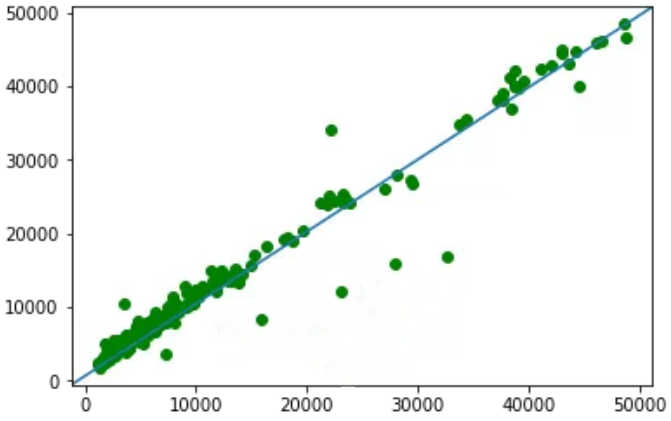
(a) Linear regression



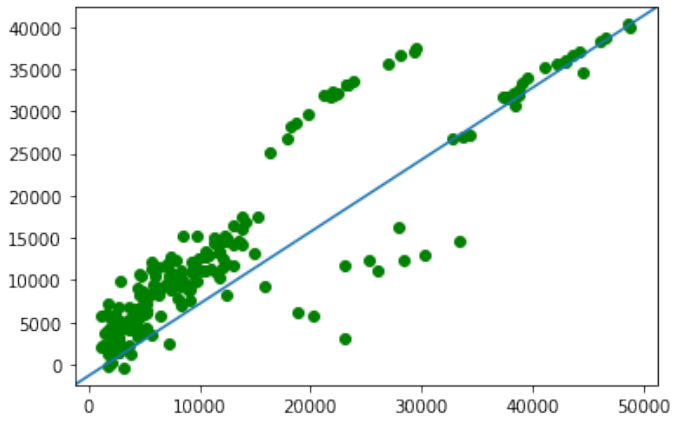
(b) Ridge regression



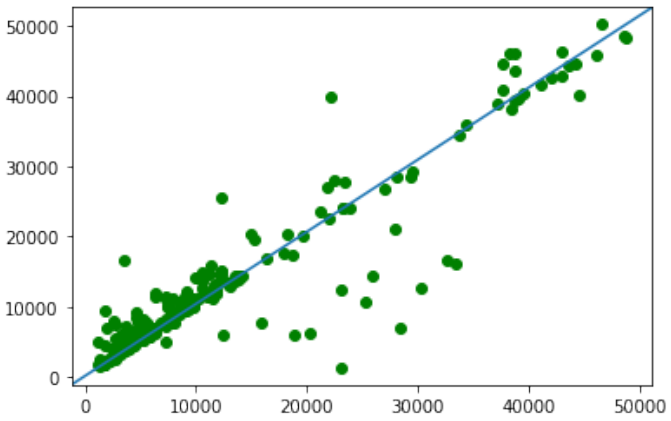
(c) Support vector regression



(d) Neural network



(e) Random Forest regression



(f) Improved random Forest regression

Fig. 6: Actual (blue) vs. predicted y (green) for different models.

[10] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.
[11] H. Nguyen, T. Nguyen, and D. T. Nguyen, "An empirical study on prediction of population health through social media," *Journal of Biomedical Informatics*, vol. 99, p. 103277, 2019.
[12] S. Rose, "Robust machine learning variable importance analyses of medical conditions for health care spending," *Health Services Research*, vol. 53, no. 5, pp. 3836–3854, 2018.
[13] B. Lantz, *Machine Learning with R*. Birmingham: Packt Publishing, Oct. 2013.
[14] J. L. Moran, P. J. Solomon, A. R. Peisach, and J. Martin, "New models for old questions: generalized linear models for cost prediction," *Journal of evaluation in clinical practice*, vol. 13, no. 3, pp. 381–389, 2007.
[15] D. Bertsimas, M. V. Bjarnadóttir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala, and G. Wang, "Algorithmic prediction of health-care costs," *Operations Research*, vol. 56, no. 6, pp. 1382–1392, 2008.
[16] B. Lahiri and N. Agarwal, "Predicting healthcare expenditure increase for an individual from medicare data," in *Proceedings of the ACM SIGKDD workshop on health informatics*, 2014, pp. 73–79.
[17] D. Gregori, M. Petrinco, S. Bo, A. Desideri, F. Merletti, and E. Pagano, "Regression models for analyzing costs and their determinants in health care: an introductory review," *International Journal for Quality in Health Care*, vol. 23, no. 3, pp. 331–341, 2011.
[18] J. A. Hanley, "Simple and multiple linear regression: sample size considerations," *Journal of clinical epidemiology*, vol. 79, pp. 112–119, 2016.
[19] E. Ostertagová, "Modelling using polynomial regression," *Procedia engineering*, vol. 48, pp. 500–506, 2012.
[20] R. M. Heiberger, E. Neuwirth, R. M. Heiberger, and E. Neuwirth, "Polynomial regression," *R Through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics*, pp. 269–284, 2009.
[21] D. W. Marquardt and R. D. Snee, "Ridge regression in practice," *The American Statistician*, vol. 29, no. 1, pp. 3–20, 1975.
[22] G. C. McDonald, "Ridge regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 1, pp. 93–100, 2009.
[23] J. D. Nelson, C. Nafornita, and A. Isar, "Generalised m-lasso for robust, spatially regularised hurst estimation," in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2015, pp. 1265–1269.
[24] S. Sushmita, S. Newman, J. Marquardt, P. Ram, V. Prasad, M. D. Cock, and A. Teredesai, "Population cost prediction on public healthcare datasets," in *Proceedings of the 5th international conference on digital health 2015*, 2015, pp. 87–94.
[25] D. Erlangga, M. Suhrcke, S. Ali, and K. Bloor, "The impact of public health insurance on health care utilisation, financial protection and health status in low- and middle-income countries: A systematic review," *PLoS ONE*, vol. 14, no. 8, Aug. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6713352/
[26] C. Yang, C. Delcher, E. Shenkman, and S. Ranka, "Machine learning approaches for predicting high cost high need patient expenditures in health care," *BioMedical Engineering OnLine*, vol. 17, no. 1, p. 131, Nov. 2018. [Online]. Available: https://doi.org/10.1186/s12938-018-0568-3
[27] M. A. Morid, K. Kawamoto, T. Ault, J. Dorius, and S. Abdelrahman, "Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation," *AMIA Annual Symposium Proceedings*, vol. 2017, pp. 1312–1321, Apr. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977561/
[28] A. M. Jödicke, U. Zellweger, I. T. Tomka, T. Neuer, I. Curkovic, M. Roos, G. A. Kullak-Ublick, H. Sargsyan, and M. Egbring, "Prediction of health care expenditure increase: how does pharmacotherapy contribute?" *BMC Health Services Research*, vol. 19, no. 1, p. 953, Dec. 2019. [Online]. Available: https://doi.org/10.1186/s12913-019-4616-x
[29] S. Chibani and F.-X. Coudert, "Machine learning approaches for the prediction of materials properties," *APL Materials*, vol. 8, no. 8, p. 080701, Aug. 2020. [Online]. Available: https://aip.scitation.org/doi/10.1063/5.0018384
[30] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.
[31] V. Jakkula, "Tutorial on support vector machine (svm)," *School of EECS, Washington State University*, vol. 37, no. 2.5, p. 3, 2006.
[32] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?" in *Machine Learning and Data Mining in Pattern Recognition*, P. Perner, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 154–168.