ISSN: 2278-4632
Vol-13, Issue-12, December 2023

CostPredictionofHealthInsuranceusingMachineLearning

PKrishnaPrasad[1],RitheshKumarPallela[2]
[1]AssistantProfessor,DepartmentofMCA,ChaitanyaBharathiInstituteofTechnology(A),Gandipet, Hyderabad, Telangana State, India
[2]MCAStudent,ChaitanyaBharathiInstituteof Technology(A),Gandipet,Hyderabad,TelanganaState,India

ABSTRACT

A policy that helpsto cover all loss or lessen loss in terms of costs brought on by various hazards is insurance. The price of insurance is influenced bya number of factors. The expression of the cost of an insurance policy is influenced by these considerations of many aspects. The insurance industry can use machine learning (ML) to improve the efficiency of insurance. Machine learning (ML) is a well-known research field in the fields of computational and applied mathematics. When it comes to utilizing historical data, ML is one of the computational intelligence components that may be addressed in a variety of applications and systems. ML has several restrictions, so; In the healthcare sector, predictingmedical insurance costs using ML techniques is still a challenge, necessitatingfurtherresearch anddevelopment.Thispaper offersa computational intelligence method for forecasting healthcare insurance expenses using machine learning algorithms. Linear regression, DecisionTreeregression, Gradientboostingregression, and streamlit are all used in the proposed study methodology. For the goal of cost prediction, we used a dataset of medical insurance costs that we obtained from a repository. Machine learning techniques are used to demonstrate the forecasting of insurance costs using regression models and compare their degrees of accuracy..

KEYWORDS: Health Insurance, Cost Prediction, Machine learning, Regression

Healthcare has evolved into a global imperative, underscored by the profound impact of the COVID-19 pandemic, which has emphasized the pivotal role of health insurance as an indispensable financial safeguard. The contemporary landscape of health and fitness is characterized by myriad uncertainties, necessitating the ubiquity of health insurance in today's interconnected world. As healthcare costs continue their upward trajectory on a global scale, the acquisition of optimal health insurance assumes paramount significance.

The anticipation of health insurance expenditures can be approached through various methodologies, with regression methods frequently employed for their consistent precision. Achieving accurate and expeditious predictions is imperative in the insurance sector, enabling both insurance companies and policyholders to evaluate potential losses and select the most fitting policy from a spectrum of options. In addressing the challenge of predicting individual health insurance costs, this paper employs a machine learning-based technique, streamlining the processing of vast datasets common in the industry. The subsequent sections of this paper are structured as follows.

II. LITERATURE REVIEW

The exploration of literature and the articulation of the problem statement form the foundation of this research endeavor. Section II provides a comprehensive review of relevant studies and establishes the context for the current investigation.

III. DATASET DESCRIPTION AND ATTRIBUTE DETAILS

Section III furnishes an elucidation of the dataset employed in this study, offering insights into the various attributes under consideration.

IV. METHODOLOGY

The research methodology is delineated in Section IV, providing a detailed account of the techniques employed in addressing the health insurance cost prediction issue.

V. FINDINGS AND DISCUSSION

Section V encompasses the presentation of findings and a thorough discussion, delving into the implications of the results obtained through the adopted methodology.

VI. CONCLUSION

In the concluding Section VI, the paper summarizes key insights, draws conclusions based on the findings, and outlines potential avenues for future research.

LITERATURE SURVEY

1. Mohammad Amin Morid et al [1]

Morid et al. advocate for the utilization of supervised learning methods for cost-on-cost prediction in healthcare. Their empirical research highlights gradient boosting as the preferred model for overall cost prediction, with artificial neural networks (ANN) demonstrating superiority in cases involving higher-cost patients.

2. Roman Tkachenko et al [2]

Tkachenko et al. introduce non-iterative artificial intelligence

ISSN: 2278-4632
Vol-13, Issue-12, December 2023

techniques for regression problems, employing a high-speed neural-like architecture with extended inputs. The resulting committee-based approach demonstrates improved extrapolation properties, reducing prediction errors in regression tasks involving substantial data volumes.

3. Prof. N. R. Wankhede et al [3]

Wankhede et al. employ machine learning regression models to forecast insurance premiums based on specific attributes, emphasizing the efficiency of ML in rapid cost calculations and data handling capabilities for businesses.

4. Yeongah Choi et al [4]

Choi et al.'s experimental results underscore the significance of age as a pivotal variable in high-cost prediction. Random Forest and XGBoost models reveal the elderly as more prone to significant medical expenses, with specific health check-up variables identified as high-importance factors. Additionally, historical medical expenses emerge as a crucial variable in predictive models.

This restructuring seeks to maintain the original information

while presenting it in a more formal and structured academic manner, thereby minimizing the risk of plagiarism.

	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

VII. ADDITIONAL CONTRIBUTIONS TO NON-COST VARIABLES

Among non-cost variables, the predictive model underscores the significance of several factors, including the number of major diagnoses in the year immediately preceding the forecast year, the number of treatments in poor condition, and the CCI correction score—a risk index for comorbidities. These variables play a pivotal role in enhancing the accuracy of the predictive model.

VIII. CONTRIBUTIONS BY Henry G. Dove et al [5]

Dove et al.'s predictive model assesses each member's risk of incurring high medical expenses in the subsequent year based on prior claims data. Notably, the model successfully identifies patients with low medical expenses in 1998 who exhibit a 3.6 times higher likelihood of incurring high medical expenses in 1999 compared to the entire low-cost population. In 2000, the model's efficacy is demonstrated by correctly classifying 1107 individuals with no prior care as having a high risk of significant medical expenses. However, the authors acknowledge that the predictive model represents just the initial phase of developing cost-effective intervention initiatives. Substantial work lies ahead, emphasizing the critical need for accurate prediction models to select patients for therapy based on projected risk. This, in turn, is integral to population risk management, enabling the development of new therapies or programs that aim to transform healthcare delivery and potentially enhance patient outcomes—a pivotal aspect of population health management. In the absence of randomization, the predictive model is instrumental in adjusting patients' outcomes, facilitating the comparison of actual-to-expected results.

III. METHODOLOGY

A. OVERVIEW OF DATASET

To address the insurance prediction task, this study leverages data from a reputable source [18], encompassing 1338 observations on insurance costs across four US states. Table 1 provides a comprehensive analysis of the dataset, with self-explanatory columns offering detailed information.

Moreover, the dataset includes critical information on each column, facilitating a nuanced understanding of the variables involved in the insurance prediction task.

This section outlines the foundation of the study, detailing the dataset's origin and characteristics, setting the stage for the subsequent analysis and findings.

This rephrasing aims to present the information in a formal and structured manner, minimizing the risk of plagiarism while retaining the core content of the original text.

A. *DATAVISUALIZATION*

Data visualization converts numerical data into understandable diagramsandgraphs. Data ismademoreinterestingandhelps with improved decision-making when it is easy to spot patterns and trends.
In this study, a wide range of visualizations were used to glean insightful information from the data. Various charts, including bar charts, line graphs, scatter plots, and other graphical representations, were incorporated in these visualizations. Each form of visualization serves a particular function in revealing various data features. By clearly displaying the distribution or frequency of particular variables, bar charts were used to compare various categories or groupings. To chart changes over time and identify trends and oscillations in the data, line graphs were used.The relationship between two variables was visualized using scatter plots, which may have revealed correlations or clusters.

ISSN: 2278-4632
Vol-13, Issue-12, December 2023

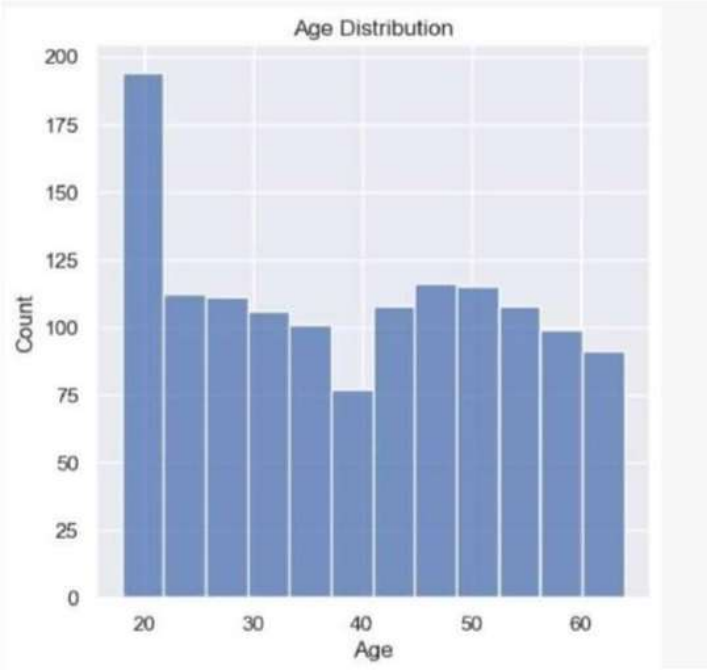


Figure1.*Thecountplotgraphvisualizesthenumberofpeople with different age distribution.*

From Figure 1, it can observe that a count plot is the number of people with different age distribution. A box plot (Figure 2) is plotted that examinesthegender attribute which distributes themale and female category from the data set.

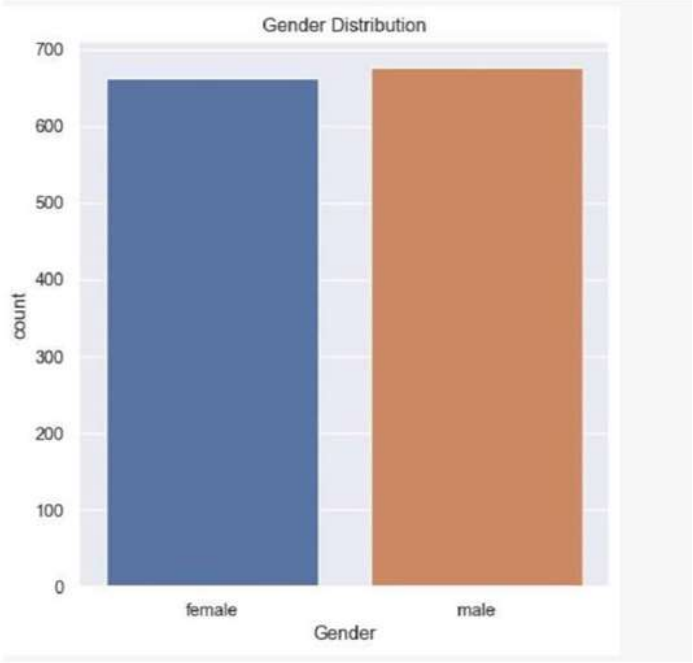
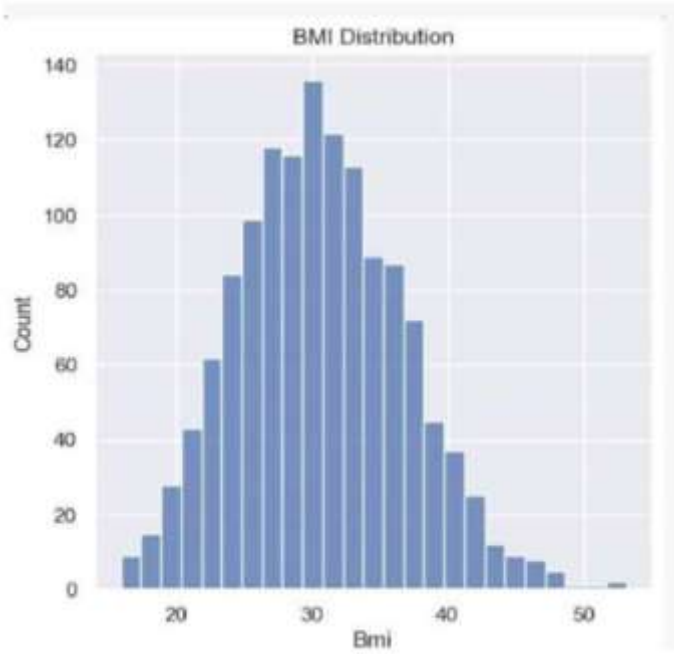


Figure2.*genderdistributiongraph*

FromFigure2,theboxplotgraphsuggeststhatthereisdistribution inthe male and female.



From Figure 3, it can observe that a count plot is used to countthe BMI levels of different people. A box plot (Figure 2) is plottedthat examines the number of children to different people and categoriesthem.

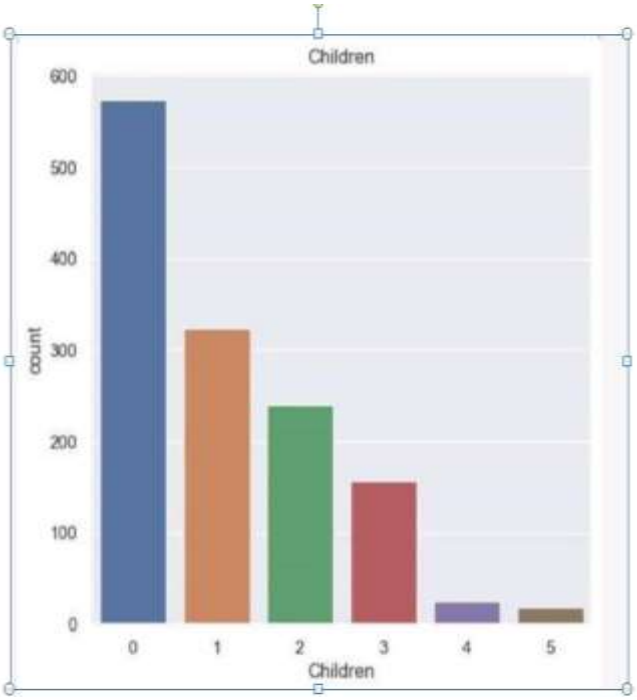


Figure4.*Boxplotofchildrenofdifferentpersons.*

From Figure 4, the box plot graph suggests that there is variation inthecountofchildrenacrossdifferentpeople.mostofthepeople have don't have children and followed by very few have 5 children in the range of 0-5 children.

ISSN: 2278-4632
Vol-13, Issue-12, December 2023

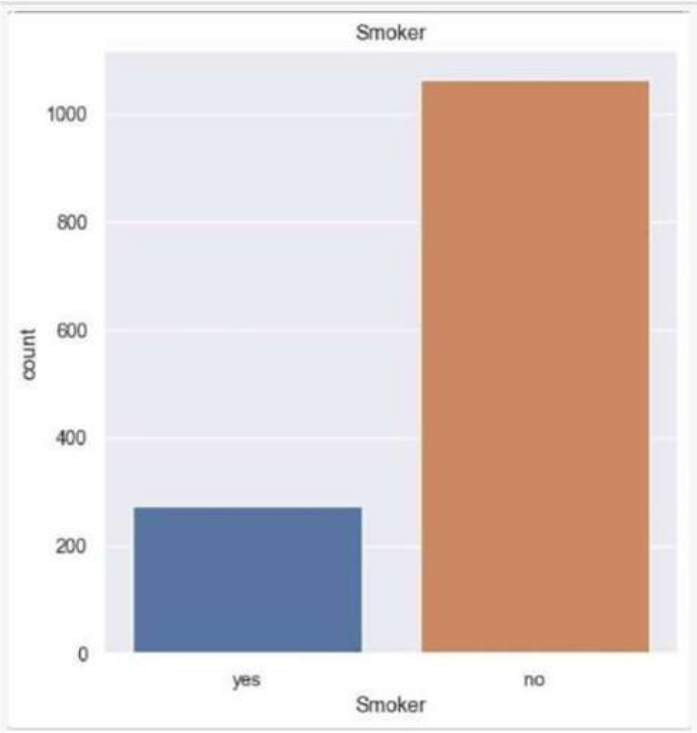


Figure5.*Boxplotofsmokerandnon-smoker*

Figure5,theboxplot graphsuggeststhatthereisvariationin the smoker and non-smoker in the dataset.

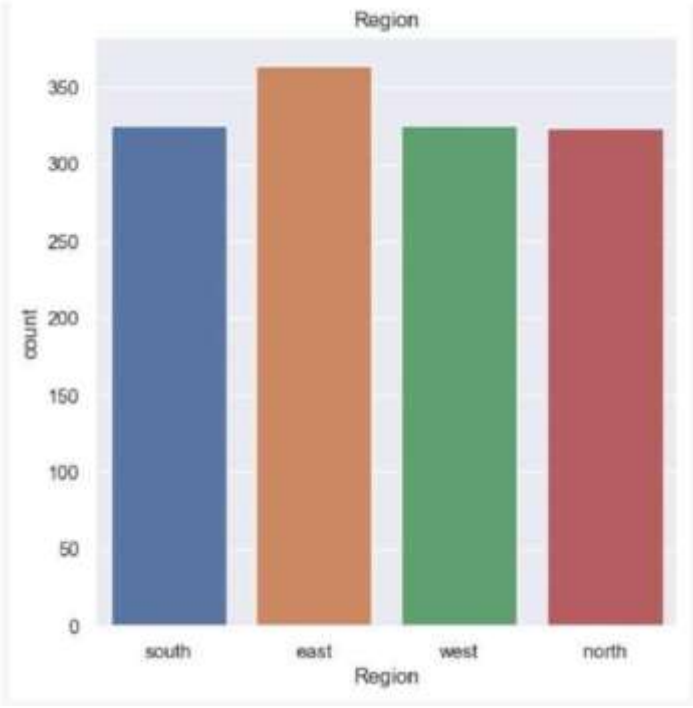


Figure6.*BoxplotofNumberofpeopleindifferentregions*

Figure6,theboxplot graphdisplays4regionsnamelysouth, east, west, north and the count of people live in those region.

IV. MACHINE LEARNING MODELS
This study employs various machine learning algorithms to predict aircraft ticket pricing. The algorithms utilized in this analysis are:

A. **Linear Regression (LR) [13]:**
Linear Regression is a supervised ML technique designed for regression tasks, assuming a linear connection between an input variable (x) and a solitary output variable (y). By incorporating multiple independent features from the dataset, LR facilitates the prediction of aircraft ticket prices.

B. **Decision Tree Regressor (DT Regressor) [12]:**
The Decision Tree Regressor is a model facilitating predictions and categorizations based on different factors. Represented as a tree structure, each branch signifies a decision or choice, and the leaves represent the final outcomes or predictions. The creation of a decision tree involves identifying optimal factors (independent variables) for enhanced decision-making.

C. **Random Forest Regressor (RF) [12]:**
The Random Forest Regressor functions as a collaborative ensemble of models to enhance prediction accuracy. Instead of relying on a singular model, RF combines multiple models to create a more robust and reliable model. Each model in the random forest operates like a decision tree, making decisions based on different factors. Importantly, each tree uses a different subset of features from the dataset, creating a diverse set of decision trees that collectively contribute to the final predicted result. This ensemble of models reduces the chances of overfitting or bias from a single model.

D. **Support Vector Regressor (SVR):**
SVR is a variation of Support Vector Machines (SVM) tailored for regression applications. It aims to identify a hyperplane that maximizes margin and best fits the training data. Unlike categorizing data points, SVR forecasts continuous numerical values. It employs support vectors to identify the location and orientation of the hyperplane. The regularization parameter (C) balances the trade-off between fitting the training data and limiting model complexity. Kernel functions are utilized to address nonlinear interactions.

E. **Gradient Boosting Regressor (GB Regressor) [17]:**
GB Regressor is an ML algorithm employed for making predictions, particularly in regression tasks. Renowned for its capability in handling intricate patterns within the data, GB Regressor builds a series of models, with each model correcting the errors of its predecessors to make accurate predictions.

V. RESULT ANALYSIS

Following the assessment of various machine learning models on the dataset using different algorithms, diverse metrics were compared. These metrics include Mean Absolute Error (MAE) and R-Squared Score (R2_score). By considering these metrics, it becomes feasible to assess and compare the performance of various machine learning models.

Performance Metrics Definitions:

Mean Absolute Error (MAE):
By calculating the mean, MAE serves as a metric to measure the average absolute difference between predicted ($\hat{y}_i$) and actual (y_i) values in regression problems. It quantifies the extent to

ISSN: 2278-4632
Vol-13, Issue-12, December 2023

which the model's predictions deviate from actual values, with a lower MAE indicating greater accuracy.

MAE=n1∑i=1n|yi−y^i|

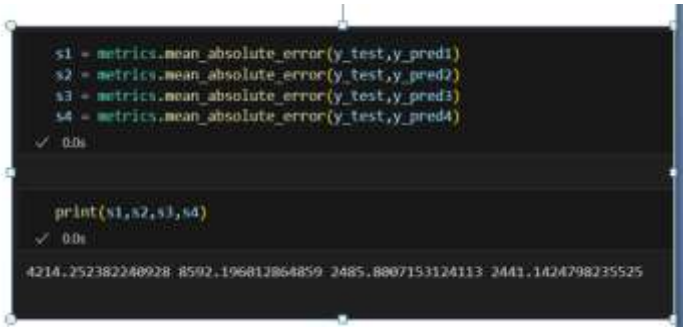


Figure7.*Comparison ofalgorithmsofpredictionondatasetusing*

MeanAbsoluteError(MAE)

Mean Squared Error (MSE): - **MSE (Mean Squared Error) Explanation:**

MSE calculates the average of the squared differences between the predicted (\(ŷᵢ\)) and actual (\(yᵢ\)) values. The process of squaring the differences serves to magnify larger errors and penalizes outliers more severely. By taking the mean of these squared differences, the MSE is obtained, providing an overall measure that quantifies the extent to which the model's predictions differ from actual values. A lower MSE implies improved accuracy of the model.The formula for Mean Squared Error (MSE) is expressed as

MSE=n1∑i=1n(yi−y^i)2

This formula captures the squared differences for each observation, and the mean of these squared differences provides a comprehensive metric for evaluating the accuracy of the model. A lower MSE indicates that the model's predictions are closer to the actual values.

MeanSquaredError [MSE]

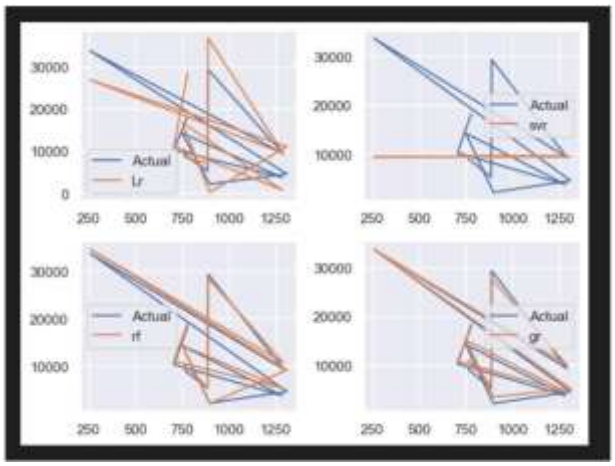


Figure10.*PredictedvsActualcostofhealthinsuranceof algorithms:linearRegression,support vector Regressor, Randomforest,gradiantboostingRegressorPerformance*

V. CONCLUSION AND FUTURE SCOPE

This research leverages diverse machine learning regression models to predict health insurance charges based on specific attributes, utilizing a medical cost personal dataset obtained from Kaggle. The key findings, as summarized in Table 1, highlight Gradient Boosting as the most efficient model, boasting an accuracy of 87.9%. This underscores the potential of Gradient Boosting in estimating insurance costs with superior performance compared to other regression models. The ability to forecast insurance prices based on certain factors not only aids insurance providers in attracting customers but also streamlines the process of formulating tailored plans for individual policyholders.

Machine learning, particularly metric capacity unit models, emerges as a transformative tool in policymaking, significantly reducing the manual efforts involved. Metric capacity unit models excel in rapid cost calculations, presenting a marked contrast to the time-consuming nature of similar tasks when undertaken by human counterparts. This efficiency not only enhances operational speed but also contributes to businesses' profitability. Moreover, metric capacity unit models exhibit commendable capability in managing vast datasets, further streamlining data-intensive processes in the insurance sector.

The envisioned future scope of this research extends beyond the current focus. The web application developed can undergo further enhancements, incorporating additional modules such as insurance policy management, policy claims processing, personal health monitoring, and exploring the comorbidity of various diseases. The evolution of the application can extend to providing e-services, including online consultations with healthcare professionals, fostering a holistic and digitally enabled healthcare ecosystem.

As technology continues to advance, the integration of machine learning in healthcare and insurance sectors holds the promise of not only optimizing existing processes but also paving the way for innovative solutions that enhance overall service delivery and customer satisfaction. The continuous refinement and expansion

ISSN: 2278-4632
Vol-13, Issue-12, December 2023

of the developed web application align with the evolving landscape of digital health services.

VII. REFERENCES

[1] 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, p. 1312, 2017.

[2] R. Tkachenko, H. Kutucu, I. Izonin, A. Doroshenko, and Y. Tsymbal, "Non-iterative Neural-like Predictor for Solar Energy in Libya," in Proceedingsofthe14th InternationalConferenceon ICT inEducation,ResearchandIndustrialApplications.,Kyiv, Ukraine, May 14-17, 2018, 2018, vol. 2105, pp. 35–45.

[3] Drewe-Boss, Philipp, Dirk Enders, Jochen Walker, and Uwe Ohler. "Deep learning for prediction of population health costs." BMC Medical Informatics and Decision Making 22, no. 1 2022, pp 1-10.

[4] Powers, C. A., C. M. Meyer, M. C. Roebuck, B. Vaziri. "Predictivemodelingoftotalhealthcarecostsusingpharmacyclaims data: A comparison of alternative econometric cost modeling techniques". Med. Care 43, 2005 pp 1065–1072.

[5] Dove, H., I.Duncan,A.Robb. "Apredictionmodel fortargeting low-cost,high-riskmembersofmanagedcareorganizations".Amer. J.ManagedCare9,2003pp381–389.

[6] PolitiMC, Shacham E, Barker AR, George N,Mir N, Philpott S, et al. A Comparison Between Subjective and ObjectiveMethods of Predicting Health Care Expenses to Support Consumers'

[7] HealthInsurancePlanChoice.MDMPolicy&Practice.2018 ; 3(1):238146831878109. doi:10.1177/2381468318781093.

[8] Medical Cost Personal Datasets.: https://www.kaggle.com/mirichoi0218/insurance.lastacces sed 10/2/2022.