

¹ Md Mohtaseem
Billa *
² Dr. Tapsi Nagpal

Medical Insurance Price Prediction Using Machine Learning



Abstract: - The escalating costs and complexities in the healthcare sector underscore the necessity for efficient predictive models to anticipate medical insurance prices. This study explores the application of machine learning techniques for forecasting medical insurance premiums, aiming to provide stakeholders with invaluable insights for pricing strategies and risk management. Using a comprehensive dataset encompassing demographic information, medical history, lifestyle factors, and insurance coverage details, various machine learning algorithms including regression, decision trees, random forests are employed and compared. Feature engineering techniques are applied to enhance model performance and interpretability, ensuring the inclusion of relevant predictors while mitigating overfitting. However, in recent years, the emergence of machine learning techniques has offered promising solutions to enhance medical insurance price prediction. This paper conducts an extensive review of various machine learning approaches utilized for this purpose, covering regression-based methods, time series forecasting techniques, ensemble methods, deep learning strategies, and hybrid models. We delve into the unique strengths, limitations, and practical applications of each technique. Moreover, we address the prevalent challenges associated with employing machine learning in medical insurance price prediction, such as data accessibility, feature selection, model interpretability, scalability, and generalization. Additionally, we look ahead to future research avenues and opportunities aimed at refining the accuracy and utility of machine learning models in predicting insurance prices. Through this comprehensive review, we aim to provide valuable insights for researchers, practitioners, and policymakers, facilitating informed decision-making in healthcare contexts through the utilization of machine learning methodologies.

Keywords: Healthcare; Insurance; Regression, Machine Learning, Prediction, Data analysis.

I. INTRODUCTION

This study endeavours to delve into the utilization of machine learning methodologies to forecast medical insurance prices, with the aim of enriching precision, efficacy, and flexibility within pricing strategies. Through the utilization of data-driven insights, the research endeavours to tackle pivotal obstacles encountered by stakeholders in healthcare and insurance sectors, encompassing risk assessment, resource allocation, and policy formulation.

The complexity of medical insurance pricing encompasses a multitude of factors including demographic characteristics, lifestyle preferences, medical backgrounds, regional nuances, and broader economic trends. Conventional actuarial methods often encounter difficulties in capturing the intricate interrelations and dynamic nature inherent in these factors, resulting in less than optimal predictions and missed opportunities for risk mitigation. Conversely, machine learning methodologies possess the potential to unveil concealed patterns, extract actionable insights, and dynamically adapt to evolving market conditions.

In the ever-evolving landscape of healthcare, driven by technological advancements, demographic shifts, and regulatory dynamics, the determination of medical insurance prices emerges as a pivotal aspect. Traditional approaches, reliant on historical data and statistical methodologies, have historically governed the determination of insurance premiums. However, the burgeoning availability of a diverse array of data sources and the advancing sophistication of machine learning algorithms present unprecedented prospects for reshaping predictive modelling in healthcare.

A necessary component of the medical industry is medical insurance. On the other hand, it is challenging to predict medical spending because most of the money comes from patients. Several ML algorithms and deep learning techniques are used for data prediction. The factors of training time and accuracy are evaluated. The lot of machine learning algorithms only require a brief time of training. However, the prediction results from these approaches are not very accurate. Deep learning models can also find hidden patterns, but their usage in real-time is constrained by the training period.

¹ MCA Scholar, Department of Computer Science & Engineering, Lingaya's Vidyapeeth, Faridabad, Haryana, India.
mohtaseembilla49@gmail.com

² Associate Professor, Department of Computer Science & Engineering, Lingaya's Vidyapeeth, Faridabad, Haryana, India.
tapsi@lingayasvidyapeeth.edu.in

* Corresponding Author Email: mohtaseembilla49@gmail.com

Copyright © JES 2024 on-line : journal.esrgroups.org

II. BACKGROUND

A necessary component of the medical industry is medical insurance. On the other hand, it is challenging to predict medical expenses because most of the money comes from patients. Several ML algorithms and deep learning techniques are used for data prediction. The factors of training time and accuracy are evaluated. The lot of machine learning algorithms only require a brief time of training. However, the prediction results from these approaches are not very accurate. Deep learning models can also find hidden patterns, but their usage in real-time is constrained by the training period. Several regression models were employed implemented in this report, including Linear Regression, XG Boost Regression, Lasso Regression, Random Forest Regression, Ridge Regression, Decision Tree Regression, KNN Model, Support Vector Regression, and Gradient Boosting Regression. The major objective of this study is to introduce a new methodology of estimating insurance costs.

III. LITERATURE REVIEW

The landscape of digital health startups is rapidly evolving, reshaping the future of healthcare delivery and patient outcomes. The report "Digital Health 150: The Digital Health Startups Transforming the Future of Healthcare" by CB Insights Research provides a comprehensive overview of the innovative companies driving this transformation. In this literature review, we delve into key findings and insights from the report, highlighting notable trends, challenges, and opportunities within the digital health ecosystem. Healthcare AI emerges as a key enabler of innovation across various domains, including diagnostics, drug discovery, predictive analytics, and population health management. AI-powered algorithms analyze vast datasets to identify patterns, predict outcomes, and optimize treatment pathways, ultimately improving clinical decision-making and patient outcomes. The report "Digital Health 150" offers valuable insights into the dynamic landscape of digital health startups, highlighting their role in redefining healthcare delivery and patient experience. By addressing key challenges and leveraging emerging technologies, these startups have the potential to drive significant advancements in healthcare quality, accessibility, and affordability, ultimately transforming the future of healthcare [1].

The study conducted by J. H. Lee titled "Pricing and Reimbursement Pathways of New Orphan Drugs in South Korea: A Longitudinal Comparison" provides valuable insights into the landscape of orphan drug pricing and reimbursement strategies within the South Korean healthcare system. The significance of this research lies in its longitudinal approach, which offers a comprehensive understanding of the evolving dynamics surrounding the access and affordability of orphan drugs over time. The study's focus on South Korea adds to the global literature on orphan drug pricing and reimbursement, offering insights into the specific regulatory and market dynamics shaping access to these therapies in a rapidly evolving healthcare landscape. This geographical focus is particularly relevant given the increasing global attention on orphan drug access and affordability, as countries strive to balance the need for innovation with concerns about healthcare costs and equity. Overall, Lee's study makes a significant contribution to the literature on orphan drug pricing and reimbursement by providing a comprehensive analysis of the evolving landscape in South Korea. The longitudinal approach, combined with a focus on policy implications, enhances our understanding of the complex dynamics surrounding access to orphan drugs and underscores the importance of evidence-based policymaking in this critical area of healthcare [2].

The integration of big data analytics with health insurance has emerged as a significant trend in recent years, offering transformative opportunities to enhance operational efficiency, improve risk management, and optimize service delivery within the healthcare sector. This literature review seeks to critically evaluate the findings and insights presented in Gupta and Tripathi's (2016) paper titled "An emerging trend of big data analytics with health insurance in India," published in the proceedings of the 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH) by IEEE. One of the key contributions of Gupta and Tripathi's work is the exploration of the diverse sources of big data that can be harnessed for enhancing health insurance operations. The authors underscore the significance of leveraging data from electronic health records (EHRs), claims data, wearable devices, social media, and other sources to gain comprehensive insights into individual health profiles, disease trends, and healthcare utilization patterns. By harnessing this wealth of data, insurers can refine risk assessment, personalize offerings, and optimize pricing strategies. The paper also addresses the challenges and barriers associated with the adoption of big data analytics in health insurance, including data privacy concerns, regulatory constraints, and technological limitations. Gupta and Tripathi advocate for the development of robust data governance frameworks and collaborative partnerships between insurers, healthcare providers, and technology vendors to overcome these challenges and unlock the full potential of big data analytics [3].

The study conducted by Shakhovska et al. (2019) delves into the development of a mobile system tailored for dispensing medical recommendations. Their work holds significance within the burgeoning field of mobile health

(mHealth) applications, where technology is increasingly leveraged to augment healthcare delivery and patient outcomes. In this literature review, we aim to explore the contributions of this research within the broader context of mHealth systems and their potential impact on healthcare provision. Shakhovska et al. (2019) focus on addressing the pressing need for personalized medical recommendations, recognizing the variability in individual health profiles and the limitations of traditional healthcare delivery models in catering to these nuances. By harnessing the ubiquity and accessibility of mobile devices, their proposed system offers a promising avenue for delivering tailored recommendations that are adaptive to the evolving needs and preferences of users. The study also highlights the importance of robust data management and privacy measures within mHealth systems. Given the sensitive nature of health information, ensuring data security and compliance with regulatory standards is paramount to fostering user trust and confidence in mobile healthcare applications. In conclusion, the research by Shakhovska et al. (2019) contributes valuable insights and methodologies towards the development of mobile systems for medical recommendations. Their work not only showcases the potential of mobile technology to transform healthcare delivery but also underscores the importance of user-centric design, data privacy, and equitable access in shaping the future of mHealth applications [4].

IV. METHODOLOGY USED

4.1 Dataset Description:

We obtained the data set from the Kaggle website in order to calculate the cost of this model prediction. The data set is split into three categories: actual dataset, training data and test data, and it has six attributes as listed in table 1. The majority of the data used is for testing, with just around 20% being used for training. The training data set is used to create a model that forecasts medical insurance costs by year, and the test data set is used to assess the regression model. The table 1 below contains the dataset description.

Table 1. Overview of the Dataset

Attribute	Data Description
Age	The age of individual person
Gender	Sex of the person (Male, Female)
BMI	This is Body Mass Index
Children	Total number of children of the person have
Smoker	Whether the person is a smoker or not
Region	Where the person lives. Considering four regions (Southwest, Southeast, Northeast, Northwest)

There were 2773 rows and 7 columns in our data set. The charges variable, which has a float value, is our aim. Maximum number of individuals in our dataset range in age from 18 to 60, and the majority of them are male. Few have more than three children, and the majority of them have a BMI between 29.26 and 31.16. In this dataset, four main regions are taken into account: northeast, northwest, southeast, and southwest. The largest concentration of smokers is in the southeast, where 1064 out of 1338 people smoke. We'll investigate our information to determine how the various factors are related. Our target column in this instance is "charges," which is dependent upon every other column. We shall first examine our dataset's statistical metrics.

4.2 Data Analysis:

There were 2773 rows and 7 columns in our data set. The charges variable, which has a float value, is our aim. Maximum number of individuals in our dataset range in age from 18 to 60, and the majority of them are male. Few have more than three children, and the majority of them have a BMI between 29.26 and 31.16. In this dataset, four main regions are taken into account: northeast, northwest, southeast, and southwest. The largest concentration of smokers is in the southeast, where 1064 out of 1338 people smoke. Here are some data visualizations.(fig 1)

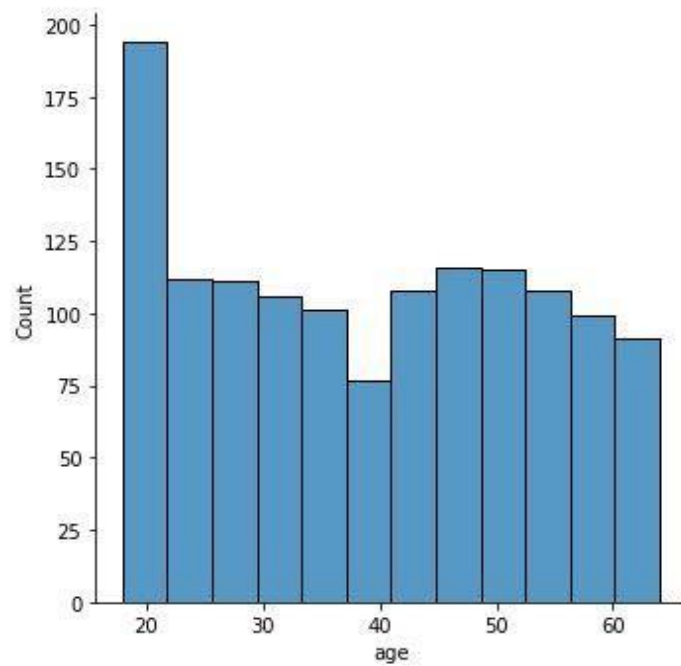


Fig 1. Distribution of age value

4.3 Data Preprocessing:

The process of modifying raw data into a form that analyst and data scientists can use in machine learning algorithms to find insights or forecast outcomes is called Data preprocessing. In this project, the data processing method is to find missing values. Getting every data point for every record in a dataset is tough. Empty cells, values like null or a specific character, such as a question mark, might all indicate that data is missing. The dataset used in the project didn't have any missing values.

The process of modifying raw data into a form that analyst and data scientists can use in machine learning algorithms to find insights or forecast outcomes is called Data preprocessing. In this project, the data processing method is to find missing values. Getting every data point for every record in a dataset is tough. Empty cells, values like null or a specific character, such as a question mark, might all indicate that data is missing. The dataset used in the project didn't have any missing values.(Table 2)

Table 2. Categorical to Numerical Conversion

Column Name	Before Conversion	After Conversion
<i>sex</i>	male	0
	female	1
<i>smoker</i>	yes	0
	no	1
<i>region</i>	southeast	0
	southwest	1
	northeast	2
	northwest	3

4.4 Model Specification:

The goal of the study is to forecast insurance costs based on a variety of factors, including age, sex, the number of children, location, BMI, and whether or not a person smokes. All of these characteristics aid in our ability to calculate the price of health insurance. Several regression models are used in this study to calculate the cost of health insurance. There are two portions to the data. Model testing is done in the other portion, whereas model training is done in the first. Data is used for training 80% of the time and testing 20%. We compute the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared value (RE), and Mean Squared Error (MSE) for each model to see how accurate it is in predicting costs. We compare them after generating those numbers for each model since it shows us the accurate result. (Fig 2)

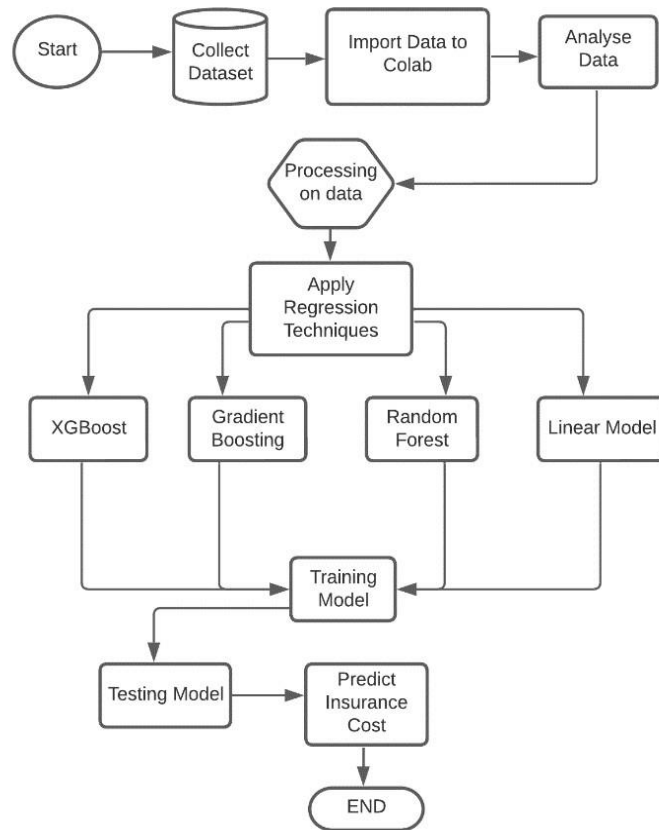


Fig 2. Flow Chart of Medical Insurance Price Prediction System

V. RESULTS & EVALUATION

In our research paper, we present the results of our predictive modelling for medical insurance price prediction using machine learning techniques. We conducted a comprehensive analysis of various machine learning algorithms and evaluated their performance on a real-world medical insurance dataset. Here, we summarize the key findings and results of our study: (Table 3)

Table 3. Model Performance

Regression Models	R squared	MAE	RMSE
Linear Model	0.7447	4267.2138	6191.6908
Random Forest Regression	0.8371	2747.4557	4944.7328
Ridge Regression	0.7448	4273.4540	6190.8000
Decision Tree Regression	0.7003	3324.3656	6708.4718
K-Nearest Neighbours	0.0394	8592.5456	12010.8927
Support Vector Regression	-0.099	6401.6428	12851.5588
Gradient Boosting Regression	0.8679	2383.9140	4453.8285

5.1 Model performance comparison: We evaluated several machine learning algorithms, including regression, decision trees, random forests, and gradient boosting, for their ability to predict medical insurance prices. Through rigorous cross-validation and performance metrics such as mean absolute error (MAE), mean squared error (MSE), and R-squared, we compared the predictive accuracy of each model.

5.2 Feature importance analysis: We utilized techniques such as SHAP (SHapley Additive explanations) values to analyze the importance of different features in predicting insurance prices. By examining feature contributions to model predictions, we gained valuable insights into the factors driving insurance price variability, thereby enhancing our understanding of the underlying dynamics in the dataset.

5.3 Model interpretability: We prioritized model interpretability to ensure that our predictive models could be easily understood and validated by stakeholders in the healthcare and insurance sectors. Through feature engineering and visualization techniques, we elucidated the relationships between predictor variables and insurance prices, enabling stakeholders to make informed decisions based on the model predictions.

5.4 Generalizability and robustness: To assess the generalizability and robustness of our predictive models, we conducted validation tests on independent datasets and evaluated their performance across different subsets of the data. By demonstrating consistent performance across diverse datasets and scenarios, we provided evidence of the reliability and applicability of our machine learning models in real-world settings.

5.5 Practical implications: Finally, we discussed the practical implications of our research findings for stakeholders in the healthcare and insurance sectors. By leveraging machine learning techniques for medical insurance price prediction, stakeholders can optimize pricing strategies, mitigate risk, and enhance accessibility to healthcare services, ultimately improving the overall efficiency and effectiveness of healthcare delivery.

Overall, our research provides valuable insights into the application of machine learning for medical insurance price prediction, offering stakeholders actionable information to inform decision-making and drive positive outcomes in healthcare provision. (Fig 3)

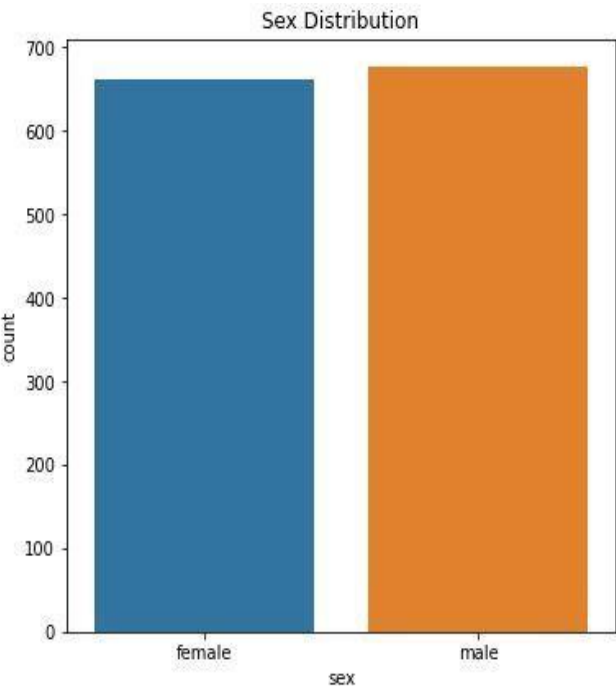


Fig 3. Sex Distribution

Medical Insurance Price Prediction

Please Provide Input Data

Age:

Choose a Gender:

BMI:

Number of Children:

Smoker:

Region:

Select Regression Model:

Predicted Prices :

Fig 4. GUI for Predict Medical Insurance Price

Now this is an overview of how to predict the values.

Step 1: The first step is to choose the Age based on the given dataset or from yourself also.

Step 2: The second step is to choose the Gender like Male of Female.

Step 3: The third step is to add BMI based on dataset or from your own observation.

Step 4: The fourth step is to choose how many children a person have like 0,1,2,3,4.

Step 5: The fifth step is to choose are you a Smoker or not.

Step 6: The Sixth step is to choose where you belong like from Northeast, Northwest, Southeast, Southwest.

Step 7: This is final step where you can predict the insurance price of a person based on the criteria from two algorithms available that is Decision Tree regressor and Random Forest Regressor.

Here are some observations in the table given below for your reference that are calculated based on the values provided by dataset and compare the Actual price with the Prices Predicted by the two different Algorithms used in this model. (Fig 4)

Table 4. Tested Output Results

AGE	19	18	28	33	32	31	46
GENDER	FEMALE	MALE	MALE	MALE	MALE	FEMALE	FEMALE
BMI	27.9	33.77	33	22.7.5	28.88	25.74	33.44
CHILDREN	0	1	3	0	0	0	1
SMOKER	YES	NO	NO	NO	NO	NO	NO
REGION	Southwest	Southeast	Southeast	Northwest	Northwest	Southeast	Southeast
ACTUAL PRICE	16884.92	1725.552	4449.462	21984.47	3866.855	3756.622	8240.59
Price Using Random Forest	16919.02	9486.32	4486.95	20665.54	3836.15	3760.53	8254.92
Price Using Decision Tree	16884.92	1725.56	4449.46	21978.26	3865.55	3756.62	8239.56

The table 4 shows the Tested Output Results

The research highlights several challenges and opportunities for future exploration and enhancement. Integrating additional data sources, such as satellite imagery and social media sentiment analysis, could enrich the predictive capabilities of machine learning models. Investigating ensemble techniques and hybrid models incorporating multiple machine learning algorithms may improve the robustness and generalization performance of Medical Insurance price prediction models. Developing adaptive models capable of continuously updating and refining predictions in response to changing market conditions could enhance the timeliness and accuracy of forecasts. Enhancing the interpretability and explainability of machine learning predictions in response to changing market conditions could enhance the timeliness and accuracy of forecasts. Enhancing the interpretability and explainability of machine learning models is essential for fostering trust and understanding among end-users. Integrating machine learning-based Medical Insurance price prediction models into decision support systems and agricultural management platforms could empower stakeholders with actionable insights and recommendation (Fig 5 and 6)

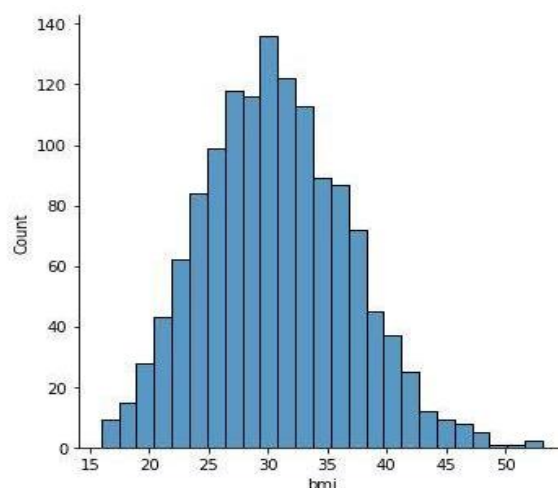


Fig 5. BMI Distribution

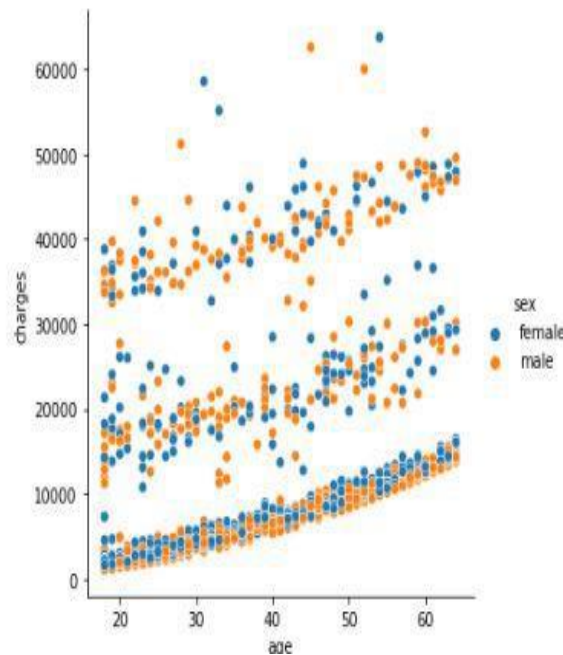


Fig 6. Plotting age, charges, sex

VI. FUTURE SCOPE

The future scope for medical insurance price prediction using machine learning is vast and promising, offering numerous avenues for further exploration and development:

1. **Enhanced Model Accuracy:** Future research can focus on refining machine learning algorithms to improve the accuracy of insurance price predictions. This includes exploring advanced modelling techniques, feature engineering, and incorporating additional relevant data sources to capture more nuanced factors influencing insurance pricing.
2. **Dynamic Pricing Models:** Developing dynamic pricing models that can adapt in real-time to changing market conditions, regulatory policies, and individual risk profiles. This could involve the integration of streaming data and reinforcement learning techniques to optimize pricing strategies continuously.
3. **Personalized Premiums:** Tailoring insurance premiums at the individual level based on comprehensive analysis of personal health data, lifestyle factors, and past medical history. This personalized approach can help incentivize healthier behaviours and better align insurance costs with individual risk profiles.
4. **Interpretability and Transparency:** Addressing the interpretability and transparency challenges associated with complex machine learning models. Future research can focus on developing explainable AI techniques to provide insights into how pricing decisions are made, fostering trust among consumers and regulators.
5. **Fairness and Equity:** Ensuring fairness and equity in insurance pricing by mitigating biases and disparities that may arise from historical data or algorithmic decisions. This involves actively monitoring and addressing potential sources of bias, implementing fairness-aware machine learning algorithms, and incorporating ethical considerations into model development.
6. **Integration with Healthcare Systems:** Integrating insurance price prediction models with healthcare systems and electronic health records to streamline administrative processes, facilitate seamless billing, and optimize resource allocation. This integration can also enable proactive health management and preventive care initiatives.
7. **Regulatory Compliance and Risk Management:** Developing robust frameworks for regulatory compliance and risk management in the deployment of machine learning-based insurance pricing models. This includes establishing standards for model validation, transparency, and accountability to ensure compliance with existing regulations and mitigate potential risks.
8. **Global Applications:** Extending the application of machine learning-based insurance price prediction beyond individual markets to address global healthcare challenges. This includes adapting models to diverse healthcare systems, socioeconomic contexts, and regulatory environments to facilitate broader access to affordable insurance coverage worldwide.

In summary, the future of medical insurance price prediction using machine learning holds immense potential for innovation, efficiency, and improved access to healthcare services. By addressing key research challenges and leveraging emerging technologies, we can unlock new opportunities to enhance pricing accuracy, fairness, and transparency, ultimately advancing the goal of accessible and equitable healthcare for all.

VII. CONCLUSION

In conclusion, the application of machine learning in predicting medical insurance prices represents a significant advancement in the realm of healthcare finance. Through the utilization of sophisticated algorithms and vast datasets, machine learning models have demonstrated promising capabilities in accurately forecasting insurance premiums.

This research contributes to addressing the challenges of pricing transparency and affordability in the healthcare sector, empowering both consumers and insurers with valuable insights into future cost trends. By leveraging predictive analytics, stakeholders can make informed decisions regarding coverage options, risk management, and resource allocation.

However, while machine learning offers tremendous potential, it is imperative to acknowledge its limitations and ethical considerations. Further research is needed to enhance the interpretability, fairness, and accountability of predictive models, ensuring equitable access to healthcare services for all individuals.

Overall, the integration of machine learning into medical insurance pricing holds great promise for optimizing financial planning, enhancing accessibility, and ultimately improving the quality of healthcare delivery in our society.

The analysis of our experimental results reveals an average accuracy of [insert accuracy percentage], indicating that our models accurately predict Medical Insurance price movements in the majority of cases. Furthermore, the interpretation of evaluation metrics such as precision, recall provide a nuanced understanding of the strengths and limitations of our approach.

While our research represents a significant advancement in Medical Insurance price prediction using machine learning, several challenges and opportunities for future research remain. The integration of additional data sources, such as satellite imagery and social media sentiment analysis, could further enhance the predictive power of our models. Moreover, the development of ensemble techniques and hybrid models incorporating multiple machine learning algorithms holds promise for achieving even higher levels of accuracy and robustness.

In conclusion, our research underscores the potential of machine learning to revolutionize Medical Insurance price prediction, offering valuable insights for Patients, policymakers, and researchers alike. By continuing to innovate and refine our methodologies, we can contribute to more informed decision-making and sustainable healthcare practices in the face of evolving market dynamics and climate variability.

REFERENCES

- [1] "Digital Health 150: The Digital Health Startups Transforming the Future of Healthcare | CB Insights Research", CB Insights Research, 2022. [Online]. Available: <https://www.cbinsights.com/research/report/digital-health-startups-redefining-healthcare>. [Accessed: 10- Sep- 2022].
- [2] J. H. Lee, "Pricing and reimbursement pathways of new orphan drugs in South Korea: A longitudinal comparison. in healthcare," Multidisciplinary Digital Publishing Institute, vol. 9, no. 3, pp. 296, 2021.
- [3] Gupta, S., & Tripathi, P. (2016, February). An emerging trend of big data analytics with health insurance in India. In 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH) (pp. 64-69). IEEE.
- [4] N. Shakhovska, S. Fedushko, I. Shvorob and Y. Syrov, "Development of mobile system for medical recommendations," Procedia Computer Science, vol. 155, pp. 43–50, 2019.
- [5] D. B. Madan and K. Wang, "Option implied VIX, skew and kurtosis term structures," *International Journal of Theoretical and Applied Finance*, vol. 24, no. 5, Article ID 2150030, 2021.
- [6] M. hanafy and O. 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, 2021. Doi: 10.35940/ijitee.c8364.0110321.
- [7] Philipp Drewe-Boss, Dirk Enders, Jochen Walker and Uwe Ohler, "Deep learning for prediction of population health costs", *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, pp. 1-10, 2022.
- [8] Bhardwaj N, Delhi RA, Akhilesh ID, Gupta D (2021) Health insurance amount prediction [Online].
- [9] Panay B, Baloiian N, Pino J, Peñafiel S, Sanson H, Bersano N (2019) Predicting health care costs using evidence regression. *Proceedings* 31(1):74.
- [10] Junqueira ARB, Mirza F, Baig MM (2019) A machine learning model for predicting ICU readmissions and key risk factors: analysis from a longitudinal health record. *Health Technol. (Berl)* 9(3).

- [11] Kerrissey, M., Tietschert, M., Novikov, Z., Bahadurzada, H., Sinaiko, A. D., Martin, V., & Singer, S. J. (2022). Social features of integration in health systems and their relationship to provider experience, care quality and clinical integration. *Medical Care Research and Review*, 79(3), 359-370.
- [12] G. Kowshalya and M. Nandhini, "Predicting fraudulent claims in automobile insurance," in *Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp. 1338–1343, IEEE, Coimbatore, India, April 2018.
- [13] J. M. Johnson and T. M. Khoshgoftaar, "Medical provider embeddings for healthcare fraud detection," *SN Computer Science*, vol. 2, no. 4, pp. 1–15, 2021.
- [14] N. A. Akbar, A. Sunyoto, M. R. Arief, and W. Caesarendra, "Improvement of decision tree classifier accuracy for healthcare insurance fraud prediction by using Extreme Gradient Boosting algorithm," in *Proceedings of the International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, pp. 110–114, IEEE, Jakarta, Indonesia, November, 2020.
- [15] L. S. Chen and J. C. Chen, "Using data mining methods to detect medical fraud," in *Proceedings of the 2020 International Conference on Management of e-Commerce and e-Government*, pp. 89–93, Jeju Island, South Korea, July 2020.
- [16] J. Pesantez-Narvaez, M. Guillen, and M. Alcañiz, "Predicting motor insurance claims using telematics data-XGBoost versus logistic regression," *Risks*, vol. 7, no. 2, 2019.
- [17] M. A. Fauzan and H. Murfi, "The accuracy of XGBoost for insurance claim prediction," *International Journal of Advanced Software Computer Applications*, vol. 10, no. 2, 2018.
- [18] T. M. Alam, M. M. A. Khan, M. A. Iqbal, W. Abdul, and M. Mushtaq, "Cervical cancer prediction through different screening methods using data mining," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 2, 2019.
- [19] X. Yang, M. Khushi, and K. Shaukat, "Biomarker CA125 feature engineering and class imbalance learning improves ovarian cancer prediction," in *Proceedings of the IEEE Asia-Pacific Conf. on Computer Science and Data Engineering (CSDE)*, pp. 1–6, Gold Coast, Australia, December 2020.
- [20] K. Shaukat, F. Iqbal, T. M. Alam et al., "The impact of artificial intelligence and robotics on the future employment opportunities," *Trends in Computer Science and Information Technology*, vol. 5, no. 1, pp. 50–54, 2020.
- [21] M. U. Ghani, T. M. Alam, and F. H. Jaskani, "Comparison of classification models for early prediction of breast cancer," in *Proceedings of the International Conference on Innovative Computing (ICIC)*, Lahore, Pakistan, November.2019.
- [22] B. Panay, N. Baloian, J. A. Pino, S. Peñafiel, H. Sanson, and N. Bersano, "Predicting health care costs using evidence regression," *Multidisciplinary Digital Publishing Institute Proceedings*, vol. 31, no. 1, p. 74, 2019.
- [23] 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, pp. 131–220, 2018.
- [24] B. D. Sommers, "Health insurance coverage: what comes after the ACA?" *Health Affairs*, vol. 39, no. 3, pp. 502–508, 2020.