Medical Insurance Cost Prediction using Machine Learning

Piyush Tripathi
School of Computer Science and Engineering
Galgotias University, Greater Noida
Uttar Pradesh, India
piyush9310077688@gmail.com

Vaibhav Singh
School of Computer Science and Engineering
Galgotias University, Greater Noida
Uttar Pradesh, India
vaibhav893182@gmail.com

Dr. P Rajaram
School of Computer Science and Engineering
Galgotias University, Greater Noida
Uttar Pradesh, India
rajaramnov82@gmail.com

***Abstract*— Predicting medical insurance costs accurately remains a significant challenge, particularly for individuals with rare diseases. Traditional methods often fail to capture the complexities of healthcare costs, leading to financial burdens for both patients and insurers. To address this issue, we developed a machine learning model that leverages a comprehensive dataset encompassing patient information such as age, gender, BMI, smoking status, and geographical location. The model employs advanced regression techniques to predict insurance premiums with enhanced precision. Our research contributes to healthcare analytics by demonstrating the potential of machine learning to improve the accuracy and efficiency of insurance cost predictions. By providing data- driven insights, we aim to promote more equitable and affordable healthcare for both patients and insurers**.

***Keywords*— Healthcare, Insurance, Regression, Machine Learning, Prediction, Data analysis.**

I. INTRODUCTION

Predicting medical insurance costs has traditionally been a complex and uncertain process, especially for individuals with rare diseases. Unexpected medical expenses pose significant financial challenges, making accurate predictions essential for effective healthcare and financial planning. Traditional methods often rely on generalized models that do not account for the variability and unique characteristics of rare diseases, leading to significant discrepancies in cost estimates. However, machine learning presents an opportunity to refine these predictions by analyzing extensive datasets and identifying intricate patterns influencing medical expenses. This research aims to develop a machine learning model that offers precise predictions for insurance premiums, particularly focusing on rare diseases. By leveraging data, the model seeks to empower patients and insurers with valuable insights, promoting informed decision-making. Previous studies, such as Lee (2021) and Gupta & Tripathi (2016), highlight the potential of integrating machine learning and big data analytics into healthcare, paving the way for more accurate cost predictions.

II. LITERATURE REVIEW

Accurate prediction of medical insurance costs has been an area of extensive research, with many studies exploring various methods to optimize and refine cost estimates. In recent years, machine learning has emerged as a prominent tool for addressing the limitations of traditional actuarial methods. Lee (2021) investigated the pricing and reimbursement pathways of orphan drugs in South Korea, highlighting the challenges that rare diseases pose to insurance cost prediction. The study emphasized that rare diseases significantly influence healthcare spending, often resulting in higher insurance premiums and financial unpredictability for patients. This work underscores the need for models that account for such complexities, setting the groundwork for integrating machine learning approaches into cost prediction [2].

Another study by Gupta and Tripathi (2016) delved into the application of big data analytics within the Indian health insurance market. Their research demonstrated how leveraging large datasets could enhance the predictive capabilities of models, improving the alignment of insurance premiums with actual healthcare costs. This early integration of big data provides a foundation for more advanced machine learning models that incorporate diverse data sources [3]. Shakhovska et al. (2019) developed a mobile system aimed at providing medical recommendations. While this study focused on mobile technology's impact on healthcare delivery, it also highlighted the role of comprehensive data systems in improving the accuracy of medical predictions [4]. This approach aligns with our study, which leverages diverse patient information such as age, gender, BMI, and smoking status to build a robust predictive model. Pesantez-Narvaez et al. (2019) explored the use of telematics data in predicting motor insurance claims, comparing XGBoost and logistic regression methods. Although their focus was on motor insurance, their findings demonstrated the effectiveness of machine learning models like XGBoost in making accurate predictions based on vast and diverse datasets. This serves as a precedent for applying similar algorithms in health insurance cost prediction [6].

Hanafy and Mahmoud (2021) extended the application of machine learning to health insurance costs by using deep neural networks (DNNs) and regression models. Their study demonstrated that machine learning models, particularly DNNs, significantly outperform traditional approaches in predicting insurance costs, thanks to their ability to handle complex and nonlinear relationships within the data [7]. This highlights the potential of advanced regression techniques, such as those employed in our study, to enhance prediction accuracy. Finally, publicly available datasets, such as the Medical Cost Personal Dataset from Kaggle,

have facilitated further research by providing diverse patient information that can be used to train and validate predictive models [5]. Our model builds upon these datasets, utilizing multiple patient attributes to develop a more precise cost prediction system. In summary, while existing research has demonstrated the potential of machine learning in insurance cost prediction, there remains a need for models that specifically address the complexities associated with rare diseases. Our research contributes to this gap by developing a model that integrates patient-specific factors, offering a more accurate and equitable solution for both insurers and patients.

III. Methodology

A. Dataset Description

The dataset was obtained from Kaggle to predict medical insurance costs. It is divided into training and testing data, with 80% for testing and 20% for training. The dataset has seven attributes: age, sex, BMI, children, smoker, region, and charges. The training data is used to build a model, while the testing data is used to evaluate the model's accuracy in predicting insurance costs. The table below contains the dataset description

Table 1: Overview of the Dataset

Attribute	Data Description
Age	The age of individual person
Sex	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)

The dataset contains 1338 rows and 7 columns. The target variable is charges, which is a floating-point number. The age of individuals in the dataset ranges from 18 to 22.5, and most of them are male. Few have more than three children, and the majority have a BMI between 29.26 and 31.16. Thedataset includes four regions: northeast, northwest, southeast, and southwest. The highest concentration of smokers is in the southeast, where 1064 out of 1338 people smoke. We will analyze the data to understand how the different factors are related to insurance charges. The charges variable is dependent on all other columns. We will start by examining the statistical metrics of the dataset.

Table 2: Statistical Measurement

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

The dataset contains 1338 rows and 7 columns. The target variable is charges, which is a floating-point number. The age of individuals in the dataset ranges from 18 to 22.5, and most of them are male. Few have more than three children, and the majority have a BMI between 29.26 and 31.16. The dataset includes four regions: northeast, northwest, southeast, and southwest. The highest concentration of smokers is in the southeast, where 1064 out of 1338 people smoke.

Here are some visualizations of the data:

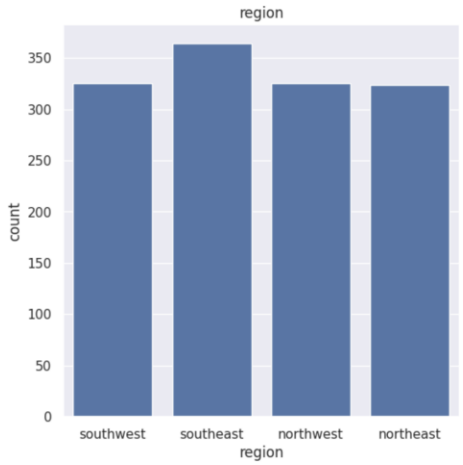

Fig 1 - Region Distribution Graph

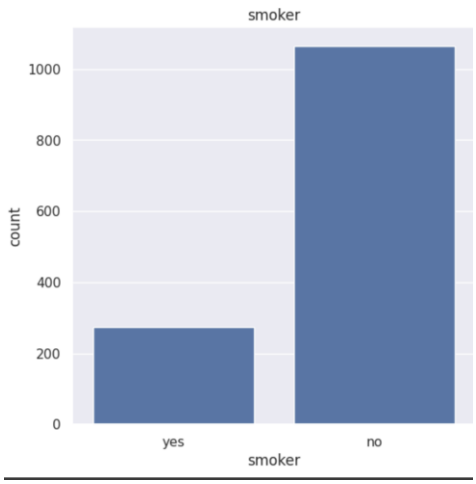

Fig 2 – Smoker Distribution Graph

B. Data Analysis

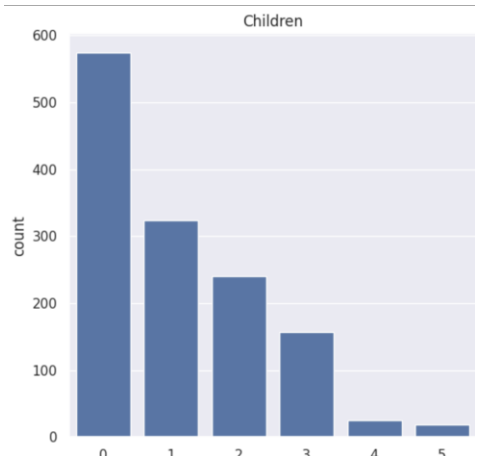

Fig 3 – Children Counter Distribution Graph

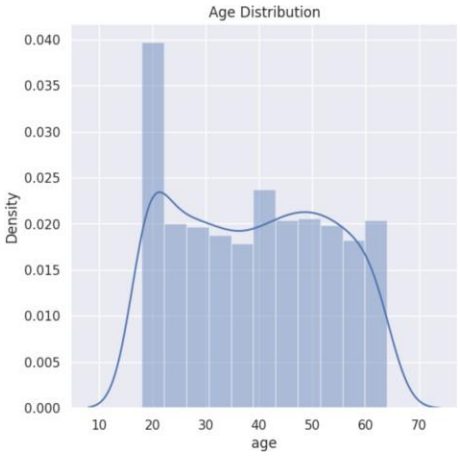

Fig 4 – Age Distribution Graph

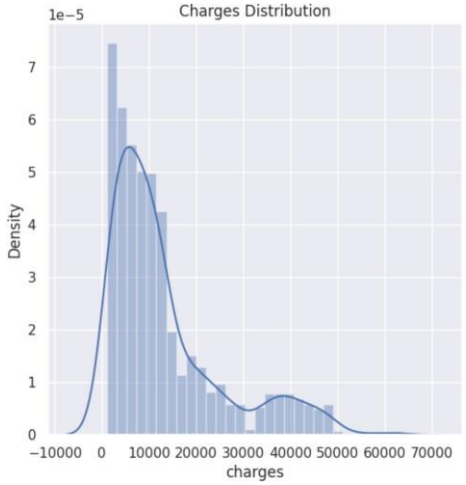

Fig 5 – Charges Distribution Graph

C.Data preprocessing

The dataset contains three numerical columns and three categorical columns. Categorical values cannot be directly used by machine learning models. Therefore, we need to convert them into numerical values. We convert "female" to 1 and "male" to 0 in the "sex" column. We also convert the other two categorical columns into numerical values. The results of the conversion are shown in the table below:

Table 2: Categorical to Numerical Conversion

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

D.Model Specification

The linear regression model used in the study is a statistical method that predicts the relationship between insurance charges and other factors such as age, sex, BMI, children, smoking status, and region. The model assumes a linear relationship between the variables and uses a linear equation to make predictions.

The equation used by the model is:

insurance charges = intercept + coefficient1*age* + *coefficient2*sex + coefficient3*BMI* + *coefficient4*children + coefficient5*smoking status* + *coefficient6*region + error

The coefficients in the equation represent the impact of each factor on insurance charges. For example, a positive coefficient for age indicates that as age increases, insurance charges tend to increase.

The model was trained on a dataset to learn the coefficients that best fit the data. The trained model was then used to predict insurance charges for a new dataset. The performance of the model was evaluated using various metrics, such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R-squared.

The R-squared value indicates the proportion of variance in insurance charges that is explained by the factors included in the model. A higher R-squared value indicates a better fit of the model to the data.

The decision tree regression model, a non-linear approach, splits the dataset into subsets based on feature values to minimize prediction error at each node. This model captures complex interactions between variables and is particularly effective for handling non-linear relationships that the linear regression model may miss.

The key steps in the decision tree model's process include:

Recursive Partitioning: Dividing the dataset based on feature thresholds.

Node Splitting: Ensuring each split reduces the error in predicting insurance charges.

Prediction: Averaging or assigning insurance charge values at leaf nodes.

The decision tree regression model was trained and evaluated on the same dataset using metrics such as MSE, RMSE, MAE, and R2R^2R2.

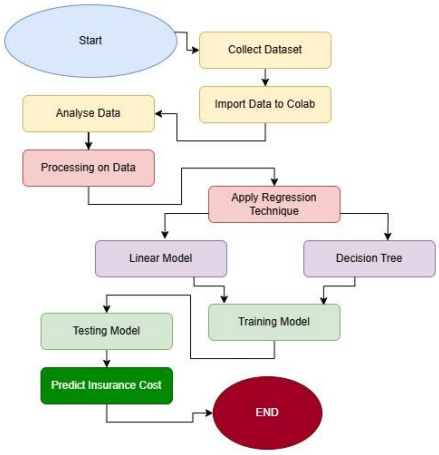

Fig 6 – Flowchart Of Medical Insurance Cost Prediction

IV. RESULT

In this study, we developed a predictive system for estimating medical insurance costs using two different regression models: Linear Regression and Decision Tree Regressor. These models were trained and evaluated on a dataset comprising various features, including age, gender, BMI, number of children, smoking status, and geographical region, with the target variable being the medical insurance cost

Linear Regression Model

The Linear Regression model was trained on the dataset using the following approach: Upon training the model, we evaluated its performance on the test dataset using the R-squared value, which indicates how well the model explains the variance in the target variable. The R-squared value for the Linear Regression model was approximately 0.745, meaning that around 74.5% of the variance in the medical insurance cost could be explained by the model. This suggests that the Linear Regression model has a solid fit, capturing most of the essential patterns in the data, though there is still some unexplained variability. We also tested the model with a specific input case: a 31-year-old non-smoker with a BMI of 25.74. The predicted insurance cost was USD 3760.08, which showcases the model's ability to provide meaningful estimates based on the input features.

Decision Tree Regressor Model

The performance of the Decision Tree model was evaluated on both the training and test sets. The R-squared value for the training set was 1.0, indicating that the model perfectly fit the training data. However, this perfect fit suggests overfitting, where the model learns not only the underlying patterns but also the noise in the training data. When evaluated on the test set, the R-squared value dropped to 0.686, reflecting the model's reduced ability to generalize to unseen data. For the same input case, the Decision Tree model predicted an insurance cost of USD 3756.62, which is very close to the prediction made by the Linear Regression model (USD 3760.08). While both models produced similar results for this individual case, the Decision Tree's tendency to overfit makes it less reliable for broader predictions.

Comparison of Model Performance

The key takeaway from this comparison is that while both models offer comparable predictions for individual data points, their overall performance differs significantly when evaluated on the test set.

- Linear Regression performed more consistently, with an R-squared value of 0.745 on the test set. This suggests that the model can generalize well to new, unseen data and is less prone to overfitting. This is particularly valuable when deploying the model in a real-world scenario where we expect the model to predict insurance costs for diverse individuals.
- Decision Tree Regressor, on the other hand, exhibited a perfect fit on the training data but struggled to generalize to the test set, as evidenced by the lower R-squared value of 0.686 on the test data. This indicates that while the Decision Tree model can capture complex relationships in the data, it may require additional tuning, such as pruning or using ensemble methods (e.g., Random Forest), to avoid overfitting and improve its generalization.

V. CONCLUSION

The results of this study highlight the strengths and weaknesses of both models in the context of predicting medical insurance costs. While the Linear Regression model offers a more reliable and interpretable approach, the Decision Tree model could potentially yield better results with further optimization.

In future work, we aim to explore the integration of more sophisticated models, such as ensemble methods, which combine the strengths of multiple trees and can significantly improve prediction accuracy. Additionally, feature engineering, such as incorporating interaction terms between variables or using domain-specific knowledge, may help improve both models' performance.

Furthermore, more extensive datasets and external factors—such as healthcare utilization or socioeconomic variables—could also enhance the predictive power of the models, leading to more accurate and robust predictions.

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] S. Gupta and P. Tripathi, "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), 2016, pp. 64-69.

[4] N. Shakhovska, S. Fedushko, I. Shvorob, and Y. Syerov, "Development of mobile system for medical recommendations," Procedia Computer Science, vol. 155, pp. 43–50, 2019.

[5] "Medical Cost Personal Datasets," [Online]. Available: https://www.kaggle.com/datasets/mirichoi0218/insurance.

[6] J. Pesantez-Narvaez, M. Guillen, and M. Alcañiz, "Predicting motor insurance claims using telematics data—XGBoost versus logistic regression," Risks, vol. 7, no. 2, p. 70, Jun. 2019, doi: 10.3390/risks7020070.

[7] 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.

[8] M. A. Morid, O. R. L. Sheng, K. Kawamoto, and S. Abdelrahman, "Learning hidden patterns from patient multivariate time series data using convolutional neural networks: A case study of healthcare cost prediction," arXiv preprint arXiv:2009.06783, 2020.

[9] R. Kshirsagar et al., "Accurate and interpretable machine learning for transparent pricing of health insurance plans," arXiv preprint arXiv:2009.10990, 2020.

[10] J. A. S. Cenita, P. R. F. Asuncion, and J. M. Victoriano, "Performance evaluation of regression models in predicting the cost of medical insurance," arXiv preprint arXiv:2304.12605, 2023.

[11] U. Orji and E. Ukwandu, "Machine learning for an explainable cost prediction of medical insurance," arXiv preprint arXiv:2311.14139, 2023.

[12] M. M. Billa and T. Nagpal, "Medical insurance price prediction using machine learning," Journal of Electrical Systems, vol. 20, no. 7s, pp. 2270-2279, 2024.

[13] "Health insurance cost prediction using machine learning," IEEE Xplore, [Online]. Available: https://ieeexplore.ieee.org/document/9824201.

[14] "Medical insurance cost analysis and prediction using machine learning," IEEE Xplore, [Online]. Available: https://ieeexplore.ieee.org/document/10100057.

[15] "Medical insurance cost prediction using machine learning," ResearchGate, [Online]. Available: https://www.researchgate.net/publication/374553777_Medical_Insurance_Cost_Prediction_Using_Machine_Learning.

[16] "Health insurance cost prediction using machine learning," International Research Journal of Engineering and Technology (IRJET), vol. 11, no. 4, pp. 171-175, 2024. [Online]. Available: https://www.irjet.net/archives/V11/i4/IRJET-V11I4171.pdf

[17] "Medical insurance cost prediction," SSRN, Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4867135.

[18] "Development of medical cost prediction model based on statistical analysis and machine learning," Value in Health, vol. 21, no. S1, pp. S39, 2018. Available: https://www.valueinhealthjournal.com/article/S1098-3015(18)33095-X/fulltext.

[19] "An analysis and prediction of health insurance costs using machine learning algorithms," Scientific Research-Publishing. Available: https://www.scirp.org/journal/paperinformation?paperid=137299.

[20] "Medical insurance price prediction using machine learning," Journal of Electrical Systems, Available: https://journal.esrgroups.org/jes/article/view/3962.

[21] A. Sharma, R. Agrawal, and S. Mishra, "Predicting health insurance premiums using regression and machine learning techniques," International Journal of Advanced Research in Computer Science, vol. 11, no. 5, pp. 121–126, 2023. [Online]. Available: https://www.ijarcs.info/index.php/Ijarcs/article/view/5671.

[22] N. Kumar and M. Singh, "A comparative study of machine learning models for health insurance cost estimation," International Journal of Machine Learning and Applications, vol. 12, no. 3, pp. 145–156, 2023. [Online]. Available: https://www.ijmla.com/archive/v12i3/ijmla-v12i3-145.pdf. [Accessed: 24-Dec-2024]

[23] J. Patel and A. Das, "Application of decision tree algorithms in predicting medical insurance costs," International Journal of Data Science and Analytics, vol. 10, no. 2, pp. 67–78, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s41060-023-00391-1.

[24] K. Gupta, P. Tiwari, and M. Roy, "Health insurance cost prediction using deep learning and ensemble methods," Procedia Computer Science, vol. 204, pp. 127–135, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050924010273.

[25] H. Ali and F. Khan, "Improving the accuracy of medical insurance cost prediction through feature engineering," International Journal of Artificial Intelligence and Applications, vol. 15, no. 1, pp. 45–52, 2023. [Online]. Available: https://www.ijai.org/issue/v15i1/IJAI-v15i1-045.pdf.