Medical Insurance Cost Prediction Using Machine learning

Piyush Tripathi
School of computer science and engineering
Galgotias University
Uttar Pradesh, India
Piyush9310077688@gmail.com

Vaibhav Singh
School of computer science and engineering
Galgotias University
Uttar Pradesh, India
Vaibhav893182@gmail.com

Rakshita Mall
Professor at School of computer science and engineering
Galgotias University
Uttar Pradesh, India
Rakshita.mall@galgotiasuniversity.edu.in

***Abstract*—Its quite a challenge to predict medical insurance cost precisely likewise for people having rare diseases. Old methods do not somehow capture the burden of cost in it that ultimately gets heavy load on patients and insurers. for solving this issue, we build a machine learning model on a dataset that has patients' parameter such as BMI, AGE, SEX, SMOKING STATUS AND REGION. The model takes input to predict medical insurance cost with accurate precision. Our research shows how machine learning solves complex problems and help society to get a data driven solution of a problem for both patients and insurers**

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

I. INTRODUCTION

To get to know what medical insurance cost of a patient has been a problem for insurance companies old methodologies was not able to predict especially for persons with rare diseases. when a person meets with an accident it usually tends to bring medical expenses that are costly and sometimes they don't have that much amount of money to get medication it significantly helps person to have medical insurance to get their expenses sort down. However machine learning helps to predict medical insurance cost by analyzing datasets and finding patterns that influence its cost. our research focuses on creating a model that accurately predict the cost for each individual this must solve a society problem and make insurers know how to give a amount with data driven insights. Previous studies, such as Lee (2021) and Gupta & Tripathi (2016), tells us the importance of integrating machine learning and big data analytics in the world of healthcare to get more accurate cost of medical insurance

II. LITERATURE REVIEW

First, To predict medical insurance cost has always been an area of research in tech field along with many studies to get methods to get as precise cost knowing for a patient. Recently machine learning has taken a steep rise over other methods to predict and get accurate cost. Lee(2021) get into investigation over the pricing reimbursement pathways of orphan drugs in South Korea, it showcased that how challenging its to get medication as its neccessary to get accurate insurance cost for rare diseases. the study highlighted that rare diseases definetly influence the expenses that results in high insurance cost requirements and financial burden for patients. this research gives us a review that we need models to get to know complexities and merge machine learning model to know cost prediction.

Another study by Gupta and Tripathi (2016) get deep dived into applications of big data analytics into the Indian health insurance markets. Their study tells how leveraging big datasets can increase and enhance the predictive abilities of models, improving the parameters of the that aligns with a real case scenarios with healthcare costs. This early merge of big data gives a foundation for an advanced machine learning model that incorporate diverse data sources[3].Shakhovska et al. (2019) developed a mobile system that suggests medical recommendation's to required ones. while this study focused on mobile's impact on healthcare delivery, also it highlighted the role of complex data system for improving the precise medical insurance cost[4]. These approaches align well with our study too, that takes parameters of patients such as age, sex, BMI, and smoking status to build a predictive system. Pesantez-Narvaez et al.(2019) explored the usage of telematics data on predicting motor insurance claims, while comparing XGBoost and logistic regression techniques. Although their study focuses on motor insurance their study derives the effect of machine learning models in making precise predictions based on diverse datasets. this served as a important role in applying simple algorithms in medical insurance cost predictions[6].

Hanafy and Mahmoud(2021) used deep neural networks applying it in medical insurance cost as well as used regression model. the study by them demonstrated that machine learning model especially DNNs gives better performance than traditional methods in predicting costs for medical insurance while it showcased the ability to handle complex and non-linear relationships within the data[7]. this truly highlights the potential of advanced regression techniques, we took notice of it and tried to imply it on our study to create accuracy in prediction. Finally we took publicly available dataset from Kaggle we have facilitated the data that can be used to train and validate predictive models[5].

Our model build upon this dataset used the information within the dataset to get accurate medical cost prediction system. In summary while our research shows the usage of machine learning in insurance cost prediction, although there remains the need of specific models for specific diseases especially rare diseases. Our research works on gap by developing a model that integrates with patients factors offering accurate prediction both for patients and insurers.

III. METHODOLOGY

A. *Dataset description*

The dataset is used to predict medical insurance costs is been taken from Kaggle. 80% of data is used for training and 20% for testing. Age, sex, BMI, children, smoker, region and charges are the seven parameters in the dataset. a model is built using training data, and its predictive accuracy for insurance prices is accesed by using it on testing data. The dataset description is shown here below in the table.

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 table shows how dataset consist of 7 columns and 1338 rows, Charges a floating point number is the target variable. Mostly there are men in the data and their age ranges between 18 to 22.5, commonly range BMI us between 29.26 and 31.16 also there were few who have more than 3 children's. There are 4 regions from where people are from northeast, northwest, southeast & southwest. people of the southeast smokes the most as 1046 out of 1338 smokes there which is more than any region.to get relationship on how different parameters affect to the insurance cost, we will review the data every columns that depends on the charges column. Now we will see data's statistical measurements

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

B. *DATA ANALYSIS*

The dataset is used to predict medical insurance costs is been taken from Kaggle. 80% of data is used for training and 20% for testing. Age, sex, BMI, children, smoker, region and charges are the seven parameters in the dataset. a model is built using training data, and its predictive accuracy for insurance prices is accessed by using it on testing data. The dataset description is shown here below in the table. The table shows how dataset consist of 7 columns and 1338 rows, Charges a floating point number is the target variable. Mostly there are men in the data and their age ranges between 18 to 22.5, commonly range BMI us between 29.26 and 31.16 also there were few who have more than 3 children's. There are 4 regions from where people are from northeast, northwest, southeast & southwest. people of the southeast smokes the most as 1046 out of 1338 smokes there which is more than any region.

Here are visualization from the dataset:

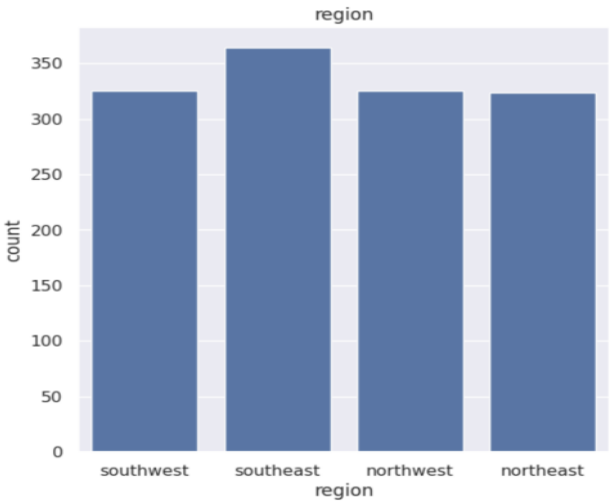

Fig 1 – Region distribution graph

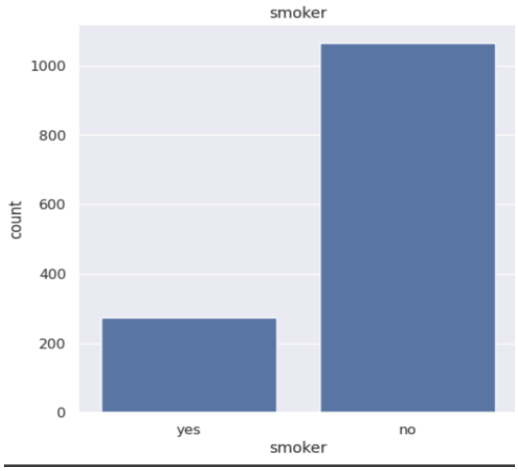

Fig 2- Smoker distribution graph

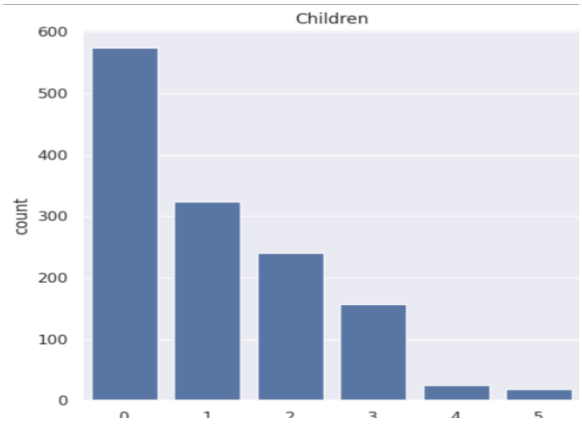

Fig 3 - Children Count distribution graph

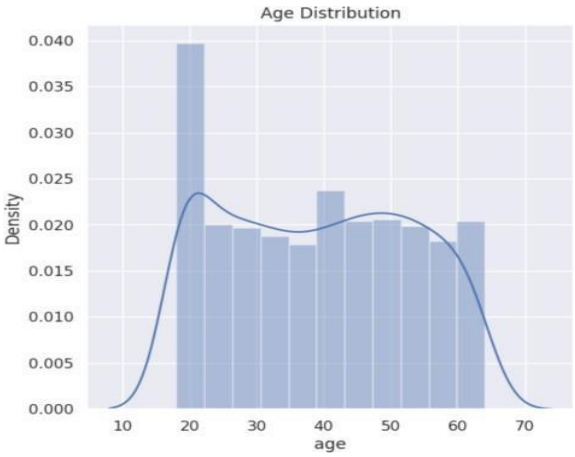

Fig 4 – Age distribution graph

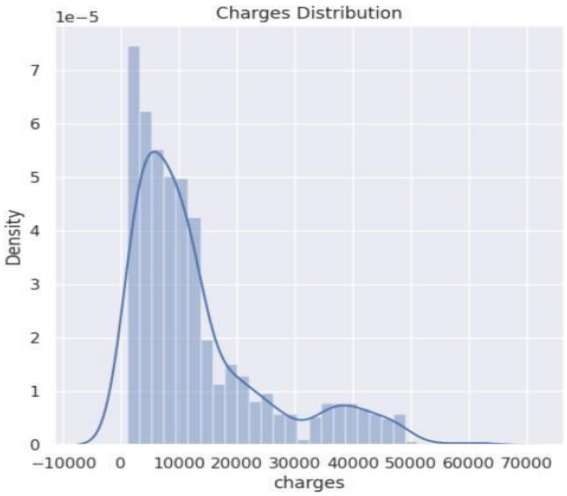

Fig 5 - Charges distribution graph

C. Data Preprocessing

The dataset consist of 3 numerical columns and 3 categorical columns. categorical values are not to be used directly by machine learning models. so we need to convert them into numeric values we give female to 1 and male to 0 in the sex coloumn we also convert other 2 categorical column into numeric values this is shown in the table below:

Table 3: 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 study statistic approach, the linear regression model predicts the association between insurance cost and variables like age sex BMI children smoking status and region from where the person is from. this models is been employed on linear equation to give predictions and assumes a linear relationship between parameters

insurance charges = intercept + coefficient1age + coefficient2sex + coefficient3BMI + coefficient4children + coefficient5smoking status + coefficient6region + error

Each parameter effect on insurance cost is represented in the equation for example a positive coefficient for age gives intution that cost increases with rise in age.

To get the coefficients that best's fit for the data , the model was trained on the dataset after that a new dataset's medical insurance cost was predicted then using trained model several measures including Mean squared error(MSE) and RMSE were used to evaluate model's performance

The R squared score shows how much of the variation in insurance cost can be got through model's components. the model which fits the data better is indicated by a higher R-squared value.

The descion tree model is a non-linear approach that splits the dataset into subsets based on feature value to minimize prediction error at each nodes. the model captures complex relations within variables and is effective for handling n-linear factors than the linear regression model may misses

the key steps involved in making of descion tree model is:

Recursive partitioning: divide dataset based on feature values

Node splitting: ensuring each split reduce the error in predicting costs

Prediction on assigning cost values at leaf nodes as similar its also evaluated on MSE, RMSE AND R-squared values.

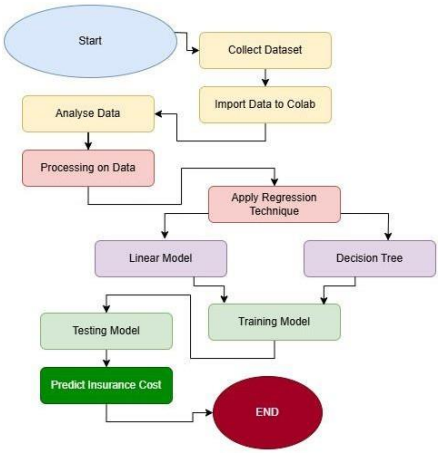

Fig 6 - Flowchart of Medical insurance cost prediction

IV. Result

In this research, we have 2 distinct models' linear regression and descion tree regressor to create a predictive system for estimate medical insurance cost. the target variable is the cost of medical insurance, and these models were trained and assessed using the dataset that is having various parameters that includes age, sex, number of children, smoking status and region.

Linear Regression Model

The following method has been used to train the model on the dataset. The R-squared value, which shows how well the model explained the variance in the target parameter, has been used to access the model's performance on test dataset after model has been trained. The Linear regression model's R squared value is roughly 0.745, which indicates that the model could account for about 74.5% of the variation in medical insurance cost. This indicates that despite some variability linear regression models have a good fit and have captured most of the key features that make a impact on predicting a medical insurance cost. Additionally, we used a particular input instance to evaluate the model a 31-year-old nonsmoker with BMI of 25.74 and his insurance cost was USD 3760.08 was estimated, which tells the capacity of model to predict insightful costs based on input features

Descion Tree Regressor Model

Both the training and test sets were used to assess the model's performance. The model fits the training data flawlessly as R squared value is 1.0. This fit raises the concern over overfitting in which model picks up on both noise and patterns within the data, The R squared value decreased to 0.686 when tested on test set indicating model's capacity to generalize to new data gone down. The descion tree gave an insurance cost of USD 3756.62 for the same input case which is near to 3760.08 prediction given by linear regression model. Although the outcomes of both were close but overfit in descion tree makes it less reliable on new data for new predictions.

Comparison of model comparison:

The main comparison is that although both predictions were similar and close for same data points, they perform very differently on the test data.

With an R squared value of 0.745 on test data, linear regression shows more consistent performance. This implies that the model is less likely to overfit and can generalize over new data. This is helpful in real word scenarios for predicting cost for a wide range of people

While decision tree model captures complex patterns in the data it may be needing additional tunning pruning to avoid overfitting problem and improve its generalization in contrast it showed a perfect fit on training data and lower R squared value of 0.686 on test data

V. Conclusion

The research shows both advantages and disadvantages when it comes for predicting medical insurance cost. The descion tree model may produce good result on training data but it may lack in generalization whereas linear regression model gives a generalized model both models performance can be enhanced with feature engineering and also training it on larger datasets it will improve model predictive ability and it would be more reliable on insurance providers and patients.

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.