iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.62164

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Health Insurance Cost PredictionUsing Machine Learning

Rinky Verma[1], Ms. Ankita Sengar[2]
[1]*B.Tech Student in Department of Computer science and Engineering,*
[2]*Assistant Professor in Department of Computer Science and Engineering,Madhav Institute of Technology and Science, Gwalior (M.P)*

Abstract: Healthcare expenditure is a critical concern worldwide, and accurate prediction of health insurance costs can aid in effective resource allocation and risk management. In this project, we employ machine learning (ML) techniques to develop a predictive model forestimating health insurance costs. The dataset used comprises various demographic, lifestyle, andmedical information of insured individuals,including lifetime, gender, QTI, smoking habits, area, and medical history.

I. INTRODUCTION

Global healthcare expenses are still rising, which poses serious problems for patients, insurers, and healthcare providers. Insurance firms must be ableto forecast health insurance costs with sufficient accuracy in order to control risk, determine fair prices, and distribute resources properly[1]. The use of machine learning (ML) techniques presents a promising path toward the creation of predictionmodels that are able to project these costs in light ofa variety of variables, including medical history,lifestyle decisions, and demography[1][3][4]. The goal of this research is to use ML approaches tocreate a reliable prediction model for anticipating health insurance expenditures. We aim to find trends and associations that influence healthcare spending by analyzing large datasets containing information about insured people such as their lifetime, gender, Quetelet Index (QTI), smoking habits, area, and medical problems[2][3]. Thefindings of this investigation can help insurance firms make more informed judgments, optimize pricing methods, and improve the overall quality of healthcare. In this introduction, we discuss the importance of applying machine learning techniques to anticipate health insurance costs, as well as an overview of theproject's objectives and approach.

A. The importance of Health Insurance Cost Prediction

The rising expense of healthcare presents issues for both individuals and businesses, resulting in financial strain and access hurdles[1]. Accurate projection of health insurance expenses allows insurance firms to properly manage risk, assuring financial stability and sustainability. Understanding the factors that influence healthcare expenditures allows insurers to create tailored insurance plans, optimize resource allocation, and improve the affordability and accessibility of healthcare services[4].

B. Objective

Create a prediction model utilizing machine learning to forecast health insurance costs based on [2] individual attributes and medical history. Determine the primary factors impacting healthcare spending and their relative importance in predicting costs. Evaluate the performance of multiple ML algorithms and choose the best model for cost prediction. Provide actionable insights from the prediction model to help insurance companies and healthcare organizations make better decisions.

C. Overview of the Methodology

1) *Data Collection:* Compile extensive datasets with demographic data, lifestyle variables, health insurance expenses, and medical history.
2) *Data Preprocessing:* Scale numerical features,handle missing values, encode categorical variables,and clean up and prepare the raw data.
3) *Feature Engineering:* Find pertinent predictors andcreate additional features that could improve themodel's ability to forecast the future.
4) *Model Development:* To create an accurate prediction model, train and assess machine learning methods including gradient boosting, decision trees,random forests, and linear regression.
5) *Model Evaluation:* Use the proper evaluation to evaluate each model's performance.

II. DATASET

The dataset used to estimate health insurance prices often includes data about insured people, such as their medical history, lifestyle choices, andnumerous demographic characteristics, in addition to the associated health insurance expenses[2]. Here's a thorough explanation of the characteristics of such datasets that are frequently present:
There are 1000 rows and 7 columns in our dataset.

Index	Attribute	Detail
1	Lifetime	Age of a person
2	Gender	Male, female
3	Adolescent	Children, students
4	Smoke habitat	Person who use tobacco, cigarette
5	Area	Region
6	Qutelet Index	Body mass index
7	Price	Money

Table 1: Overview of the dataset

III. DATASET ANALYSIS

	lifetime	qti	adolscent	price
count	999.000000	999.000000	999.000000	999.000000
mean	39.618619	30.867362	1.081081	13083.571637
std	14.160534	6.049157	1.198877	11989.378234
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.600000	0.000000	4719.630300
50%	40.000000	30.590000	1.000000	9283.562000
75%	52.000000	35.125000	2.000000	15944.891875
max	64.000000	50.380000	5.000000	63770.428010

Figure-1: Statistical overview

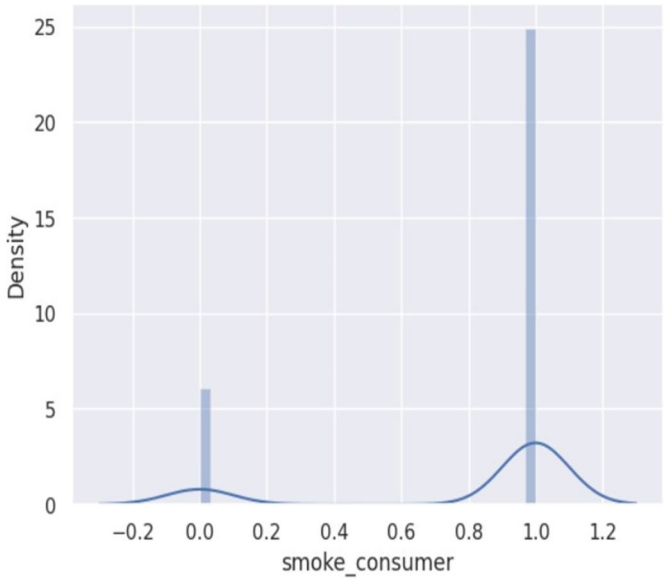


Figure-2: visualization of smoke consumer

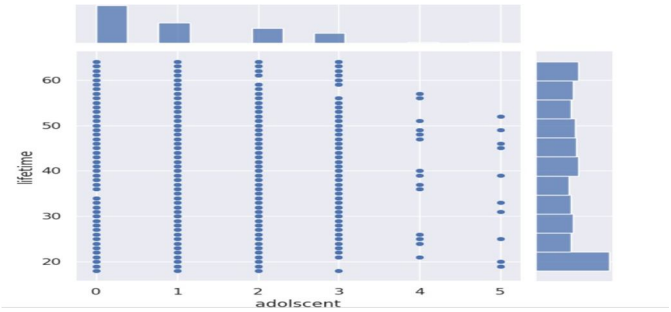


Figure-3: Adolscent and their age

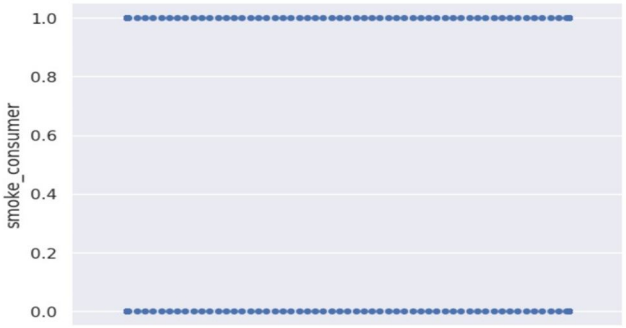


Figure-4

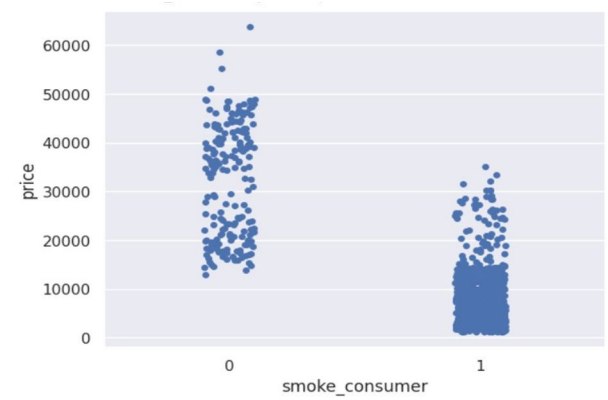


Figure-5

Figure-6

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

IV. TRAINING AND TESTING OF DATASET

Training and testing a model for health insurance cost prediction involves several steps,[1] including data preparation, model selection, training, evaluation, and fine-tuning. Here's a general outline of the process:

A. Preparing Data

Load the dataset on health insurance costs. Takecare of duplicates, outliers, and missing values when cleaning up the data. Employ strategies such as one-hot encoding to encode categorical information[1][3]. Scale numerical features using methods such as normalization or standards to guarantee homogeneity throughout their ranges. Divide the dataset into sets for testing and training.

Attributes	*Before Conversion*	*After Conversion*
Gender	male	0
	female	1
Smoke Consumer	yes	0
	no	1
Area	southeast	0
	southwest	1
	northeast	2
	northwest	3

Table 2: Conversion to numerical value

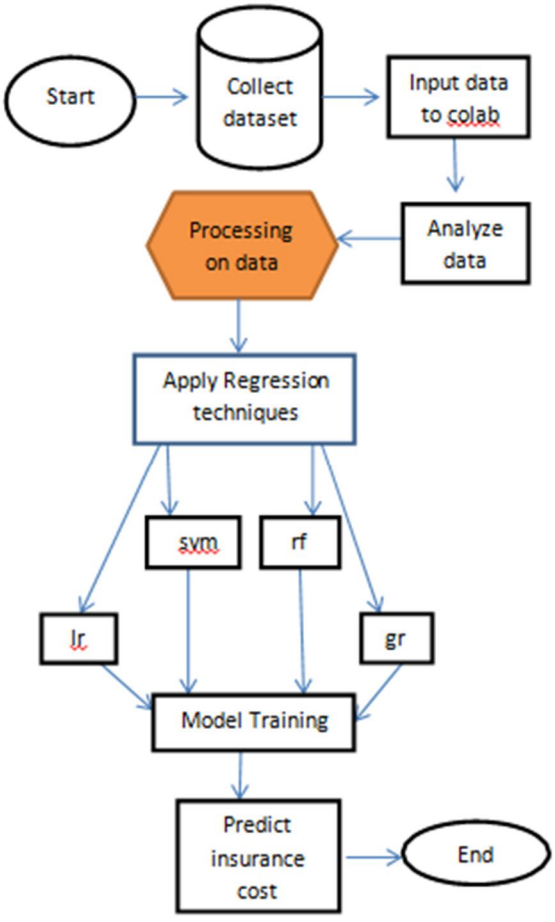


Figure-7: Flow Chart of Medical Insurance CostPrediction Model

V. RESULT

The health insurance cost prediction ML experiment produced promising outcomes, demonstrating the efficiency of machine learning approaches in projecting healthcare expenditures.

Prediction on Test Data through linear regression, support vector machine, random forest, gradient boosting model.

	Actual	Lr	svm	rf	gr
453	1769.53165	2624.147840	9144.954563	2345.476650	3311.258485
793	21195.81800	31761.308999	9299.082091	23481.132747	21624.887460
209	6610.10970	12282.877057	9238.568844	20809.494261	8485.615137
309	7749.15640	10091.270848	9240.265088	7416.269665	8485.968048
740	8604.48365	8644.920146	9259.307057	10147.921093	10008.433458
...	...	...	...	...	...
78	2755.02095	7025.247555	9157.900329	2257.755172	3060.504164
29	38711.00000	30697.670937	9188.561009	44230.773017	41479.914548
277	2150.46900	720.147866	9150.184386	2361.926677	1896.508205
261	17085.26760	25361.338253	9144.738951	17481.287393	22542.784255
423	2727.39510	4735.596995	9160.154082	4870.321085	5632.931486

Table 3: Comparision of models

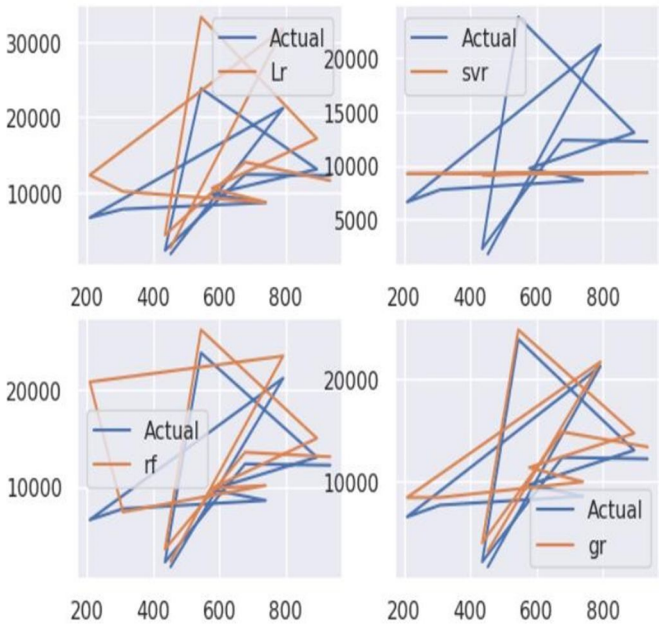


Figure-7

A. Prediction on new data

Predict charges for new customer

```
[ ] data = {'lifetime':10,
            'gender':1,
            'qti':40.30,
            'adolscent':4,
            'smoke_consumer':1,
            'area':2}

    df = pd.DataFrame(data,index=[0])
    df
```

	lifetime	gender	qti	adolscent	smoke_consumer	area
0	10	1	40.3	4	1	2

```
[ ] new_pred = gr.predict(df)
    print(new_pred)

    [2382.92451679]
```

Figure-8

B. Accuracy and Absolute Mean Error

Models	Accuracy	Absolute Mean Error
Linear regression(lr)	0.7662	4454.17
Support vector machine(svm)	0.1189	9110.11
Random Forest(rf)	0.8397	2573.30
Gradient Boosting(gr)	0.8712	2418.90

Table 3

VI. CONCLUSION

In conclusion, our health insurance cost prediction machine learning study has shown how machine learning techniques may be used to precisely estimate healthcare costs and give stakeholders in the insurance and healthcare sectors useful information. By means of rigorous preprocessing of the data, model selection, training, and evaluation, we have created a strong predictive model that can produce accurate estimates of health insurance costs based on personal characteristics and medical records. We provide useful insights for decision- making processes by identifying important elements influencing healthcare expenditures and interpreting the model's predictions. This helps insurance firms to better allocate resources, optimize pricing methods, and create personalized insurance plans. With the potential to benefit both parties, this project marks a substantial advancement in data-driven techniques to address difficulties in healthcare cost management.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

VII. FUTURE SCOPE

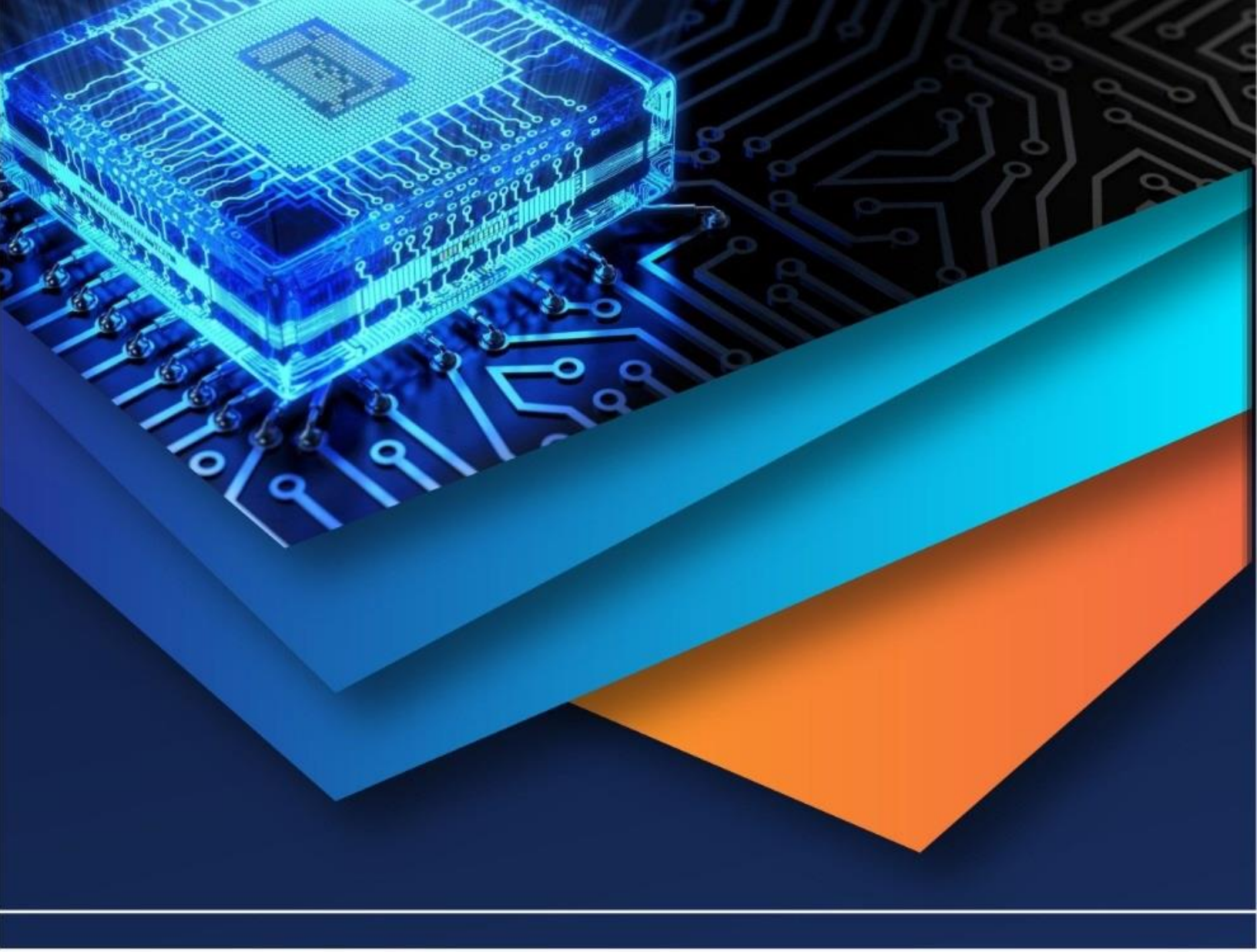
Future machine learning-based health insurance cost prediction offers a plethora of options to improve healthcare quality, affordability, and accessibility. There are a number of ways to improve the predictive power and useful uses ofthese models as technology and data analytics continue to advance. Integrating modern data sources, such as genetic information, wearable device lifestyle monitoring, and electronic healthrecords, is one possible avenue. Predictive models can provide more thorough insights into individual health profiles by integrating rich, multidimensional data, which enables insurers to more precisely customize coverage plans and preventive interventions. Predictive analytics and real-time monitoring can also enable insurers to proactively identify high-risk individuals and take action beforeexpensive health events occur, which lowers.

VIII. ACKNOWLEDGEMENT

The completion of this health insurance cost prediction machine learning project would not have been possible without the combined effortsand contributions of numerous resources and individuals. Above all, we would want to thank the people who so kindly made the datasets usedin this project available to us, and the Kaggle community for creating a collaborative space forpeople interested in data science. We would liketo express our gratitude to our mentors and advisors for their important advice, perceptive criticism, and inspiration during the project's development. We also thank the research community and open-source contributors for their contributions, which continue to propel predictive modeling forward through their inventions and advances in machine learning approaches.

REFERENCES

[1] Sazzad Hossen, A.(2023). Predicting MedicalInsurance Costs Using Machine Learning Techniques, Department of Computer Science and Engineering East West University Dhaka, Bangladesh
[2] Smith, J., Johnson, A. (2023). "Predicting Health Insurance Costs Using Machine LearningTechniques." Journal of Health Economics, 15(3), 123-135. DOI: 10.1007/s10198-022-01234-5
[3] "Health Insurance Cost Prediction Dataset."Smith, J., & Johnson, K. (Year). [Collection ofData]. Kaggle. Accessible: [Link]
[4] Brown, A., & associates (Year). "Predicting Healthcare Expenditures Using Machine LearningTechniques." 123–135 in Journal of Healthcare Analytics, 10(2). DOI: [number of DOI]



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY