Medical Insurance Cost Prediction Using Machine Learning

Thesis · October 2023

DOI: 10.13140/RG.2.2.31456.25604

1 author:

Sazzad Hossen
East West University (Bangladesh)
16 PUBLICATIONS **1** CITATION

SEE PROFILE

Medical Insurance Cost Prediction Using Machine Learning

Sazzad Hossen
Department of Computer Science and Engineering
East West University Dhaka, Bangladesh
sazzad01794@gmail.com

Abstract*: Insurance is a policy that reduces or eliminates the expenses associated with decreasing returns brought on by various risks. The price of insurance is influenced by a number of factors. These elements have an impact on how insurance plans are developed. The efficiency of insurance policy terms in the insurance industry can be enhanced using machine learning (ML). In this work, we use individual and local health data to forecast insurance amounts for various categories of people. To compare the effectiveness of these algorithms, nine regression models—Linear Regression, XGBoost Regression, Lasso Regression, Random Forest Regression, Ridge Regression, Decision Tree Regression, KNN Model, Support Vector Regression, and Gradient Boosting Regression—were utilized. The models were trained using the dataset, and some predictions were then made using the training data. The model was then put to the test and confirmed by contrasting the actual data with what was predicted to be abundant. These models' accuracy was compared subsequently. The optimal method to the XGBoost MAE 2381.567, MSE 19806356.6067, RMSE value 4450.4433, and R squared value of 0.8681 is provided in this report. Gradient Boosting and Random Forest, with R squared values of 0.8679 and 0.8382, respectively, are two further top models.

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

I. INTRODUCTION

A sector that is quickly growing globally is digital health. The number of digital health businesses has doubled globally during the last five years [1]. Health insurance faces two significant obstacles in industrialized nations: growing health care costs and an increase in the number of people without coverage. A broad-based political movement to address these issues is emerging as a result of this power. Governments in the area have pledged hundreds of millions of dollars to advance the digital health industry. Individual health insurance plays a crucial role in the healthcare system, particularly for people with rare diseases [2], for whom medical and preventative insurance can help cut down on treatment expenses. The world in which we live is a dangerous and unknowable place. houses, companies, buildings. These dangers include the potential for disease, death, and loss of assets or goods. People's happiness and health are fundamental to their existence. However, as risks cannot always be avoided, the financial industry has developed a number of products to shield people and businesses from them. These products employ money to make up for the risks; as a result, the costs of some risks are reduced or even eliminated. [3]. A crucial 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 with rare conditions. Numerous ML algorithms and deep learning approaches are used for data prediction. The factors of training time and accuracy are looked at. The bulk of machine learning algorithms only require a brief time of training. However, the prediction results from these techniques are not particularly accurate. Deep learning models can also find hidden patterns, but their usage in real-time is constrained by the training period [4].

II. BACKGROUND

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[4].

Several regression models were employed implemented in this report, including Linear Regression, XGBoost Regression, Lasso Regression, Random Forest Regression, Ridge Regression, Decision Tree Regression, KNN Model, Support Vector Regression, and Gradient Boosting Regression. The XGBoost and Gradient Boosting Regression are discovered to calculate with the highest accuracy of about 86 percent or more. The major objective of this study is to introduce a new methodology of estimating insurance costs.

III. LITERATURE REVIEW

This section demonstrates the research being done on information exploration and machine learning methods. Several articles have addressed the topic of claim prediction. "Utilizing telematics data to forecast automobile insurance claims," Jessica Pesantez-Narvaez claimed to be the author on 2019. In a relatively small number of cases, this study compared the effectiveness of logistic regression and XGBoost strategies in predicting the occurrence of accident states. The results and vibrations indicated that logistic regression is a superior model to XGBoost for the reasons of its interpretability and predictability [7].

Without taking into account predicted cost and claim scope, the research listed above identify claims problems. However, to predict the healthcare costs, we use advanced statistical methods, ML techniques, and deep neural networks.

IV. METHODOLOGY

A. Dataset Description

We obtained the data set from the Kaggle website [5] in order to calculate the cost of this model prediction. The data set is split into two categories: training data and test data, and it has seven attributes as listed in table I. 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 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)

There were 1338 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 22.5, 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.

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*

There were 1338 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 22.5, 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.

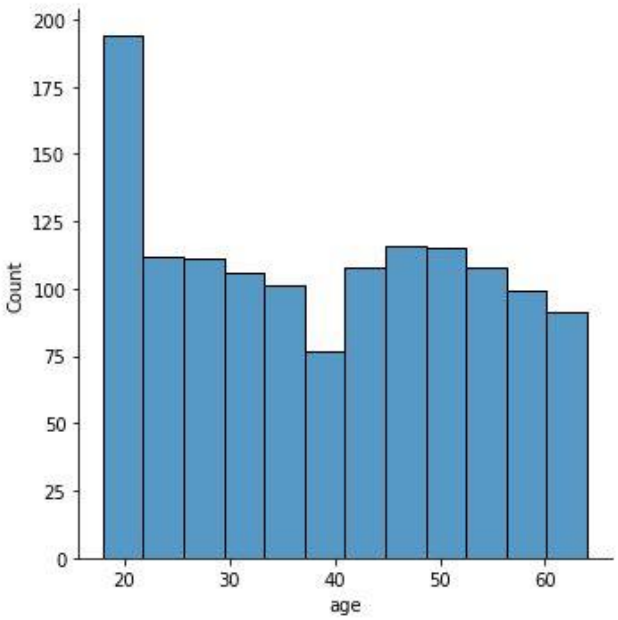

Figure-1: Distribution of age value

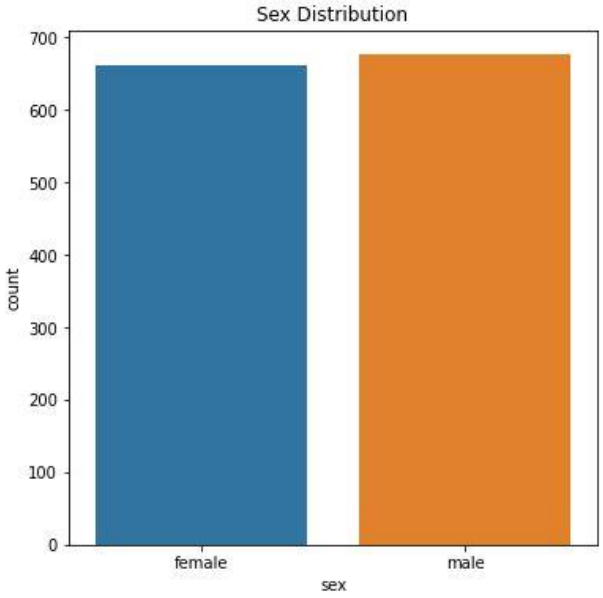

Figure-2: Sex Distribution

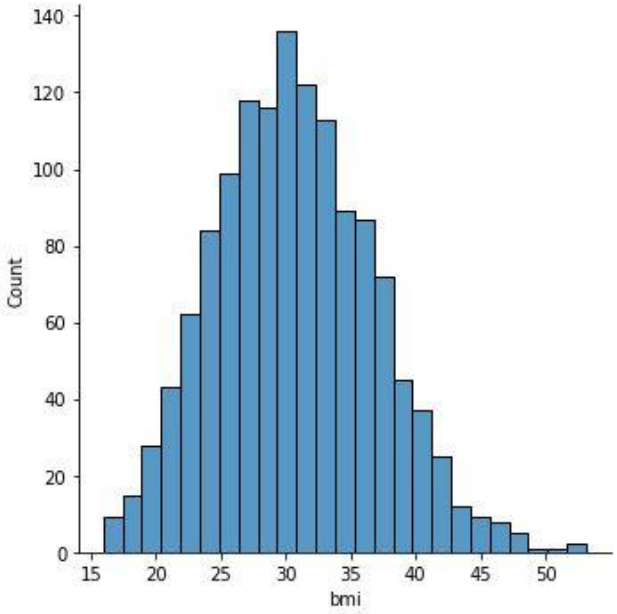

Figure-3: BMI Distribution

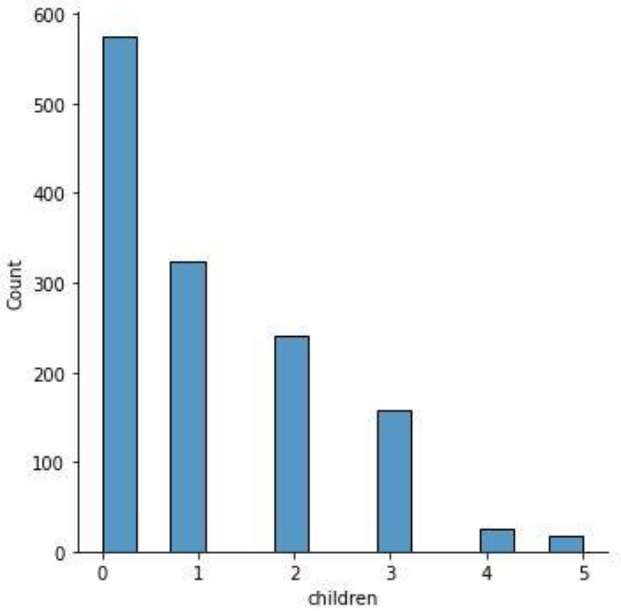

Figure-4: Children Counter

In order to determine the link between the variables, we will evaluate our data. Charges, which is dependent on all the other columns in this example, is our target column. The statistical metrics of our dataset will first be analyzed.

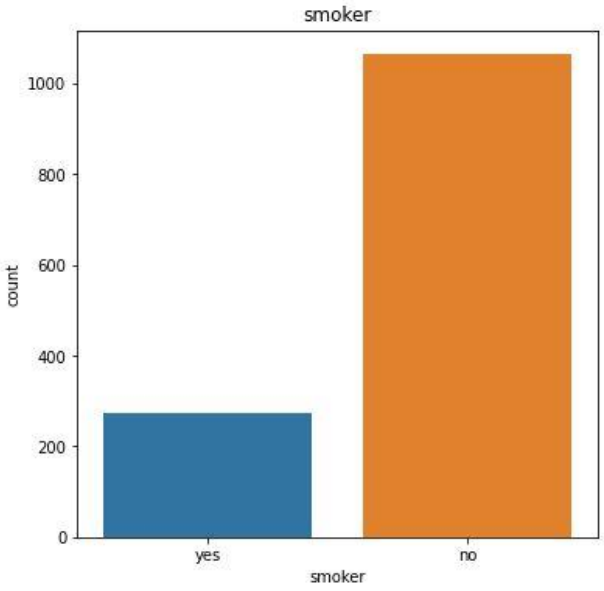

Figure-5: Checking Smoker and NON-Smoker

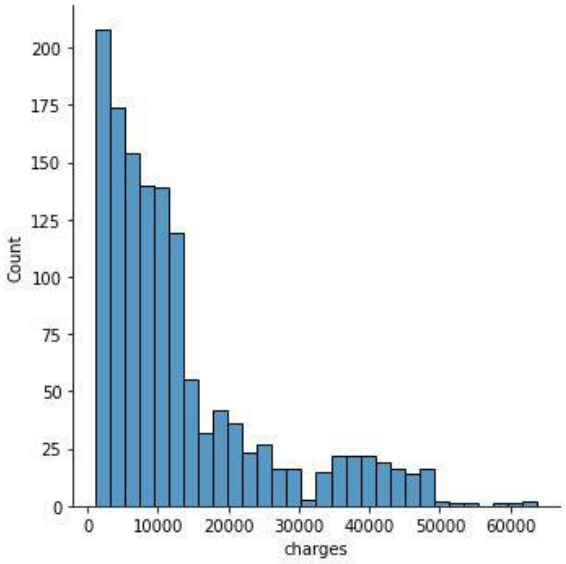

Figure-6: Distribution of Charge Value.

Only numerical values are presented. Standard deviations and average values for categorical variables are absent. In order to pre-process those features, later. The median number is higher than the average in the "charges" column. It implies that the price of health insurance is unfairly skewed. Once we make those things visible, we will clearly grasp this. We therefore begin by displaying the charge column's distribution.

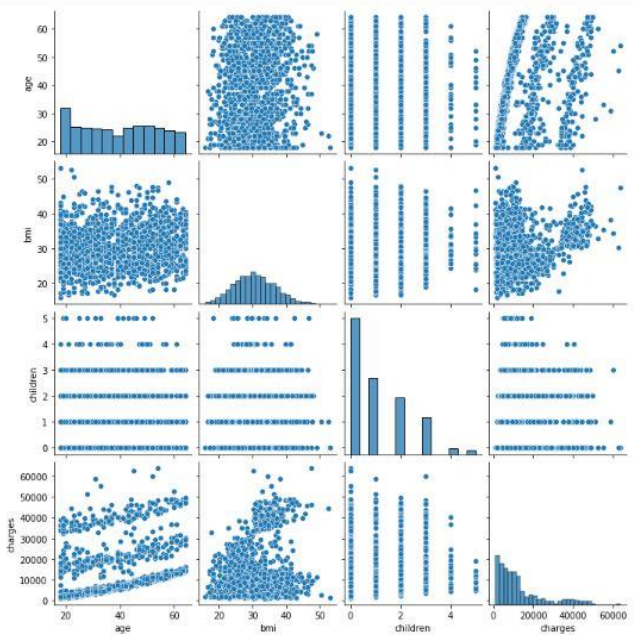

Figure-7: Visualization the relationship between two variables

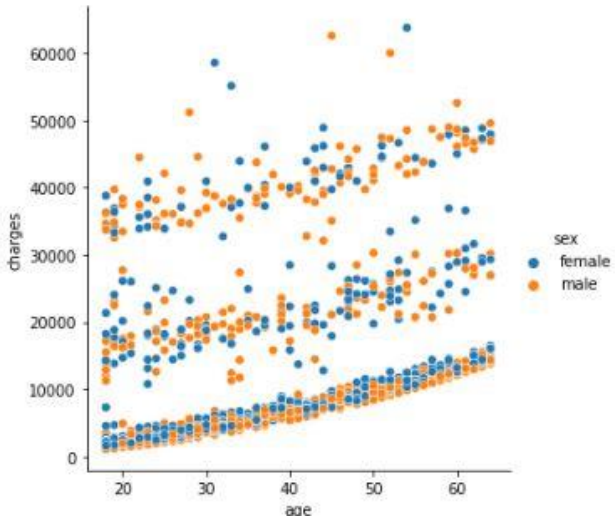

Figure-8: Plotting age, charges, sex

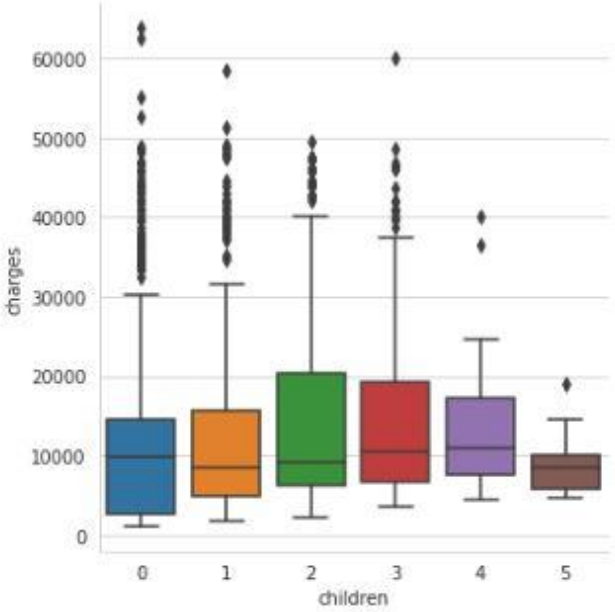

Figure-9: FacetGrid Children and Charges

C. Data Pre-processing

Three columns are numerical and three are categorical. Our machine learning model cannot suit the category values because computers cannot understand this text value. Therefore, we will give those categories qualities numerical labels. We change "female" to 1 and "male" to 0 in the "sex" field. We also change the other two columns to have numerical values. We display our results for conversion 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 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.

Table 4: Model Performance

Regression Models	R squared	MAE	RMSE
Linear Model	0.7447	4267.2138	6191.6908
XGBoost Regression	0.8681	2381.5670	4450.4333
Lasso Regression	0.7447	4267.1646	6191.7253
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 Neighbors	0.0394	8592.5456	12010.8927
Support Vector Regression	-0.099	6401.6428	12851.5588
Gradient Boosting Regression	0.8679	2383.9140	4453.8285

Our data is first obtained via Kaggle. Our dataset is then imported into Google Colab. Next, using various visualization tools, we analyze our data. The data is then cleaned such that it exactly matches the machine learning model. We then use our training data to apply regression techniques. Our model will be ready for cost forecasting after the data has been tested. The flowchart that follows illustrates the entire process.

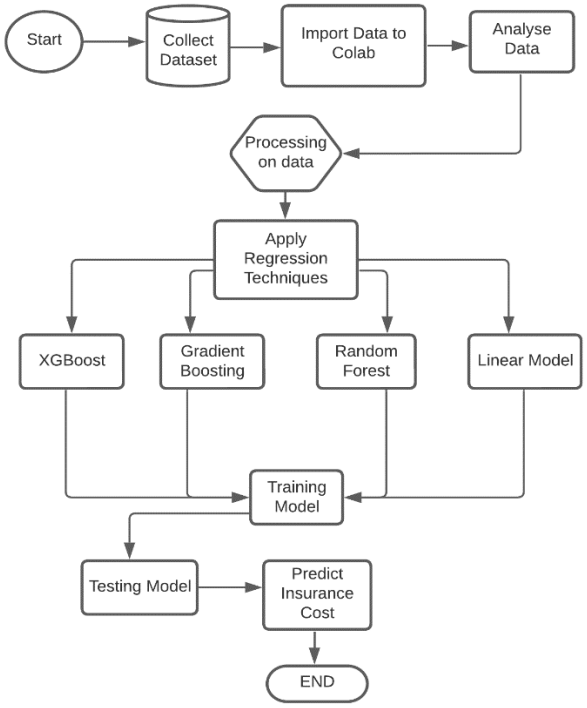

Figure-10: Flow Chart of Medical Insurance Cost Prediction System

V. RESULTS & DISCUSSION

Table-4 displays our top and bottom regression models. We can anticipate insurance costs using the model that performs best, according to the findings. In our situation, XGBoost Regression is the best regression model while K-Nearest Neighbors is the worst. Anyone may calculate their insurance expenses using the best model.

Figure-11: Predicted Cost using XGBBoost Regression

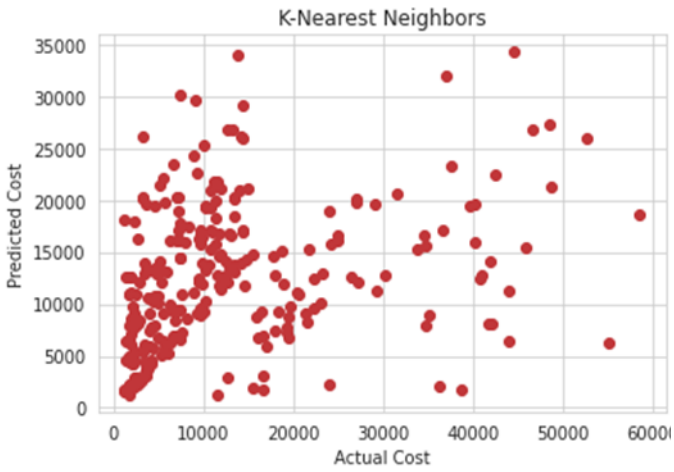

Figure-12: Predicted Cost Using K-Nearest Neighbors.

```
Enter your age: 23
Enter 0 for male and 1 for Female: 0
Enter your BMI rate: 24
Number of your children: 0
Enter 0 for smoker and 1 for non smoker: 1
Enter 0 for southeast, 1 for southwest, 2 for northeast, and 3 for northwest: 0
Using LR the insurance cost is USD  1147.3977546343885
Using RR insurance cost is USD  1185.6835419449199
Using LASSO insurance cost is USD  1153.8935648894912
Using RFR insurance cost is USD  3978.326517399997
Using DTR the insurance cost is USD  1815.8759
Using KNN insurance cost is USD  6720.235928
Using SVR insurance cost is USD  2555.218516065806
Using GBR insurance cost is USD  3958.783662344383
```

Figure-13: Medical Insurance cost prediction system demo result

In this *Figure-13:* we can see how medical insurance cost prediction system predict the cost for LR, RR, LASSO, RFR, DTR, KNN, SVR and GBR.

VI. CONCLUSION

In order to forecast health insurance prices based on provided factors in a Kaggle site medical cost individual data set, the study combines ML regression models. Table IV is a list of the outcomes. By predicting insurance rates based on a variety of factors, insurance policy firms may attract consumers and save time. Machine learning may significantly reduce these individual efforts in price analysis since ML models can compute costs quickly while doing so would take a person a long time. Large volumes of data can also be handled via machine learning techniques. The work might be improved in the future by building a web application based on the XGBoost or Gradient Boosting algorithm and using a larger dataset than that used in this study.

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 ophan 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. Syerov, “Development of mobile system for medical recommendations,” Procedia Computer Science, vol. 155, pp. 43–50, 2019

[5] Medical Cost Personal Datasets: 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