International Journal of Innovative Research in Engineering
Volume 6, Issue 2 (March - April 2025), PP: 122-128.
https://www.doi.org/10.59256/ijire.20250602015
www.theijire.com
ISSN No: 2582-8746

Medical Insurance Cost Prediction Using Machine Learning

Aditya Muluk[1], Prasadkumar Pandey[2], Priya Naik[3], Somnath Shanbhag[4], Naina Kaushik[5]
[1,2,3,4]Department of Computer Engineering, Rajiv Gandhi Institute of Technology Mumbai University Mumbai, Maharashtra, India
[5]Professor, Department of Computer Engineering, Rajiv Gandhi Institute of Technology Mumbai University Mumbai, Maharashtra, India

How to cite this paper:
Aditya Muluk[1], Prasadkumar Pandey[2], Priya Naik[3], Somnath Shanbhag[4], Naina Kaushik[5] "Medical Insurance Cost Prediction Using Machine Learning", IJIRE-V6I2-122-128.

Abstract *This paper presents a machine learning-based system for predicting medical insurance costs. The system utilizes a dataset from Kaggle containing 1,338 entries with features such as Age, Gender, BMI, Smoking Habit, and number of children. The study aims to enhance the efficiency of insurance policies through advanced predictive modeling, particularly in the context of the Covid-19 pandemic. Various regression models, including Linear Regression, Random Forest, and Gradient Boosting, were employed. The models were trained on a 70-30 dataset split and evaluated for accuracy. Random Forest emerged as the top performer with an R-squared value of 0.87, followed closely by Gradient Boosting (0.85). The results underscore the potential of machine learning to refine insurance pricing by leveraging personal and regional health data for more accurate predictions.*

Key Words: *Medical Insurance, Cost Prediction, Machine Learning, Regression, Random Forest, Gradient Boosting*

I.INTRODUCTION

Healthcare has become a global necessity, especially highlighted by the Covid-19 pandemic, which has underscored the importance of health insurance as a critical financial backup. With rising healthcare costs, accurate prediction of insurance costs is essential for both insurers and customers to manage potential risks and select the best policies.

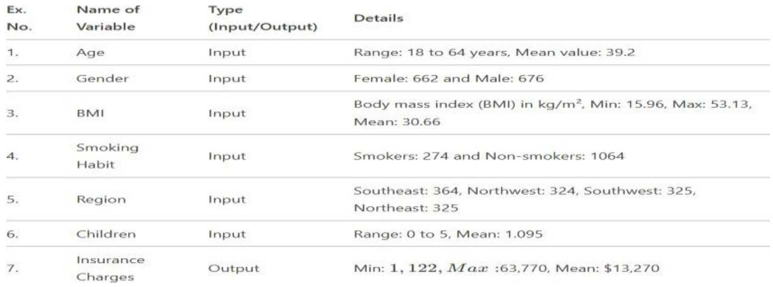
This paper explores the use of machine learning (ML) to predict individual health insurance costs. ML methods offer accuracy and efficiency in processing large volumes of data, benefiting both customers and insurance companies. The paper is structured as follows: Section II reviews the literature and problem statement, Section III describes the dataset and features, Section IV outlines the methodology, Section V presents the results and discussion, and Section VI concludes the study.

II.LITERATURE REVIEW AND PROBLEM STATEMENT

In recent years, several attempts have been made to predict health insurance costs using machine learning. A study by Kashish Bhatia et al. [1] proposed a machine learning-based health insurance prediction system, achieving an accuracy of 81.3% using linear regression. Another study by Mohamed Hanafy [2] demonstrated how different regression models could forecast insurance costs, highlighting the importance of accurate and reliable predictive tools. The primary challenge in health insurance cost prediction lies in the complexity of the data and the dynamic nature of healthcare policies. Existing systems often struggle with data quality, model complexity, and generalization issues. This project aims to address these challenges by developing a machine learning system that can accurately predict medical insurance costs using a dataset from Kaggle with 1,338 entries.

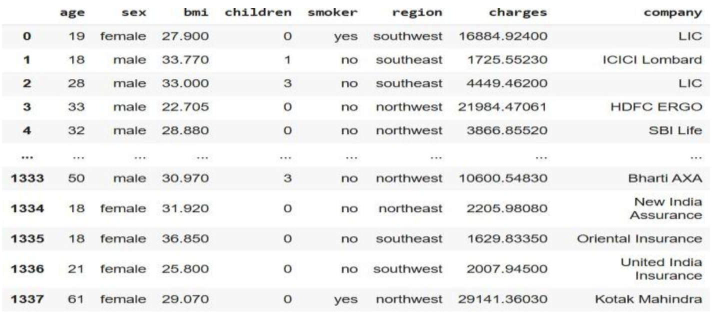
III.DATASET DESCRIPTION

We utilize the data from [18] to solve the insurance prediction task. The dataset contains 1,338 observations on insurance costs in four India regions. A detailed analysis of the dataset is given below in Table I.



Ex. No.	Name of Variable	Type (Input/Output)	Details
1.	Age	Input	Range: 18 to 64 years, Mean value: 39.2
2.	Gender	Input	Female: 662 and Male: 676
3.	BMI	Input	Body mass index (BMI) in kg/m², Min: 15.96, Max: 53.13, Mean: 30.66
4.	Smoking Habit	Input	Smokers: 274 and Non-smokers: 1064
5.	Region	Input	Southeast: 364, Northwest: 324, Southwest: 325, Northeast: 325
6.	Children	Input	Range: 0 to 5, Mean: 1.095
7.	Insurance Charges	Output	Min: 1, 122, *Max* :63,770, Mean: $13,270

Table I. Characteristics of Dataset Used



	age	sex	bmi	children	smoker	region	charges	company
0	19	female	27.900	0	yes	southwest	16884.92400	LIC
1	18	male	33.770	1	no	southeast	1725.55230	ICICI Lombard
2	28	male	33.000	3	no	southeast	4449.46200	LIC
3	33	male	22.705	0	no	northwest	21984.47061	HDFC ERGO
4	32	male	28.880	0	no	northwest	3866.85520	SBI Life
...	...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830	Bharti AXA
1334	18	female	31.920	0	no	northeast	2205.98080	New India Assurance
1335	18	female	36.850	0	no	southeast	1629.83350	Oriental Insurance
1336	21	female	25.800	0	no	southwest	2007.94500	United India Insurance
1337	61	female	29.070	0	yes	northwest	29141.36030	Kotak Mahindra

Fig. 1. Sample of data from the used dataset

IV.METHODOLOGY

The methodology involved a systematic approach to predicting medical insurance costs using machine learning. The dataset was preprocessed to handle missing values, outliers, and feature engineering. Three models—Multiple Linear Regression, Random Forest, and Gradient Boosting— were trained and evaluated using a 70-30 dataset split. The models were evaluated using R-squared and MSE, with Random Forest emerging as the best-performing model. The results were interpreted to understand the impact of different features on insurance costs, and the model was deployed for real-world use. This approach ensures accurate and reliable predictions, which can help insurance companies and customers make informed decisions.

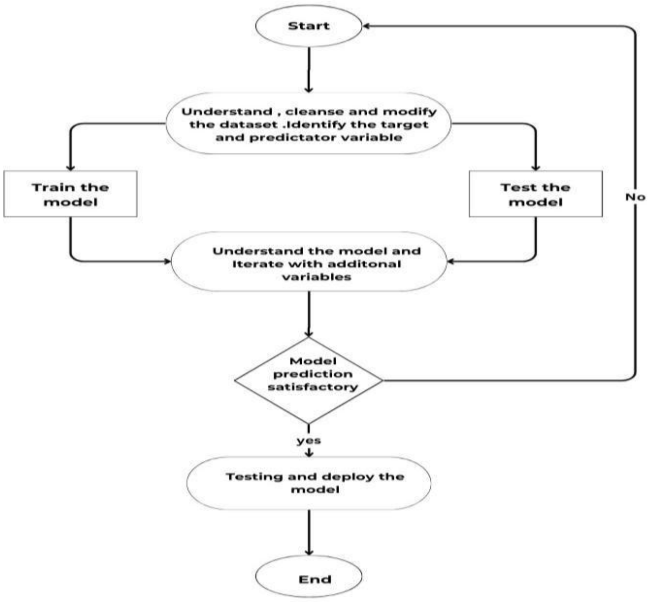

Start
Understand , cleanse and modify the dataset .Identify the target and predictator variable
Train the model
Test the model
No
Understand the model and iterate with additonal variables
Model prediction satisfactory
yes
Testing and deploy the model
End

Fig. 2. Proposed methodology

1. **Multiple Linear Regression:** Linear regression is a supervised machine learning algorithm that models the linear relationship between the dependent variable (insurance cost) and one or more independent features (e.g., age, BMI, smoking habit). It fits a linear equation to the observed data, allowing for the prediction of the dependent variable based on the input features.

○ **Mathematical Representation:**

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta nxn + \epsilon$$

Where:

- y= predicted insurance cost (dependent variable),
- $\beta 0$ = intercept (constant term),
- $\beta 1, \beta 2, \ldots, \beta n$ = coefficients of the independent variables,
- $x1, x2, \ldots, xn$ = independent variables (e.g., age, BMI, smoking habit),
- ϵ = error term (residuals)

2. **Random Forest:** Random Forest is an ensemble learning method that constructs multiple decision trees during the training phase. Each tree is built using a random subset of the dataset and a random subset of features. The final prediction is made by averaging the predictions of all trees, which reduces the risk of overfitting and improves overall prediction accuracy.

- **Mathematical Representation:**

$$\hat{y}_i = \frac{1}{K}\sum_{k=1}^{K} f_k(x_i)$$

Where:

- y^i = predicted value for data point xi,
- K = number of trees in the Random Forest,
- fk(xi) = prediction from the k-th tree for xi.

3. **Gradient Boosting:** Gradient Boosting is another ensemble learning technique that combine the predictions of multiple weak models (typically decision trees) to produce a stronger prediction. It works by sequentially building trees, where each tree corrects the errors of the previous one. This method is particularly effective for handling large datasets and achieving high accuracy.

- **Mathematical Representation:**

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

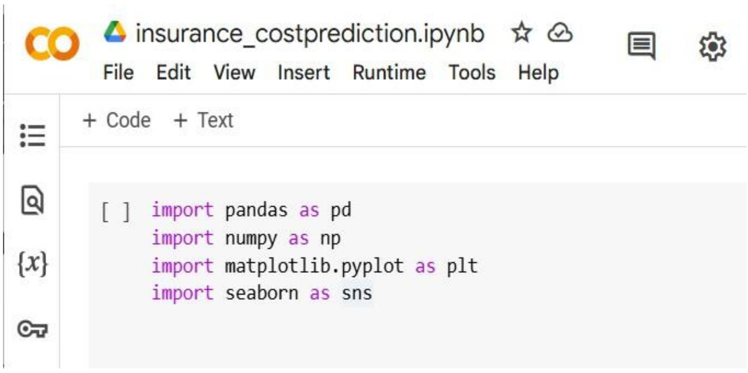
Where:

- y^i = predicted value for data point xi,
- K = number of trees or base models,
- fk(xi)= output from the k-th base learner for input xi.

4. **GOOGLE COLLAB:** Collaboratory, or simply Colab, is a Google Research tool that allows developers to create and run Python code directly from their browser. For deep learning tasks, Google Colab is a good tool.

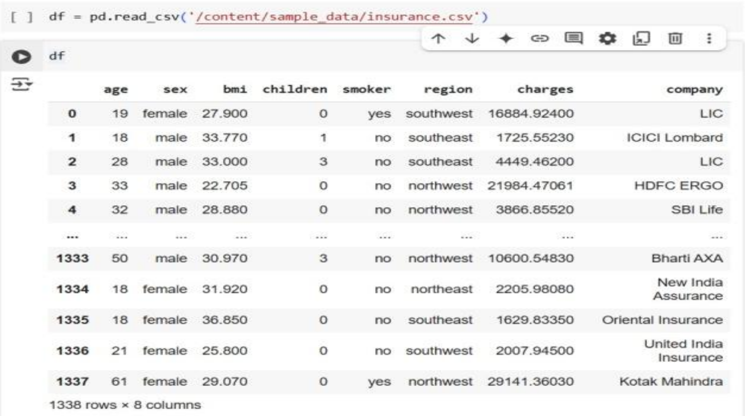
Steps for predicting the insurance cost: The above technologies and datasets were used to implement the research paper. Here, we will discuss steps to carry out this work alongwith supporting screenshots.

1. **Loading the libraries and modules:** All the required libraries and modules were loaded into the Google Collab.



```
[ ] import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    import seaborn as sns
```

2. **Loading Data:** Next, data downloaded from Kaggle is uploaded in the project after cleaning and filtrations.



```
[ ] df = pd.read_csv('/content/sample_data/insurance.csv')
```

```
df
```

	age	sex	bmi	children	smoker	region	charges	company
0	19	female	27.900	0	yes	southwest	16884.92400	LIC
1	18	male	33.770	1	no	southeast	1725.55230	ICICI Lombard
2	28	male	33.000	3	no	southeast	4449.46200	LIC
3	33	male	22.705	0	no	northwest	21984.47061	HDFC ERGO
4	32	male	28.880	0	no	northwest	3866.85520	SBI Life
...	...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830	Bharti AXA
1334	18	female	31.920	0	no	northeast	2205.98080	New India Assurance
1335	18	female	36.850	0	no	southeast	1629.83350	Oriental Insurance
1336	21	female	25.800	0	no	southwest	2007.94500	United India Insurance
1337	61	female	29.070	0	yes	northwest	29141.36030	Kotak Mahindra

1338 rows × 8 columns

V.RESULTS AND DISCUSSION

This section represents the results obtained, the first six graphs of this section shows the correlation between charges and various features like age, sex, BMI, Number of children, Smoker, charges, region, company

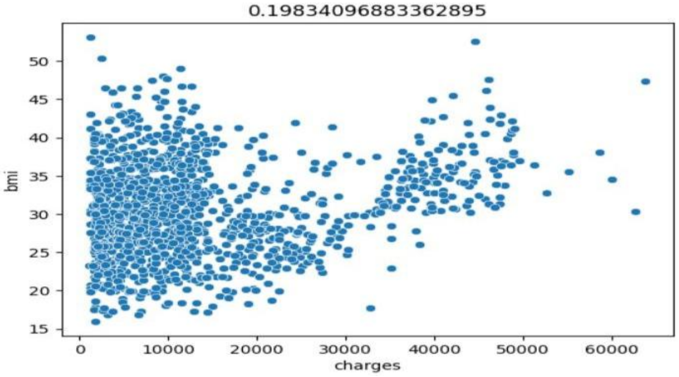


Fig. 3. Correlation between insurance charges and bmi.

.Figure 3 presents the correlation between insurance charges and BMI. As shown, there is a weak positive correlation (0.198) between BMI and insurance charges. While an increase in BMI is generally associated with higher charges, the relationship is not strong.

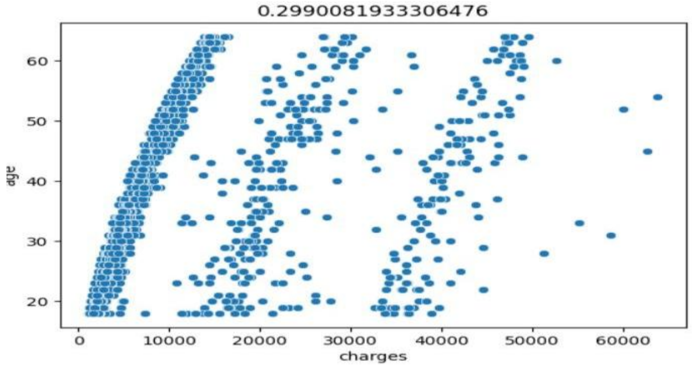


Fig. 4. Correlation between insurance charges and age

Figure 4 presents the correlation between insurance charges and age. As expected, charges increase exponentially with age. Higher age groups tend to have significantly higher insurance charges.The correlation coefficient (0.299) indicates a moderate positive correlation.

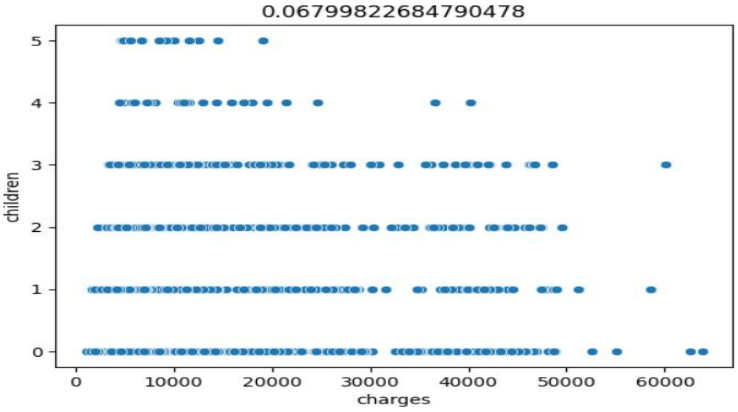


Fig. 5. Correlation between insurance charges and children

Figure 5 presents the correlation between insurance charges and number of children. As shown, there is no significant correlation between the number of children and insurance charges, as indicated by the low correlation coefficient (0.068).

Numerical column analysis

1. age is the strongest predictor in dataset
2. bmi moderately signify the cost predication
3. childern has weak correlation with cost predication

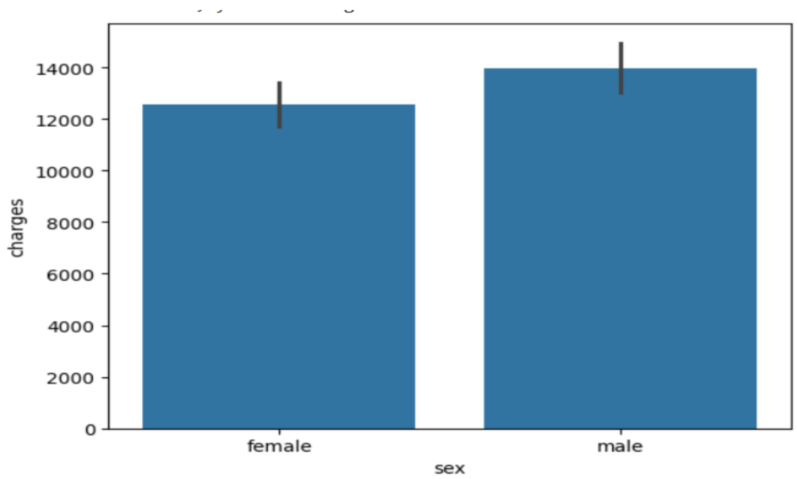


Fig. 6. Barplot between insurance charges and sex

Figure 6 presents the relationship between insurance charges and sex. As shown, there is no significant difference in insurance charges between males and females. Both genders have similar average charges, with males having slightly higher costs.The bar chart suggests that sex is not a strong factor in determining insurance charges.

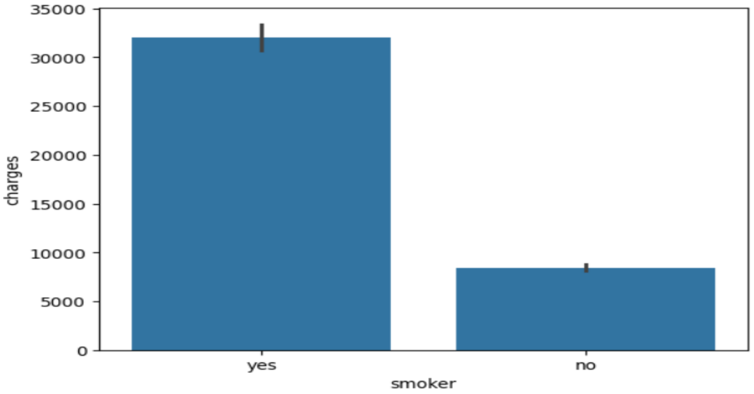


Fig. 7. Barplot between insurance charges and smoker

Figure 7 presents the relationship between insurance charges and smoking status. As shown, smokers have significantly higher insurance charges compared to non- smokers. The difference is substantial, indicating that smoking is one of the most influential factors affecting insurance costs.The bar chart suggests that smoking status has a strong impact on medical expenses.

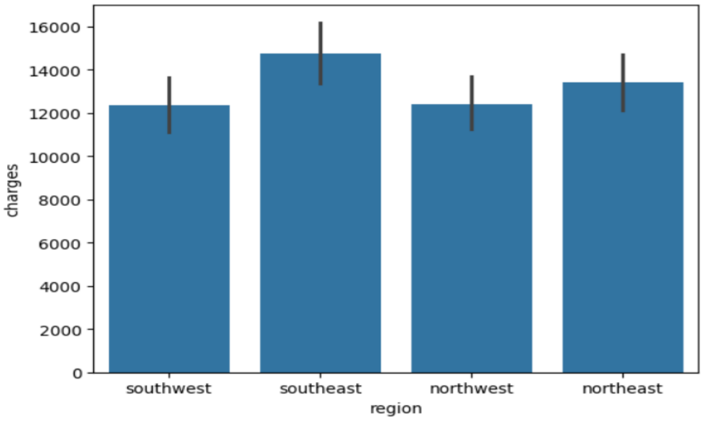


Fig. 8. Barplot between insurance charges and region

Figure 8 presents the relationship between insurance charges and region. As shown, insurance charges vary slightly across different regions, with the Southeast region having the highest average charges and the Southwest and Northwest regions having slightly lower costs.The bar chart suggests that region does not have a significant impact on insurance charges, as the differences are relatively small.

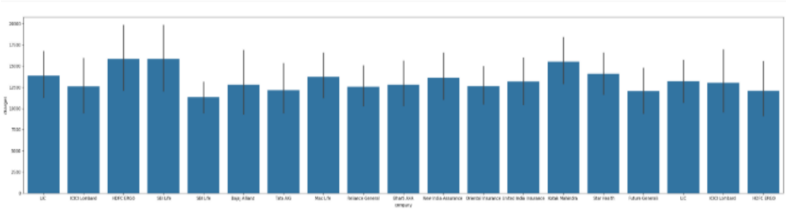
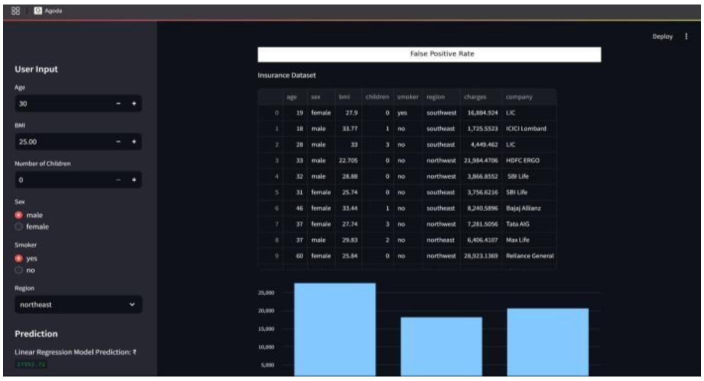


Fig. 9. Barplot between insurance charges and companies

Figure 9 presents the relationship between insurance charges and companies. As shown, insurance charges remain relatively consistent across different companies, with minor variations in average costs.

Categorical -Numerical column analysis:

1. gender wise effect cannot signify much in cost predication
2. smoker is highly significant in cost predication
3. region wise does not much signify much in cost prediction
4. companies like SBI LIFE , HDFC ERGO and KOTAK MAHINDRA moderately signify in the cost predication



User Input

Age

30

BMI

25.00

Number of Children

0

Sex

● male
○ female

Smoker

● yes
○ no

Region

northeast

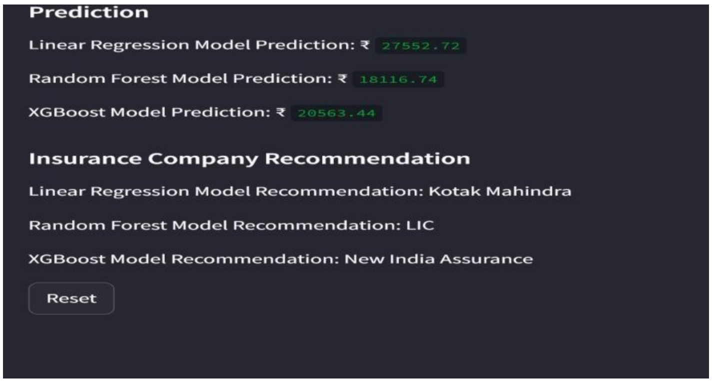
Prediction

Linear Regression Model Prediction: ₹ 27552.72

Insurance Dataset

	age	sex	bmi	children	smoker	region	charges	company
0	19	female	27.9	0	yes	southwest	16,884.924	LIC
1	18	male	33.77	1	no	southeast	1,725.5523	ICICI Lombard
2	28	male	33	3	no	southeast	4,449.462	LIC
3	33	male	22.705	0	no	northwest	21,984.4706	HDFC ERGO
4	32	male	28.88	0	no	northwest	3,866.8552	SBI Life
5	31	female	25.74	0	no	southeast	3,756.6216	SBI Life
6	46	female	33.44	1	no	southeast	8,240.5896	Bajaj Allianz
7	37	female	27.74	3	no	northwest	7,281.5056	Tata AIG
8	37	male	29.83	2	no	northeast	6,406.4107	Max Life
9	60	female	25.84	0	no	northwest	28,923.1369	Reliance General

Fig. 10. Website page



Prediction

Linear Regression Model Prediction: ₹ 27552.72

Random Forest Model Prediction: ₹ 18116.74

XGBoost Model Prediction: ₹ 20563.44

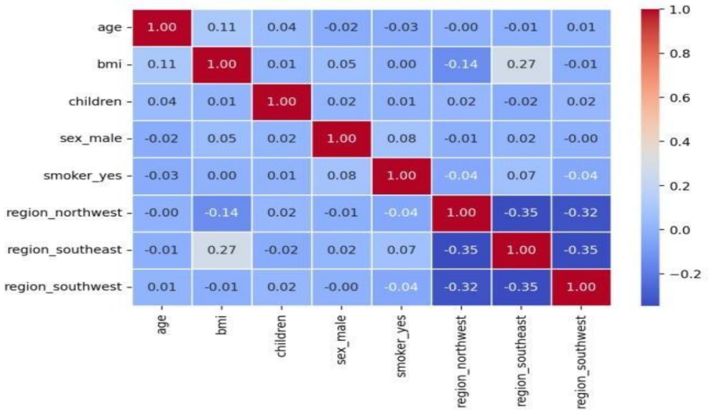
Insurance Company Recommendation

Linear Regression Model Recommendation: Kotak Mahindra

Random Forest Model Recommendation: LIC

XGBoost Model Recommendation: New India Assurance

Fig. 11. Output



	age	bmi	children	sex_male	smoker_yes	region_northwest	region_southeast	region_southwest
age	1.00	0.11	0.04	-0.02	-0.03	-0.00	-0.01	0.01
bmi	0.11	1.00	0.01	0.05	0.00	-0.14	0.27	-0.01
children	0.04	0.01	1.00	0.02	0.01	0.02	-0.02	0.02
sex_male	-0.02	0.05	0.02	1.00	0.08	-0.01	0.02	-0.00
smoker_yes	-0.03	0.00	0.01	0.08	1.00	-0.04	0.07	-0.04
region_northwest	-0.00	-0.14	0.02	-0.01	-0.04	1.00	-0.35	-0.32
region_southeast	-0.01	0.27	-0.02	0.02	0.07	-0.35	1.00	-0.35
region_southwest	0.01	-0.01	0.02	-0.00	-0.04	-0.32	-0.35	1.00

Fig. 12. Correlation Matrix

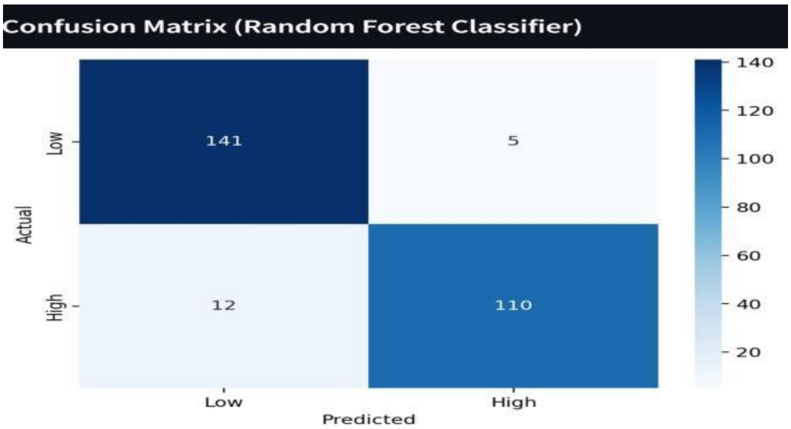


Fig. 13. Confusion Matrix

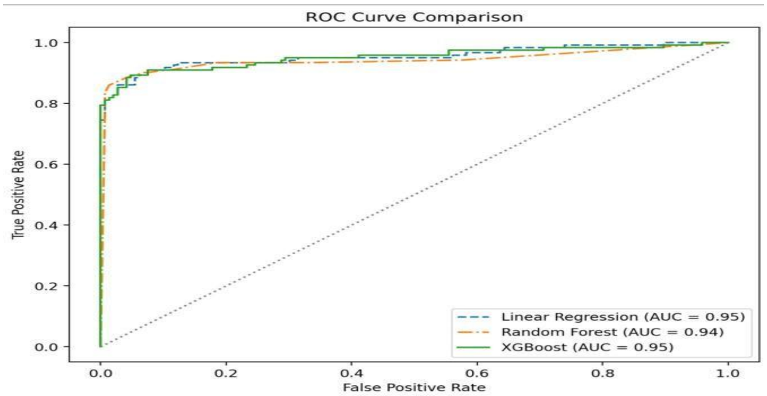


Fig. 13. ROC Curve

VI.CONCLUSION

In this medical insurance cost prediction project, we developed a model that effectively estimates insurance costs based on key factors: age, BMI, sex, region, and smoking status. Our findings indicate that each of these variables significantly influences insurance premiums, with age and smoking having particularly strong correlations. This highlights the importance of a data-driven approach in designing insurance products that are both fair and tailored to individual risk profiles.

References

The references cited in this project are presented below. This is the list of references that were utilized to support the research and development of this project.

[1] Digital Health 150: The Digital Health Startups Transforming the Future of Healthcare | CB Insights Research", CB Insights Research, 2022. [Online].https://www.cbinsights.com/research/report/digital-health-startups- redefining-healthcare. [Accessed: 10-Sep- 2022]

[2] Kashish Bhatia, Manish Kumar, Shabeg Singh Gill, Rajesh Kumar Bhatia, and Navneet Kamboj, "Health Insurance Cost Prediction using Machine Learning," IEEE, 2022.

[3] Mohamed Hanafy, "Predict Health Insurance Cost by using Machine Learning and DNN Regression Models," IJITEE, 2021.

[4] Medical Cost Personal Datasets, Kaggle, [Online]. Available: https://www.kaggle.com/mirichoi0218/insurance.