MEDICAL INSURANCE
PREMIUM PREDICTION USING MACHINE LEARNING

B.Ashok kumar 1, N.Venkata Surendra 2, A.Ganesh 3, B.Venkatesh 4

1, 2, 3, 4Student, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Guntur (D.T) A.P, India

5Professor, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Guntur (D.T) A.P, India

ashokkbk789@gmail.com[1], venkatasurendra70@gmail.com[2], ganeshakarapu1@gmail.com[3], battulavenkatesh140@gmail.com[4]

Abstract: Medical insurance can play a crucial role in protecting individuals and families from the financial burden of unexpected healthcare expenses. With the rising cost of healthcare, having a medical insurance policy can provide much-needed financial security during a medical emergency. When selecting a policy, individuals should consider various factors, such as the type of coverage, deductibles, copayments, and network providers. Many insurance companies offer a range of policy options to meet the diverse needs of individuals and families.

Keywords : *Machine learning, Linear regression, KNN, XGBoost, Random forest, Decision tree, Gradient boosting.*

I. INTRODUCTION

Health insurance is a crucial aspect of the healthcare industry in India, but its effectiveness has been hindered by inefficiencies in the government-run system. Additionally, the Indian bureaucracy's lack of innovation and control has been a deterrent for private players in the health insurance sector.

This project report aims to provide a comprehensive understanding of health insurance in India. The main objective of this research is to analyze the current state of the company and identify areas of improvement to help the company achieve its business goals.

The second part of the report focuses on data analysis, which was collected through a survey of 30 people using a questionnaire. The objective of these interviews was to gather first-hand insights into the experiences of customers and understand their preferences when it comes to insurance products. The data collected from visits was analyzed to identify best practices and potential areas of improvement for HDFC Life

I. DATASET DESCRIPTION

we require a dataset with various attributes that can be used to train and test the system. In this project, we obtained the dataset from Kaggle, which contains crucial details of several attributes necessary for the prediction.

The dataset includes factors such as age, gender, BMI, number of dependents, smoking habits, region, and charges incurred. We used this dataset to train and test our machine learning algorithms to accurately predict the insurance premium for individuals.

III. DATA PREPROCESSING

Preprocessing involves transforming raw data into clean, organized data that is suitable for analysis. Properly formatted data is essential for obtaining accurate results when using machine learning models. Each algorithm has specific requirements for data format. For instance, if we use the Random

Forest algorithm, it does not support null values. Therefore, to apply this algorithm, we must manage any null values in the raw data beforehand.

Preprocessing can involve various tasks such as data cleaning, data integration, data transformation, and data reduction. The purpose of these tasks is to improve the quality of data and eliminate any errors or inconsistencies that can affect the accuracy of machine learning models. By performing pre-processing, we can ensure that our machine learning models are trained on clean, organized data, leading to better results and more informed decision-making.

IV. EXISTING SYSTEM

The accuracy of health insurance premium prediction algorithms can be impacted by a variety of factors, including inadequate preprocessing techniques, inadequate consideration of relevant attributes, and the use of suboptimal algorithms. To address these challenges and ensure more accurate predictions, a new system has been proposed. This system incorporates advanced preprocessing techniques and considers a broader range of attributes to enhance the accuracy of predictions. Additionally, the system utilizes efficient algorithms that have been rigorously tested to minimize errors and ensure reliable results. Overall, the proposed system represents a significant improvement over traditional methods for health insurance premium prediction and promises to deliver more accurate and reliable predictions for insurers and customers alike.

Disadvantages

- Firstly, these algorithms often fail to generate accurate and efficient results,
- The computation time required to process these algorithms is often very high.
- Another significant issue is that these algorithms may fail to take all relevant attributes into account
- lack of accuracy in health insurance premium prediction can result in significant financial losses for insurers and customers alike

V.PROPOSED SYSTEM

Our proposed system aims to assist financial institutions in predicting premiums based on various attributes, such as age, gender, smoking habits, and other relevant factors. This decision system will provide a faster and more efficient method for predicting policy prices, while also reducing manual work for financial institutions.

To achieve this goal, we have utilized machine learning algorithms, such as Linear Regression, Random Forest, Ridge Regression, and Lasso Regression. These algorithms have proven to be highly accurate and reliable, allowing our system to generate efficient and accurate results.

Advantages:

- our system greatly reduces computation time, making it a more efficient solution for financial institutions.
- our automated prediction system, financial institutions can save time and resources
- our proposed system offers accurate and efficient results.
- Reduced computation time
- Efficient prediction

VI.METHODOLOGY OF RESEARCH

The research methodology used in this study involved the collection of both Direct data and Indirect data

Direct Data:

Direct data was collected through the use of questionnaires, individual interactions were conducted with employees of the company. Current Customers and remote interviews. A specimen of the questionnaire used in this study is attached on the last page

Indirect Data:

Indirect data was collected from various sources such as journals, census reports, web links and company prospectuses.

Tools and Techniques:

Questionnaires were used as a source to collect primary data, while secondary data was collected from journals, websites, and company prospectuses. Percentage analysis was performed on the collected data to draw conclusions.

Sampling Plan:

After determining the research approach and instruments, a sampling plan was designed to ensure representativeness of the population.

Sampling Unit:

By selecting this specific target population, the study aimed to gain insights into the experiences, behaviors, and preferences of this group of clients.

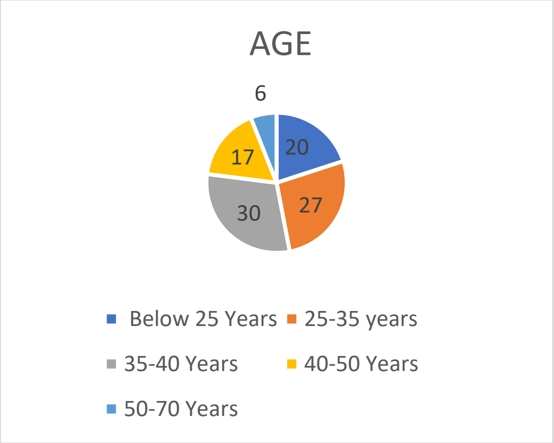
Sample Size:

A sample sizes from individuals
Customers was selected using a random sampling technique. This approach was taken to ensure that the sample represented the target population as
accurately as possible.

VII. DATA ANALYSYS

Aimed to gather information about the age group of the participants.

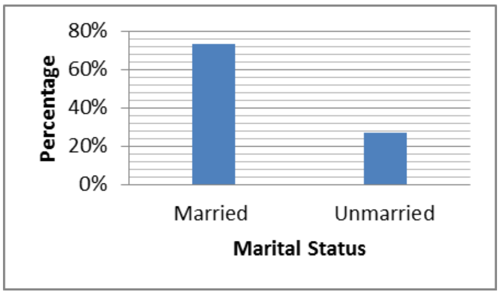
Response	Response	Percentage
Below 25 yrs.	6	20%
25-35yrs	8	27%
• 35-40yrs	9	30%
• 40-50yrs	5	17%
• 50yrs-70 years	2	6%
Total	30	100%



Analysis: - we can observe that the age group of 35-40 years old had the highest percentage of participants at 30%.

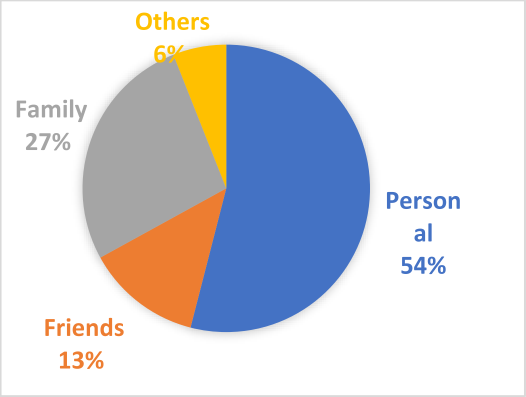
Aimed to know about Marital Status

Response	Respondents	Percentage
Married	22	73%
Unmarried	8	27%
Total	30	100



Aimed to gather information on the individuals who take health insurance policies

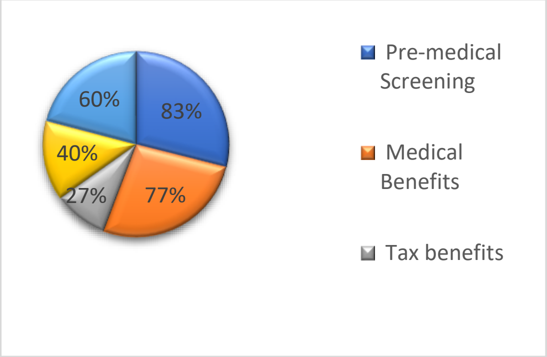
Response	Respondents	Percentage
Personal	16	54%
Friends	4	13%
Family	8	27%
Others	2	6%
Total	30	100%



Analysis: -Many people tend to prioritize personal and family health risks before other health concerns.

Aimed to gather information on the people's perspectives on the significance of health insurance.

Response	Respondents	Percentage
Pre-medical Screening	25	83%
Medical Benefits	23	77%
Tax benefits	8	27%
Convalescence benefit	12	40%
Lump sum for critical illnesses	18	60%
Total	30	100%

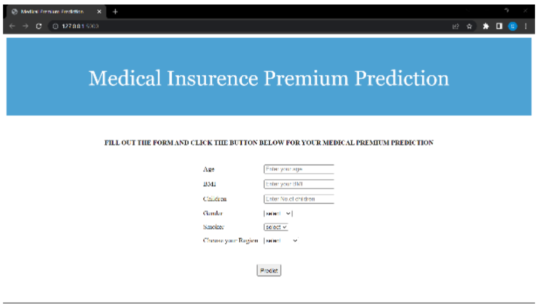


Analysis: - we can observe that many people place a high value on pre-medical screening benefits

VIII. RESULTS AND ANAYLYSIS

Linear Regression	-78.9796
Gradient Boosting	-85.9726
Decision Tree	-83.9796
KNN	-77.2665
XGBoost	-77.2608
RandomForest Regressor	-84.9282

HOME PAGE

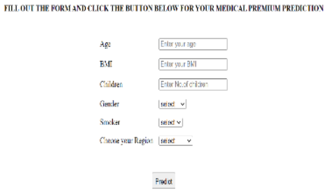


Medical Insurence Premium Prediction

FILL OUT THE FORM AND CLICK THE BUTTON BELOW FOR YOUR MEDICAL PREMIUM PREDICTION

Age: Enter your age
BMI: Enter your BMI
Children: Enter No.of children
Gender: select
Smoker: select
Choose your Region: select

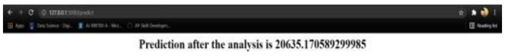
PREDICTION FORM PAGE



FILL OUT THE FORM AND CLICK THE BUTTON BELOW FOR YOUR MEDICAL PREMIUM PREDICTION

Age: Enter your age
BMI: Enter your BMI
Children: Enter No.of children
Gender: select
Smoker: select
Choose your Region: select

DISPLAYING THE RESULT



Prediction after the analysis is 20635.170589299985

IX. CONCLUSION

We have used algorithms like Linear Regression, Random Forest Regressor, Gradient Boosting, Decision Tree, KNN, XGBooost in-order to predict the medical insurance premium. The accuracyvaries for different algorithms. The accuracy for Linear Regression algorithm is 78.48. The accuracy for Random Forest Regressor algorithm is 84.33. The accuracy for, Gradient Boosting algorithm is 85.27. The accuracy for Decision Tree algorithm is 83.87. The accuracy for, KNN algorithm is 77.26. The accuracy for, XGBoost algorithm is 77.26. The highest accuracyis achieved when we have Gradient Boosting used algorithm.

X. REFERENCES

Dataset:

https://www.kaggle.com/mirichoi0218/insurance

OtherSources:

- https://www.geeksforgeeks.org
- https://www.w3schools.com
- https://openaccess.thecvf.com
- https://www.mdpi.com/2075-4418/10/6/417/html