AMITY UNIVERSITY ONLINE, NOIDA, UTTAR PRADESH

In partial fulfilment of the requirement for the award
of degree of **Master of Computer Applications**

MEDICAL INSURANCE PRICE PREDICTION USING MACHINE LEARNING

Guide Details: -

Name: Deepak Singh Bisht

Designation: Quality Analyst Lead

Submitted By: -

Name of the Student: Ashish Kanojia

Enrollment No: A9929723001033(el)

ABSTRACT

Accurate medical insurance price prediction is crucial for both insurance providers and policyholders, ensuring fair pricing and efficient risk management. Conventional pricing models frequently face challenges when dealing with extensive, high-dimensional data, which is why machine learning serves as a valuable approach to improve precision and flexibility.

This project focuses on developing a machine learning-based model to predict medical insurance prices based on various factors such as age, gender, medical history, lifestyle habits, and socioeconomic data. The model employs techniques like regression analysis, random forest, and visualizing techniques to identify patterns and predict insurance costs more effectively.

Key components of the project include data collection, preprocessing, feature selection, model training, and evaluation. Various supervised learning algorithms will be explored, with the performance assessed using metrics like R-squared and Accuracy values. The project aims to enhance pricing transparency and help insurance companies optimize their pricing strategies while providing consumers with better estimates for their premiums.

This method facilitates dynamic, data-informed decision-making through the integration of machine learning techniques, thereby enhancing both the efficiency and equity of medical insurance pricing.

Keywords*: Machine Learning (ML), Artificial Intelligence (AI), Medical insurance, Premium Price, Regression Algorithm.*

AMITY UNIVERSITY

DECLARATION

I, **Ashish Kanojia**, a student pursuing **Master of Computer Applications (Semester IV)** at Amity University Online, hereby declare that the project work entitled "**MEDICAL INSURANCE PRICE PREDICTION USING MACHINE LEARNING**" has been prepared by me during the academic year 2025 under the guidance of Deepak Singh Bisht, Quality Analyst Lead, Xceedance Consulting India Private Limited.

I assert that this project is a piece of original bona-fide work done by me. It is the outcome of my own effort and that it has not been submitted to any other university for the award of any degree.

Ashish Kanojia

Signature of Student

AMITY UNIVERSITY

CERTIFICATE

This is to certify that **Ashish Kanojia** of Amity University Online has carried out the project work presented in this project report entitled "**MEDICAL INSURANCE PRICE PREDICTION USING MACHINE LEARNING**" for the award of **Master of Computer Applications (General)** under my guidance. The project report embodies results of original work, and studies are carried out by the student himself/herself.

Certified further, that to the best of my knowledge the work reported herein does not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Signature

(Deepak Singh Bisht)

(Quality Analyst Lead)

AMITY UNIVERSITY

PLAGIARISM STATEMENT & REPORT

I certify that this Project Report/Dissertation is my own work, based on my personal study and/or research and that I have acknowledged all material and sources used in its preparation, whether they be books, articles, reports, lecture notes, and any other kind of document, electronic or personal communication.

I also certify that this Project Report/Dissertation has not previously been submitted for any assessment in any other unit, except where specific permission has been granted from all unit coordinators involved, or at any other time in this unit, and that I have not copied in part or whole or otherwise plagiarized the work of other students and/or persons/or any entity.

The plagiarism in this report has been checked using the tool https://plagiarisma.net/ and it came out to be **100%** unique.

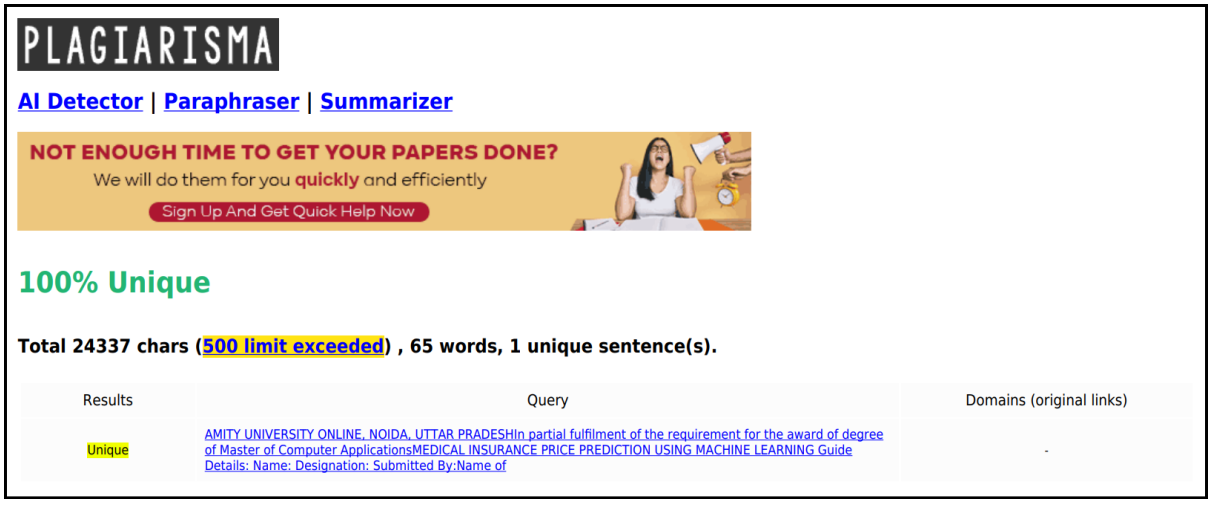

TABLE OF CONTENTS

1. INTRODUCTION

Healthcare systems in developing nations rely significantly on out-of-pocket payments, which act as a barrier to achieving universal health coverage, as they lead to inefficiency, inequity, and increased costs. Health insurance functions as a tool for individuals across various countries to mitigate the financial risks associated with medical expenses. It offers protection against the costs arising from medical treatments and related services. Nevertheless, due to the elevated rates imposed by insurance providers, a considerable number of individuals remain uninsured, resulting in their inability to access timely healthcare services, which in turn leads to elevated mortality rates.

A health insurance policy is designed to cover or reduce the financial burden of losses incurred from a range of risks. The accurate forecasting of individual healthcare costs through predictive models is essential for various stakeholders and health departments, as numerous elements affect the pricing of insurance or healthcare.

Precise cost assessments can aid health insurers and, increasingly, healthcare delivery organizations in future planning and in prioritizing the distribution of limited care management resources. Additionally, having advance knowledge of their potential future expenses can help patients select insurance plans with suitable deductibles and premiums. These considerations play a vital role in the creation and evolution of insurance policies.

However, determination of health insurance premiums is frequently complex, as it must establish rates that are acceptable to both insurance providers and policyholders; insurance companies must generate profit by collecting more in premiums than they disburse for the

medical expenses of their clients, thereby ensuring their continued operation. These companies determine premium prices based on the likelihood of specific events occurring within a population.

Nevertheless, estimating medical expenses and related costs proves challenging due to the rarity and seemingly arbitrary nature of the most expensive conditions. Another complex aspect of assessing medical expenses is the variability in the occurrence of certain diseases among individuals and across different population segments. Consequently, there is a necessity for a premium calculation model that accurately reflects the distinct factors of the population.

In this context, this study aims to leverage advanced computational techniques to estimate the cost of medical insurance premiums based on various factors like ***Age, Sex, Health Conditions, Location, and Lifestyle Habits***. By analyzing historical insurance data and identifying key patterns, the model provides accurate predictions, helping both insurers and policyholders make informed decisions. Traditional pricing methods rely on statistical analysis and expert judgment, but **Machine Learning (ML)** offers a more accurate and data-driven approach to predicting insurance costs.

*As an IT professional specializing in the insurance domain, I have had the opportunity to work with esteemed client such as **Berkshire Hathaway Specialty Insurance (BHSI)**. Given my background and expertise in insurance technology. I chose this topic because it closely aligns with my professional interests and industry relevance. By focusing on this subject, I aim to leverage my knowledge of insurance systems, risk assessment models, and healthcare analytics to explore innovative solutions in the domain.*

2. LITERATURE REVIEW

In the literature review, numerous studies are examined concerning the forecasting of health insurance premiums through the application of machine learning algorithms. Consequently, the study advocates for the utilization of Extreme Gradient Boosting (XGBoost) and Random Forest Regression to create more precise models for predicting premiums. Furthermore, the study utilizes gradient-boosting models to forecast medical insurance prices. A computational intelligence strategy employing regression-based machine learning algorithms is suggested for estimating healthcare insurance prices.

Various regression models are examined to predict insurance expenses, with comparisons conducted among them. Additionally, innovative ranking methods utilizing machine learning algorithms are employed to categorize cost predictions in health insurance. The approach entails training and assessing a regression model based on machine learning to forecast medical insurance premiums according to individual characteristics.

The objective is to forecast future high-cost patients by employing machine learning algorithms such as Random Forest, Gradient Boosting and Support Vector Machine. Additionally, the charges in a specific Region are analysed using predictive machine learning models, especially focusing on smokers and non-smokers. Another approach that leverages machine learning algorithms is suggested for estimating medical costs based on the individual's gender. Ultimately, an artificial or machine learning model is created to assess the significance of features derived from the dataset.

In summary, these investigations underscore the significance of machine learning techniques in forecasting medical insurance premiums and costs, with various algorithms and methodologies being examined to enhance accuracy and efficiency.

3. RESEARCH OBJECTIVES AND METHODOLOGY

The primary objective of this project is to develop an ***accurate, efficient, and data-driven model*** that can predict medical insurance premiums based on various personal, health, and lifestyle factors. By leveraging ***machine learning algorithms***, the goal is to improve transparency, optimize pricing strategies, and assist both insurance providers and policyholders in making informed decisions.

Specific Objectives: -

- ***Enhance Pricing Accuracy:*** Use historical data and predictive analytics to estimate insurance costs with higher precision.

- ***Identify Key Risk Factors:*** Analyze the impact of parameters such as *age, BMI, smoking status, medical history, and geographic location* on premium calculations.

- ***Optimize Insurance Strategies:*** Help insurers refine policy pricing models based on data-driven insights.

The dataset on *Medical Insurance Price* was obtained from the Geeks for Geeks repository. It consists of seven attributes as outlined in ***Table 1***. Provided by over 1000 customers, the data encompasses a range of health-related parameters. The fees, indicated in INR (₹), reflect the annual charges incurred by the customers.

Attribute	Data Description
Age	The age of the customer
Sex	Gender of the customer (male or female)
BMI	Body Mass Index, measure of body fat based on height and weight
Children	No. of children/dependents covered under the insurance policy
Smoker	Smoking status of the customer
Region	Geographic region of the customer
Charges	Medical insurance premium charged to the customer

Table 1. Overview of the Dataset

➢ RESEARCH PROBLEM

The challenge of accurately predicting ***medical insurance premiums*** lies in the complexity of healthcare costs, individual risk factors, and evolving insurance policies. Traditional actuarial methods rely on statistical models, but they often fail to capture non-linear relationships between variables such as *age, medical history, lifestyle habits, and regional healthcare costs.*

Key Research Challenges: -

1. ***Data Quality & Availability:*** Insurance datasets often contain missing or inconsistent information, affecting model reliability.
2. **Feature Selection & Engineering** – Identifying the most relevant factors that influence premium pricing.
3. ***Model Accuracy & Interpretability:*** Balancing predictive performance with transparency in pricing decisions.
4. ***Bias & Fairness:*** Ensuring that ML models do not reinforce discriminatory pricing based on sensitive attributes.
5. ***Regulatory Compliance:*** Adhering to legal frameworks governing insurance pricing and consumer protection.

➢ RESEARCH DESIGN & VALIDATION

The project follows a ***structured approach*** to developing an accurate and efficient machine learning model for predicting medical insurance premiums. It involves ***data collection, preprocessing, model selection, evaluation, and deployment*** to ensure reliability and practical applicability.

1. **Exploratory Data Analysis (EDA):** Understanding trends and distributions in medical insurance price dataset. EDA is a crucial step in understanding the dataset before applying machine learning models. It helps identify patterns, correlations, and anomalies that influence medical insurance pricing.

   The dataset includes features like *age, BMI, smoking status, number of dependents, region, and medical charges*. Examined for columns, counts, missing values & data types as mentioned in ***Figure 1*** and based on the results, we can conclude that no-null values exist in the dataset.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Age       1338 non-null   int64  
 1   Sex       1338 non-null   object 
 2   BMI       1338 non-null   float64
 3   Children  1338 non-null   int64  
 4   Smoker    1338 non-null   object 
 5   Region    1338 non-null   object 
 6   Charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

```
df.isnull().sum()
```

```
Age         0
Sex         0
BMI         0
Children    0
Smoker      0
Region      0
Charges     0
dtype: int64
```

Figure 1. Info of the Dataset

2. **Feature Analysis & Visualization:** Distributed Sex, Smoker & Region columns in the form of ***Pie Chart*** to analyze the trend. According to the ***Figure 2***, we can summarize that data is almost equally distributed among the Sex and the Region column but there is a large difference in the Smoker column.

   From the middle graph which show the Smoker trend, we can determine a ratio of almost **80:20** in between the smokers & non-smokers.

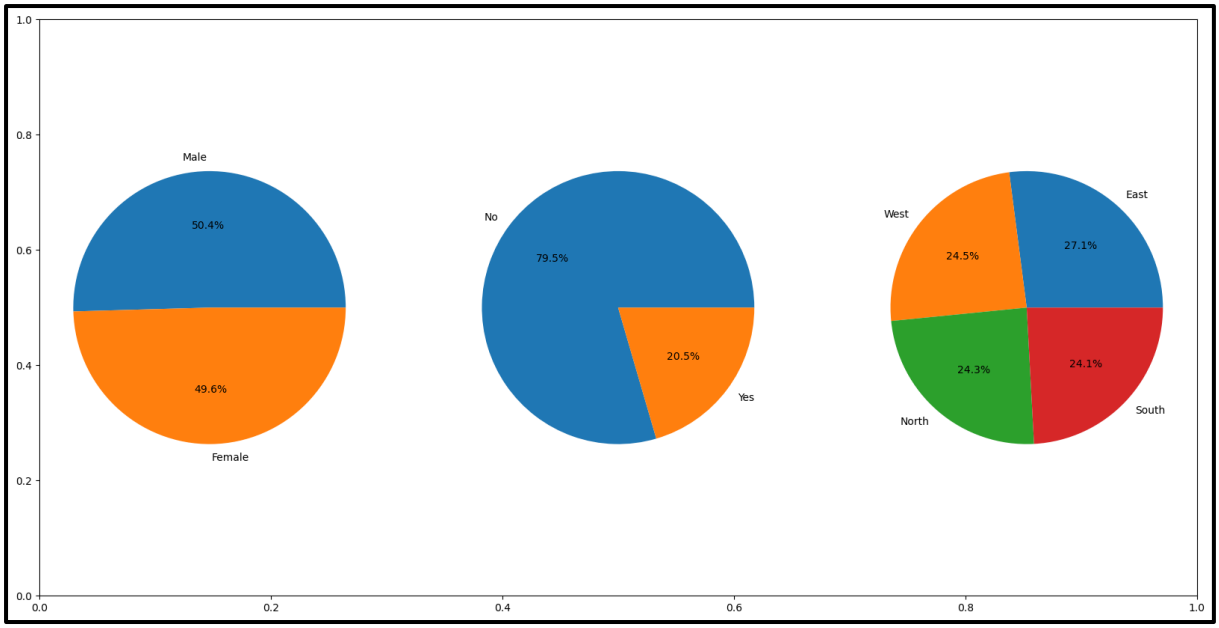

Figure 2. Data Distribution of Sex, Smoker & Region

The following are some conclusions that can be drawn from the graphs as mentioned in ***Figure 3***:

- The charges are higher for Males as compared to Females, but the difference is not much. Hence, we can say approximate equal charges for both the genders.
- The charges are nearly uniform across the four specified Regions.
- ***The premium imposed on Smokers is thrice than that for non-smokers.***

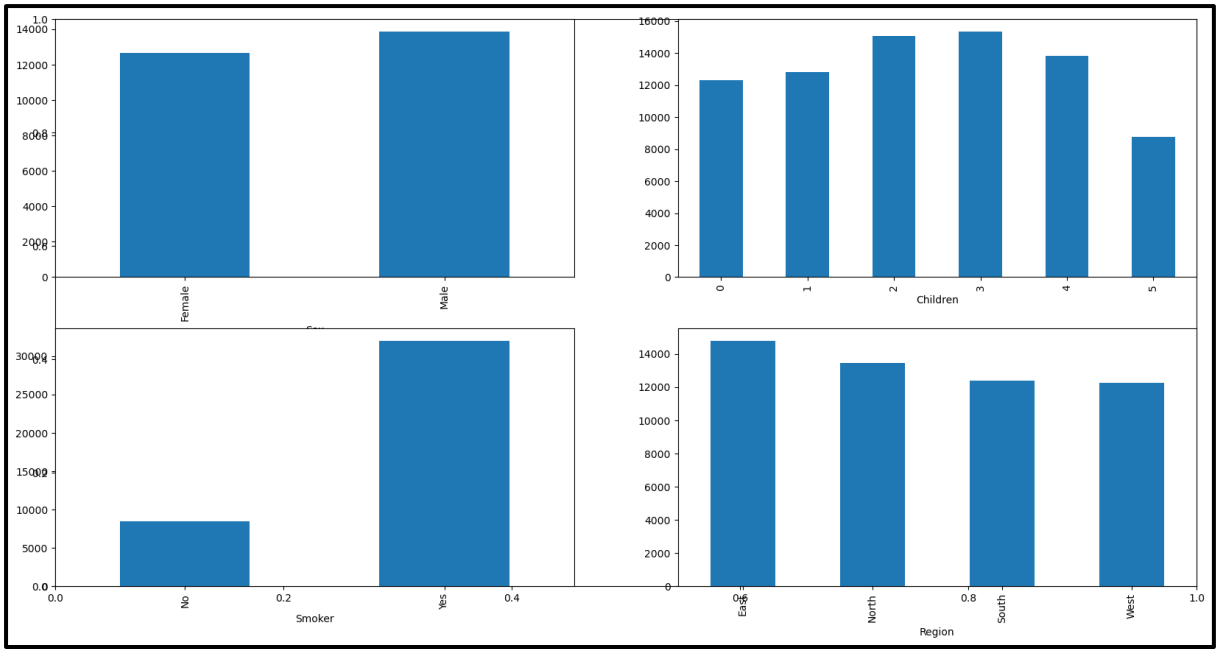

Figure 3. Categorical Data Distribution of Charges

Similarly, we can also observe the trend with the combination of age & smoking status as mentioned in the ***Figure 4***. The trend seems the same for smokers but also premium prices increase with the age.

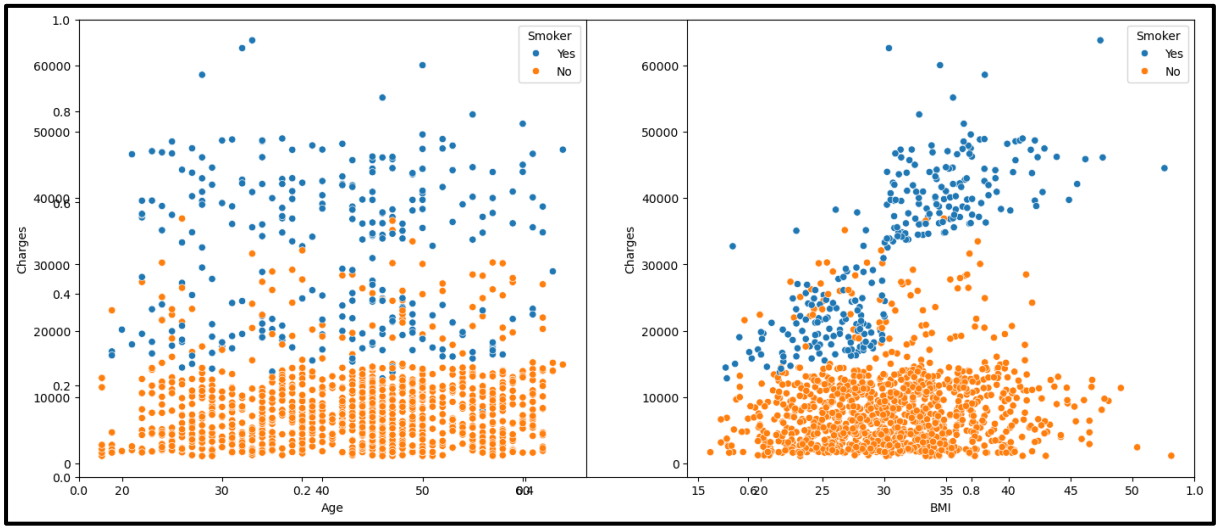

Figure 4. Distribution of Charges with Age

Outliers are the data point that significantly deviates from the overall pattern or central tendency of a dataset, therefore it's important to check the outliers to avoid or reduce any kind of discrepancy in the model. In the dataset, Outliers are only present in BMI column & can be seen in ***Figure 5***.

These outliers can arise due to ***measurement errors, data entry mistakes, or genuine variations in body composition***.

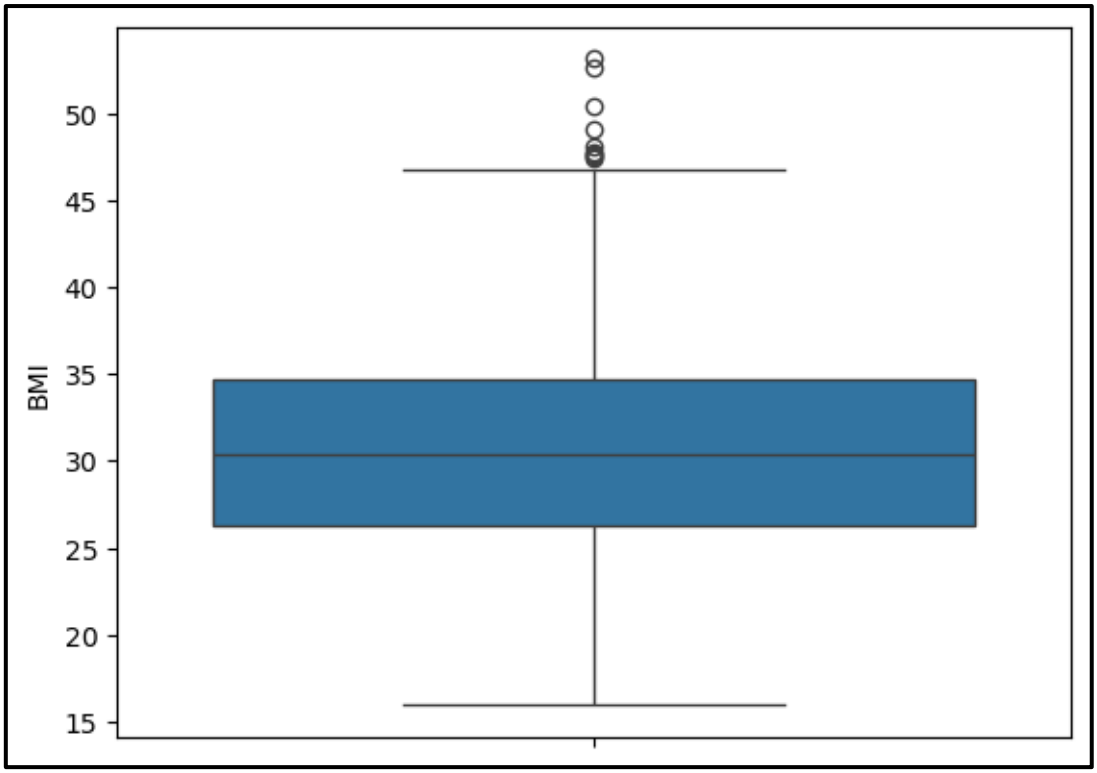

Figure 5. Boxplot of BMI With Outliers

Due to the existence of outliers only in BMI column as mentioned in ***Figure 5***, it is necessary to address these outliers by substituting their values with the mean, as the BMI column contains continuous data. Therefore, using quartile ranges to avoid the outliers and the result can be seen in ***Figure 6***.

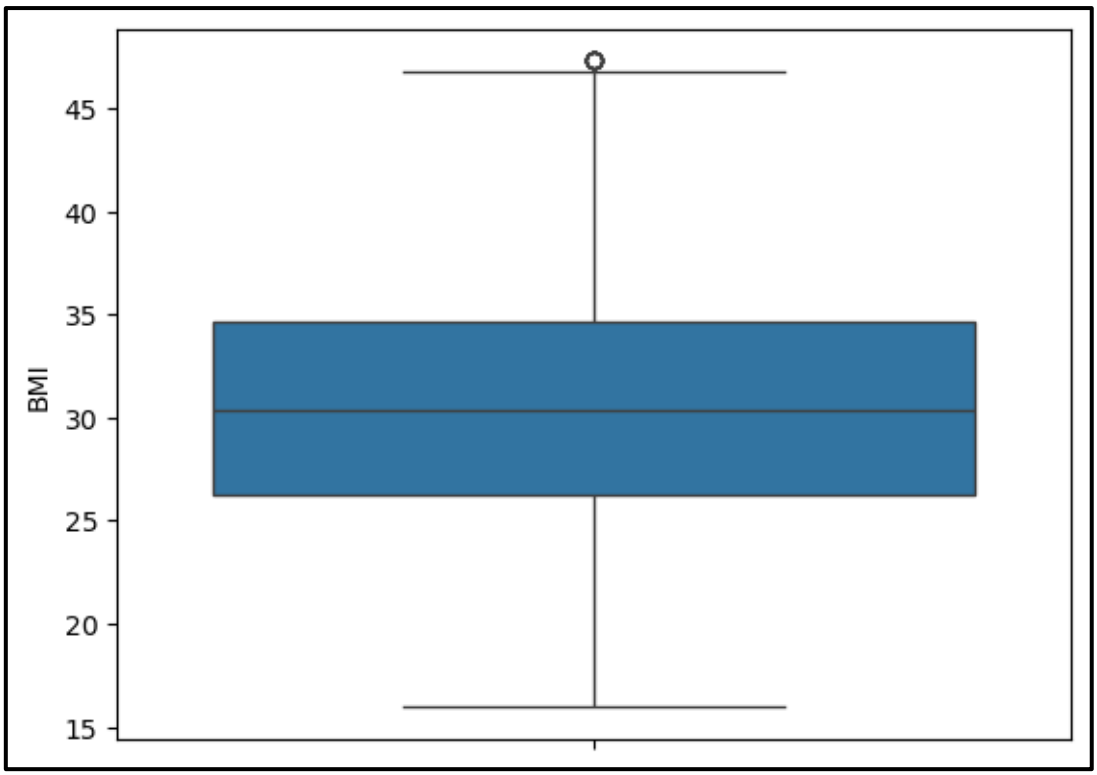

Figure 6. Boxplot of BMI Without Outliers

3. **Feature Engineering:** Feature engineering in machine learning is the process of deriving significant features from unprocessed data, frequently utilizing domain expertise to improve the efficacy of machine learning algorithms.

   Since the success of machine learning models is largely contingent upon the quality of the input data used in training, feature engineering is essential as a preliminary step. It involves identifying the most relevant attributes from the raw training data, customized to suit both the predictive objective and the specific model being employed.

   In the dataset concerning medical insurance prices, the attributes identified are considered essential elements that affect the premium amount. Feature scaling is a

widely used standardization method that normalizes features and reduces the influence of significant scale variations on the models.

In contrast to feature transformation, which alters data from one type to another, feature scaling modifies data regarding its range and distribution while maintaining its original data type.

4. **Machine Learning Algorithms, Training and Validation:** A machine learning model serves as a computational tool that employs algorithms to forecast outcomes based on input data. In contrast to conventional techniques that depend on established equations, these models acquire knowledge directly from the provided dataset.

   Through the examination of a recognized set of input data along with their associated responses (outputs), the model undergoes training to produce predictions for new or previously unencountered data.

   In the ongoing analysis, the regression models used are as follows:

   - Linear Regression
   - Support Vector Machine
   - Random Forest
   - Gradient & Extreme Gradient Boost.

   These models are applied to predict continuous variables based on the provided dataset.

5. **Training & Testing Algorithm:** A training algorithm in Machine Learning (ML) are methods used to optimize model parameters so that the model can learn patterns from data and make accurate predictions. The training process involves ***adjusting weights and biases to minimize errors*** using optimization techniques. The goal is to minimize the error between the predicted outputs and actual values, ensuring the model generalizes well to unseen data.

   Testing algorithms in Machine Learning (ML) ensure that models perform accurately, efficiently, and reliably across different datasets. Machine Learning testing involves evaluating ***data quality, model performance, and generalization ability***.

➢ DATA ANALYSIS TOOL

There are several data analysis tools used for medical price prediction, leveraging machine learning and statistical techniques. Some of the most effective tools include:

- NumPy
- Pandas
- Matplotlib
- Seaborn
- Boxplot

- NumPy & Pandas are essentially for data manipulation and numerical computations.

- Matplotlib, Seaborn & Boxplot are helpful to visualize trends in medical insurance prices.

4. DATA ANALYSIS RESULTS AND INTERPRETATION

After data analysis phase in a medical insurance price prediction which involves examining key variables, identifying patterns, and interpreting insights that influence premium costs. Based on existing studies and models, here are some common findings:

1. **Key Findings from Data Analysis: -**
   - ***Age and Smoking Status:*** Older individuals and smokers tend to have significantly higher insurance costs due to increased health risks.
   - ***Regional Differences:*** Insurance costs vary based on geographic location, reflecting differences in healthcare accessibility and costs.
   - ***Dependents and Policy Costs:*** The number of dependents covered under a policy influences premium pricing.

2. **Model Performance & Evaluation: -**
   - Regression Models (***Random Forest& Extreme Gradient Boosting***) show high accuracy in predicting insurance costs.
   - ***Feature Importance Analysis*** highlights smoking status, BMI, and age as the most influential factors.

3. **Interpretation of Results: -**
   - ***Smokers pay significantly higher premiums*** due to increased health risks.
   - Young, healthy individuals with lower BMI tend to have lower insurance costs.
   - ***Machine learning models improve pricing transparency*** by identifying key cost-driving factors.

Summary, the dataset contains health-related parameters of the customers. By using the health parameters attributes, a machine learning (ML) model is built by performing the training and validation on the train feature dataset.

On comparing all the models as mentioned in ***Figure 7***, we can conclude that out of all the models Random Forest & Gradient Boost is giving the highest accuracy which means predictions made by these models are close to the real values as compared to the other models.

Based on the dataset which is quite similar with real time dataset, we conclude that this model will work properly in real-life scenarios and help insurers to estimate patterns in the relation between the independent features and the medical premium for insureds.

Model	Train Accuracy	Test Accuracy	CV Score
Linear Regression	0.648	0.698	0.656
Support Vector Machine	-0.098	-0.074	-0.104
Random Forest	0.959	0.750	0.734
Gradient Boost	0.819	0.785	0.756
XG Boost	0.985	0.710	0.668

Figure 7. Comparison of All Models

At last, the ***accuracy of Train, Test & CV Score*** of the model comes out to be ***approximate 79%*** which signifies that the performance of model is ***HIGH***.

5. CONCLUSION

In the field of healthcare, ***Machine Learning (ML) emerges as an impressive tool that can execute tasks more swiftly than human counterparts***. The incorporation of machine learning into health insurance operations have significant potential in improving accuracy, efficiency, and transparency in premium estimation. By leveraging historical data and advanced algorithms, ML models can identify ***key factors influencing insurance costs such as age, BMI, smoking status, medical history, and geographic location.***

By automating repetitive tasks, ML allows insurance professionals to focus their efforts on improving the overall experience for policyholders. This transition not only optimizes administrative functions but also enhances the effectiveness of patient care, benefiting all stakeholders within the healthcare ecosystem, including patients, hospitals, physicians, and insurance companies.

This paper has examined ***various machine learning regression models designed to predict medical insurance prices based on specific characteristics. By utilizing these models, insurance providers can accelerate the development of customized plans for individuals, thus conserving significant time and resources in policy formulation.***

In summary, the integration of machine learning has the potential to transform the health insurance sector, making processes quicker, more economical, and ultimately more responsive to the requirements of both policyholders and insurers.

➦ **Click Here to Access Project Link.**

7. RECOMMENDATIONS AND LIMITATIONS OF THE STUDY

➢ RECOMMENDATIONS

To improve the accuracy, efficiency, and fairness of medical insurance price prediction, consider the following recommendations:

- Ensure ***clean, complete, and unbiased datasets*** to avoid skewed predictions.
- Use ***feature engineering*** to extract meaningful insights from variables like ***age, BMI, smoking status, and medical history***.
- Apply ***normalization*** and ***encoding*** techniques to handle categorical and numerical data effectively
- Experiment with ***ensemble learning techniques*** (like: ***Random Forest, Gradient Boosting, XGBoost***) for better predictive performance.
- Incorporate ***wearable health data*** and ***electronic health records*** for dynamic premium adjustments.
- Explore ***reinforcement learning*** for adaptive pricing models based on evolving health conditions.
- Align ML-driven pricing models with ***insurance regulations and consumer protection laws***.
- Ensure ***data privacy and security*** in handling sensitive medical records.

➢ LIMITATIONS OF THE STUDY

While Machine Learning (ML) has significantly improved medical insurance price prediction, several challenges and limitations remain:

- ***Incomplete or biased datasets*** can lead to inaccurate predictions.
- ***Privacy concerns*** restrict access to detailed medical records, limiting model effectiveness.
- ML models may unintentionally ***reinforce biases*** based on demographic factors like age, gender, or location.
- ***Ethical concerns*** arise when pricing disproportionately affects certain groups.
- ***Healthcare costs fluctuate***, requiring frequent model retraining to stay relevant.
- ***Real-time predictions demand*** efficient processing, which can be costly.

8. BIBLIOGRAPHY

1. Medical Insurance Price Prediction using Machine Learning - Python, Geeks for Geeks, 05 Sep, 2024.
2. Health Insurance Cost Prediction Using Machine Learning, International Research Journal of Engineering and Technology (IRJET), 04 Apr,2024.
3. Medical Insurance Premium Prediction Using Regression Models, International Journal for Research Trends and Innovation (IJRTI), 2023.
4. Insurance-Premium-Prediction, GitHub, 2024.
5. Health Insurance Dataset EDA | Excel, Kaggle, 2023.
6. BMI Dataset, Kaggle, 2022.
7. Medical-Insurance-Price-Prediction-System, GitHub, 2024
8. How to Unpack a PKL File in Python, 18 Jun, 2024.
9. upGrad Certification in Data Science with AI, 2023.
10. Dataset Link, Geeks for Geeks.