A Project report on

Medical Insurance Premium Prediction

A Dissertation submitted to JNTU Hyderabad in partial fulfillment of the academic requirements for the award of the degree.

Bachelor of Technology

in

Computer Science and Engineering

Submitted by

T. Bhagya Laxmi
(20H51A05F7)

Yasmeen
(20H51A05G1)

B. Sree Harsha
(20H51A05M9)

Under the esteemed guidance of

Mr. B. Sivaiah
(Associate Professor)

Department of Computer Science and Engineering

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

(UGC Autonomous)

*Approved by AICTE *Affiliated to JNTUH *NAAC Accredited with A^{+} Grade

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD - 501401.

2020- 2024

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Major Project Phase I report entitled**"Medical Insurance Premium Prediction"** being submitted by T.Bhagya Laxmi (20H51A05F7),Yasmeen (20H51A05G1), B.Sree Harsha(20H51A05M9) in partial fulfillment for the award of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide work carried out his/her under my guidance and supervision. The results embodies in this project report have not been submitted to any other University or Institute for the award of any Degree.

Mr.B.Sivaiah	**Dr. Siva Skandha Sanagala**	**External Exmainer**
Associate Professor	**Associate Professor and HOD**	
Dept. of CSE	**Dept. of CSE**	

ACKNOWLEDGEMENT

With great pleasure we want to take this opportunity to express my heartfelt gratitude to all the people who helped in making this project work a grand success.

We are grateful to **Mr.B.Sivaiah, Associate Professor** , Department of Computer Science and Engineering for his valuable technical suggestions and guidance during the execution of this project work.

We would like to thank **Dr. Siva Skandha Sanagala,** Head of the Department of Computer Science and Engineering, CMR College of Engineering and Technology, who is the major driving forces to complete my project work successfully.

We are very grateful to **Dr. Vijaya Kumar Koppula**, Dean-Academics, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Major Dr. V A Narayana,** Principal, CMR College of Engineering and Technology, for giving permission to carry out this project in a successful and fruitful way.

We would like to thank the **Teaching & Non- teaching** staff of Department of Computer Science and Engineering for their co-operation

We express our sincere thanks to **Shri. Ch. Gopal Reddy**, Secretary, CMR Group of Institutions, for his continuous care.

Finally, We extend thanks to our parents who stood behind us at different stages of this Project. We sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project work.

T. Bhagya Laxmi 20H51A05F7
Yasmeen 20H51A05G1
B. Sree Harsha 20H51A05M9

TABLE OF CONTENTS

List of Figures

List Of Tables

ABSTRACT

Insurance is a policy that helps to cover up all loss or decrease loss in terms of expenses incurred by various risks. A number of variables affect how much insurance costs. These considerations of different factors contribute to the insurance policy cost expression. Machine Learning (ML) in the insurance sector can make insurance more effective. In the domains of computational and applied mathematics the machine learning (ML) is a well-known research area. ML is one of the computational intelligence aspects when it comes to exploitation of historical data that may be addressed in a wide range of applications and systems. There are some limitations in ML so; Predicting medical insurance costs using ML approaches is still a problem in the healthcare industry and thus it requires few more investigation and improvement. Using the machine learning algorithms, this study provides a computational intelligence approach for predicting medical insurance costs. The proposed research approach uses Linear Regression, Decision Tree Regression and Gradient Boosting Regression and also stream lit as a framework. We had used a medical insurance cost dataset that was acquired from the KAGGLE repository for the cost prediction purpose, and machine learning methods are used to show the forecasting of insurance costs by regression model comparing their accuracies.

CHAPTER 1
INTRODUCTION

CHAPTER 1
INTRODUCTION

1.1 Problem Statement

In the modern world, it is essential to have health insurance, The goal of this project is to help individuals understand the amount of money they may need for health insurance based on their personal health status. This can assist individuals in focusing more on the health related aspects of insurance rather than the unnecessary ones. Most people have a relationship with a public or private health insurance provider. The factors that influence insurance costs vary from company to company.

The aim of the medical insurance premium prediction project is to develop a predictive model that accurately estimates the insurance premiums for individual policyholders based on various demographic, lifestyle, and health-related factors. By leveraging machine learning techniques and historical insurance data, the goal is to create a model that can assist insurance companies in determining appropriate premium rates for potential customers, thereby optimizing risk management and pricing strategies. The model should take into account factors such as age, gender, BMI, smoking status, region, pre-existing medical conditions, and other relevant variables to predict the annual or monthly insurance premiums with a high level of accuracy. Additionally, the model should be interpretable and transparent, providing insights into the factors influencing premium calculations to enhance decision-making processes for both insurers and policyholders.

1.2 Research Objective

To develop a predictive model that accurately estimates medical insurance costs for individuals based on relevant demographic, health- related, and lifestyle factors. The goal is to provide insurance companies and policyholders with a reliable tool for pricing and planning, ultimately improving the transparency and fairness of the insurance market.

This objective outlines the main aim of your research, which is to create a predictive model for medical insurance costs, and highlights the potential benefits of such a model for

both insurance providers and individuals seeking coverage. The research objective of the medical insurance premium prediction project is to develop a robust predictive model that accurately estimates individual medical insurance premiums based on various demographic, health, and lifestyle factors. By leveraging advanced machine learning techniques and statistical analysis, the project aims to provide insurance companies with a tool that can effectively assess risk and determine fair premiums for policyholders. Additionally, the research seeks to explore the potential impact of different variables such as age, gender, pre-existing conditions, geographic location, and lifestyle choices on insurance premiums. Through comprehensive data analysis and model training, the project aims to enhance the transparency and fairness of the insurance pricing process while optimizing risk management strategies for insurers. Ultimately, the objective is to improve the accessibility and affordability of medical insurance for individuals while ensuring the financial sustainability of insurance providers.

1.3 Project Scope and Limitations

Scope:

- Personalization:

  Advanced prediction models can consider individual characteristics, such as age, gender, medical history, and lifestyle factors, to personalize premium rates.

- Cost Control:

  Accurate prediction models help insurance companies manage costs by aligning premiums with expected claims, reducing underwriting losses, and minimizing adverse selection.

- Customer Retention:

  Personalized premium rates can enhance customer satisfaction and retention, as policyholders feel their premiums are fair and tailored to their needs.

- Portfolio Management:

  Insurers can use premium prediction to effectively manage their portfolios, balancing risk and revenue to optimize their business.

- Regulatory Compliance:

  Premium prediction must comply with local and national insurance regulations, ensuring fairness and non-discrimination.

Limitation:

- Data Availability:

  prediction relies heavily on data, and the accuracy of the predictions is limited by the quality and availability of data. Inaccurate or incomplete data can lead to unreliable predictions.

- Privacy Concerns:

  Collecting and using personal data for premium prediction may raise privacy concerns. Insurers need to handle customer data responsibly and in accordance with relevant privacy regulations.

- Regulatory Constraints:

  Insurance companies must be to regulatory guidelines when determining premium rates. This can limit the flexibility in setting rates based on predictive models.

- Data Quality:

  Accuracy of the predictive model heavily depends on the quality and completeness of available data. Incomplete or biased datasets may lead to inaccurate predictions and unfair premium assessments.

- Complexity of Factors:

  Insurance premiums are influenced by a complex array of factors including socio-economic status, cultural differences, and evolving healthcare trends. Accounting for all these factors accurately within the model may be challenging.

CHAPTER 2
BACKGROUND WORK

CHAPTER 2
BACKGROUND WORK

2.1 Health Insurance Cost Prediction by Using Machine Learning

2.1.1. Introduction

The goal of this research is to help individuals understand the amount of money they may need for health insurance based on their personal health status. This can assist individuals in focusing more on the health related aspects of insurance rather than the unnecessary ones. In the modern world, it is essential to have health insurance, and most people have a relationship with a public or private health insurance provider. The factors that influence insurance costs vary from company to company. Additionally, some people in rural areas may not be aware that the Indian government offers free health insurance to those who are below the poverty line. However, the process can be complex, and some rural residents either get private health insurance or make no investment at all. Additionally, people may be vulnerable to being misled into paying for expensive health insurance that they don't need.

Our research does not provide an exact amount required by any specific health insurance provider, but it does give a general sense of the cost a person may incur for their own health insurance. This is a preliminary estimate and does not adhere to any particular company, so it should not be the only factor considered when choosing health insurance. Early estimation of health insurance costs can help individuals consider the required amount more thoughtfully.

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.

2.1.2 Merits

- Risk Management:
  Predictive modeling helps insurance companies assess and manage risk effectively.
- Personalization:
  Premiums can be tailored to individual health and lifestyle factors, making insurance more affordable for many.
- Cost Control:
  Predictive models enable the identification of high-risk policyholders and the implementation of preventive measures to reduce claims.
- Competitive Edge:
  Accurate premium predictions can give insurance companies a competitive advantage in the market.
- Customer Satisfaction:
  Tailored premiums can lead to higher customer satisfaction as policyholders feel their premiums are more fairly priced

Demerits

- Privacy Concerns:
  Extensive data collection for predictive modeling can raise privacy concerns and potential data breaches.
- Bias and Fairness:
  Predictive models may introduce bias if training data contains historical disparities, leading to unfair premium pricing.
- Complexity:
  Developing and maintaining accurate predictive models requires substantial computational resources and expertise.

2.1.3. Implementation

INPUT DATA

The following article discusses a dataset that can be accessed on the Kaggle website for the purpose of training and testing. This dataset is saved in a CSV file and is well organized. It is available at the specified link for those interested in using it. In order to accurately predict the cost of health insurance, it is necessary to clean the dataset before applying regression algorithms. The data shows that age and smoking status have the most significant impact on the amount of insurance, with smoking having the greatest effect. However, factors such as family medical history, BMI, marital status, and geography also play a role. Children's property was found to have little impact on the prediction, so it was removed from the input for the regression model to improve efficiency and accuracy.. The data shows that age and smoking status have the most significant impact on the amount of insurance, with smoking having the greatest effect.

This is a preliminary estimate and does not adhere to any company. These algorithms are designed to make classifications or predictions using statistical techniques, which can uncover key insights in data mining processes. The outcomes from these insights can be seen in the given figure key growth indicators in businesses and applications, if used correctly. The data shows that age and smoking status have the most significant impact on the amount of insurance, with smoking having the greatest effect. They will be able to make a more informed decision. Additionally, it may suggest.

```
In [7]: data.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
```

Fig – 2.1.3 – Data set of model

METHODOLOGY

1. Machine Learning:

Machine Learning is a subset of computer science and AI that involves using data and algorithms replicate the way that humans learn. These algorithms are designed to make classifications or predictions using statistical techniques, which can uncover key insights in data mining processes. The outcomes from these insights can have a significant impact on key growth indicators in businesses and applications, if used correctly. The data shows that age and smoking status have the most significant impact on the amount of insurance, with smoking having the greatest effect. However, factors such as family medical history, BMI, marital status, and geography also play a role. The data shows that age and smoking status have the most significant impact on the amount of insurance, with smoking having the greatest effect. However, factors such as family medical history, BMI, marital status, and geography also play a role.

2. Linear Regression Algorithm:

Linear regression is a machine learning algorithm that is based on the concept of "supervised learning." It is used to predict the value of a dependent variable (y) based on the value of an independent variable (x). Essentially, this means that linear regression is used to determine how closely a dependent variable is related to an independent variable, and then make predictions based on that relationship. This is a very useful tool for data analysis, as it allows analysts to understand complex patterns and relationships in data, and to make more accurate predictions about future outcomes.

3. Multiple Linear Regression:

Similar to simple linear regression, multiple regression is a statistical procedure that examines the degree of association between a set of independent variables and a dependent variable. There is just one independent and one dependent variable in basic linear regression, but there are numerous predictor variables in multiple linear regression and the value of dependent variable(Y) is now calculated depending on the values of the predictor variables. it is assumed that there is no dependency among the predictor variables. Suppose if the target value is dependent on “n” independent variables then the regressor fits

the regression line in a N dimensional space. The regressor line equation is now modified where “a” is the Y-intercept value and < b1, b2, b3, …., bn > are the regression coefficients associated with the n independent variables and is the error term.

4. Support Vector Machine Algorithm:

SVM, or Support Vector Machine, is a widely used supervised learning algorithm for solving classification and regression problems, with a focus on classification in machine learning. The goal of SVM is to determine the best line or decision boundary that can separate a multi-dimensional space into different classes, enabling the efficient classification of new data points in the future. This optimal decision boundary is known as a hyperplane, which is created using extreme vectors and points called support vectors. SVM is a popular choice due to its ability to effectively classify data and handle high dimensional spaces. This is a preliminary estimate. This optimal decision boundary is known as a hyperplane.

5. Random Forest Regression:

The Random Forest approach utilizing bootstrapping involves the use of multiple decision trees generated from the data and combined through ensemble learning techniques. This method often leads to accurate predictions and classifications by averaging the results of the randomly select.

6. Gradient Boosting Algorithm:

Gradient boosting is a highly popular machine learning technique for analyzing tabular data sets. It is well-known for its ability to handle missing values, outliers, and large categorical values in the features, as well as its ability to detect nonlinear relationships between the target and the features. This makes it a powerful tool for data analysis and prediction.

Training And Predicting of Dataset

1. **Training:** After the necessary data has been formatted and prepared, the model can begin its training and testing phases. A key focus during the training phase is choosing the appropriate model for the task at hand. This may involve deciding on the optimal modelling strategy or determining the best parameter values for a particular model. In some cases, this process is referred to as model selection because various models may be tested and the one that performs the best, is ultimately chosen, which is created using extreme vectors and points called support vectors.

2. **Prediction:** The model used for predicting the insurance sum for health was based on the relationship between certain features and the label. The accuracy of this prediction was determined by the extent to which the expected value matched the actual amount. In order to improve the accuracy, the model employed various characteristics, methods, and train-test split sizes. It was found that the amount of data used for training had a significant impact on the accuracy, with a larger train size leading to better results. The model also employed multiple algorithms in order to forecast the premium amount, and showed how each attribute affected the outcome.

2.2 Medical Insurance Cost Prediction using Machine Learning

2.2.1 Introduction

We live on a planet full of threats and uncertainty. Including People, households, durables, properties are exposed to different risks and the risk levels can vary. These risks range from risk of health diseases to death if not get protection, and loss in property or assets. But risks cannot usually be avoided, so the world of finance has developed numerous products to shield individuals and organizations from these risks by using financial capital to shield the Therefore, Insurance is one of the policies that either decreases or removes loss costs incurred by various risks. The value of insurance in the lives of individuals. That's why it becomes important for insurance companies to be sufficiently precise to measure the amount covered by this specific policy and the insurance charges which must be paid for it. Various parameters or factors play an important role in estimating the insurance charges and Each of these is important. If any factor is omitted or changed when the amounts are computed then, the overall policy cost changes. It is therefore very critical to carry out these tasks with high accuracy. So, the possibility of human mistakes is high so insurance agents also use different tools to calculate the insurance premium. And thus, ML is beneficial here. ML may generalize the effort or method to formulate the policy. These ML models can be learned by themselves. The model is trained on insurance data from the past. The model can then accurately predict insurance policy costs by using the necessary elements to measure the payments as its inputs. This decreases human effort and resources and improves the company's profitability. Thus, the accuracy can be improved with ML. Our goal is to predict insurance costs. The value of insurance fees is based on different variables. As a result, insurance fees are continuous. Regression is the best choice available to fulfill our needs. We use multiple linear regression in this analysis since there are many independent variables used to calculate the dependent(target) variable. For this study, the dataset for cost of health insurance is used. Preprocessing of the dataset done first. Then we trained regression models with training data and finally evaluated these models based on testing data. In this article, we used several models of regression, for example, multiple linear regression, Decision Tree Regression and Gradient Boosting Regression.

2.2.2 Merits

- Customer Satisfaction:

  Tailored premiums can lead to higher customer satisfaction as policyholders feel their premiums are more fairly priced.

- Fraud Detection:

  Predictive models can help detect fraudulent claims by identifying unusual patterns and behaviors.

- Efficiency:

  Automation of premium prediction processes can improve efficiency within insurance companies.

- Data-Driven Decision Making:

  Predictive modeling encourages data-driven decision- making, improving underwriting processes.

Demerits

- Ethical Concerns:

  Predictive modeling raises ethical questions about fairness, transparency, and the potential for discrimination.

- Resistance from Policyholders:

  Some may be resistant to premium predictions using models, as they may not fully understand the process or have concerns about its accuracy.

- Limited Data for Some Factors:

  Not all relevant factors affecting an individual's health and insurance risk are easy to capture in data, limiting prediction accuracy.

- Maintenance Costs:

  Continuous maintenance and updates of predictive models can be costly.

- Overreliance on Models:

  Overreliance on predictive models may lead to a lack of human judgment and expertise in underwriting and risk assessment.

2.2.3 Implementation

The objective of the study is to prophetic the insurance cost supported age, BMI, kid number, the region of the person living, sex, and whether or not a shopper is smoking or not, drinks alcohol or not, having diabetics or not . These options contribute to our target variable prediction of insurance costs. For the measuring of the value of insurance, many regression models are applied during this study. The dataset is split into 2 sections. One half for model training and also the other part for model analysis or testing.

During this study, the info set is separated into two-part the first half is termed coaching knowledge and also the second called take a look at data, training data makes up for eighty percent of the whole data used, and the rest for test data. all of those models are trained with the training data part and so evaluated with the test data. The accuracy is checked with the assistance of r2 score.

1. **Machine Learning:**

Machine learning is a branch of computer science and artificial intelligence that uses data and algorithms to replicate the way humans learn. These algorithms are designed to use statistical methods for classification or prediction to reveal important insights during the data mining process. When used correctly, the results of these insights can significantly contribute to business and application grow. Data shows that age and smoking have the biggest impact on insurance coverage, while smoking has the biggest impact. However, factors such as family medical history, body weight, marital status, and region of residence also play a role. Data shows that age and smoking have the biggest impact on insurance premiums, while smoking has the biggest impact. However, factors such as family history, body measurements, marital status, and region of residence also play a role

2. **Linear Regression Algorithm:**

   Linear regression is a machine learning algorithm based on the concept of "supervised learning". It is used to predict the value of variable (y) based on the value of variable (x). Essentially, this means that linear regression is used to determine how well the variables are related to the independent variables, and then the prediction is based on that relationship. Again, prediction of future outcomes.

3. **Support Vector Machine Algorithm:**

   SVM or Support Vector Machine is a widely used supervised learning algorithm used to solve classification and regression problems, focusing on classification in machine learning border. The site can be divided into different classes, making it useful for new information for the future. This well-defined boundary is called the hyperplane and is created using points called extreme vectors and support vectors. SVM is a popular choice due to its ability to classify data well and handle high density. This is a preliminary predict

4. **Decision Tree Regression**

   Decision trees are a category of supervised ML models employed by the Train Using AutoML tool. They classify or regress data based on true or false answers to specific questions. When visualized, the resulting structure takes the form of a tree, featuring various types of nodes, including the root, internal nodes, and leaf nodes. The root node serves as the initial point in the decision tree, from which branches extend to internal nodes and eventually lead to leaf nodes. The leaf nodes represent the ultimate classification categories or real values within the decision tree. Notably, decision trees are renowned for their simplicity and interpretability, making them easy to understand and explain.

5. **Random Forest Regression:**

   The bootstrapped random forest approach uses multiple decision trees constructed from data and combined through the learning process. This method generally provides accurate prediction and classification by averaging the results of selected trees.

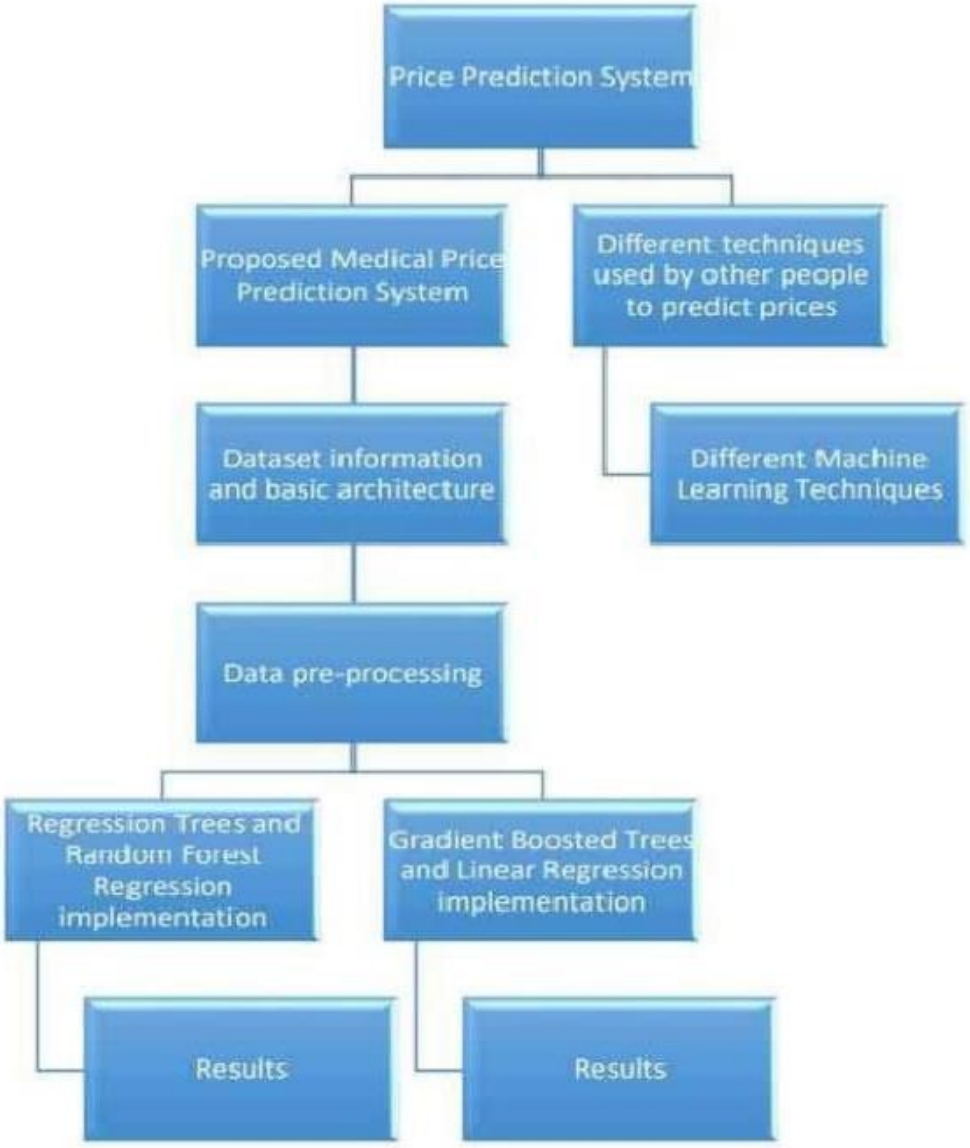

Fig – 2.1.2 – Flow chart

2.3 Health Insurance Cost Prediction Using Regression Models

2.3.1 Introduction

Public health is an integral part of the society, of the country and of the world we live in and is an important matter of concern. Human lives and public health may be endangered by natural calamities, global epidemic and pandemics, global crisis of medical aids, etc. which increases the vulnerability of public health. People can meet with unavoidable and unforeseen circumstances at any point of their lifetime. Individuals, families, companies and properties are uncovered and uninsured to diverse hazard forms, natural calamities, and the likelihood can shift. These perils include the possibility of mortality, health, and property disaster or resource depletion. People's lives revolve around two main elements: life and prosperity. However, keeping a safe and sound distance from unforeseen events is impossible.

Today, data has evolved drastically since the recent past decade and insurance carriers have access to it. The health insurance system is exploring ways to use predictive modelling to boost their business operations and services. Computer algorithms and Machine Learning (ML) are used to study and analyze the historical insurance data and predict new output values based on trends in customer behavior, insurance policies and data-driven business decisions, and supports in formulation of new schemes. Besides, most insurance companies use conventional databases to store their data which is primarily structured data. Moreover, merely 10-15 percent of the total data available is processed for gaining insights. Thus, transformation of the data is necessary to gain valuable insights that may be very crucial for the growth of such companies. Therefore, analyzing the structured data as unstructured data for making better decisions requires statistical and Machine Learning (ML) techniques.The main advantage of ML is that it can be effectively applied to a massive volume of structured, semi-structured, or unstructured datasets. The ML model can be used across multiple value chains to understand the weight-age of risk involved, claims made and customer behavior with greater predictive accuracy. ML applications in the health insurance sector include various tasks such as understanding risk tolerance and premium leakage leading to inaccurately pricing of premiums, loss deterrence, claims handling, expense management, subrogation, litigation, and fraud identification.

The calculation of health insurance charges in the traditional process are a hefty task for the insurance companies. Therefore, ML may generalize the exertion or strategy to define such an approach. These models can perform self-learning to predict the cost of insurance using past insurance data of the companies. The model inputs are the main parameters that are utilized to calculate the instalments made. This enables the algorithm to precisely estimate the disbursement of insurance coverage. In this way, the correctness can be progressed with ML. The objective of the proposed model is to perform rapid estimation and prediction of insurance charges at a hospital incurred by a patient, using ML models upon the Kaggle dataset. Thus, this paper develops a real-time insurance cost price prediction system named ML Health Insurance Prediction System (MLHIPS) using ML algorithms which will aid the insurance companies in the market for easy and rapid determination of values of premiums and thereby curb down health expenditure. The proposed model incorporates and demonstrates different regression models such as Multiple Linear Regression, Ridge Regression, Simple Linear Regression, Lasso Regression and Polynomial Regression to predict the insurance costs and compares the models based on their results.

2.3.2 Merits:

- Diverse Model Selection:

  By considering multiple algorithms such as linear regression, support vector machine, decision tree, random forest, and gradient boosting, the project benefits from a diverse set of modeling approaches. This allows for exploration of various modeling techniques to identify the most suitable one for the dataset and problem at hand.

- Robustness and Generalization:

  Utilizing a combination of algorithms can enhance the robustness of the predictive model. Each algorithm has its own strengths and weaknesses, so ensemble techniques like random forest and gradient boosting can help mitigate individual algorithm weaknesses and improve overall predictive performance. This can lead to better generalization and performance on unseen data, which is crucial for the reliability of the predictive model in real-world applications.

- Interpretability and Transparency:

  Algorithms such as linear regression and decision trees offer interpretability, providing insights into the factors influencing medical insurance costs. Interpretability is essential for stakeholders such as healthcare providers, insurers, and policy-makers to understand and trust the predictions made by the model.

- Scalability and Efficiency:

  Some algorithms, like linear regression and decision trees, are computationally efficient and scalable to large datasets. This ensures that the predictive model can handle the potentially vast amount of medical insurance data efficiently, making it suitable for real-world deployment.

Demerits:

- Complexity and Model Selection:

  Managing and selecting from multiple algorithms can introduce complexity to the project. The process of comparing and evaluating different algorithms requires expertise and computational resources. Choosing the most appropriate algorithm for the specific dataset and problem domain may be challenging and time-consuming.
- Computation and Resource Requirements: Some algorithms, particularly ensemble methods like random forest and gradient boosting, can be computationally expensive and resource-intensive. Training and tuning these models may require substantial computational resources, including processing power and memory.
- Overfitting and Model Tuning: Ensemble methods such as random forest and gradient boosting are susceptible to overfitting, especially if not properly tuned. The process of hyperparameter tuning for these algorithms can be iterative and require careful optimization to achieve optimal performance without overfitting.

2.3.3 Implementation

The regression techniques used are the statistical methods that establishes the association between a target or dependent variable and a set of independent or predictor variables. It assumes that both the target and the predictor variables are having numerical values and there exists some kind of corelation between the two.

The models that we are implementing in our problem are discussed below,

A. Model Selection:

1) Simple Linear Regression:

In simple linear regression, the target variable(Y) is dependent on a single independent variable(X) and the model establishes a linear relationship among these two variables. The linear regression model tries to fit the regressor line between the independent(X) and dependent(y) variable. The equation of the line is given by: $Y = a + bX$

where “a” and “b” are the models parameters called as regression coefficients, “a” is the value of the Y intercept that the line makes when X is equal to zero and “b” is the slope that signifies the change of Y with the change of X. More the value of “b” means a small change in X causes a significant change

in Y, and vice versa. The value of “a” and “b” can be found by Ordinary Least Square method. In linear regression models the values predicted may not be accurate always, there will always be some difference hence we add an error term to the original equation that accounts for the difference and thus help in making better predictions.

$Y = a + bX$

Assumptions in Linear Regression

- The sample size of data should exceed the number of available parameters.
- Only over a restricted range of data the regression can be valid.
- Error term is normally distributed. This also means that the mean of the error has expected value of 0.

2) Multiple Linear Regression:

Similar to simple linear regression, multiple regression is a statistical procedure that examines the degree of association between a set of independent variables and a dependent variable. There is just one independent and one dependent variable in basic linear regression, but there are numerous predictor variables in multiple linear regression and the value of dependent variable(Y) is now calculated depending on the values of the predictor variables. it is assumed that there is no dependency among the predictor variables. Suppose if the target value is dependent on "n" independent variables then the regressor fits the regression line in a N dimensional space. The regressor line equation is now modified where "a" is the Y-intercept value and the regression coefficients associated with the n independent variables and is the error term.

3) Polynomial Regression:

Polynomial Regression is yet another a special case of linear regression. In Linear regression the model tries to fit a straight regression line between the dependent and independent variable. In scenarios where there doesn't exist linear relationship between the target and predictor variable then instead of a straight line a curve is being fitted against the two variables. This is accomplished by fitting a polynomial equation of degree n on the non-linear data which establishes a curvilinear relationship among the dependent and independent variables. In polynomial regression the assumption that the independent variables must be independent of each other is not mandatory.

The following are some of the benefits of applying polynomial regression:

- Polynomial Regression offers the best estimate of the relationship between the dependent and independent variable.
- Higher degree polynomial generally provides a good fit on the dataset.
- Polynomial Regression basically fits a wide range of curves of varying degree to the dataset.

Drawbacks of applying Polynomial Regression:

- These are too sensitive to the existence of outliers in the dataset, as outliers cause the model's variance to rise**.** When the model comes across an unknown data item, it underperforms.

4) Ridge Regression:

Ridge regression is a standard model tuning process used to analyse the data suffering from multicollinearity. This is a way to approximate the coefficients of the regression model when the independent variables are firmly related or there exists an association between them. Ridge Regression's main objective is to take the dataset and fit a new line into it in a way that does not overfit the model. For this purpose, ridge regression adds an insignificant amount of bias that determines the fitting of the line into the data. We obtain a substantial reduction in variance, which leads to an increase in the accuracy value by the addition of bias. The least Square determines the values of the parameters for the equation which diminishes the sum of squared residuals. But in contrast the Ridge Regression regulates the value for parameters that results in minimization of the sum of squared residuals along with an additional term λ*b2. Ridge Regression performs L2 regularization. Where y* = a + bX is the predicted value. In this ridge regression method, the coefficients are penalised by a value lambda, this acts as a control parameter, which determines how severe is the penalty and how much significance should be given to Xi. The higher the values of the lambda the bigger is the penalty and therefore the magnitude of coefficients is reduced.

When the slope (b) of the line is steep then the target variable(Y) is very sensitive to relatively small change in the predictor variable variable(X). In ridge regression by the addition of the lambda value the sensitivity decreases. If lambda is zero then ridge regression reduces to linear regression and when lambda increases gradually the slope the line decreases asymptotically. To know which value of lambda is to choose we try different values of lambda and use cross validation to determine which one result in the lowest variance.

5) Lasso Regression:

Least Absolute Shrinkage and Selection Operator (LASSO) is similar to ridge regression that penalizes the regression model. It performs L1 regularization. The Lasso regression process is usually used in machine learning for the selection of the significant subset of variables. The prediction accuracy of this model is higher when compared to other model interpretations. Like ridge regression lasso regression also results in a line with little amount of bias added to it which thereby decreases the variance of the model.

The major difference between lasso and ridge regression is, ridge regression decreases the slope asymptotically close to zero but lasso regression can reduce the slope all the way down to zero, thereby eliminating the useless parameters from the line equation that don't have any significant role for predicting the value of the target variable.

Lasso regression usually works better under conditions where some predictors have high coefficients, and the rest have low coefficients. Ridge regression performs better when the result is a function of many predictors, all of which have coefficients of approximately the same size.

B. Description of the Dataset:

From the Kaggle site obtained our dataset for developing the ML Health Insurance Prediction System (MLHIPS). The data set obtained contains seven attributes or features and 1338 rows; out of the seven attributes or features three of them contains categorical values and the rest contain numerical values. The data set is then divided into two halves. The first component is referred to as training data, while the second is referred to as test data. The more data that is supplied to the model during its training period, the more accurate the model will be when making predictions on unseen data. Data is typically split at a ratio of 80:20 for testing and training purposes. Training datasets are used to build models as predictors of health insurance costs, and test sets are used to evaluate regression models displays the dataset's description.

The dataset contained missing values in certain fields. After reviewing the distributions, it was decided to replace the missing variables with new attributes, implying that the data is missing. This is only possible if the data is lost completely at random.Multilevel structure and hidden dependencies are present in medical data. It is vital to figure out these hidden patterns and use various fundamental analysis techniques present in a combined fashion. This is the reason why in the field of medical data analysis, many researchers use different ensemble Machine Learning models.Many of them have used different ensemble learning techniques such as Random Forests, Adaboost, GBM, and XGBM. to forecast the modulus of elasticity of the collective recycled concrete, an ensemble of Random Forests (RF) and Support Vector Machines (SVM) is utilised. This classical ensemble has a much higher level of precision. To take advantage of the heterogeneity of distinct sets of meta-features,

an ensemble of K-Nearest Neighbor (KNN) classifiers was created for recommendation purposes is analysed.

For the diabetic retinopathy dataset, researchers used an ensemble-based machine learning model that included ID3, Random Forests, Adaboost, Logistic Regression and KNN

C. Data Pre-processing:

The dataset contains seven variables, as shown in the table above. While calculating the cost of the Charges of a customer which is our target variable the values of the rest six of the variables are taken into consideration. In this phase, the data is reviewed, properly reconstructed, and properly applied to machine learning algorithms. The dataset was first checked for missing values. The dataset was found containing missing values in the bmi and charges columns. The missing values were imputed by the mean values of the respective attribute values.

As regression models accept only numerical data, the categorical columns in our case the sex, smoker and region columns containing categorical columns were converted into numerical values using label encoding. Then the updated dataset was partitioned into training and testing dataset. And the model was trained using the training dataset.

The data set is then divided into two halves. The first component is referred to as training data, while the second is referred to as test data. The more data that is supplied to the model during its training period, the more accurate the model will be when making predictions on unseen data. Data is typically split at a ratio of 80:20 for testing and training purposes. Training datasets are used to build models as predictors of health insurance costs, and test sets are used to evaluate regression models. Displays the dataset's description. The dataset contained missing values in certain fields. After reviewing the distributions, it was decided to replace the missing variables with new attributes, implying that the data is missing. This is only possible if the data is lost completely at random; thus, the missing data mechanism, which determines the best method to data processing, must first be developed. Multilevel structure and hidden dependencies are present in medical data.

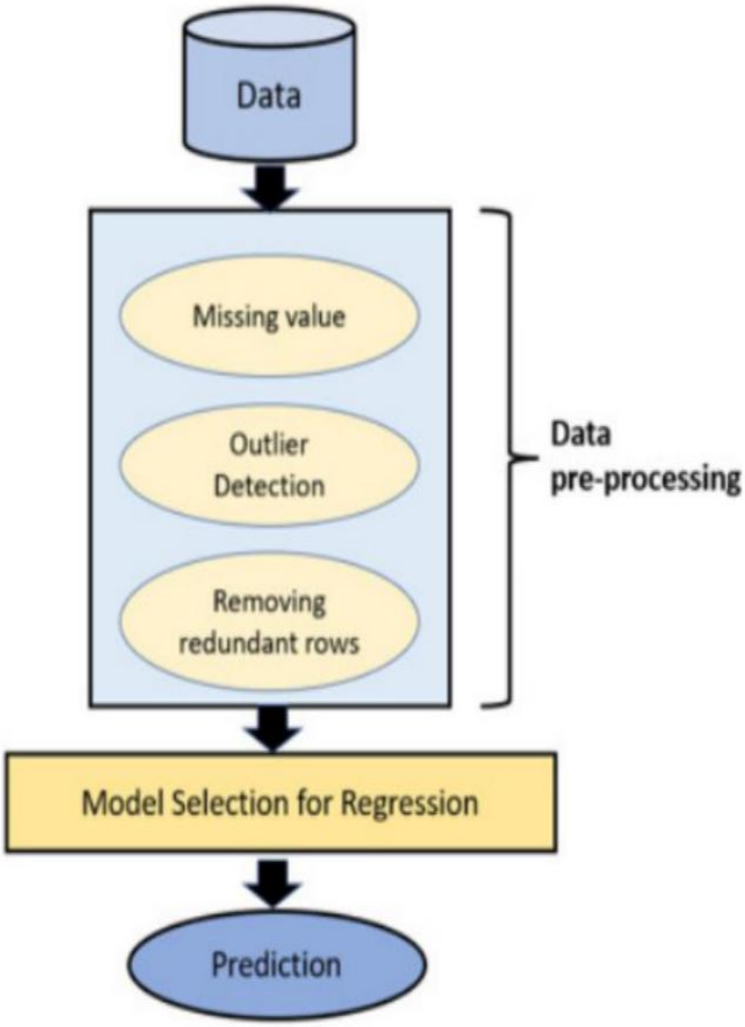

Fig - 2.3.3 – Block Diagram

CHAPTER 3
PROPOSED SYSTEM

3.1. Objective of Proposed Model

Our Project Medical Insurance Premium Prediction estimate the cost of an insurance premium based on certain features like age, BMI, number of dependents, smoking habits, etc. Along with existing algorithms we are using xgboost-extreme-gradient-boosting algorithm for more accuracy. We also aim to test experiments and compare results using different learning models (e.g., gradient boosted trees and linear regression, Support Vector Machine) on the same data. The goal is to provide insurance companies and policyholders with a reliable tool for pricing and planning.

3.2. Algorithms of Proposed Model

A. Linear Regression

Linear regression is a machine learning algorithm rooted in concept of "supervised learning." It is employed to forecast the value of a dependent variable (y) depending on the value of an independent variable (x). Essentially, this implies that linear regression is utilized to ascertain how closely a dependent variable is related to an independent variable, and then make prediction according to that correlation. This is an invaluable asset for data analysis, as it allows analysts to understand complex trends and correlations in data, and to enhance the precision of predictions regarding future outcomes.

Linear Regression Formula:

Y=mx+b

Here

m=slope
x=Value of first dataset
Y=Value of second dataset
The values of x and y in this scenario represents the training datasets used for illustrating rating the Linear Regression models.

B. **Support Vector Machine Algorithm:**

SVM, is a commonly employed supervised learning technique for solving classifying and predicting tasks, with a focus regarding classification in machine learning. The objective of Support Vector Machines (SVM) is to identify the optimal line or boundary of decision that can effectively divide a multi-dimensional space into distinct classes. This facilitates the accurate classification of new data points in the future. This optimal decision boundary is known as a hyperplane, which is formed utilizing extreme vectors and points called support vectors. SVM is favored for its capability to effectively classify data and manage high dimensional spaces. This is preliminary estimate. This optimal the decision boundary is known as a hyperplane.

SVM Formula

$$\boldsymbol{y = sign(\sum_{i=1}^{n}(w_i . x_i) + b)}$$

Where
wi is a weights
xi is a input data.
b is the bias term.

C. Random Forest Regression:

The Random Forest methodology utilizing bootstrapping involves the use of multiple of decision tress generated from the data and amalgamated via ensemble learning technique This approach often leads to accurate predictions and classifications taking the average of the outcomes from the randomly selected tress.

Random Forest Formula

$$y = 1/N \sum_{i}^{N} tree_i(x)$$

N represents the quantity of trees in the random forest

D. Decision Tree Regression

Decision trees are a category of supervised ML models employed by the Train Using Auto ML tool. They classify or regress data based on true or false answers to specific questions. When visualized, the resulting structure takes the form of a tree, featuring various types of nodes, including the root, internal nodes, and leaf nodes. The root node serves as the initial point in the decision tree, from which branches extend to internal nodes and eventually lead to leaf nodes. The leaf nodes represent the ultimate classification categories or real values within the decision tree. Notably, decision trees are renowned for their simplicity and interpretability, making them easy to understand and explain.

To build a decision tree, begin by designating a feature at the root node. Usually, no single feature can perfectly predict the final classes, leading to what is termed impurity. Methods such as Gini, entropy, and information gain are employed to quantify this impurity and determine how effectively a feature classifies the provided data. The feature with the least impurity is chosen as the node at any given level.

For calculating Gini impurity with numerical entries within a feature, start by sorting the information within ascending order and determining the averages of adjacent values. Next, determine the Gini impurity of each selected average value by arranging the data points according to whether the feature values are smaller than or greater than the chosen threshold. Evaluate the effectiveness of this selection by assessing how accurately it classifies the data according to the specified conditions. The Gini impurity is then computed using the equation below, where K represents the quantity of classification categories and p denotes the proportion of instances of those categories.

Determine the weighted average of Gini impurities for the leaves associated with each selected value. Select the value with the minimum impurity for that feature. Iterate this procedure for various features to choose the feature and value that will serve as the node. Iterate this procedure at every node and depth level until all data is classified.

Once the tree is constructed, to predict for a data point, traverse the tree using the conditions at each node to reach the final value or classification. For decision trees used in regression, the measure of impurity shifts from Gini impurity to using the sum of squared residuals or variance.

D. Gradient Boosting Algorithm

Gradient boosting is a highly popular machine learning method for analyzing tabular data sets. It is well-known for its ability to handle missing values, outliers, and large categorical values in the features, in addition to its ability to detect nonlinear relationships between the target and the features. This attribute renders it a potent tool for conducting data analysis and making predictions.

Gradient Boosting is a potent boosting algorithm that amalgamates multiple weak learners into a robust learner. In this method, each new model is trained to reduce the loss function, such as average squared deviation or cross-entropy, of the preceding model using gradient descent. This iterative process enhances the overall predictive capability of the model. During each iteration, the algorithm determines the gradient of the loss function concerning the predictions generated by the current ensemble. Subsequently, a new weak model undergoes training to minimize this gradient, refining the overall predictive accuracy of the ensemble. The forecasts of the new model are integrated into the ensemble, and this procedure is iteratively repeated until a predetermined stopping criterion is fulfilled.

Unlike AdaBoost, Gradient Boosting do not adjust the weights of training instances. Instead, each predictor is trained using the residual errors of the preceding model as labels. A specific technique within Gradient Boosting known as Gradient Boosted Trees utilizes CART (Classification and Regression Trees) as its base learner.

Gradient Boosting Algorithm Formula

$$y = \sum_{i=1}^{N} tree_i$$

N represents the number of trees in the gradient boosting ensemble.

D. **Xg Boost extreme gradient boosting algorithm**

XGBoost, or Extreme Gradient Boosting, is a state-of-the-art machine learning algorithm well known for its exceptional predictive performance. It's the gold standard in ensemble learning, especially when it comes to gradient-boosting algorithms. It develops a series of weak learners one after the other to produce a reliable and accurate predictive model. Fundamentally, XGBoost builds a strong predictive model by combining predictions of several weak learners, usually decision trees. It uses a boosting technique to construct an extremely accurate ensemble model by having each weak learner after it correct the mistakes of its predecessors. The optimization method (gradient) minimizes the cost function by repeatedly changing the model's parameters in response to the gradients of the errors. The algorithm also presents the idea of "gradient boosting with decision trees," in

$$y = \sum^{N}{}_{i=1} tree_i(x)$$

N represents the number of trees in the xg-boost ensemble

3.3 Designing

3.3.1 Flow Chart

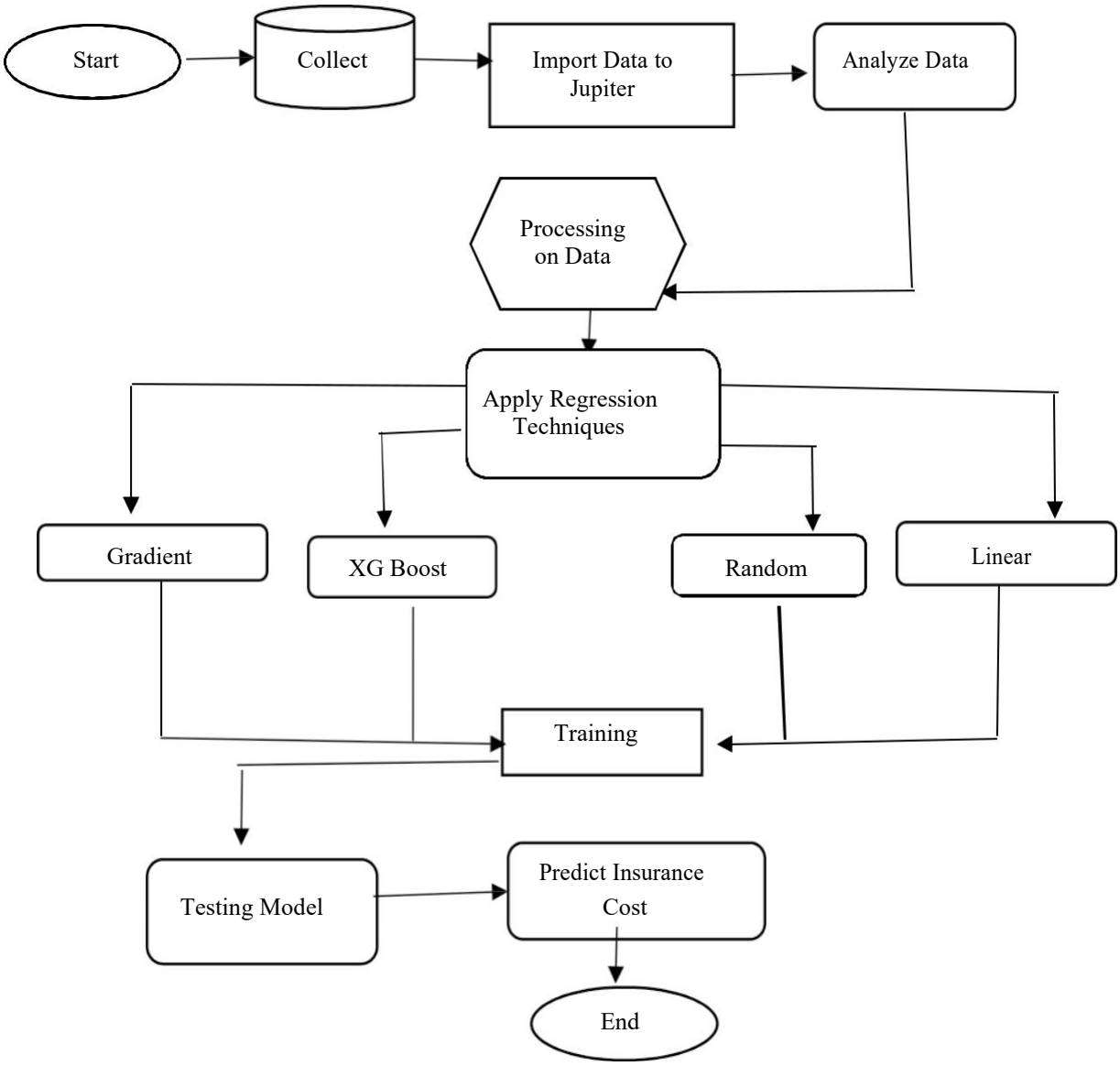

Fig 3.3.1 - FlowChart

DATASET

We had used a dataset from Google for creating our prediction model. This data set includes six attributes and the data set has splitted into two-parts: training data and testing data. For training the model, 80% of total data is used and the rest for testing. To build a predictor model of medical insurance premium prediction the training dataset is applied and to evaluate the regression model, test set is used. The following table shows the Description of the Dataset.

Table 3.3.1 Dataset overview

TABLE	DESCRIPTION
NAME	NAME OF PERSON
BMI	BODY MASS INDEX
AGE	AGE OF PERSON
GENDER	MALE/FEMALE
SMOKER	WHETHER A PERSON IS A SMOKER OR NOT
ALCOHOLIC	WHETHER A PERSON IS DRINK ALCOHOL OR NOT
NO.OF KIDS	NUMBER OF CHILDREN THE CLIENT HAVE
REGION	WHETHER THE PERSON LIVES IN SOUTHWEST, NORTHEAST, SOUTHEAST OR NORTHEAST
CHARGES (TARGET VARIABLE)	MEDICAL COST THE PERSON NEEDS TO PAY

DATA PREPROCESSING

The dataset includes six variables, as shown in table . From these variables each one of these attributes has some contribution to estimate the cost of the insurance, which is our dependent variable. In this stage, the data is scrutinized and updated properly to efficiently apply the data to the Machine Learning algorithms.

Now the categorical variables are converted into numeric or binary values to represent either 0 or 1. For example, instead of "SMOKER" with yes or no, the "SMOKER" with no variable would be considered as false (0) . And "SMOKER" with yes variable would be considered as true.

Table **3.3.2 Data set**

TABLE	**DESCRIPTION**
NAME	NAME OF PERSON
BMI	BODY MASS INDEX
AGE	AGE OF PERSON
GENDER	MALE/FEMALE
SMOKER	WHETHER A PERSON IS A SMOKER OR NOT 0 = no 1 = yes
ALCOHOLIC	WHETHER A PERSON IS DRINK ALCOHOL OR NOT
NO.OF KIDS	NUMBER OF CHILDREN THE CLIENT HAVE
REGION	WHETHER THE PERSON LIVES IN SOUTHWEST, NORTHEAST, SOUTHEAST OR NORTHEAST
CHARGES (TARGET VARIABLE)	MEDICAL COST THE PERSON NEEDS TO PAY

Linear Regression Model

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error,
r2_score import numpy as np
# Fit a linear regression model on the training set
model = LinearRegression().fit(X_train, y_train)
predictions = model.predict(X_test)
np.set_printoptions(suppress=True)
print('Predicted labels: ',
np.round(predictions)[:10]) print('Actual labels :
' ,y_test[:10])
```

```
Predicted labels:  [31412. 13350.  9590. 36380.  6511. 30265.  1693. 26250. 10695. 12194.
Actual labels   :  1304    21259.38
446    12731.00
459     7682.67
252    44260.75
979     4889.04
153    19964.75
529     1708.00
994    16420.49
242    35160.13
382    20781.49
Name: expenses, dtype: float64
```

Fig – 3.3.2 - Linear regression Prediction values

Visualizing Regression Line

```
plt.xlabel('Actual Labels')
plt.ylabel('Predicted Labels')
plt.title('Daily Bike Share Predictions')
# overlay the regression line
z = np.polyfit(y_test, predictions, 1)
p = np.poly1d(z)
plt.plot(y_test,p(y_test), color='magenta')
plt.show()
mse = mean_squared_error(y_test, predictions)
print("MSE:", mse)
rmse = np.sqrt(mse)
print("RMSE:", rmse)
r2 = r2_score(y_test, predictions)
print("R2:", r
```

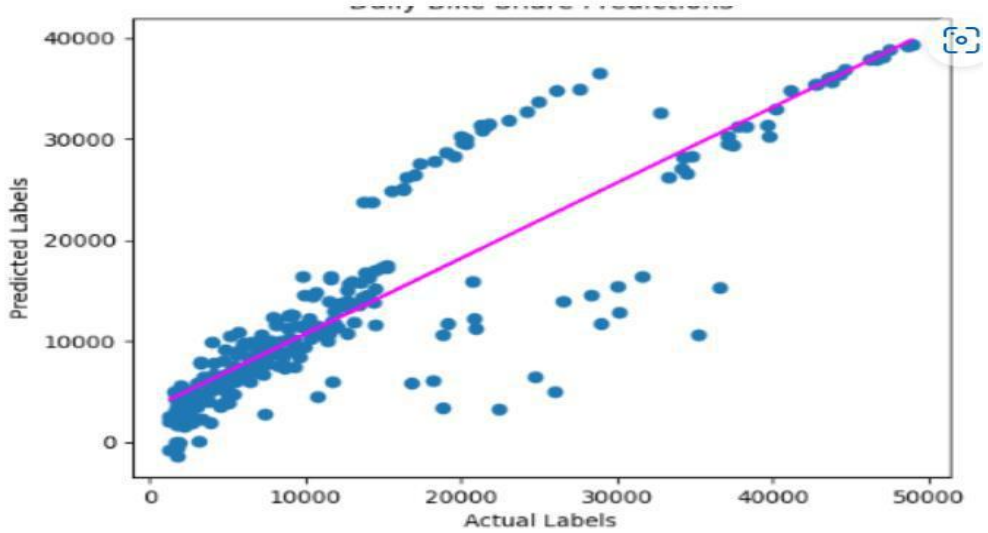

Fig – 3.3.3 – Visualizing regression Prediction values

The linear regression model is pretty average having an R2 score of 0.75. Next, we will try some advanced regression models to see if we can improve the performance of the model.

Random Forest Regression

```
from sklearn.ensemble import
RandomForestRegressor # Train the model
model_rf = RandomForestRegressor().fit(X_train, y_train)
print (model_rf, "\n")
# Evaluate the model using the
test data predictions =
model_rf.predict(X_test)
mse = mean_squared_error(y_test,
predictions) print("MSE:", mse)
rmse =
np.sqrt(mse)
print("RMSE:
", rmse)
r2 = r2_score(y_test,
predictions) print("R2:",
r2)
# Plot predicted vs actual
plt.scatter(y_test, predictions)
plt.xlabel('Actual Labels')
plt.ylabel('Predicted Labels')
plt.title('Daily Bike Share Predictions')
# overlay the regression line
z = np.polyfit(y_test, predictions, 1)
p = np.poly1d(z)
plt.plot(y_test,p(y_test), color='magenta')
plt.show()
```

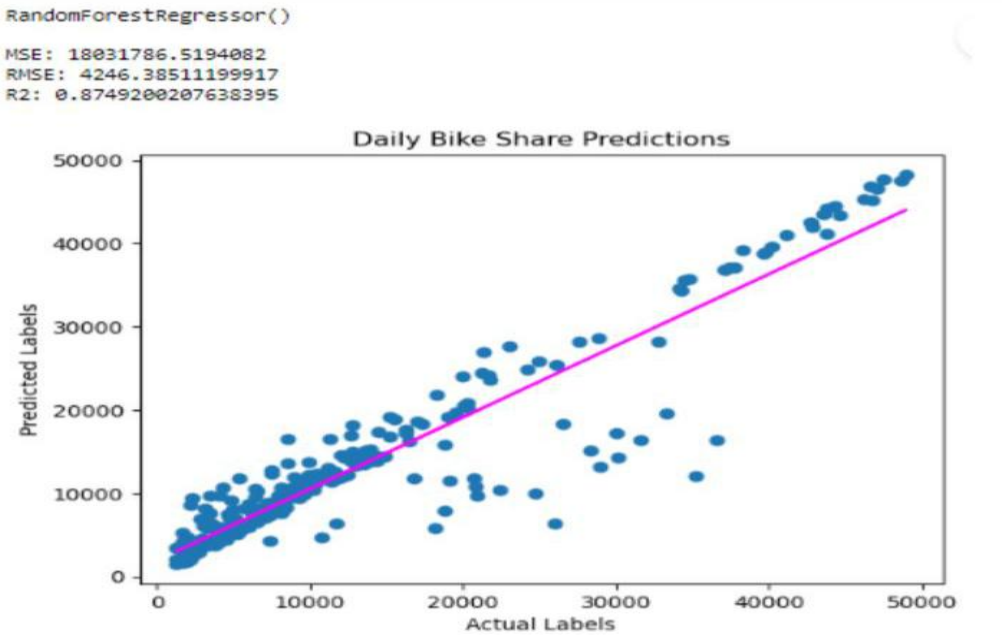

Fig - 3.3.4 – Random Forest regression Prediction values

random forest regressor has shown significant performance by attaining an R2 score of **0.87** whereas RMSE of 4246.38 and MSE of 18021786 are also lesser than linear regression RMSE and MSE respectively.

3.3 Stepwise Implementation and Code

```
import pandas as pd
import seaborn as sns
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib as pt
import warnings
warnings.filterwarnings("ignore")
df=pd.read_csv("insurance.csv")
df
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080
1335	18	female	36.850	0	no	southeast	1629.83350
1336	21	female	25.800	0	no	southwest	2007.94500
1337	61	female	29.070	0	yes	northwest	29141.36030

1338 rows × 7 columns

```
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)
```

```
df.describe()
```

age	bmichildren		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

```
df.isnull().sum()
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
df['smoker'].value_counts()
smoker
no      1064
yes      274
Name: count, dtype: int64
df['children'].value_counts()
children
0    574
1    324
2    240
3    157
4     25
```

```
df['region'].value_counts()
region
southeast    364
southwest    325
northwest    324
northeast    32
<Axes: ylabel='age'>
```

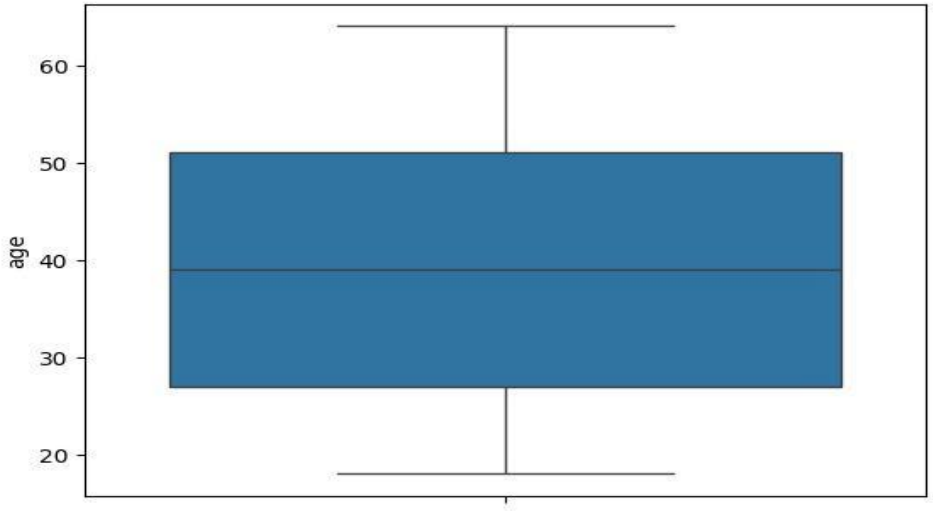

Fig -3.4.1- Boxplot of Medical charges per Age

<Axes: ylabel='bmi'>

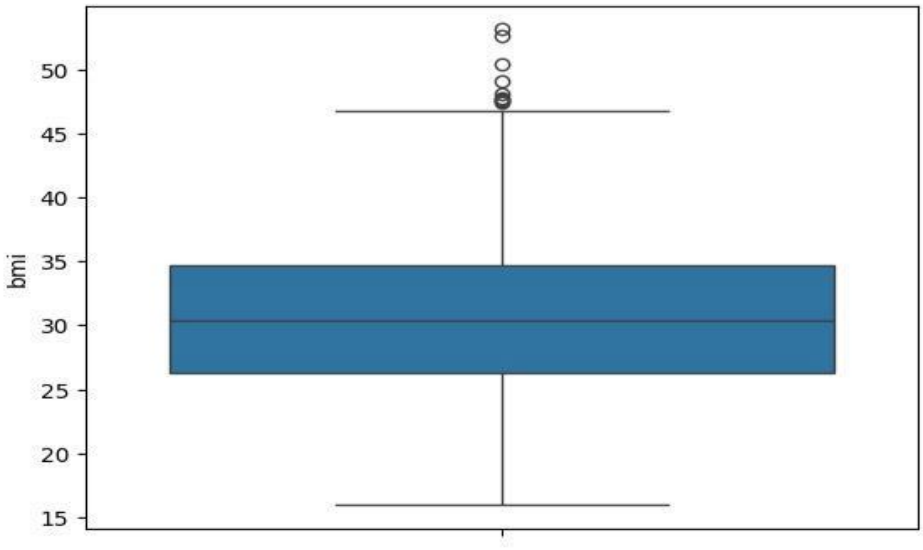

Fig – 3.4.2- Boxplot of Medical charges per BMI

```
Q1=df['bmi'].quantile(0.25)
Q2=df['bmi'].quantile(0.5)
Q3=df['bmi'].quantile(0.75)
iqr=Q3-Q1
lowlim=Q1-1.5*iqr
upplim=Q3+1.5*iqr
print(lowlim)
print(upplim)
from feature_engine.outliers import ArbitraryOutlierCapper
arb=ArbitraryOutlierCapper(min_capping_dict={'bmi':13.6749},max_capping_dict={'bmi':47.315})
df[['bmi']]=arb.fit_transform(df[['bmi']])
from feature_engine.outliers import ArbitraryOutlierCapper
arb=ArbitraryOutlierCapper(min_capping_dict={'bmi':13.6749},max_capping_dict={'bmi':47.315})
df[['bmi']]=arb.fit_transform(df[['bmi']])
sns.boxplot(df['bmi'])
<Axes: ylabel='bmi'>
```

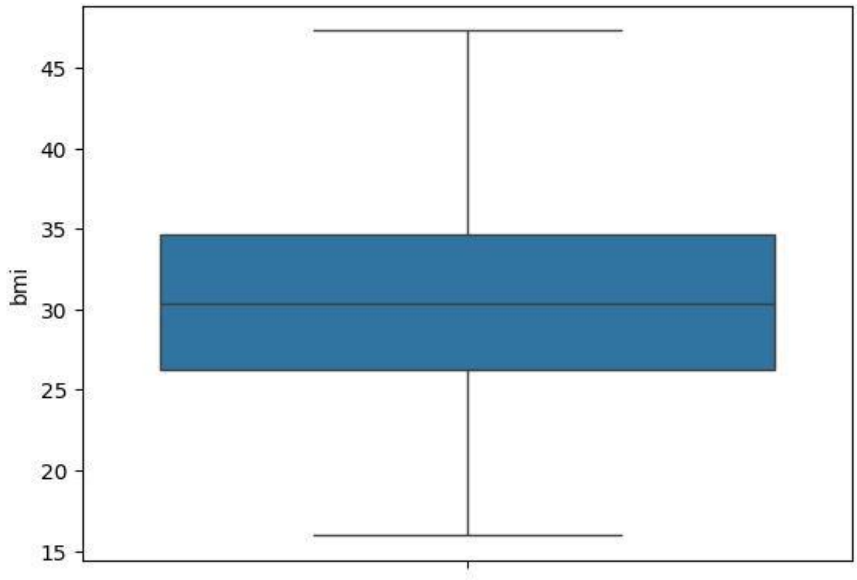

Fig – 3.4.3 – Boxplot of Medical Charges per BMI

```
df['bmi'].skew()
df['age'].skew()
```

```
df['age'].skew()
0.054780773126998195
df['sex']=df['sex'].map({'male':0,'female':1})
df['smoker']=df['smoker'].map({'yes':1,'no':0})
df['region']=df['region'].map({'northwest':0, 'northeast':1,'southeast':2,'southwest':3})
df
```

	age	sex	bmi	children	smoker	region	charges
0	19	1	27.900	0	1	3	16884.92400
1	18	0	33.770	1	0	2	1725.55230
2	28	0	33.000	3	0	2	4449.46200
3	33	0	22.705	0	0	0	21984.47061
4	32	0	28.880	0	0	0	3866.85520

```
sns.heatmap(df.corr(),annot=True)
<Axes: >
```

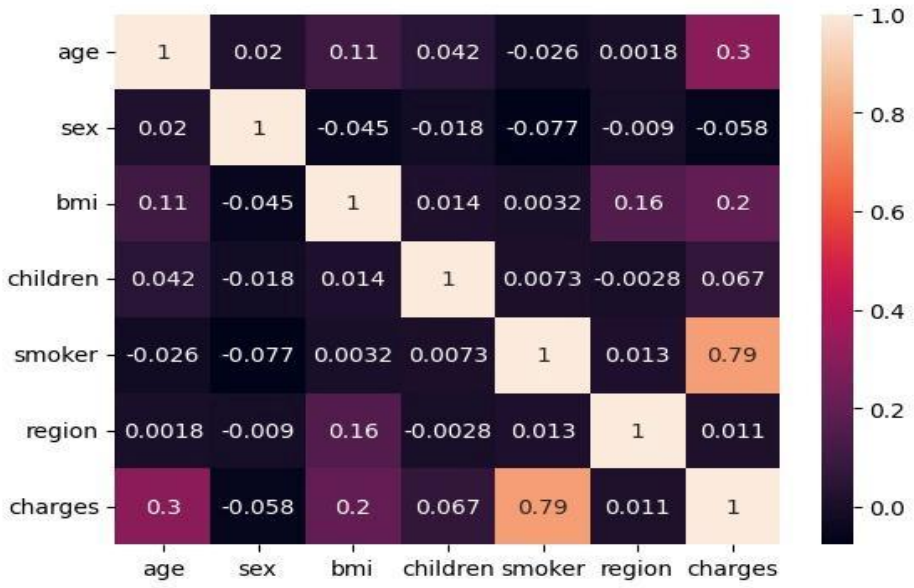

Fig – 3.4.3 - Values of various parameters

```
from sklearn.linear_model import
LinearRegression,Lasso from sklearn.svm import SVR
from sklearn.ensemble import RandomForestRegressor
```

```
from sklearn.ensemble import
GradientBoostingRegressor from xgboost import
XGBRegressor
from sklearn.linear_model import
LinearRegression,Lasso from sklearn.svm import SVR
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
from xgboost import XGBRegressor
gbmodel=GradientBoostingRegressor()
gbmodel.fit(xtrain,ytrain)
ypredtrain3=gbmodel.predict(xtrain)
ypredtest3=gbmodel.predict(xtest)
print(r2_score(ytrain,ypredtrain3))
print(r2_score(ytest,ypredtest3))
print(cross_val_score(gbmodel,X,Y,cv=5,).mean())
xgmodel=XGBRegressor()
xgmodel.fit(xtrain,ytrain)
rfmodel.fit(xtrain,ytrain)
ypredtrain2=rfmodel.predict(xtrain)
ypredtest2=rfmodel.predict(xtest)
print(r2_score(ytrain,ypredtrain2))
print(r2_score(ytest,ypredtest2))
print(cross_val_score(rfmodel,X,Y,cv=5,).mean())
from sklearn.model_selection import GridSearchCV
estimator=GradientBoostingRegressor()
param_grid={'n_estimators':[10,15,19,20,21,50],'learning_rate':[0.1,0.19,0.2,0.21,0.8,1]}
grid=GridSearchCV(estimator,param_grid,scoring="r2",cv=5)
grid.fit(xtrain,ytrain)
gbmodel=GradientBoostingRegressor(n_estimators=19,learning_rate=0.2)
gbmodel.fit(xtrain,ytrain)
ypredtrain3=gbmodel.predict(xtrain)
ypredtest3=gbmodel.predict(xtest)
```

```
print(r2_score(ytrain,ypredtrain3))
print(r2_score(ytest,ypredtest3))
print(cross_val_score(gbmodel,X,Y,cv=5,).mean())

important_features=feats[feats['Importance']>0.01]
important_features
X=df.drop(df[['charges']],axis=1)
xtrain,xtest,ytrain,ytest=train_test_split(Xf,Y,test_size=0.2,random_state=42)
finalmodel=XGBRegressor(n_estimators=15,max_depth=3,gamma=0)
finalmodel.fit(xtrain,ytrain)
ypredtrain4=finalmodel.predict(xtrain)
ypredtest4=finalmodel.predict(xtest)
print(r2_score(ytrain,ypredtrain4))
print(r2_score(ytest,ypredtest4))
print(cross_val_score(finalmodel,X,Y,cv=5,).mean())
new_data=pd.DataFrame({'age':19,'sex':'male','bmi':27.9,'children':0,'smoker':'yes','region':'northeast'},
inde
x=[1])
finalmodel.predict(new_data)
array([18035.828], dtype=float32)
from pickle import dump
dump(finalmodel,open('insurancemodelf.pkl','wb'))
```

GUI.ipynb code

```
from tkinter import *;
from pickle import load
import pandas as pd
def show_entry():
    p1=int(e1.get())
    p2=float(e3.get())
    p3=int(e4.get())
```

```
 p4=int(e5.get())
    model=load(open('insurancemodelf.pkl','rb'))
    new_data=pd.DataFrame({'age':p1,'bmi':p2,'children':p3,'smoker':p4},index=[0])
    cost=model.predict(new_data)
 Label(master,text="insurance cost").grid(row=7)
    Label(master,text=cost).grid(row=8
master=Tk()
master.title("Insurance cost prediction")
label=Label(master,text='Insurance cost
prediction',bg="black",fg="white").grid(row=0,columnspan=2)
Label(master,text = "enter your age").grid(row=1)
Label(master,text="enter male or female").grid(row=2)
Label(master,text="enter your bmi value").grid(row=3)
Label(master,text="enter number of children").grid(row=4)
Label(master,text="smoker yes/no [1/0]").grid(row=5)
Label(master,text="enter your region").grid(row=6)
e1=Entry(master)
e2=Entry(master)
e3=Entry(master)
e4=Entry(master)
e5=Entry(master)
e6=Entry(master)
e1.grid(row=1,column=1)
e2.grid(row=2,column=1)
e3.grid(row=3,column=1)
e4.grid(row=4,column=1)
e5.grid(row=5,column=1)
e6.grid(row=6,column=1)
Button(master,text='predict insurance cost',command=show_entry).grid()
mainloop()
```

CHAPTER 4
RESULTS AND DISCUSSION

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Performance metrics

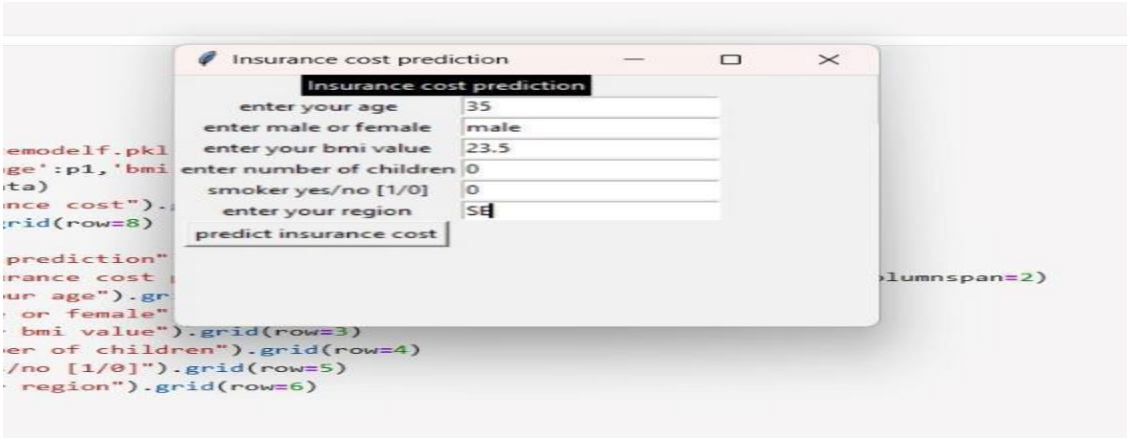

Fig 4.1.1 –Parameters for calculating the insurance cost

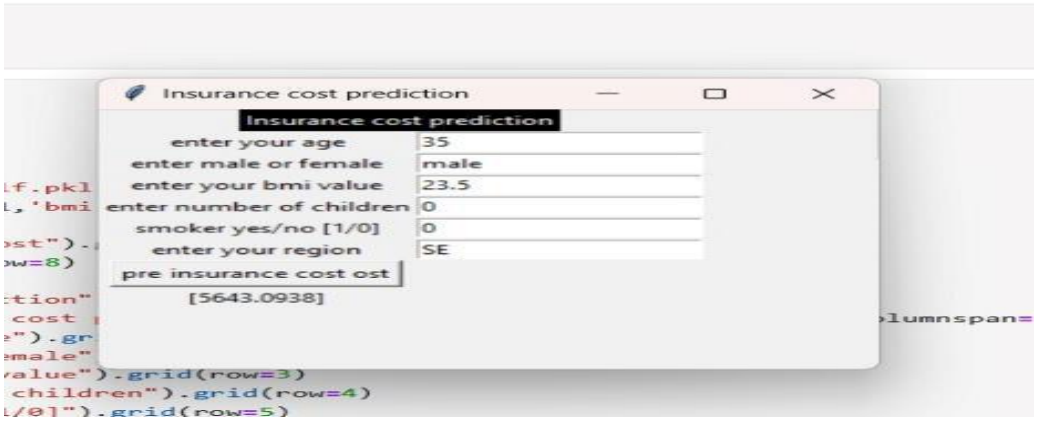

Fig - 4.1.2 – Prediction of Insurance cost

Results:

The following results can be seen in Prediction:

1)Linear Regression Algorithm:

The accuracy of the Linear Regression

Algorithm is 78.334%

2)Support Vector Machine:

The accuracy of the Support Vector Machine Algorithm is

7.229%

3)Random Forest Regression:

The accuracy of Random Forest Regression is 87.006%

It was found that the amount of data used for training had a significant impact on the accuracy, with a larger train size leading to better results

4)Gradient Boosting Algorithm :

The Accuracy of the Gradient Boosting Algorithm is 87.776%.

Table – 4.1 – Accuracy values of Regression Algorithms

ALGORITHMS	**ACCURACY %**
LINEAR REGRESSION ALGORITHM	78.334%
SVM	70.229%
RANDOM FOREST	79.004%
GRADIENT BOOSTING	81.006%
XG-BOOST EXTREME GRADIENT BOOSTING ALGORITHM	95%

CHAPTER 5
CONCLUSION & FUTURE ENHANCEMENT

CHAPTER 4
CONCLUSION

- Our project has practical implications for healthcare providers, insurers, and individuals seeking insurance coverage.
- It can guide fairer pricing strategies and informed decision-making in the healthcare finance sector.
- Continued data collection and model refinement could enhance prediction accuracy.
- Exploring the impact of policy changes and new data sources can further enrich our insights.

FUTURE ENHANCEMENT

- The use of the Random Forest algorithm allows for the introduction of unpredictability in the feature selection process, which can improve prediction accuracy.
- In order to assess the scalability of the system, it would be beneficial to test it on a dataset with at least a million records in the future.
- Distributed frameworks like Spark and Hadoop can be utilized to handle large amounts of data and enhance the scalability of the system.
- Currently, the algorithm is being trained and tested using thousands of records

REFERENCES

REFERENCES

[1] Gupta, S., & Tripathi, P. (2016, February). A leading trend of data analytics with health insurance in India. In 2016 International Conference on Innovation and Challenges in CS (ICICCS-INBUSH) (pp. 64-69). IEEE.

[2] [Yerpude, P., Gudur, V.: Prediction modeling of crime dataset using data mining. Int. J. Data Min. Knowl. Manage. Process (IJDKP) 7(4) (2017)

[3] Grosan, C., Abraham, A.: Intelligent Systems: A Modern Approach, Intelligent Systems Reference Library Series. Springer, Cham (2011)

[4]https://www.ijraset.com/research-paper/medical-insurance-cost-prediction-using-machine-learning

[5] A. Tike and S. Tavarageri. (2017). A Medical Price Prediction System using Hierarchical Decision Trees. In: IEEE Big Data Conference 2017. IEEE.

[6] "National Health Expenditures 2015 Highlights," CMS.gov, 2015. [Online]. Available:https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trendsand-Reports/NationalHealthExpendData/downloads/highlights.pdf

[7] J. Cubanski, C. Swoop and T. Neuman,"How Much Is Enough? Out-of-Pocket Spending Among Medicare Beneficiaries: A Chartbook," The Henry J. Kaiser Family Foundation, Menlo Park, CA, 2014. [Online]. Available: http://files.kff.org/attachment/ how-much-isenough-out-of-pocket-spending-among-medicare-beneficiaries-a-chartbook-

[8] J. Cubanski and T. Neuman,"The Facts on Medicare Spending and Financing," The Henry J. Kaiser Family Foundation, Menlo Park, CA, 2017. [Online]. Available: http://files.kff.org/attachment/Issue-Brief-The-Facts-on-Medicare-Spending-and-Financing.

[9] The Henry J. Kaiser Family Foundation. (2017). Total Number of Medicare Beneficiaries. [Online].Available: https://www.kff.org/ medicare/state-indicator/total-medicarebeneficiaries/. [Accessed Nov. 2, 2017].

[10] Scikit-learn.org. (2018). About us — scikit-learn 0.19.1 documentation. [online] Available at: http://scikit-learn.org/stable/about.html#people [Accessed 23 Apr. 2018].

[11] R. Hafezi, J. Shahrabi, and E. Hadavandi, "A bat-neural network multi- agent system (bnnmas) for stock price prediction: Case study of dax stock price," Applied Soft Computing, vol. 29, pp. 196–210, 2015.

[12] H. Lee, M. Surdeanu, B. MacCartney, and D. Jurafsky, "On the importance of text analysis for stock price prediction." in LREC, 2014, pp. 1170–1175.

[13] P.-C. Chang and C.-H. Liu, "A tsk type fuzzy rule based system for stock price prediction," Expert Systems with applications, vol. 34, no. 1,pp. 135–144, 2008.

[14] K. Kohara, T. Ishikawa, Y. Fukuhara, and Y. Nakamura, "Stock price prediction using prior knowledge and neural networks," Intelligent systems in accounting, finance and management, vol. 6, no. 1, pp. 11–22, 1997.

[15] K.-j. Kim and I. Han, "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index," Expert systems with Applications, vol. 19, no. 2, pp.
125–132, 2000

Githublink

https://github.com/yasmeen-10/Major_Project_Batch-14