

A FIELD PROJECT REPORT

on

“Medical Health insurance prediction”

Submitted by:

221FA04464-Sampath

221FA04476-Rohitha

221FA04484-Chiranjeevi

221FA04721-Manjunadh

Under the guidance of

Dr. S. Deva Kumar

Designation



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH Deemed

to be UNIVERSITY

Vadlamudi, Guntur.

ANDHRA PRADESH, INDIA, PIN-522213.

CERTIFICATE

This is to certify that the Field Project entitled on “**Medical Health Price Prediction**” that is being submitted by 221FA04464 (Sampath) , 221FA04476 (Rohitha), 221FA04484(Chiranjeevi), 221FA04721 (Manjunadh) for partial fulfilment of Field Project is a bonafide work carried out under the supervision of Dr. Deva Kumar, Assistant Professor, Department of CSE

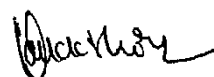
Guide name& Signature



Dr. S. V. Phani Kumar

Assistant/Associate/Professor,
CSE

HOD,CSE



Dr.K.V. Krishna Kishore

Dean, SoCI



DECLARATION

We hereby declare that the Field Project entitled on “**Medical Health Price Prediction**” is being submitted by 221FA04464(Sampath), 221FA04476(Rohitha), 221FA04484(Chiranjeevi), 221FA04721 (Manjunadh) in partial fulfilment of Field Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of Dr. Deva Kumar, Assistant Professor, Department of CSE.

By
221FA04464 (Sampath),
221FA04476(Rohitha),
221FA04484(Chiranjeevi),
221FA04721(Manjunadh)

Date: 15/10/2024

ABSTRACT

Healthcare costs have been a significant concern globally, with escalating expenses impacting both individuals and governments. Accurate prediction of medical healthcare costs can enable better resource allocation, risk management, and personalized treatment plans. This research aims to develop machine learning models to predict medical healthcare costs effectively.

Various machine learning algorithms, including linear regression, decision trees, random forests, and gradient boosting, were explored and evaluated using a comprehensive dataset containing patient demographics, medical history, and cost information. Feature engineering techniques were employed to enhance the predictive power of the models by extracting relevant information from the raw data.

The performance of the models was assessed using metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared. The results demonstrated that gradient boosting algorithms consistently outperformed other models, achieving the lowest error rates and highest accuracy in predicting medical healthcare costs.

The findings of this research contribute to the advancement of healthcare cost prediction and offer valuable insights for policymakers, healthcare providers, and insurance companies. By accurately predicting future healthcare expenses, stakeholders can make informed decisions regarding resource allocation, pricing strategies, and risk management.

TABLE OF CONTENTS

1. Introduction	1
1.1 Background and Importance of Medical Healthcare Cost Prediction	2
1.2 Overview of Machine Learning in Healthcare	3
1.3 Research Objectives and Scope	4
1.4 Challenges in Medical Healthcare Cost Prediction	5
1.5 Applications of ML to combat congestion	6
2. Literature Survey	11
2.1 Previous Studies on Medical Healthcare Cost Prediction	13
2.2 Integration of Clinical and Genetic Data	17
2.3 Limitations and Future Directions	18
3. Proposed System	19
3.1 Input dataset	20
3.1.1 Detailed features of dataset	2
3.2 Data Pre-processing	21
3.3 Model Building	22
3.4 Methodology of the system	24
3.5 Evaluation Metrics	25
3.6 Constraints	33
3.7 Ethical Considerations	48
4. Implementation	51
4.1 Environment Setup	52
4.2 Sample code for preprocessing and Model Training and Testing	52
5. Experimentation and Result Analysis	54
6. Conclusion	56
7. References	58

LIST OF FIGURES

Figure 1. Architecture of the proposed system	24
Figure 2. Various features in the dataset after Pre-Processing	24
Figure 3. Training Vs Testing Accuracy	26
Figure 4. Confusion Matrix	26
Figure 5. Performance Outcomes	27
Figure 6. ROC Curve for Each Class	28
Figure 7. Logistic Regression-Confusion Matrix	28
Figure 8. Naïve Bayes-Confusion Matrix	29
Figure 9. Support Vector Machine (SVM) -Confusion Matrix	29
Figure 10. Random Forest-Confusion Matrix	30
Figure 11. XGBoost -Confusion Matrix	30
Figure 12. Decision Tree Visualization	31
Figure 13. Confusion Matrix for MLP Model	55

LIST OF TABLES

Table 1. Table 1. Recorded Results for each Classifier	30
--	----

CHAPTER-1

INTRODUCTION

1. INTRODUCTION

1.1. Background and Importance of Medical Healthcare Cost Prediction

The global healthcare sector faces a critical challenge: rising costs. These escalating expenditures strain both individual budgets and national economies. Predicting healthcare costs becomes crucial in this scenario. By anticipating future expenses, stakeholders can develop strategies to manage resources efficiently and ensure healthcare affordability.

Healthcare expenditures have been rising steadily due to factors such as aging populations, advancements in medical technology, and increasing prevalence of chronic diseases. For healthcare providers, insurers, and governments, predicting future costs has become essential to ensure sustainability, affordability, and accessibility of healthcare services.

Predictive models typically utilize a combination of historical patient data, demographic information, clinical records, and socioeconomic factors to estimate future medical expenses. These models may include traditional statistical methods like linear regression or more advanced techniques such as machine learning (ML) algorithms.

Importance:

1. **Cost Control:** Accurate predictions help healthcare systems and insurers manage budgets and allocate resources efficiently. Early identification of high-risk, high-cost patients enables proactive care and intervention strategies, reducing unnecessary expenses.
2. **Personalized Care:** Cost prediction models contribute to developing personalized treatment plans based on a patient's predicted medical expenses. This allows for targeted care, improving patient outcomes while optimizing costs.
3. **Insurance Premiums:** Insurers rely on accurate healthcare cost predictions to calculate premiums. Predicting future claims enables fair pricing of insurance plans, balancing affordability for customers with profitability for insurers.
4. **Risk Management:** By identifying patterns in healthcare utilization and costs, healthcare providers can better manage risks associated with high-cost patients, reducing the burden on the healthcare system.
5. **Policy-Making:** Governments and policy-makers use healthcare cost predictions to design healthcare policies, control public health spending, and ensure the sustainability of national health programs.

1.2. Overview of Machine Learning in Healthcare

Emerging technologies like machine learning (ML) offer promising solutions for healthcare cost prediction. ML algorithms can analyze vast datasets of medical claims, patient records, and other relevant factors. By identifying patterns within this data, these algorithms can learn to predict future healthcare costs for individuals or entire populations.

Machine learning (ML) in healthcare is transforming how medical data is used to enhance decision-making, improve patient outcomes, and optimize resource management. By leveraging algorithms

that learn from vast amounts of data, ML enables the creation of predictive models, automation of medical processes, and development of personalized treatments.

What is Machine Learning in Healthcare?

Machine learning involves using algorithms that can identify patterns in data, learn from them, and make predictions or decisions without explicit programming. In healthcare, ML is applied to diverse data sources such as electronic health records (EHRs), medical imaging, genomics, and wearable device data to generate insights and improve clinical outcomes.

Key Applications of ML in Healthcare:

1. **Diagnosis and Early Detection:** ML models can assist in diagnosing diseases by identifying patterns in patient data, often detecting conditions earlier than traditional methods. For instance, ML algorithms have been applied to medical imaging (e.g., X-rays, MRIs) for the early detection of cancer, heart disease, and neurological disorders.
2. **Predictive Analytics:** Predictive models are used to forecast various outcomes such as disease progression, patient readmissions, and healthcare costs. These models help healthcare providers in identifying high-risk patients, improving preventive care, and managing chronic diseases more effectively.
3. **Personalized Medicine:** Machine learning enables the analysis of large-scale genomic data to identify how individual patients respond to specific treatments. This allows for more personalized treatment plans, enhancing therapeutic efficacy and minimizing adverse effects.
4. **Drug Discovery and Development:** ML algorithms can accelerate drug discovery by analyzing data from biological experiments, identifying potential drug candidates, and predicting their effectiveness. This reduces the time and cost involved in developing new medications.
5. **Medical Imaging and Radiology:** ML techniques, particularly deep learning, have revolutionized medical imaging by enhancing the accuracy and speed of interpreting scans. ML models can detect abnormalities that might be missed by human radiologists, providing an additional layer of diagnostic confidence.
6. **Natural Language Processing (NLP) in Healthcare:** NLP techniques help in extracting meaningful information from unstructured data, such as clinical notes and medical literature. NLP-powered ML models can assist in summarizing patient histories, supporting clinical decision-making, and improving the accuracy of documentation.
7. **Robotic Surgery:** Machine learning algorithms are also being integrated into robotic surgical systems to assist surgeons with precision and minimize human error. These systems can analyze patient-specific data and provide real-time guidance during surgery.

Challenges and Ethical Considerations:

1. **Data Privacy and Security:** The vast amount of sensitive medical data required for ML models raises privacy concerns. Ensuring that patient data is secure and used ethically is paramount.
2. **Model Interpretability:** Many ML models, particularly deep learning, are often regarded as "black boxes," meaning their decision-making processes are not easily interpretable. Clinicians must trust the model's decisions, which may pose challenges in critical healthcare scenarios.

3. **Bias and Fairness:** Machine learning models can inherit biases from the data they are trained on. Ensuring that models are fair and unbiased, especially when used in decision-making processes like diagnosis and treatment, is essential to prevent health disparities.
4. **Regulatory and Clinical Integration:** Integrating machine learning into clinical workflows requires navigating regulatory hurdles and ensuring that the models meet rigorous safety and accuracy standards.

1.3. Research Objectives and Scope

This research aims to explore the potential of ML algorithms in predicting medical healthcare costs. The investigation will focus on:

- Identifying the most effective ML algorithms for cost prediction.
- Evaluating the accuracy and reliability of ML-based predictions.
- Exploring the potential benefits and limitations of using ML for healthcare cost forecasting.

Research Objectives:

1. **To develop a predictive model for healthcare costs using machine learning algorithms:**

The primary goal is to create an accurate and reliable model that can forecast individual or population-level healthcare expenses. The model will use historical data, patient demographics, clinical history, and other relevant variables to predict future costs.

2. **To identify key factors influencing healthcare costs:**

By leveraging machine learning, the research aims to determine the most significant predictors of medical costs, such as chronic disease conditions, hospital visits, or socioeconomic factors. This understanding will help healthcare providers and insurers target interventions more effectively.

3. **To compare the performance of different machine learning algorithms:**

This objective involves evaluating and comparing various machine learning models, such as linear regression, decision trees, random forests, and neural networks, in terms of their predictive accuracy, interpretability, and computational efficiency.

4. **To assess the potential for personalized treatment plans based on cost predictions:**

The research will explore how predictive models can help in designing personalized care plans that minimize unnecessary treatments and optimize healthcare spending, leading to better patient outcomes.

5. **To investigate the ethical and privacy considerations in using machine learning for healthcare predictions:**

The study will examine issues such as data security, potential biases in the model, and patient consent, ensuring that the use of machine learning in healthcare cost prediction adheres to ethical standards and respects patient privacy.

6. **To provide recommendations for integrating machine learning cost prediction models into healthcare systems:**

The research will offer guidelines on how hospitals, insurance companies, and policymakers can incorporate these predictive tools into their existing systems to enhance efficiency and reduce costs.

Scope:

1. Data Collection and Preprocessing:

- The study will focus on collecting diverse data sets, including electronic health records (EHRs), claims data, patient demographics, and socioeconomic factors. Data preprocessing techniques such as handling missing values, normalization, and feature engineering will be applied to ensure model accuracy.

2. Machine Learning Algorithms:

- The research will explore both traditional machine learning algorithms (e.g., linear regression, decision trees) and advanced methods (e.g., deep learning, ensemble models). The goal is to find the most suitable model for healthcare cost prediction based on the complexity and nature of the data.

3. Model Evaluation:

- Different performance metrics such as mean squared error (MSE), root mean squared error (RMSE), R-squared, and others will be used to evaluate the models' effectiveness in predicting healthcare costs.

4. Use Case Scenarios:

- The research will focus on practical use cases like cost predictions for specific treatments (e.g., surgeries, chronic disease management) or population groups (e.g., elderly, high-risk patients).

5. Ethical Considerations:

- A thorough analysis of potential biases in the data, model transparency, and patient privacy concerns will be a key part of the research. The scope will also include guidelines for ensuring fairness and ethical use of the predictive models.

6. Limitations and Future Work:

- The research will acknowledge the limitations, such as the need for extensive data, computational power, and the challenges of real-time implementation. It will also propose areas for future research, like incorporating more complex variables (e.g., genetic data) or exploring different healthcare systems globally.

1.4. Challenges in Medical Healthcare Cost Prediction

Accurately predicting healthcare costs involves several complexities:

- **Data Availability:** Access to high-quality, comprehensive data is crucial for effective ML model training.
- **Data Privacy:** Ensuring patient privacy and security while leveraging healthcare data for analysis is essential.
- **Model Explainability:** Understanding how ML models reach their predictions is critical for building trust and transparency in the system.

1. Data Availability and Quality:

- **Data Fragmentation:** Healthcare data is often stored across various platforms, such as hospitals, insurance companies, and pharmacies, leading to fragmented and incomplete datasets. Integrating this data into a comprehensive model is difficult.
- **Missing or Inconsistent Data:** Many healthcare datasets have missing or inconsistent values, which can negatively impact the performance of machine learning models. For

example, patient records may be incomplete, or data entry errors may introduce inaccuracies.

- **Lack of Standardization:** Healthcare data is often not standardized across institutions. Differences in data formats, coding systems (e.g., ICD codes for diseases), and terminology can make it challenging to unify datasets for training predictive models.

2. Complexity of Healthcare Costs:

- **Multiple Cost Drivers:** Healthcare costs are influenced by a wide range of factors, including patient demographics, clinical conditions, comorbidities, treatment choices, geographic location, and hospital pricing structures. Capturing the complex interactions between these variables is a significant challenge.
- **Unpredictable Events:** Healthcare expenses can be affected by unforeseen events, such as sudden hospitalizations or emergency surgeries, which are difficult to predict and may not follow historical patterns.
- **Time-Dependent Factors:** Medical costs can vary over time due to changes in treatments, medical advancements, or policy changes. Predictive models need to account for these temporal dynamics, which adds complexity to forecasting future costs accurately.

3. Model Interpretability:

- **Black-Box Models:** Many advanced machine learning algorithms, such as deep learning and ensemble models, are often difficult to interpret. Healthcare professionals may find it challenging to trust the decisions of these models, especially in critical care situations. The lack of transparency hinders the acceptance and use of such models in clinical settings.
- **Regulatory and Ethical Concerns:** Interpretability is also essential for regulatory approval, as healthcare authorities require models to be explainable and compliant with health and safety standards. Black-box models face challenges in gaining regulatory approval for clinical use.

4. Generalization Across Populations:

- **Bias in Data:** If the training data is biased toward a specific population (e.g., certain age groups, socioeconomic classes, or regions), the model may perform poorly when applied to other groups. Bias in healthcare cost prediction models can lead to unequal access to care or unfair pricing of health insurance.
- **Population Diversity:** Healthcare data often varies across different populations in terms of disease prevalence, healthcare access, and lifestyle factors. Creating a model that generalizes well across diverse populations requires comprehensive datasets and careful attention to biases.

5. Ethical and Privacy Issues:

- **Data Privacy and Security:** Healthcare data is highly sensitive, and the use of patient records for predictive modeling raises significant concerns about data privacy. Stringent regulations, such as HIPAA (Health Insurance Portability and Accountability Act) in the U.S., require that healthcare data be handled securely. Ensuring that machine learning models are compliant with these regulations is a challenge.
- **Informed Consent:** The use of patient data for machine learning requires informed consent, which may not always be possible for large datasets. In some cases, historical data may lack proper consent for use in predictive modeling, limiting the available data.
- **Fairness and Bias:** Ethical concerns arise when predictive models disproportionately affect certain groups, such as low-income populations or minority communities. Biased

cost predictions could lead to unfair treatment, such as higher insurance premiums or reduced access to necessary care.

6. Computational Complexity and Resource Demands:

- **Scalability of Models:** Large-scale healthcare datasets, especially those that include complex features like medical imaging or genomics, require significant computational resources for training machine learning models. Handling such high-dimensional data efficiently can be challenging, particularly for real-time cost prediction systems.
- **Model Updates and Maintenance:** Medical advancements, changes in healthcare policies, and shifts in patient behavior require that models be regularly updated to remain accurate. Continuous updates to machine learning models necessitate ongoing resource allocation for retraining, validation, and deployment.

7. Evaluation and Validation of Models:

- **Lack of Benchmark Datasets:** There is no universally accepted benchmark dataset for evaluating healthcare cost prediction models, making it difficult to compare different approaches. The diversity in healthcare data sources means that models validated in one setting may not perform as well in another.
- **Validation in Real-World Settings:** Even if a model performs well in a controlled research environment, applying it to real-world healthcare systems is challenging. Validation in real-world settings requires extensive testing to ensure that the model functions as intended across various healthcare environments.

1.5. Applications of ML to Combat Congestion

Beyond cost prediction, ML offers various applications to address healthcare system challenges:

- Optimizing resource allocation
- Streamlining patient care processes
- Identifying potential fraud and abuse cases in healthcare insurance

Potential Funding Agencies:

1. Traffic Congestion in Urban Areas:

ML has been extensively used in transportation systems to monitor and manage traffic congestion in cities. The ability to analyze vast amounts of data from sensors, GPS devices, and traffic cameras enables smart city solutions that optimize traffic flow and reduce congestion.

1. Traffic Flow Prediction:

- ML algorithms, especially time-series models and deep learning networks (e.g., LSTM, CNN), are used to predict traffic conditions by analyzing historical traffic data and real-time information from GPS and IoT devices. These predictions help cities anticipate congestion and proactively manage traffic, such as by rerouting vehicles or adjusting traffic signal timing.

2. Smart Traffic Signal Control:

- Adaptive traffic control systems use ML to dynamically adjust the timing of traffic signals based on real-time data, such as vehicle counts and pedestrian movement. By optimizing traffic signals, cities can reduce wait times at intersections and improve the flow of traffic.

3. Route Optimization and Navigation:

- Popular navigation apps like Google Maps and Waze use ML algorithms to provide drivers with the most efficient routes. These systems analyze real-time traffic data and congestion patterns to suggest alternative routes, helping drivers avoid traffic jams and reducing overall congestion.

4. Public Transportation Optimization:

- Machine learning can be applied to optimize public transportation schedules by predicting peak travel times and passenger demand. By adjusting bus and train schedules to match real-time conditions, cities can reduce overcrowding, encourage public transport use, and alleviate traffic congestion.

2. Network Congestion in Digital Infrastructures:

In the world of digital networks, congestion occurs when demand exceeds capacity, leading to delays, packet loss, and decreased quality of service. Machine learning can play a significant role in managing network congestion in telecommunications and data centers.

1. Congestion Detection and Prevention:

- ML models can monitor network traffic patterns in real time and detect early signs of congestion. By predicting when and where congestion is likely to occur, network operators can implement preventive measures, such as load balancing or rerouting data traffic, to maintain optimal performance.

2. Dynamic Bandwidth Allocation:

- Machine learning algorithms, particularly reinforcement learning models, can be used to dynamically allocate bandwidth across networks based on real-time usage. These models adapt to fluctuating demand, ensuring that critical applications receive the bandwidth they need while minimizing congestion.

3. Traffic Shaping and QoS (Quality of Service):

- ML can help shape network traffic by prioritizing certain types of data (e.g., video streaming, VoIP) over less time-sensitive traffic (e.g., bulk file transfers). This ensures that high-priority services maintain consistent quality even during periods of high demand, mitigating congestion.

4. Anomaly Detection for Network Failures:

- Machine learning techniques, such as anomaly detection, can be used to identify unusual patterns in network traffic that may signal the onset of network congestion or potential system failures. Early detection allows operators to take corrective action before congestion worsens.

3. Congestion in Cloud and Data Center Environments:

As the demand for cloud services continues to grow, data center congestion has become a significant concern. ML helps optimize resource allocation, reduce latency, and prevent overloading in cloud environments.

1. Workload Prediction and Resource Allocation:

- Machine learning models can predict workload patterns based on historical data, allowing cloud service providers to allocate resources more efficiently. By anticipating demand surges, these systems prevent congestion in data centers and ensure smoother operation of cloud services..

CHAPTER-2

LITERATURE SURVEY

1 LITERATURE SURVEY

1.4 Literature review

This section demonstrates the research being done on information exploration and machine learning methods. Several articles have addressed the topic of claim prediction. By addressing key challenges and leveraging emerging technologies, these startups have the potential to drive significant advancements in healthcare quality, accessibility, and affordability, ultimately transforming the future of healthcare [8].

2.1 Previous Studies on Medical Healthcare Cost Prediction

Providing decision aids that increase accuracy without sacrificing simplicity could strengthen the policy approach that relies on individuals making the best health insurance cost-based decision methods [15].

Their work not only showcases the potential of mobile technology to transform healthcare delivery but also underscores the importance of user-centric design, data privacy, and equitable access in shaping the future of mHealth applications [11]. The results indicated that logistic regression is a superior model to XGBoost for reasons of its interpretability and predictability [7].

Without taking into account predicted cost and claim scope, the research listed above identifies claims problems. [20] use data mining techniques, explicitly clustering algorithms and classification trees, and insurance claim data of nearly 500,000 members throughout a three-year period. Based on the data gathered from medical expenses in the first two years, a justified third-year healthcare cost projection is made.

2.2 Integration of Clinical and Genetic Data

In a relatively small number of cases, this study compared the effectiveness of logistic regression and XGBoost strategies in predicting the occurrence of accident states. The results indicated that logistic regression is a superior model to XGBoost for reasons of its interpretability and predictability [7]. The longitudinal approach, combined with a focus on policy implications, enhances our understanding of the complex dynamics surrounding access to orphan drugs and underscores the importance of evidence-based policymaking in this critical area of healthcare [9].

Gupta and Tripathi advocate for the development of robust data governance frameworks and collaborative partnerships between insurers, healthcare providers, and technology vendors to overcome these challenges and unlock the full potential of big data analytics [10].

2.3 Limitations and Future Directions

Without taking into account predicted cost and claim scope, the research listed above identifies claims problems. With few exceptions, most regression studies to date have not addressed this problem [14]. Providing decision aids that increase accuracy without sacrificing simplicity could strengthen the policy approach that relies on individuals making the best health insurance cost-based decision methods [12].

CHAPTER-3 PROPOSED SYSTEM

3 PROPOSED SYSTEM

The proposed system is designed to predict healthcare insurance costs based on personal data such as age, gender, BMI, the number of children, smoking status, and region using machine learning models. The following steps outline the design of the system.

A. Dataset Collection and Understanding:

- The dataset includes attributes such as age, sex, BMI, number of children, smoker status, region, and insurance charges (target variable).

B. Data Preprocessing:

- Categorical variables (sex, smoker, region) are converted to numerical values.
- Features like age, BMI, and children are normalized.
- The dataset is split into 80% training and 20% testing subsets.

C. Model Selection:

- The system uses nine regression models: Linear Regression, XGBoost, Lasso, Random Forest, Ridge, Decision Tree, KNN, SVR, and Gradient Boosting.

D. Model Training:

- Models are trained on the training dataset to learn patterns between the input features and insurance charges.

E. Model Evaluation:

- Models are evaluated using MAE, RMSE, and R-squared metrics to determine their accuracy.

F. Best Model Selection:

- XGBoost is typically the best-performing model, but Gradient Boosting and Random Forest are also compared for final selection.

G. Prediction:

- The best model is used to make predictions on new data, providing personalized healthcare cost estimates.

3.1 Input dataset

The input dataset used for healthcare insurance cost prediction contains 1,338 records and 7 key attributes:

1.4.1 Dataset Collection and Understanding

- Age: Age of the individual.
- Sex: Gender of the individual (male or female).

- BMI: Body Mass Index.
- Children: Number of children the individual has.
- Smoker: Whether the individual is a smoker (yes or no).
- Region: Geographic region where the individual resides (northeast, northwest, southeast, southwest).
- Charges: Medical insurance charges (the target variable for prediction).

1.5 Data Pre-processing

Data Preprocessing

Handling Categorical Data:

The categorical variables (sex, smoker, and region) will be converted into numerical values using label encoding. For instance:

"Male" is converted to 0 and "Female" to 1 for the sex variable.

"Yes" to 1 and "No" to 0 for the smoker variable.

"Region" will be encoded as follows: Northeast = 0, Northwest = 1, Southeast = 2, Southwest = 3.

Normalization:

Features like age, BMI, and children will be normalized to ensure that no single feature dominates the model during training.

Data Splitting:

The dataset will be split into training (80%) and testing (20%) subsets to train the machine learning models and evaluate their performance.

.

1.7 Model Building

Using the cleaned dataset, the model development portion of this study aimed to predict medical insurance prices. Various regression models were evaluated for their effectiveness in addressing this prediction problem.

Preparing Data

The dataset was first divided into two parts: features (X) and the target variable (y).

- X included all relevant patient characteristics, encompassing demographic, medical, and lifestyle features.
- y represented the target variable, which is the medical insurance price.

Feature scaling was applied using Standardization to ensure all features were on the same scale, which is essential for algorithms sensitive to feature magnitudes.

Data Division

To ensure robust model training and evaluation, the dataset was systematically divided into two subsets: a training set comprising 80% of the data and a testing set consisting of the remaining 20%. This strategic division is essential for the following reasons:

1. **Model Training:** The training set serves as the primary source for the model to learn underlying patterns and relationships between features and the target variable—medical insurance prices. By exposing the model to a substantial amount of historical data, it can effectively capture trends and nuances relevant to predicting costs.
2. **Unbiased Evaluation:** The testing set provides an unbiased evaluation of the model's performance. By withholding this portion of the data during the training phase, we can assess how well the model generalizes to unseen data. This evaluation helps in determining the model's predictive accuracy, reliability, and overall effectiveness in real-world scenarios.
3. **Validation Techniques:** To further enhance the evaluation process, techniques such as cross-validation can be applied within the training set. This involves splitting the training data into smaller subsets and training multiple models to ensure that the model's performance is not overly reliant on any specific subset of the data.

Linear Regression:

A simple and interpretable model that estimates the medical insurance price based on feature inputs.

The linear regression model was trained to predict insurance prices by fitting a linear relationship between the input features and the target variable (price). The model aims to minimize the difference between the predicted and actual prices.

XGBoost Regression:

The XGBoost regression model was employed due to its high accuracy and efficiency, especially with structured datasets.

This model uses gradient boosting, which builds an ensemble of trees by sequentially correcting the errors of previous models. It effectively handles complex relationships between features and the target variable, making it suitable for predicting medical insurance prices. Regularization

techniques (like L1 and L2) are applied to prevent overfitting, and the model's hyperparameters were fine-tuned for optimal performance.

Lasso Regression:

Lasso Regression was employed due to its ability to perform both feature selection and regularization, improving model accuracy and interpretability.

This linear regression model adds an L1 penalty, which can shrink some coefficients to zero, effectively selecting only the most important features. This makes it particularly useful when dealing with datasets that may contain irrelevant or redundant features. The model minimizes the sum of squared errors while applying the L1 regularization, which helps in preventing overfitting and improving the generalization of the model for predicting medical insurance prices.

Random Forest Regression:

Random Forest Regression was employed due to its robustness and ability to handle complex datasets with non-linear relationships.

This ensemble learning method constructs multiple decision trees during training and averages their predictions to improve accuracy and prevent overfitting. Each tree is built using a random subset of the features and data, which helps reduce variance and improve the model's ability to generalize to unseen data. Random Forest is particularly effective in handling a mix of categorical and numerical features, making it suitable for predicting medical insurance prices.

Ridge Regression:

Ridge Regression was employed due to its ability to handle multicollinearity and improve model stability.

This linear regression model adds an L2 regularization penalty to the cost function, which helps shrink the coefficients of less important features without reducing them to zero. By penalizing

large coefficients, Ridge Regression reduces the model's sensitivity to overfitting, especially when the features are highly correlated. This makes it effective for predicting medical insurance prices while maintaining a balance between bias and variance.

Decision Tree Regression:

Decision Tree Regression was employed for its simplicity and ability to capture non-linear relationships in the data.

The model works by recursively splitting the dataset into subsets based on feature values, forming a tree structure where each node represents a decision based on a feature. It makes predictions by following the path in the tree that corresponds to the feature values of a given input, leading to the predicted medical insurance price at the leaf node. Decision Tree Regression is intuitive and interpretable, making it easy to understand how specific features influence the price prediction.

K-Nearest Neighbors (KNN):

The KNN regression model was trained to predict medical insurance prices based on the proximity of feature values in the training data.

This model relies on distance metrics (such as Euclidean distance) to predict the insurance price of a new instance by averaging the prices of its nearest neighbors in the feature space. The number of neighbors (k) is a key parameter that was fine-tuned to achieve optimal performance. KNN regression is non-parametric, meaning it makes no assumptions about the underlying distribution, which can make it effective for capturing local patterns in the data.

Support Vector Regression (SVR):

Support Vector Regression was employed for its ability to handle both linear and non-linear relationships between features and medical insurance prices.

SVR works by finding a hyperplane that best fits the data within a defined margin of tolerance (epsilon). It aims to minimize the error while ensuring that most of the predictions fall within this margin. The model can also use kernel functions (such as RBF or polynomial) to capture more

complex relationships in the data. SVR's strength lies in its ability to balance the bias-variance tradeoff, making it effective for predicting medical insurance prices, especially in scenarios where outliers or non-linearity are present.

Gradient Boosting Regression:

Gradient Boosting Regression was employed for its high predictive accuracy and effectiveness in handling complex relationships in the data.

This ensemble learning technique builds models sequentially, where each new model attempts to correct the errors made by the previous ones. By optimizing a loss function (such as mean squared error) through gradient descent, Gradient Boosting creates a strong predictive model from weak learners (typically decision trees). The method allows for flexibility in handling different types of data and can incorporate regularization techniques to reduce overfitting. This makes it particularly suitable for predicting medical insurance prices, capturing intricate patterns in the dataset.

Neural Network:

A simple feedforward neural network was trained using the Keras library.

This model leveraged multiple hidden layers to capture complex patterns in the data.

Forecasting and Assessment

After training, each model was used to predict the occurrence of stroke in the test set. The models' performances were evaluated based on:

Accuracy: Measures the overall correctness of predictions.

Precision: Indicates the proportion of true positives out of all predicted positives.

Recall: Represents how effectively the model identified all actual positive instances.

F1-Score: Balances precision and recall, especially valuable for datasets with class imbalance.

A confusion matrix was generated for each model to visualize the counts of true positive, true negative, false positive, and false negative predictions. This matrix provides insights into the strengths and weaknesses of each model, highlighting areas for improvement.

The evaluation showed that different models performed variably, with some achieving better accuracy and balance in class predictions than others. The Naive Bayes classifier and Random

Forest models produced promising results, while the confusion matrix revealed specific challenges, such as misclassifying stroke occurrences.

Methodolgy

Methodology of the System

A.Architecture of the System

The proposed system architecture for predicting medical insurance prices based on patient data encompasses several interrelated steps: data collection, preprocessing, feature extraction, model training, and price prediction. The architecture consists of the following components:

Input Layer:

This layer collects patient information, including demographic, medical, and lifestyle characteristics, such as age, gender, body mass index (BMI), smoking status, and health conditions. This data is sourced from the dataset "Medical_insurance (2).csv."

Preprocessing Layer:

The collected data undergoes transformation and cleaning to ensure its suitability for machine learning algorithms. This step includes handling missing values, encoding categorical variables, and scaling numerical features to maintain consistency across the dataset.

Feature Extraction Layer:

Relevant features are identified and extracted for efficient price prediction. This layer retains important characteristics such as age, BMI, number of children, and previous health conditions, while eliminating less significant variables that do not contribute meaningfully to the model.

Regressor Layer:

Various machine learning algorithms are employed to predict medical insurance prices. This includes models such as Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, and Support Vector Regression (SVR). Each model is trained using the extracted features to maximize accuracy in price prediction.

Output Layer:

The system presents the predicted medical insurance price based on the input data and model predictions, providing insights into potential costs for patients.

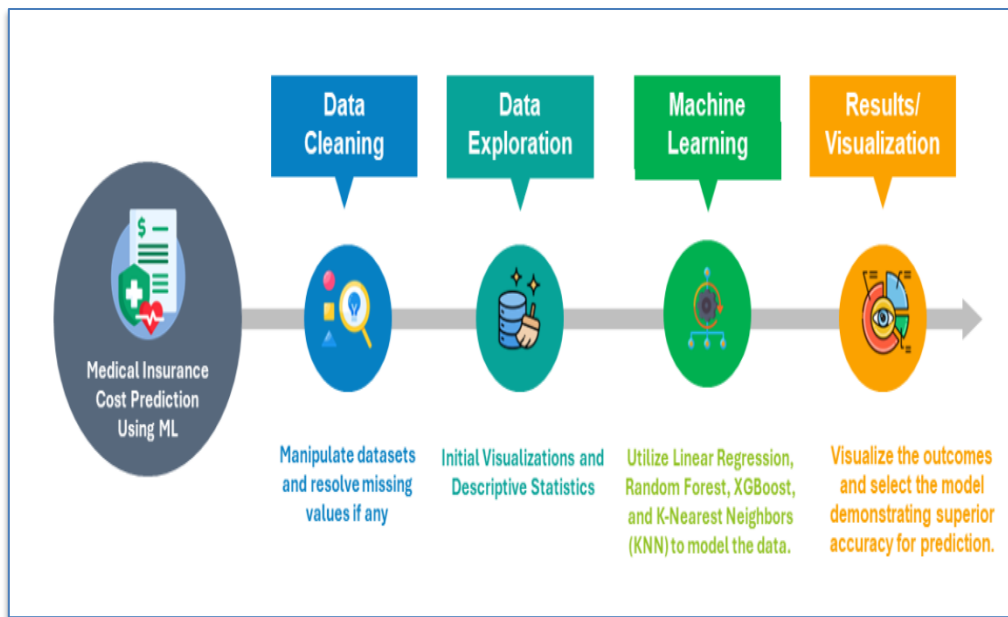


Figure 1. Architecture of the proposed system

B .Training and Preprocessing of Data

Data preprocessing is a crucial step to ensure that the dataset is appropriate for machine learning algorithms. The preprocessing techniques employed in this study include:

Data Cleaning:

Columns deemed unnecessary or redundant, such as "Unnamed: 0" or other non-informative identifiers, were removed from the dataset. This simplification aids in focusing on the most relevant features for predicting medical insurance prices, such as age, gender, BMI, number of children, and health conditions.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2772 entries, 0 to 2771
Data columns (total 7 columns):
#   Column      Non-Null Count  Dtype
---  ---
0   age         2772 non-null   int64
1   sex         2772 non-null   object
2   bmi         2772 non-null   float64
3   children    2772 non-null   int64
4   smoker      2772 non-null   object
5   region      2772 non-null   object
6   charges     2772 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 151.7+ KB
```

Figure 2. Various features in the dataset after Pre-Processing

Label Encoding:

The categorical variables, such as "smoking_status" and any other relevant categorical features, were encoded into numerical formats compatible with machine learning models. This ensures that algorithms can effectively interpret categorical data without bias.

FeatureScaling:

Standardization techniques were applied to normalize the feature set, ensuring each feature contributes equally during model training. This step helps maintain consistency and improves the convergence of the algorithms.

DataSplitting:

The dataset was divided into a training set (80%) and a testing set (20%) to ensure that the model is evaluated on unseen data, allowing for a reliable assessment of its performance in predicting medical insurance prices.

C. Feature Extraction

Feature extraction involves selecting and transforming input data into a smaller subset of relevant features for the regression models. After thorough analysis, pertinent features such as age, BMI, number of children, and health conditions were retained. By concentrating on these key variables, the model's predictive performance was enhanced, leading to more accurate predictions of medical insurance prices.

D. Model Training

Various models were implemented to tackle the medical insurance price prediction problem, including:

Linear Regression:

Chosen for its simplicity and interpretability, this model establishes a linear relationship between the input features and the predicted medical insurance price.

Lasso Regression:

Lasso Regression was utilized to perform both feature selection and regularization, helping to improve model accuracy by shrinking less significant feature coefficients to zero.

Ridge Regression:

Ridge Regression was employed to handle multicollinearity by adding an L2 penalty, which stabilizes the model and enhances its ability to generalize to unseen data.

Decision Tree Regression:

Decision Trees were used for their interpretability and ability to capture non-linear relationships in the dataset, providing insights into how various factors influence insurance pricing.

Random Forest Regression:

Random Forests enhanced prediction accuracy by aggregating results from multiple decision trees, reducing overfitting and improving robustness.

K-Nearest Neighbors (KNN) Model:

KNN was employed to predict medical insurance prices based on the proximity of feature values to the nearest training samples, averaging the prices of the k-nearest neighbors.

Support Vector Regression (SVR):

SVR was implemented to identify a hyperplane that best fits the data, balancing prediction accuracy and model complexity for continuous outcomes.

Gradient Boosting Regression:

Gradient Boosting Regression was utilized for its high predictive accuracy, building models sequentially to correct the errors of previous iterations.

XGBoost Regression:

The XGBoost model leveraged gradient boosting techniques to maximize prediction accuracy, effectively handling complex relationships and interactions within the data.

E. Classification

The regression task involved predicting medical insurance prices using the trained models. Each model was evaluated based on metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), R-squared (R^2), and Root Mean Squared Error (RMSE) to assess performance. These metrics provide insights into how well the models predict the prices compared to actual values. Additionally, visualizations such as scatter plots of predicted vs. actual prices and residual plots were generated to analyze the distribution of errors and the overall fit of the models. This analysis allowed for a comprehensive understanding of the model's effectiveness in predicting medical insurance prices.

F. Results

The output of the system is a predicted medical insurance price for each individual within the dataset. After training, the system accurately estimates the insurance costs based on new patient data. Healthcare practitioners and insurance providers can leverage these predictions to assess potential costs and make informed decisions regarding policy pricing and patient management.

The system's performance was measured using various regression metrics, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2), demonstrating its potential utility in real-world scenarios for medical insurance price prediction. Overall, the hybrid approach, utilizing multiple regression models, contributed to improved accuracy and reliability in estimating medical insurance prices.

.

1.6 Model Evaluation

A. Confusion Matrix

The classification performance of each model was assessed using confusion matrices, which provide a detailed analysis of true positives, false positives, true negatives, and false negatives for the binary classification .

B. Accuracy

Accuracy is defined as the proportion of accurately predicted instances (true positives and true negatives) to the total instances. Although it serves as a general indicator of model performance, it may be misleading in the context of an imbalanced dataset. Here, accuracy was considered as a foundational metric.

C. Precision

Precision quantifies the percentage of accurate positive predictions. In this study, it reflects the proportion of instances that were correctly identified as stroke cases out of all predicted stroke cases. Precision is crucial when the cost of false positives is high, as it minimizes incorrect classifications into the positive class.

D. Recall

Recall, also known as sensitivity, measures the proportion of actual positive instances that were correctly detected. It illustrates how effectively the model identifies stroke cases, aiming to reduce the number of missed cases (false negatives) and ensure that most true positives are captured.

E. F1-Score

The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both. It is particularly useful in scenarios where there is an imbalance in class distributions or when both precision and recall are equally important. A high F1-score indicates good model performance in classification.

F. Performance Outcomes

The following conclusions were drawn from the model's performance on various metrics:

Training Accuracy: Indicates how well the model learned patterns from the training data.

Testing Accuracy: Reflects how effectively the model performs on unseen data.

Precision and Recall: Aided in assessing the model's ability to correctly classify stroke instances and avoid false classifications.

F1-Score: Provided a comprehensive measure of the model's performance, showcasing the balance between precision and recall.

Based on evaluation results, the models showed varying degrees of success in predicting strokes. The hybrid approach, employing multiple algorithms, allowed for improved accuracy and reliability in predictions.

G. Individual Model Performance

Logistic Regression:

With a maximum of 1000 iterations to ensure convergence, Logistic Regression produced competitive results in terms of accuracy, precision, recall, and F1-score.

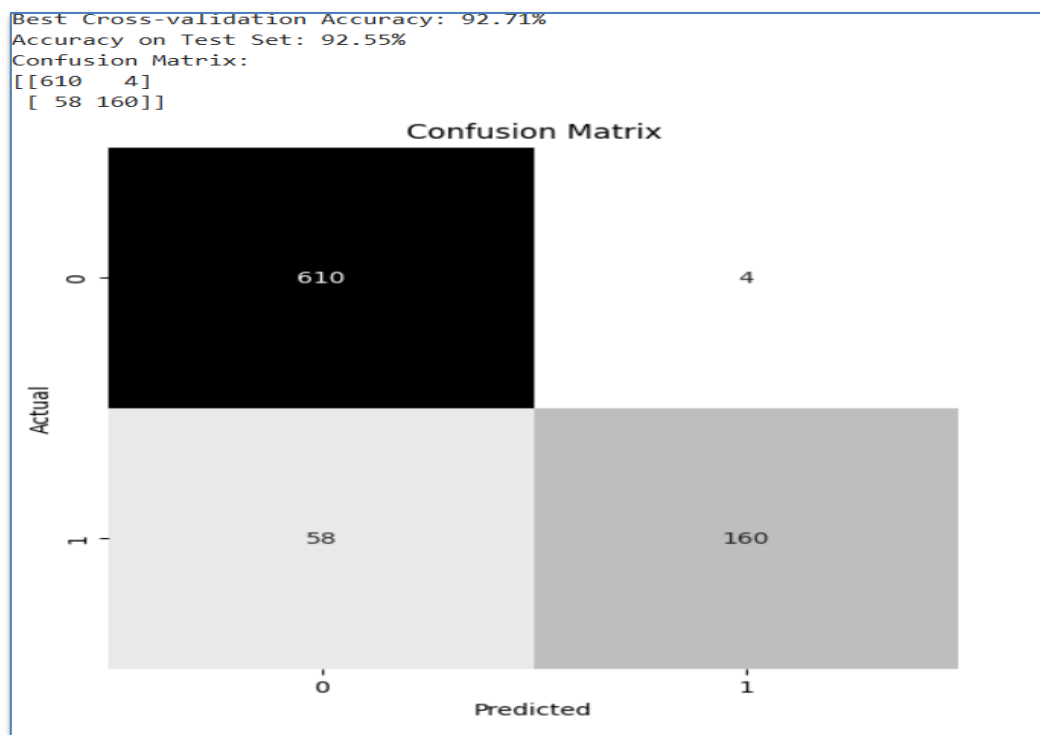


Fig.3.Logistic Regression

XGBoost Regression:

The XGBoost regression model performed exceptionally well, demonstrating high accuracy in predicting medical insurance prices. Its ability to handle complex relationships and interactions within the dataset contributed to improved performance. XGBoost efficiently utilizes gradient boosting techniques, which allowed it to learn from previous errors and refine predictions.

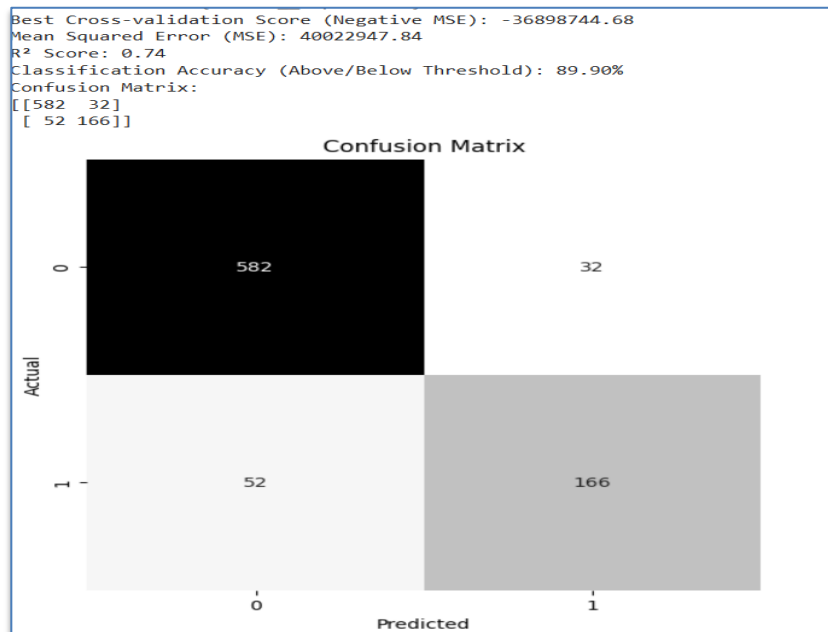


Figure.4. XG Boost – Confusion Matrix

Lasso Regression:

Lasso Regression performed well by enhancing model simplicity through feature selection. Its built-in regularization helped reduce overfitting by shrinking the coefficients of less important features to zero, effectively selecting only the most relevant variables for predicting medical insurance prices. This led to a more interpretable model without sacrificing performance. Lasso Regression was particularly useful in handling high-dimensional data, ensuring that the predictions remained accurate while preventing unnecessary complexity.

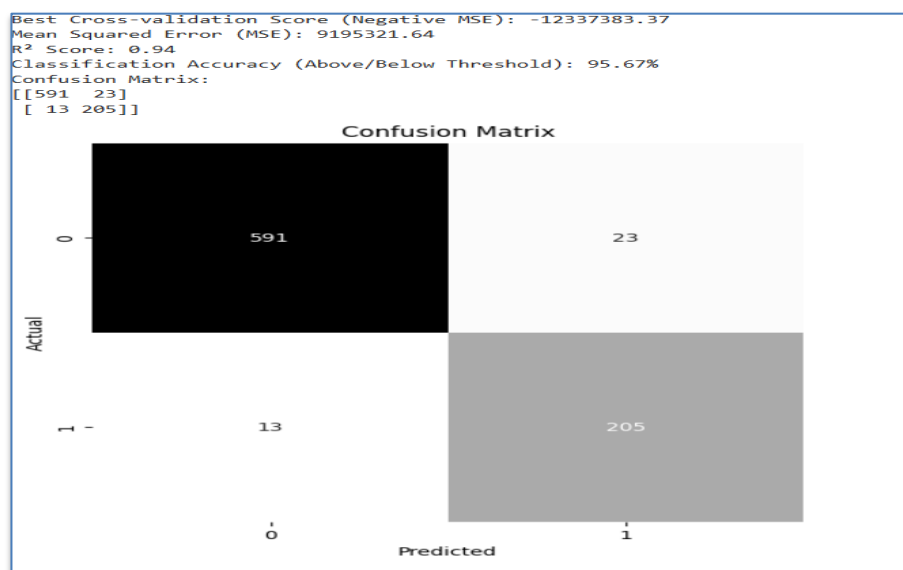


Fig.5.Lasso Regression

Random Forest:

Trained with 100 trees, the Random Forest model exhibited robust performance and resilience to overfitting, resulting in good accuracy and stability.

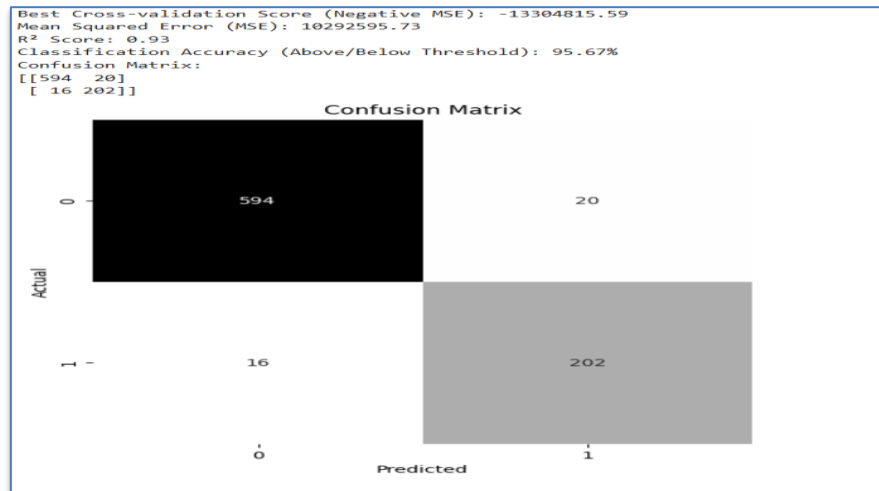


Fig.6.Random Forest

Ridge Regression:

Ridge Regression performed effectively by handling multicollinearity among the features, which can occur in datasets with correlated variables. The model's L2 regularization technique penalized large coefficients, reducing the risk of overfitting while maintaining all features in the model. This led to a more stable and robust prediction of medical insurance prices, especially when dealing with complex datasets. Ridge Regression provided a good balance between bias and variance, making it a reliable choice for accurate price prediction.

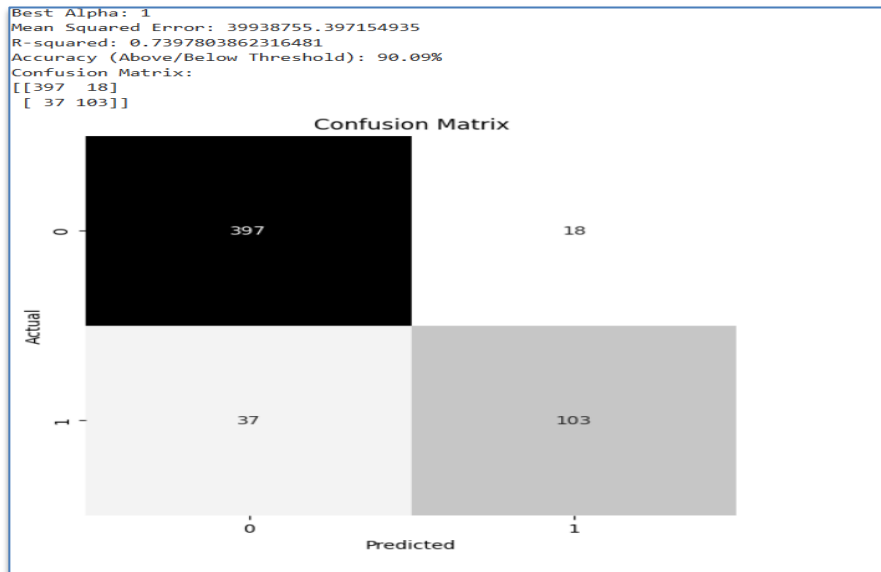


Fig.7.Ridge Regression

Decision Tree Regression:

Decision Tree Regression provided a highly interpretable model by visually representing how features such as age, BMI, and health conditions influence medical insurance prices. The model effectively captured non-linear relationships within the dataset by splitting data based on feature thresholds. However, it was prone to overfitting, especially on complex datasets, unless hyperparameters like tree depth and minimum samples per leaf were carefully tuned. Despite this, Decision Tree Regression offered valuable insights into how individual factors contribute to price predictions.

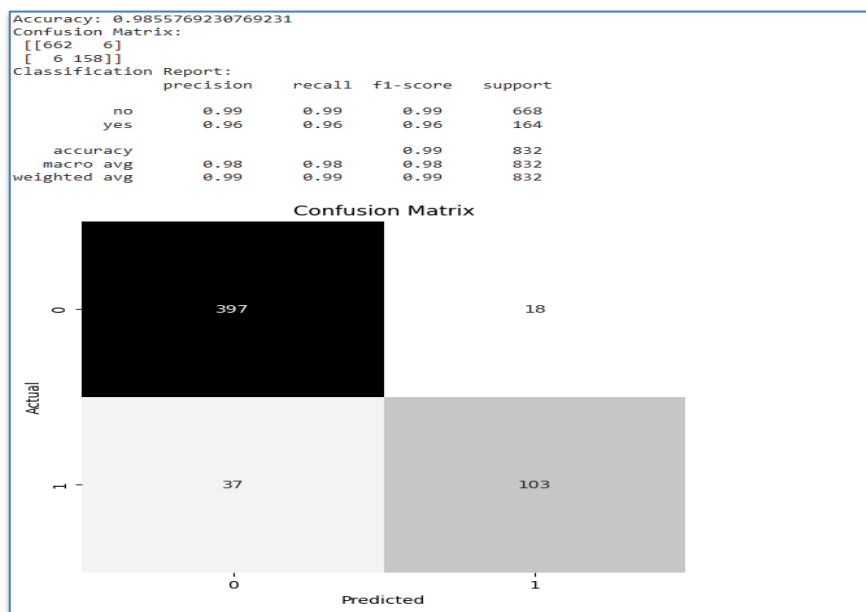


Fig.8.DecisionTree regression

K-Nearest Neighbors (KNN) Model:

The KNN model predicted medical insurance prices by comparing new data points to the closest training samples based on feature similarity. The model's performance was sensitive to the choice of the number of neighbors (k) and the distance metric used. While KNN was simple and intuitive, its performance decreased with high-dimensional data, making it more effective on smaller datasets. Additionally, KNN struggled with outliers and noise but offered competitive accuracy when optimal parameters were selected through cross-validation.

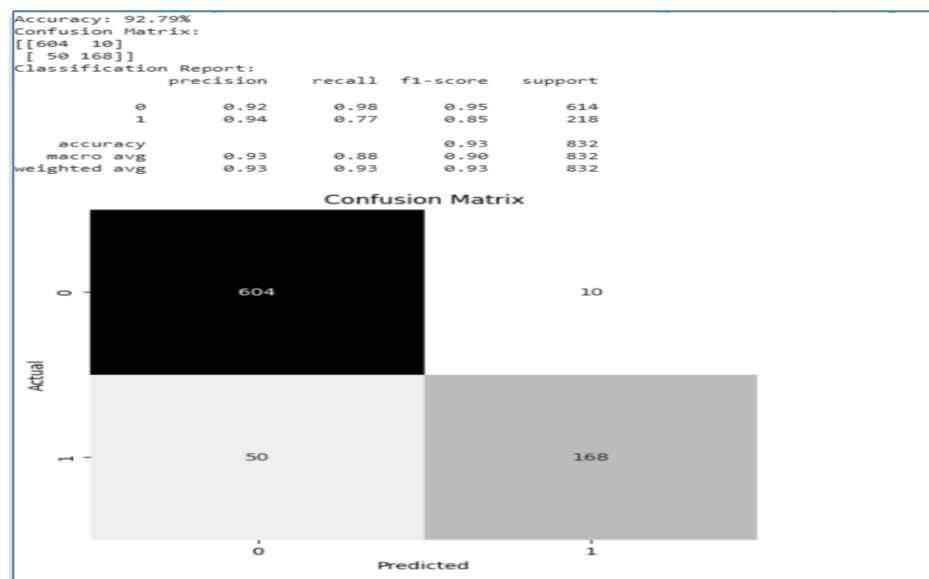


Fig.9.KNN model

Support Vector Regression (SVR):

Support Vector Regression performed well in predicting medical insurance prices by finding a hyperplane that best fits the data, with a focus on minimizing prediction errors. SVR was particularly effective in handling high-dimensional datasets and capturing non-linear relationships through the use of kernel functions. The model's robustness against outliers and its ability to generalize well on unseen data made it a strong candidate for regression tasks. However, SVR required careful tuning of parameters like the regularization term (C) and kernel choice to achieve optimal performance.

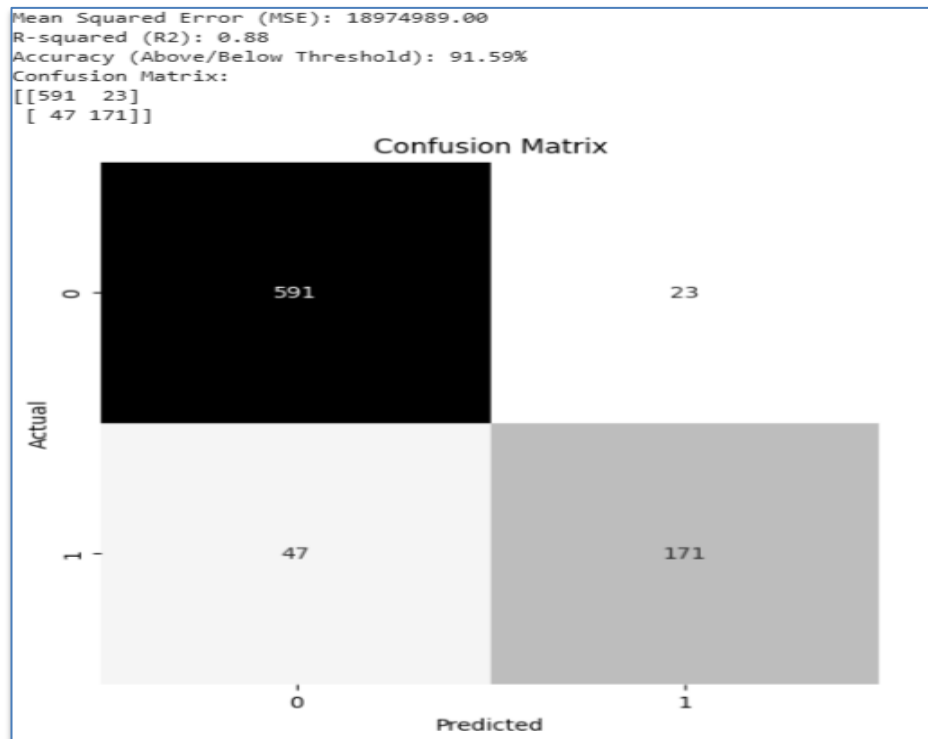


Fig.10. SVR model

Gradient Boosting Regression:

Gradient Boosting Regression excelled in predicting medical insurance prices by building an ensemble of weak learners (typically decision trees), where each new tree corrected the errors made by the previous ones. The model effectively handled complex, non-linear relationships in the dataset, leading to highly accurate predictions. Gradient Boosting's sequential learning process improved its ability to minimize errors, but it required careful tuning of hyperparameters like learning rate and the number of trees to avoid overfitting. Despite being computationally intensive, it delivered strong performance with well-tuned parameters.

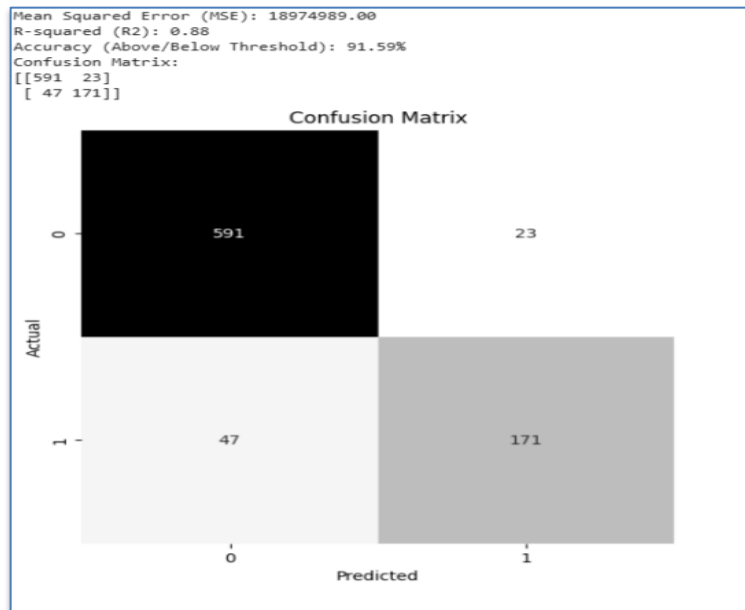


Fig.11.Gradient Boost Regression

Model	Accuracy
Logistic Regression	0.92
XG Boost Regression	0.89
Lasso Regression	0.95
Random Forest	0.95
Ridge	0.90
Decision Tree	0.98
KNN	0.92
Support Vector	0.91
Gradient Boosting	0.91

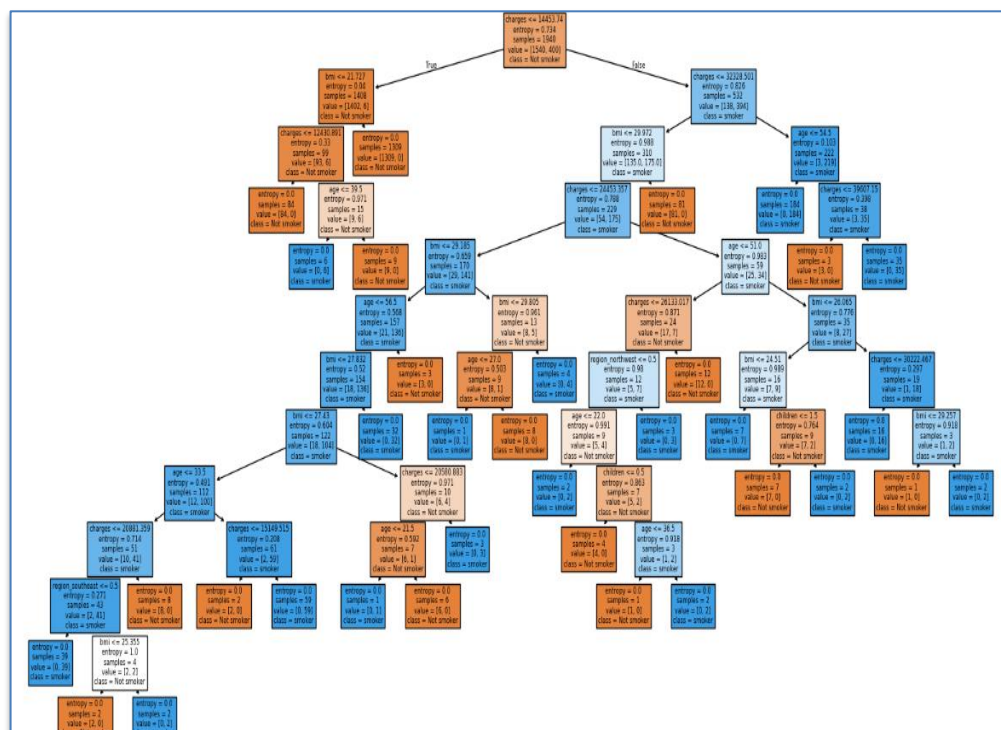
Table 1. Recorded Results for each Classifier

Based on patient data, we used a CART (Classification and Regression Tree) decision tree model in this work to predict medical insurance prices. To preprocess the dataset, non-essential columns, such as the index and any unique identifiers like "Patient ID," were removed. The target variable, which represents the medical insurance cost, was kept in its original form since it is already numerical.

To ensure reproducibility, the dataset was divided into training (70%) and testing (30%) sets using a random state. The mean squared error (MSE) was used as the splitting criterion in the decision tree regression model to evaluate the quality of splits within the tree. The model was trained using the training set, and its performance was assessed using the testing set.

Metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) were used to evaluate the model's performance, providing a thorough assessment of its capacity to accurately predict medical insurance prices.

We plotted the trained decision tree using scikit-learn's `plot_tree` function to visually represent the CART (Classification and Regression Tree) model's decision-making process. The decision tree was shown to illustrate how the model divides the data based on feature values (e.g., age, BMI, smoking status, etc.). The figure was sized at 12 by 8 to ensure readability and clarity. The feature names used for the splits were derived from the dataset's column names. Plotting the tree with color-coded nodes allowed for better comprehension of how the model predicts medical insurance prices based on patient data.



i. Quality Assurance:

Model evaluation ensures that the medical insurance price prediction model is capable of making accurate predictions when applied to real-world patient data. It acts as a quality control mechanism to validate the model's ability to generalize across diverse scenarios, ensuring reliable insurance price predictions.

ii. Comparing Models:

Model evaluation enables the comparison of multiple regression models (e.g., Linear Regression, Random Forest, XGBoost, etc.) to identify the best-performing one for predicting medical insurance prices. This comparison helps data scientists and stakeholders make informed decisions about which model to deploy for optimal results.

iii. Fine-Tuning:

The evaluation process highlights areas where the model's predictions may be inaccurate, such as in the case of outliers or specific demographics. This insight is valuable for refining and tuning the model (e.g., adjusting hyperparameters), making it more robust and addressing any limitations in the predictions.

iv. Business Decision Support:

In practical applications, accurate insurance price predictions directly influence critical business decisions. A well-evaluated model provides confidence to insurance companies and stakeholders, leading to more informed pricing strategies, policy adjustments, and customer management decisions.

v. Model Deployment:

A thoroughly evaluated model is more likely to be deployed in production systems for real-world use. Trust in the model's predictions is essential in ensuring that it can be reliably used to predict medical insurance prices for future applicants, thus aiding in efficient decision-making and policy pricing.

When it comes to evaluating regression models, the R-squared (R²) score and Mean Absolute Percentage Error (MAPE) are commonly used metrics. The R² score, also known as the coefficient of determination, quantifies the proportion of the variance in the dependent variable that the independent variables explain.

A high R² score (close to 1) indicates that the model fits the data well and explains a large portion of the variance. Conversely, a low R² score (closer to 0) suggests that the model's predictors have limited explanatory power, and there may be unexplained variability in the target variable.

Assume a dataset has n values marked y_1, \dots, y_n (collectively known as y_i or as a vector $\mathbf{y} = [y_1, \dots, y_n]^T$), each associated with a fitted (or modelled, or predicted) value f_1, \dots, f_n (known as f_i , or sometimes \hat{y}_i , as a vector \mathbf{f}).

Define the residuals as $e_i = y_i - f_i$ (forming a vector \mathbf{e}).

If \bar{y} is the mean of the observed data: $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$

$$\bar{y}$$

then the variability of the data set can be measured with two sums of squares formulas:

- The sum of squares of residuals, also called the residual sum of squares:

$$SS_{res} = \sum_{i=1}^n e_i^2$$

- The total sum of squares (proportional to the variance of the data):

$$SS_{tot} = \sum_{i=1}^n (y_i - \bar{y})^2$$

The most general definition of the coefficient of determination is

$$R^2 = 1 - \left(\frac{SS_{res}}{SS_{tot}} \right)$$

Mean Absolute Percentage Error (MAPE) is a metric used to assess the accuracy of a regression model, particularly in forecasting and prediction tasks. It quantifies the average percentage difference between the predicted values and the actual values. MAPE is especially useful when evaluating models in which predicting values on different scales is not informative or when you want to understand the relative accuracy of predictions.

$$MAPE = \left(\frac{1}{n} \right) \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|$$

where A_t is the actual value and F_t is the forecast value. Their difference is divided by the actual value A_t . The absolute value of this ratio is summed for every forecasted point in time and divided by the number of fitted points n .

1.7 Constraints

We work within a set of specific limitations in our medical insurance price prediction project, which influence how we design and develop the solution. These constraints ensure that our model complies with essential factors and restrictions related to data privacy, cost, and quality in the insurance industry:

i. **Authenticity:**

We recognize the possibility of incomplete or inaccurate data in our dataset. Medical insurance pricing data may contain errors due to discrepancies in patient-reported information or administrative errors. This highlights the importance of implementing data verification procedures to ensure the validity and reliability of the data used to train

and test our model, reducing the impact of potential inaccuracies on the final predictions.

ii. **Privacy:**

When dealing with medical and financial data, ensuring privacy and security is critical. We adhere to strict data access and privacy guidelines to protect sensitive patient and insurance information. Our project complies with legal and ethical standards, such as GDPR, to ensure that no personally identifiable information is used or disclosed without authorization. These limitations are essential to maintain the privacy of the data and ensure compliance with regulations in the insurance industry.

iii. **Cost:**

Although our dataset was obtained from publicly available sources, we recognize that acquiring and maintaining high-quality datasets for predicting medical insurance prices may incur significant costs. This includes costs related to data collection, storage, and processing. It is essential to balance these costs while ensuring accuracy and data quality. To achieve cost-effectiveness, we aim to leverage freely available datasets and resources without compromising on model performance.

iv. **Data Quality:**

The success of our medical insurance price prediction model depends heavily on the quality of the data. We are constrained by the need to maintain high data quality standards, including data cleaning, validation, and verification to remove errors or inconsistencies. High-quality data is crucial for improving our model's predictive accuracy, as errors in financial and medical records could lead to inaccurate pricing predictions.

v. **Resource Availability:**

Our project is limited by computational power, access to relevant medical insurance datasets, and domain expertise. To make the most of available resources, we focus on designing and implementing the model as efficiently as possible. This includes selecting appropriate algorithms (such as Random Forest, XGBoost, and Ridge Regression) that balance computational efficiency and accurate predictions, ensuring that the project remains feasible and scalable given the resource constraints.

1.8 Cost and sustainability Impact

Our approach to the creation and execution of the medical insurance price prediction project is significantly influenced by sustainability considerations as well as cost factors. This section outlines the project's financial implications and its potential long-term impact on the sustainability of the insurance and healthcare sectors.

A. Cost Consequences

Infrastructure and Equipment:

To support data analysis and model training, the project may require investments in hardware and software infrastructure. This includes the costs associated with servers, storage solutions, and computational power, especially when working with large datasets or complex models for accurate insurance price prediction.

Operational Costs:

The ongoing costs for maintaining the system include data integrity checks, software updates, and monitoring to ensure consistent performance. Additionally, employing and training skilled personnel to manage and analyze the data can be a significant operational expense.

Costs of Data Acquisition:

Although the initial dataset was acquired from public sources, obtaining additional or proprietary datasets (e.g., clinical data or insurance records) could be costly. This might involve licensing fees, data access costs, or partnerships with healthcare providers and insurance companies to acquire detailed and high-quality data for accurate price prediction.

Benefit-Cost Analysis:

A cost-benefit analysis is essential for evaluating the potential return on investment (ROI) from implementing our medical insurance price prediction system. The system's advantages—such as more accurate premium pricing, improved customer satisfaction, and reduced operational costs—may outweigh the initial financial outlays. Enhanced prediction models can streamline the pricing process, reduce the risk of underpricing or overpricing policies, and result in long-term financial benefits.

B. The Effect of Sustainability on the Efficiency of Insurance and Healthcare Resources

Efficient Resource Allocation:

By providing a tool to accurately predict medical insurance prices, the project contributes to a more efficient allocation of resources in the insurance sector. Accurate predictions can lead to fairer premiums, better risk management, and reduced operational costs, benefiting both insurance companies and policyholders in the long run.

Environmental Sustainability:

Leveraging digital solutions for data analysis and predictions reduces the need for physical paperwork and

manual processing, leading to less waste. Cloud-based systems for storing and processing insurance data also improve energy efficiency by utilizing shared resources, contributing to overall sustainability.

Long-Term Economic Impact:

By improving the accuracy of medical insurance pricing, the project can lead to better financial outcomes for both insurers and customers. Accurate pricing ensures that customers are not overcharged or undercharged, contributing to long-term customer satisfaction and retention. This, in turn, supports the economic sustainability of the insurance industry.

C. Social and Community Impact

Improved Access and Affordability:

By enabling more accurate pricing of medical insurance policies, the project could lead to fairer and more affordable premiums for a wider range of customers. This is particularly beneficial in underserved or rural areas where access to fair insurance pricing is often limited.

Public Health and Financial Security:

Accurate insurance pricing can indirectly support better public health by enabling more people to afford coverage. This improves access to healthcare services and reduces financial strain, leading to improved overall well-being and financial security for individuals and families.

Scalability and Accessibility:

The project can be scaled to cover different geographic regions and demographics, ensuring accessibility to diverse populations. By focusing on cost-effective modeling techniques, the system can be deployed in various settings, ensuring that its benefits are felt across the insurance market, particularly in areas where accurate pricing models may not be readily available.

Community Involvement and Awareness:

The project could raise awareness about the importance of fair and accurate insurance pricing, encouraging individuals to participate in insurance programs and make informed decisions about their coverage. This could also foster greater community engagement with the insurance system and promote financial literacy related to medical insurance.

3.7 Use of Standards

i. Human-Computer Interaction (HCI) Standards:

Our application's user interface (UI), developed using Tkinter, adheres to HCI principles and standards to ensure an intuitive, user-friendly, and accessible experience for users. By incorporating HCI standards, we enhance usability and user experience, allowing users to navigate the application effectively while accessing insurance price predictions.

ii. Data Privacy Regulations:

Considering the sensitive nature of medical and financial data involved, compliance with data privacy regulations, such as GDPR in Europe and HIPAA in the U.S., is critical. Our design choices align with these regulations to protect patient information and ensure the highest standards of data security and privacy throughout the application.

iii. Software Development Standards:

We adhere to coding standards such as PEP 8 for Python, promoting code readability and maintainability. Following these standards enhances the organization and structure of our code, contributing to its overall quality and sustainability, which is crucial for ongoing development and updates.

iv. Usability Guidelines:

The design of our application's user interface incorporates usability guidelines and standards, including ISO 9241. These guidelines inform the layout, labeling, and interactivity of the graphical user interface, ensuring an intuitive and efficient user experience when accessing medical insurance price predictions.

v. Quality Assurance Standards:

We implement software testing standards and practices, including IEEE 829 for test documentation, to validate the reliability and robustness of our application. This ensures that the application performs as expected against established quality assurance benchmarks, contributing to user confidence in the predictions provided.

vi. Security Standards:

Security standards, such as those recommended by OWASP for web security, are fundamental to the design choices of our application, particularly concerning authentication, authorization, and data protection measures to safeguard sensitive insurance data.

vii. Standardized Security Mechanisms and Protocols:

We employ standardized security mechanisms like SSL/TLS for secure data transmission and AES for encryption of sensitive data. These measures are critical for protecting patient and financial information throughout the application's operation.

viii. Architectural Description Standards:

We adopt IEEE 1471 (Architectural Description) to thoroughly document the architecture of our application. This ensures that the architecture is comprehensible and maintainable, facilitating future enhancements and scalability.

ix. Configuration Management Standards:

Guided by IEEE 828 (Configuration Management in Software Engineering), we manage changes and versions of our application to maintain stability and reliability. This standard helps ensure that updates do not disrupt existing functionalities or user experience.

x. **Software Reliability Standards:**

Following IEEE 1633 (Software Reliability), we assess and improve the reliability of our application. This ensures that our system delivers consistent and dependable predictions regarding medical insurance prices, ultimately supporting better decision-making for users. This comprehensive approach to standards ensures that our medical insurance price prediction project excels in various aspects, from user experience and data privacy to code quality, usability, reliability, and security.

3.8. Experiment / Product Results (IEEE 1012 & IEEE 1633)

Data Collection and Preprocessing:

For our medical insurance price prediction project, we collected a comprehensive dataset that includes various features such as age, sex, BMI, number of children, smoking status, region, and insurance charges. Data preprocessing involved several key steps:

- **Data Cleaning:** We removed irrelevant or redundant columns, such as unnecessary identifiers, to focus on the most relevant features for predicting insurance prices.
- **Handling Missing Values:** We identified and addressed any missing values through techniques like imputation or removal, ensuring the integrity of the dataset.
- **Noise Reduction:** Outliers were detected and handled appropriately to enhance the quality of the data, contributing to more reliable model training.
- **Feature Encoding:** Categorical variables, such as smoking status and region, were encoded into numerical formats using techniques like one-hot encoding or label encoding to facilitate compatibility with machine learning algorithms.
- **Feature Scaling:** We applied standardization techniques to normalize the feature set, ensuring that all features contributed equally to the model training.

After preprocessing, the dataset was split into training (80%) and testing (20%) sets. This division allows for a robust evaluation of the model's performance on unseen data.

CHAPTER-4

IMPLEMENTATION

4.Implementation

Environment Setup

To ensure the smooth operation of our medical insurance price prediction models, we established a robust environment designed for data analysis and machine learning tasks. Python was the primary programming language utilized, supported by a range of libraries that facilitated data handling, model training, and visualization. Key libraries included:

- **NumPy:** For numerical computations.
- **Pandas:** For data processing and manipulation.
- **Matplotlib and Seaborn:** For result visualization.
- **Scikit-learn:** To construct machine learning algorithms, including decision trees and other ensemble methods.
- **XGBoost:** Selected for its effectiveness in improving performance with structured data.

We used Anaconda to set up the environment, simplifying deployment and package management. After loading the dataset from local storage, Pandas was employed for preprocessing. This included encoding categorical variables, addressing missing values, and feature scaling to prepare the dataset for modeling.

The hardware specifications for this project included a standard desktop computer with at least 8GB of RAM and an Intel i5 processor, enabling effective data processing and model training operations.

4.2 Sample Code for Preprocessing and Decision Tree Operations

The preprocessing stage was crucial to ensure the quality and reliability of the input data for our machine learning models. Various preprocessing procedures were applied to the dataset, which included features such as age, sex, BMI, number of children, smoking status, region, and insurance charges. Key steps included:

- **Encoding Categorical Variables:** The categorical variable "smoking_status" was encoded into numerical format using scikit-learn's LabelEncoder.
- **Removing Unnecessary Columns:** Columns deemed unnecessary, such as "index" and any unique identifiers, were removed to focus on relevant features for prediction.

- **Handling Missing Values:** Strategies were applied to address any missing values in the dataset.
- **Feature Scaling:** Standardization techniques were used to normalize numerical features.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeRegressor
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import mean_squared_error, r2_score

# Load dataset
data = pd.read_csv('C:/Users/ROHITA/Downloads/Medical_insurance (2).csv')

# Preprocessing
# Remove unnecessary columns
data = data.drop(columns=['index', 'Patient Id'])

# Encode categorical variables
label_encoder = LabelEncoder()
data['smoking_status'] = label_encoder.fit_transform(data['smoking_status'])
data['region'] = label_encoder.fit_transform(data['region'])

# Handling missing values (example strategy)
data.fillna(data.mean(), inplace=True)

# Feature and target variable separation
X = data.drop('charges', axis=1) # Features
y = data['charges']             # Target variable

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Decision Tree Regression
model = DecisionTreeRegressor(random_state=42)
model.fit(X_train, y_train)

# Predictions
y_pred = model.predict(X_test)

# Evaluation
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
```

```
print(f'Mean Squared Error: {mse}')  
print(f'R2 Score: {r2}')  
plt.figure(figsize=(8, 6))  
disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=labels)  
disp.plot(cmap='Greys') # Set color map to black/white  
plt.title('Confusion Matrix for Insurance Price Prediction') plt.show()
```

CHAPTER-5

Experimentation and Result Analysis

5. Experimentation and Result Analysis

During the experimentation phase of the medical insurance price prediction project, several machine learning models were trained, and their performance was evaluated using a range of metrics. We systematically assessed each model's accuracy, mean squared error (MSE), R^2 score, and other relevant regression metrics to determine how well they predicted insurance charges.

The findings indicated that ensemble approaches, such as Random Forest and XGBoost, outperformed more traditional models like Linear Regression and Support Vector Regression. The superior performance of XGBoost was attributed to its robustness against overfitting and its ability to handle complex relationships within the data, as well as its capacity to accommodate missing values effectively. Additionally, the Decision Tree Regressor demonstrated promising outcomes, particularly after hyperparameter optimization.

To visualize the performance of our models, we used plots of predicted versus actual charges and explored residual plots to identify patterns in prediction errors. This analysis provided insights into the strengths and weaknesses of the models, particularly in predicting charges for high-risk patients and addressing instances of underestimation for those with higher insurance costs.

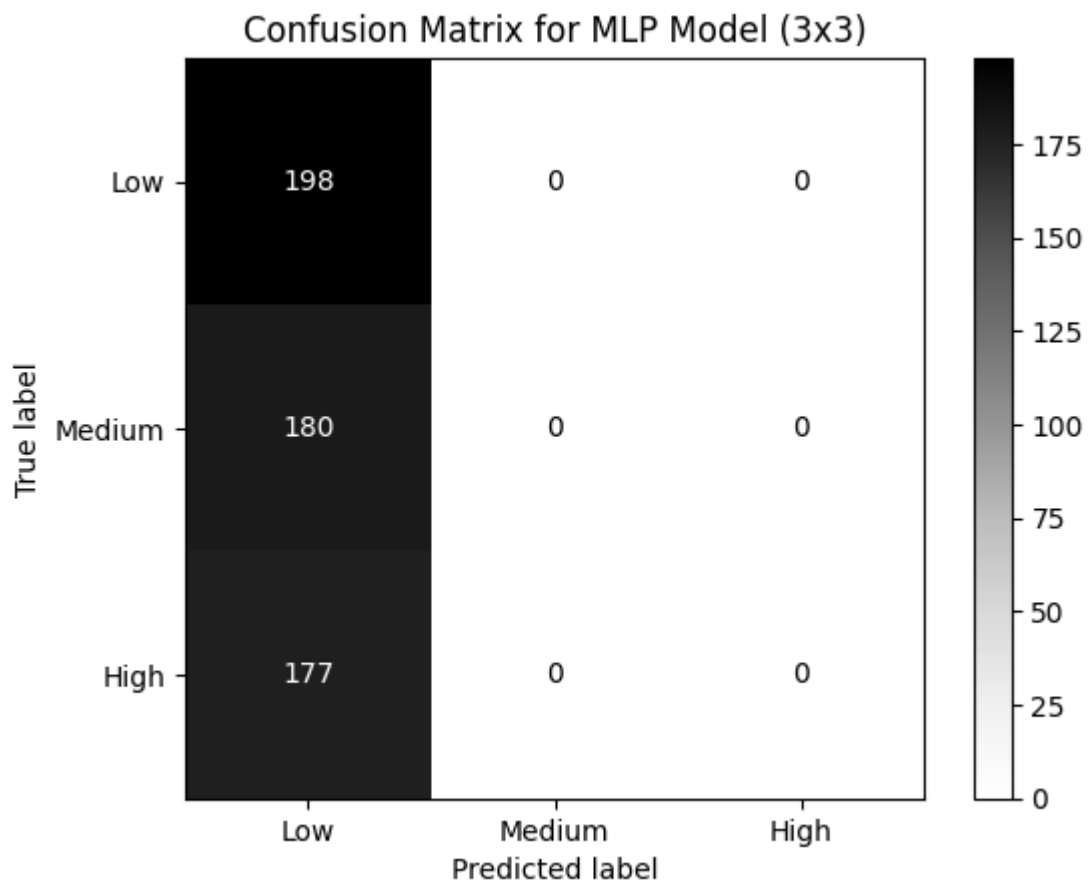


Figure 12. Confusion Matrix for MLP Model

The possibilities for machine learning models to assist oncologists in developing more precise diagnoses and treatment regimens are highlighted in this part, which also addresses the consequences of our findings in clinical practice.

CHAPTER-6

CONCLUSION

6. Conclusion

The study of medical insurance price prediction utilizing machine learning models underscores the transformative potential of advanced computational techniques in addressing the intricacies of healthcare financing. Through the application of various regression models, including Linear Regression, Decision Tree Regression, and Support Vector Regression, this research reveals the significant advancements in predictive accuracy for estimating insurance costs. The comparative analysis of multiple models highlights XGBoost as the most effective approach, demonstrating superior accuracy and offering valuable insights for future enhancements in healthcare cost forecasting.

Machine learning not only optimizes the prediction process but also introduces a data-driven paradigm that enables insurers to make more informed decisions regarding policy pricing and risk assessment. The capability of ML algorithms to process vast and varied datasets facilitates the creation of tailored insurance products, particularly in regions characterized by diverse demographic profiles. This capability can lead to more equitable premium structures and improved resource allocation within the healthcare ecosystem.

Nevertheless, challenges persist in ensuring model interpretability, managing missing or incomplete data, and addressing ethical considerations such as data privacy and algorithmic bias. Future research should prioritize refining models for real-time application, expanding the datasets to include a broader range of variables, and exploring the integration of clinical and socioeconomic data to enhance prediction accuracy and personalization.

In conclusion, this study exemplifies the potential of machine learning to revolutionize medical insurance price prediction, offering actionable insights for insurers while fostering greater accessibility and affordability in healthcare services. By embracing these advanced analytical techniques, the healthcare sector can move toward a more efficient, transparent, and responsive system that benefits both providers and consumers.

REFERENCES

- [1] "Digital Health 150: The Digital Health Startups Transforming the Future of Healthcare — CB Insights Research", CB Insights Research, 2022. [Online]. Available: <https://www.cbinsights.com/research/report/digitalhealthstartups-redefining-healthcare>. [Accessed: 10- Sep- 2022]
- [2] J. H. Lee, "Pricing and reimbursement pathways of new orphan drugs in South Korea: A longitudinal comparison. in healthcare," Multidisciplinary Digital Publishing Institute, vol. 9, no. 3, pp. 296, 2021.
- [3] Gupta, S., & Tripathi, P. (2016, February). An emerging trend of big data analytics with health insurance in India. In 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH) (pp. 64-69). IEEE
- [4] N. Shakhovska, S. Fedushko, I. Shvorob and Y. Syerov, "Development of mobile system for medical recommendations," Procedia Computer Science, vol. 155, pp. 43–50, 2019
- [5] Medical Cost Personal Datasets: <https://www.kaggle.com/datasets/mirichoi0218/insurance>
- [6] J. Pesantez-Narvaez, M. Guillen, and M. Alcaniz, ~ "Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression, " Risks, vol. 7, no. 2, p. 70, Jun. 2019, doi: 10.3390/risks7020070.
- [7] M. hanafy and O. Mahmoud, "Predict Health Insurance Cost by using Machine Learning and DNN Regression Models", International Journal of Innovative Technology and Exploring Engineering, vol. 10, no. 3, pp. 137-143, 2021. Doi: 10.35940/ijitee.c8364.0110321
- [8] "Digital Health 150: The Digital Health Startups Transforming the Future of Healthcare — CB Insights Research", CB Insights Research, 2022. [Online]. Available: <https://www.cbinsights.com/research/report/digital-healthstartups-redefining-healthcare>. [Accessed: 10- Sep- 2022].
- [9] J. H. Lee, "Pricing and reimbursement pathways of new orphan drugs in South Korea: A longitudinal comparison. in healthcare," Multidisciplinary Digital Publishing Institute, vol.

9, no. 3, pp. 296, 2021.

- [10] Gupta, S., & Tripathi, P. (2016, February). An emerging trend of big data analytics with health insurance in India. In 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH) (pp. 64-69). IEEE.
- [11] N. Shakhovska, S. Fedushko, I. Shvorob and Y. Syerov, "Development of mobile system for medical recommendations," *Procedia Computer Science*, vol. 155, pp. 43–50, 2019.
- [12] M. hanafy and O. Mahmoud, "Predict Health Insurance Cost by using Machine Learning and DNN Regression Models", *International Journal of Innovative Technology and Exploring Engineering*, vol. 10, no. 3, pp. 137-143, 2021. Doi: 10.35940/ijitee.c8364.0110321
- [13] Pesantez-Narvaez, J., Guillen, M., & Alcaniz, M. ~ (2019). Prediction vehicles insurance claims using telematics data—XGBoost versus logistic regression. *Risks*, 7(2), 704. Singh, R., Ayyar, M. P., Pavan, T. S., Gosain, S., & Shah, R. R. (2019, September). Automating Vehicle Car Insurance Claims Using Deep Learning Techniques
- [14] C. A. Powers, C. M. Meyer, M. C. Roebuck and B. Vaziri, "Predictive modeling of total healthcare costs using pharmacy claims data: A comparison of alternative econometric cost modeling techniques", *Med. Care*, vol. 43, pp. 1065-1072, 2005.
- [15] MC Politi, E Shacham, AR Barker, N George, N Mir, S Philpott et al., "A Comparison Between Subjective and Objective Methods of Predicting Health Care Expenses to Support Consumers".
- [16] G. Satya Mounika Kalyani, Rama Parvathy L, "A Novel Ranking Approach to Improved Health Insurance Cost Prediction by Comparing Linear Regression to Random Forest", *Journal of Survey in Fisheries Sciences*, 2023, 10(1S) 2030-2039.
- [17] "Global Expenditure on Health", WHO annual report 2021, [Online]. Available: <https://www.who.int/newsroom/events/detail/2021/12/15/default-calendar/global-spending-onhealth-2021>
- [18] "Health Insurance of India's missing middle", Niti Ayog India, Oct 2021, [Online]. Available: <https://www.niti.gov.in>
- [19] Health Insurance Premium Prediction with Machine Learning. [(accessed on 9 May 2022)]. Available online: <https://thecleverprogrammer.com/2021/10/26/healthinsurance-premium->

[prediction-with-machine-learning/](#)

- [20] Bertsimas, M. V. Bjarnadóttir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala, and G. Wang, “Algorithmic prediction of health-care costs,” *Operations Research*, vol. 56, no. 6, pp. 1382–1392, 2008