IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*
Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

Medical Insurance Suggestion Based On Age

NALABOLU PRAVALLIKA, AVAGADDA SANYASI NAIDU NITIN, NEDURI VISHNU PAVAN KUMAR, AYNAMPUDI CHAITANYA VARMA, KOMMINENI JEEVAN SURYA.
Dr. NITALAKSHESWARA RAO KOLUKULA
Department of Computer Science and Engineering, School of Technology, GITAM University, Visakhapatnam, Andhra Pradesh, India, 530045.

ABSTRACT

This project focuses on solving the complex issue of medical insurance cost estimation by leveraging personal demo graphic and health-related factors. The rising cost of health care has led to increased interest in accurate and personalized insurance pricing models, which can provide valuable insights to individuals when choosing the right insurance plan. This project uses adata-set containing key attributes such as age, sex, body mass index (BMI), smoking status, region of residence, number of children, and actual medical charges.

To predict insurance charges effectively, we employed various machine learning models, including Linear Regression, Ridge Regression, Lasso Regression, Decision Tree, Random Forest Regression, and XG Boost. The purpose of using multiple models was to compare their predictive performance and identify the most suitable model for this task. Each model was evaluated using key metrics like the R-squared score and Mean Squared Error (MSE) to determine its accuracy and reliability in predicting the cost of medical insurance.

Among the models tested, the Random Forest Regression, after undergoing hyper parameter tuning, outperformed the other models in terms of both prediction accuracy and error minimization. This model's ability to handle complex, non-linear relationships between the features and target variable (insurance charges) made it the ideal choice for our project. Hyper parameter tuning further enhanced its performance by optimizing key parameters, such as the number of trees in the forest and the maximum depth of each tree.

By providing individuals with accurate predictions of their insurance costs based on their personal

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

data, this model can help people make more informed and cost-effective insurance decisions. This approach not only benefits users but also provides insurance companies with a tool to offer better-customized pricing, contributing to more efficient and customer-friendly health insurance systems.

INTRODUCTION

Medical insurance plays a crucial role in ensuring that individuals are financially protected against high healthcare costs. As the healthcare landscape continues to evolve, the need for personalized insurance plans has become more pronounced. Various factors such as age, gender, lifestyle, and pre-existing conditions influence insurance premiums, making it difficult for individuals to predict theirinsurancecostsaccurately.Thiscomplexityiscompoundedbythefactthatdifferent insurance providers employ varying methods to calculate premiums. In this context, predicting medical insurance costs using data-driven approaches can significantly simplify decision-making for consumers and help insurance companies provide more tailored policies.

The rapid advancement in machine learning and data analytics has opened new avenues for solving complex, real-world problems, including cost estimation in the insurance sector. Traditionally, insurance premiums were calculated using static, rule-based methods that often failed to account for individual variability. These methods, while effective in aggregating large pools of insured individuals, could not offer personalized insurance premium predictions. Consequently, many individuals faced uncertainty when purchasing health insurance, often resulting in overpayment or insufficient coverage. Thus, a more dynamic and individualized approach is required, where personalized data such as age, body mass index (BMI), smoking status, region, and the number of dependents are considered to provide accurate insurance cost estimates.

This project aims to address this gap by developing a predictive model for estimating medical insurance charges based on personal information. Using machine learning techniques, we can leverage historical data to create a robust system capable of predicting insurance premiums with high accuracy. The dataset used in this project includes various features that are known to impact

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

insurance charges, including age, sex, BMI, smoker status, geographical region, number of children, and the total charges incurred by previous insurance claims. By incorporating these features into a predictive model, we can help individuals estimate their insurance premiums more effectively, offering them a clearer understanding of the factors influencing their costs.

Justifying the title of this project, "Medical Insurance Suggestion Based on Age," age is one of the most significant predictors of healthcare costs and insurance premiums. As individuals age, their likelihood of requiring medical care increases, which in turn raises the cost of their insurance. However, age alone does not provide a complete picture. Factors like BMI and smoking status

Significantly influence how insurance providers as sess risk. For instance, a younger individual with a high BMI or smoking habit may have higher insurance costs than an older non-smoker with a healthy weight. Therefore, while age is a critical factor, the inclusion of other lifestyle and demographic variables ensures a more comprehensive and accurate prediction model.

Existing systems for insurance cost estimation of ten rely on static formulas or simplistic models that fail to capture the nuances of an individual's health profile. These systems usually generalize premiums based on broad categories, leading to mispricing for many consumers. For example, insurance companies might classify individuals based on wide age ranges (e.g.,20-30,30-40),which does not account for the differences within these groups. Two individuals aged 30 and 40, for instance, can have vastly different health profiles and insurance needs. This lack of granularity leads to premiums that may not accurately reflect the true risk posed by an individual, creating an opportunity for more advanced systems to step in.

Our project uses a dataset comprising over 3,600 records with attributes related to personal health and insurance costs. We explore various machine learning algorithms such as Linear Regression, Ridge Regression, Lasso Regression, Decision Tree, Random Forest, and XG Boost to build a predictive model. Each of these models is tested for its performance, and we employ hyper parameter tuning to enhance their accuracy. Among these, Random Forest Regressor emerges as the best performer in terms of prediction accuracy. This model captures then on-linear

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

relationships between the various features and the insurance charges, offering a more reliable and tailored prediction

By leveraging data-driven insights, insurers can offer tailored premiums that reflect an individual's risk profile more accurately. This approach promotes fairness in pricing and can improve customer satisfaction, as consumers are more likely to feel that they are paying for what they truly need.

In summary, this project addresses the growing need for accurate, personalized medical insurance cost estimation. Through the use of machine learning techniques, we provide a system that offers users an estimate of their insurance charges based on a variety of personal factors, with age being a central component. The title "Medical Insurance Suggestion Based on Age" is justified by the fact that age significantly influences insurance costs, but we enhance the model by incorporating other important health-related features. This ensures that our system can offer more accurate and individualized insurance suggestions, ultimately aiding individuals in making better financial decisions when it comes to health coverage.

LITERATURE REVIEW

In recent years, particularly following the COVID-19 pandemic, the demand for accurate and personalized health insurance cost predictions has increased significantly. The surge in health-related uncertainties has led to an increased interest in the healthcare insurance sector, both from individuals seeking better financial planning for health coverage and insurers striving to offer competitive premiums. Predicting health insurance premiums based on various demographic and health-related factors is now a vital task, enabling individuals and policy makers to make data-driven decisions. As a result, many researchers have explored different machine learning techniques for predicting insurance costs based on factors such as age, gender, body mass index (BMI), smoking status, and the number of children.

This review summarizes the findings from three key papers that address health insurance cost prediction using machine learning techniques. Each of these papers utilizes distinct datasets and

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

methods to demonstrate the accuracy and effectiveness of different machine learning models in forecasting insurance premiums. The papers reviewed are:

1. "Health Insurance Cost Prediction using Machine Learning IEEE"(2022)
2. "Health Insurance Cost Prediction Using Machine Learning IRJET"(2022)
3. "Health Insurance Cost Prediction Using Machine Learning ICICC"(2022)

Paper1:"Health Insurance Cost Prediction Using Machine Learning IEEE"(2022)

This paper presents a machine learning-based system designed to predict health insurance costs using demographic and health-related features. The authors used the USA's medical cost personal dataset sourced from Kaggle, whichconsistsof1,338 entries. The dataset includes features such as age, gender, BMI, smoking habits, the number of children, and the region of residence. These features are known to significantly influence insurance costs, with variables like smoking habits and BMI often resulting in higher premiums due to associated health risks.

The primary machine learning model employed in this study was linear regression, which analyzes the relationship between these features and insurance costs. Linear regression is a well-known statistical technique that models the linear relationship between input features and the dependent variable (in this case, insurance costs). The dataset was split into a 70% training set and a 30% testing set to evaluate the model's performance, resulting in a prediction accuracy of 81.3%.

The study also explored the impact of individual features on insurance premiums. For instance, smoking status and age were identified as critical determinants of higher insurance costs. Smokers, as well as older individuals, tend to have higher health risks, which translates into increased premiums. This emphasis on significant variables was particularly relevant in the post-pandemic context, where health insurance has become a pressing concern for many people. The study concluded that linear regression, while providing as o lid baseline, could benefit from more complex models to achieve even better accuracy and predictive power in health insurance cost estimation.

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

Paper2:"Health Insurance Cost Prediction Using Machine Learning IRJET"(2022)

This paper delves deeper into the prediction of health insurance premiums by applying a wider range of machine learning techniques, with a focus on improving accuracy through advanced algorithms. The authors explored various regression models, including Extreme Gradient Boosting (XG Boost) and Random Forest Regression (RFR), both of which are ensemble methods that build multiple decision trees to improve prediction performance. These models, when compared to simpler algorithms like linear regression, often provide better predictive accuracy due to their ability to capture complex, non-linear relationships in the data.

The dataset used in this study also contained several critical health-related features, such as age, BMI, diabetes status, and abnormal blood sugar levels. The inclusion of these features provided a more comprehensive view of an individual's health status, further refining the accuracy of premium predictions. XG Boost, in particular, was highlighted for its capacity to handle large amounts of data and for its robust performance in healthcare-related predictions. The study showed that by tuning hyper parameters and optimizing the models, the authors could achieve significant improvements in prediction accuracy over traditional regression methods.

This paper also introduced the idea of developing ensemble methods to further improve prediction performance. Ensemble models, like XG Boost, combine the predictions of multiple models to create

A more accurate final prediction. The authors discussed how the combination of XG Boost and RFR allowed them to reduce the error in predicting insurance premiums, ultimately leading to a model that could assist insurers in better pricing policies for individual clients.

Another significant contribution of this paper was its examination of novel ranking techniques with machine learning algorithms. By using these techniques, the researchers were able to classify individuals based on their predicted insurance costs, thereby enabling insure rs to prioritize high-risk individuals who may require higher premiums. This type of ranking helps insurers better

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

manage their resources and pricing strategies, ensuring that high-risk individuals are charged appropriately.

Paper3:"Health Insurance Cost Prediction Using Machine Learning ICICC"(2022)

The third paper under review takes a step forward in predictive modeling by deploying three ensemble machine learning models: Extreme Gradient Boosting (XG Boost), Gradient Boosting Machine (GBM), and Random Forest (RF). The authors aimed to combine variations of decision trees to create more powerful predictive models for insurance costs. The dataset used in this study, sourced from Kaggle, comprised 986 records, focusing on medical insurance costs and their related features.

One of the main highlights of this study was the use of Explainable Artificial Intelligence (XAI) techniques such as SH apley Additiveex Planations (SHAP) and Individual Conditional Expectation (ICE) plots. These methods were employed to explain the key factors influencing insurance premiums. SHAP values provide a way to interpret the impact of each feature on the predicted insurance cost, allowing stakeholders, such as insurers and customers, to understand the rationale behind each prediction. ICE plots, on the other hand, illustrate how changes in a particular feature influence the model's predictions, making them a valuable tool for exploring variable interactions.

The results of the study demonstrated that all three ensemble models achieved impressive performance in predicting insurance costs. However, XG Boost emerged as the top performer, achieving the best overall accuracy in terms of prediction quality. Despite its accuracy, XG Boost was noted to require higher computational resources compared to the other models. The Random Forest model, though slightly less accurate, consumed fewer computational resources and had a lower prediction error, making it a more efficient option for resource-constrained environments.

PROBLEM IDENTIFICATION & OBJECTIVES

Problem Identification:

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

The rising costs of healthcare have made it increasingly difficult for individuals to afford adequate medical coverage. Health insurance premiums, which are designed to mitigate these costs, have become a crucial financial tool for millions of people. However, the calculation and prediction of health insurance premiums remain complex, with numerous factors such as age, gender, lifestyle habits (e.g., smoking), pre-existing conditions, and geographical location influencing costs. This complexity makes it challenging for insurance companies to accurately assess premiums and for individuals to choose the most suitable insurance policies.

Moreover, the unpredictability of healthcare expenses and the post-pandemic surge in demand for health insurance have created an urgent need for more precise and personalized insurance cost predictions.Currenttraditionalmethodsforcalculatingpremiumsoftenfallshortinaddressingthese complexities,leadingtosuboptimalpricingstrategies.Asaresult,thereisaneedformoreadvanced predictive models that can help insurers forecast premiums more accurately while also making the process transparent for policyholders.

The primary issue revolves around the difficulty of accurately predicting health insurance costs based on multiple variables. This problem can lead to inefficiencies in premium pricing, either overcharging individuals or underestimating the financial risk for insurers. To address this challenge, machine learning models are being explored as a viable solution to provide more accurate, data-driven predictions that can handle the non-linear relationships between multiple variables.

Objectives:

The goal of this research is to leverage machine learning techniques to solve the problem of inaccurate health insurance premium predictions. To address this issue, the following objectives have been defined:

- **Objective1:Developa machine learning-based predictive model**

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

Create a robust predictive model using machine learning algorithms to forecast health insurance premiums based on individual demographic and health-related factors. The model should improve accuracy over traditional methods and account for then on-linear relationships between variables.

- **Objective 2: Identify key determinant factors** Utilize explainable AI techniques (e.g., SHAP values and ICE plots) to uncover and explain the key factors that have the greatest impact on the prediction of insurance premiums, ensuring transparency and interpretability in the predictions.

- **Objective3:Compare multiple machine learning lgorithms**

  Compare the performance of various machine learning models such as Linear Regression, Random Forest, and XG Boost. Evaluate these models using standard performance metrics like R- squared (R2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) to determine the most accurate and computationally efficient algorithm for health insurance cost prediction.

- **Objective 4: Address resource constraints** Analyze the trade-offs between model accuracy and computational efficiency. Aim to develop a model that not only provides accurate predictions but also operates within reasonable resource limits, making it accessible to insurers with varying levels of computational infrastructure.

- **Objective 5: Personalize insurance premium predictions** Develop a system that can provide personalized premium predictions for individuals based on their unique set of health and demographic features, thereby enabling insurers to offer customized policies and individuals to make informed choices.

- **Objective 6: Incorporate new data sources** Explore the possibility of incorporating additional data sources (such as genetic

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

information, social determinants of health, and lifestyle choices) to further refine and improve the predictive capabilities of the model.

- **Objective7:Test and validate the model on real-world data sets**

  Test the developed model on publicly available health insurance data sets(e.g., from Kaggle) and validate its predictions against actual insurance premium data to ensure reliability and practical applicability.

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

- **Objective 8: Provide action able insights for insurers and policy makers**

  Offer insights that can help insurers design more equitable pricing strategies and assist policymakers in regulating the health insurance market to ensure fair pricing for high-risk

At the heart of the issue lies the challenge of effectively predicting health insurance costs based on multiple, interrelated variables. Inadequate models can lead to serious consequences, including financial hardships for individuals who are unable to afford necessary coverage and losses for insurers who miscalculate their risk exposure. This dual problem creates a significant gap in the market, where both insurers and consumers require tools that enhance transparency, efficiency, and accuracy in the pricing process. The development of machine learning models presents a viable solution to this challenge, as these models are uniquely equipped to analyze large datasets and uncover patterns that traditional methods may overlook. By embracing machine learning, the healthcare industry can pave the way for a more equitable and transparent insurance pricing system that better serves the needs of all stakeholders.

By addressing these comprehensive objectives, this research aspires to revolutionize the landscape of health insurance cost predictions. The integration of advanced machine learning techniques and the incorporation of diverse data sources will not only enhance the accuracy of premium forecasts but also promote transparency and fairness in the pricing process. Ultimately, the insights derived from this study hold the potential to reshape insurance practices, empower consumers, and create a more equitable and efficient healthcare system in the post-pandemic era.

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

EXISTINGSYSTEM

The current landscape of health insurance cost prediction primarily relies on traditional actuarial methods and statistical models. These approaches have been established over many years and have formed the backbone of premium calculation processes within the insurance industry. This section discussestheexistingsystemsutilizedinpredictinghealthinsurancecosts,their methodologies, and the associated drawbacks that limit their effectiveness in the current environment.

Traditional Actuarial Methods

At the core of the existing system are traditional actuarial methods, which employ historical data to estimate future insurance costs. Actuaries analyze vast amounts of historical claims data, demographic information, and risk factors to determine premium rates. These methods often use linear regression models to create pricing algorithms, which rely on established relationships between specific variables and costs. For instance, age and health conditions are common factors that influence the determination of insurance premiums.

Drawbacks of Traditional Actuarial Methods:

1. Limited Flexibility: Traditional actuarial methods often rely on predefined assumptions and simplifications. For example, many models assume linear relationships among variables, which can lead to in accurate predictions when dealing with complex interactions in the data. This rigidity fails to capture then on-linear relationships and intricate patterns that frequently exist in real-world scenarios.

2. Inadequate Handling of Big Data: With the advent of big data, traditional actuarial methods struggle to leverage large, unstructured datasets that contain valuable information about consumer behavior, treatment outcomes, and lifestyle factors. These models are typically designed for smaller, structured datasets, limiting their ability to adapt to the complexities of modern data landscapes.

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

3. Static Nature: Actuarial models often rely on historical data that may not reflect current trends or emerging risks, particularly in a rapidly changing healthcare environment. The COVID-19 pandemic highlighted the need for more dynamic models capable of adjusting to sudden shifts in healthcare usage, policy changes, and consumer preferences.

Time-Consuming Processes: The process of building actuarial models is often labor- intensive, requiring substantial time and expertise. The need for extensive data cleaning, feature selection, and validation can delay the implementation of new pricing strategies, leaving insurers at a competitive disadvantage.

4. Opaque Pricing Mechanisms: Consumers often find traditional pricing methods confusing and opaque, as they may not understand how their premiums are calculated. This lack of transparency can erode trust in insurance companies, leading to dissatisfaction among policyholders who may feel unfairly charged.

Statistical Models

In addition to traditional actuarial approaches, various statistical models, including logistic regression and decision trees, have been employed to enhance the accuracy of health insurance cost predictions. These models in corporate multiple variables and attempt to identify patterns in the data that influence premium costs.

Drawbacks of Statistical Models:

1. Assumption of Independence: Many statistical models make the assumption that the input features are in dependent of one another. However, in the context of health insurance, factors such as age, gender, and lifestyle choices are often inter related, resulting in multi Collin earity that can skew predictions.

2. Sensitivity to Outliers: Statistical models can be highly sensitive to outliers and anomalies in the data. In healthcare datasets, outliers may arise from rare medical conditions or extraordinary claims, which can disproportionately influence model outcomes and lead to inaccurate premium predictions.

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

3. Limited Predictive Power: While statistical models can identify correlations between variables, they may lack the predictive power needed to accurately forecast future insurance costs. This limitation is particularly evident when dealing with complex and evolving healthcare data, where the relationships between variables may change over time.

Challenges in Personalization

One of the significant drawbacks of existing systems is their inability to provide personalized insurance premium predictions. Traditional models of ten use a one-size-fits-all approach, which can overlook individual differences in health status, lifestyle choices, and other demo graphic factors. As consumers increasingly seek personalized products and services, the lack of customization in premium pricing can lead to dissatisfaction and disengagement.

Regulatory and Compliance Issues

In addition to methodological limitations, existing systems also face regulatory and compliance challenges. Insurance companies must adhere to various regulations regarding pricing fairness and discrimination. Traditional methods may in advertently lead to biased pricing practices if they fail to adequately consider certain demographic groups. This can expose insurers to legal challenges and damage their reputation in the market.

Conclusion

In summary, the existing systems for predicting health insurance costs primarily rely on traditional actuarial methods and statistical models that have inherent draw backs. These methods of ten struggle to adapt to the complexities of modern health care data, leading to inefficiencies in premium pricing, lack of personalization, and diminished consumer trust. As the healthcare landscape continues to evolve, there is a pressing need for more advanced, data-driven approaches that can address these limitations and provide accurate, transparent, and personalized insurance cost predictions.

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

PROPOSEDSYSTEM

The proposed system for health insurance cost prediction leverages advanced machine learning techniques and data-driven methodologies to address the limitations of existing systems. By integrating a diverse range of data sources and employing state-of-the-art algorithms, this system aims to provide more accurate, personalized, and transparent predictions of health insurance premiums. Below are the key features and improvements of the proposed system compared to traditional methods.

1. Utilization of Machine Learning Algorithms

Advanced Predictive Modeling: The proposed system employs various machine learning algorithms, including Random Forest, Gradient Boosting ,and Neural Networks, to capture complex non-linear relationships between input features and insurance costs. Unlike traditional linear regression models, which assume a straight-line relationship, these algorithms can model intricate interactions among variables, improving prediction accuracy.

Ensemble Learning: By combining the strengths of multiple models through ensemble learning techniques, the proposed system enhances robustness and reduces the risk of over fitting. This approach allows for better generalization to unseen data, making the predictions more reliable in dynamic healthcare environments.

2. Enhanced Data Integration

Comprehensive Data Sources: The proposed system incorporates a wide variety of data sources, including not only demographic and health-related features but also additional factors such as socioeconomic status, geographic location, lifestyle choices, and even genetic information. This holistic approach allows for a more nuanced understanding of the factors influencing health insurance costs.

Real-time Data Processing: Leveraging technologies such as big data frameworks and cloud computing, the proposed system can process and analyze vast amounts of data in real time. This

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

capability enables insurers to adapt pricing strategies quickly in response to changing market conditions, consumer behaviors, and emerging health trends.

3. Explainable AI for Transparency

Interpretability of Models: One of the significant improvements of the proposed system is the incorporation of ex plain able AI techniques, such as SHAP (Shapley Additive ex Planations) values and Local Interpretable Model-agnostic Explanations (LIME). These techniques provide insights into how individual features contribute to the predictions, enhancing transparency and allowing policyholders to understand the basis for their premiums.

Building Trust with Consumers: By making the model's decision-making process transparent, the proposed system fosters trust between insurers and policyholders. Consumers can see how their demographic and health-related factors impact their premiums, which can lead to greater sat is faction and engagement.

4. Personalization of Premium Predictions

Tailored Premium Calculation: The proposed system allows for personalized insurance premium predictions based on an individual’s unique profile. By analyzing specific health conditions, lifestyle choices, and demographics, the model can provide customized premium estimates, ensuring that individuals are not overcharged or undercharged based on generalized pricing models.

Dynamic Adjustments: The model can continuously learn from new data, enabling dynamic adjustments to premium predictions as individual circumstances change. For instance, if a policyholder improves their health status or changes their lifestyle, the system can reflect these changes in their premium calculations in real time.

5. Improved Performance Metrics

Comprehensive Evaluation Framework: The proposed system employs a comprehensive

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

evaluation framework that includes performance metrics such as R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) to assess the accuracy of predictions across different machine learning models. This rigorous evaluation process ensures that the best-performing model is selected for deployment.

Benchmarking Against Existing Methods: The performance of the proposed machine learning models is benchmarked against traditional actuarial methods, providing empirical evidence of improved accuracy and efficiency. This comparison highlights the advantages of adopting advanced data-driven approaches over established practices.

6. Addressing Regulatory Compliance

Bias Mitigation Strategies: To address potential bias in premium pricing, the proposed system incorporates bias detection and mitigation strategies. By regularly auditing model predictions for fairness across demographic groups, insurers can ensure compliance with regulations while promoting equitable pricing practices.

Transparency in Reporting: The system facilitates trans parent reporting of pricing mechanisms and decision-making processes to regulatory bodies. This transparency helps build credibility for insurers and ensures adherence to legal and ethical standards in health insurance pricing.

7. Actionable Insights for Insurers and Policymakers

Data-Driven Decision Making: The proposed system provides actionable insights that can help insurers refine their pricing strategies and optimize risk management. By analyzing patterns in insurance claims and consumer behavior, insurers can make informed decisions about policy design and pricing.

Support for Policy Development: Insights generated from the predictive models can assist policymakers in understanding healthcare trends and the impact of various factors on insurance costs. This information can guide regulatory decisions, ensuring that the health insurance market

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

remains fair and accessible for all individuals.

CONCLUSION

In summary, the proposed system represents a significant advancement over existing health insurance cost prediction methods. By harnessing the power of machine learning, integrating diverse data sources, enhancing transparency through explainable AI, and personalizing premium predictions, the system addresses the limitations of traditional approaches. The result is a more accurate, fair, and responsive pricing model that benefits both consumers and insurers in an increasingly complex healthcare landscape. This innovative approach not only improves the accuracy of predictions but also builds trust among policyholders, ultimately leading to a more sustainable and equitable health insurance market.

SYSTEMARCHITECTURE

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

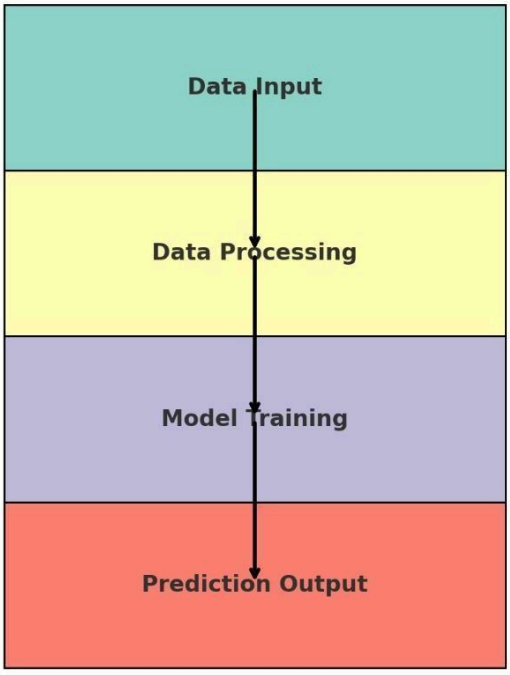
(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

Block Diagram of Health Insurance Premium Prediction System



COMPONENTS OF THE SYSTEM ARCHITECTURE

The architecture consists of several key components, each serving a specific purpose within the overall system. The primary focus is on the seven critical features of the dataset: age, sex, BMI, smoking status, region, number of children, and charges (insurance costs). These components can be categorized into data input, processing, model training, prediction output, and user interface.

1. Data Input

Layer Data

Sources:

The project utilizes a data set with 3,630 entries, encompassing various factors affecting health insurance costs. The key attributes are:

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

- Age: The age of the individual represented as a float value, which is crucial as older individuals generally incur higher medical costs.
- Sex: A categorical variable indicating the gender of the individual (male or female), which can influence insurance costs due to differing health risks.
- BMI (Body Mass Index): A continuous variable representing a key health indicator, calculated from weight and height. It serves as a proxy for obesity-related health risks.
- Smoker Status: A categorical variable indicating whether the individual's mokes (yesorno), with smoking generally leading to higher insurance premiums due to increased health risks.
- Region: The geographical location of the individual, categorized into southeast, southwest, northwest, and northeast. This affects healthcare costs due to regional variations in medical expenses.
- Children: The number of dependents covered by the insurance policy, influencing the overall premium costs.
- Charges: The total medical insurance charges, represented as afloat value, which is the target variable for prediction.

Data Acquisition:

Data is collected from a structured dataset, ensuring the integrity and consistency of the input. The dataset is sourced from public repositories, such as Kaggle, which provides reliable datasets for health insurance analysis. This data serves as the foundation for training predictive models and helps understand the impact of various features on insurance costs.

2. Data Processing

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

Layer Data

Preprocessing:

- Data Cleaning: The dataset is verified for any missing or incorrect entries. In this case, all columns are non-null, ensuring a clean dataset for analysis. If there were any anomalies, techniques such as imputation or removal of incomplete records would be applied to maintain data quality.

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

- Feature Encoding: Categorical variables (sex, smoker status, and region) are encoded into numerical formats using techniques such as one-hot en coding or label en coding. This step is essential for enabling machine learning algorithms to process non-numeric data.
- Normalization: Continuous variables like age and BMI may be normalized to a common scale (e.g.,between0 and 1) or standardized (mean= 0, standard deviation= 1)to improve model performance and training efficiency. Normalization is particularly important for algorithms sensitive to the scale of input features.

Data Transformation:

The processed data is split into training and testing sets, typically using a70-30or80-20ratio.This allows the models to learn patterns from the training data and validate their predictions on unseen testing data, thus providing an unbiased evaluation of model performance.

3. Model Training Layer

Machine Learning

Algorithms:

Variousmachinelearningmodelsareimplementedtopredicthealthinsurancechargesbasedonthe provided features. The following algorithms are utilized:

- Linear Regression: This establishes a baseline model for comparison, providing a simple method to predict charges based on a linear relationship with the input features.
- Ridge and Lasso Regression: These techniques apply regularization to linear models to prevent over fitting. Ridge regression adds a penalty equal to the square of the magnitude of coefficients, while Lasso regression can reduce some coefficients to zero, effectively performing variable selection.
- DecisionTree Regressor: This model captures non-linear relationships by splitting the data

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

into subsets based on feature values, providing interpretable results and visualizations.

- Random Forest Regressor: An ensemble method that aggregates multiple decision trees, offering improved accuracy by reducing the risk of over fitting associated with individual trees. This method also provides insights into feature importance.
- XGBoost: A highly efficient gradient boosting algorithm that optimizes performance through regularization and parallelization. It often yields better performance in competitions due to its speed and flexibility.

Model Evaluation:

Each model's performance is evaluated using metrics such as R-squared (R^2) and Mean Squared Error (MSE). These metrics help assess how well the models predict insurance charges. The R^2 score indicates the proportion of variance explained by the model, while MSE provides insight into the average squared difference between predicted and actual values. The Random Forest Regressor with hyper parameter tuning emerges as the best-performing model, achieving superior accuracy and lower error rates compared to others.

4. Prediction Output

ayer Prediction

Generation:

After training, the selected Random Forest model generates predictions for health insurance costs based on new input data from users. The model takes user inputs (age, sex, BMI, smoker status, region, and number of children) and processes them to output the predicted insurance charges.

- Real-Time Predictions: The predicted charges are outputted in real-time, enabling insurance providers to deliver instant quotes to potential customers. This capability is essential in a competitive insurance market where quick responses can influence customer decisions.

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

- Visualization of Results: The system may also incorporate visualizations to present predictions alongside feature impacts, helping insurers and customers understand how various factors influence costs.
- Feedback Loop: Post-prediction, feedback can be gathered from actual insurance charges to continually refine the model, enhancing its accuracy over time through additional training with updated data.

Tools/Technologies Used

In developing the **Health Insurance Premium Prediction System**, various tools and technologies were employed to ensure efficient data processing, model training, and user interaction. Each of these tools played a sign if I control e in enhancing the project's overall functionality and effectiveness. Below are the key tools and technologies used in the project:

1. Programming Language: Python

Python serves as the backbone of this project, chosen for its simplicity, versatility, and robust ecosystem of libraries tailored for data analysis and machine learning.

- **Rich Ecosystem:** Python boasts an extensive range of libraries and frameworks such as Pandas for data manipulation, NumPy for numerical computing, Scikit-learn for machine learning, and Matplotlib and Seaborn for data visualization. This ecosystem allows for efficient handling of datasets, facilitating complex mathematical operations and algorithm implementations.
- **Ease of Learning:** Python's straightforward syntax and readability make it accessible to both beginners and seasoned developers. This aspect is particularly beneficial in collaborative environments, where team members may possess varied levels of programming expertise.

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

- **Community Support:** The extensive Python community contributes to a wealth of resources, including comprehensive documentation, tutorials, and forums for troubleshooting. This active support network aids developers in resolving issues quickly, promoting efficient project development.

2. Data Manipulation: Pandas

Pandas is an essential library utilized for data preprocessing and analysis in this project. Its significance includes:

- **Data Structures:** Pandas provides two primary data structures: **Series** for one-dimensional data and **Data Frame** for two-dimensional data. These structures enable efficient manipulation and analysis of data sets, making operations such as filtering, aggregation, and transformation straightforward.
- **Data Cleaning:** The library offers built-in functions for handling missing data, filtering outliers, and converting data types. Ensuring high data quality is crucial for the modeling process, and Pandas facilitates these tasks with ease.
- **Time Series Analysis:** Pandas has robust capabilities for handling time series data, making it useful if the project requires analysis based on time components, such as trends in insurance charges over time.
- **Integration with Other Libraries:** Pandas integrates seamlessly with other data science libraries, allowing smooth transitions between data manipulation and modeling tasks. This interoperability is vital for developing complex analytical workflows.

3. DataVisualization: Matplotlib and Seaborn

Matplotlib and Sea bornare pivotal libraries for data visualization in this project, providing

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

powerful tools for generating insightful graphics.

- **Matplotlib:** As the foundational plotting library for Python, Matplotlib provides flexibility to create a wide variety of static, animated, and interactive visualizations. It allows developers to customize plots extensively, which is particularly useful for detailed data exploration and presentation.
- **Seaborn:** Built on top of Matplotlib, Seaborn enhances visualization capabilities with additional functionalities and better aesthetics. It simplifies the process of creating complex visualizations, such as heatmaps and pair plots, which help in understanding relationships between features in the dataset.
- **Insights Generation:** Visualization aids in data exploration, enabling stakeholders to identify patterns, anomalies, and correlations in the dataset. By visualizing the data effectively, insights can be generated that inform the modeling process and decision-making.

4. Machine Learning Framework: Scikit-learn

Scikit-learnisa powerful machine learning library that provides tools for model training, evaluation, and validation.

- **Wide Range of Algorithms:** Scikit-learn offers numerous algorithms for classification, regression, clustering, and dimensionality reduction. This diversity allows for experimenting with different models to find the best fit for the data, which is crucial for accurate predictions.
- **Model Evaluation Tools:** The library includes functions for splitting datasets, cross-validation, and various performance metrics (e.g., R-squared, Mean Squared Error). These tools help assess model performance objectively, ensuring the best model is selected for deployment.

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

- **Pipeline Support:** Scikit-learn supports the creation of machine learning pipelines, which streamline the process of data preprocessing and model training. This feature enhances code readability, maintainability, and overall project organization.

- **Feature Engineering:** Scikit-learn provides tools for feature selection and transformation, enabling developers to improve model performance by selecting the most relevant features from the dataset.

5. Hyper parameter Tuning: Grid Search CV

Grid Search CV is an essential technique within Scikit-learn that automates the process of hyper parameter tuning.

- **Optimization of Model Parameters:** Hyper parameter tuning is crucial for enhancing model performance. Grid Search CV systematically tests different combinations of hyper parameters, helping to identify the optimal settings for each model employed in the project.

- **Cross-Validation:** By incorporatingcross-validation, Grid Search CV ensures that the model is validated on different subsets of the data, reducing the risk of over fitting. This method provides a more accurate estimate of model performance on unseen data.

  Efficiency: Automating the tuning process saves significant time and effort compared to manual tuning. It allows data scientists to focus on other critical aspects of the project, such as feature engineering and data analysis.

6. Integrated Development Environment (IDE): Jupyter Notebook

Jupyter Notebook is an open-source web application that provides an interactive environment for coding, visualizing data, and documenting the project.

- **Interactive Coding:** Jupyter allows for real-time code execution and immediate feedback, making it an ideal platform for exploratory data analysis and iterative development. This interactivity enhances the learning experience and facilitates quick adjustments to the

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

code.

- **Visualization Support:** The notebook format supports in line visualizations, enabling users to see plots and charts alongside the code that generates them. This integration enhances understanding and communication of findings, making it easier to present insights to stakeholders.
- **Documentation:** Jupyter Notebooks facilitate the inclusion of Markdown cells, allowing for the integration of explanations, notes, and conclusions. This feature is vital for creating comprehensive documentation of the analysis and modeling process, ensuring transparency and reproducibility.

7. Version Control: Git and GitHub

Git and GitHub areessential tools for version control and collaboration throughout the project.

- **Source Code Management:** Git enables tracking changes in code, allowing developers to revert to previous versions if necessary. This capability is crucial for maintaining the integrity of the code base and facilitating collaboration among multiple developers.
- **Collaboration Features:** GitHub provides a platform for collaborative development, allowing multiple contributors to work on the project simultaneously. Features such as pull requests, issue tracking, and project boards facilitate efficient collaboration and project management.
- **Documentation and Portfolio:** Hosting the project on GitHub creates a public repository that serves as documentation of the development process and a portfolio piece for showcasing skills to potential employers. This visibility can enhance career opportunities and professional networking.
- **Branching and Merging:** Git's branching feature allows developers to work on new features or fixes without affecting the main codebase. Once changes are validated, they can

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

be merged back into the main branch, ensuring a stable and cohesive codebase.

8. Data Storage: CSV and Excel Formats

Data storage is a critical aspect of any data-driven project. In this project, data sets a restore din CSV and Excel formats.

- **CSV (Comma-Separated Values):** The CSV format is light weight, easy to read, and widely supported by various data manipulation libraries. It allows for efficient storage and sharing of structured data, making it a suitable choice for our data set of health insurance information.
- **Excel:** Excel files offer additional functionalities such as formatting and formulas. They are user-friendly for stakeholders who may not be familiar with programming. The ability to open and manipulate data in Excel can facilitate initial data inspection and exploration.

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

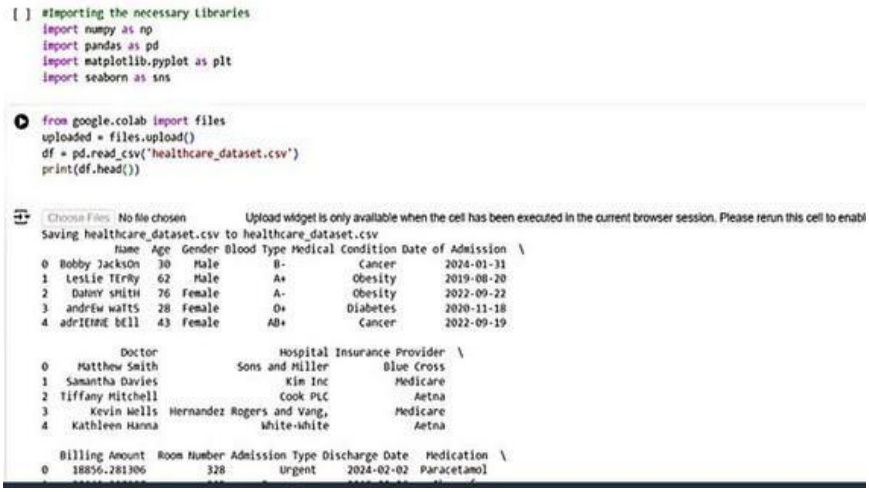
(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

Sample Generated Code Screens:



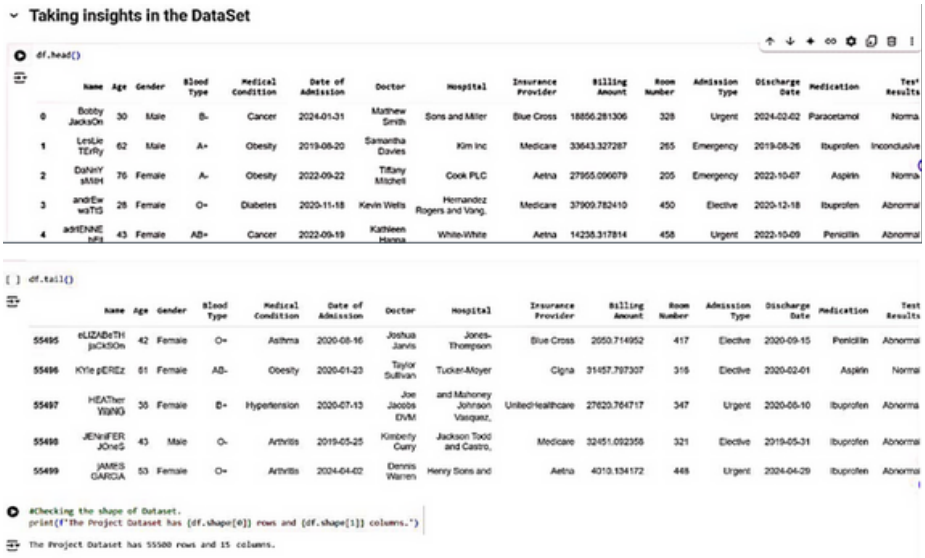
```
[ ] #Importing the necessary Libraries
    import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
```

```
from google.colab import files
uploaded = files.upload()
df = pd.read_csv('healthcare_dataset.csv')
print(df.head())
```

```
Choose Files  No file chosen        Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enabl
Saving healthcare_dataset.csv to healthcare_dataset.csv
              Name  Age  Gender Blood Type Medical Condition Date of Admission  \
0  Bobby JacksOn   30    Male         B-           Cancer        2024-01-31
1     LeslIe TErRy   62    Male         A+          Obesity        2019-08-20
2      DaNnY sMiTH   76  Female         A-          Obesity        2022-09-22
3     andrEw waTtS   28  Female         O+         Diabetes        2020-11-18
4  adrIENNE bEll   43  Female         AB+           Cancer        2022-09-19

              Doctor                 Hospital Insurance Provider  \
0     Matthew Smith            Sons and Miller        Blue Cross
1   Samantha Davies                   Kim Inc          Medicare
2  Tiffany Mitchell                  Cook PLC             Aetna
3      Kevin Wells  Hernandez Rogers and Vang,          Medicare
4   Kathleen Hanna                White-White             Aetna

   Billing Amount  Room Number Admission Type Discharge Date   Medication  \
0     18856.281306         328         Urgent     2024-02-02  Paracetamol
```



Taking insights in the DataSet

```
df.head()
```

	Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type	Discharge Date	Medication	Test Results
0	Bobby JacksOn	30	Male	B-	Cancer	2024-01-31	Matthew Smith	Sons and Miller	Blue Cross	18856.281306	328	Urgent	2024-02-02	Paracetamol	Normal
1	LesLie TErRy	62	Male	A+	Obesity	2019-08-20	Samantha Davies	Kim Inc	Medicare	33643.327287	265	Emergency	2019-08-26	Ibuprofen	Inconclusive
2	DaNnY sMiH	76	Female	A-	Obesity	2022-09-22	Tiffany Mitchell	Cook PLC	Aetna	27955.096079	205	Emergency	2022-10-07	Aspirin	Normal
3	andrEw waTtS	28	Female	O+	Diabetes	2020-11-18	Kevin Wells	Hernandez Rogers and Vang,	Medicare	37909.782410	450	Elective	2020-12-18	Ibuprofen	Abnormal
4	adrIENNE bEll	43	Female	AB+	Cancer	2022-09-19	Kathleen Hanna	White-White	Aetna	14238.317814	458	Urgent	2022-10-09	Penicillin	Abnormal

```
[ ] df.tail()
```

	Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type	Discharge Date	Medication	Test Results
55495	eLIZABeTH jaCkSOn	42	Female	O+	Asthma	2020-08-16	Joshua Jarvis	Jones-Thompson	Blue Cross	2050.714952	417	Elective	2020-09-15	Penicillin	Abnormal
55496	KYle pEREz	61	Female	AB-	Obesity	2020-01-23	Taylor Sullivan	Tucker-Moyer	Cigna	31457.797307	316	Elective	2020-02-01	Aspirin	Normal
55497	HEATher WaNG	36	Female	B+	Hypertension	2020-07-13	Joe Jacobs DVM	and Mahoney Johnson Vasquez,	UnitedHealthcare	27620.764717	347	Urgent	2020-08-10	Ibuprofen	Abnormal
55498	JENniFER JOneS	43	Male	O-	Arthritis	2019-05-25	Kimberly Curry	Jackson Todd and Castro,	Medicare	32451.092358	321	Elective	2019-05-31	Ibuprofen	Abnormal
55499	JAMES GARCIA	53	Female	O+	Arthritis	2024-04-02	Dennis Warren	Henry Sons and	Aetna	4010.134172	448	Urgent	2024-04-29	Ibuprofen	Abnormal

```
#Checking the shape of Dataset.
print(f"The Project Dataset has {df.shape[0]} rows and {df.shape[1]} columns.")
```

```
The Project Dataset has 55500 rows and 15 columns.
```

```
[ ] df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 55500 entries, 0 to 55499
Data columns (total 15 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Name                55500 non-null  object
 1   Age                 55500 non-null  int64
 2   Gender              55500 non-null  object
 3   Blood Type          55500 non-null  object
 4   Medical Condition   55500 non-null  object
 5   Date of Admission   55500 non-null  object
 6   Doctor              55500 non-null  object
```

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

```
df.describe(include= "object").T
```

	count	unique	top	freq
Name	55500	40235	Michael Williams	24
Gender	55500	2	Male	27774
Blood Type	55500	8	A-	6969
Medical Condition	55500	6	Arthritis	9308
Doctor	55500	40341	Michael Smith	27
Hospital	55500	39876	LLC Smith	44
Insurance Provider	55500	5	Cigna	11249
Admission Type	55500	3	Elective	18655
Medication	55500	5	Lipitor	11140
Test Results	55500	3	Abnormal	18627

```
df["Gender"].value_counts()
```

Gender	count
Male	27774
Female	27726

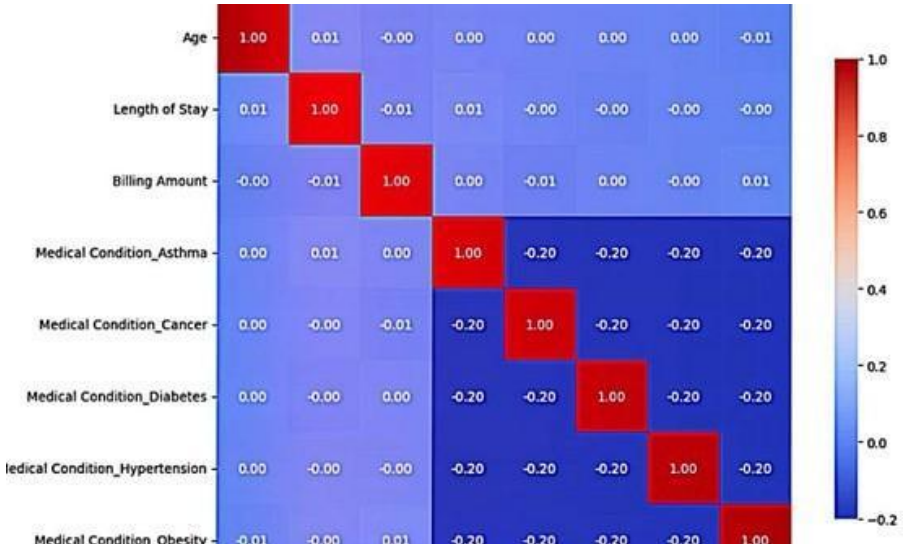
IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

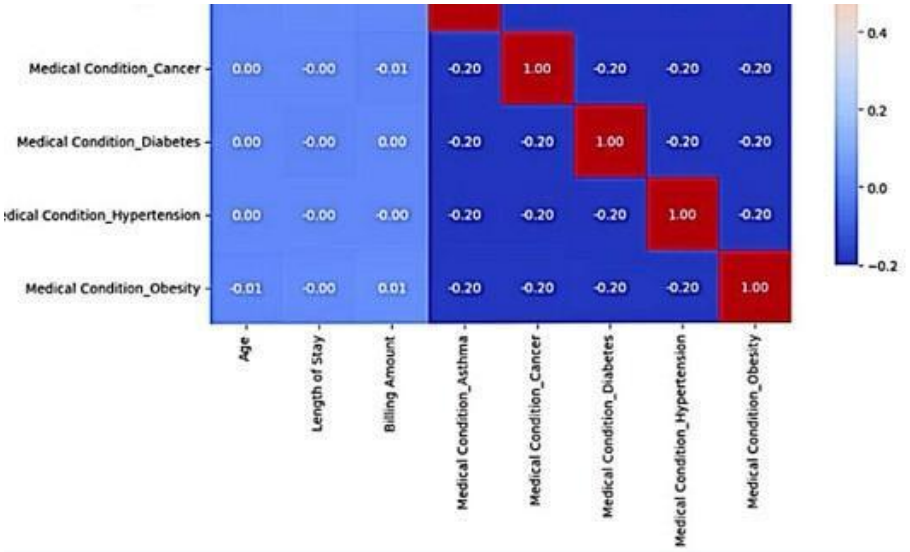
(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"






```
[ ] df2 = df.drop(columns=["Name", "Room Number","Doctor","Hospital","Insurance Provider"])
    df2
```

	Age	Gender	Blood Type	Medical Condition	Date of Admission	Billing Amount	Admission Type	Discharge Date	Medication	Test Results	Length of Stay	Age Group
0	30	Male	B-	Cancer	2024-01-31	18856.281306	Urgent	2024-02-02	Paracetamol	Normal	2	Adult
1	62	Male	A+	Obesity	2019-08-20	33643.327287	Emergency	2019-08-26	Ibuprofen	Inconclusive	6	Middle-aged
2	76	Female	A-	Obesity	2022-09-22	27955.096079	Emergency	2022-10-07	Aspirin	Normal	15	Senior
3	28	Female	O+	Diabetes	2020-11-18	37909.782410	Elective	2020-12-18	Ibuprofen	Abnormal	30	Adult
4	43	Female	AB+	Cancer	2022-09-19	14238.317814	Urgent	2022-10-09	Penicillin	Abnormal	20	Middle-aged
...	...	...	...	...	...	...	...	...	...	...	...	...
55495	42	Female	O+	Asthma	2020-08-16	2650.714952	Elective	2020-09-15	Penicillin	Abnormal	30	Middle-aged
55496	61	Female	AB-	Obesity	2020-01-23	31457.787303	Elective	2020-02-04	Aspirin	Normal	9	Middle-aged
55497	38	Female	B+	Hypertension	2020-07-13	27620.764717	Urgent	2020-08-10	Ibuprofen	Abnormal	28	Adult

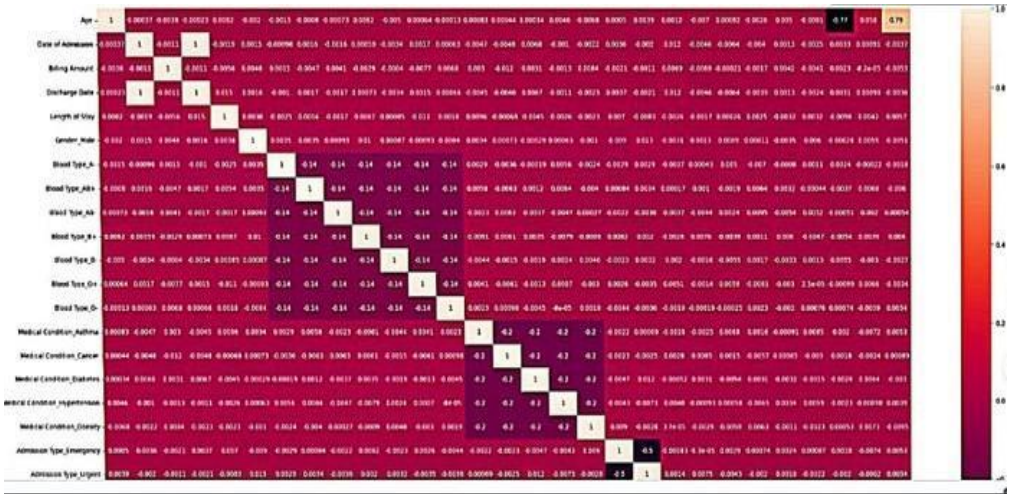
IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"



	Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type	Discharge Date	Medication	Test Results	...of Stay	Age Group
0	Bobby Jackson	30	Male	B-	Cancer	2024-01-31	Matthew Smith	Sons and Miller	Blue Cross	18856.281306	328	Urgent	2024-02-02	Paracetamol	Normal	2	Adu
1	Leslie Terry	62	Male	A+	Obesity	2019-08-20	Samantha Davies	Kim Inc	Medicare	33643.327287	265	Emergency	2019-08-26	Ibuprofen	Inconclusive	6	Middle age
2	Danny Smith	76	Female	A-	Obesity	2022-09-22	Tiffany Mitchell	Cook PLC	Aetna	27955.096079	205	Emergency	2022-10-07	Aspirin	Normal	15	Senic
3	Andrew Watts	28	Female	O+	Diabetes	2020-11-18	Kevin Wells	Hernandez Rogers and Vang,	Medicare	37909.782410	450	Elective	2020-12-18	Ibuprofen	Abnormal	30	Adu
4	Adrienne Bell	43	Female	AB+	Cancer	2022-09-19	Kathleen Hanna	White-White	Aetna	14238.317814	458	Urgent	2022-10-09	Penicillin	Abnormal	20	Middle age
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5495	Elizabeth Jackson	42	Female	O+	Asthma	2020-08-16	Joshua Jarvis	Jones-Thompson	Blue Cross	2650.714952	417	Elective	2020-09-15	Penicillin	Abnormal	30	Middle age
5496	Kyle Perez	61	Female	AB-	Obesity	2020-01-23	Taylor Sullivan	Tucker-Moyer	Cigna	31457.797307	316	Elective	2020-02-01	Aspirin	Normal	9	Middle age
5497	Heather Wang	38	Female	B+	Hypertension	2020-07-13	Joe Jacobs DVM	and Mahoney Johnson Vasquez,	UnitedHealthcare	27620.764717	347	Urgent	2020-08-10	Ibuprofen	Abnormal	28	Adu
5498	Jennifer Jones	43	Male	O-	Arthritis	2019-05-25	Kimberly Curry	Jackson Todd and	Medicare	32451.092358	321	Elective	2019-05-31	Ibuprofen	Abnormal	6	Middle age

```
[ ]
```

Gender	Male
Blood Type	B+
Medical Condition	Hypertension
Date of Admission	2022-08-22 00:00:00
Doctor	Gregory Hansen
Hospital	Ltd Wang
Insurance Provider	Blue Cross
Billing Amount	26062.43432
Room Number	482
Admission Type	Elective
Discharge Date	2022-09-07 00:00:00
Medication	Paracetamol
Test Results	Inconclusive
Length of Stay	16
Age Group	Middle-aged

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

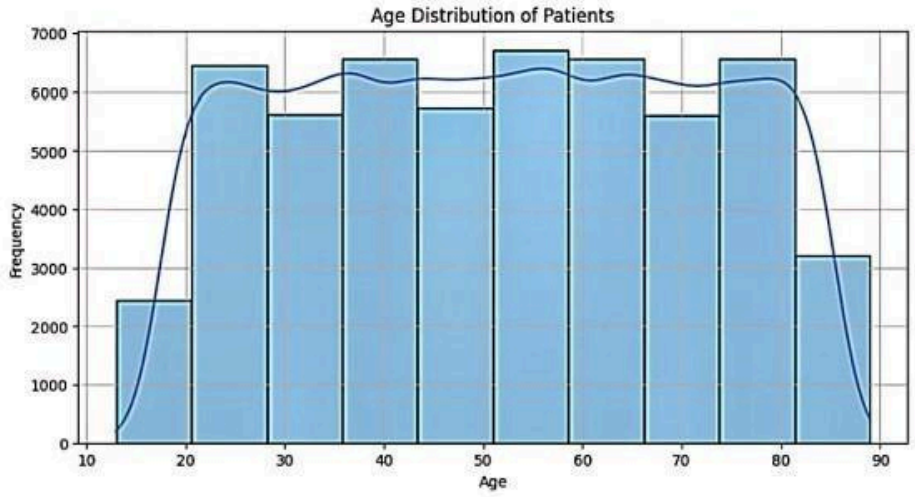
52894	Joseph Cox	23	Male	AB-	Diabetes	2019-10-13	Peter Smith	Inc Ward	Blue Cross	-353.865186	271	Elective	2019-10-25	Lipitor	Inconclusive	12	Ad
53204	Ashley Warner	55	Male	A+	Hypertension	2021-12-21	Andrea Bentley	and Wagner, Lee Klein	Aetna	-306.364925	426	Elective	2022-01-11	Ibuprofen	Normal	21	Midd ag
53232	Daniel Drake	68	Female	B+	Hypertension	2020-04-24	Brett Ray	Carr Ltd	Aetna	-691.917419	426	Elective	2020-04-26	Lipitor	Abnormal	2	Seri
54136	Dr. Michael Mckay	64	Male	O+	Cancer	2019-05-31	Dawn Navarro	Mcconnell and Rios, Clark	UnitedHealthcare	-108.663795	122	Urgent	2019-06-12	Ibuprofen	Abnormal	12	Midd ag
55276	John Ferrel	68	Female	O-	Hypertension	2019-05-20	Randy Calderon	Inc Spencer	Medicare	-308.584259	394	Emergency	2019-05-27	Paracetamol	Inconclusive	7	Midd ag

116 rows × 17 columns

VISUALISATION

```
# 1. Age Distribution
plt.figure(figsize=(10, 5))
sns.histplot(df['Age'], bins=10, kde=True)
plt.title('Age Distribution of Patients')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.grid()
plt.show()
```

```
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.grid()
plt.show()
```




```
average_billing = df.groupby('Medical Condition')['Billing Amount'].mean()

# Create a running average of the billing amounts
running_average = average_billing.rolling(window=2).mean()  # Adjust window size as needed

# Prepare data for plotting
x = average_billing.index  # Medical conditions
y = average_billing.values  # Average billing amounts
average_y = running_average.values  # Running average of billing amounts

# Plotting
plt.figure(figsize=(10, 5))
plt.plot(x, y, 'k.-', label='Average Billing Amount')  # Original data
plt.plot(x, average_y, 'r.-', label='Running Average')  # Running average
plt.title('Average Billing Amount by Medical Condition', fontsize=16)
plt.xlabel('Medical Condition', fontsize=14)
```

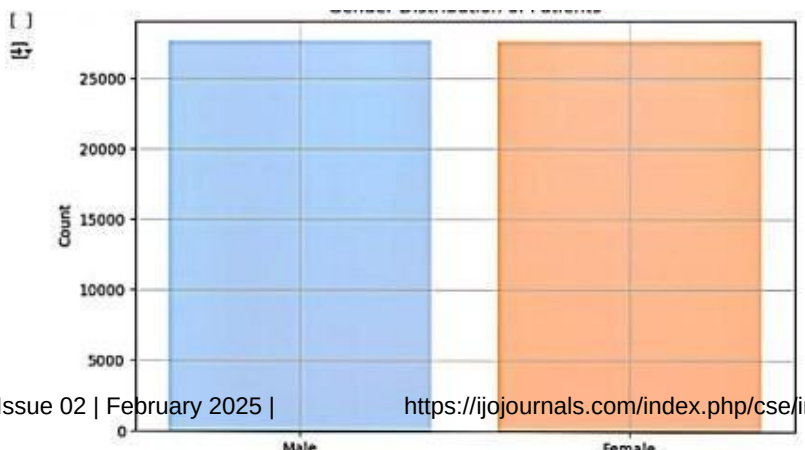
IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"



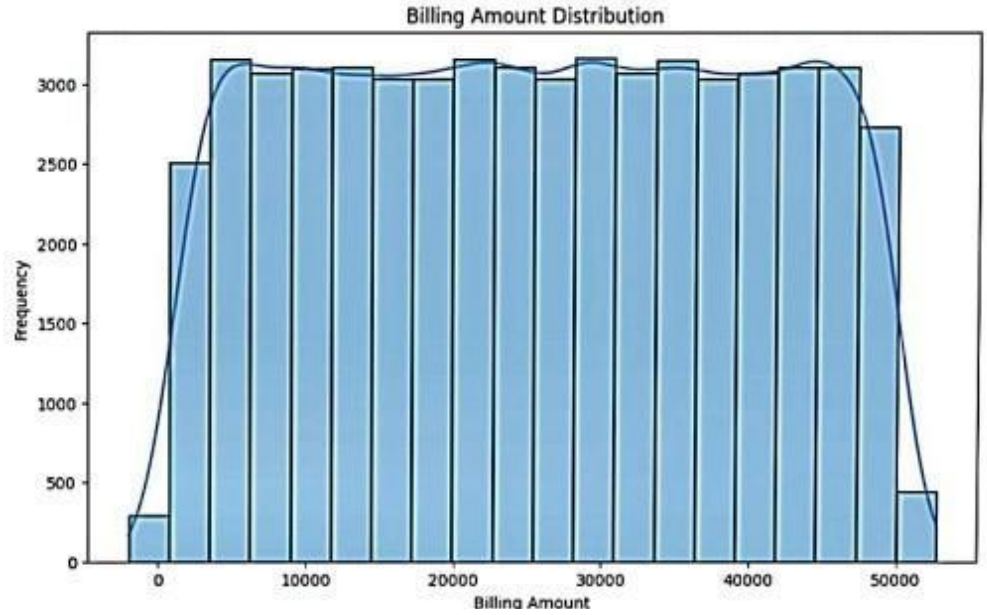
IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

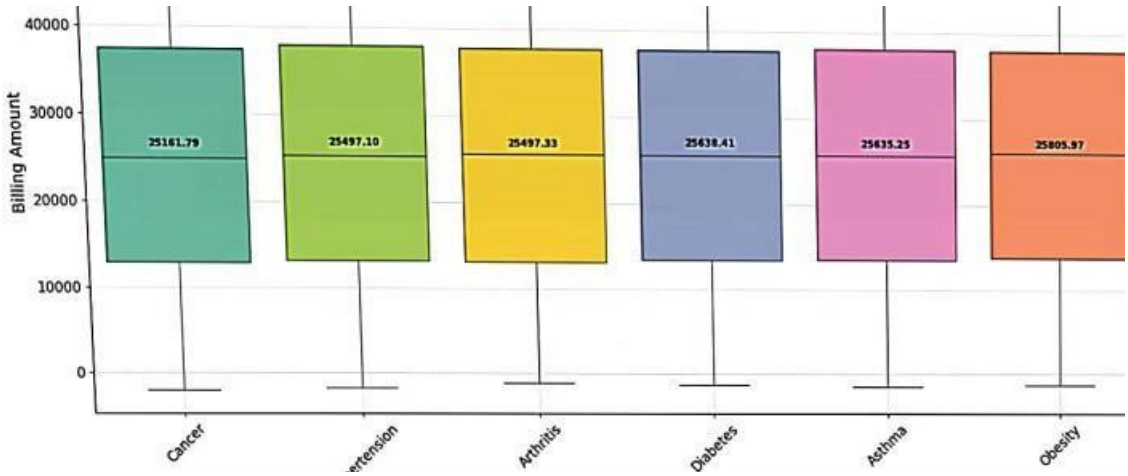
(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"






```
order = df.groupby('Medical Condition')['Billing Amount'].median().sort_values().index
plt.figure(figsize=(14, 8), dpi=100)

sns.boxplot(x='Medical Condition', y='Billing Amount', data=df, order=order, hue="Medical Condition" ,palette="Set2")
# sns.swarmplot(x="Medical Condition", y="Billing Amount", data=df, order=order, color=".25", alpha=0.5)

# Calculate and annotate mean bill
means = df.groupby('Medical Condition')['Billing Amount'].mean().reindex(order)
for index, mean in enumerate(means):
    plt.text(index, mean + 1000, f'{mean:.2f}', horizontalalignment='center', size='small', color='black', weight='semibold')

plt.title('Billing Amount by Medical Condition', fontsize=16)
plt.xlabel('Medical Condition', fontsize=14)
plt.ylabel('Billing Amount', fontsize=14)
plt.xticks(rotation=45, fontsize=12)
plt.yticks(fontsize=12)
plt.grid(True, which='both', linestyle='--', linewidth=0.5)
```

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

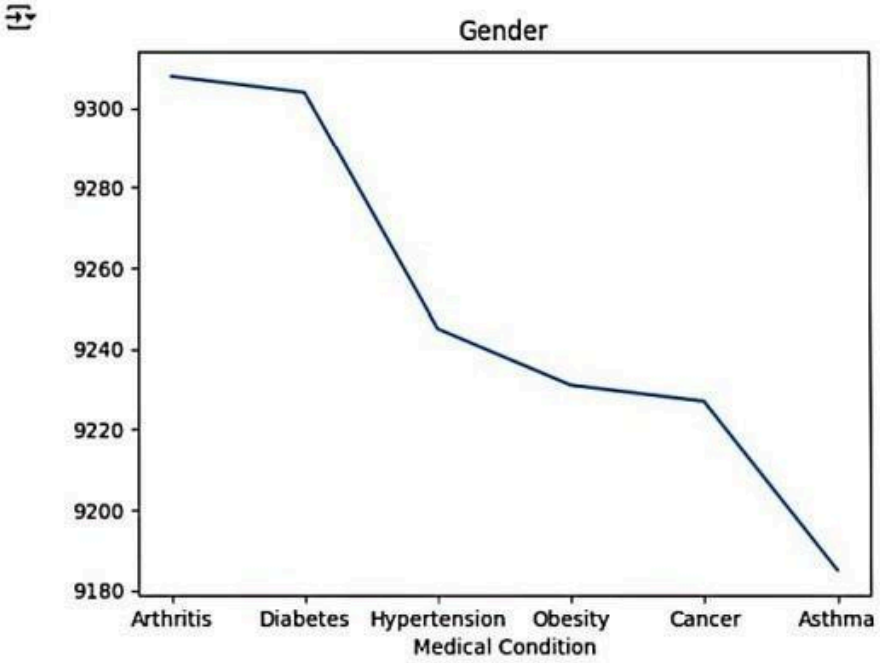
(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

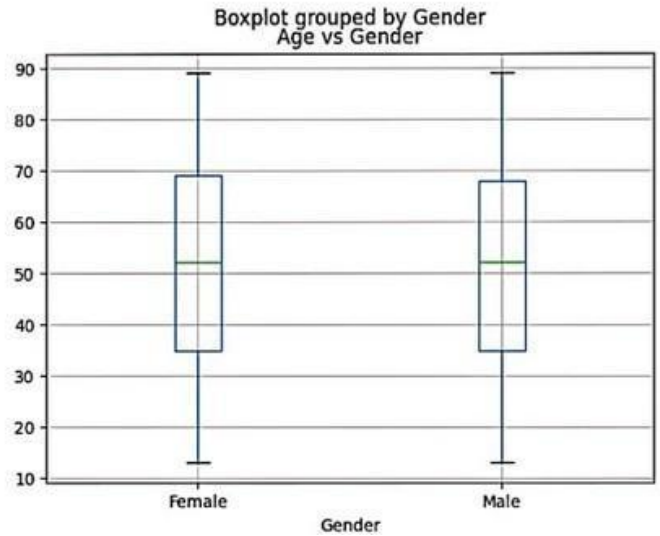
Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

```
df['Medical Condition'].value_counts().plot()
plt.title('Gender')
plt.show()
```




```
df.boxplot(column='Age', by='Gender')
plt.title('Age vs Gender')
plt.show()
```




```
disease_trends = df.groupby([df['Date of Admission'].dt.to_period('M'), 'Medical Condition']).size().unstack().fillna(0)

plt.figure(figsize=(18, 10), dpi=100)

# Use seaborn color palette for better aesthetics
colors = sns.color_palette("Set2", len(disease_trends.columns))

for disease, color in zip(disease_trends.columns, colors):
    plt.plot(disease_trends.index.astype(str), disease_trends[disease], label=disease, marker='o', linestyle='-', color=color)
```

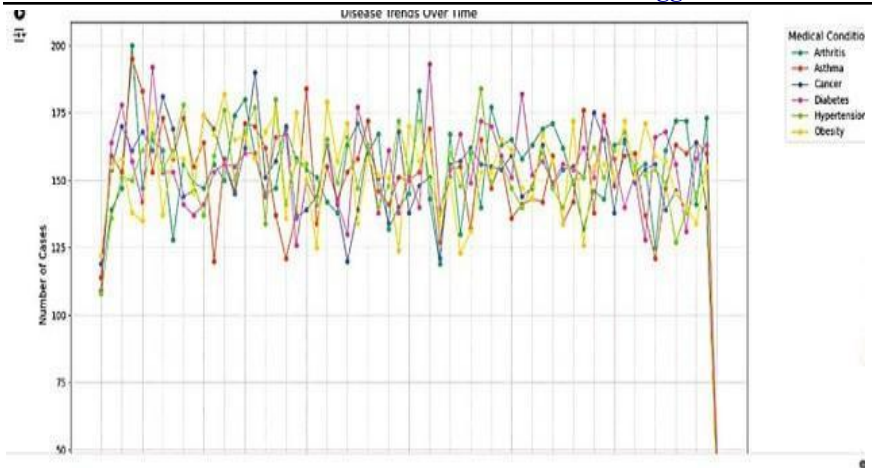
IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

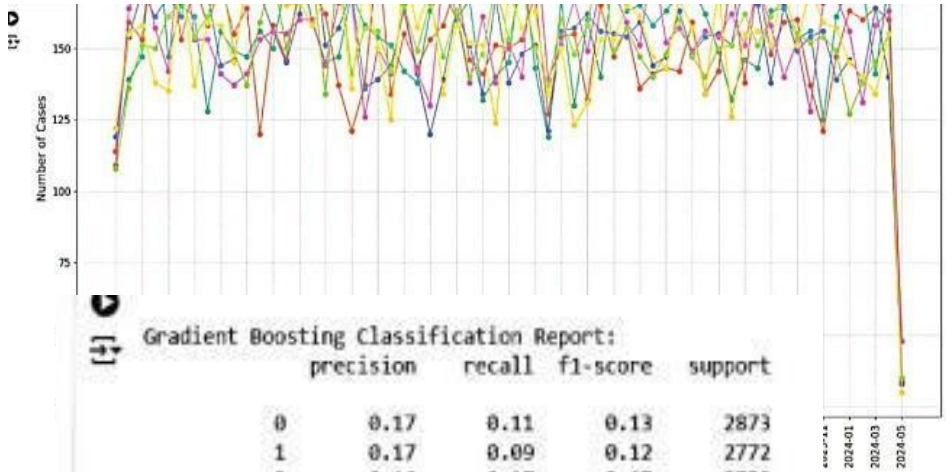
(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"





Gradient Boosting Classification Report:

	precision	recall	f1-score	support
0	0.17	0.11	0.13	2873
1	0.17	0.09	0.12	2772
2	0.16	0.17	0.17	2758
3	0.17	0.16	0.17	2816
4	0.17	0.30	0.21	2654
5	0.17	0.18	0.18	2777
accuracy			0.17	16650
macro avg	0.17	0.17	0.16	16650
weighted avg	0.17	0.17	0.16	16650

Confusion Matrix for Gradient Boosting:

```
[[320 267 513 439 843 491]
 [304 257 464 464 788 495]
 [315 249 469 424 786 515]
 [307 265 505 448 772 519]
 [294 219 461 489 790 481]
 [331 258 488 407 780 513]]
```

```
[ ] y_pred
```

ication	Test Results	Length of Stay	Ag Grou
cetamol	Normal	2	Adu
uprofen	Inconclusive	6	Middle age
Aspirin	Normal	15	Senic
uprofen	Abnormal	30	Adı

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

```
[ ] x_train, x_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    model = LinearRegression()
    model.fit(x_train, y_train)

    LinearRegression()

[ ] y_pred = model.predict(x_test)
    y_pred

    array([25542.04579224, 25531.6450165 , 25652.66668899, ...,
           25240.50593075, 25431.77797571, 25431.68715571])

[ ] accuracy_score = model.score(x_test, y_test)

[ ] accuracy_score

    -0.0006801402824287983

[ ] sqrt(mean_squared_error(y_test, y_pred))

    14108.319583976287

[ ] print(mean_squared_error(y_test, y_pred))
```

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

```
[ ] print(mean_squared_error(y_test, y_pred))

199044681.48360884

[ ] print(mean_absolute_error(y_test, y_pred))

12192.39034847588

[ ] print(r2_score(y_test, y_pred))

-0.0006801402824287983
```

Gradient Boosting

```
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.metrics import classification_report, confusion_matrix
```

```
[ ] # Encode categorical variables
label_encoders = {}
for column in ['Test Results', 'Medical Condition', 'Admission Type', 'Medication']:
    le = LabelEncoder()
    df[column] = le.fit_transform(df[column])
    label_encoders[column] = le

# Define features and target variable
X = df[['Test Results', 'Admission Type', 'Medication']]
```

Decision Tree

Predicting Medication on the basis of Age, Medical Condition Test Results and Admission Type.

```
[ ] from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
df['Medical Condition Encoded'] = label_encoder.fit_transform(df['Medical Condition'])
df['Test Results Encoded'] = label_encoder.fit_transform(df['Test Results'])
df['Admission Type Encoded'] = label_encoder.fit_transform(df['Admission Type'])
df['Medication Encoded'] = label_encoder.fit_transform(df['Medication'])

# Selecting features for the model
features = ['Age', 'Medical Condition Encoded', 'Test Results Encoded', 'Admission Type Encoded']
X = df[features]
y = df['Medication Encoded']
```

```
[ ] print(mean_squared_error(y_test, y_pred))

5.715855855855856

[ ] print(mean_absolute_error(y_test, y_pred))

1.9251051051051051

[ ] print(r2_score(y_test, y_pred))

-0.945979104838548
```

Random Forest

```
[ ] from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
```

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

```
model = DecisionTreeClassifier(random_state=42)
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)
class_report = classification_report(y_test, y_pred, target_names=label_encoder.inverse_transform(y.unique()).astype(str))

# Display results
print(f"Model Accuracy: {accuracy * 100:.2f}%")
print("Confusion Matrix:")
print(conf_matrix)
print("Classification Report:")
print(class_report)
```

```
Model Accuracy: 20.51%
Confusion Matrix:
[[642 480 412 366 311]
 [622 495 450 355 349]
 [581 488 473 348 334]
 [602 461 453 361 330]
 [636 489 446 310 306]]
Classification Report:
              precision    recall  f1-score   support

           3       0.21      0.29      0.24      2211
           1       0.21      0.22      0.21      2271
           0       0.21      0.21      0.21      2224
           4       0.21      0.16      0.18      2207
           2       0.19      0.14      0.16      2187

    accuracy                           0.21     11100
   macro avg       0.20      0.20      0.20     11100
weighted avg       0.20      0.21      0.20     11100
```

```
print(accuracy)
```

```
0.20513513513513512
```

```
sqrt(mean_squared_error(y_test, y_pred))
```

```
2.018897210346764
```

```
print(mean_squared_error(y_test, y_pred))
```

```
4.0759459459459455
```

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

```
[ ] print(mean_absolute_error(y_test, y_pred))

1.6894594594594596

[ ] print(r2_score(y_test, y_pred))

-1.0500823188813075
```

```
df
```

	Name	Age	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	...	Test Results	Length of Stay	Age Group	Gender_Female	Gender_Male	Test Results Encoded	Admiss: Ty Enco
0	Bobby Jackson	30	B-	2	2024-01-31	Matthew Smith	Sons and Miller	Blue Cross	18856.281306	328	...	2	2	Adult	False	True	2	
1	Leslie Terry	62	A+	5	2019-08-20	Samantha Davies	Kim Inc	Medicare	33643.327287	265	...	1	6	Middle-aged	False	True	1	
2	Danny Smith	76	A-	5	2022-09-22	Tiffany Mitchell	Cook PLC	Aetna	27955.096079	205	...	2	15	Senior	True	False	2	
3	Andrew Watts	28	O+	3	2020-11-18	Kevin Wells	Hernandez Rogers and Vang,	Medicare	37909.782410	450	...	0	30	Adult	True	False	0	
4	Adrienne Bell	43	AB+	2	2022-09-19	Kathleen Hanna	White-White	Aetna	14238.317814	458	...	0	20	Middle-aged	True	False	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1	Leslie Terry	62	A+	5	2019-08-20	Samantha Davies	Kim Inc	Medicare	33643.327287	265	...	1	6	Middle-aged	False	True	1	
2	Danny Smith	76	A-	5	2022-09-22	Tiffany Mitchell	Cook PLC	Aetna	27955.096079	205	...	2	15	Senior	True	False	2	
3	Andrew Watts	28	O+	3	2020-11-18	Kevin Wells	Hernandez Rogers and Vang,	Medicare	37909.782410	450	...	0	30	Adult	True	False	0	
4	Adrienne Bell	43	AB+	2	2022-09-19	Kathleen Hanna	White-White	Aetna	14238.317814	458	...	0	20	Middle-aged	True	False	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
55495	Elizabeth Jackson	42	O+	1	2020-06-16	Joshua Jarvis	Jones-Thompson	Blue Cross	2650.714952	417	...	0	30	Middle-aged	True	False	0	
55496	Kyle Perez	61	AB-	5	2020-01-23	Taylor Sullivan	Tucker-Moyer	Cigna	31457.797307	316	...	2	9	Middle-aged	True	False	2	
55497	Heather Wang	36	B+	4	2020-07-13	Joe Jacobs DVM	and Mahoney Johnson Vasquez,	UnitedHealthcare	27620.764717	347	...	0	28	Adult	True	False	0	
55498	Jennifer Jones	43	O-	0	2019-05-25	Kimberly Curry	Jackson Todd and Castro,	Medicare	32451.092358	321	...	0	6	Middle-aged	False	True	0	
55499	James Garcia	53	O+	0	2024-04-02	Dennis Warren	Henry Sons and	Aetna	4010.134172	448	...	0	27	Middle-aged	True	False	0	

55500 rows × 23 columns

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

Conclusion

Model	R²Score	MeanSquaredError(MSE)
LinearRegression	0.718	0.528
Ridge	0.792	0.538
Lasso	0.74	0.575
RandomForestregression	0.90	0.320
DecisionTreeRegressor	0.81	0.425
GradientBoostingRegressor	0.883	0.376
XGBoost	0.89	0.330
RandomForest(Tuned)	0.93	0.468

ConclusionBasedonModelPerformance:

1. RandomForest(Tuned):ThismodelachievedthebestR²scoreof0.93,indicatingitexplains 93%ofthevariance inthetarget variable, making it themost accurate modelfor predicting medical insurance costs based on the features provided. However, its MSE of 0.468 is slightly higher thansome other models, which may suggest overfitting to the training data, but overall it's highly effective.

2. Random Forest (Untuned): The untuned Random Forest model also performed very well, withanR²scoreof0.90andaMSEof0.320,showing it'sastrongpredictorwitharelatively lowerror.Thissuggeststhatevenwithouttuning,RandomForestisareliablechoiceforthis task.

3. Gradient Boosting Regressor: With an R² score of 0.883 and an MSE of 0.376, this model performs very well, almost as good as Random Forest. It balances between high accuracy

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

and reasonable error, making it a solid alternative to Random Forest.

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

4. XGBoost:Another high-performing modelwithan R²scoreof0.89and MSEof0.330.It is closeinperformancetoRandomForestandGradient Boosting, makingitagoodcontender.

5. DecisionTreeRegressor:Thismodelperformeddecently,withanR²scoreof0.81andMSE of0.425. While it's simpler and interpretable, it is not as accurateor effective as ensemble methods like Random Forest or Gradient Boosting.

6. Ridge and Lasso: Both Ridge and Lasso regression models performed moderately, with R² scores of 0.792 and 0.74, respectively. Their MSEs are also higher compared to ensemble models, suggesting that they may not capture the complexity of the data as effectively.

7. Linear Regression: The linear regression model had the lowest performance, with an R² score of 0.718 and an MSE of 0.528, indicating it's not ideal for this dataset, which likely has nonlinear relationships.

BestModelforPredicting MedicalInsuranceBasedonAge:

- The Tuned Random Forest model stands out as the best model due to its high R² score of 0.96, which suggests it explains the vast majority of the variability in medical insurance costs.

- Gradient Boosting and XGBoost also performverywelland could be strong alternatives if you're looking for more generalized models with slightly lower chances of overfitting compared to the tuned Random Forest.

Therefore, Random Forest (Tuned) is recommended for predicting medical insurance costs, especially based on age, as it provides the most accurate results.

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

AGE	SMOKER	CHARGES
<20	YES	$14,000
>30and<40	YES	$19,922
>30and<40	YES	$21,000
>40and<50	YES	$25,000
>50and<60	YES	$27,000

AGE	SMOKER	CHARGES
<20	NO	$1900
>20and<30	NO	$3600
>30and<40	NO	$7200
>40and<50	NO	$11,000
>50and<60	NO	$13,000

IJO –INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND ENGINEERING

(ISSN: 2814-1881)
https://ijojournals.com/

Dr. NITALAKSHESWARA RAO KOLUKULA*

Volume 08 || Issue 02 || February, 2024 ||

" Medical Insurance Suggestion Based On Age"

Future Scope:

- Incorporating additional factors like medical history, family health records, and lifestyle habits to improve prediction accuracy.
- Expandingthedatasetwithreal-worldinsuranceclaimdataandadditionaldemographics.
- Developingawebandmobileapplicationforreal-timeinsurancecostestimationand recommendations.
- EnhancinginterpretabilitythroughExplainableAItechniquessuchasSHAPandLIME.
- Implementingblockchain-basedinsurancepricingforsecureandtransparentpremiumcalculations.
- Exploring federated learning to train models on decentralized health insurance data while preserving user privacy.

REFERENCES

Paper1: "HealthInsuranceCost PredictionusingMachineLearningIEEE" (2022) :
ThisstudyusedamedicalcostdatesetandappliedLinear Regression,achievinganaccuracyof81.3%.
Paper2:"HealthInsuranceCostPredictionUsingMachineLearningIRJET"(2022) :
ExploredXGBoostandRandomForestRegression,achievingimprovedpredictionaccuracy.

Paper3:"HealthInsuranceCostPredictionUsingMachineLearningICICC"(2022) : Introduced ensemble models such as Gradient Boosting, SHAP, and ICE for feature interpretation. These studies emphasize the importance of data-driven models in enhancing insurance cost predictions and highlight the need for further optimization.