Health Insurance Cost Predictions Using Machine Learning Algorithms

R.P.C. Thenuka*, W.M.G.A.S Weerakoon†, T.N. Colombage†, W.W.K.N. Ranathunga†, S.A.N. Nimsara†
Faculty of Information Technology, Horizon Campus, Malabe, Sri Lanka
Emails: *cthenuka09@gmail.com, †aweerakoon86@gmail.com, †Nimantha1020@gmail.com,
†kalananadun13@gmail.com, †nimsaranadupa@gmail.com

***Abstract*—Medical insurance is a vital component of healthcare systems that helps individuals overcome the financial burden of medical expenses. Health insurance costs have been rising, creating a need for accurate ways to predict costs. Machine learning techniques are increasingly being applied to forecast these costs with greater accuracy. This paper investigates the use of machine learning models for effective prediction of medical insurance costs based on various individual features and compares model performance to identify the most accurate algorithms.**

***Index Terms*—Health Insurance, Machine Learning, cross-validation, Hyper-parameter Tuning, Streamlit**

I. Introduction

Medical insurance is a vital component of healthcare systems that helps the individual overcome the financial burden of medical expenses. In recent years, people have been tending to buy medical health insurance. However, health insurance costs have been rising, leading to the need for more accurate ways to predict costs to prevent undercharging or overcharging customers. With increasing health costs, insurance companies have found it necessary to predict the insurance premium using customer profiles. In Sri Lanka and the World, there is a common problem called "How to predict the actual cost of Medical Insurance?" Some companies offer packages but these packages do not match with all customers. Because some customers have different types of diseases but others do not. Sometimes these Insurance Companies charge money by analyzing and using actuarial tables for the relevant factors. However, these companies predict the cost by using traditional methods. These Traditional methods are often time-consuming, causing pricing errors, and inefficient. Because they cannot analyze all the relevant factors or features related to the insurance cost. Insurance cost estimations have now become quite important, and machine learning techniques are increasingly being applied to forecast these costs with heightened accuracy, thus enabling more personalized premiums and improving overall efficiency within the insurance industry. In general, this project investigates the following research question:

- How can machine learning models be used for effective prediction of medical insurance costs based on individual features?
- Which feature(s) will have the most significant impact on predicting the cost of medical insurance, such as age, gender, BMI, etc.?
- Which machine learning algorithm will result in the highest accuracy for predicting insurance premiums?

It thus has practical implications both for insurers and insured people. To insurance companies, it will mean that better cost-predicting models translate into better pricing and risk management strategies. In return, this may mean fairer insurance premiums based on individual health and lifestyle factors that could encourage healthier behaviors. More efficient and effective.

II. Literature Review

Most studies now focus on predicting medical insurance costs using machine learning, since there is an emerging need for accurate and efficient models that estimate the premium of insurance. Many works aimed at solving this problem using different machine learning algorithms. Most of these studies are focused on regression models, principally Linear Regression, Decision Tree Regression, and Gradient Boosting Regression, among others. Most of the works depended on age, BMI, parity, and smoking status as some of the main demographic features in building the prediction model. These are, however, characterized by varying degrees of success concerning model performance and sometimes relatively lower performance of traditional machine learning techniques when compared to advanced methods such as deep learning.

A document titled "Medical Insurance Cost Prediction using Machine Learning," by Mukund Kulkarni et al., published in the International Journal for Research in Applied Science & Engineering Technology, IJRASET, in December 2022, focuses on the use of machine learning methods to forecast medical insurance costs. Their study applied various models, namely Linear Regression, Decision Tree Regression, and Gradient Boosting Regression, on a dataset from Kaggle. Their goal was to predict the insurance expenses as accurately as possible, considering some personal factors such as age, BMI, number of children, smoking, etc. The authors carried out the model evaluation process using Gradient Boosting Regression to obtain a maximum accuracy value of 86.86%. One of the limitations in their work is that their contribution relies mostly on regression techniques and does not emphasize deep learning models, which are capable of enhancing accuracy even more through better handling of complex dependencies in the data [1].

In 2023, Sazzad Hossen, a Bangladesh-based researcher from East West University, wrote a paper titled "Medical Insurance Cost Prediction Using Machine Learning." The study aimed to use machine learning (ML) algorithms to design a predictive model of insurance cost in order to better forecast the insurance premium rates. Hossen utilized multiple regression models, including Linear Regression, XGBoost, Lasso, Random Forest, Ridge, and Gradient Boosting Regression in making analyses and predictions. The dataset from Kaggle incorporated characteristics such as age, sex, BMI, number of children, smoking status, and area. The best model in the study was the XGBoost model, with an R-squared value of 0.8681. However, models such as Gradient Boosting and Random Forest also performed well. Some concerns arose from this study, including dealing with one specific dataset and the lack of real-time applications [2].

The research done by Sabarinath U S and Ashly Mathew (2024) discussed predicting medical insurance costs using machine learning models. This research analyzed how demographic elements like BMI, age, number of dependents, and region predict insurance premiums. This study uses different machine learning algorithms such as Support Vector Regression (SVR), Decision Trees, and Linear Regression to create accurate models. By performing accuracy tests and using error metrics like mean absolute error to analyze model performance, they found that Support Vector Regression is the best algorithm due to its ability to handle non-linear relationships among variables. However, their model is limited to the dataset they used, and complex socio-economic or lifestyle factors were not considered [3].

A study by Shoroog Albalawi et al. (2023) discusses the boundary of features for predicting health insurance costs. This research not only discusses age-related factors but also includes social and economic lifestyle factors such as smoking, drinking behaviors, type of employment, and health conditions. In this study, they used advanced machine learning algorithms like XGBoost, Random Forest, Gradient Boosting, and deep learning algorithms. They also discussed how hyperparameter optimization improves model performance. The research found that XGBoost is the best method for achieving accurate predictions in healthcare insurance costs. The use of socio-economic and lifestyle factors helps to enhance the models' performance compared to models that use only demographic data. However, a significant challenge identified was the computational complexity of these models, which is challenging for real-time insurance systems due to high computational costs. Collecting sensitive data like socio-economic and lifestyle data is another challenge [4].

The research titled "Machine Learning-Based Regression Framework to Predict Health Insurance Premiums" investigates the ability to predict the premiums of health insurance and mitigate the challenges of time-consuming and often inaccurate traditional methods. In this research, the authors used multiple algorithms such as Linear Regression, Decision Trees, Multiple Linear Regression, and Gradient Boosting. They found that Multiple Linear Regression and Gradient Boosting algorithms performed similarly well. However, the Gradient Boosting algorithm required less computational time than the Multiple Linear Regression algorithm. Their methodology includes data analysis, feature engineering, data training, and evaluation of a linear regression model, as well as training and evaluating an ANN-based (Artificial Neural Networks) regression model. This research is relevant to the current research because the authors developed a machine learning-based predictive system for predicting medical insurance costs [5].

The study by Orji and Ukwandu (2024) investigates machine learning models for predicting insurance costs, focusing on the models' accuracy and performance. The authors used three machine learning models: XGBoost, Gradient Boosting Regressor, and Random Forest Regressor. They used R-squared, MAE, RMSE, and MAPE for model evaluation and hyperparameter tuning to optimize model performance. Among the three models, the XGBoost model performed better than the other two. The research team used an open-source dataset from Kaggle to train and develop the prediction system. This research is also related because the research team developed a machine learning model for predicting the cost of insurance [6].

The research paper "Predicting the Cost of Health Insurance" was authored by Chintala Srikar, Maloth Kiran, Dubbudu Sumanth, and Preeti Jeevan from the Department of Computer Science and Engineering, Hyderabad, India. It was published in 2023 under the International Research Journal of Engineering and Technology (IRJET). The authors' research work focuses on medical insurance cost predictions using various machine learning regression models, such as Linear Regression, Multiple Linear Regression, Polynomial Regression, and Ridge Regression. The study examines the increasing relevance of health insurance in India due to rising healthcare costs, life expectancy, and non-communicable diseases. The researchers performed data preprocessing, handling missing values, and converting categorical variables into numerical ones using a dataset from Kaggle. One major limitation of their work was that Polynomial Regression achieved an accuracy of 80.97%. However, more advanced models such as Support Vector Machines, XGBoost, or Random Forest were not tested. These techniques will be explored in this research to address the gap for better prediction performance [7].

The research paper "Medical Health Insurance Cost Prediction" was presented by students Vaahini Reddy K, Kanisetti Uday Kalyani, Keerthy Reddy K, and Dr. D. Shravani from Stanley College of Engineering and Technology for Women, Osmania University, Hyderabad, India. The paper was published in April 2024 and researched predicting health insurance costs using various machine learning models. The adopted technologies used by the authors include SVMs, deep learning, linear regression, decision trees, and random forests. The data, collected on demographics, medical history, and lifestyle factors, were pre-processed using techniques such as one-hot encoding and normalization to ensure the data's quality. Their analysis revealed that the neural network model

gave the most accurate predictions, followed by the random forest model. However, the study had limitations, including its focus on a specific geographic region and the omission of important factors like pre-existing health conditions, which could influence insurance costs [8].

In the study "A Computational Intelligence Approach for Predicting Medical Insurance Cost" (2021), Ch. Anwar ul Hassan et al. focused on using machine learning for medical insurance cost forecasting. The researchers used a freely available dataset from Kaggle and employed machine learning algorithms such as Linear Regression, Support Vector Regression, Ridge Regressor, Stochastic Gradient Boosting, XGBoost, Decision Tree, Random Forest Regression, and K-Nearest Neighbor. The data was split into 70% training data and 30% test data. Stochastic Gradient Boosting (SGB) emerged as the most accurate model with a cross-validation score of 0.858 and a Root Mean Squared Error of 0.340. The researchers identified key attributes influencing insurance costs, such as age, BMI, and smoking status. Other attributes like sex, number of children, and region showed less correlation with costs. The researchers emphasized the need for continuous improvement of forecasting models to keep up with the evolving healthcare landscape and data availability [9].

The study "Medical Insurance Price Prediction Using Machine Learning" (2024) by Md Mohtaseem Billa and Dr. Tapsi Nagpal focused on the escalating costs and complexities in healthcare. The study emphasized the need for effective tools to forecast medical insurance costs. This study aimed to improve insurance pricing using a freely available dataset. Features included age, gender, BMI, number of children, smoking status, and region. The dataset contained 2773 rows. The methodology involved data cleaning and preprocessing to ensure accurate predictions. The researchers used several machine learning algorithms, including regression models, decision trees, and ensemble methods. The data was split into 80% for training and 20% for testing. Gradient Boost Regression emerged as the most accurate model, with an R-squared value of 0.8679 and an RMSE of 4453.83. The study showed that age, BMI, and smoking status significantly influenced insurance prices. Moreover, the study emphasizes the importance of accurate forecasting of insurance prices [10].

A. Key Theories or Concepts

From the literature, the following key concepts were identified:

1) **Linear Regression and Multiple Linear Regression**: Widely used in cost predictions but lacking in complex and nonlinear modeling relationships.
2) **Gradient Boosting Regression and XGBoost**: Ensemble methods frequently used, outperforming simple linear models with better predictive performance when optimized with hyperparameters.
3) **Support Vector Regression (SVR)**: Demonstrated improved capability in handling non-linear data relationships.
4) **Deep Learning Models**: While this is not a widely explored area, it is nonetheless increasingly noted as the next step towards capturing those relations that depend on intricate dependencies and non-linearities in the data, especially in neural networks, which had promise in such works as Vaahini Reddy K et al.

B. Gaps or Controversies in the Literature

During the literature review, the research team could identify the Gaps or Controversies in the Literature. These are,

1) **Absence of Deep Learning Techniques**: Most of the studies rely on basic machine learning models, such as regression and decision trees. none consider exploring the deep learning technique for possible benefits. Some of the studies note this as a limitation of the work and indicate that using neural networks may increase the accuracy of some predictions, especially when the dataset is very complex, as mentioned in Mukund Kulkarni et al. (2022) and Sazzad Hossen (2023).
2) **Limited Datasets**: This also seems to be an important limitation, such that specific datasets were selected-received, in most cases, from Kaggle. Considering the nature of these datasets, they may not carry complete information on the variability of influences on the cost of insurance-for instance, previous health conditions, geographical changes, or even real-time data.
3) **High Computational Costs**:The works of Shoroog Albalawi et al. show that studies which have used advanced models, including XGBoost and Random Forest, faced computational complexity challenges. This begets the major problem of scalability in dynamic environments because these models are applied for real-time purposes.
4) **Underrepresentation of Socio**: While some studies, like the one by Shoroog Albalawi, include socio-economic and lifestyle factors, most depend on parameters related to demographic data. In that case, this will lead to simplified models that might not represent all the complexities of the insurance cost prediction.

III. Methodology

This research was a quantitative study, using machine learning to analyze and predict the cost of medical insurance. It will follow the standard pipeline in data science: data collection, preprocessing, model training and evaluation, and testing.

A. Data Collection Method and load the dataset

The features of the dataset included age, sex, BMI, number of children, smoker or not, region, and charges. This dataset is publicly available in the Kaggle [11]. To train the Machine Learning Model, the research team used Google Colab [12] because Colab provides a vast amount of features without any charges and Python is a programming language. As a first step, the Research team opened the Google Colab notebook and imported the necessary libraries for data visualization and preprocessing. Such as pandas, NumPy, matplotlib, and seaborn. After importing the libraries, the dataset was loaded, and print the first five rows from the dataset. (Fig 1)

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Fig. 1. first five rows of the dataset

B. Data Preprocessing

When it comes to talking about data preprocessing when training the Machine Learning Models, the first step should be Data preprocessing also known as data cleaning. In this process the research team checks the missing values, checks the data types of the features, handles the duplicated values, handles outliers, data visualization, converts categorical values into numeric values, etc.

C. Checks the datatypes

Having an idea about the data types of the features is important. When training the machine learning models, every feature should be in a numeric format (int, float). Because the machines cannot identify or deal with the categorical values such as objects. The chosen dataset has both numeric and categorical features (Table 1).

Columns	Datatypes
Age	int64
Sex	object
BMI	float64
Children	int64
smoker	object
Region	object
Charges	float64

TABLE I
DATATYPES BEFORE PREPROCESSING

D. Convert Categorical values into Numeric values

There are three columns in categorical format. Such as sex, smoker, and region. Before the training of the machine learning model, the research team converted these three categorical columns into numeric columns by using the encoding technique called Label Encoding.

E. Handling Outliers

Especially in regression-based model training, handling outliers is very important. Because the outliers directly affect to the model's performances like accuracy and efficiency. When machine learning model training with the outliers the accuracy of the model can be less and the predictions can be wrong. Therefore, the research team handled the outliers by using the method called Inter Quartile Range Method (IQR). When checking the dataset, the research team can identify the outliers in one column called BMI (Fig 2 and 3). Boxplot is used to identify the outliers.

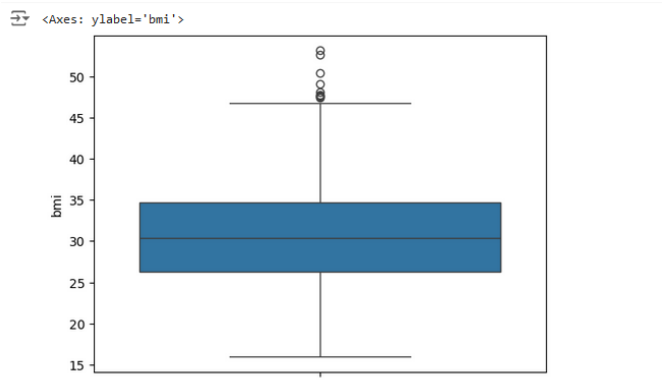

Fig. 2. before handling the outliers in BMI

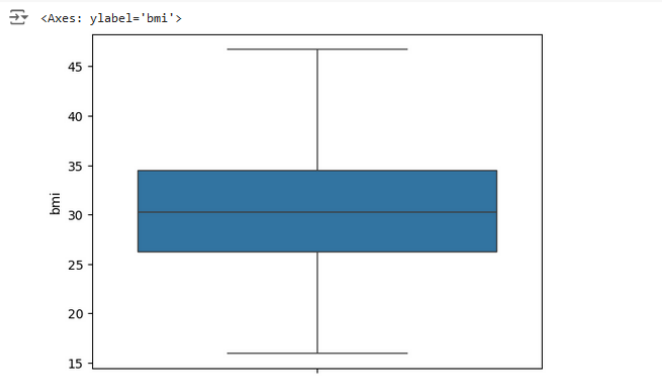

Fig. 3. after handling the outliers in BMI

F. Data Visualization

Complex relationships between the data points can simply be identified by using data visualization. The famous two data visualization libraries used by the research team. Such as matplotlib and seaborn. by using these two libraries the research team creates histograms, distplots, boxplots, distplots, bar charts, pie charts, count plots, co-relation heatmaps, etc.

G. Model Selection and Training

Before training the Machine Learning model, the appropriate model should be chosen. As a first step, the whole data set is split into a training dataset and a testing dataset. For the training dataset, the team used 80% and for the testing dataset used 20%. The research team used eight machine-learning models. Such as Linear Regression, Random Forest Regressor, Decision Tree Regressor, Support Vector Regressor (SVR), Ridge Regressor, K-Neighbors Regressor, Gradient Boosting Regressor, and XGB Regressor. After calculating each model's performance, the research team chose the **Gradient Boosting Regressor** as the best model. To identify the performance of the models the research team used the R^2 score, Mean Squared

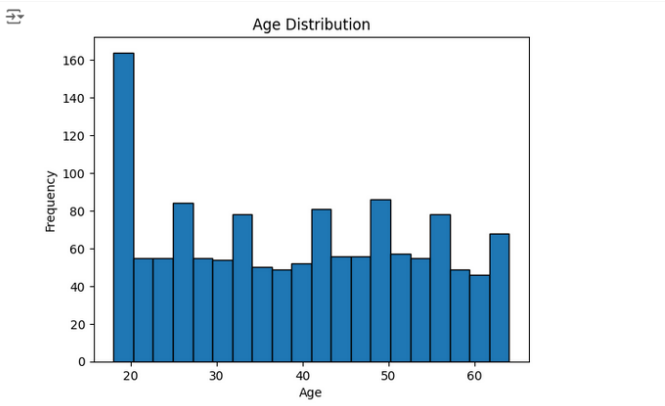

Fig. 4. histogram of age distribution

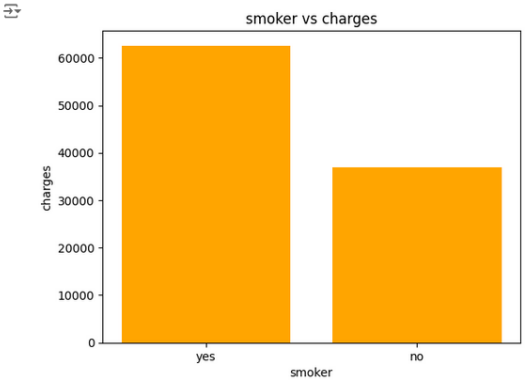

Fig. 5. histogram of smoker vs charges

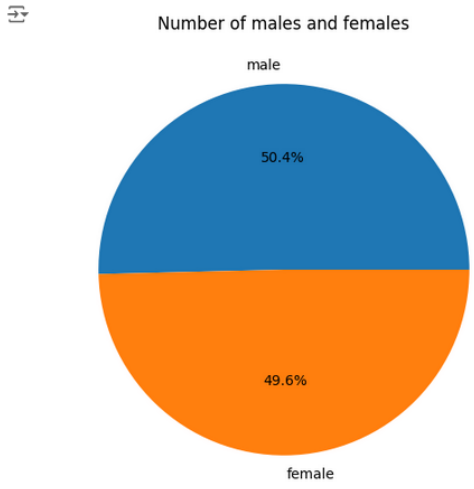

Fig. 6. pie chart of number of males and females

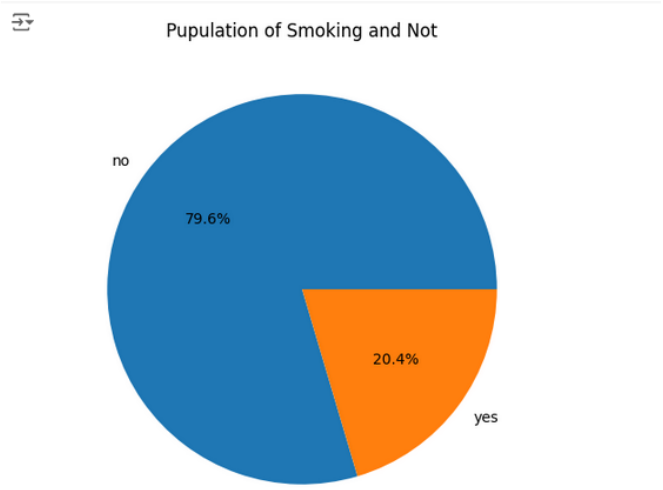

Fig. 7. pie chart of smoking or not

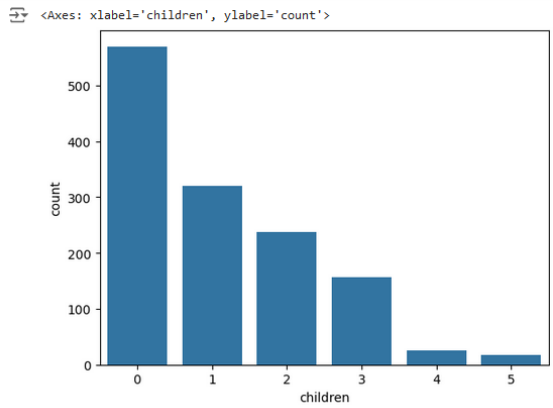

Fig. 8. count plot of number of children

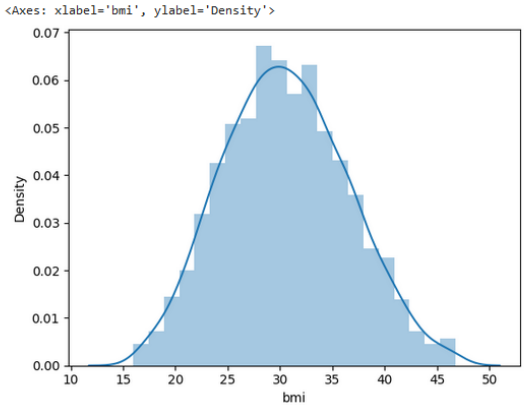

Fig. 9. distplot of BMI

Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) for cross-validation. After selecting the best performance model, the model training process began and to increase the model performance and accuracy used a technique called **Hyperparameter Tuning**.

H. Deploy to the Streamlit cloud

Streamlit [13] is an open-source framework used to design and share web applications, especially for machine learning and data science projects. It enables developers to create user-friendly interfaces and users can visualize data, images, and also users can predict the predictions. Before deploying the Models to the streamlit cloud [14], the research team created an environment using the Anaconda CMD on their laptop. To create the environment, import necessary libraries like Streamlit, scikit-learn, etc. After that using the Python language, the research team created the Streamlit application and deployed it [15] to the Streamlit cloud via GitHub [16]. The deployed Streamlit application is shown in Fig 10 and 11.

Fig. 10. Streamlit Application

Fig. 11. Streamlit Application

IV. Results

A. Presentation of Findings

This section will summarize the performance of each model concerning prediction results for the cost of medical insurance. This research used eight models, to identify each model's performance the research team used evaluation metrics such as R2 score, Mean Square Error (MSE), Mean Absolute Error (MAE), and RMSE (Root Mean Square Error). The summarization of each model performance is shown in Table II.

According to the performances of each model, the research team selected the Gradient Boost Regressor as the best model. Because when comparing other model's performances, the research team identifies the Gradient Boost Regressor to perform well in both training and testing data. After that, the research team apply the K-Fold cross-validation technique to confirm the model selection. The results of K-Fold cross-validation are provided in RMSE (Table III).

After the cross-validation, the research team could confirm the Gradient Boost Regressor model is the best model among the other seven models.

After selecting the best performance model, the research team began the model training and Hyperparameter tuning. The technique called Hyperparameter Tuning is used to increase the accuracy of the model. Table IV represents the model accuracy, before and after the Hyperparameter Tuning.

- Co-Relation Heatmap A co-relation heatmap is used to identify the relationships between the features or columns. By using this heatmap anyone can get an idea about the important features and not important features (Fig 12). To draw the heatmap the research team used the Seaborn visualization library.

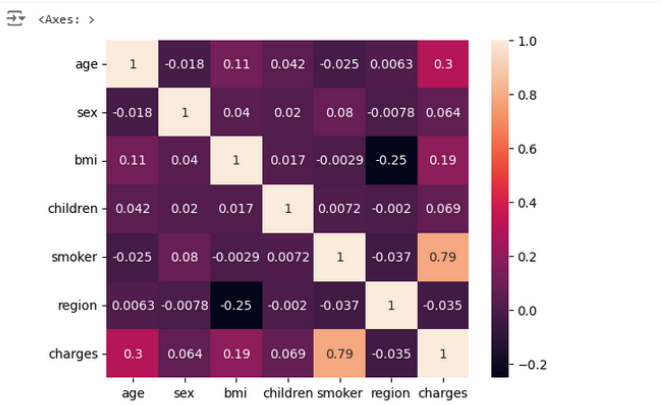

Fig. 12. co-relation heatmap

B. Data Analysis and Interpretation

Then, after training the models, the feature importance of each model will be reviewed. For example, it may be observed that smoker, BMI, age, and children are more influential in predicting costs than other factors like sex, and region (Fig 13).

Models	Train MSE	Train MAE	Train R2	Test MSE	Test MAE	Test R2
Linear Regression	36238736.33	4169.83	0.75	35146101.14	4077.28	0.75
Random Forest Regression	3420686.90	984.89	0.98	27092683.18	2929.50	0.81
Decision Tree	172501.05	18.02	1.00	56627755.43	3571.05	0.59
Support Vector	158631534.47	8296.85	-0.09	151782751.69	8089.46	-0.09
Ridge Regression	36246118.97	4164.85	0.75	35072121.67	4064.85	0.75
K-Neighbors	19903939.08	2639.50	0.86	33049365.18	3501.26	0.76
Gradient Boost	13488925.20	1982.43	0.91	22192338.17	2611.59	0.84
XGBoost	522099.93	387.42	1.00	27563828.95	2990.52	0.80

TABLE II
MODEL PERFORMANCE METRICS

Models	RMSE
Linear Regression	6078.01
Random Forest Regressor	4809.28
Decision Tree Regressor	6489.26
Support Vector Regressor (SVR)	12610.14
Ridge Regressor	6078.18
K-Neighbors Regressor	5650.44
Gradient Boosting Regressor	4553.34
XGB Regressor	5170.18

TABLE III
CROSS-VALIDATION RESULTS

Metric	Before Hyperparameter Tuning	After Hyperparameter Tuning
R2 Score (Test data)	84%	86%
MSE (Test data)	22192338.17	20148914.03

TABLE IV
MODEL PERFORMANCE BEFORE AND AFTER HYPERPARAMETER TUNING

According to the scatter plot (Fig 14), the model performed well below the $20000.

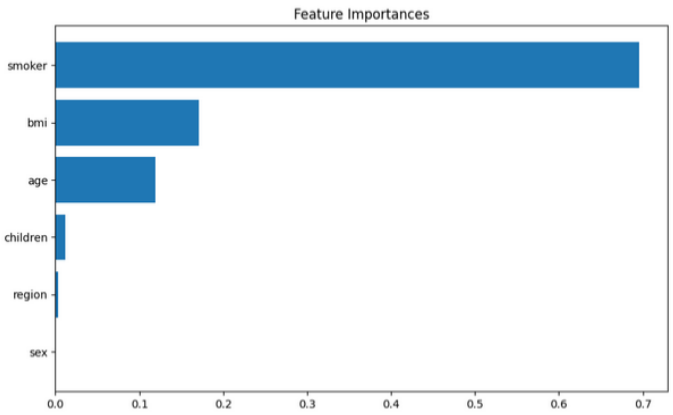

Fig. 13. feature selection

C. Support for Research Questions or Hypotheses

All three research questions were addressed through this research. According to the evaluation metrics, the best model was the Gradient Boost Regressor. This model perfume is well compared to the other seven models. But in some data, this model did not perform well like predictions above $20000. The reason behind that is the dataset was small and there are few features used as input data.

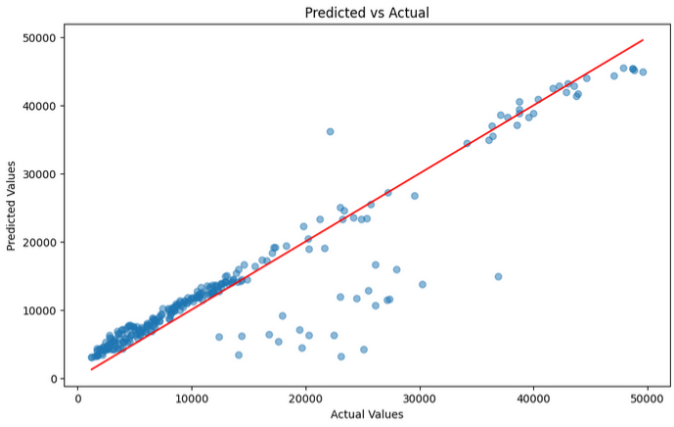

Fig. 14. scatter plot

V. DISCUSSION

A. Interpretation of Results

These results mean that, once the hyperparameters are optimized, Gradient Boosting Regressor is the top model on the prediction of medical insurance costs, with an R-squared

score of 86%. From these, the model will be reliable in making the predictions since it explains a large portion of the variance in the insurance premiums. The features that were important and identified include age, number of children, BMI, and whether or not one smokes; these suggest that they are strong drives of medical costs. That was expected since these factors have conventionally been considered major in determining health insurance premiums.

Yet, there were some limitations regarding this model' performance, especially in the forecasting of expenses greater than $20,000. About this, it could reflect a characteristic of the distribution of the dataset, where greater charges were fewer, which in turn would have an impact on the generalization of the model for those extreme values.

B. Comparison with Existing Literature

The findings are supportive of existing literature that highlights machine learning algorithms, especially ensemble methods such as Gradient Boosting, as prospective solutions to forecast health care-related expenditures. Previous studies have documented the efficiency of machine learning methods applied to health economics, often referring to and highlighting their advantages over traditional statistical models. This might hint that it's much better performance, w.r.t. simpler models, such as a linear regression or decision trees also evaluated in this paper, can be explained by the capability of a gradient boosting regression to cope with complex feature interactions.

C. Implications and Limitations of the Study

This also means a lot to large stakeholders in the health industry, like insurers and policy makers. More accurate projections of insurance costs help to price better while considering risks, hence making the delivery of care more effective.

Yet, this is not without its limitations. While comprehensive, this dataset may not include all the possible independent variables that can influence the cost of medical insurance, such as particular medical history or socioeconomic variables. Generalization also might be limited in view of the fact that, for this research study, one has to rely on just one data set. This work is to be verified in diversified data sets and further investigations with other available variables that lead to improved predictive accuracy.

VI. Conclusion

A. Summary of Key Findings

In summary, this research has successfully shown the usage of machine learning in predicting medical insurance costs with Gradient Boosting regressor. The reliability of the model was maintained using its performance metrics and analysis output from Feature Importance gives us some interesting leads into what part of insurance charges are driving it.

B. Contributions to the Field

The present research contributes to the emergent field of predictive modeling in healthcare by demonstrating eight machine-learning techniques that can be effective for estimating the cost of insurance. This also develops the feature selection process, which enhances the accuracy of the model.

C. Recommendations for Future Research

This could include larger and more varied datasets, trying deep learning on harder prediction problems, or examining the impact of new input variables such as medical history or genetic predispositions

References

[1] M. Kulkarni, D. D. Meshram, B. Patil, R. More, M. Sharma, and P. Patange, "Medical Insurance Cost Prediction using Machine Learning," Int. J. Res. Appl. Sci. Eng. Technol., vol. 10, no. 12, pp. 449–456, Dec. 2022, doi: 10.22214/ijraset.2022.47923.

[2] S. Hossen, "Medical Insurance Cost Prediction using Machine Learning," Int. J. Res. Appl. Sci. Eng. Technol., vol. 10, no. 12, pp. 449–456, Dec. 2022, doi: 10.22214/ijraset.2022.47923.

[3] Department of Computer Science, St. Albert's College, Kochi (Kerala), India., Sabarinath U S, A. Mathew, and Department of Computer Science, St. Albert's College, Kochi (Kerala), India., "Medical Insurance Cost Prediction," Indian J. Data Commun. Netw., vol. 4, no. 4, pp. 1–4, Jun. 2024, doi: 10.54105/ijdcn.D5037.04040624.

[4] S. Albalawi, "Prediction of healthcare insurance costs," 2023.

[5] K. Kaushik, A. Bhardwaj, A. D. Dwivedi, and R. Singh, "Machine Learning-Based Regression Framework to Predict Health Insurance Premiums," Int. J. Environ. Res. Public. Health, vol. 19, no. 13, p. 7898, Jun. 2022, doi: 10.3390/ijerph19137898.

[6] U. Orji and E. Ukwandu, "Machine learning for an explainable cost prediction of medical insurance," Mach. Learn. Appl., vol. 15, p. 100516, Mar. 2024, doi: 10.1016/j.mlwa.2023.100516.

[7] C. Shreekar, M. Kiran, D. Sumanth, and P. Jeevan, "Cost Prediction of Health Insurance," vol. 10, no. 01, 2023.

[8] V. Reddy, "Medical Health Insurance Cost Prediction." IJMRSET. [Online]. Available: DOI:10.15680/IJMRSET.2024.0704145

[9] Ch. A. Ul Hassan, J. Iqbal, S. Hussain, H. AlSalman, M. A. A. Mosleh, and S. Sajid Ullah, "A Computational Intelligence Approach for Predicting Medical Insurance Cost," Math. Probl. Eng., vol. 2021, pp. 1–13, Dec. 2021, doi: 10.1155/2021/1162553.

[10] Md Mohtaseem Billa, "Medical Insurance Price Prediction Using Machine Learning," J. Electr. Syst., vol. 20, no. 7s, pp. 2270–2279, May 2024, doi: 10.52783/jes.3962.

[11] "Medical Cost Personal Datasets." Accessed: Oct. 15, 2024. [Online]. Available: https://www.kaggle.com/datasets/mirichoi0218/insurance

[12] "Welcome To Colab - Colab." Accessed: Oct. 21, 2024. [Online]. Available: https://colab.research.google.com/

[13] “Streamlit Docs.” Accessed: Oct. 23, 2024. [Online]. Available: https://docs.streamlit.io/

[14] “Streamlit Community Cloud • Streamlit.” Accessed: Oct. 23, 2024. [Online]. Available: https://streamlit.io/[...id]

[15] “Streamlit.” Accessed: Oct. 23, 2024. [Online]. Available: https://medical-insurance-cost-prediction-bil9m4wb5rtxtos4z43oru.streamlit.app/

[16] C. Thenuka, Thenuka09/Medical-Insurance-Cost-Prediction. (Oct. 23, 2024). Jupyter Notebook. Accessed: Oct. 23, 2024. [Online]. Available: https://github.com/Thenuka09/Medical-Insurance-Cost-Prediction