Health Insurance Cost Prediction Using Machine Learning

Sathyam Reddy Mothe
Asst.Professor
Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopet, Andhra Pradesh
sathyamreddym@gmail.com

Anjali Rudrajada
Student
Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopet, Andhra Pradesh
rudrajadaanjali@gmail.com

Devisri Jangala
Student
Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopet, Andhra Pradesh
jangaladevi10@gmail.com

Thanusha Batchu
Student
Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopet, Andhra Pradesh
thanushabatchu9703@gmail.com

***Abstract*— This study presents a data-driven approach to predict health insurance costs, leveraging machine learning techniques on comprehensive healthcare datasets. The rising complexity and variability in healthcare expenses necessitate accurate prediction models to assist insurers, policyholders, and healthcare providers in estimating future costs and making informed decisions. The proposed model incorporates diverse features such as demographics, medical history, lifestyle factors, and regional healthcare trends to capture the multifaceted determinants of healthcare expenses. We used the 1338-entry medical cost personal dataset from Kaggle for the United States. The age, gender, bmi, smoking habit, number of children, and other features of the dataset are used to forecast the cost of insurance. To improve the model's prediction ability, feature engineering approaches are utilized to derive significant insights from unprocessed data. The best model for predicting health insurance costs is determined by evaluating a number of machine learning methods, such as gradient boosting, random forest, decision trees, and regression. Appropriate assessment criteria, including mean absolute error (MAE), mean squared error (MSE), root mean square error (RSME), and R-squared, are used to evaluate each model's performance. We used a 70-30 split to train the system, and the accuracy we obtained was 88.9%.**

Keywords—Health Insurance, Decision Tree, Regression, Random Forest, Cost Prediction, Gradient Boosting and Accuracy

I. INTRODUCTION

Healthcare costs continue to escalate globallydriven by variables including the rising incidence of chronic illnesses, demographic changes, and technology breakthroughs. Among the components of healthcare expenditure, health insurance costs represent a significant financial burden for individuals, families, and healthcare systems alike. Predicting these costs accurately is paramount for insurers, policymakers, and healthcare providers to allocate resources effectively, set premiums fairly, and ensure the sustainability of healthcare financing mechanisms.

Traditional approaches to estimating health insurance costs often rely on actuarial techniques and historical data analysis [1]. While these methods provide valuable insights, they may overlook the intricate interplay of diverse factors influencing healthcare expenses, such as individual health profiles, lifestyle choices, and regional healthcare dynamics. Moreover, the dynamic nature of healthcare systems necessitates adaptive and data-driven approaches to capture evolving trends and mitigate risks effectively.

In recent years, the proliferation of electronic health records, advances in data analytics, and the availability of large-scale healthcare datasets have paved the way for more sophisticated predictive modeling techniques [2].

For deciphering intricate healthcare data and finding trends that conventional approaches can miss. By leveraging machine learning algorithms, researchers and practitioners can develop predictive models capable of accurately forecasting health insurance costs, thereby enabling proactive decision-making and targeted interventions.

This study aims to contribute to the field of healthcare economics by presenting a data-driven approach to predict health insurance costs. The main objective is to create a forecasting tool that encompasses the various factors that influence healthcare spending. Such a model holds the potential to enhance risk assessment, improve financial planning, and inform policy formulation in the realm of health insurance.

In this paper, the methodology employed to construct the predictive model, including data preprocessing, feature selection, model selection, and evaluation metrics. We also discuss the implications of our findings for insurers, policymakers, and healthcare stakeholders, emphasizing the importance of leveraging predictive analytics to address the challenges of rising healthcare costs and promote access to affordable and equitable healthcare services.

The remainder of the work is put in order into sections. The review of literature survey is covered in the Section II. Section III includes a description of the dataset as well as specifics on its main properties. We have talked about the suggested System in Section IV. In the final section, V, we conclude the paper.

II. LITERATURE SURVEY

The prediction of health insurance costs has attracted a lot of attention lately, and academics are using a variety of data-driven approaches and machine learning strategies to improve knowledge and accuracy in this field.

- To shed light on the complexities of cost estimates in healthcare contexts, Abdelmoula (2021) [3] developed a machine learning-based prediction tool specifically for hospitalization expenses.
- Building on this framework, Baro (2022) [5] concentrated on forecasting hospitalization episodes, making a valuable contribution to enhanced risk assessment techniques that are essential for healthcare planning and resource distribution.
- Parallel to this, Subroto et al (2022) [5] investigated the use of tree-based algorithms to forecast the readiness of informal workers to pay for the national health insurances in order to identify personal preferences and behaviours with relation to the participation of health insurance.
- Furthermore, research on cost analysis of medical insurance and prediction using ML was conducted by Thejeshwar (2023) [6]. This research improved our knowledge of the variables affecting medical insurance costs and enabled better informed decision-making.
- Moreover, Vijayalakshmi (2023) [7] put into practice on price prediction system of medical insurance that used algorithms related to regression models to improve cost estimation accuracy. This helped policymakers and insurers select suitable coverage plans and rates.

This review of the literature emphasizes the variety of techniques and strategies used to forecast health insurance costs, highlighting the significance of data-driven insights in guaranteeing effective and fair funding and delivery of healthcare.

To tackle this issue, we have reviewed the whole body of research and employed an out-of-sample methodology. The paper's major contribution is its use of gradient boosting to achieve 88.9% accuracy. Additionally, the thorough graphical analysis of the outcome is covered in Section IV.

III. DATASET DESCRIPTION

In this Project, the utilized dataset is from Kaggle website [8] through the internet to complete the duty of insurance forecast. There were 1338 observations made about insurance prices throughout four US areas. Table 1 provides a thorough examination of the dataset.

The dataset was transformed into factors with numerical values desginated at each level, containing four numerical characteristics (charges, age, children and age) and three categorical characteristics (sex, region, and smoker).

Column Descriptions:
• Age: The age of the principal beneficiary
• Sex: The gender of the major beneficiary
• BMI: body mass index, which gives information on weights that are comparatively high or low in relation to height.
• Children: Total number of dependents / number of children with health insurance
• Smoker: Does one smoke?
• Region: Area of residence of the beneficiary in the United States (northwest, southwest, southeast, and northeast)
• Charges: Individual medical bills sent by health insurance.

A few dataset items are also displayed in below figure 1. The columns also explain themselves. Table 1 provides each column's specifics.

	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. Dataset that was utilized

TABLE I. DATASET CHARACTERISTICS

Sr. No.	Name of Variable	Type (Input/Output)	Details
1.	Age	Input	Range: 18 to 64 years, Mean value is 39.2
2.	Gender	Input	Female:662 and male: 676
3.	BMI	Input	Body mass index (BMI) in kg/m2 min. value: 15.96; max. value: 53.13; mean value: 30.66
4.	Smoking Habit	Input	Smokers: 1064 and no-smokers: 274
5.	Beneficiary's residential area	Input	In USA: southeast: 364 Northeast: 324 Southwest: 325 Northwest: 325
6.	Children	Input	Range: 0 to 5 years, Mean value is 1.095
7.	Insurance charges	Output	In $, min. value: $ 1122; max. value: $63770; mean value:$13270

IV. PROPOSED SYSTEM

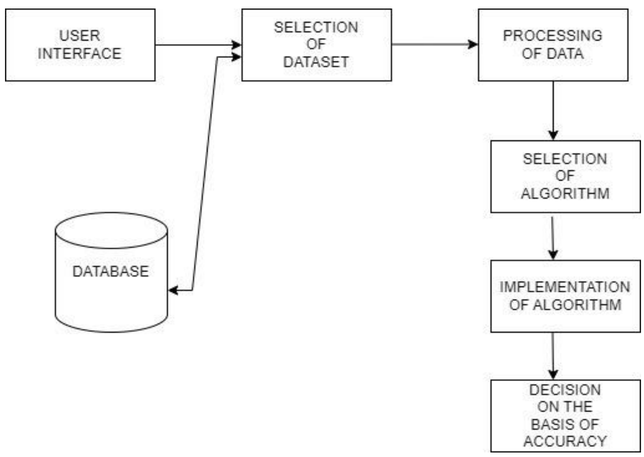

Fig. 2. Steps involved in a model

The first step involves the user interacting with the database system through a user interface (UI). This UI could be a web application, a desktop application, or even a command line interface (CLI). Figure 14 represents the user interaction html form, which contains the some fields that has to be filled by the user.

The second step involves selecting a dataset. This dataset is the collection of data that the database system will use to make predictions. In the case of a medical cost insurance prediction system, the dataset might include information about patients' medical history, demographics, and insurance claims.

The third step involves selecting an algorithm. Algorithm is a collection of guidelines that the database system will use to process the data and make predictions. In the case of a medical cost insurance prediction system, the algorithm might be a ML model that trained on a large dataset of medical insurance data.

The fourth step involves implementing the algorithm. This involves running the algorithm on the selected dataset. In the case of a medical cost insurance prediction system, this would involve feeding the patients' data into the machine learning model and generating predictions about their future medical costs.

The fifth step involves making a decision on the basis of the accuracy. This step is likely referring to the process of evaluating the accuracy of the predictions made by the algorithm. In the case of a medical cost insurance prediction system, this might involve comparing the model's predictions to actual medical costs incurred by patients. This Whole process is represented in figure 2.

Our Model is Proposed based on certain criteria as follows

- ➢ Data Visualization
- ➢ Preprocessing Techniques
- ➢ Model Creation and Evaluation
- ➢ Accuracy

A. Data Visualization

In the health insurance cost prediction project, a graphical representation serves as a visual aid to understand and Data Visualization [9] of the dataset attributes such as sex, age, children, region, BMI, and charges. This graphical representation could take the form of various types of plots and charts to provide a comprehensive understanding of the data.

- *Boxplot*: A boxplot [10], also referred to as a box-and-whisker plot, is a visual depiction of a dataset's distribution. Through the use of quartiles, it illustrates the distribution of numerical data, emphasizing any potential outliers as well as the central tendency and variability. A box that covers the data's interquartile range (IQR) and a line that represents the median make up a boxplot. The data range is represented by whiskers that extend outside the box, with the exception of outliers, which are depicted as individual points. Boxplots are helpful for displaying any extreme values as well as the data's skewness and dispersion. We started by creating a boxplot to identify any outliers in the 'age', 'bmi', 'children' and 'charges' attributes. This allowed us to understand the distribution of these variables and identify any extreme values that might impact our analysis. In figure 3, for the age attribute a boxplot is created and identified that it doesn't have any outliers present in it. In figure 4, the boxplot is established for the attribute bmi, representing a clear visualization. In this case, the y-axis displays the distribution of frequency, while the x-axis indicates the data to be shown.

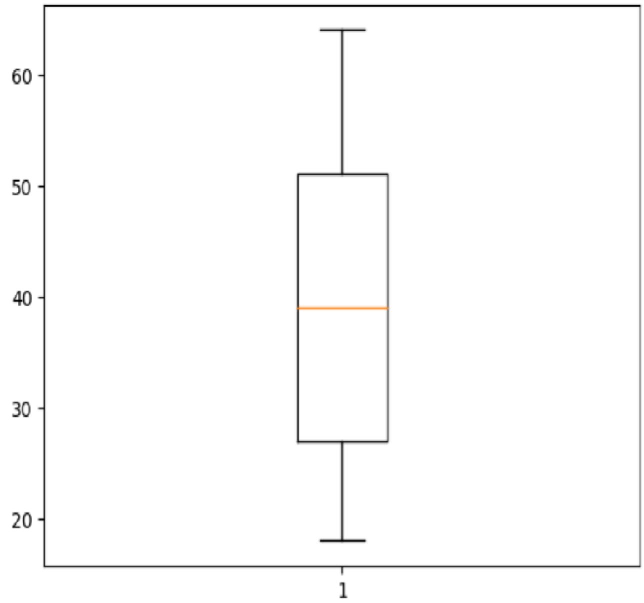

Fig. 3. Boxplot for the Attribute Age

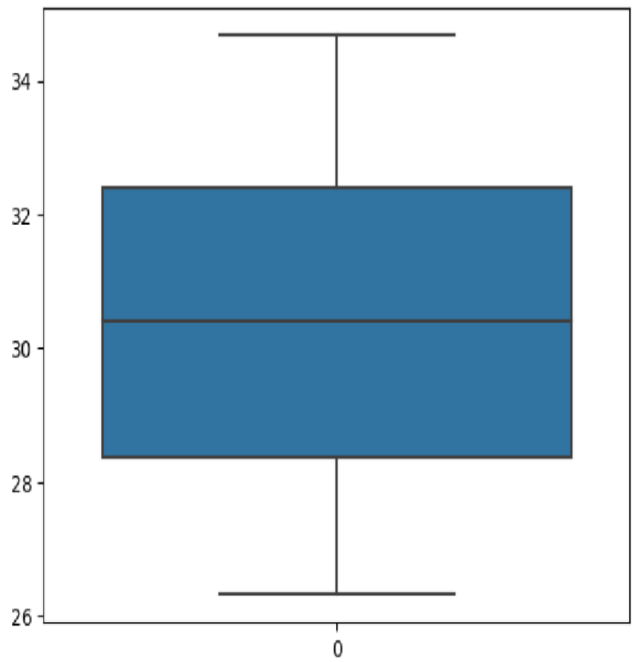

Fig. 4. Boxplot for the Attribute bmi

- *Displot:* A displot [11], short for distribution plot, is a type of visualization that displays the dataset attributes data distribution. It provides a visual representation of the frequency (or density) of values within the dataset across different intervals or bins. Displots are often used to visualize the shape of the distribution, including its central tendency, spread, skewness, and presence of multiple peaks. Common types of displots include histograms, kernel density estimation (KDE) plots, and rug plots. Displots are helpful for figuring out the data's underlying distribution and spotting trends or abnormalities. For the attributes like ‘age’,’bmi’,’children’ and ‘charges’ we utilized displot (distribution plot) to visualize their distributions and understand their impact on insurance charges. Observe the example of how the data has been distributed by children and charges attributes in figure 5,6. Where the y-axis denotes the count of the data distributed.

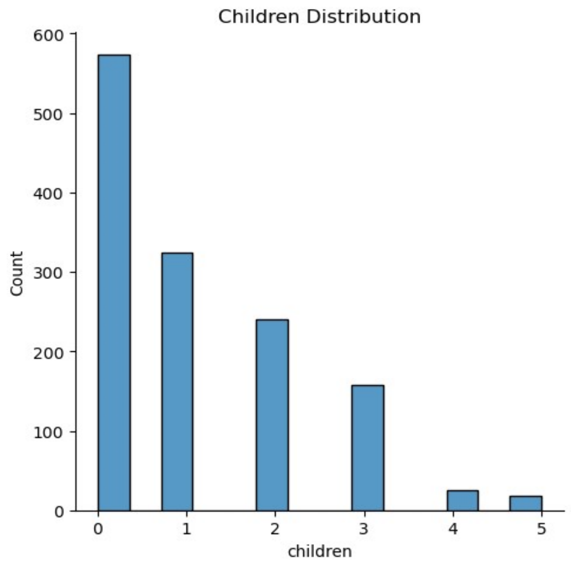

Fig. 5. Distributed data of children attribute in the Dataset

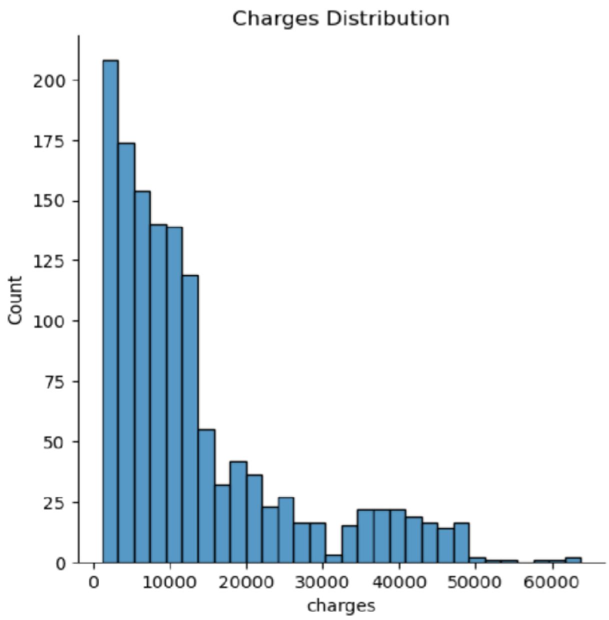

Fig. 6. Distribution of charges data

- *Countplot:* One kind of bar plot that shows the number of observations in each category of a categorical variable is called a countplot [12]. It provides a visual representation of the frequency of different categories within the dataset, allowing for easy comparison between groups. Countplots are particularly used for visualizing the data distribution of categorical attributes and detects any imbalances or patterns within the data. They can be customized with additional parameters to display proportions or percentages instead of raw counts, and to arrange the bars in a specific order based on the frequency or other criteria. Exploratory data analysis frequently uses countplots to understand the distribution of categorical variables and how they relate to other variables. Countplots are to visualize the distribution of 'sex', 'region', and 'smoker' within the dataset. This helped us to understand the gender distribution, regional distribution, and the proportion of smokers in our dataset.

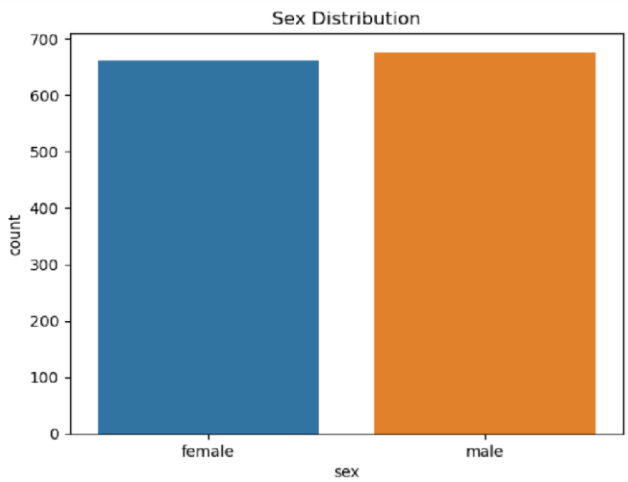

Fig. 7. Countplot for the sex data in the Dataset

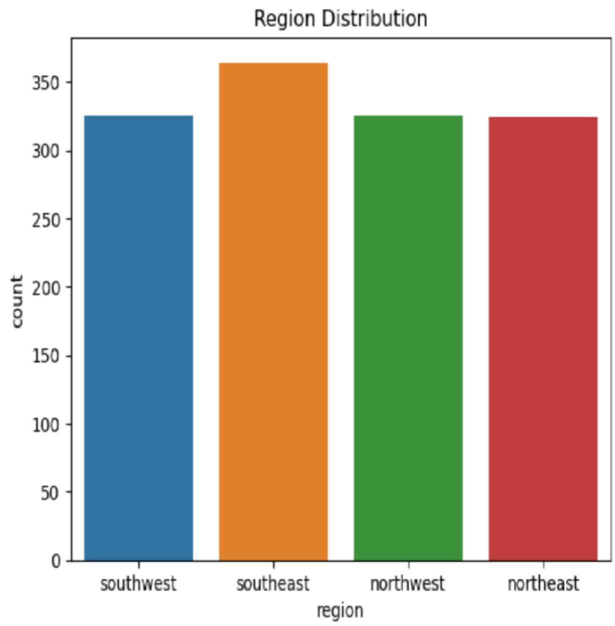

Fig. 8. Countplot for the region data in the Dataset

This Project is able to fully comprehend our information and spot any patterns or trends that can provide guidance for the health insurance cost prediction model by utilizing these visualization approaches. In figure 9, there is a detailed visualization information about the correlation between the attributes and the charges in the dataset. Using a correlation matrix, which can be observed in figure 11. We can assess how two variables are related to one another: A relationship is strong if the relationship value is 1. The connection is neutral if the relationship value is 0. In the event that the relationship's value is -1, it is weak or negative.

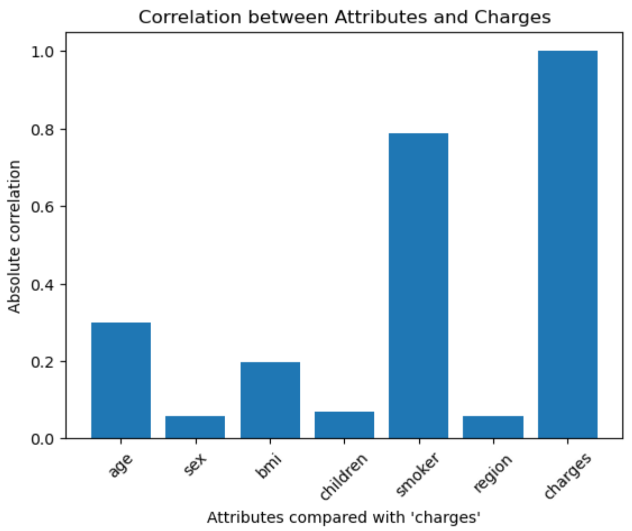

Fig. 9. Correlation between the Attributes in the Dataset

B. Preprocessing Techniques

The changes we make to our data before supplying it to the algorithm are known as pre-processing. Since the dataset we're using has a lot of instances, there will inevitably be some noise and outliers, which might have an impact on the model's performance. Therefore, data preprocessing [13] is used to prepare the process for fitting the data into a mode. Pre-processing refers to steps like replacing null values which is present in our data, those null values can be replaced with mean, mode and median another method for removing the noise is using clustering.

There is another technique called OneHotEncoder [14] which is also used in preprocessing of data. Encoding in Machine Learning helps you convert the categorical data into the numerical data as most of the models works better with numerical data also column transformer which converts the categorical and numerical seperately. Figure 1, explains how the data is presented in the dataset before performing the encoding technique. Figure 10, explains how the categorical data converted into numerical data and the shows how the data in the dataset is converted after performing the encoding.

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

Fig. 10. Dataset after performing encoding.

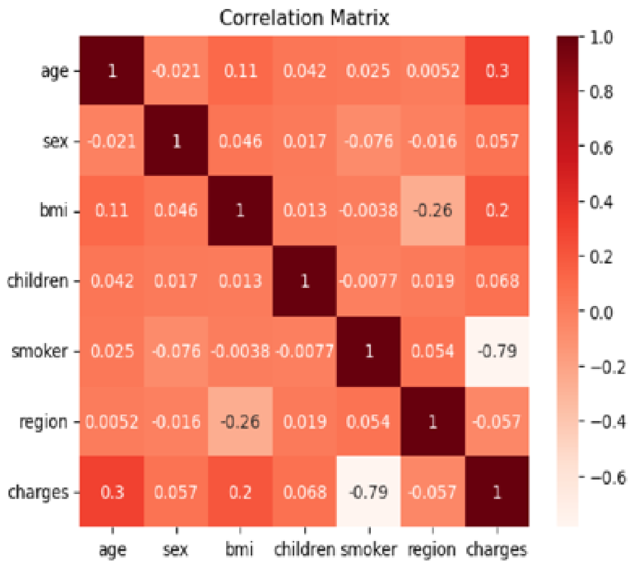

Fig. 11. Correlation Matrix

C. Model Creation and Evaluation

In this step, Using the prepared, an appropriate machine learning algorithm is chosen and instructed. In order to reduce the discrepancy between the model's anticipated output and the actual output found in the training set, the parameters are optimized. The model is tested on a different validation dataset to gauge its performance after training. The performance criteria and problem type determine the assessment measures that are employed. To evaluate the performance of the model, divide the data into 70-30 training and testing sets. Utilizing the training data, train the chosen model. Utilize techniques like cross-validation to ensure the model's generalizability.

Regression evaluation metrics [15] include Accuracy, K-Fold Cross-Validated [16] Accuracy etc. The simple-to-understand MAE stands for the average absolute difference between the actual and anticipated values. Larger mistakes are penalized more severely by MSE and RMSE than by MAE, which may be helpful depending on the demands of the prediction job. The percentage of the target variable's volatility that can be predicted from the independent variables is expressed as the R2 Score.

Based on the evaluation results, model can further refined by adapting its parameters or selecting a different algorithm. Visualize the predicted versus actual values to evaluate the accuracy of the model and find any trends or patterns. Perform additional analysis, such as feature importance, to understand which variables contribute most to the predictions. If the model's performance isn't up to pace, think about enhancing its features, modifying its hyperparameters, or fine-tuning its architecture. In particular, if the dataset is noisy or tiny, regularize the model to avoid overfitting.

The model's performance on recent, unknown data will be enhanced by doing this crucial step. The model may be used in a production setting to make predictions or judgments based on fresh data once it has been trained and assessed.

Overall, this process of creating and evaluating a machine learning model involves several steps that require careful consideration and attention to detail. By following these steps, it is possible to create models that can learn and make accurate predictions on new data.

Example Code:

```
from sklearn.ensemble import GradientBoostingRegressor
gradient_boosting= GradientBoostingRegressor()
mdl=gradient_boosting.fit(x_train, y_train)
y_pred = mdl.predict(x_test)
gradient_boosting_accuracy = random_forest.score(x_test, y_test)
mse = mean_squared_error(y_test, y_pred)
gradient_r2 = r2_score(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
gradient_rsme=np.sqrt(mse)
print(f"MSE: {mse}")
print(f"R-squared (R2): {gradient_r2}")
print(f"MAE: {mae}")
print(f"RSME:{gradient_rsme}")
kf = KFold(n_splits=5, shuffle=True, random_state=42)
print(f"Gradient Boosting Regression Accuracy: {gradient_boosting_accuracy*100}")
cv_gradient_reg= cross_val_score(GradientBoostingRegressor(), x, y, cv=kf)
cvgr=cv_gradient_reg.mean()
print(f"Cross-validated Gradient Boosting Regression Accuracy: {cvgr*100}")
plt.scatter(y_test, y_pred)
plt.xlabel("Actual Values")
plt.ylabel("Predicted Values")
plt.title("Gradient Boosting Regression: Actual Values vs Predicted Values")
plt.show()
```

Output for the above code:

MSE: 17619481.005488485

R-squared (R2): 0.8895118824573768

MAE: 2486.239183786422

RSME:4197.5565517915875

Gradient Boosting Regression Accuracy:

86.77526033602389

Cross-validated Gradient Boosting Regression Accuracy: 85.34540800483353

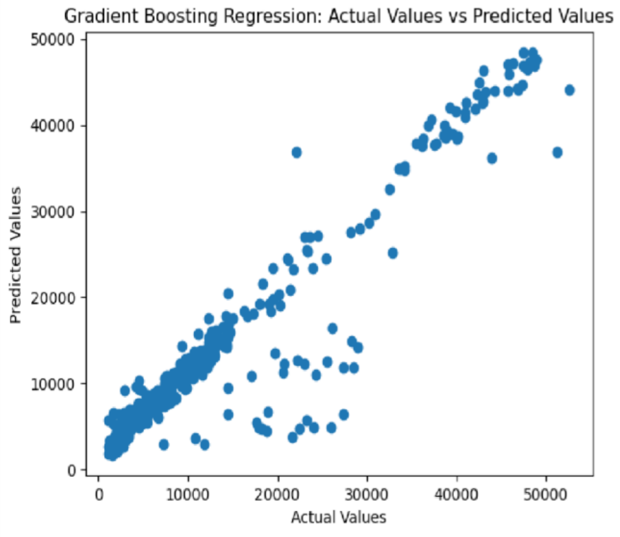

Fig. 12. Actual vs Predicted Values

In this project, the dataset is trained through many different algorithms to produce the better and accurate results. Among all the algorithms gradient boosting obtained highest accuracy of 88.9%. Performance of each model is mentioned in table 2.

TABLE 2. MODEL PERFORMANCE

Algorithm	MAE	RSME	Accuracy
Linear	4018.0	5776.5	0.7907
Polynomial	2831.9	4432.9	0.8767
Ridge	4032.6	5782.9	0.7902
Lasso	4018.6	5776.9	0.7907
Random Forest	2788.5	4683.4	0.8624
Decision Tree	3109.0	6776.4	0.7120
Gradient Boosting	2485.9	4197.6	0.8895

D. Accuracy

The ratio of the number of accurate forecasts to the total number of predictions produced is commonly used to quantify accuracy. It may be stated mathematically as follows in equation (1):

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100\% \quad (1)$$

A popular indicator for assessing a machine learning algorithm's performance is accuracy [17]. It calculates the percentage of each instance in the test dataset that is properly categorized. When it comes to predicting health insurance costs, accuracy is a crucial metric, but it is vital to consider the specific context and requirements of the prediction task. While accuracy is a common metric for classification tasks, for regression tasks like health insurance cost prediction, regression metrics are more appropriate. It's essential to interpret the chosen metric in the context of health insurance cost prediction. For example, a low MAE indicates the average of the model's predictions, which are close to the actual insurance costs. By using different models we obtain different accuracy, which can be observed in figure 13, and representing that the gradient boosting produce the highest accuracy of 88.9% comparing with the other models.

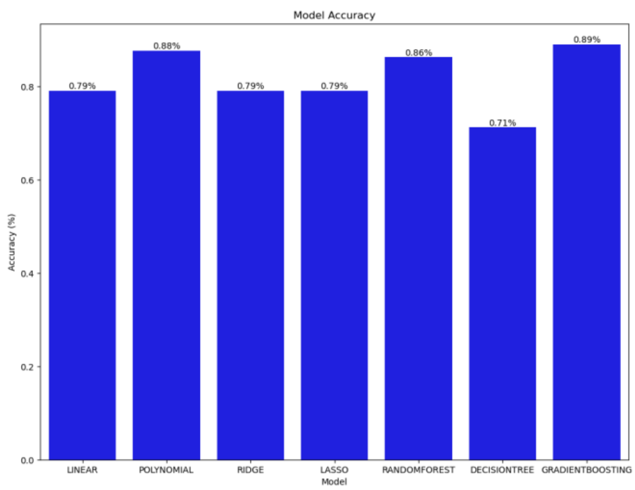

Fig. 13. Accuracy of different Algorithms through Visualization

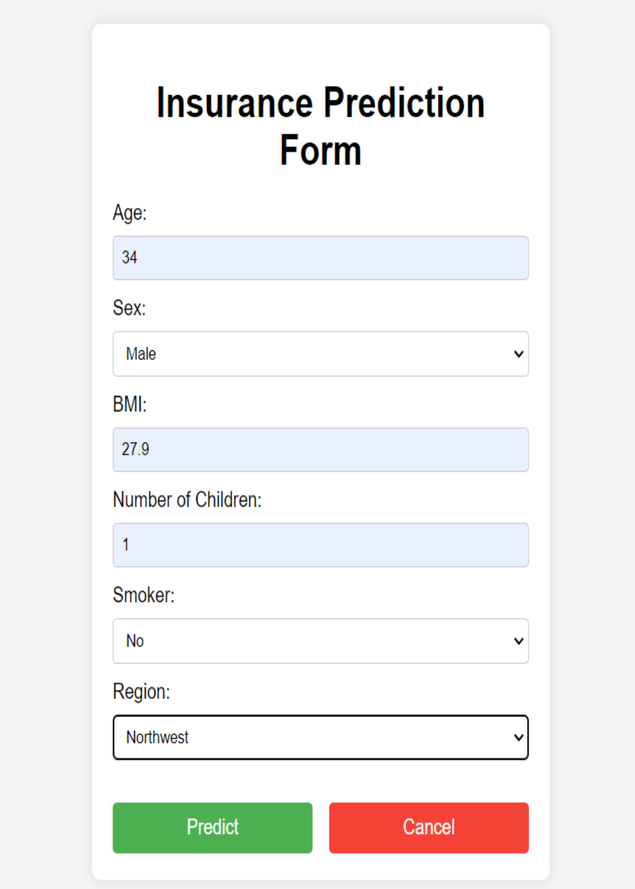

Fig. 14. Insurance Prediction Form

V. CONCLUSION

In conclusion, the use of machine learning to forecast health insurance costs has great potential to enhance the healthcare sector's decision-making, resource allocation, and general efficiency. Through the development and evaluation of predictive models, we can gain valuable insights into the factors influencing insurance costs and better tailor insurance plans to individual needs. Predictive models for health insurance cost estimation offer several benefits than the existing model, including:

- Improved accuracy in estimating individual insurance costs.
- Enhanced risk assessment, enabling insurance companies to better manage their portfolios.
- Personalized pricing and policy recommendations based on individual characteristics and risk factors.
- Optimization of resource allocation and strategic planning within insurance companies.
- In the existing proposed model [18], we have obtained 81.3% accuracy by using Linear Regression Algorithm [19]. But, In the proposed system, we have obtained 88.9% accuracy by using Gradient Boosting Algorithm.

TABLE 3. DIFFERENT MODEL ACCURACIES OF EXISTED VS PROPOSED SYSTEM

Algorithm	Existed Accuracy	Proposed Accuracy
Linear	0.7447	0.7907
Polynomial	--	0.8767
Ridge	0.7448	0.7902
Lasso	0.7447	0.7907
Random Forest	0.8371	0.8624
Decision Tree	0.7003	0.7120
Gradient Boosting	0.8679	0.8895

This study explored the application of popular ML algorithms, Decision Tree, Linear, Random Forest, Lasso, Gradient Boosting, Polynomial, and Ridge Regression for predicting health insurance cost prediction. The results obtained indicate that the algorithms can be effective in generating accurate predictions, with Gradient Boosting outperforming the remaining algorithms in terms of efficiency and accuracy. The predictive model developed in this study can be useful for insurers, healthcare providers, and policyholders alike. Overall, this project highlights the potential of machine learning in the health care industry and how it can enhance the accuracy and reliability of health insurance cost predictions. Future research can expand on this work by incorporating more complex features, exploring alternative algorithms, and analyzing the model's performance on a larger dataset.

VI. REFERENCES

[1] Kaushik K, Bhardwaj A, Dwivedi AD, Singh R. “Machine Learning-Based Regression Framework to Predict Health Insurance Premiums”. Int J Environ Res Public Health. 2022 Jun 28;19(13):7898. doi: 10.3390/ijerph19137898. PMID: 35805557; PMCID: PMC9265373.

[2] Ghosheh GO, Thwaites CL, Zhu T. “Synthesizing Electronic Health Records for Predictive Models in Low-Middle-Income Countries (LMICs)”. Biomedicines. 2023 Jun 18;11(6):1749. doi: 10.3390/biomedicines11061749. PMID: 37371844; PMCID: PMC10295936.

[3] Abdelmoula, B., Torjmen, M., & Abdelmoula, N. B. (2021). “Machine learning based prediction tool of hospitalization cost”. 2021 22nd International Arab Conference on Information Technology (ACIT). https://doi.org/10.1109/acit53391.2021.9677110.

[4] Baro, E. F., Oliveira, L. S., & de Souza Britto Junior, A. (2022). “Predicting hospitalization from health insurance data. 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)”. https://doi.org/10.1109/smc53654.2022.9945601.

[5] Mansur, R., & Subroto, A. (2022). “Using tree-based algorithm to predict informal workers’ willingness to pay national health insurance after Tele-Collection. 2022 10th International Conference on Information and Communication Technology (ICoICT)”. https://doi.org/10.1109/icoict55009.2022.9914901.

[6] T, Thejeshwar., T, S. Harsha., V, V. Krishna., & R, Kaladevi. (2023). “Medical Insurance Cost Analysis and prediction using machine learning. 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)”. https://doi.org/10.1109/icidca56705.2023.10100057.

[7] Vijayalakshmi, V., Selvakumar, A., & Panimalar, K. (2023). “Implementation of medical insurance price prediction system using regression algorithms. 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)”.https://doi.org/10.1109/icssit55814.2023.10060926 .

[8] Kaggle Medical Cost Personal Datasets. Kaggle Inc. https://www.kaggle.com/datasets/mirichoi0218/insurance.

[9] https://www.javatpoint.com/data-visualization-in-machine-learning

[10] https://www.machinelearningplus.com/plots/python-boxplot/

[11] https://www.digitalocean.com/community/tutorials/seaborn-distplot

[12] https://www.geeksforgeeks.org/countplot-using-seaborn-in-python/

[13]https://www.geeksforgeeks.org/data-preprocessing-machine-learning-python/

[14]John McCullagh implemented preprocessing techniques OneHotEncoder and replacing null values IC31 27-29 Nov 2015.

[15] https://www.geeksforgeeks.org/regression-metrics/

[16] Sireesha Moturi, S.N.Tirumala Rao, Srikanth Vemuru,"Grey wolf assisted dragonfly-based weighted rule generation for predicting heart disease and breast cancer, Computerized Medical Imaging and Graphics," Volume 91, 2021, 101936, ISSN 0895-6111,https://doi.org/10.1016/j.compmedimag.2021.101936.

[17] Baratloo A, Hosseini M, Negida A, El Ashal G." Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity. Emerg (Tehran)". 2015 Spring;3(2):48-9. PMID: 26495380; PMCID: PMC4614595.

[18] K. Bhatia, S. S. Gill, N. Kamboj, M. Kumar and R. K. Bhatia, "Health Insurance Cost Prediction using Machine Learning," 2022 3rd International Conference for Emerging Technology (INCET), Belgaum, India, 2022, pp. 1-5, doi: 10.1109/INCET54531.2022.9824201.

[19] M. S. Acharya, A. Armaan and A. S. Antony, "A Comparison of Regression Models for Prediction of Graduate Admissions," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 2019, pp. 1-5, doi: 10.1109/ICCIDS.2019.8862140.