Grenze International Journal of Engineering and Technology, June Issue

Forecasting Medical Insurance Costs: A Data-Driven Approach

Chandrashekhar H. Patil[1], Ashish Garud[2], Ajinkya Sonawane[3], Akash Gavhane[4] and Mayur Jadhav[5]
[1, 3-5]Dr. Vishwanath Karad MIT World Peace University / Department of Computer Science and Application, Pune, India
Email: chpatil.mca@gmail.com, ajusonawane07@gmail.com, akash1106gavhane@gmail.com, mayurjadhav3232@gmail.com
[2] MAEER'S MIT Arts, Commerce and Science College, Alandi(D), Pune, Maharashtra, India 412105
Email: aashish.garud@gmail.com

***Abstract*—This project's primary goal is to forecast medical insurance costs using SVM, XGBoost, Random Forest, and Linear Regression machine learning algorithms. The process involves creating a dataset from scratch that is similar to real health data, preparing it, applying various algorithms to it, and evaluating its performance. The Target variable in the dataset is medical insurance expenses, while other parameters include age, BMI, children, gender, smoking habits, region, blood pressure, cholesterol levels, and diabetes. Min-Max scaling is used to normalize numerical characteristics, whereas one-hot encoding is used for categorical variables. To optimize performance, the models are trained using the pre-processed data. Model evaluation uses evaluation metrics like R-squared and Mean Squared Error (MSE). Cross-validation assures consistency in model prediction. The results contain data extracted from best-fit lines illustrating the predictions generated by every program. The program provides an in-depth review of medical insurance cost forecasts for the benefit of healthcare and insurance decision-makers.**

***Index Terms*— Medical insurance, Machine Learning, BMI, Healthcare.**

I. INTRODUCTION

The ability to accurately predict medical insurance costs is crucial to effective financial planning and resource allocation in a time where healthcare factors are constantly changing. To predict medical insurance prices, this research uses four algorithms—Linear Regression, XGBoost, Random Forest, and SVM [1] —all related to machine learning. The first step is to start from scratch and generate a dataset that is intended to look like actual health data. The dataset serves as the foundation for further predictive modeling as it encompasses a wide range of characteristics, including age, BMI, children, gender, smoking habits, region, blood pressure, cholesterol, diabetes, and the final target variable is medical insurance costs. To improve the data's suitability for machine learning methods, one-hot encoding is used for categorical variables, and Min-Max scaling is used to normalize numerical features. This guarantees that the models can efficiently absorb knowledge from the wide range of data that is contained in the dataset. The application of machine learning algorithms to the preprocessed data is the central idea. Models learn about the intricate connections between an individual's specific health characteristics and the associated costs of insurance. In model evaluation, accuracy metrics like Mean Squared Error (MSE) and R-squared are utilized. They help in determining the predictive capability of the models. The

Grenze ID: 01.GIJET.10.2.559_1

models' reliability is enhanced by employing cross-validation techniques, ensuring their ability to generalize effectively to fresh data. The project's conclusions are communicated through insights derived from best-fit lines, which visually represent the predictions generated by each algorithm. The present study improves the field of medical insurance cost prediction and provides healthcare and insurance policyholders with significant perspectives to facilitate well-informed decision-making.

II. RELATED WORK

[2]This major contribution delves into the domain of predicting health insurance costs using machine learning. The authors utilize a Kaggle dataset featuring variables such as Age, Gender, BMI, Smoking Habit, and number of children. Through the use of linear regression, the study achieves an 81.3% accuracy in forecasting insurance expenses. This research aligns with broader trends in utilizing machine learning for healthcare decision-making. [3]This study employs three different algorithms on a Kaggle dataset that includes age, gender, BMI, smoking habits, and number of children: linear regression, decision tree regression, and gradient boosting regression. Utilizing Streamlite to navigate the field of artificial intelligence, the study achieves an astounding accuracy rate of 81.3%. The focus on healthcare prediction issues underscores the necessity of further research and development. This research demonstrates the potential of machine learning in healthcare decision-making and makes a significant contribution to emerging trends. [4]This research, presented at ICICC 2022, delves into the economic impact of healthcare expenses, constituting 30% of GDP. Led by Ajay Sahu and the team from GNIT, the study employs Random Forest Regression to forecast medical care costs, aiming to guide patients and aid policymakers. The exploration of gradient-boosted trees and Linear Regression underscores the importance of early cost estimation, preventing unnecessary expenditures. While not offering specific figures, the study provides a valuable understanding of potential health insurance costs, contributing to the broader discourse on healthcare expense prediction. [5]Jiayuan Wang and other co-authors investigate the impact of data preprocessing techniques on the performance of machine learning models in predicting medical insurance costs. The study evaluates the effectiveness of techniques such as Min-Max scaling and one-hot encoding, providing insights into best practices for dataset preparation. [6]Prediction has been introduced previously in medicine. Clinical predictions based on data are becoming commonplace in medicine. Ranging Risk categorization of patients in the critical care unit ranges from risk scores to anticoagulant treatment (CHADS2) and cholesterol medicine usage (ASCVD) (APACHE). You may easily develop prediction models for hundreds of similar clinical questions using clinical data sources and contemporary machine learning. These approaches might be used for everything from sepsis early warning systems to superhuman diagnostic imaging.The real data source, on the other hand, has an issue. Unlike traditional techniques, new data sources are sometimes unstructured because they were developed for various purposes (clinical care, billing, etc.). Patient self-selection, indication biases, and even racist programming in machine prediction. As a result of this understanding, and discussing the potential of data analysis to aid medical decision-making. [7]Identifying and managing patients most at risk within the healthcare system is vital for governments, hospitals, and health insurers but they use different metrics for identifying the patients they perceive to be at most risk. Hospitals focus on readmission rates and cumulative risk of death during hospitalization. Accurately predicting these indicators could assist in allocating limited resources and thus improve the hospital’s operational efficiency. Health insurers are mostly concerned with insurance risk because they agree to reimburse health-related services in exchange for a fixed monthly premium. Poor risk measures could result in exceeding a financial budget. Therefore, one of the most obvious goals for health insurers is to Various predictive models have been developed to identify high-risk customers by predicting Healthcare expenses. [8]Pavan and his co-authors delved into the fundamentals of the linear regression model, exploring its utility in forecasting charges and comparing anticipated versus actual outcomes for estimating medical expenses. They also advocated for a machine learning approach, employing regression techniques such as Linear Regression. Notably, their observations highlighted age and BMI as critical features influencing the dependent variable. Among their experiments, this particular model showcased superior performance. [9]The study employs various machine learning techniques to predict health insurance prices using specific information from a medical expenses dataset. The most effective method identified is Stochastic Gradient Boosting, boasting an accuracy of 85.82%, with an MAE of 0.17448 and an RMSE of 0.380189. Consequently, in the realm of insurance cost estimation, stochastic gradient boosting surpasses alternative approaches. Policymakers stand to gain from machine learning, as ML models can swiftly compute costs—a task that would be time-consuming for humans. This efficiency could potentially boost revenue for companies. Moreover, ML models excel at managing large datasets. [10]While linear regression and its regularizations showed acceptable precision through five-fold cross-validation, more complex algorithms like XGBoost, Random Forest, Simple Tree, and KNR

performed exceptionally well when optimal parameters were determined using GridsearchCV. Notably, Simple Tree demonstrated competitive computational efficiency with an R2 value of 0.88, making it a preferred choice in certain organizational contexts despite slightly lower accuracy compared to Random Forest and XGBoost. [11]Three new ensemble of supervised learning classifiers are presented in this research with the goal of controlling health insurance expenditures. An open dataset is used to create these ensembles in order to develop data analysis techniques. This dataset contains multiple implementations of weak predictors. Two of the four ensembles that were created are enhanced ensembles. It is observed that when compared to bagging, the boosting and stacking ensembles show better accuracy.

III. PROBLEM STATEMENT

The goal is to develop and implement a machine learning model for predicting health insurance costs according to personal characteristics. Age, BMI, number of children, gender, smoking habits, region, blood pressure, cholesterol, and diabetes status are among the variables included in the dataset. The main goal is to produce accurate insurance cost estimates by applying various machine learning techniques, including SVM, XGBoost, Random Forest, and Linear Regression.

IV. PROPOSED METHOD

A. Data Collection

Generate a comprehensive dataset by employing a simple Python code to create synthetic data with relevant features such as Age, BMI, Children, Gender, Smoker, Region, Cholesterol, Blood Pressure, Diabetes, and Insurance Cost. Subsequently, save this generated dataset into a CSV file for further utilization in the machine learning model development.

B. Data Preprocessing

To handle missing data, utilize imputation or removal techniques. Additionally, convert categorical variables such as Gender, Smoking, Region, Cholesterol, Blood Pressure, and Diabetes into numerical representations like one-hot encoding. Conduct exploratory data analysis (EDA) to gain insights into feature distributions and identify potential outliers.

C. Feature Engineering

Explore interactions between features. Consider generating new features that may enhance predictive power.

D. Data Scaling

Apply Min-Max scaling or Standard scaling to normalize numerical features.

E. Train-Test Split

Divide the dataset into training and testing sets (e.g., 80% training, and 20% testing).

F. Model Building

Implement the following machine learning algorithms:

Linear Regression
XGBoost
Random Forest
Support Vector Machine (SVM)
Train each model using the training dataset.
Fine-tune model parameters using techniques like Grid Search or Random Search to optimize performance.

G. Prediction

Utilize the trained models to predict insurance costs on the testing dataset.

H. Evaluation

Assess model performance using standard metrics:
Mean Squared Error (MSE)
R-squared
Visualize the best-fit line for the Linear Regression model.

I. Correlation Analysis

Examine the correlation between predicted and actual insurance costs.

J. Comparison of Models

Evaluate and compare the performance of each algorithm.

K. Documentation and Reporting

Document the entire process, including data preprocessing steps, model-building parameters, and evaluation metrics. Summarize findings and insights. Consider creating visualizations to present results effectively.

L. Deployment(Optional)

If applicable, deploy the best-performing model for practical use. Create a user-friendly interface for end-users to input their characteristics and receive predicted insurance costs.

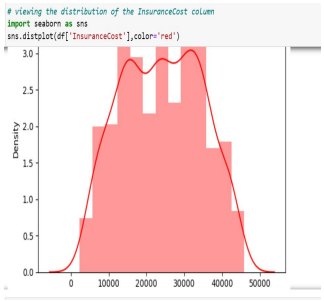
M. Continious Improvement

Identify areas for further research and enhancement. By following this methodology, the project aims to develop accurate and robust machine learning models for predicting medical insurance costs, providing actionable insights for decision-makers in the healthcare and insurance sectors.

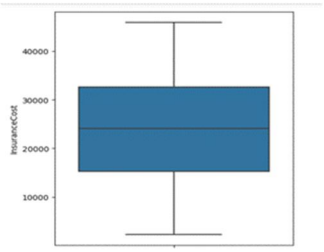
N. Data Scaling

Data Generation: Create synthetic data with Python code.
Data Storage: Save the generated data into a CSV file.
Data Preprocessing: Handle missing values, convert categorical data into numerical data, and scale features.
Model Training: Implement and train machine learning models.
Evaluation: Assess model performance and compare algorithms.
Correlation Analysis: Examine the relationship between predicted and actual values.
Documentation: Document the process and findings.
Continuous Improvement: Identify areas for enhancement and iterate.

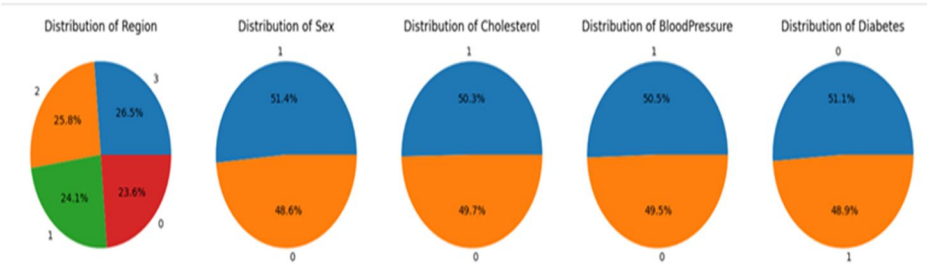
V. GRAPHS AND TABLES



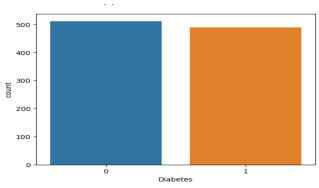
Here we use a distplot for Univariate Analysis. distribution plots in seaborn which is used for examining univariate and bivariate distributions.



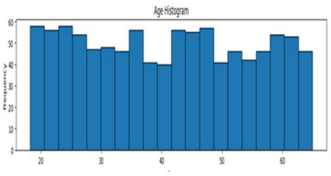
Here we use a box plot for analyzing the outlier but in the above figure there is not an outlier in Insurance cost the given feature has a maximum range is about 50,000 & a minimum range is 0 & 25000 is the median value. Here we Analyze the insurance cost over the Density. also Given data is normally distributed.



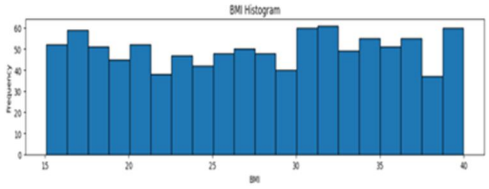
In the above graph, we analyze the Categories(0 and 1) in Percentages (%).



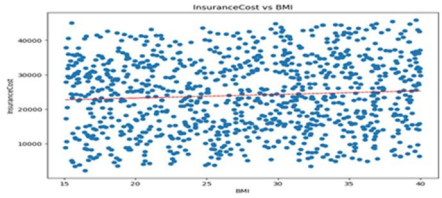
Reveals the frequency or count of each category. Shows which categories are more or less common for each variable. Differences in bar heights indicate relative frequency.



Here we Analyzed the average Age of our data points is 45 age. The highest is 60 years old and the Minimum is 40 years Old.



The above graph shows the BMI of data points. the average BMI is 50 which is not good for health. It produces a set of vertically stacked histograms, each representing the distribution of a different feature in your dataset. The histograms provide insights into the frequency and distribution of values for each variable.



Each scatter plot illustrates the relationship between a numerical variable and 'InsuranceCost'. The red dashed line represents the linear regression fit to the data, showing the general trend. Scatter plots can help visualize the correlation between variables, and the regression line provides a simplified model of the relationship. This kind of visualization is useful for understanding whether there is a linear trend between the numerical variables and the target variable ('InsuranceCost'). If the points follow a clear linear pattern, it suggests a potential linear relationship, and the regression line serves as a visual representation of this relationship. The scatter plot of "BMI" versus "Insurance Cost" visually represents the relationship between these variables. Each point on the plot corresponds to an individual, and the pattern of the points can indicate the potential correlation between higher BMI values and higher insurance costs.

```
#used to get concise summary of dataframe
data_encoded.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             1000 non-null   float64
 1   BMI             1000 non-null   float64
 2   Children        1000 non-null   int32
 3   Sex             1000 non-null   int64
 4   Smoker          1000 non-null   int64
 5   Region          1000 non-null   int64
 6   Cholesterol     1000 non-null   int64
 7   BloodPressure   1000 non-null   int64
 8   Diabetes        1000 non-null   int64
 9   InsuranceCost   1000 non-null   float64
dtypes: float64(3), int32(1), int64(6)
memory usage: 74.3 KB
```

Above dig. Shows the Basic information about the dataset.

Find the Nan Values

```
: # missing values

missing_values = data_encoded.isna().sum()
print("Missing values in the dataset:")
print(missing_values)
```

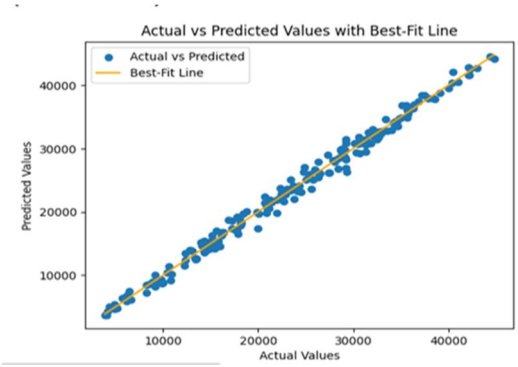
```
Missing values in the dataset:
Age              0
BMI              0
Children         0
Sex              0
Smoker           0
Region           0
Cholesterol      0
BloodPressure    0
Diabetes         0
InsuranceCost    0
dtype: int64
```

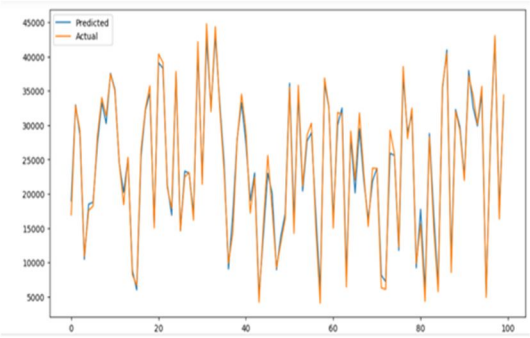
Model and Accuracy Interpretation Predictive Analysis
Step1: Start
Step2: Open Interface
Step3: Upload data sets
Step4: Pre-Processing
Step5: Training and Testing

Step6: Enter Constraints and Submit
Step7: result
Step8: Stop
We have used various Regression models for predicting medical insurance costs based on individual characteristics and also check which model fits best to our data. After analysis, we got accuracy for various Regression models as below :

Models	Mean Absolute Error	R Squared Error
Linear Regression	721.05	0.99
xgboost	4684320.39	0.95
Random Forest Regressor	914.98	0.98
SVM	8259.44	0.14
Gradient boosting regressor	875.91	0.98

Prediction Analysis





VI. Results

We can say that the XGBoost Regressor is the most suitable for predicting medical insurance costs based on individual characteristics since we have achieved almost 99.65% accuracy after tuning the parameters for the XGBoost Regressor Model. Discussion: Utilizing machine learning, we aim to develop a system capable of predicting both the medical treatment expense of a disease and the corresponding insurance claim amount based on user inputs. By leveraging advanced algorithms and analysing relevant features such as patient characteristics, medical history, and insurance details, we intend to create a model that can accurately predict these expenses. This system holds the potential to streamline the healthcare process, providing valuable insights for both patients and insurance providers, ultimately leading to more informed decision-making and improved cost management.

VII. Conclusion

Regression analysis is like the backbone of machine learning and data science. It helps us understand how different things relate to each other. Knowing about different types of regression analysis is really important because it helps us choose the right method for our data. After looking at everything closely, we've found that the XGBoost Regressor is the best option for our data. It's like the superhero of our analysis! With its special powers, we can dig deep into our data and make really accurate predictions. By using XGBoost, we can unlock hidden insights and understand our data better. It's like having a powerful tool in our hands to solve complex problems and make smarter decisions.

Future Scope

XGBoost provides numerous parameters that can be adjusted to improve accuracy or generalization in our models. In this discussion, we'll focus on a selection of key parameters that have significant influence and require close attention. By understanding these parameters, we can grasp how different values impact model performance. Additionally, we'll demonstrate how to utilize grid search to discover the best parameter values

within a specified range for our model. Furthermore, leveraging AI models can optimize feature selection, ensuring that we include the most relevant features for achieving optimal results. By utilizing AI-driven feature selection techniques, we can streamline the model-building process and maximize predictive accuracy.

REFERENCES

[1] S. Bhoite, C. H. Patil, S. Thatte, V. J. Magar and P. Nikam, "A Data-Driven Probabilistic Machine Learning Study for Placement Prediction," *2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, Bengaluru, India, 2023, pp. 402-408, doi:10.1109/IDCIoT56793.2023.10053523.

[2] 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

[3] 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.

[4] Sahu, Ajay and Sharma, Gopal and Kaushik, Janvi and Agarwal, Kajal and Singh, Devendra, Health Insurance Cost Prediction by Using Machine Learning (February 22, 2023). Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2022

[5] Fan, Cheng & Chen, Meiling & Wang, Xinghua & Wang, Jiayuan & Huang, Bufu. (2021). A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. Frontiers in Energy Research. 9. 10.3389/fenrg.2021.652801.

[6] Chen JH, Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. N Engl J Med. 2017 Jun 29;376(26):2507-2509. doi: 10.1056/NEJMp1702071. PMID: 28657867; PMCID: PMC5953825.

[7] Xie Y, Schreier G, Chang DCW, Neubauer S, Liu Y, Redmond SJ, Lovell NH. Predicting days in hospital using health insurance claims. IEEE J Biomed Health Inform. 2015 Jul;19(4):1224-1233. doi: 10.1109/JBHI.2015.2402692. Epub 2015 Feb 11. PMID: 25680222.

[8] E.Puneeth Kumar, P.Pavan Krishna, M.Raja Vardhan(2022) , Medical Expense Prediction Using Machine Learning.

[9] Hanafy, Mohamed. (2021). Predict Health Insurance Cost by using Machine Learning and DNN Regression Models. International Journal of Innovative Technology and Exploring Engineering. Volume-10. 137. 10.35940/ijitee.C8364.0110321.

[10] Thais Carreira Pfutzenreuter , Edson Pinheiro de Lima (2021) , Machine Learning In HealthCare Management For Medical Insurance Cost Prediction

[11] Shakhovska, Natalya & Melnykova, Nataliia & Chopiyak, Valentyna & ml, Michal. (2022). An Ensemble Methods for Medical Insurance Costs Prediction Task. Computers, Materials and Continua. 70. 3969-3984. 10.32604/cmc.2022.019882.