Predicting Medical Insurance Costs

K. Nikitha[1], G. Rohitha[2], Fazal Ur Rahman[3], K. Kushal[4], Dr.Padmaja Pulicherla[5]
[1,2,3,4]*Student of Computer Science and Engineering, HITAM Hyderabad, India*
[5]*Professor and HOD, Department of CSM, HITAM,Hyderabad, India*

***Abstract*—In healthcare industry It is very important for both the insurance companies and the insured insurance to predict as accurately as possible the cost of medical cover for an individual by using machine learning. This project centres around applying a machine learning model to predict the medical premium of an individual, based on individual characteristics-demeanour, age, gender, Body Mass Index, smoking status, number of dependents, and region of residence. Considering a dataset with earlier mentioned variables, then a predictive model can be built, which will be highly accurate in estimating insurance costs. It is employed in this research to predict medical insurance costs using Machine learning techniques like Linear Regression, Support Vector Regression (SVR), and Random Forest Regressor. Numerous models will thus be built and evaluated on how they estimate insurance premiums. Linear Regression is the simplest way to ascertain the linear relationships, while SVR captures even more complex non-linear patterns, and Random Forest Regressor uses ensemble learning to improve accuracy in predictions. The achieved results will show the capability of these models in predicting costs. At the same time, they will tell how well the models perform with respect to each other in that and which area they were good with respect to their applicability.**

***Index Terms*—Medical Insurance, Python, Machine Learning, Predictive Modeling, Insurance Premiums, Healthcare.**

I. INTRODUCTION

Digital health is a rapidly growing sector worldwide, with the number of digital health businesses having doubled over the last five years. This growth reflects a global shift towards integrating technology and healthcare, enabling innovative solutions for patients and providers alike. Health insurance, a vital component of the healthcare ecosystem, faces significant challenges in industrialized countries, including escalating healthcare costs and a rising number of uninsured individuals. This has spurred a broad-based political movement advocating for reform, with governments pledging hundreds of millions of dollars to accelerate advancements in digital health. Individual health insurance is particularly critical for individuals with rare diseases, who often face prohibitive treatment costs. Medical and preventative insurance can reduce these expenses, providing a safety net for those who need specialized care. In our unpredictable world, where risks to personal well-being and assets are constant—from disease and death to potential property losses—the financial industry has developed a range of products to protect individuals and businesses. These insurance products provide a monetary buffer against such risks, helping to mitigate or even eliminate the financial impact of certain adverse events. This project focuses on accurately predicting medical insurance costs is crucial for both insurers and customers, as it helps in better financial planning and managing risks. This project focuses on building a machine learning model that can estimate insurance premiums based on various factors like age, gender, BMI, smoking habits, number of dependents, and location. By using different techniques—Linear Regression, which identifies straightforward relationships; Support Vector Regression (SVR), which captures more complex patterns; and Random Forest Regressor, which combines multiple decision trees—the project compares these models. The goal is to determine which model is best for accurately predicting insurance costs and helping people make better financial decisions

II. RELATED WORK

Now in this related work part, we will discuss some work that has been done in this field.

A. Predict health insurance cost by using Machine Learning and DNN Regression Models. [1]

The aim of this study has been to primarily base the forecast on defining various machine learning regression models together with deep neural networks to predict charges based on predetermined attributes-specific ones that can define the medical cost personal data set at kakgle.com. The results have been summarized in Table IV. indicates that Stochastic Gradient Boosting has the highest efficiency with an RMSE value of 0.380189, an MAE value of 0.17448, and an accuracy of 85.82. Hence, stochastic gradient boosting could be adopted for estimating insurance costs much better than other regression methods. Predicting insurance costs based on some parameters would assist the consumers in tending toward insurance policy providers, thus saving time to draft plans for each individual. Machine learning reduces effort in formulating individual policies as ML models can perform cost calculation rapidly with time-consuming tasks performed by humans. This will allow companies to increase profitability. ML models can handle huge amounts of data as well.

Mrithikainectediebtbootanely computeted in k-mean clusters the medical cost personal data set in order to enhance its reliability. Comparison of input features was possible using the following nine features: gender, age, children, bmi, smoker, work type, education, marriage status, and income. Overall, the model achieved an RMSE value of 0.380189, MAE value of 0.17448, and an accuracy of 85.82 for stochastic gradient boosting. Therefore, this model can indeed be used for estimating insurance costs much better than other regression models. It helps predict the insurance costs based on some parameters that may enable consumers to steer towards insurance policy providers save time in drafting individualized plans.

B. Machine-Learning-Based Regression Framework to predict Health Insurance premiums. [2]

The authors suggested that machine learning could be used for health insurance tasks that use to be done by human beings but much slower. With the help of intelligent tools and machine learning, massive quantity of data can be analyzed more quickly and efficiently, facilitating the operation of health insurance by making it easier for both policyholders and insurers. Time and money can be saved using machine learning. With the machine learning tasks being performed quicker with fewer costs compared to humans, all of these would be for the benefit of patients, physicians, hospitals, and insurance companies alike. The authors recommended the use of machine learning in particular ANN artificial neural networks to cater to issues regarding the analysis of historical data for prediction of health insurance. Development of an ANN-based regression model for predicting health insurance premiums and evaluation of its performance using RMSE, MSE, MAE, and R^2 as evaluation metrics. The authors pursue plotting a correlation matrix to illustrate the interrelationship of various factors and their influence on insurance charges. This particular insurance prediction field is still lacking in studies and needs to have deeper explorations in this area.

C. Predicting Health Insurance Premiums with Machine Learning Techniques [3]

The authors propose the utilization of machine learning regression models for predicting health insurance prices on medical cost individual dataset from Kaggle. The findings are summarized in Table IV. With prediction of insurance premiums using various factors, insurance companies could save time as well as attract customers. The authors said, "By using machine learning, one could lessen the physical effort of price analysis because machines compute costs much faster than humans." Furthermore, these models can also process very large quantities of data. By future work application of either XGBoost or Gradient Boosting would serve a significant improvement to generalize with a more extensive data set than that applied in this study.

D. Machine Learning for an explainable cost prediction for Medical Insurance [4]

This study persuades the model-appropriate accuracy level available to be exploited at the associated computation resources and offers an explanation on the value accorded to explainability, which forms an essential part in building trust among stakeholders towards the models developed in healthcare insurance. XGBoost had the great accuracy; however, it was very expensive regarding computations. On the other hand, Random Forest served as less costly. Yet this study cannot be generalized because it has a small sample that may not be representative. It also does not compare to very advanced or very simple models, it has no validation against a realistic scenario, and does not have provisions that show how things will change as the nature of health care and the fairness issues regarding ethics change from time to time. While SHAP and ICE allow one to understand the models

more, they do not provide much about either level or comfort. They deviate from the focus of this research, which is much richer toward the prediction of premiums, to less favored themes such as estimating claim loss or evaluating risk.

E. Medical Insurance Premium Prediction with Machine Learning [5]

This is a machine learning technique that applied learning datasets consisting of patient's demographics such as age, gender, body mass index (BMI), smoking habit, area of residence, and personal medical records for the study of the prediction of health insurance premiums in this paper. Several techniques have been ensemble methods, regression techniques, and an ann using Tensorflow Keras. For this exercise, datasets were preprocessed, split into training and testing sets, and R-squared and the Mean absolute error were used to measure model performance. The models thus provided reasonable estimates of the relevant determinants that significantly influenced insurance premiums, including BMI and smoking. This method has its own model-related shortcomings, as it also helps in better-pricing risk reassessments and customization of insurance coverage. Structure dependence rather than difference is less generalized and scalable as it approaches the problem with only one dataset. Further, it does not include comparisons with other models, exhaustive hyperparameter tuning, and ethical concerns like bias and equity, making the methodological critique weak. Hence, in-depth methods cannot really be hammered out without involving actual field situations or motivational data.

F. Medical Insurance cost prediction. [6]

The main objective of the research as mentioned above is to forecast the expenses incurred by a client regarding the medical insurance and to do this by employing regression models specifically the linear regression, the result in precision coming at 74.45%, and also Ridge regression and Support Vector Regression which achieved 82.59% accuracy. It examines parameters like age, sex, chronic past illness, lifestyle, and geographical location-all informative and determinant factors for estimating costs-to assist the insurance company in premium formulation and budgeting. However, along with many useful outputs, this study has a few drawbacks like time accuracy, very limited dataset, and did not take prices into consideration and biases and fairness. Further, it does not compare the outcomes with current machine learning algorithms and also lacks real-life implementations, which limits the applicability and scalability of the model.

G. Health Insurance Amount Prediction. [7]

Here we analyze the individual's personal health data to derive the insurance amount. Different regression methods, such as the Multiple Linear Regression, the Decision tree Regression, and the Gradient Boosting Decision Tree Regression, were applied to contrast the performance of these algorithms. Dataset was used to train the models and train the models helped come up with some predictions. Then, the predicted amount was compared to actual data to test and model verification. Later, the accuracy of the models was compared against another. It was found that multiple linear regression and gradient boosting outperformed the linear regression and decision tree. In this case, gradient boosting fits the bill the best, as it takes far less time than computing time to reach the same performance metric, while it performs similarly to multiple regressions.

H. Approach for medical insurance costs Prediction using SGTM Neural-like structure [8]

This paper proposes a method for predicting medical costs incurred due to insurance. It is based on a piecewise-linear approach that uses the SGTM neural-like structure. Piecewise-linear gives very high processing efficiency in case of large data, and then SGTM neural-like structure gives very high accuracy with a high-speed training procedure. The suggested method was simulated with actual health insurance cost data on two SGTM neural-like structure cascades. Here, experimental results were derived to ascertain the high speed and accuracy of the proposed method. For comparing the proposed method, existing methods were included, particularly multilayer perceptron and the Common SGTM neural-like structure, which solved the task according to the entire dataset. Meanwhile, it was found that the worst results show a multilayer perceptron: its accuracy of operation according to MAPE is more than 23% less than the accuracy of the proposed method, whereas the time of the training procedure lasted 51 times longer. The baseline method shows a relatively higher learning speed but lower accuracy: 11% more error than that from the developed method. These results gave evidence of feasibility for the proposed approach for the possible processing of huge data, especially in the

fields of medicine, economics, materials science, and service sciences**.**

III. PROBLEM STATEMENT

Traditional methods for calculating insurance costs often rely on simplified models or heavily emphasize historical data, leading to potential inaccuracies. These models typically use only a few straightforward factors, such as age and gender, without accounting for the complex interactions among various other elements that can significantly affect premiums. For instance, while traditional methods may consider age and gender, they might neglect the intricate interplay between these factors and other variables, like Body Mass Index (BMI) or smoking habits. Additionally, lifestyle choices, regional healthcare costs, and the number of dependents are often overlooked, which can result in estimates that don't accurately reflect an individual's unique risk profile or the true cost of coverage. Consequently, these methods may not capture emerging trends or accurately predict future expenses, limiting their effectiveness for both insurers and policyholders.

This project proposes to develop and evaluate advanced machine learning models to predict medical insurance costs, incorporating a comprehensive set of features. These include age, gender, BMI, smoking habits, number of dependents, and geographical location. By leveraging machine learning, we aim to enhance the precision of cost predictions and uncover insights into how these factors collectively influence insurance premiums. Machine learning techniques offer the ability to model complex relationships among variables and dynamically adjust to new data, which promises to deliver more accurate and individualized insurance cost estimations than traditional methods.

IV. PROPOSED METHODOLOGY

Typical methodologies have been to evaluate the annual premium costs incurred by insurers using simplified models or data from historical databases. This has given rise to biases because such models rely on only few factors, like age and gender, and do not consider the emerging signals in premiums coming from interactions of many others that can considerably affect the premiums. For instance, at age and sex, the traditional measure may take BMI or smoking as an afterthought, ignoring other factors of its intersection, including lifestyle, regional care costs, and a dependent count, which would bring estimates even farther from the individual's risk profile or real coverage cost. Such methodologies can hence prove quite gross, as they fail to pick some emergent trends and cannot predict future expenses accurately. Develop and evaluate advanced machine learning models for predicting medical insurance costs on a very comprehensive set of features, including but not limited to age, sex, BMI, number of dependents, smoking habits, and the geographical location. If they could improve on traditional methodologies in reliably predicting costs and educating them as to how these factors could be used together to affect premiums, machine learning models promise to do the same with regard to much more complex interactions among variables that constantly adapt to new data.

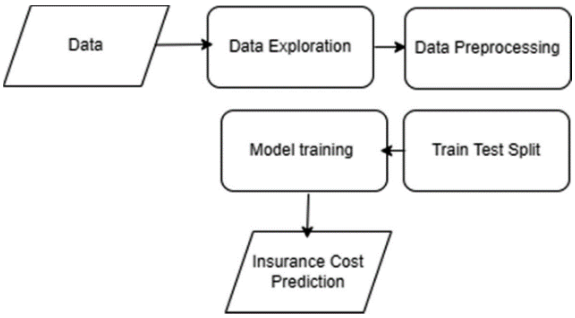

Fig 1: Proposed Diagram

A. Data Pre-processing

Data preprocessing is an essential process that helps to safeguard the quality of the data to be used for training any machine learning model. The following actions were taken:

- The dataset, likely containing insurance-related data with features like age, BMI, smoking status, and region, was loaded into the Python environment using the Pandas library.
- Irrelevant or redundant columns are removed to streamline the dataset. This step ensures that only significant features contributing to the prediction of the target variable (charges) are included.
- The independent features were assigned to X, which represents all columns except charges (the target variable).
- The dependent variable (charges) was assigned to Y, which represents the column to be predicted.

- The dataset was split into training and test sets multiple times using a loop with varying random state values (from 40 to 50) to evaluate model performance under different random splits.
- A test size of 20% was used, meaning 80% of the data was used for training, and 20% was used for testing.
- Standardization of the feature set was done to make the model's performance better. The StandardScaler function from scikit-learn was utilized to standardize the features by removing the mean and scaling to unit variance. This step ensures all numerical features contribute equally to the model and avoids biases caused by varying feature magnitudes.

B. Model Selection

A few machine learning algorithms were selected to identify the most suitable model for medical insurance cost prediction:

Basic Algorithms: Neural Networks, Linear Regression, Polynomial Regression, Decision Trees, Support Vector Machines (SVM) with Regression (SVR), Random Forest Regressor

Advanced Algorithms**:** Gradient Boosting Machines (GBM): XGBoost, LightGBM

C. Model Training and Evaluation

Each model was trained on the standardized training set and evaluated using various metrics:

- Each classifier was fitted to the training data, followed by predictions on the test set.
- Performance Metrics: The Training Accuracy, Testing Accuracy and Cross-Validation Score are calculated for each model.

D. Prediction on New Data

The most accurate machine learning model, identified through evaluation, is used to predict insurance cost for new customer. Finally, the prediction will provide an estimated insurance cost for the individual in the new data.

V. RESULT AND DISCUSSION

The performance of various machine learning models in predicting medical insurance costs was evaluated based on the training accuracy, testing accuracy and cross – validation(cv) score.

Model	Training Accuracy
Linear Regression	0.729
Support Vector Machine	-0.105
Random Forest	0.974

Table 1

Table 1 shows the Training Accuracy for the Machine learning models evaluated.

Model	Testing Accuracy
Linear Regression	0.806
Support Vector Machine	-0.314
Random Forest	0.882

Table 2

Table 2 shows the Testing Accuracy for the Machine Learning models.

Model	CV Score
Linear Regression	0.747
Support Vector Machine	0.103
Random Forest	0.836

Table 3

Table 3 shows the Cross-Validation score for the evaluated models.

Among the models evaluated, Table 1 shows that Random Forest achieved the highest training accuracy of 97.4% followed by Linear Regression whereas, Support Vector Machine showcased a very poor performance. When we take Testing Accuracy and CV Score also the Random Forest model achieved the highest accuracy of 88.2%, 83.6% respectively.

Also, when we consider the visualizing of data in pie charts, scatterplots and boxplots, Visualizing the categorical features helps to understand the dataset's composition and any potential biases that may exist. By considering the distribution of these features, we can enhance the model's performance, reduce bias, and improve the accuracy of predictions for medical insurance costs.

Fig 2, Fig 3 , Fig 4 gives the Pie charts for the sex, smoker, and region column based on the dataset provided.

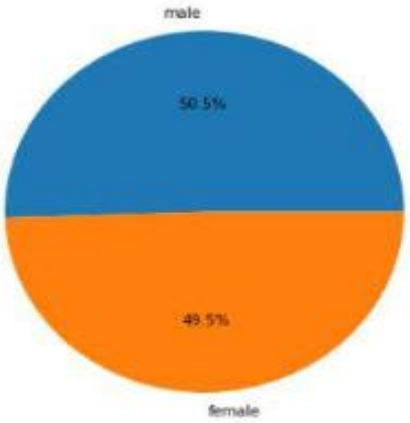

Fig 2 Pie chart for sex column

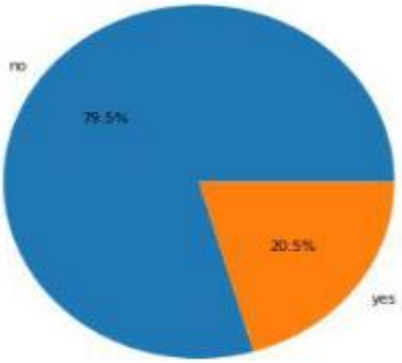

Fig 3 Pie Chart for smoker column

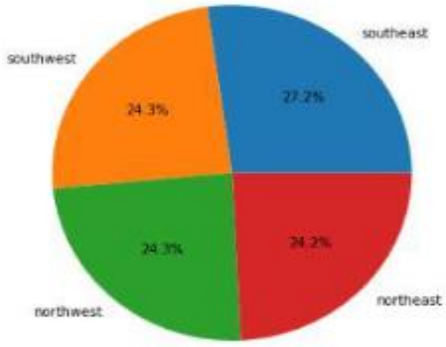

Fig 4 Pie Chart for region column

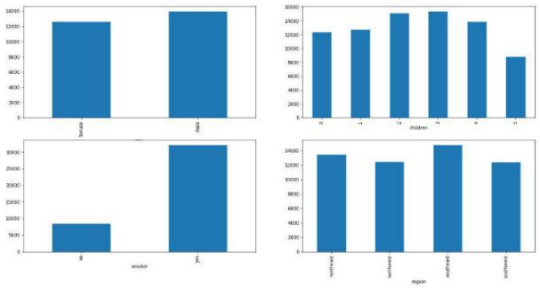

Fig 5 Comparison between Charges paid between different groups

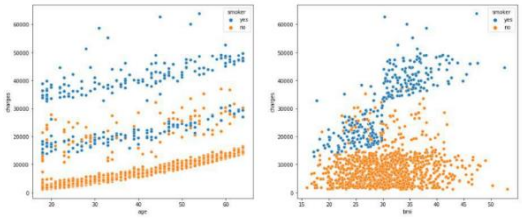

Fig 6 Scatter plot of the charges paid v/s age and BMI respectively

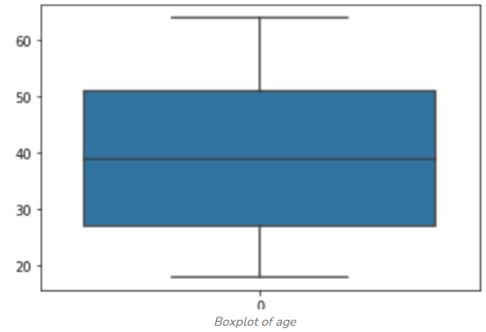

Fig 7 Boxplot of age

Here, we can see there are no outliers present in the age column.

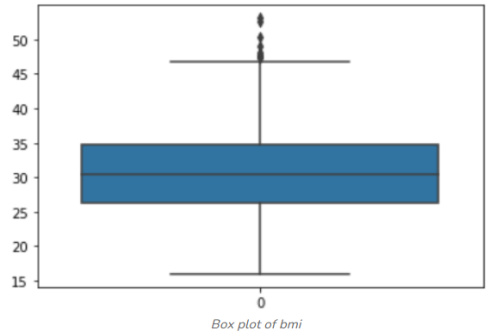

Fig 8 Box plot of BMI

Due to the presence of outliers present in BMI column we need to treat the outliers

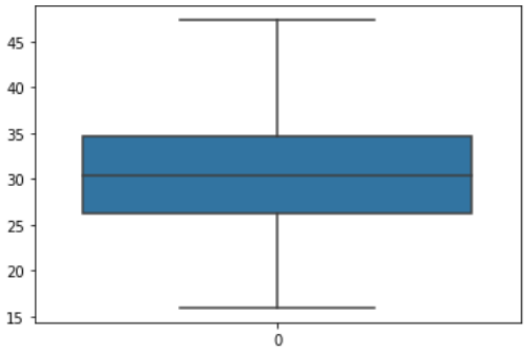

Fig 9 Boxplot of BMI

We successfully treated the outliers by replacing the values with the mean

REFERENCES

[1] Mohamed hanafy, Omar M.A. Mahmoud - Predict Health Insurance Cost by using Machine Learning and DNN Regression Models.
[2] Keshav Kaushik, Akashdeep Bhardwaj, Ashutosh Dhar Dwivedi and Rajani Singh - Machine Learning-Based Regression Framework to Predict Health Insurance Premiums
[3] Sazzad Hossen - Predicting Health Insurance Premiums with Machine Learning Techniques.
[4] Ugochukwu Orji a, Elochukwu Ukwandu - Machine learning for an explainable cost prediction of medical insurance
[5] Prof. M. S. Patil, Kulkarni Sanika, Khurpe Sanjana - Medical Insurance Premium Prediction with Machine Learning
[6] Sabarinath U S, Ashly Mathew - Medical Insurance Cost Prediction
[7] Nidhi Bhardwaj, Rishabh Anand -Dr. Akhilesh Das Gupta - Health Insurance Amount Prediction
[8] Roman Tkachenko, Ivan Izonin, Natalia Kryvinska, Valentyna Chopyak, Nataliia Lotoshynska, Dmytro Danylyuk - Approach for Medical Insurance Costs Prediction using SGTM Neural-Like Structure