MEDICAL INSURANCE PRICE PREDICTION USING MACHINE LEARNING IN PYTHON

[1] Mrs.S.Yoga., M.Sc(CS&IT).,M.Sc(Maths).,M.Phil(CS)., M.Phil(Maths)., Assistant Professor

[2] J. Logeshwari, M.Sc (Computer Science),

[1] [2] Department Of Computer Science, Sakthi College of Arts and Science for Women, Oddanchatram.

ABSTRACT

Medical costs are one of the most common reoccurring expenses in a person's life. It is general known that a person's lifestyle and numerous physical factors determine the diseases or disorders they may get, and that these conditions determine medical expenses. According to several research, there are several significant reasons that lead to greater expenditures. smoking, age, and BMI are all factors in personal medical care. The goal of this study is to examine and identify a link between personal medical costs and other characteristics. Then, by generating linear regression models and comparing them using ANOVA, we use the significant traits as predictors to forecast medical expenditures. In our research, we discovered that smoking, age, and a higher BMI all have a significant connection with higher medical expenditures, showing that they are key contributors to the charges, and that the regression can predict the charges with more than 75% accuracy. According to the World Health Organization, personal medical and healthcare spending is growing faster than the global economy This rise in spending has been related to a variety of factors, the most prominent of which are smoking, ageing, and higher BMI. Using insurance data from diverse persons with variables such as smoking, age, number of children, area, and BMI, we hope to uncover a link between medical expenditures and other parameters.

I INTRODUCTION

The expense of health care is rising every day. There is a need to forecast health costs as the number of novel viruses infecting humans grows. This form of forecasting aids governments in making health-related decisions. People are also aware of the significance of health-care spending. Machine Learning is a field that touches all aspect of life. Machine learning models are also used in the health-care system for a variety of health-related applications. We conducted a predicate analysis on medical health insurance expenses in this study. We create a model to forecast a person's medical insurance costs depending on gender. The dataset comes from Kaggle and

comprises 1338 rows of data with the following attributes: age, gender, smoker, BMI, children, region, and insurance charges. Medical information and expenditures billed by health insurance companies are included in the data. To forecast medical expenses, we used a variety of regression techniques on this dataset. The Python programming language was utilized to implement the project.

II LITERATURE SURVAY

Machine learning is a technique that allows computer to learn from past data and anticipate fresh samples. Machine Learning models may be used in any sector. Medical records are likewise not exempt from machine learning. For numerous years, the medical industry has used models in various settings. Many of the studies used machine learning approaches to forecast medical costs B. Nithya [1] et.al In Predictive Analytics in Health Care, machine learning models were used. For predictive analysis, they used a variety of supervised and unsupervised models. They also claimed that machine learning tools and techniques are crucial in health-care sectors, and that they are exclusively employed in the detection and prognosis of various malignancies. Ahuja Tike[2] et.al applied hierarchical decision trees for the medical price prediction systems. Their experiments showed that the price prediction system achieves high accuracy. Moran et al. [3] utilized linear regression techniques to anticipate Intensive Care Unit (ICU) expenses and utilize understanding socioeconomics, DRG (Diagnostic Related Group), length of stay in the clinic, and a couple of others as highlights. Gregory [4] et.al applied various regression models for analyzing medical costs in the health care system. They mainly concentrated on reducing the bias in the cost estimates to achieve good results. Dimitris Bertsimas[5] et.al applied different data mining techniques which provided an accurate prediction of medical costs and represent a powerful tool for the prediction of healthcare costs.

2.1 Medical Expense Prediction System using Machine learning Techniques and Intelligent Fuzzy Approach (2020, H. Chen Jonathan, M. Asch Steven) Prediction isn't a new concept 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, which rely on data from cohorts that have been thoroughly prepared to prevent bias, new data sources are sometimes unstructured due to the fact that they were developed for various purposes (clinical care, billing, etc.). Patient self-selection, indication misunderstanding, and inconsistent outcome data can all contribute to unintentional biases and even racist programming in machine prediction. As a result of this understanding, discussing the potential of data analysis to aid medical decision-making isn't just wishful thinking.

2.2 Predicting Days in hospital using health Insurance Claims (2020, Yang Xie, C.W. Chang, Sandra Neubauer) Identifying and managing patients most at risk within the health care 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[1]. Hospitals focus on readmission rate [2], [3] and cumulative risk of death during hospitalization [4]. 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 measure 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 health-care expenses [5]. Traditional prediction models used demographic information and prior costs to predict future costs [1]. More sophisticated models that incorporate diagnoses [6], [7], drug claims [6], [7] and self-reported health status data [8], have been shown to improve prediction performance. Zhao et al. [6] reported a coefficient of determination (R2) of 0.168 when both drug and diagnosis were used to estimate costs for the coming year. Bertsimas et al. proposed two models: a decision tree model and a clustering model [7], to improve the performance of an earlier model based on classical regression models [6], [9]. Since hospitalization is usually the largest component of health expenditure [10], a separate identification of subpopulations at higher hospitalization risk could improve current underwriting processes and pricing methodology. Moreover, insurance companies also manage insurance risk by using specific interventions in different sub-populations (especially high-risk groups) to minimize the resources they require [11]. Programs that use case and disease management have been developed, which target different sub-groups of customers, such as aged care programs, chronic disease programs, and more recently

telehealth programs, all of which have been shown to improve health care outcomes. identify high risk customers by predicting their health care expenditures.

III THEORETICAL BACKGROUND

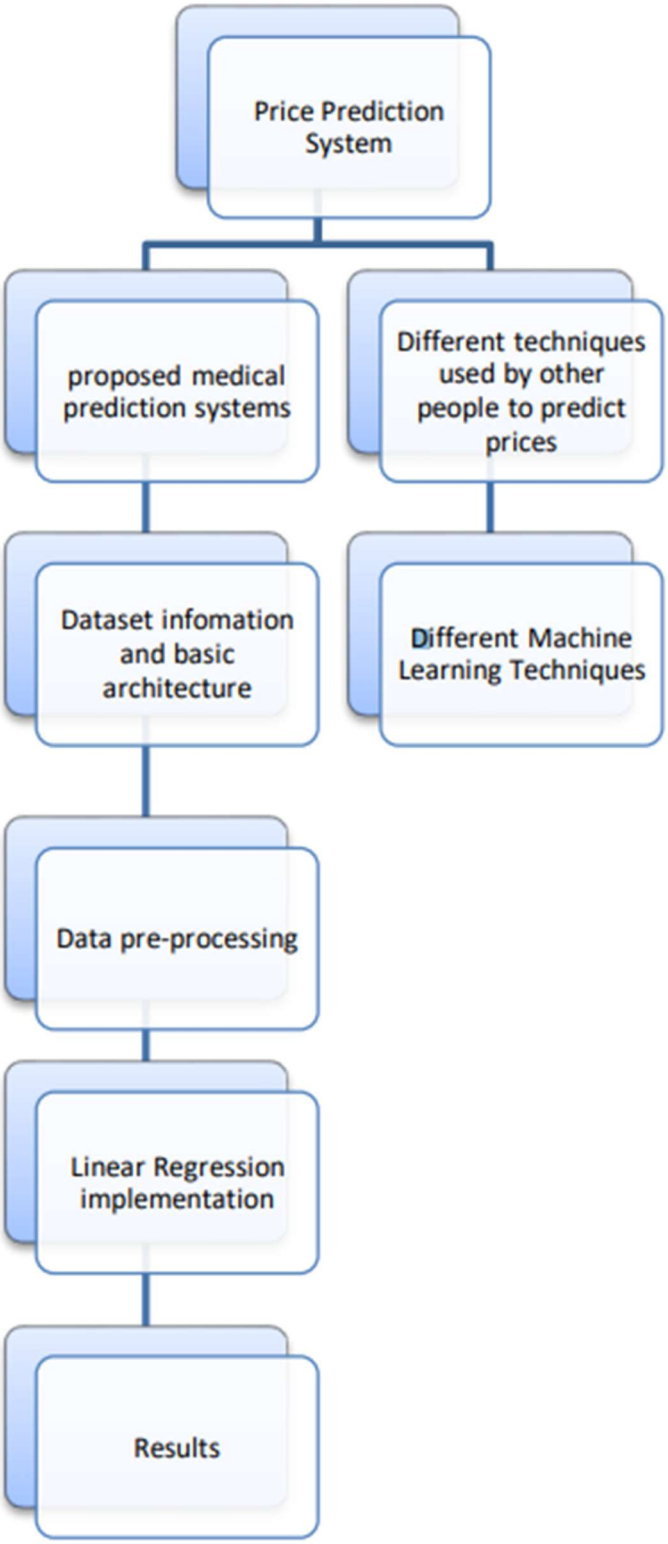
3.1 PROBLEM IDENTIFICATION

- In sample sizes ranging from small to large, statistical approaches (E.g., Lenear regression) suffer due to the zsero point spike and skewed distribution of health care expenses with a strong right-hand tail. To address this issue, advanced methods have been proposed, such as The data source, on the other hand, has a flaw. Unlike traditional techniques, which rely on data from cohorts that have been thoroughly prepared to eliminate bias, new data sources are sometimes unstructured since they were developed for various purposes (clinical care, billing, etc.). A range of issues, ranging from patient self-selection to indication uncertainty and inconsistent outcome data, can induce unintentional biases and even racist programming in computer prediction. As a result of this understanding, speculating on the potential of data analysis to aid medical decision-making isn't far-fetched.

3.2 PROBLEM SOLVING

- The Medical information and expenditures billed by health insurance companies are included in the data. It has 1338 rows of information with the following columns: age, gender, BMI, illnesses, smokers, and insurance costs In these features, insurance charge is a dependent variable and the remaining features are called independent variables. In regression analysis, we need to predict the value of the dependent variable using independent variables. First, we collected the dataset and applied various data preprocessing methods. Data preprocessing is a technique in which we can remove missing values in the data. Because of these missing values, it is not possible to apply machine learning algorithms. After removal of missing values, we need to apply label encoding, one hot encoding data to the categorical features.

3.3 SYSTEM ARCHITECTURE



IV SYSTEM IMPLEMETATION

4.1. MODULE

- Dataset Information
- Selection of features from the dataset
- Data preprocessing

MODULE DESCRIPTION

- **Dataset Information:**
  Our suggested system's input dataset will be a dataset. That combines two different datasets. A series of inpatient Medicare payment data and a column of Zillow data will be displayed. be included in our final input dataset. In the following part, we'll look at these columns in further depth. The first dataset, Hospital-level payments to about 30,000 hospitals in the 100 most often billed Diagnosis Related Groups are included in the Medicare payment data collection (DRGS). The top one-hundred DRGS account for 60% of total inpatient Medicare payments. expenditures and accounts for 7 million discharges. Each row in the payment dataset comprises 10 columns, as seen in the preceding section. Each of these columns denotes a characteristic of machine learning. A feature is a significant attribute that influences the prediction variable under consideration. Every problem under investigation has a collection of independent characteristics that aid in the construction of an accurate machine learning
- **Selection of features from the dataset:**
  Selecting the key features from a large range of features that are more relevant and building a strong model is a critical effort. As a result, we will choose just those variables from our Medicare dataset that will independently assist us in predicting medical prices. The location of the provider is represented by columns such as provider address, ZIP code, state, city, and hospital area referral description. As a result, rather than examining all of them, we will just include one of them in our feature set. We'll go with 'hospital region referral description' because it's not as particular as a provider's location or city, but it's also not as wide as a state. DRG Definition and Total Discharges are two further independent characteristics we'll pick from medical payment data. In addition to these features, the Medicare payment dataset now includes a new feature: real estate values. These are the prices of real estate in the hospital's immediate vicinity for the same year as the medical data. This element, in our opinion, can potentially be a prominent component in price prediction. The notion is that a hospital's operating costs are factored into the price it charges. Among the several expenditures that a hospital must face, one of the most significant is the cost of real estate - the cost of owning or renting a facility. Real estate expenses are also a proxy for other costs, in the sense that if a location

has high real estate costs, it is likely to have higher costs elsewhere 19 Salary given to physicians, personnel, and other categories are also included.

- **Data preprocessing:**
  The data that will be utilized to answer the problem is one of the most significant aspects of machine learning difficulties. Data preparation accounts for around sixty to seventy percent of the overall time spent on a typical machine learning project. In order to get successful outcomes, it is critical to have the proper data for the situation at hand. In general, data preparation consists of selecting characteristics and pre-processing those features. As a result, after selecting features from a vast quantity of data, the following step is to pre-process those features. Because the data is useless in its raw form. The goal of pre-processing in this case is to make features appropriate for the machine learning model we'll use. If the characteristics are set up correctly, the model can produce better results. In addition, the data formats for various models varies.

V CONCLUSION & FUTURE WORK

5.1 CONCLUSION

We've looked at the fundamentals of the linear regression model, how to use it forecast charges, and how to compare anticipated and real outcomes. I hope you found this post helpful and that you now have a basic understanding of how a linear regression model works. For estimating medical expenditures, we suggested a machine learning approach. We applied regression techniques Linear Regression and observed that age, BMI are features that decide the dependent variable. Out of all experiments, this model gave a better result.

REFERENCE

1. Y. Bengio, "Deep Learning," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798-1828, 2013.

2. A.Ng, "Machine Learning Yearning," 2018. [Online]. Available: http://www.mlyearning.org.

3. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. Springer, 2016.

4. P. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

5. B. Baesens et al., "Predictive Analytics in Healthcare," International Journal of Healthcare, vol. 23, no. 3, pp. 124-132, 2015.

6. A.Geron, Hands-On Machine Learning with Scikit- Learn and TensorFlow, O'Reilly Media, 2017.

7. P. Kumar, V. Gupta, and R. Bhardwaj, "Medical Insurance Prediction using Machine Learning Algorithms," Journal of Artificial Intelligence Research, vol. 9, no. 2, pp. 345-352, 2019.

8. M. Shwartz, I. Mahajan, and S. Patel, "Ensemble Learning Techniques in Healthcare Predictions," Machine Learning and Applications, vol. 33, pp. 98- 105, 2020.

9. I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed., Elsevier, 2016.

10. P. Refaeilzadeh, L. Tang, and H. Liu, "CrossValidation," Encyclopedia of Database Systems, Springer, 2009.