Introduction

Health Care is a fundamental need, and Health Insurance plays a key role in making medical services accessible and affordable for everyone. No one plans to get injured or sick, but as we know our lives are very unpredictable. A sudden illness or accident or even a normal routine check-up can lead to expensive healthcare costs which can drain your savings overnight and can lead to financial burdens. However, in recent years, the number of people with health insurance has declined due to rising costs, changes in employer covered coverages and shifting government policies. This trend is continuing to rise and creates a serious problem because of more people are at risk and raises concerns about whether insurance pricing is fair and transparent for all individuals.

Insurance companies usually decide their premiums based on various factors which are age, gender, body mass index (BMI), smoking habits and even geographic region. For example, smokers are usually charged more because they are at a higher risk of health problems. But sometimes, these pricing models are not fair. Two individuals with similar health risks might be charged differently because of their gender or their geographic location. This leads to creation of inequality in the system. In this case, people may end up paying more than they should and sometimes even denied the proper insurance claim.

In this project, the main aim is to use machine learning models to predict the cost of medical insurance based on the chosen dataset which has features like age, BMI, smoking status, and region. The goal is to build a model which predicts fair and accurate results which means the model should not make predictions that are biased or unfair to certain group of people.

Related work:

In today’s world, the insurance industry is also growing and modernizing with the help of Artificial Intelligence, Machine Learning and automation tools. These technologies can help to reduce and eliminate unfair biases and decisions and improve how people are charged for medical services. In one study, the reasearchers worked on predicting when insurance companies may deny medical claims. They tested six AI models using data from a hospital with 60000 medical claims. They focused on fairness, data privacy and accountability with keeping main aim as stopping the insurance denial before they happen. Their approach recommended using more data from many hospitals to improve the system;s fairness and usefulness. Another study, focused on prediction of healthcare expenditures. They found in this study that insurance companies sometimes design their policy plans in such a way that underrepresented groups were often underpaid by these insurance companies. Here the authors proposed the use of new regression models which include these underrepresented groups and reduced the unfairness by 98% which showcased that small changes in the model can greatly increase the fairness.

One of the other studies also raised concerns about how health insurers use AI. They shared real examples where AI systems were used to quickly deny claims, even discharging patients too early. Some companies are now facing legal action. The authors said that even though algorithms can be dangerous, some may be required to make the system fairer. They suggested that more transparency and national rules are needed so that insurers use AI responsibly and fairly.

After careful consideration of these studies, I was motivated to build a system that predict insurance costs, checks for fairness and explains how the model makes its decisions using SHAP values.

Approach

Here the dataset used is the Medical insurance Cost dataset from Kaggle. This dataset contained 2,772 entries across 7 columns like age, sex, BMI, children, smoker status, geographic region and medical charges. At first, basic data exploration was conducted where I checked the data types, ensured there were no null values, and calculated the statistics of the dataset. I found that the average age was 39 years, where the age range lied from 18 to 64 years. The BMI (body mass index) ranged from 15.96 to 53.13 with a mean value of 30.66, which suggested that many patients fall near or into the obese category.

I also wanted to know if the male : female ratio was balanced or not, where I found that there were 675 males and 662 females, from which I can say it was equally balanced. Then further I wanted to know what is the average number of children which was 1. Coming to the insurance charges, they ranged from $1121.87 to $63770.43, which showcased a highly skewed distribution, where a small number of high cost cases raised the overall average for the charges.

To have a more cleaner dataset, data duplicates were identified and removed. There were 1435 records in total which were redundant, and after removal 1337 unique records were kept. This was important step in data cleaning and preprocessing to make sure that data fed into the model was unbiased and accurate.

A region wise analysis was also done, where it was revealed that the South – East region had the highest total insurance charges and South west region had the lowest total insurance charges.

To understand the data distribution more accurately and detect patterns, we visualized different columns:

Firstly, Age Distribution was visualized where I could see how ages were fairly distributed with slightly fewer entries in the 30s.

Secondly, BMI distribution was a bell-shaped curve centered around 30, indicating borderline obesity.

Thirdly, Medical charges was right-skewed distribution with significant high cost outliers.

Fourthly, Boxplots were generated to visualize smokers had much higher medical charges compared to non – smokers.

A correlation heatmap was also generated. It showed smoker status and age were most strongly associated with medical charges. BMI also had a noticeable correlation, while region and gender had minimal influence on charges.

In terms of Data preprocessing, categorical columns like “sex” and “smoker” were converted into numerical values (Binary encoding). The “regio” column was transformed using one-hot encoding to capture regional variance in a machine readable format.

After data cleaning and pre-processing, the dataset was saved in a cleaned format with no missing or duplicate values. Despite the presence of outliers in charges, they were kept as it is due to their importance in real world prediction scenarios, especially when they reflect genuinely high-cost individuals like chronic patients and smokers with high risk.

Once the cleaned dataset was ready, Exploratory Data Analysis (EDA) was performed. In this phase , The focus was on improving the quality and usability of data by creating new features , standardizing inputs and preparing the dataset for training the machine learning model.

Categorical Binning:

Here the BMI and age were categorized into new groups to gain more insights and potential improvements in prediction. For BMI, a new column was created using the standard BMI classification:

1. Underweight category where BMI is less than 18.5.
2. Normal category where BMI lies in the range 18.5 to 24.9
3. Overweight category where BMI lies in the range 25 to 29.9
4. Obese category where BMI is greater than 30

Similarly, age was also categorized into 3 groups which were young ( 18 -30), Middle-aged (31 – 50) and Senior (51 – 64).

This categorization can now assist in interpretation of predictions more effectively.

Derived Feature – Risk Factor:

A new feature called risk factor was created by multiplying the smoker status ( 1 or 0) with bmi. This is a synthetic features which was specially designed to capture the association effect of smoking with obesity which are the two major risk contributors in one metric. For example, a high BMI with a smoker status of 1 indicates a high risk patient which associates with high charges.

Feature Scaling:

To prepare the data for ML models, especially which are sensitive to feature scale, standardization was applied. The age and bmi columns were normalized so that they have zero mean and unit variance.

Clustering with K-Means and Risk Segmentation:

Before building any machine learning models, I wanted to understand the patterns in the data more clearly. So an unsupervised learning technique called Kmeans Clustering was applied. This technique helped us group similar patients together based on their personal features, without using the charges column. The aim was to find hidden groups or segments in the population who have similar health and risk profiles.

The three important features used for clustering were : age, BMI, and smoker ( 1 for smoker, 0 for non smoker)

These three features are known to influence medical expenses heavily.

Clustering algorithms like Kmeans are very sensitive to the scale of features. For example, age values range from 18 to 64, BMI from 15 to 53, while smoker is only 0 or 1. If we don’t scale the data, features with bigger numbers might have affected the clustering more than smoker, even though smoking is an important factor in prediction of health risks.

To solve this, Standard Scaler from scikit – learn was used, which transforms each feature so it has a mean of 0 and a standard deviation of 1.

Next, the KMeans clustering was applied with number of clusters equal to 3. The algorithm tried to make each group as different as possible from others while keeping patients inside a group as similar as possible.

Once the clustering was performed, every patient was given a cluster ID (0,1,2). To make these labels easier to understand, a new column called Cluster label was created and mapped these cluster IDs to more meaningful names such as :

Cluster 0 : High Risk Smokers

Cluster 1 : Average Risk Non Smokers

Cluster 2 : Low Risk Healthy Individuals

Cluster Analysis:

These cluster labels were grouped and calculated the mean of key features to understand each group better.

In High- Risk Smokers (Cluster 0), all patients in this group were smokers. They also had a slightly above average BMI and paid the highest medical insurance charges on average which were around $32050. Which accurately shows smoking combined with BMI contributes to heavy medical costs.

In Average Risk Non Smokers ( Cluster 1 ), in this group the patients were non smokers but had a little higher average age and BMI. Their charges were moderate around $11,624. This group represented the average population who were older in age but were not involved in behaviors like smoking.

In Low Risk Healthy Individuals (Cluster 2) , this group had people who were younger and non smokers with low BMI. As expected, their insurance charges were the lowest around $5,065 which indicates this was a healthy and low risk group.

This cluster analysis, we could observe the clear differences in behaviour and risk across the dataset. This step was crucial with respect to the fairness analysis because it helped to see if certain groups are more likely to be heavily charged than others.

Visualizing clusters with Principle Component Analysis :

Principle Component Analysis is a technique that reduces the number of dimensions ( features) while retaining the most important features. Since there were 3 features which were age, BMI and smoker status, PCA converted them into 2 principle components so we could visualize them into a 2D scatter plot.

Each dot on this plot, represented one patient and the color showed which cluster they belonged to. The resulting plot clearly separated the 3 distinct groups, with very little overlap. This means that the Kmeans clustering algorithm worked very well and was able to separate patients based on the selected features.

This step was valuable because:

1. Helped understand data diversity which tells that patient population is not uniform.
2. Validated the features which were age, BMI, and smoking which are strong features in influencing insurance charges.
3. By looking at these clusters, we can later ask whether there is fairness among all groups when we predict costs. And if smokers or the elderly were unfairly penalized.

Next steps:

Now that dataset has been successfully explored, cleaned , features were engineered and patient clustering was perfomed. The next phase of the project would be to focus on building a machine learning model which will help ensure fairness and interpreting model decisions in a transparent and ethical way.

Firstly, we would build and train multiple regression models like XGBoost Regressor and Random Forest Regressor to predict medical insurance charges. Also compare to see which one performs the best.

Second, model evaluation would be performed using metrics like R square Score, Root Mean Square Error, and Mean Absolute Error. These metrics would be helpfult o understand the accuracy and consistency of the model across different patients.

Also, SHAP (Shapley Additive Explanations) would be used to analyze every feature and their importance. This would help us understand how the model is making predictions for different individuals.

Also a fairness audit could be conducted to chect whether certain groups like smokers vs non-smokers, men vs women, regions are being unfairly charged.

Further, Outlier Handling could be investigated whether very high cost patients should be handled differently to avoid skewing the model’s predictions.