STAT 4600 Project Proposal

Project: **Medical Insurance Forecasting**

Team Members: Kongara Bargavi, Sharma Shailja, Adigarla Srivalli, Ahuja Prakhar Kumar

1. The research problem

For a health insurance company to make money, it needs to collect more in yearly premiums than it spends on medical care for its beneficiaries. It is crucial to forecast medical expenses for the insured population. Medical expenses are difficult to estimate because the costliest conditions are rare and seemingly random. Still, some conditions are more prevalent for certain segments of the population. We decided to research the effect of people's age, sex, BMI, Children, smoking, and region on medical expenses such as diagnosis, treatment and drug costs.

Objectives: To assess the average medical care expenses of people across 4 different regions (northeast, northwest, southeast and southwest) of the United States of America, who are aged between 18 and 64.

2. Why is this research problem important

To begin with, the purpose of this research problem is to predict medical expenses through variables shown in the data: age, sex, BMI, Children, smoking, and region – to increase accurate results for medical insurance companies. This research is related to finding the average medical care expenses and it is beneficial for insurance companies. For example, we've chosen the sum of people's age in all regions and the sum of medical charges to generate a comparative line chart, shown in Figure 1. We clearly see a trend that people who smoke are having more medical expenses. Also, the visualization result tells us age is also a driving factor to determine the premiums.

Additionally, by analyzing the correlation between a smoker and medical charges, we did further research on medical expenses. Meanwhile, the dataset also allows us to inspect the relationship between BMI and charges. This study helps medical insurance companies to predict medical expenses and decide the insurance premiums and therefore reduce losses.

The article (Forecasting Health expenses to U.S Medicare system, 2002) treats the uncertainty of long-term forecasts for health spending more systematically than in the past through a stochastic approach. Also, it uses the relation of health care spending to time until death to incorporate the changing health status of the population in the forecasts. Similarly, our project utilizes a dataset that establishes a dependency on different parameters like people's age, sex, BMI, Children, smoking, and the region on medical charges.

Furthermore, changes in healthcare expenditure appear quickly after changes in smoking behavior. A 10% relative drop in smoking in every state is predicted to be followed by an expected $63 billion reduction (in 2012 US dollars) in healthcare expenditure the next year. (Smoking Behavior and Healthcare Expenditure in the United States,2016). In our project, the research also focuses on smoker's expenditure on medical charges.

Lastly, predicting medical insurance costs is still a problem in the healthcare industry that needs to be investigated and improved. In this paper (A Computational Intelligence Approach for Predicting Medical Insurance Cost, 2021), by using a set of ML algorithms, a computational intelligence approach is applied to predict healthcare insurance costs. We will also be doing regression analysis on our dataset for various variables to determine the medical expenses. This will help the insurance companies to decide their premium charges.

3. Type/source/content of data used in the project

The type of dataset that will be used in this project is CSV format.
Data Source: https://www.kaggle.com/code/shrutidandagi/medical-insurance-forecast-by-linear-regression/data
The dataset contains charges, age, BMI, sex, children, and smoking habits of people in 4 regions. We do see that no values are missing in the dataset. Hence, we decided to consider the entire data set which has 1338 rows and 7 columns.

Variables	Description	Values/Range	Type
age	age of beneficiary	18 - 64	int64
Sex	gender of beneficiary	Male, Female	object
BMI	Body mass index, providing an understanding of the body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight	15.96 - 53.13	float64
children	Number of children covered by health insurance / Number of dependents	0 - 5	int64
smoker	the smoking habit of the beneficiary	yes, no	object

region	the beneficiary's residential area in the US, northeast, southeast, southwest, northwest	northwest, northeast, southeast, southwest	object
charges	Medical Insurance Cost	1121.8739 - 63770.42801	float64

For our analysis, we decided to choose all 7 variables. A variable named Smoker tells us about the smoking habits of a person and it is a major cause of diseases such as cancer, heart disease, cerebrovascular and respiratory disease. The health hazards of smoking lead smokers to consume more medical resources than non-smokers and thus greater medical expenditure.
Other variables like age, children, and BMI also determine the cost of the treatment. Medical expenses would be changed based on the insured person age as well.

4. Analysis and Modeling methods

In this data set, several factors which affect medical expenses are taken into consideration.
The following 2 methods will be used to analyze this dataset:
Multiple Linear Regression: In the data set, we have multiple independent variables and one dependent variable. Hence, we will be using Multiple Linear Regression. The dependent variable(Y) is charges, and the independent variables(X) are age, sex, BMI, children, smoker and region.
Random Forest Model: The random forest model provides a higher level of accuracy in predicting outcomes over a single mode.

5. The software that you use

We, as a team decided to use python and excel for our data analysis as they are widely used for data analysis. Python is a popular multi-purpose programming language widely used for its flexibility, as well as its extensive collection of libraries, which are valuable for analytics, visualizations and complex calculations.
Excel allows us to quickly modify rows and columns before our analysis. We can even visualize the data by using the default chart options that are provided in excel.

6. Summary or descriptive statistics to support this proposal

We decided to work on insurance dataset to research the relationship between different independent variables including age, sex, BMI, children, smoker etc., and dependent variable charges. we will run the regression and random forest models on our dataset. Last, we use descriptive, diagnostic, predictive, and prescriptive methods to analyze our results.

We have a few initial results we would like to show as in **Figure 1, Figure 2, Table 1, and Table 2.** In Figure 1, we see that the smoking status of the insured is the main variable to predict the medical cost as it shows more difference in the charges. In Figure 2, we see that age is also a driving factor for medical expenses. In Tables 1 and 2, the linear regression model can significantly explain the relationship between smoker and charges based on the 0.000000 p-value and R Square is 0.619, which means at least 60% of data can be explained by the model. This indicates the relationship between smoker and charges is positive. We now have moderate confidence to say that there are relationships among all variables, and we will analyze more to finalize our project.

Last, our descriptive data is shown in **Table 2**. For the dataset that we have taken, the Mean value of charges is 13270.42227, while the Median is 9382.033 and the Mode resulted in 1639.5631. Kurtosis is 1.606298653, which means the distribution is longer or fatter, while the Skewness is 1.515879658, which means that data is highly skewed.

Figure 1:

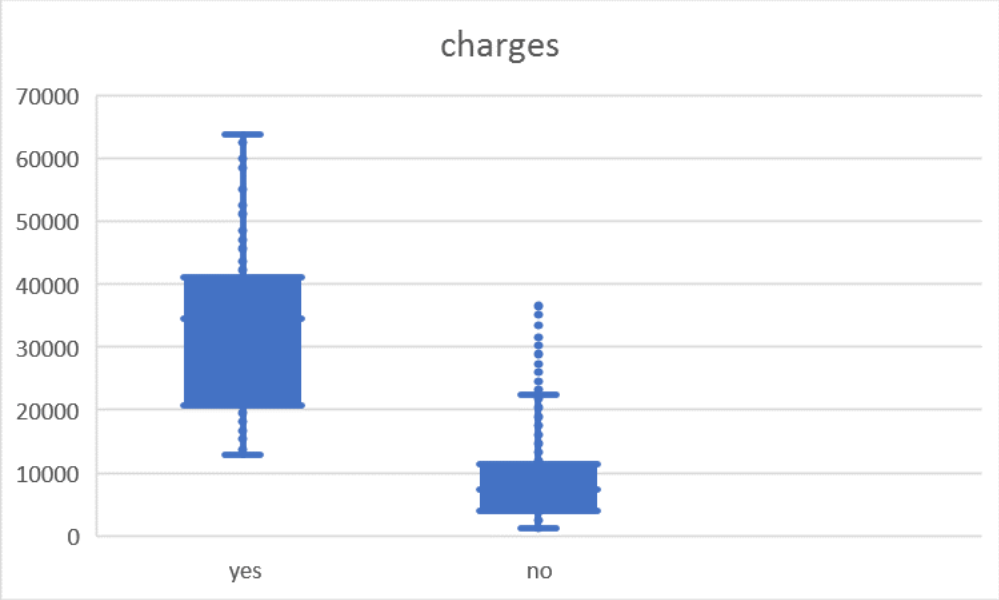

Figure 2:

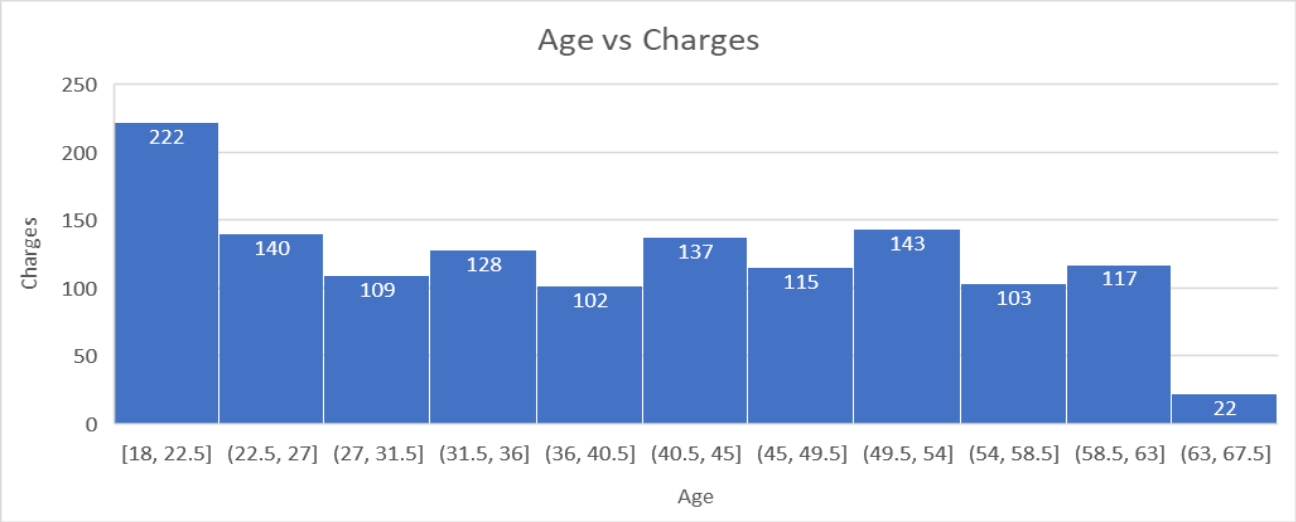

Table 1:

Regression Model of smoker and Medical Charges

Regression Statistics	
Multiple R	0.787233058
R Square	0.619735888
Adjusted R Square	0.619451046
Standard Error	7470.755405
Observations	1337

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	121431569120.11	121431569120.11	2175.72	0.00
Residual	1335	74509268736.88	55812186.32		
Total	1336	195940837856.98			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	8440.579414	229.1384049	36.83616205	0.00	7991.06885	8890.08997	7991.06885	8890.089973
smoker	23609.65242	506.1605667	46.64459062	0.00	22616.6957	24602.6091	22616.6957	24602.60914

Table 2: Regression Model of Medical Insurance Forecast

Descriptive Statistics

charges	
Mean	13270.42227
Standard Error	331.0674543
Median	9382.033
Mode	1639.5631

Standard Deviation	12110.01124
Sample Variance	146652372.2
Kurtosis	1.606298653
Skewness	1.515879658
Range	62648.55411
Minimum	1121.8739
Maximum	63770.42801
Sum	17755824.99
Count	1338

References:

Ch. Anwar ul Hassan, Jawaid Iqbal, Saddam Hussain, Hussain AlSalman, Mogeeb A. A. Mosleh, Syed Sajid Ullah, "A Computational Intelligence Approach for Predicting Medical Insurance Cost", Mathematical Problems in Engineering, vol. 2021, Article ID 1162553, 13 pages, 2021. https://doi.org/10.1155/2021/1162553

Lightwood J, Glantz SA (2016) Smoking Behavior and Healthcare Expenditure in the United States, 1992–2009: Panel Data Estimates. PLoS Med 13(5): e1002020. https://doi.org/10.1371/journal.pmed.1002020

Ronald Lee, Ph.D., An Approach to Forecasting Health Expenditures, with Application to the U.S. Medicare System, 2002, https://doi.org/10.1111/1475-6773.01112