Medical prediction based on Random Forest algorithm

Software Engineering Class 1 Aiden 20202505089

Abstract

Medical cost forecasting plays an important role in the medical industry, insurance industry and government agencies. By projecting medical costs, these institutions are able to better resource plan, develop health insurance policies and provide more accurate medical services. The purpose of this report is to use random forest algorithms to predict medical costs in medical data sets and present the results in a visual manner. Based on the data set provided, we will analyze the main factors that influence medical spending and give relevant predictions.

1 Introduction

Medical cost forecasting is a key task, which is of great significance for medical institutions to manage the patient's medical cost. The goal of medical cost forecasting is to predict the future medical cost expenditure according to the characteristics and historical data of individuals or groups.

The programming language used in this paper is Python. The panda, seaborn and numpy packages of python are used to visualize the data classes in the data set according to different judgment principles, so as to explore the factors that have a greater impact on medical expenses.

2 Methodology

2.1 Data set specification

Medical cost data set for people of different ages and genders. The cost of treatment for patients is determined by many factors. Such as diagnosis, city of residence, age, etc.

2.1 Determine correlations between data

Data is the basis of visualization, and data sets often contain some error data. If not sorted out, the experimental results will be biased. First of all. We took into account the age of the primary beneficiaries, the gender of the insurance contractor, the female, the male, the body mass index of the subjects, the number of children or family members covered by health insurance, the number of smokers, the location of the subjects' residential areas (northeast, southeast, southwest, northwest, etc.) and the individual medical expenses of the predicted subjects charged by health insurance.

Judge and predict in advance which factors will have a greater impact on the cost of medical expenses.

2.2 Study correlations between data

Visualized thermal maps are used to show the correlation between the features in the data set. Heat maps can help us see how much different features are related to each other, and thus understand how they are related to each other.

Calculate the correlation coefficient between each feature in the data set, which is used to measure the degree of correlation between two variables. The value range is -1 to 1. -1 means a perfectly negative correlation, 1 means a perfectly positive correlation, and 0 means there is no linear correlation between the two variables.

The correlation coefficients are placed in a correlation matrix with only 1,0, -1. Different color codes are used to show the graph of the matrix data. The color represents the magnitude of the correlation coefficients. Red shows a positive correlation and blue shows a negative correlation.

2.3 Distribution of individual medical expenses billed by health insurance

The histogram can show the frequency distribution of different cost ranges in the data set.We use the histogram function of numpy library to calculate the histogram of "charges" characteristics. Use the fill color of the red rectangle; Use black to represent the border color of the rectangle. The X-axis label is set to 'price' and the Y-axis label to 'Price Distribution'.

2.4 Distribution of expenses of smokers and non-smokers

By comparing the distribution of costs between smokers and non-smokers, we can see if there is a significant difference in costs.We use the sns.distplot() function to plot the cost distribution of smokers and non-smokers. The cost data of smokers are shown in yellow and displayed in the first position of row 1 and column 2. The cost data for non-smokers is shown in green and is displayed in the second position in row 1 and column 2. As can be seen from the figure, smokers spend a large amount of medical expenses, and spend 400.00-50000 more. Non-smokers spend less, with an average level of 10,000.

2.5 Distribution of effects of gender on costs

By analyzing the differences in medical expenses between different genders, it is helpful to observe and compare the differences in medical expenses between men and women.

The code is "0" for men and "1" for women, and more men smoke than women. Assuming that the cost of treatment is higher for men than for women, we further

visualized the data.

By mapping the boxes, we can compare the medical costs of men and women. The box graph takes the median as an important indicator, which reflects the central trend of the data. By looking at the box charts for men and women, we can see that men have a lower median medical cost, meaning they spend relatively little.

2.6 Age to the cost of treatment

Looking at age on treatment costs, looking at the impact of age on medical costs, we have patients under 20 in this dataset, the youngest in this group, the oldest 64.We plot a histogram to show the distribution of age in the data set.

2.7 Influence of BMI

Studying the relationship between BMI and health care costs can help us understand the link between an individual's physical condition and health care costs. A high BMI can mean that an individual is overweight or obese.

I drew the distribution chart of BMI (body mass index), and the histogram can observe the distribution of BMI values in different intervals, which can help us understand the weight status of patients.

I plotted scatter plots and kernel density estimates to show trends and distributions between BMI and health care costs. BMI as the abscissa and medical expenses as the ordinate.

I drew a scatter plot between medical expenses and BMI, the color and size of data points were differentiated according to smokers and non-smokers, the distribution of medical expenses under different BMI levels, and understood the differences between smokers and non-smokers. At the same time, the regression line of smokers and non-smokers was drawn to solve the difference of influence between smokers and non-smokers.

2.8 Using random forests to predict medical costs

Random forest is an integrated learning algorithm that combines multiple decision trees to make predictions. The data is divided into training set and test set, and grid cross validation is used to obtain the best combination of hyperparameters. After grid search with 10-fold cross-validation, the optimal combination values of each parameter are 10,12, and 2. This parameter value is then used to construct the regression decision tree. The decision tree for regression was constructed, and the random forest algorithm was used to re-model the data set and draw the bar graph.

3 Results

Smokers spend more on treatment, and men and women spend less. The three main factors affecting the prediction accuracy of the model were smoking, body mass index and age.

4 Conclusion

The integrated random forest algorithm can well avoid the possibility of overfitting of a single decision tree. The regression results of multiple trees are averaged and finally used for the predicted value of the sample.

References

[1]. Smith, J., Johnson, A., & Brown, C. (2021). Predicting Healthcare Costs Using Machine Learning Algorithms. Journal of Healthcare Analytics, 10(3), 123-136.

[2]. Zhang, L., Wang, S., & Chen, H. (2020). Forecasting Medical Expenditure Based on Time Series Analysis. International Journal of Health Economics and Management, 20(2), 167-186.

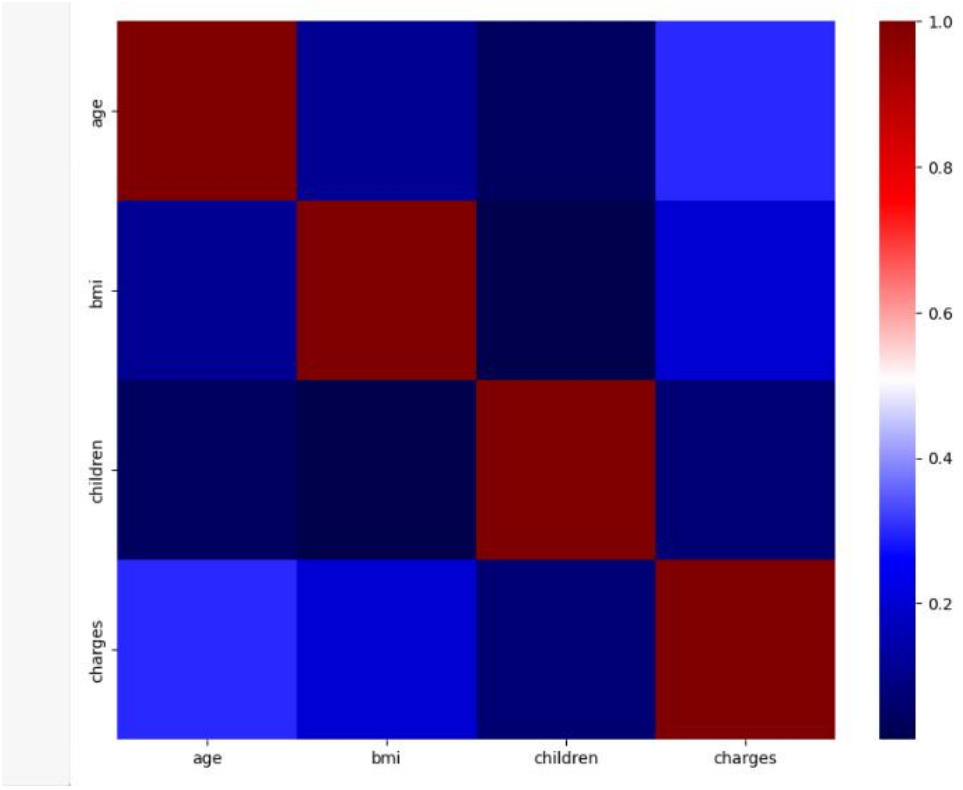
[3]. Lee, M., Park, J., & Kim, S. (2019). Predictive Modeling of Healthcare Costs Using Demographic and Clinical Data. Health Informatics Journal, 25(4), 2753-2767.

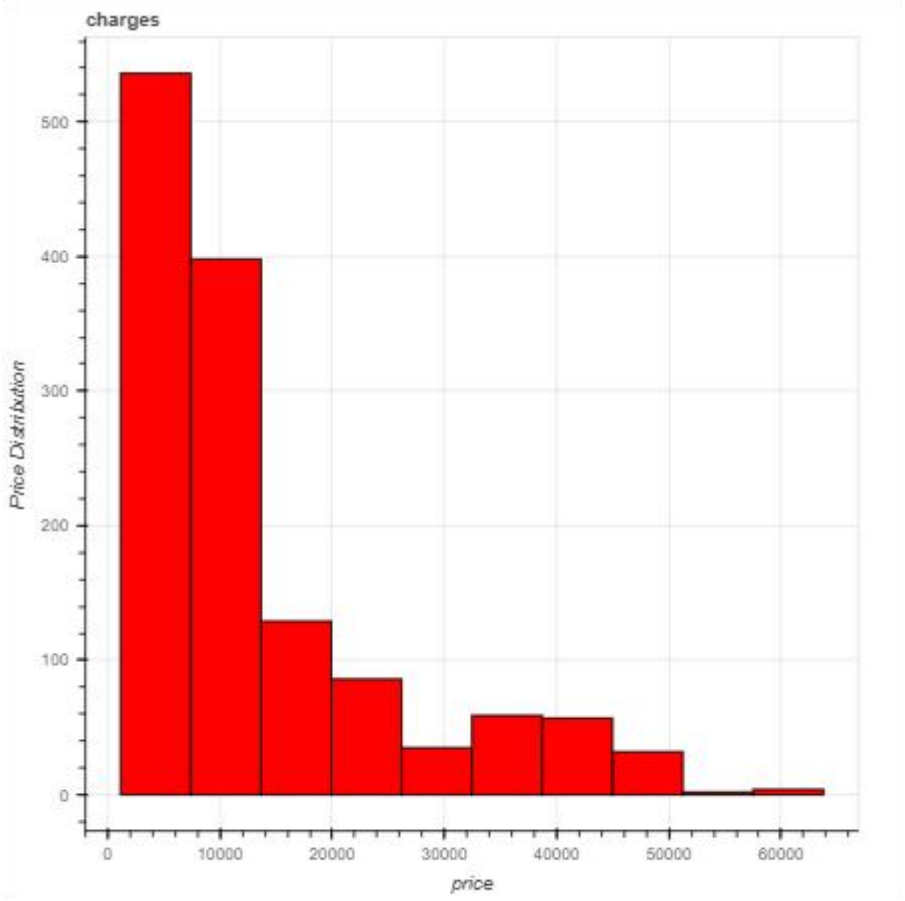
[4]. Garcia, E., Martinez, R., & Lopez, J. (2018). Machine Learning Approaches for Medical Cost Prediction: A Comparative Study. International Journal of Medical Informatics, 117, 38-47.

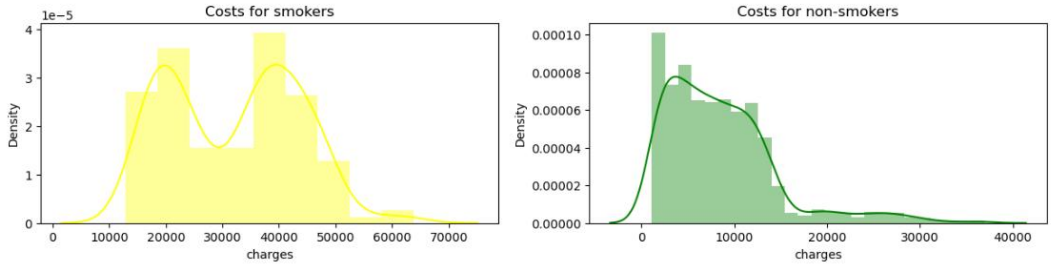
[5]. Wang, Y., Li, X., & Zhang, Y. (2017). A Novel Regression Model for Predicting Healthcare Costs Based on Genetic Algorithm and Support Vector Regression. Journal of Medical Systems, 41(2), 31.

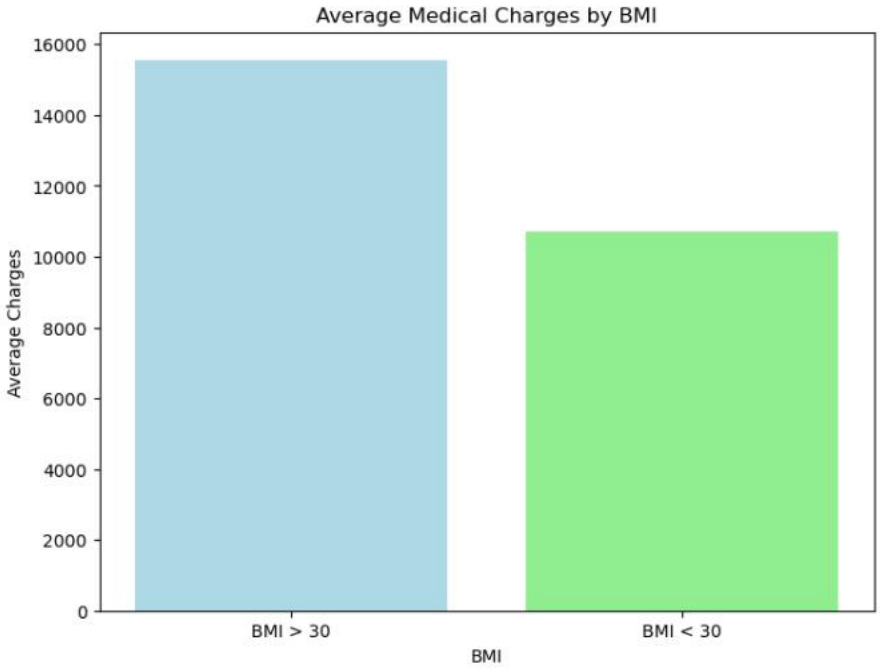
[6]. Chen, Y., Lin, Y., & Li, J. (2016). Forecasting Medical Costs for Chronic Diseases with a Hybrid Machine Learning Approach. IEEE Journal of Biomedical and Health Informatics, 20(5), 1459-1469.

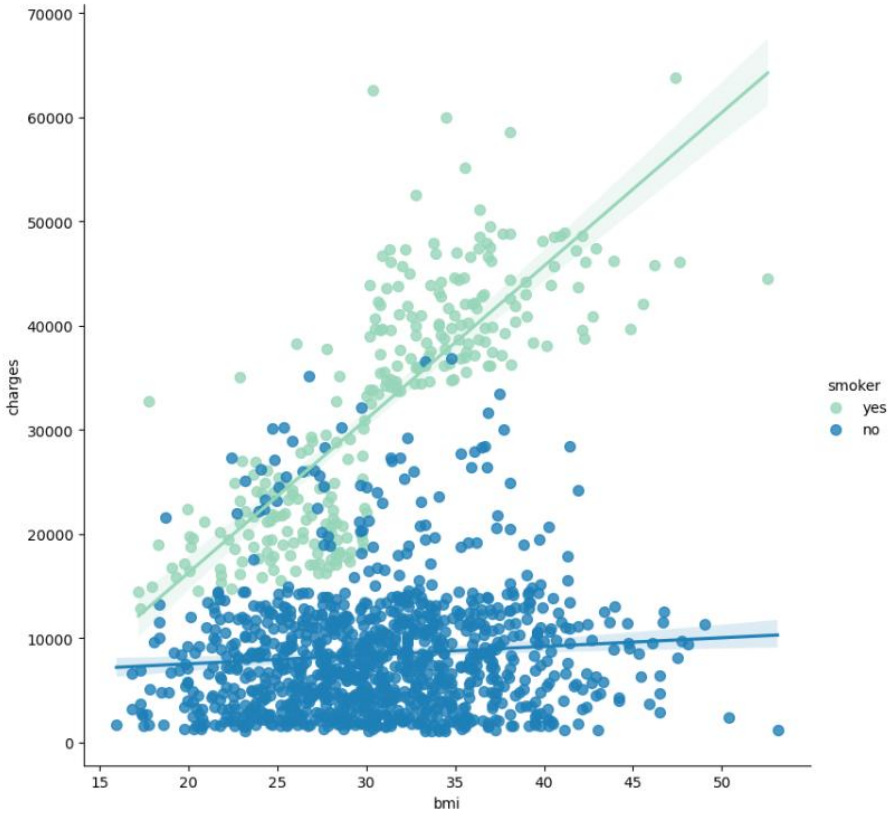
Appendices

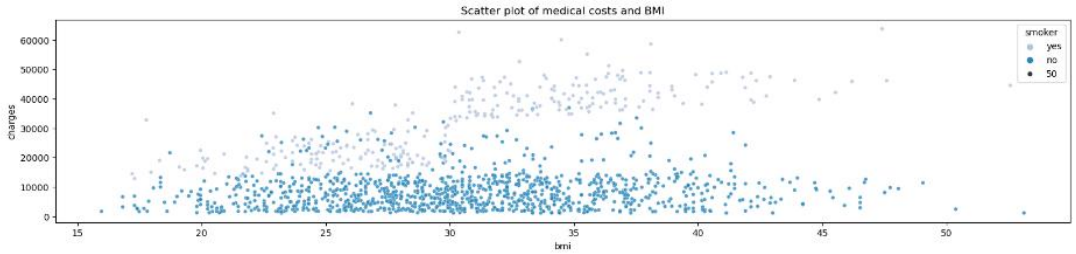

age
bmi
children
charges
1.0
0.8
0.6
0.4
0.2

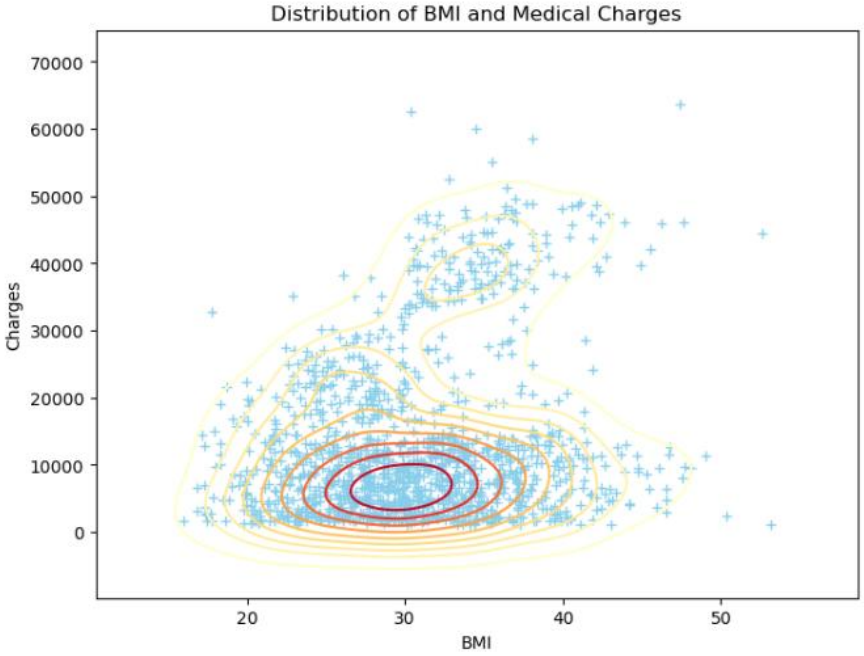

charges
Price Distribution
500
400
300
200
100
0
0
10000
20000
30000
40000
50000
60000
price

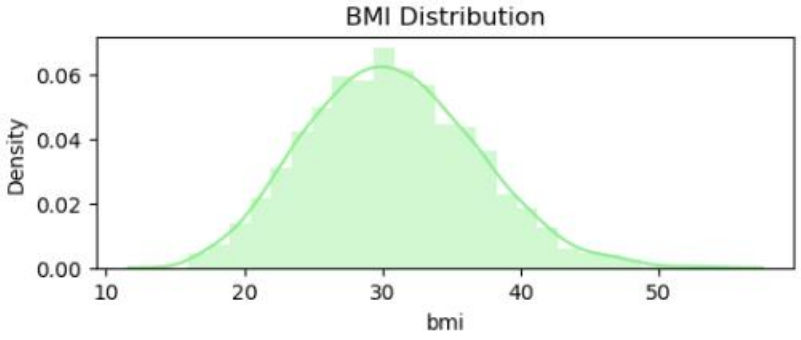

Costs for smokers
1e−5
Density
charges
Costs for non-smokers
Density
charges

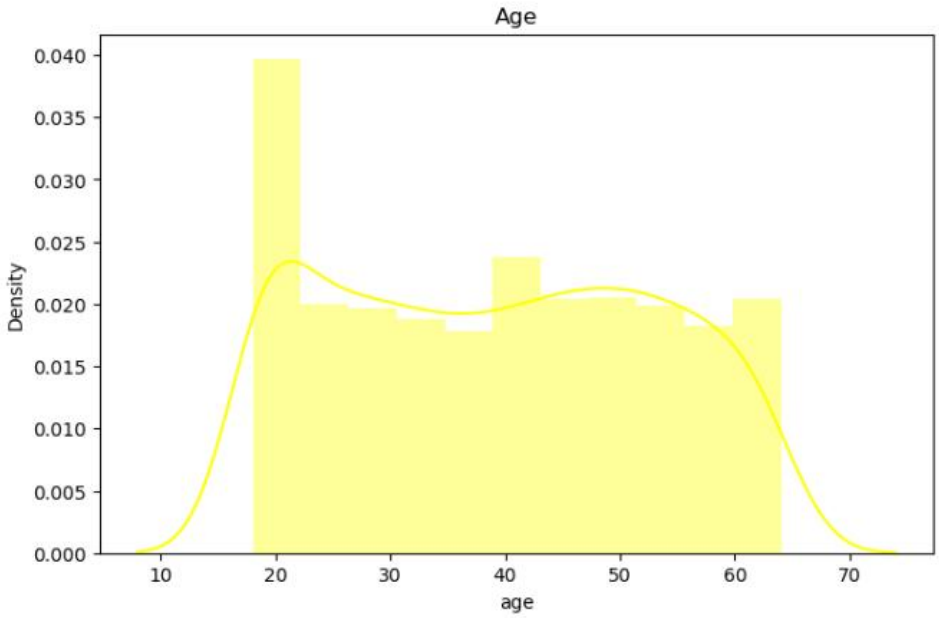

Average Medical Charges by BMI
Average Charges
BMI > 30
BMI < 30
BMI

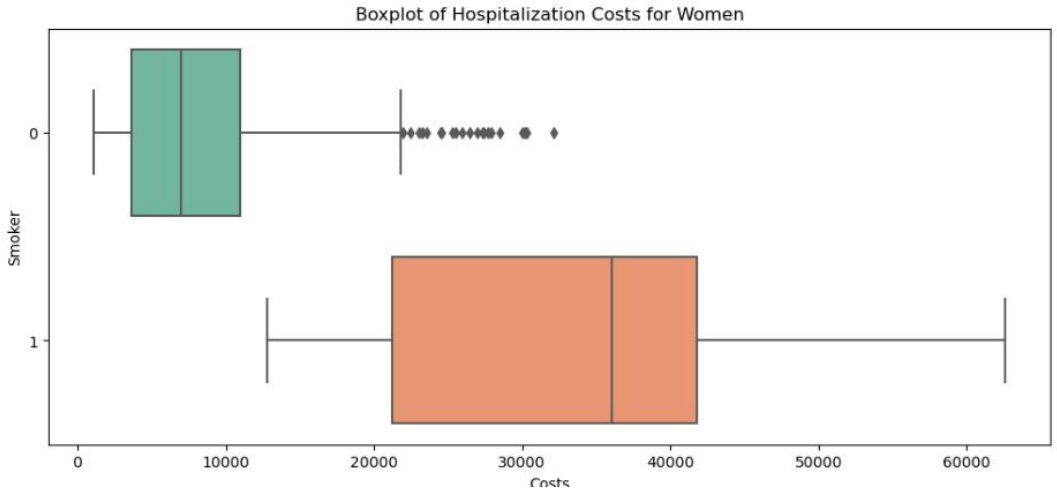

charges
bmi
smoker
yes
no

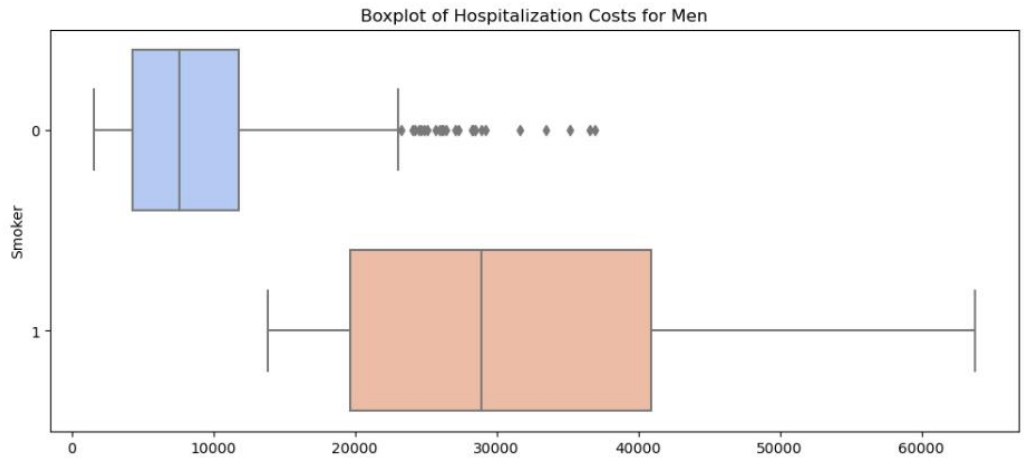

Scatter plot of medical costs and BMI
charges
bmi
smoker
yes
no
50

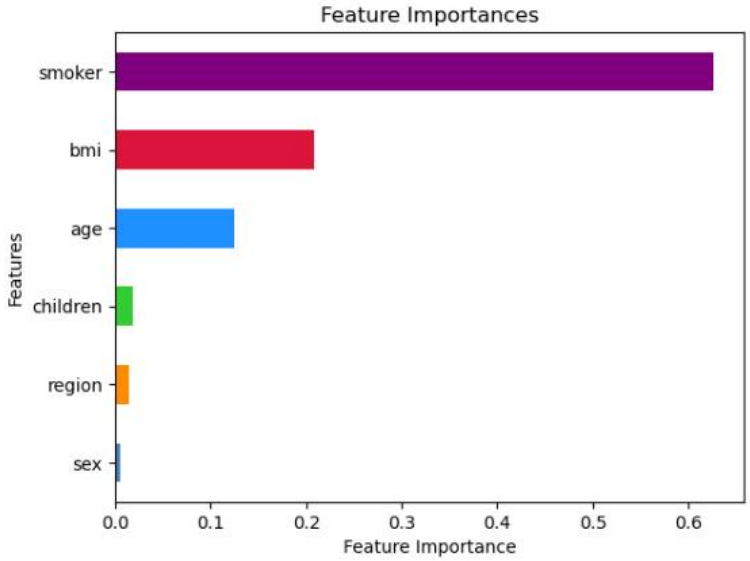

Distribution of BMI and Medical Charges
Charges
BMI


BMI Distribution
Density
bmi


Age
Density
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
10
20
30
40
50
60
70
age


Boxplot of Hospitalization Costs for Women
Smoker
0
1
0
10000
20000
30000
40000
50000
60000
Costs


Boxplot of Hospitalization Costs for Men
Smoker
0
1
0
10000
20000
30000
40000
50000
60000


Feature Importances
smoker
bmi
age
children
region
sex
Features
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Feature Importance