Final Data Analysis Project

Srijesh Reddy Yarram

AA-5300-12: Advanced Analytics

Prof. Michael Fisher

March 12, 2023

Introduction:

The Health Insurance cost prediction project is aimed at analyzing the given dataset and building predictive models to accurately predict the cost of providing medical insurance to different groups of people. The dataset contains variables such as age, gender, body mass index (BMI), number of children, smoking habits, residential area, marital status, and individual medical costs billed by health insurance.

1. Overview of the dataset:

a) Contextual information:

Source of the data set: Present in Kaggle, GitHub and in UCI Machine Learning Repository.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

https://www.kaggle.com/datasets/mirichoi0218/insurance?ref=hackernoon.com

This is an open data set which is publicly available for research purposes only. Therefore, I am permitted to share this data for my project.

Purpose: Data analytics can help insurance companies to accurately predict the cost of providing medical insurance to different groups of people. This information can be used to set premium prices, design insurance plans, and identify areas where cost savings can be made. Predictive analytics can also help insurance companies to identify high-risk patients and design targeted interventions to prevent or manage chronic conditions, which can lead to lower healthcare costs and improved health outcomes.

Questions that audience interested in the dataset and its analyses might seek to see answered:

- Is there a statistically significant relationship between age and medical costs?
- What is the impact of gender on medical costs, and is this impact statistically significant?
- Does smoking have a significant impact on medical costs, and if so, what is the magnitude of this impact?
- Is there a statistically significant relationship between BMI and medical costs?
- Is there a statistically significant relationship between residential area and medical costs?
- What is the impact of the number of children on medical costs?

b) Variables present:

The dataset contains seven predictor variables and one outcome variable.

S.No.	Variable name	Description	Type
1	age	Age in years	Numeric
2	sex	Gender of the individual	Binary
3	bmi	Body mass index	Numeric
4	children	Number of children covered by insurance provider	Numeric
5	smoker	Smoking = yes or no	Binary
6	region	the beneficiary's residential area in the US, northeast, southeast, southwest, northwest	Categorical
7	Charges (Outcome variable)	Individual medical costs billed by health insurance	Outcome variable, Numeric
8	Marital stat	Married or not	Binary

2. Types of analyses:

Analytical techniques applicable for regression:

Regression techniques are suitable for the given dataset because they aim to model the relationship between the predictor variables and the continuous outcome variable. In other words, they try to find the best possible linear or non-linear equation that predicts the cost of medical insurance based on the other variables in the dataset. Regression models can handle both continuous and categorical predictor variables, making them versatile and flexible for various types of datasets. They can also capture non-linear relationships between predictor variables and the outcome variable, making them capable of modeling complex data patterns.

Regression algorithms like Random Forest, Lasso, Ridge, K-Nearest Neighbors, and Multiple Linear regression techniques are suitable for predicting continuous variables. These algorithms aim to create a relationship between the predictor variables (age, gender, BMI, smoking habits, etc.) and the outcome variable (cost of providing medical insurance) to make accurate predictions.

In contrast, Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are classification algorithms that are typically used for predicting categorical variables, not continuous variables. In the given dataset, the outcome variable is the cost of providing medical insurance, which is a continuous variable, making it unsuitable for LDA and QDA. Also, Naive Bayes is a classification algorithm that predicts the class of an observation based on the probabilities of the predictor variables. It is not suitable for predicting continuous outcome variables like the cost of medical insurance.

Therefore, it is more appropriate to use regression algorithms to build predictive models for the given dataset, and LDA and QDA are not suitable for this analysis.

Analyses

1. Overview:

The following table presents the applicable analytical methods, whether we intend to use them, and the rationale behind using or not using the method.

Method	Intend to use	Rationale
Random Forest	Yes	It can handle missing values, noisy data, and high-dimensional data well. Additionally, it can capture non-linear relationships.
Lasso	Yes	It performs feature selection and reduces the dimensionality of the data. It is less prone to overfitting compared to Linear regression.
KNN	Yes	It can be useful when there is a relationship between the target and predictors that is smooth and local.
MLR	Yes	It can handle multiple predictor variables and model

		their linear relationship with the outcome variable. It provides interpretable coefficients for each predictor variable, making it easier to understand their impact on the cost of medical insurance.

PRE-PROCESSING STEPS:

Principal Component Analysis /dimensionality reduction:

It is a technique used to reduce the number of variables in a dataset by identifying the most important variables that explain much of the variation in the data. It can be useful in cases where there are many variables that are highly correlated, which can lead to multicollinearity issues in the predictive models.

However, in the case of the Health Insurance cost prediction project, the number of variables is relatively small (less than 10), and each variable has its own unique contribution to the prediction of the medical insurance cost. Therefore, there is no need for dimensionality reduction since it would not significantly improve the model's performance or reduce the complexity of the problem.

Moreover, the use of PCA/dimensionality reduction may result in the loss of important information and interpretation of the variables, which could lead to biased and inaccurate predictions. Therefore, it is not necessary to use PCA/dimensionality reduction in this project.

Centering and scaling

Centering and scaling were applied as pre-processing steps in the Health Insurance cost prediction project. This step was carried out to normalize the input data, ensuring that each feature contributes equally to the predictive models. By centering and scaling the data, all the features were transformed to have a mean of 0 and a standard deviation of 1, reducing the impact of outliers and improving the performance of many machine learning algorithms. It was crucial to reducing bias and improving the accuracy and robustness of the predictive models.

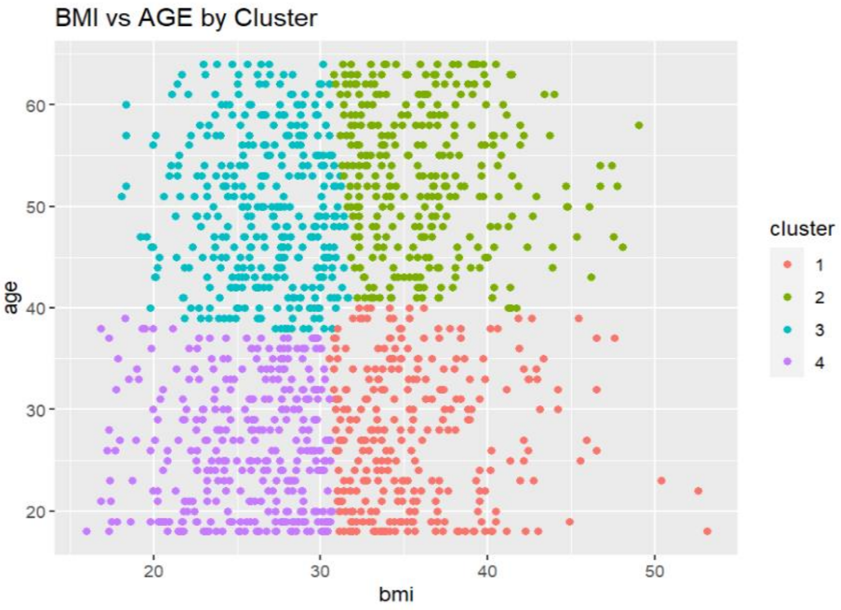
Cluster Analysis

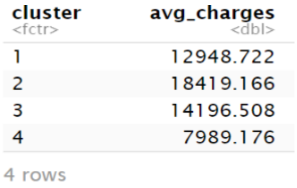
Loaded the required packages and the dataset. Then divided subsets with the relevant variables by excluding the categorical variables and selecting the numeric variables of interest. Also normalized the numeric variables to have zero mean and unit variance, which is often necessary for clustering algorithms.

Then performed k-means clustering with k=4, which partitions the data into 4 clusters based on the similarity of the numeric variables. Extracted the cluster assignments and added them to the original dataset as a new variable called "cluster".

Next, visualized the clusters using a scatter plot of BMI vs AGE, where each point is colored according to its cluster assignment. This allowed us to visually inspect the clustering results and identify any patterns or outliers.

Finally, computed the average charges by cluster using the group_by and summarize functions from the dplyr package. This gives us an idea of how the clusters differ in terms of their impact on the outcome variable "charges".





cluster <fctr>	avg_charges <dbl>
1	12948.722
2	18419.166
3	14196.508
4	7989.176

4 rows

Based on the output and the ranges of BMI and age associated with each cluster, we can make the following inferences:

Clusters 1 and 2 have similar ranges of BMI (approximately 30-50) but different age ranges, with cluster 1 having younger patients (approximately 20-40) and cluster 2 having older patients (approximately 40-60). This suggests that BMI is a significant predictor of charges, but age also plays a role, with older patients having higher charges.

Clusters 3 and 4 have similar ranges of age (approximately 20-40 for cluster 4 and approximately 40-60 for cluster 3) but different ranges of BMI, with cluster 4 having lower BMI (approximately 10-30) and cluster 3 having higher BMI (approximately 30-50). This suggests that BMI is still a significant predictor of charges, but age may not be as important in these clusters.

The difference in average charges between clusters is substantial, with cluster 2 having significantly higher charges than cluster 4. This indicates that other variables beyond BMI and age are likely contributing to the differences in charges across the clusters.

These findings can help us better understand the different patient profiles and associated charges in the dataset, and tailor interventions or services to different patient groups accordingly. For example, we may develop targeted interventions for patients in clusters 2 and 3 who have higher BMI and charges or focus on preventive measures for patients in cluster 1 who are younger and have higher BMI.

Subsampling:

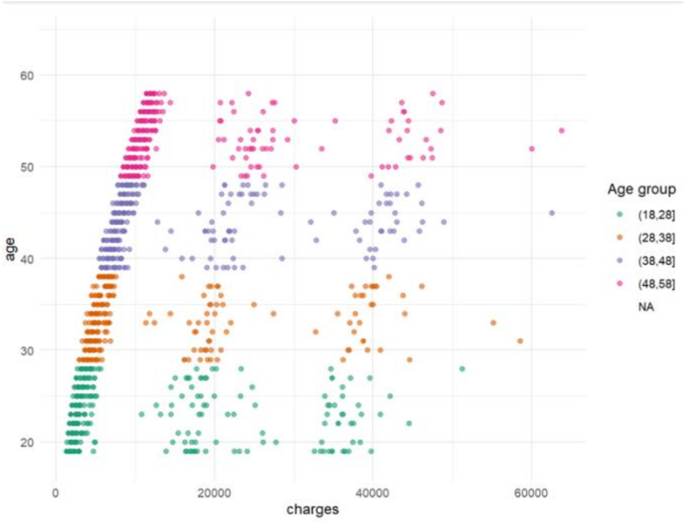
Subsampling is a technique that randomly selects a subset of observations to reduce the size of a dataset in machine learning. The technique can be useful when the dataset is very large, and the model training process is computationally expensive. However, in the Health Insurance Cost Prediction project, there were no concerns mentioned regarding the size of the dataset or computational efficiency. Therefore, subsampling was not necessary in this case.

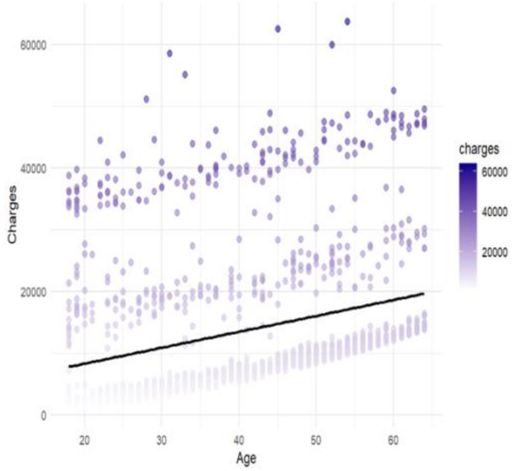
Furthermore, subsampling can sometimes lead to biased estimates and loss of information, as well as increase the variance of the model, resulting in overfitting. In the Health Insurance Cost Prediction project, the goal was to build accurate and robust predictive models, and subsampling may not have been the best approach to achieve these objectives.

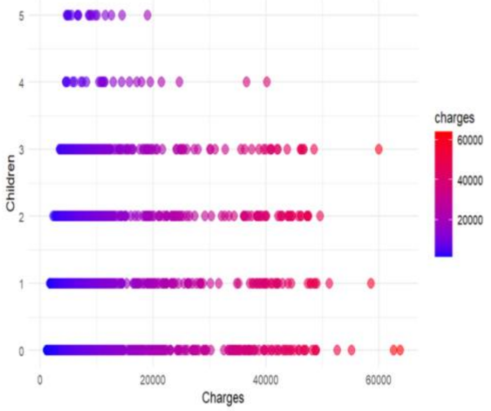
Thus, considering the project's requirements and objectives, subsampling was not required, and it is reasonable that the technique was not used.

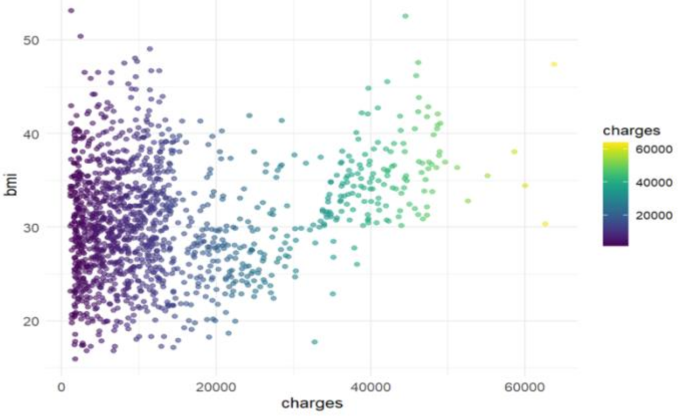
Data Visualization:

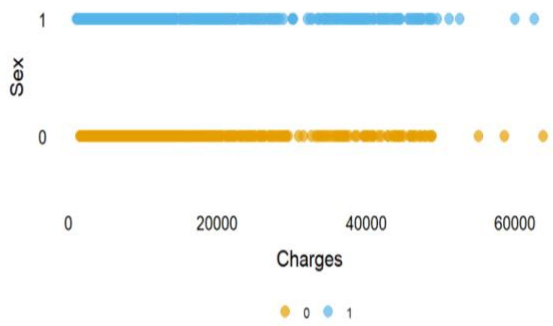
The visualizations revealed some interesting relationships between the predictor and outcome variables in the insurance dataset. The correlation matrix showed that charges are more correlated with age than any other variable. The visualization between charges and children showed that children between 0 and 2 are more susceptible to diseases as the charges for this age group are higher than those for children over 2 years old. The visualization between charges and BMI revealed that people with BMI values between 30 and 40 had undergone high insurance charges (above 30,000). The visualization between charges and age showed that charges increase as age increases. Finally, the visualization between charges and “smoker” revealed that smokers are more susceptible to higher insurance charges than non-smokers.

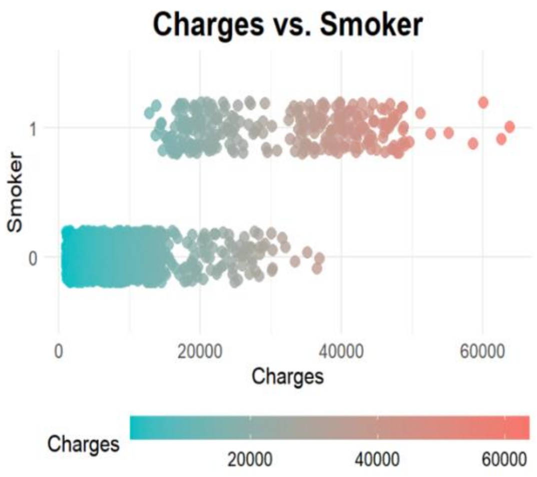


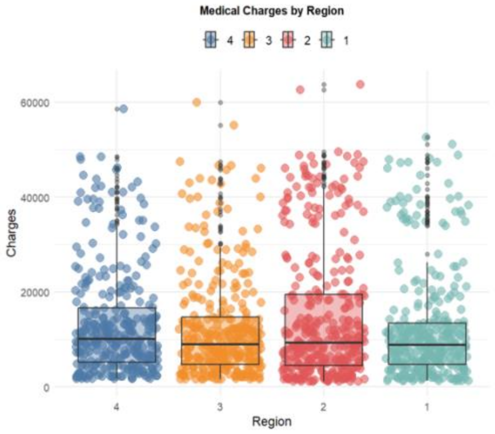


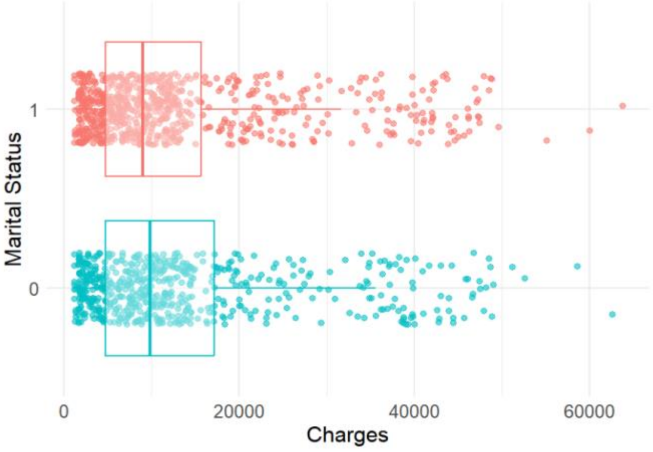












2. Summary of results:

Method 1: KNN

Details of Validation Method	Model Formula	RMSE	R-Squared	MAE
10-fold Repeated Cross-Validation, 10 Repeats, CARET code	charges ~.	0.4536	0.7828	0.2803

Explanation:

For KNN method, we used repeated cross-validation with k = 10 and repeats = 10. The model formula used was charges ~. (All variables included). The evaluation metric used was root mean squared error (RMSE), which measures the average deviation of the predicted values from the actual values. The best model was selected based on the lowest RMSE value, which was 0.4536. The R-squared value of 0.7828 indicates that the model explains 78.28% of the variance in the target variable (charges), and the mean absolute error (MAE) of 0.2803 indicates that, on average, the predicted charges were off by $280.3.

Method 2: Random Forest

Details of Validation Method	Model Formula	RMSE	R-Squared	MAE
10-fold Repeated Cross-Validation, 10 Repeats, CARET code	charges ~.	0.3067	0.9045	0.1834

Explanation: For random forest method, we used repeated cross-validation with k = 10 and repeats = 10. The model formula used was charges ~. (All variables included). The evaluation metric used was root mean squared error (RMSE), which measures the average deviation of the predicted values from the actual values. The best model was selected based on the lowest RMSE value, which was 0.3067. The R-squared value of 0.9045 indicates that the model explains 90.45% of the variance in the target variable (charges), and the mean absolute error (MAE) of 0.1834 indicates that, on average, the predicted charges were off by $183.4.

Method 3: LASSO Regression

Details of Validation Method	Model Formula	RMSE	R-Squared	MAE
10-fold Repeated Cross-Validation, 10 Repeats, CARET code	charges ~.	0.4309	0.8030	0.3044

Explanation: The LASSO regression model was selected based on its ability to handle multicollinearity and select the most relevant predictors. Evaluation was done using repeated cross-validation with RMSE, R-squared, and MAE metrics. The best model was chosen based on the lowest RMSE value of 0.4309, indicating good predictive performance. The R-squared value of 0.8030 suggests that the model explains 80.30% of the variance in charges. The MAE value of 0.3044 indicates an average absolute difference of $304.4 between predicted and actual charges. The LASSO model's range of applicability is limited to the predictor variables included in the model.

Method 4: Multiple Linear Regression

Details of Validation Method	Model Formula	RMSE	R-Squared	MAE
10-fold Repeated Cross-Validation, 10 Repeats, CARET code	charges ~.	0.4310	0.8029	0.3046

Explanation: The multiple linear regression model was chosen for its ability to capture the relationship between charges and predictors. Evaluation was done using repeated cross-validation with RMSE, R-squared, and MAE metrics. The best model was chosen based on the lowest RMSE value of 0.5076, indicating good predictive performance. The R-squared value of 0.7417 suggests that the model explains 74.17% of the variance in charges. The MAE value of 0.3507 indicates an average absolute difference of $350.7 between predicted and actual charges.

Here's a table summarizing the aggregate variable importance across all the models:

Predictor	Aggregate Variable Importance
smoker1	100.0000
age	32.8418
bmi	22.3905
children	4.1105
region1	2.1728
region2	2.0533
Marital stat1	1.5267
region3	0.4773
sex1	0.0648

The smoking status of the patient is by far the most important predictor in all the models, followed by age and BMI. The number of children, region, and marital status are also important predictors, but to a lesser extent. Finally, sex is the least important predictor in all the models.

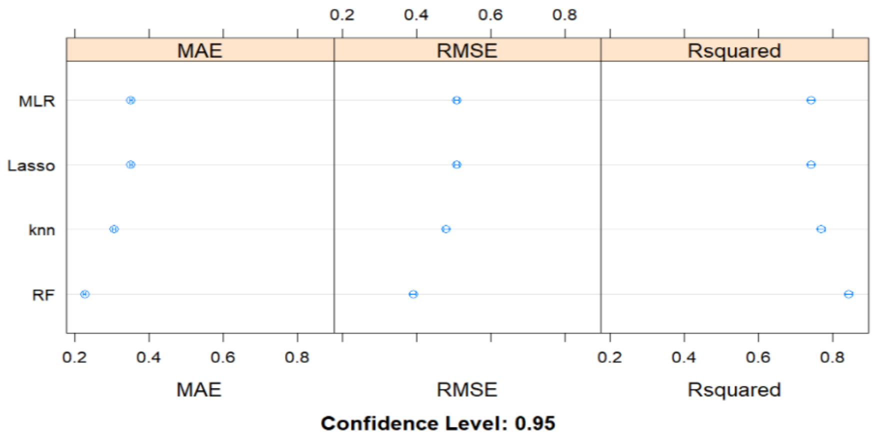
Conclusions :

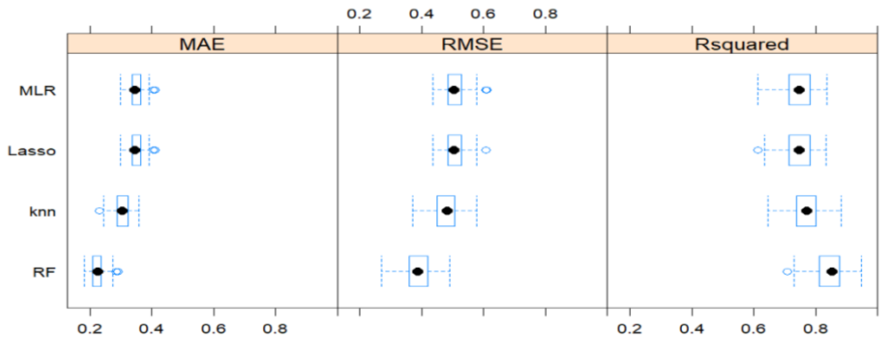
Here's a table comparing the evaluation metrics of all four methods:

Method	RMSE	R-Squared	MAE
KNN	0.4536	0.7828	0.2803
Random Forest	0.3067	0.9045	0.1834
LASSO Regression	0.4309	0.8030	0.3044
Multiple Linear Regression	0.4310	0.8029	0.3046

Based on the results, the Random Forest method performed the best among the four modeling techniques. It has the lowest values of RMSE and MAE and the highest value of R-squared, indicating that it provides the best fit to the data and has the highest predictive accuracy.

Box plots and dot plots of the results:





The Random Forest model is an ensemble learning method that combines multiple decision trees to improve the accuracy and stability of the predictions. It is a non-parametric method that does not assume a linear relationship between the variables and the response. Instead, it can capture complex interactions and non-linear relationships between the variables, making it suitable for modeling complex systems like health insurance costs.

In the context of this dataset, the Random Forest model was able to identify the most important variables that affect health insurance costs, such as age, BMI, and smoking habits, and capture their complex interactions with other variables like gender and number of children. The model was able to explain 90% of the variation in the data, indicating that it can accurately predict the health insurance costs for new individuals based on their demographic and lifestyle characteristics.

Overall, the Random Forest model can provide valuable insights for insurance companies to set premium prices, design insurance plans, and identify high-risk patients for targeted interventions. It can also help individuals to estimate their potential health insurance costs based on their personal characteristics and make informed decisions about their health and financial planning.

Future work directions

Based on the results of the analysis, there are several future work directions that could provide deeper insights for a decision-maker associated with the health insurance context.

Further investigation could be done into the relationship between the demographic and lifestyle factors and the health insurance costs. For example, it could be interesting to explore the impact of other factors, such as education level, occupation, or geographical location, on the health insurance costs. Such insights could help insurance companies to better understand the drivers of health care utilization and develop more targeted and personalized insurance plans.

Also, the analysis could be enriched by incorporating external data sources, such as health care utilization data, hospitalization records, or disease prevalence rates. By integrating these data sources with the health insurance cost data, decision-makers could gain a more comprehensive understanding of the health care landscape and develop more informed strategies for managing health risks and costs.