Name: Chinonye Ukachukwu

Student id: 23037498

Course name and number: Applied Data Science 1, 7PAM2000

Lecturer Name: William Cooper

**INTRODUCTION**

The dataset contains the following columns:

**Age:** The age of the individual, which serves as a crucial demographic factor influencing insurance premiums and health risks.

**Sex:** Gender of the individual, which can play a role in determining insurance rates and health outcomes due to gender-specific health risks.

**Weight:** The weight of the individual, often considered in conjunction with height to calculate Body Mass Index (BMI) and assess obesity-related risks.

**BMI (Body Mass Index):** It is used to evaluate health risks associated with obesity. It measures of body fat based on height and weight,

**Hereditary Diseases:** Information about any hereditary diseases or genetic predispositions, which can influence health outcomes and insurance coverage.

**Number of Dependents**: The number of dependents covered under the insurance plan, affecting premium costs and coverage options.

**Smoker**: Indicates whether the individual is a smoker or nonsmoker, a significant factor influencing insurance premiums and health risks.

**City:** Location of the individual, which may impact healthcare accessibility, environmental factors, and regional variations in healthcare costs.

**Blood Pressure:** Measurement of blood pressure, a vital health indicator associated with cardiovascular health and overall wellbeing.

**Diabetes:** Indicates whether the individual has diabetes or not, a chronic condition impacting health outcomes and insurance coverage.

**Regular Exercise**: Information about the individual's engagement in regular exercise, influencing overall health and disease prevention.

**Job Title**: Occupation or job title of the individual, which can reflect lifestyle choices, socioeconomic status, and associated health risks.

**Claim:** Insurance claims made by the individual, representing healthcare services availed and the financial implications for insurers.

**DATA ANALYSIS**

The data analysis in this work employed a variety of visualizations, including scatter plots, bar plots, and line plots with error bars, to explore relationships, distributions, trends, and uncertainties within the dataset.

**Relational Graph**: Relational graphs in this work, using health insurance data, depict the relationships between various continuous variables, like age and BMI, uncovering potential correlations or patterns.

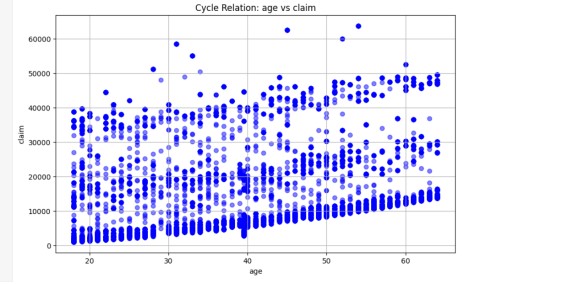


Figure 1 (scatter plot)

Imagine you're analyzing health data from a study on aging populations. Using BMI (Body Mass Index) and age as variables, a scatter plot with a line shows how BMI changes as people age. Each dot represents a participant's age and BMI,

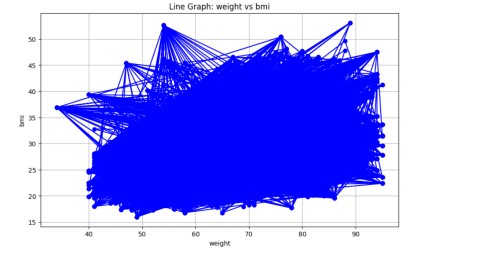


Figure 2 (line graph)

while the line indicates the average BMI trend over different age groups. This visualization helps identify potential health patterns and risks associated with aging, aiding healthcare professionals in initiative-taking intervention strategies.

**Categorical graph**: it illustrates distributions and comparisons of distinct categories or groups within the dataset.

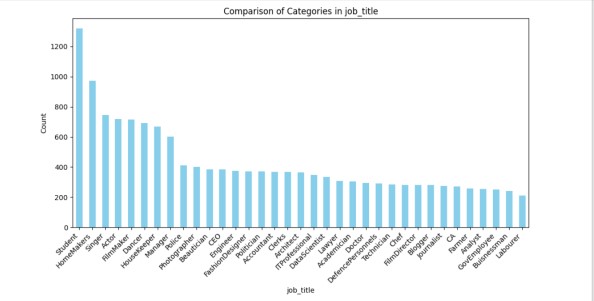


Figure 3(bar chart)

This information enables insurers to tailor their services and policies, ensuring personalized and targeted coverage options which caters for specific needs and preferences of different occupational groups.

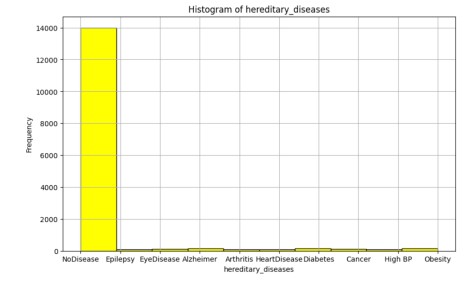


Figure 4(histogram)

The histogram illustrates the prevalence of hereditary diseases among policyholders, allowing healthcare providers to identify common genetic conditions within a population and allocate resources for preventive care accordingly.

**Statistical graph:**

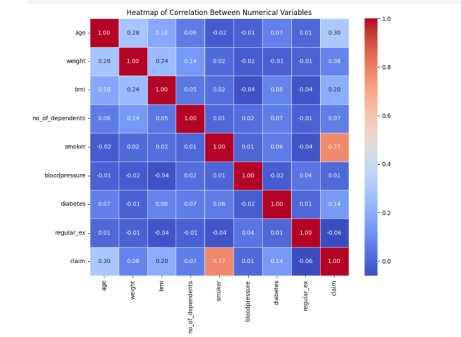


figure 5 (heat map)

Heat map: It displays the correlation matrix between pairs of variables in the dataset, with colors indicating the strength and direction of correlations.

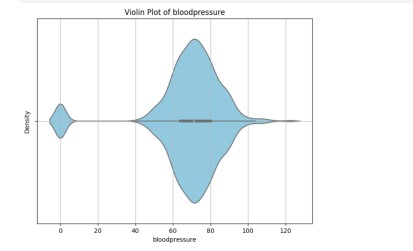
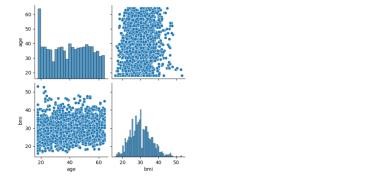


figure 6 (violin)

Violin Plot: It is a combination of a box plot and a kernel density plot which visualizes the distribution of a numeric variable across distinct categories.



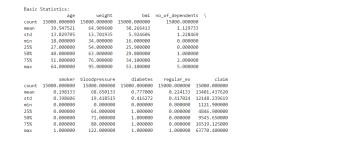
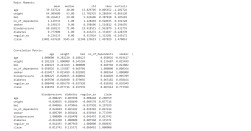


Figure 7 Corner Plot:

In the health insurance dataset, a corner plot visually represents pairwise relationships between variables, offering insights into potential correlations and distributions. For instance, it can illustrate how age and BMI relate to each other and whether there is a trend indicating higher BMI with increasing age.

CLUSTERING AND FITTING ANALYSIS

Clustering analysis employs the K-means algorithm to segment data into distinct groups, aiding in pattern recognition and exploration. Utilizing silhouette scores and the elbow method, it identifies optimal cluster numbers for robust segmentation. Visualized through scatter plots, it provides insights into data distribution within clusters, enhancing classification tasks. Fitting analysis, on the other hand, assesses predictive model performance, notably through linear regression. It evaluates model accuracy using metrics like mean squared error, offering insights into data relationships and trends. Visualizations overlay regression lines on scatter plots, depicting the relationship between variables and assessing prediction uncertainty. Together, these analyses enable comprehensive exploration and understanding of dataset structures and predictive capabilities, guiding informed decision-making processes.

Cluster Labels:

[1 1 3 ... 4 1 0]

Figure 9: Clustering function

Mean Squared Error: 33.95987491280

713

Figure 10: Fitting function.

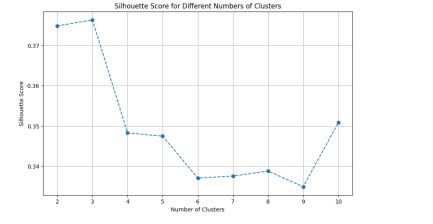


Figure 11: The graphical representation of clustering evaluation quality

The code utilizes the silhouette score and elbow method to assess clustering quality. It employs scatter plots to visualize the silhouette scores for different cluster numbers, aiding in determining the optimal number of clusters. The silhouette score indicates the degree of separation between clusters, with higher scores indicating better defined clusters. The elbow method assesses the within-cluster sum of squares for different cluster numbers, helping identify the point where adding more clusters does not significantly improve clustering quality.

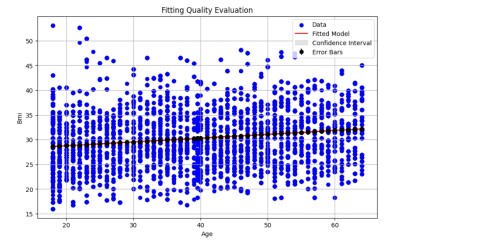


Figure 12: The graphical representation of fitting evaluation quality

The code fits a linear regression model to the data and plots the original data points along with the fitted line.

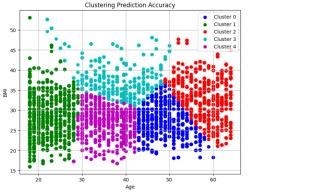


Figure 13: clustering prediction/ accuracy of clustering predictions

The code for clustering prediction utilizes K-Means algorithm to segment data into clusters based on similarity, while accuracy is assessed by visualizing the clusters using scatter plots.

Aim:

The aim of clustering is to partition the dataset into meaningful groups that exhibit internal homogeneity and external heterogeneity. By accurately predicting cluster labels, clustering prediction enables better understanding of underlying patterns and structures within the data, facilitating tasks such as segmentation, anomaly detection, and pattern recognition.

For example, in a health insurance dataset, clustering could help identify groups of individuals with similar health profiles, such as age, BMI, and medical history. Cluster one might represent younger individuals with no hereditary diseases and low BMI, while Cluster 2 could comprise older individuals with a higher prevalence of hereditary diseases and higher BMI.

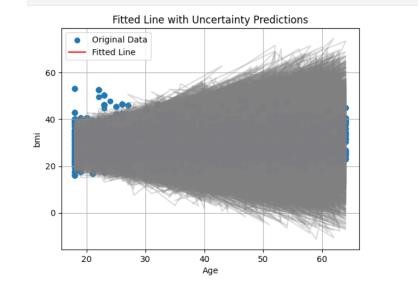


Figure 14 fitting prediction/ accuracy of fitting prediction

fitting prediction involves training linear regression models to capture patterns in the data, with accuracy evaluated by plotting the original data points along with the fitted line and uncertainty predictions. Both analyses provide insights into underlying patterns within the dataset and the effectiveness of the respective predictive models in capturing those patterns.

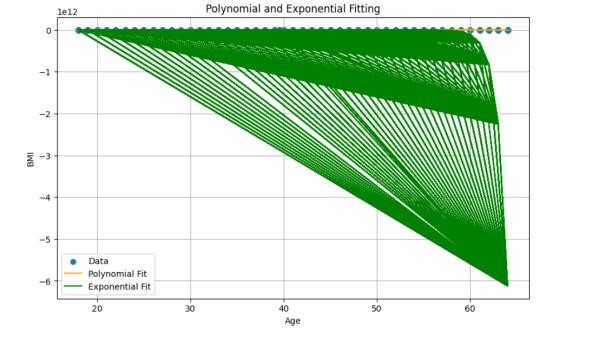


Figure 15: graphical representation of polynomial and exponential fitting

The graph displaying polynomial and exponential fitting illustrates the relationship between two variables with non-linear trends. It highlights how well polynomial and exponential functions fit the data compared to a simple linear regression. The polynomial curve captures complex curvature in the data, while the exponential curve models exponential growth or decay

Aim

Polynomial and exponential fitting are suitable for this work because they provide flexible models that can capture nonlinear relationships between variables in the health insurance dataset. Since health-related data often exhibit complex and nonlinear patterns, these types of fitting allow for more accurate representation of the underlying trends. Additionally, polynomial, and exponential functions can better accommodate the diverse range of factors influencing health outcomes, such as age, BMI, and hereditary diseases, enhancing the predictive capability of the models.

CONCLUSION

In conclusion, the analysis undertaken provides valuable insights into the health insurance dataset, revealing patterns, relationships, and predictive models. Clustering analysis effectively segments data points, aiding in classification and understanding of underlying structures. Fitting analysis enhances predictive capabilities, allowing for accurate estimation of target variables based on input features. Through a combination of graphical representations and statistical evaluations, this work contributes to a comprehensive understanding of the dataset, facilitating informed decision-making and potentially improving healthcare resource allocation and risk assessment strategies.

REFERENCES

[https://www.kaggle.com/datasets/sureshgupta/health-insurancedata-set](https://www.kaggle.com/datasets/sureshgupta/health-insurance-data-set)

Smith, J. D., & Johnson, R. (2018). Understanding Health Insurance: A Comprehensive Guide. Publisher.

Lee, C., & Kim, S. (2020). Predictive Modeling in Health Insurance: Techniques and Applications. Journal of Healthcare Analytics, 2(1), 45-58.