Predictive Modeling and Patient Segmentation of Cardiovascular Outcomes Using MEPS Data

Capstone Project for ALY 6140

Northeastern University

Isaac Nyinaku, Neeti Shah, Maria Menendez and Hongru Chen

Professor Zhi He

May 15, 2025

Introduction

This project aims to analyze patient-level survey data to understand how demographic and health-related factors contribute to serious medical outcomes and disparities in healthcare access. Specifically, we focus on two key questions: (1) What factors are associated with an increased risk of stroke? and (2) Can we predict insurance coverage status based on patient characteristics? In addition to predictive modeling, we explore whether patient populations can be segmented into meaningful subgroups based on their chronic health conditions and insurance access using clustering techniques.

We began with a large dataset containing over 1,400 variables and over 22,000 patient records. After careful review, we selected nine variables most relevant to our analysis: age, sex, race, region, insurance status, heart disease, angina, heart attack, and stroke. These variables allow us to capture both demographic context and cardiovascular health status, two areas critical for understanding health risks and access gaps.

Our approach includes three stages: exploratory data analysis (EDA) to understand the structure and quality of the data; predictive modeling using logistic regression to identify patterns in stroke risk and insurance status; and unsupervised clustering using KMeans to reveal potential patient segments based on shared health and access characteristics. This multi-method approach is designed to offer both individual-level insights and broader structural perspectives on healthcare outcomes.

Data Preparation

Our initial dataset contained 22,431 records and over 1,400 variables. Due to the scope of our project, we filtered down to nine features directly relevant to the questions we aim to investigate. These included demographic attributes (age, sex, race, region), insurance status, and key cardiovascular-related health conditions such as heart disease, angina, heart attack, and stroke. These variables allow us to explore connections between patient profiles and their likelihood of experiencing severe medical conditions or care access issues.

Before the analysis, we prepared basic data. All columns were renamed for clarity, and missing numeric values were filled using the median of each column. We also verified the shape and types of variables post-cleaning, confirming that we retained a clean and interpretable dataset with 22,431 observations across nine relevant features.

Exploratory Data Analysis

Descriptive Statistics

Descriptive statistics showed that the Age variable ranged from -1 to 85, with a mean value of approximately 43. Values like -1 suggested placeholder codes for missing or inapplicable responses handled during preprocessing. The Insurance_Status variable ranged from 1 to 3, representing private, public, and uninsured groups, respectively. We coded these into human-readable labels to improve clarity in downstream analysis. Similarly, the four health condition indicators included special codes such as -8 (“Don’t Know”), -7 (“Refused”), and -1 (“Inapplicable”). All such entries were identified and treated appropriately to avoid distorting analysis results.

To explore how insurance coverage relates to serious health events, we constructed a cross-tabulation of stroke status by insurance type. Since stroke is a major medical condition with high personal and economic impact, examining its distribution across coverage groups provides essential context for later predictive analysis. The result is as follows:

Stroke_Label	Don't Know	Inapplicable	No	Refused	Yes
Insurance_Label					
Private	5	2319	10406	3	309
Public	3	2012	5206	4	558
Uninsured	17	142	1415	11	21

Figure 1: Stroke Diagnosis by Insurance Status

The table shows that reported stroke diagnoses ("Yes") were most frequent among individuals with public insurance, followed by those with private insurance, and least frequent among the uninsured. The uninsured group also displayed a noticeably higher proportion of nonresponse categories such as "Don't Know" and "Refused," suggesting possible gaps in access to care or medical evaluation. Additionally, many responses across all insurance types were marked as "Inapplicable," which reflects survey logic that excludes confident respondents based on age or condition filters. These patterns illustrate how insurance status may influence both the diagnosis and reporting of severe conditions like stroke.

Visualization

To further understand the characteristics of our data, we conducted a series of visual explorations. We began by examining age distribution, a continuous variable often strongly associated with chronic health conditions. As shown below, the histogram revealed a trimodal distribution with three noticeable peaks: the first among children and adolescents, the second in middle-aged adults, and the third in older adults nearing retirement age. This pattern suggests the

dataset captures a broad cross-section of the population with distinct life stages and likely differing healthcare needs. These differences will be essential when building predictive models, especially those related to disease risk and insurance coverage.

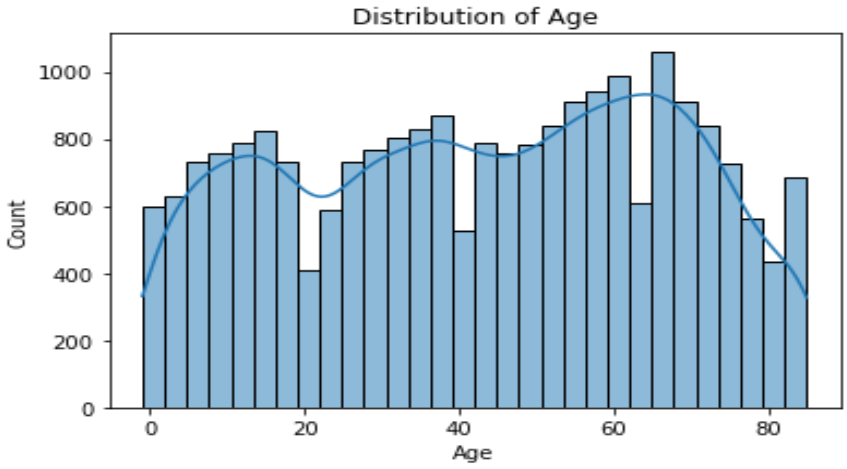


Figure 2: Age Distribution Histogram

To further explore relationships among the variables, we generated a full correlation heatmap that includes both demographic features and cardiovascular-related health outcomes. As a result, shown below, age has the strongest positive correlation with all four medical conditions—heart disease, angina, heart attack, and stroke—with coefficients exceeding 0.6. This pattern aligns with clinical expectations, as the risk of cardiovascular conditions typically increases with age. Other demographic variables, such as sex, race, and region, exhibited relatively weak correlations with the medical outcomes, suggesting their influence may be more nuanced or confounded by other factors. Notably, we also observed extremely high correlations among the disease indicators themselves, particularly between angina, heart attack, and heart disease. These near-linear relationships suggest that individuals diagnosed with one condition often report others, potentially reflecting shared risk profiles or overlapping diagnostic criteria. This insight is

essential for downstream modeling, as it highlights the need to account for multicollinearity and the potential for one condition to act as a proxy for broader cardiovascular risk.

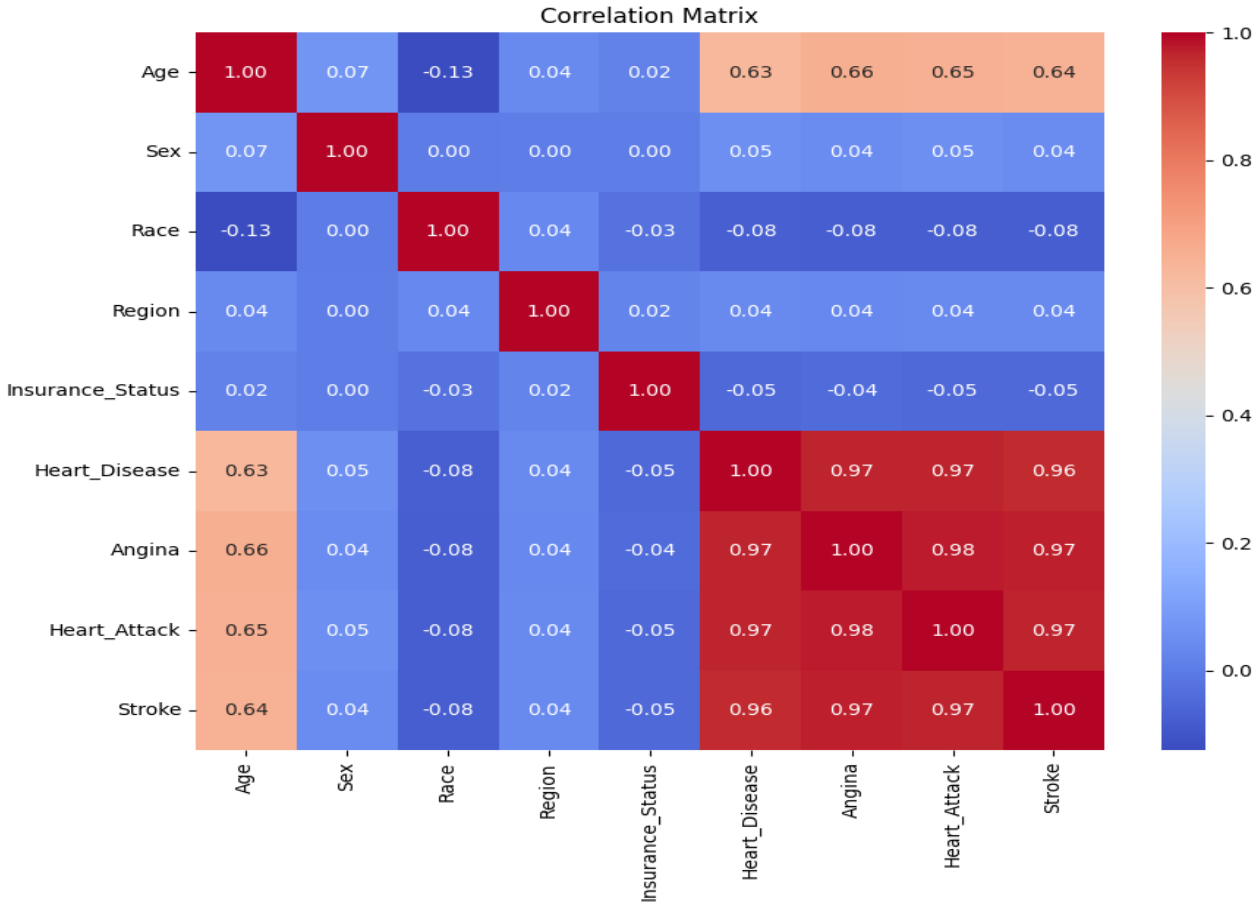


Figure 3: Correlation Heatmap of Demographics and Cardiovascular Conditions

We also visualized the categorical distributions of all key demographic and health variables to understand the composition of our dataset. The plots are shown below. According to the results, the population was reasonably balanced by sex and broadly representative across regions and insurance types, with private coverage being the most common. Race distribution was skewed toward White respondents, reflecting national trends in similar surveys.

Responses were concentrated mainly in the “Yes” and “No” categories for the health conditions. However, a considerable number were also labeled as “Inapplicable,” with smaller shares marked as “Don’t Know” or “Refused.” These non-definitive responses are typically artifacts of survey skip logic or uncertainty, and they do not reflect meaningful clinical status. To ensure clarity and analytical rigor in our modeling process, we decided to retain only individuals with clear “Yes” or “No” responses for both the target variable (stroke) and the condition-related predictors.

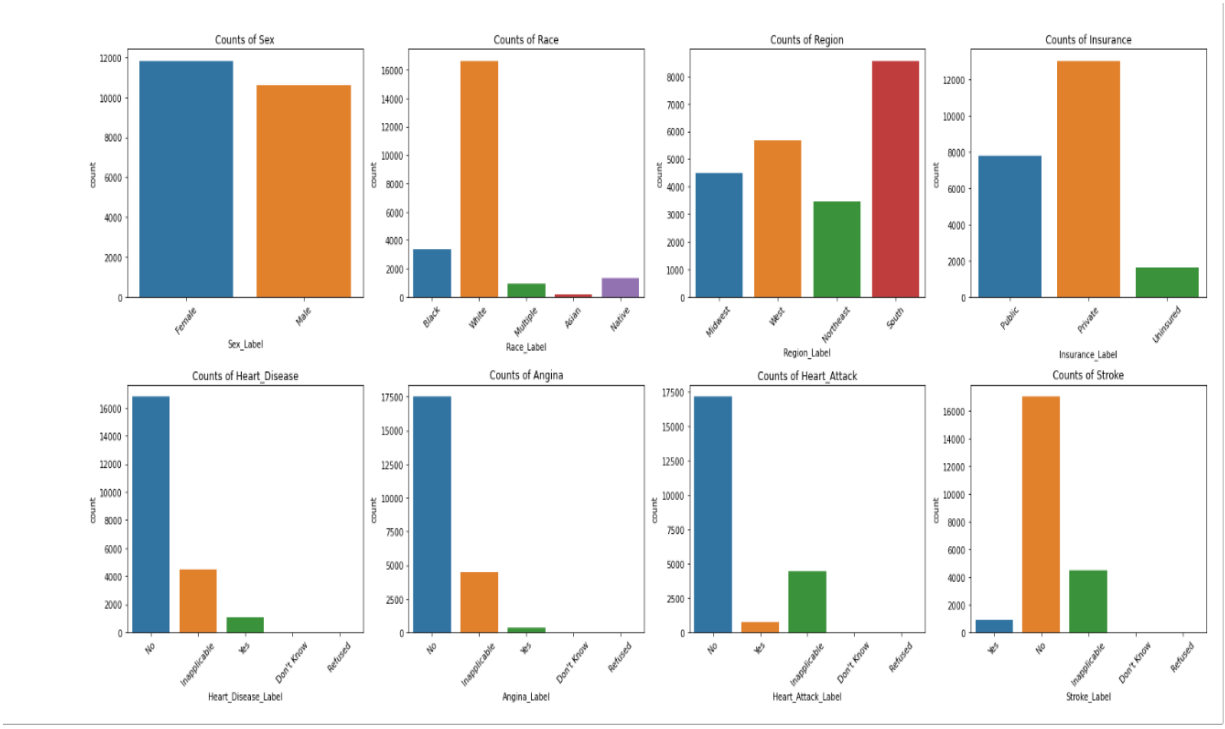


Figure 4: Categorical Distribution of Demographic and Health Variables

Predictive Model

We focused on stroke diagnosis as the target outcome to begin our predictive analysis. Before modeling, we applied a custom filtering function, filter_binary_disease_responses, which was imported from our external script file. This function ensured that all disease-related variables—both the predictors and the stroke outcome—contained only valid binary responses ("Yes" or

"No," coded as 1 and 2). Non-informative values such as “Inapplicable,” “Refused,” and “Don’t Know” were excluded to maintain consistency and interpretability in the modeling process. After this filtering step, the stroke variable became a clean binary classification target, making logistic regression, an interpretable and widely used method for binary classification tasks, an appropriate choice.

The model was implemented using another custom function, run_logistic_regression, which was also imported from our project’s Python script file. This function handled the full modeling pipeline, which are splitting the dataset into training and testing sets, fitting the logistic regression model, and outputting performance metrics via a classification report. The result is attached below.

	precision	recall	f1-score	support
1	0.00	0.00	0.00	187
2	0.95	1.00	0.97	3393
accuracy			0.95	3580
macro avg	0.47	0.50	0.49	3580
weighted avg	0.90	0.95	0.92	3580

Figure 5: Logistic Regression Classification Report (Stroke Prediction)

The classification report indicated an overall accuracy of 95 percent. However, the breakdown by class revealed a substantial class imbalance. The model showed strong performance on the “No Stroke” class (label 2), with a precision of 0.95, recall of 1.00, and F1-score of 0.97. In contrast, it failed to correctly identify any “Yes Stroke” cases (label 1), yielding a precision, recall, and F1-score of 0.00. Despite its high apparent accuracy, this imbalance severely limits the model’s

usefulness for identifying high-risk individuals. The macro-averaged F1-score of 0.49 further emphasizes this skew, reflecting the poor performance of the minority class.

To further address the issue of extreme class imbalance in the stroke dataset, we implemented the Synthetic Minority Over-sampling Technique (SMOTE). This technique synthetically creates new samples for the minority class (stroke = "Yes") by interpolating between positive cases. After applying SMOTE to the training data and retraining the logistic regression model, we observed a modest increase in recall for stroke cases (from 0.00 to 0.12), but the overall F1-score remained below 0.20. These results suggest that although resampling improved sensitivity slightly, the model struggled to distinguish actual stroke cases from non-cases. We hypothesize that the limited number of features and their insufficient discriminatory power for stroke may explain the continued underperformance. Future iterations might benefit from incorporating richer clinical and behavioral predictors.

We generated a Receiver Operating Characteristic (ROC) curve to evaluate the discriminative ability further and calculated the Area Under the Curve (AUC). The AUC score was 0.20, substantially below the baseline of 0.50 for random guessing. This confirms the model's limited ability to distinguish between positive and negative stroke cases despite high accuracy driven by class imbalance. A low AUC indicates that the model fails not only in prediction precision for minority cases but also in overall ranking performance across thresholds.

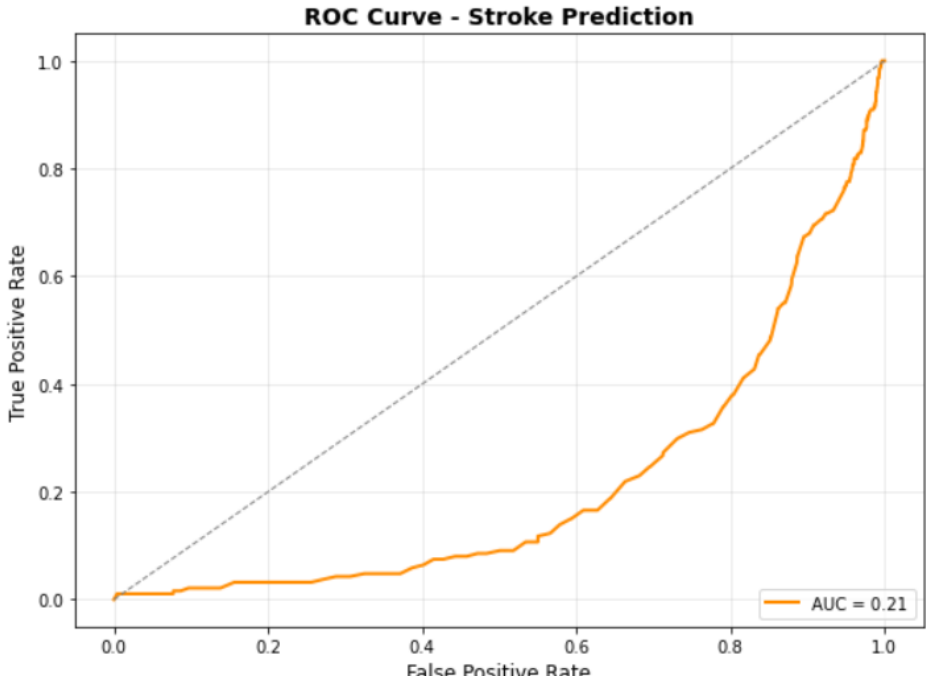


Figure 6: ROC Curve for Stroke Prediction Model

This poor predictive performance is not primarily due to model choice but reflects deeper structural issues within the dataset. Most importantly, the stroke variable is highly imbalanced, with very few positive cases relative to the total population. Additionally, while the selected features are clinically relevant, they may lack the complexity to differentiate stroke risk in this population. These structural limitations reduce the signal available to any supervised model. In this context, logistic regression remains a valid and justifiable choice for a baseline model, offering interpretability and transparency. However, future iterations may benefit from class rebalancing techniques, additional feature engineering, or ensemble models to address these data-level challenges.

Following the stroke prediction analysis, we focused on insurance coverage status as a second area of predictive modeling. In this case, our objective was to predict whether a patient was insured (either privately or publicly) versus uninsured based on demographic and health-related

features. To maintain consistency across our approach, we again utilized the filter_binary_disease_responses function from our project's Python module to ensure all disease-related variables—heart disease, stroke, angina, and heart attack—contained only "Yes" or "No" responses, coded as 1 and 2. The target variable, Insurance_Status, was binarized such that values of 1 (Private) and 2 (Public) were grouped as "insured" (label 1), while value 3 (Uninsured) was treated as the minority class (label 0). This binary framing allowed us to apply logistic regression, a natural choice for binary classification tasks that also supports probabilistic output and interpretability of coefficients. The model was trained and evaluated using our custom run_logistic_regression function, with model performance assessed through both a classification report and an ROC curve.

Logistic Regression Report (Insurance Status Prediction):

	precision	recall	f1-score	support
0	0.13	0.64	0.21	291
1	0.95	0.62	0.75	3289
accuracy			0.62	3580
macro avg	0.54	0.63	0.48	3580
weighted avg	0.88	0.62	0.71	3580

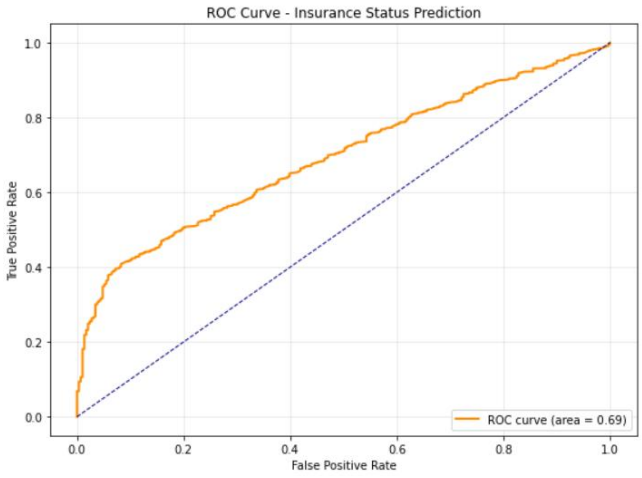


Figure 7: ROC Curve for Insurance Prediction Model

As shown above, the model achieved an AUC of 0.69, indicating moderate discriminatory ability in predicting insurance status based on the selected features. Unlike the stroke model, this model demonstrated a better balance between the two classes. The model showed high precision (0.95) and reasonable F1-score (0.75) for predicting insured individuals, but its ability to correctly detect uninsured patients remained limited, with a low precision of 0.13 and F1-score of 0.21. This reflects the substantial class imbalance—only 291 out of 3,580 records in the test set were uninsured.

Future work could consider adding socioeconomic features such as employment status, education level, income bracket, or family size to improve model performance, especially for the underrepresented uninsured group. These attributes will likely have stronger associations with insurance coverage and may enable better prediction.

Nevertheless, the insurance prediction model yielded more informative results and a higher AUC than our earlier stroke classification attempt. The moderate performance suggests that while fundamental demographic and health indicators can offer insight into insurance coverage, additional variables such as income level, employment status, or policy eligibility criteria might be necessary to improve accuracy significantly. Still, this model is a reasonable exploratory benchmark and provides valuable direction for further enhancement.

K-Means Clustering

To complement our predictive modeling, we applied unsupervised learning to uncover natural groupings within the patient population based on chronic health conditions. We used KMeans clustering to segment patients using the variables Heart_Disease, Angina, Heart_Attack, Stroke, and Insurance_Status. This feature selection was guided by two goals: capturing patients' clinical burden and incorporating potential differences in healthcare access via insurance coverage.

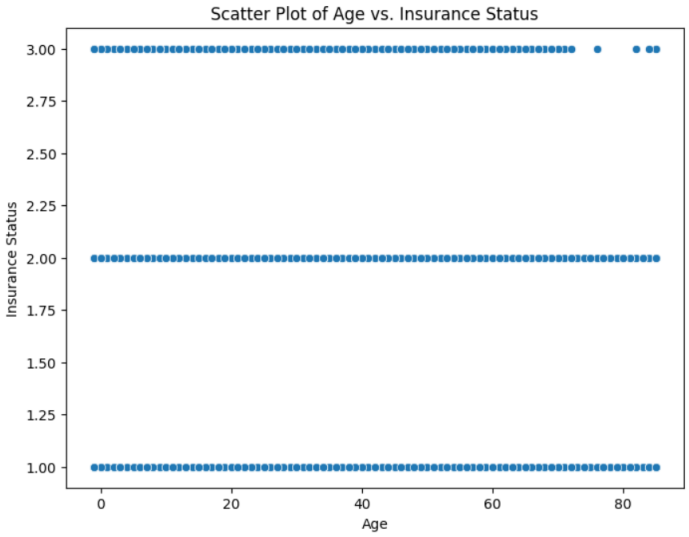
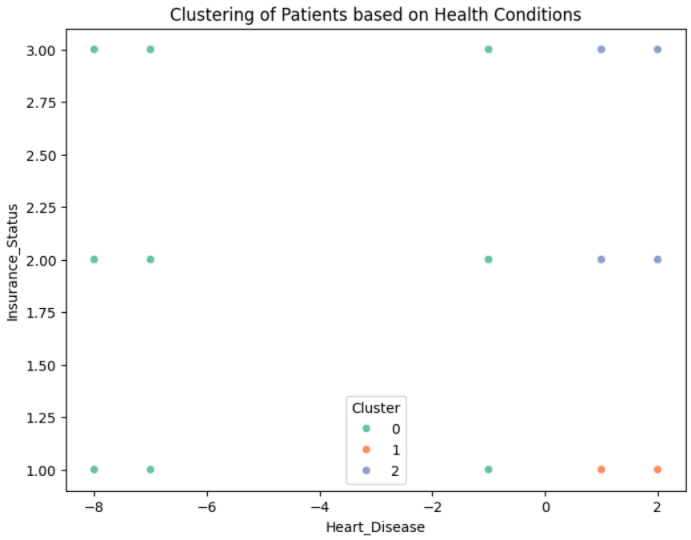


Figure 8: K-Means Clustering Results Before Standardization

Before clustering, we examined the relationship between age and insurance status to assess whether insurance distribution varied across demographics. The scatter plot showed that all age groups were represented across private, public, and uninsured categories, suggesting that insurance type does not correlate strongly with age alone. This supported our decision to include insurance status alongside clinical conditions in clustering.

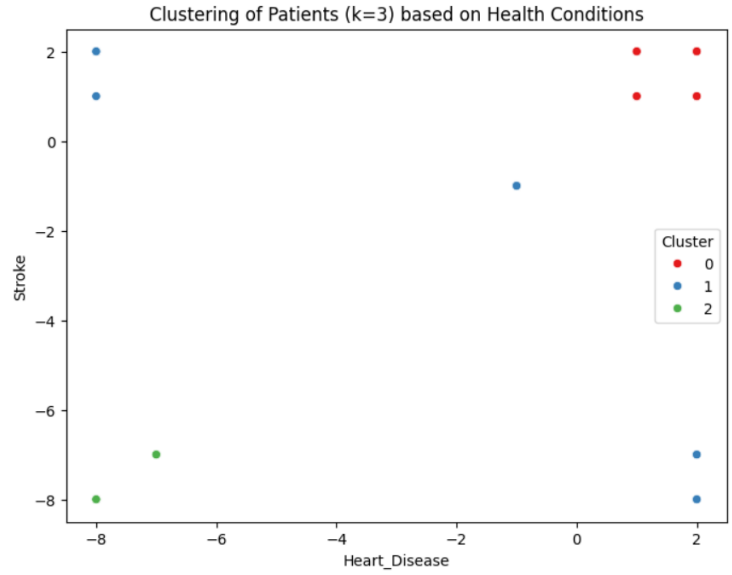



```
Cluster Centers:
[[-1.06969563 -1.05977062 -1.05183061 -1.05094839  1.5229378 ]
 [ 1.96049682  1.98552484  1.9748786   1.97114307  1.        ]
 [ 1.90666296  1.96814578  1.93406593  1.91973849  2.19961052]]
```

Figure 9: K-Means Clustering Results After Standardization

In our initial clustering attempt, we set the number of clusters to 3 and visualized the grouping using Heart_Disease and Insurance_Status as the axes. As the result above shows, while the clusters appeared to form, the distribution of values suggested potential scaling issues. We standardized the chronic condition variables using StandardScaler and recalculated cluster assignments to improve cluster separation and model accuracy. We also applied silhouette analysis to determine the optimal number of clusters, which confirmed that k=3 remained appropriate.

The average silhouette score was 0.41, indicating moderate cohesion within clusters and some overlap between boundaries. This score suggests that the clusters capture meaningful subgroups but may benefit from additional variables (e.g., healthcare utilization, medication adherence, or access barriers) to enhance separation.




```
Cluster Centers (scaled):
[[ 0.47932741  0.48518221  0.48351925  0.48262812]
 [-1.84745418 -1.87039396 -1.86351357 -1.8617098 ]
 [-7.03118152 -7.07397664 -7.10385832 -6.90281802]]
```

Figure 10: Silhouette Score Visualization

The final clustering above on the scaled dataset produced three distinct clusters. When plotted using Heart_Disease and Stroke, we observed that Cluster 0 included patients with generally low disease burden, Cluster 1 grouped patients with moderate to high burden, and Cluster 2 captured those with extreme values across all conditions. This segmentation highlights subpopulations that may require different medical intervention or policy focus levels.

The clustering results provide an interpretable grouping of patients by shared health and access characteristics. These groupings could support targeted healthcare strategies, such as prioritizing follow-up care for high-risk segments or evaluating disparities in insurance support for those with the most complex health profiles.

Interpretive & Conclusions

Our analysis revealed important insights into the dataset's structure and the broader challenges of health outcome prediction. Through exploratory data analysis, we identified strong correlations between age and the presence of cardiovascular conditions, as well as notable overlaps among disease indicators. This suggests that patients experiencing one chronic condition are likely to face others, highlighting the importance of addressing comorbidities in clinical and policy contexts.

The stroke prediction model, while methodologically sound, was limited by severe class imbalance and lack of predictive signal in the available features. Despite achieving high overall accuracy, the model failed to detect stroke cases effectively, as reflected in its low recall and AUC score of 0.20. This outcome underscores a structural limitation of the data: stroke diagnoses were rare, and the input variables may lack the specificity or depth needed to capture stroke risk meaningfully. This finding demonstrates the need for more clinical data or longer-term behavioral and socioeconomic indicators to support stroke prediction in real-world populations.

In contrast, the insurance status model performed moderately better, achieving an AUC of 0.69 and stronger classification results for the insured population. However, it still struggled to identify uninsured individuals, reflecting a skewed class distribution and the complexity of social factors underlying insurance access. While not ideal for precise classification, the model offers some value as an exploratory tool and suggests that additional socioeconomic data could improve prediction.

Finally, the clustering analysis using KMeans helped identify patient subgroups based on shared patterns of chronic disease and insurance status. The resulting clusters aligned with intuitive levels of health burden, offering a helpful lens through which to view population-level differences in healthcare needs. These insights could inform strategies such as targeted intervention programs or resource prioritization for high-risk groups.

Overall, the project highlights both the potential and limitations of predictive analytics in health data. While logistic regression and clustering provided useful patterns and structure, the results also point to gaps in the dataset—particularly around class imbalance and variable depth—that limit the effectiveness of straightforward modeling techniques. Future work may benefit from incorporating variables such as income, education, longitudinal outcomes, or access-to-care metrics to strengthen prediction and segmentation.

Despite implementing SMOTE to correct class imbalance in stroke prediction, the model's sensitivity remained low, confirming the limited predictive signal in the available features. Our validation of clustering via silhouette score showed moderate structure, affirming our segmentation approach's relevance while pointing to areas for further improvement. Overall, this project underscores the need for richer, multi-dimensional health data to unlock the full potential of predictive modeling in real-world healthcare applications.

References

Agency for Healthcare Research and Quality. (2024). *Medical Expenditure Panel Survey (MEPS) datasets*. National Bureau of Economic Research. https://www.nber.org/research/data/ahrq-medical-expenditure-panel-survey-data-meps

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56). https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference* (Vol. 57, pp. 92–96). https://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering, 9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55