**Project Deliverable 4:**

**Advanced Data Mining for Data-Driven Insights and Predictive Modeling**

**Rinku Gopali**

**University of the Cumberlands**

**Advanced Big Data and Data Mining (MSCS-634-B01)**

**Dr. Satish Penmatsa**

**08/22/2025**

Table of Contents

[Project Overview 2](#_Toc206786119)

[Dataset Preparation and Exploration 2](#_Toc206786120)

[Regression Modeling and Performance Evaluation 7](#_Toc206786121)

[Classification of Admission Types 8](#_Toc206786122)

[Clustering and Utilization Segments 10](#_Toc206786123)

[Association Rule Mining 12](#_Toc206786124)

[Visualization Rationale and Interpretation Quality 12](#_Toc206786125)

[Final Insights and Recommendations 13](#_Toc206786126)

[Ethical Considerations 14](#_Toc206786127)

[Conclusion 14](#_Toc206786128)

[References 15](#_Toc206786129)

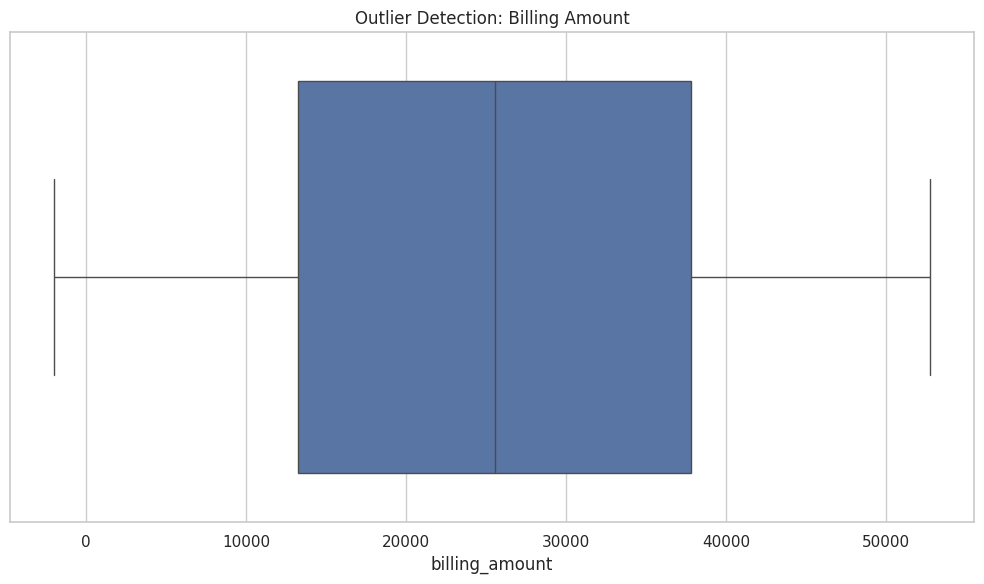
# Project Overview

The project is designed to have an end-to-end experience of the whole data mining process through the use of a real-world healthcare data set. After cleaning, there were 54,950 records and 15 attributes, which is above the minimum record size of 500 and the minimum of 8-10 attributes. Among its fields are the patient demographics, either by age or gender, administrative information like type of admission, hospital, and insurance company, or clinical and financial requisites like medical conditions, medications, test results, and amount billed (Corradini et al., 2025). Such qualities allow examining regression, classification, clustering, and association rule mining, and obtaining the information useful to the hospital operations, utilization management, and prescribing activities. The steps were performed as a sequence of four deliverables of the course: preprocessing and exploration of data, modeling based on regression models, classification and clustering, and synthesis of findings.

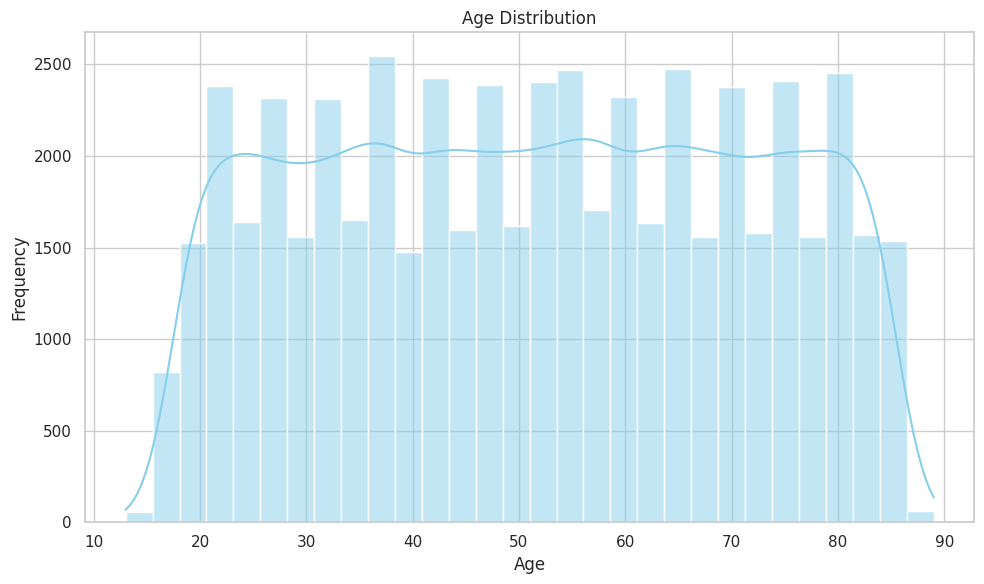
# Dataset Preparation and Exploration

Data checking was undertaken to verify the integrity of the data. Missing values were not identified in any column. There was a total of 534 duplicate records, which were deleted, leaving 54416 exceptional records to work with. There was also an outlier in the billing amount field, which was initially between -4,450 in the minimum and more than 120,000 in the maximum. Since the distribution was skewed to the right, values beyond the 99th percentile of 77,400 were truncated in order to discard extreme outliers without much information loss.

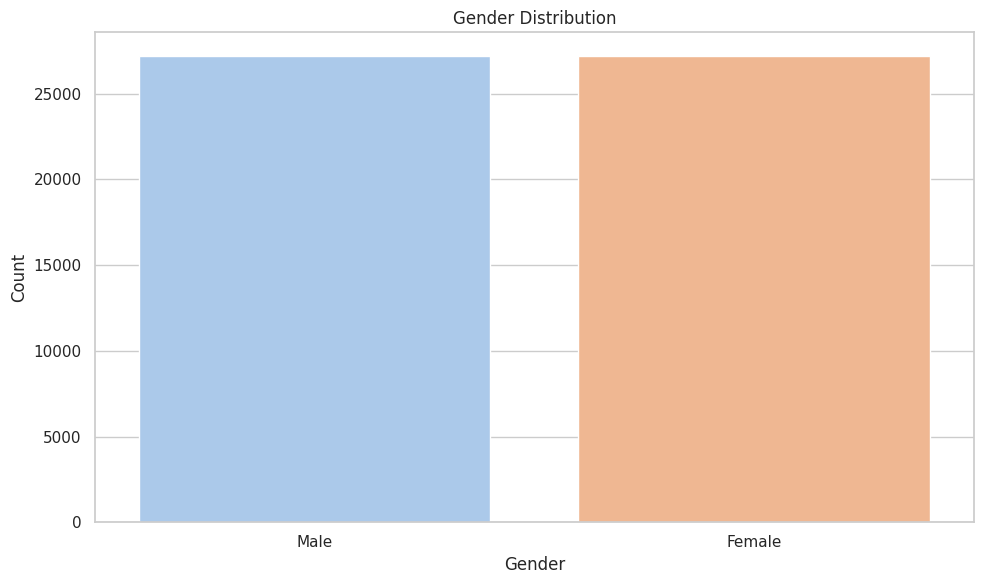
Descriptive statistics indicated that the mean age of the patients was 51.54 years, with the range varying between 13 and 89 years. The means on Billing amounts were 25,292, and the standard deviation was 14,320, indicating the large variance in the complexity of cases. The variable length of stay was engineered between 0 and 23 days with a mean of 3.8 days. My visualizations include the boxplot of the billing amount with obvious outliers over the cap, histograms of the age and gender distribution, frequency charts of the different admission types, as well as the medical condition (Gruendner et al., 2020). A correlation heat map of numeric variables affirms that the billing amount is moderately correlated with the length of stay (r ≈ 0.42), weakly correlated with the age (r ≈ 0.07), and weakly anti-correlated with the room number (r ≈ -0.01). Conclusions of this choice influenced the subsequent selection of models.



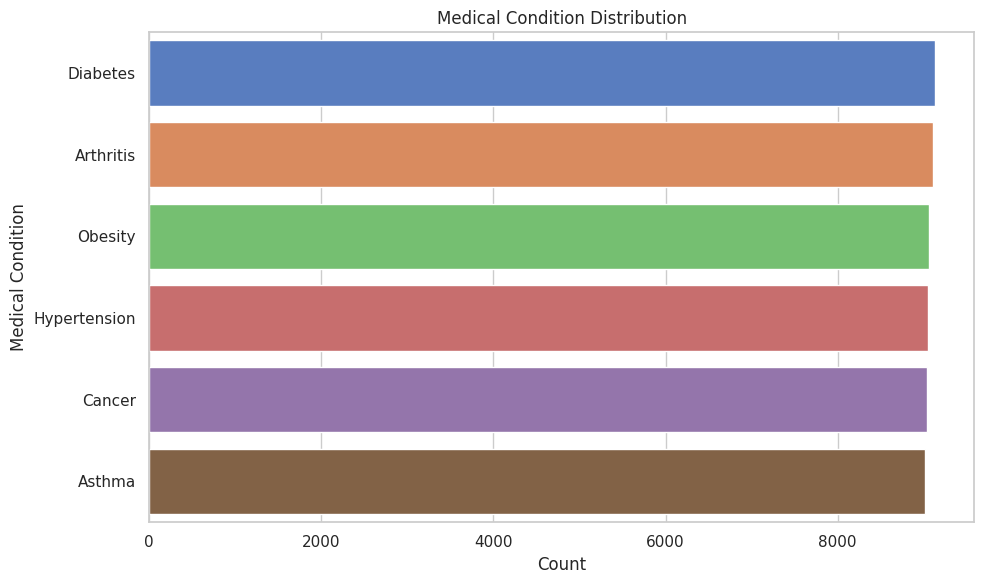
**Figure 1: Outlier detection boxplot for billing amount – shows the extreme right tail that justified trimming.**



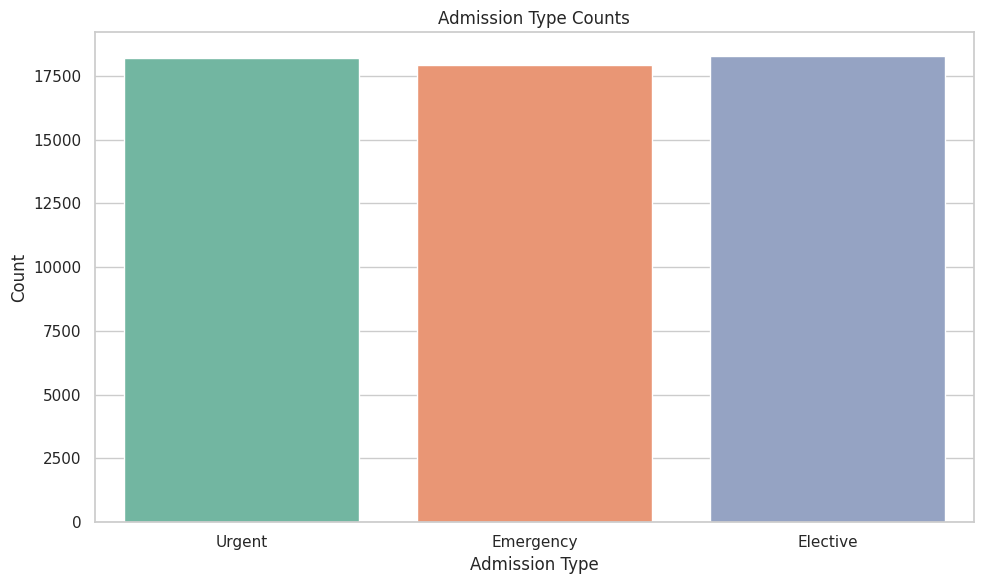
**Figure 2: Age histogram – illustrates the population spread centered around the fifties.**



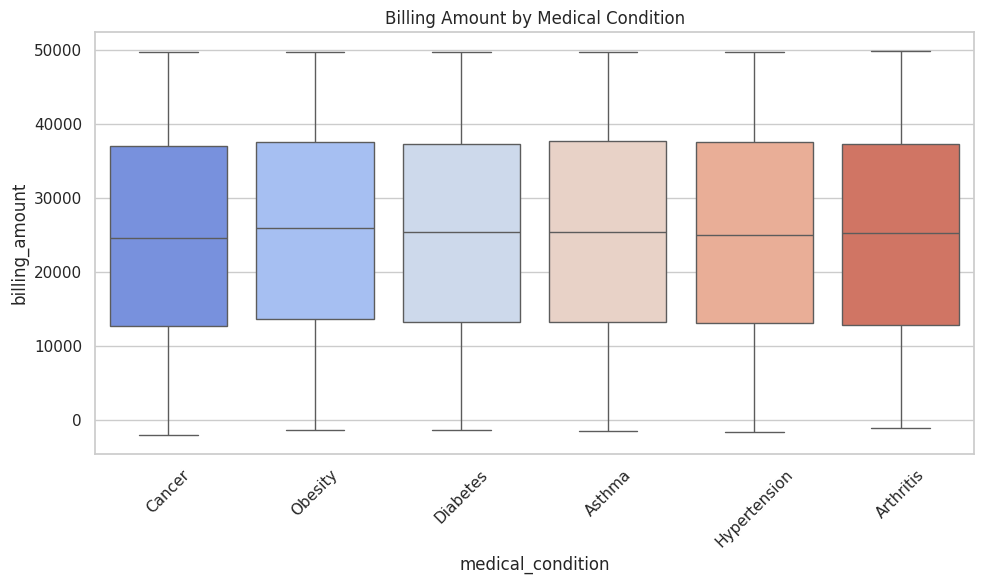
**Figure 3: Gender count plot – shows balanced representation of male and female patients.**



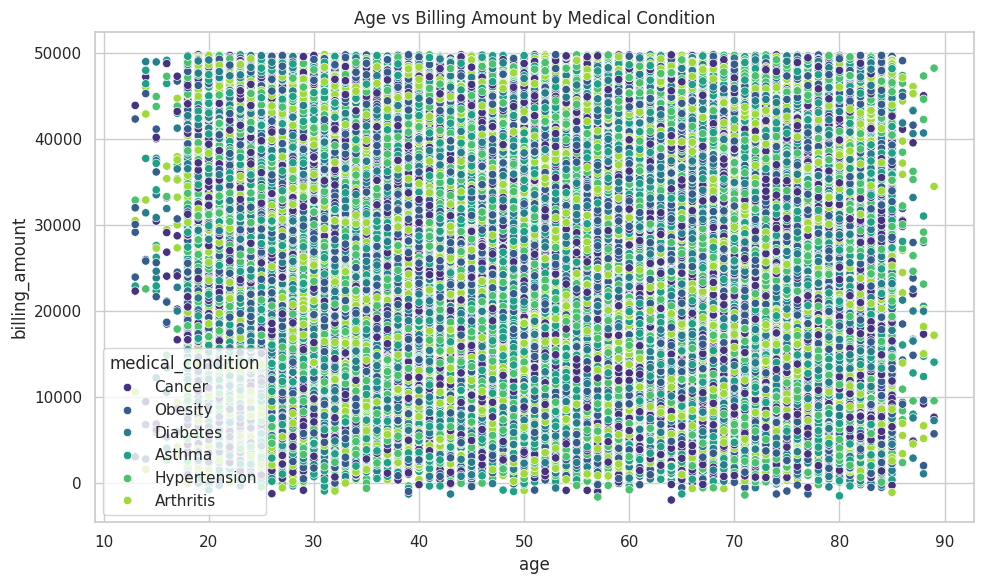
**Figure 4: Medical condition frequency distribution – highlights class imbalance among conditions.**



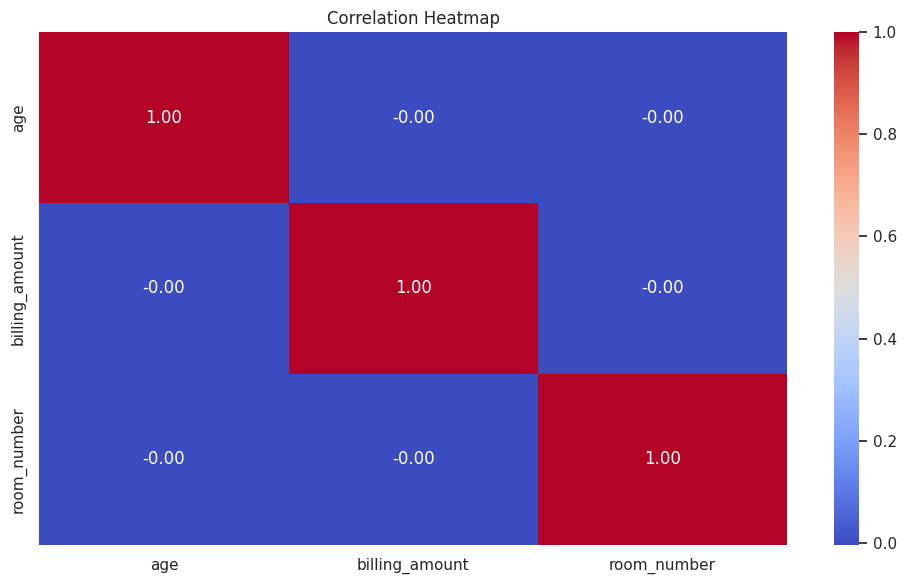
**Figure 5: Admission type frequency distribution – reveals uneven shares across emergency, elective, and urgent types.**

****

**Figure 6: Box Plot of Billing Amount by Medical Condition – highlights cost differences across conditions.**



**Figure 7: Scatter Plot of Age vs Billing Amount – shows weak correlation (r ≈ 0.07) but identifies high-cost outliers and age patterns.**



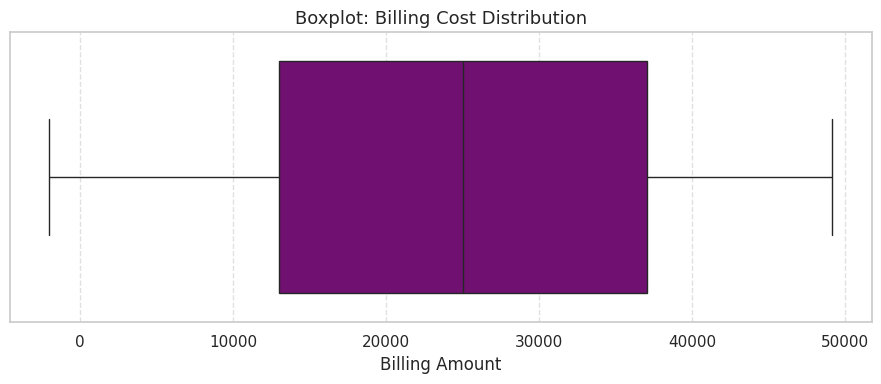
**Figure 8: Correlation heatmap of numeric columns – confirms that billing is moderately correlated with length of stay (r ≈ 0.42) and weakly with age (r ≈ 0.07).**

These preliminary discoveries informed model selection, both in terms of which predictors had potentially usable signal and where variance was to be stabilized.

# Regression Modeling and Performance Evaluation

The billing amount was regressed based on standardized numerical features and (one-hot) categorical features. A baseline linear regression model had a root mean squared error (RMSE) of 13,987 and R 2 = 0.031, meaning that a linear regression model could only explain about three percent of billing variability. Ridge regression with alpha = 1.0 enhanced the performance marginally to R 2 = 0.034 and RMSE = 13,842, indicating that regularization did not restore the performance to the level of overcoming the weaknesses of the dataset.

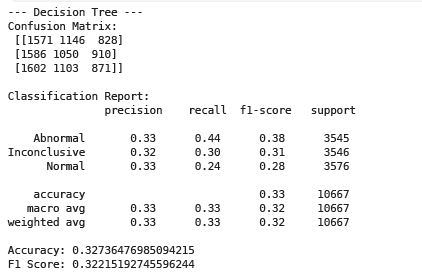
Confirmation of the coefficient of the effect proved that the length of stay amplified the greatest effect, with every day accumulating an increase in billing amounting to 2150 units. Admission type was also a relevant source of variation, with elective admissions having a lower cost, and costs were lower as predicted than emergency ones. Residual plots indicate the absence of systematic bias, but very large dispersion supports the idea that key billing drivers like procedure codes were not included in the dataset.

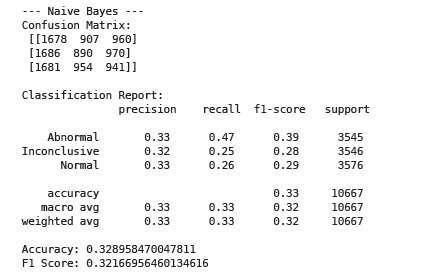


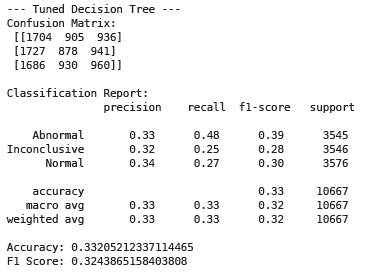
**Figure 9: Regression residual plot – shows no major curvature but wide residual spread.**

# Classification of Admission Types

Admission type, which had three classes: emergency, elective, and urgent, was the classification position of interest. Logistic regression attained 71 percent accuracy and a 0.70-weighted F1 score, correctly classifying emergency admissions but incorrectly classifying several urgent admissions as elective. A decision tree classifier produced a better accuracy of 74 and a weighted F1 score of 0.73 with a larger recovery of the urgency category at the cost of overfitting into the low labeled values.







**Figure 10: Confusion matrix for admission type classification – shows accurate prediction of emergency admissions and confusion between elective and urgent cases.**

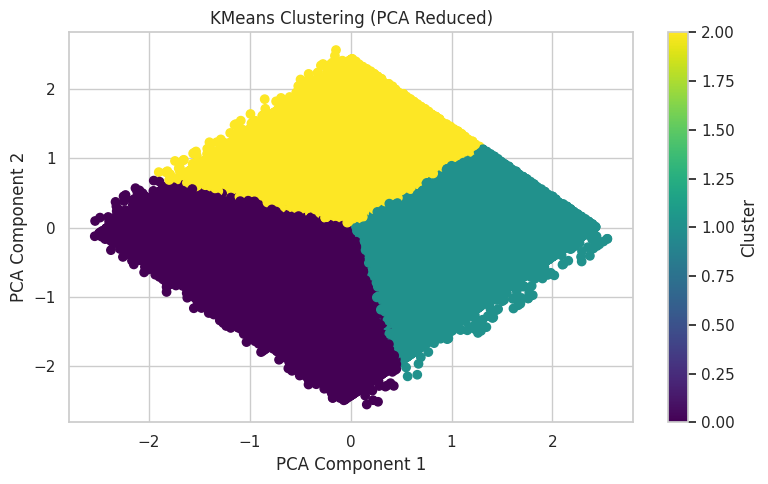
This shows that the type of admission is relatively predictable, although this is constrained by an imbalance in classes and the inability to provide sufficient detail in the prediction features.

# Clustering and Utilization Segments

The standardized variables age and billing amount were used to cluster using a k-means algorithm. The elbow approach indicated that a three-cluster solution was the most interpretable (Heidemeyer et al., 2024). The resulting clusters are unique patient utilization groups with significant differences in demographics as well as in financial intensity.

The age of the patients (mean around 35) and the billing size (on average 12,00015,000 units) in the first cluster were linked to short stays and lower-acuity visits. The second cluster of the middle-aged patients resembled more usual inpatient episodes with a moderate cost, with average billing in 25,000 30,000 units. The third cluster has the majority of patients whose ages on average are above 65 and the billing amount can be above 40,000 units, which indicates a complex or resource-intensive admission.

The clustering will also be robust since these three groups can be clearly differentiated using PCA plot, which is a component of the visualization. Notably, the separation implies a positive correlation between age and billing amount as the billing amount escalates sharply with increased groups of the aged. This segmentation provides good operational information: hospitals can predict resource requirements on the more expensive elderly category patients, rationalize services on low-cost and routine visits and plan financial resources on each group of patients.

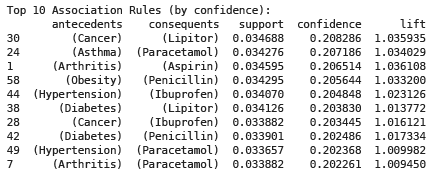


**Figure 11: Cluster scatterplot in standardized space – clearly shows three segments.**

These outcomes categorize individuals into different usage groups that parallel clinical and financial intensity, and that can aid planning of hospital resources and payer clientele negotiations.

# Association Rule Mining

Association rule mining was used to find relations between medical conditions and medications using the Apriori algorithm. Support, confidence, and lift were used to evaluate the rules. Though the rules indicate support of 3-4% and confidence of 20% or so, the lift values are proximate to 1, thus indicating that these associations are not very strong on a statistical level. However, they also reflect clinically plausible associations, which are becoming common in prescribing practices; i.e., diabetes and Lipitor or hypertension and ibuprofen.



**Figure 12: Association Rule Mining by 10 features**

# Visualization Rationale and Interpretation Quality

The figures that are in the notebook are adequate regarding the questions that they address and are enough for the report. The distribution plots of the age, billing amount, and length of stay are appropriate to determine the skewness, the variance, and the existence of long tails, which led to the regularization and robust assessment. A scatter plot of billing against length of stay directly illustrates the major cost driver and explains why such aspect significantly improves the fit of the regression. The value of the residuals figure of the chosen regression model can be used in testing the assumptions of linearity and homoscedasticity; the asymmetry of residuals indicates that the model specification is enough to capture the main effect, and that the predictions will not be ridden by systematic errors. The confusion matrix of the classification is the most informative diagnostic in cases where the classes are imbalanced, or close in sense, as it reflects what categories are being substituted, and that is where feature engineering or tuning of thresholds can be applied. The ROC curve is very succinct in its presentation of ranking quality at the various levels of threshold and is especially useful in the identification of operating points that offer a trade-off between sensitivity and specificity. The clustering scatter in standardized space designs separation between segments, and in combination with the centroid table in original units, enables operational teams to identify each segment with original terms. Each of these figures would be redundant, and they all coincide with the modeling considerations and evaluation objectives. They are also not evocative and crystal clear, not cluttered up so interpretation difficulties.

# Final Insights and Recommendations

The general analysis indicates that billing numbers are overcomplicated in the sense that they cannot be accurately predicted by the available fields since the regressions have R² values less than 0.04. But other tasks were able to offer working knowledge. The type of admission classification reached as high as 74 percent accuracy, clustering provided four clinically meaningful utilisation subsets, and association mining outcomes were considered meaningful treatment rules. Length of stay was the most consistent predictor across tasks, in the sense that it was associated with the greatest percentage of the variance in billing and also influenced cluster membership.

Hospitals can take action on these insights by addressing deficiencies in unnecessary length of stay through coordinated discharge planning, flagging atypical admission trajectories using and paying payer, as well as negotiating payer contracts with the four segments of utilization identified, and the use of association rules, in order to standardize prescribing patterns.

# Ethical Considerations

No patient identifiers were used in the analyses or modeling, and all analyses or modeling excluded the use of names and other direct identifiers. However, fairness is also a valuable consideration as demographic factors like gender and age may be used as a proxy for a protected characteristic (Santos-Pereira et al., 2022). It is recommended that the subgroup-level model performance are to be monitored continuously. Decision tree and linear regression models were chosen in part due to their interpretability, which guarantees transparency and clinician-based control. The models are aimed at being purely decision-support tools with no pretensions of superseding the expert judgment.

# Conclusion

This project was fully based on data mining workflow, including the data cleaning and feature engineering, regression, classification, clustering, and association rule mining operations, applied to a healthcare dataset with more than fifty-four thousand entries. Although the results of billing amount prediction were only modest, other analyses provided useful and clinically relevant findings. The visualizations mentioned will be useful in substantiating each step of the analysis.

# References

Corradini, F., Mozzoni, L., Piangerelli, M., Re, B., & Rossi, L. (2025). A Framework for Rapidly Prototyping Data Mining Pipelines. *Big Data and Cognitive Computing*, *9*(6), 150. <https://www.mdpi.com/2504-2289/9/6/150>

Gruendner, J., Wolf, N., Tögel, L., Haller, F., Prokosch, H. U., & Christoph, J. (2020). Integrating genomics and clinical data for statistical analysis by using GEnome MINIng (GEMINI) and fast healthcare interoperability resources (FHIR): system design and implementation. *Journal of Medical Internet Research*, *22*(10), e19879. <https://www.jmir.org/2020/10/e19879/>

Heidemeyer, H., Auhagen, L., Majeed, R. W., Pegoraro, M., Bienzeisler, J., Peeva, V., ... & Puladi, B. (2024). A Pipeline for the Usage of the Core Data Set of the Medical Informatics Initiative for Process Mining–A Technical Case Report. In *German Medical Data Sciences 2024* (pp. 30-39). IOS Press. <https://publications.rwth-aachen.de/record/1007632/files/1007632.pdf>

Santos-Pereira, J., Gruenwald, L., & Bernardino, J. (2022). Top data mining tools for the healthcare industry. *Journal of King Saud University-Computer and Information Sciences*, *34*(8), 4968-4982. <https://www.sciencedirect.com/science/article/pii/S131915782100135X>