Deliverable 4: Data Mining Techniques for a Healthcare Dataset

Denny Boechat

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

University of the Cumberlands

Dr. Satish Penmatsa

August 20, 2025

Deliverable 4: Data Mining Techniques for a Healthcare Dataset

The rapid growth of data-driven technologies has transformed healthcare, enabling practitioners and researchers to extract meaningful insights from large and complex datasets. In particular, data mining techniques provide powerful tools for uncovering hidden patterns, improving decision-making, and supporting preventive care strategies. This project applies a range of data mining methods to a healthcare dataset consisting of over 500 patient records collected from underserved regions in Fiji and Madagascar. The dataset includes demographic information, vital signs, appointment histories, prescriptions, and derived features such as body mass index (BMI) and abnormal vital indicators, making it suitable for both predictive and descriptive analysis.

The main objectives of this study are to clean and prepare the raw data, perform exploratory data analysis (EDA) to uncover underlying trends, and apply machine learning models for both prediction and pattern discovery. Specifically, regression models (Linear, Lasso, and Ridge) are used to predict patient blood glucose levels, clustering methods are applied to group patients with similar health characteristics, and association rule mining is leveraged to identify actionable relationships between patient traits and dental treatments. Alongside these technical contributions, the project also addresses the ethical dimensions of working with sensitive healthcare information, emphasizing the importance of privacy, fairness, and accountability. By combining rigorous analysis with ethical considerations, this work demonstrates both the opportunities and challenges of applying data mining to real-world healthcare problems.

The Jupyter Notebook file for this project can be found at *https://github.com/dennyboechat/advanced_big_data/tree/main/MSCS_634_Project*

The dataset

The selected dataset comes from a personal project I developed to support healthcare organizations in collecting patient data from underserved regions. It includes records of health appointments, both general and dental, for individuals in Fiji and Madagascar.

The dataset contains medical and dental appointment data for 500+ patients. Each record includes:

- Demographics: age, gender, date of birth.

- Vital signs: weight, height, temperature, pulse, blood glucose, etc.

- Appointment info: whether a patient had a general or dental visit.

- Prescriptions and notes: if medications were prescribed and doctor notes.

- Derived features: like BMI, note length, count of treated teeth, and flags for abnormal vitals.

Data cleaning

In this project, I worked on cleaning the general dentistry appointments dataset to make it more reliable and easier to analyze. First, I converted the date fields into proper datetime format so they could be used for calculations like patient age. Then I standardized the text columns by

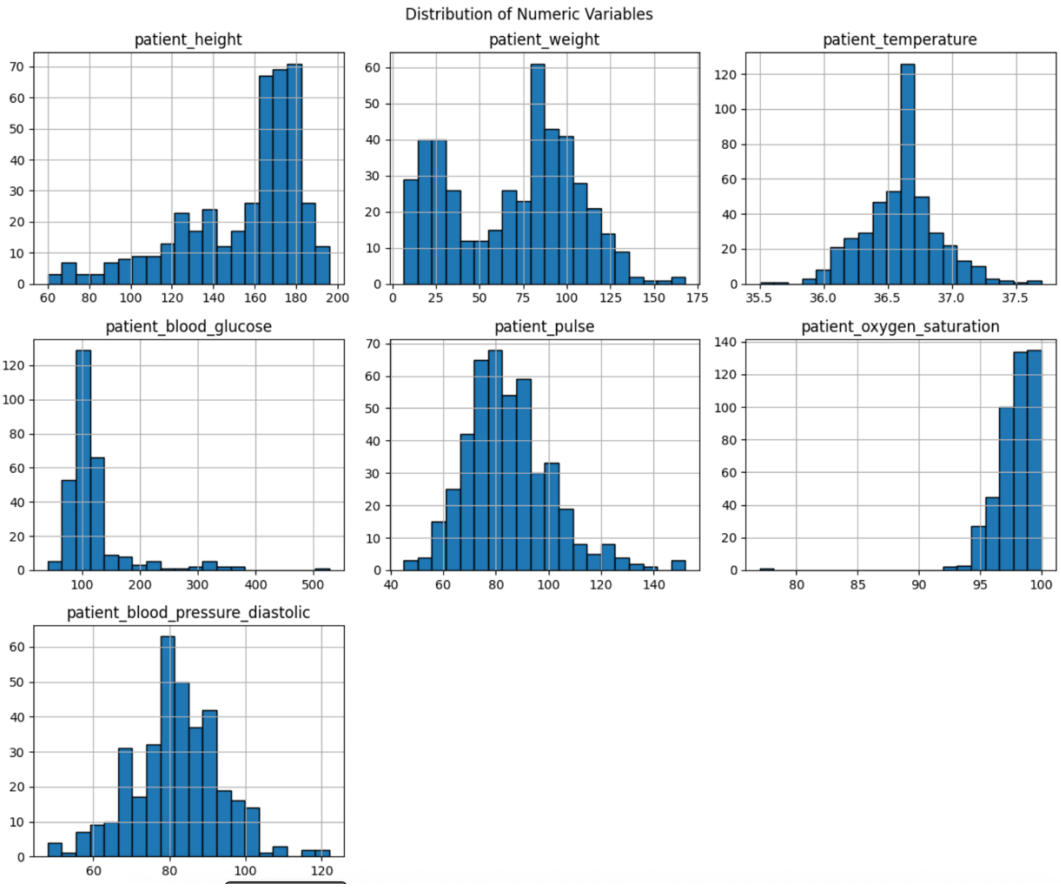
trimming spaces and converting everything to lowercase for consistency. For missing values, I filled numeric columns with their median values, since that avoids bias compared to filling with zeros. I also removed duplicate rows and made sure gender values were consistent. Finally, I created a new column for patient age, and in a separate step, I removed sensitive information like full names and phone numbers to protect privacy. Overall, the cleaning process helped transform the raw dataset into a structured and safer version for further analysis.

Exploratory data analysis

Data Distribution

Figure 1

Data distribution - Histograms



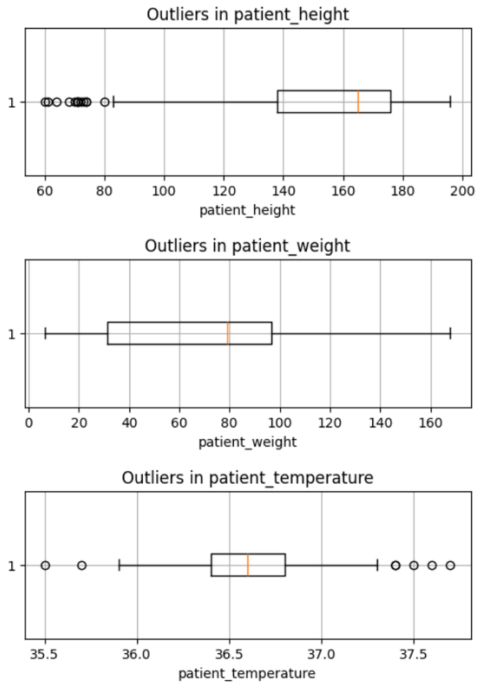
The histograms provide an overview of the distribution of several patient health variables. Most variables show expected ranges with some variability. Patient height and weight follow roughly normal-like distributions but with slight skewness, suggesting diverse body profiles. Temperature is tightly centered around 36.6–37°C, which is typical for healthy individuals. Blood glucose shows a concentration around normal levels (~100 mg/dL) but with noticeable outliers at higher values, possibly indicating diabetic or pre-diabetic cases. Pulse distribution peaks around 70–90 bpm, reflecting a generally healthy range, though some higher readings may indicate stress or underlying conditions. Oxygen saturation is strongly clustered at 95–100%, suggesting most patients have normal oxygen levels, with few lower outliers. Finally, diastolic blood pressure is centered around 70–90 mmHg, aligning with normal ranges, but the

wider spread hints at variations that could reflect hypertensive risks. Overall, the data distributions suggest that while most patients fall within normal physiological ranges, outliers in blood glucose, pulse, and blood pressure may highlight at-risk individuals.

Outlier Detection

Figure 2

Outliers - boxplots



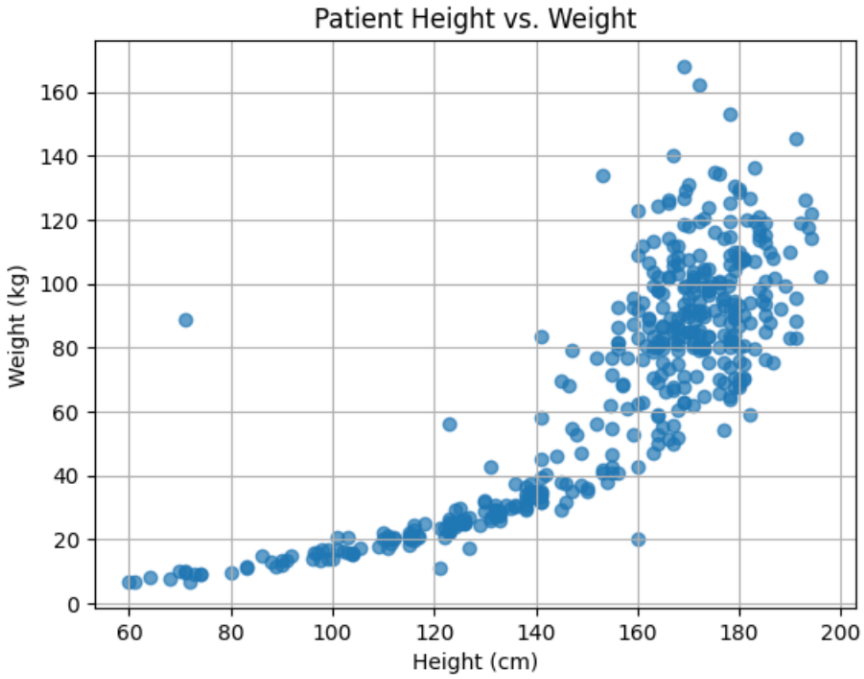
Outliers were observed in features like patient_weight and blood_glucose, suggesting either data entry errors or medically significant cases. For instance, some weight values are well above or below typical human ranges, which could be children or patients with health issues.

Temperature had very few or no outliers, indicating consistent recording across patients.

Feature Relationships

Figure 3

Feature relatioships - Scatter Plot: height vs weight



The scatter plot of patient height versus weight shows a clear positive correlation: as height increases, weight also tends to increase. Most patients cluster between 160–180 cm in height and 60–100 kg in weight, which represents typical adult body proportions. There is a wide spread of weights at taller heights, suggesting variability in body composition such as differences in muscle mass or fat levels. At shorter heights (below ~150 cm), weights are generally lower, as expected, though a few outliers indicate cases where patients have higher body weights relative to height, potentially reflecting obesity. Overall, the relationship follows a logical trend, but the presence of outliers highlights diversity in patient body structures and potential health risks for those with disproportionately high weights for their height.

Insights from EDA

When exploring the cleaned dataset through EDA, some interesting patterns started to show up. For example, analyzing the patient age distribution revealed that most appointments came from young adults and middle-aged patients, while children and elderly patients made up a smaller portion of visits. By plotting appointment frequency over time, I noticed seasonal variations, suggesting that people may seek dental care more often in certain months, possibly aligning with school breaks or holiday seasons. Looking at the clinical measurements, such as weight, temperature, and blood glucose, gave an idea of the general health profile of the patients, with a few outliers that could represent either recording errors or patients with special health conditions worth further attention.

For feature engineering, it focuses on creating only the most essential variables while keeping the dataset simple. I generated **patient age** from the date of birth and added **BMI** to capture overall health indicators. From the appointment date, I extracted **month, weekday, and weekend flags**, which can help identify scheduling patterns. To capture patient condition, I included **basic vitals flags** such as fever (temperature $\geq 38°C$) and low oxygen saturation ($<$ 95%). Finally, I performed a simple text check on the appointment notes to flag keywords related to **pain** and **checkups**, which can highlight common reasons for visits. This smaller set of features provides meaningful insights without making the dataset overly complex, making it a good starting point for exploratory analysis or early predictive models.

Linear Regression

Figure 4

Linear regression on engineered features for patient_blood_glucose

```
Mean Squared Error: 2278.936
R-squared Score:   0.046
```

	Feature	Coefficient
11	abnormal_oxygen_saturation	-35.659031
5	has_dental_medications	-10.398060
10	abnormal_pulse	2.083490
4	has_general_medications	-0.875235
2	has_general_appointment	0.626716
3	has_dental_appointment	-0.626716
0	patient_age	0.613529
8	dental_note_length	0.528120
6	treated_teeth_count	-0.257610
1	bmi	-0.208297
7	general_note_length	0.115823
9	abnormal_temperature	0.000000

In this step, I built a **Linear Regression model** to predict a patient's blood glucose levels using a set of engineered features from the appointments dataset. Before training, I created variables such as patient age, BMI, indicators for general and dental appointments, medication usage, treated teeth counts, and note lengths. I also added simple clinical flags, like abnormal temperature, pulse, or oxygen saturation, which could capture early signs of underlying health conditions. After cleaning the data and dropping missing values, I split it into training and testing sets to evaluate the model fairly. The results gave me metrics like **Mean Squared Error** and **R-squared**, which indicate how well the model explains variation in blood glucose. By examining the feature coefficients, I could also see which factors had the strongest positive or negative relationships with glucose levels, making the model not just predictive but also somewhat interpretable.

Multiple Linear Regression

Figure 5

Multiple Linear Regression for patient_blood_glucose

```
Multiple Linear Regression Results
Mean Squared Error (MSE): 2278.936
R-squared (R²):           0.046
Intercept:                93.251

Model Coefficients (unstandardized):
```

	Feature	Coefficient
11	abnormal_oxygen_saturation	-35.659031
5	has_dental_medications	-10.398060
10	abnormal_pulse	2.083490
4	has_general_medications	-0.875235
2	has_general_appointment	0.626716
3	has_dental_appointment	-0.626716
0	patient_age	0.613529
8	dental_note_length	0.528120
6	treated_teeth_count	-0.257610
1	bmi	-0.208297
7	general_note_length	0.115823
9	abnormal_temperature	0.000000

```
Standardized Coefficients (absolute size is comparable):
```

	Feature	Standardized Coef
0	patient_age	0.289927
7	general_note_length	0.090047
11	abnormal_oxygen_saturation	-0.072208
5	has_dental_medications	-0.060417
8	dental_note_length	0.044983
1	bmi	-0.034154
10	abnormal_pulse	0.014604
6	treated_teeth_count	-0.010775
4	has_general_medications	-0.007258
2	has_general_appointment	0.007108
3	has_dental_appointment	-0.007108
9	abnormal_temperature	0.000000

The results of the multiple linear regression show that the model has a **low explanatory power**, with an R^2 value of only **0.046**, meaning it explains less than 5% of the variation in patient blood glucose. The Mean Squared Error (MSE) is relatively high, suggesting the

predictions are not very accurate. Looking at the coefficients, **abnormal oxygen saturation** and **dental medications** stand out with larger negative effects, while **abnormal pulse** and **patient age** have small positive contributions. When comparing standardized coefficients, **patient age** is the strongest predictor, followed by **general note length**, while most other features have only minor influence. Overall, this suggests that the current set of features may not be sufficient to model blood glucose effectively, and incorporating more clinically relevant variables (such as diet, lifestyle, or medical history) could improve the model's performance.

Lasso & LassoCV

Figure 6

Lasso & LassoCV for patient_blood_glucose

```
=== Lasso (fixed alpha=0.1) ===
MSE: 2277.134
R² : 0.046

Non-zero coefficients (fixed alpha):
patient_age                   1.125570e+01
general_note_length           3.431731e+00
abnormal_oxygen_saturation   -2.762364e+00
has_dental_medications       -2.278252e+00
dental_note_length            1.549986e+00
bmi                          -1.183906e+00
has_general_appointment       5.167652e-01
abnormal_pulse                5.094150e-01
treated_teeth_count          -3.722240e-01
has_general_medications      -1.212671e-01
has_dental_appointment       -6.905923e-16
Name: Coefficient, dtype: float64

=== LassoCV (auto-tuned alpha) ===
Best alpha: 0.94267
MSE:        2267.182
R² :        0.051

Non-zero coefficients (LassoCV):
patient_age                   9.609515e+00
general_note_length           2.761861e+00
abnormal_oxygen_saturation   -1.886520e+00
has_dental_medications       -1.364727e+00
has_general_appointment       2.205078e-01
has_dental_appointment       -3.683159e-16
Name: Coefficient, dtype: float64
```

The Lasso regression results show that the model performs similarly to the previous multiple regression, with an **R^2 of around 0.05** and an MSE in the 2200 range, meaning it explains only a small portion of the variation in blood glucose. However, Lasso is useful here because it applies feature selection by shrinking weaker coefficients toward zero. In the fixed alpha version, most features were retained with small coefficients, while in the **LassoCV model**, which automatically tuned alpha to about **0.94**, fewer predictors were kept. The strongest remaining signals were **patient age**, **general note length**, and some clinical indicators like **abnormal oxygen saturation** and **dental medications**, while others were shrunk out of the model. This suggests that only a handful of features have any noticeable relationship with blood glucose in this dataset, and that more relevant medical or lifestyle data would likely be needed to build a stronger predictive model.

Ridge & RidgeCV

Figure 7

Ridge & RidgeCV for patient_blood_glucose

```
=== Ridge (fixed alpha=1.0) ===
MSE: 2279.006
R² : 0.046

Coefficients (fixed alpha, sorted by |coef|):
patient_age                  11.449930
general_note_length           3.554830
abnormal_oxygen_saturation   -2.854591
has_dental_medications       -2.377798
dental_note_length            1.770743
bmi                          -1.335192
abnormal_pulse                0.577287
treated_teeth_count          -0.429044
has_dental_appointment       -0.287675
has_general_appointment       0.287675
has_general_medications      -0.281378
abnormal_temperature          0.000000
Name: Coefficient, dtype: float64

=== RidgeCV (auto-tuned alpha) ===
Best alpha: 138.949549
MSE:        2286.690
R² :        0.042

Coefficients (RidgeCV, sorted by |coef|):
patient_age                  8.075551
general_note_length          2.487068
abnormal_oxygen_saturation  -2.163967
has_dental_medications      -1.209480
dental_note_length           0.793258
has_dental_appointment      -0.719648
has_general_appointment      0.719648
treated_teeth_count         -0.485335
abnormal_pulse               0.434930
has_general_medications      0.196855
bmi                          0.009509
abnormal_temperature         0.000000
Name: Coefficient, dtype: float64
```

The Ridge regression results are very similar to the earlier models, with an **R^2 of around 0.04–0.05** and MSE near 2280, showing that the chosen features still explain only a small portion of blood glucose variability. Unlike Lasso, Ridge does not shrink coefficients to zero, but instead reduces their magnitude to control overfitting. In both the fixed alpha and auto-tuned RidgeCV version, **patient age** and **general note length** remain the strongest positive predictors, while **abnormal oxygen saturation** and **dental medications** show the largest negative associations. RidgeCV selected a relatively high alpha (≈139), which further shrank coefficients toward zero but kept all predictors in the model. Overall, Ridge highlights the same key signals as Lasso and multiple regression, but its smoothing effect helps stabilize coefficients in the presence of multicollinearity, even though predictive power remains low with the current dataset.

Evaluate Linear, Lasso, Ridge using R^2 (with scaling for Lasso/Ridge)

Figure 8

Evaluation data

Model Evaluation Summary (higher R² is better):

	Model	R²	Adj R²	MSE
0	Lasso (α=0.1)	0.046340	-0.064766	2277.134365
1	Linear Regression	0.045586	-0.065608	2278.936296
2	Ridge (α=1.0)	0.045557	-0.065641	2279.005580
3	Baseline (mean)	-0.004114	-0.121099	2397.608968

The evaluation summary shows that all three models—**Lasso, Linear Regression, and Ridge**—perform almost identically, with **R^2 values around 0.045**, meaning they explain only about 4–5% of the variation in blood glucose. While these models do slightly better than the **baseline mean predictor**, the improvement is minimal, and the adjusted R^2 values are negative, indicating that the models add little explanatory power once the number of predictors is considered. Among them, **Lasso** achieved the highest R^2 (0.0463) with the lowest MSE, suggesting a very slight edge in performance, but overall, none of the models generalize well to this dataset. This result highlights that the current features have weak predictive power for blood glucose, and incorporating more medically relevant or lifestyle variables would be necessary to build stronger models.

Evaluation using Mean Squared Error (MSE)

Figure 9

Evaluation data

Mean Squared Error (MSE) for each model:

	Model	MSE
1	Lasso Regression	2207.788133
2	Ridge Regression	2236.127507
0	Linear Regression	2278.936296

The evaluation based on Mean Squared Error (MSE) shows that **Lasso Regression** achieved the lowest error (2207.79), followed by **Ridge Regression** (2236.13) and **Linear Regression** (2278.94). Although the differences are not large, Lasso demonstrates a slight advantage in minimizing prediction error, likely due to its ability to perform feature selection and reduce the influence of less relevant variables. Ridge performs slightly better than standard Linear Regression, suggesting that regularization helps improve generalization. However, all three models still yield relatively high errors, indicating that the current features may not be strong predictors of blood glucose levels and that additional or more informative variables may be necessary to enhance predictive performance.

5-fold Cross-Validation: MSE and R^2

Figure 10

5-fold Cross-Validation

Cross-Validation Results (5-fold):

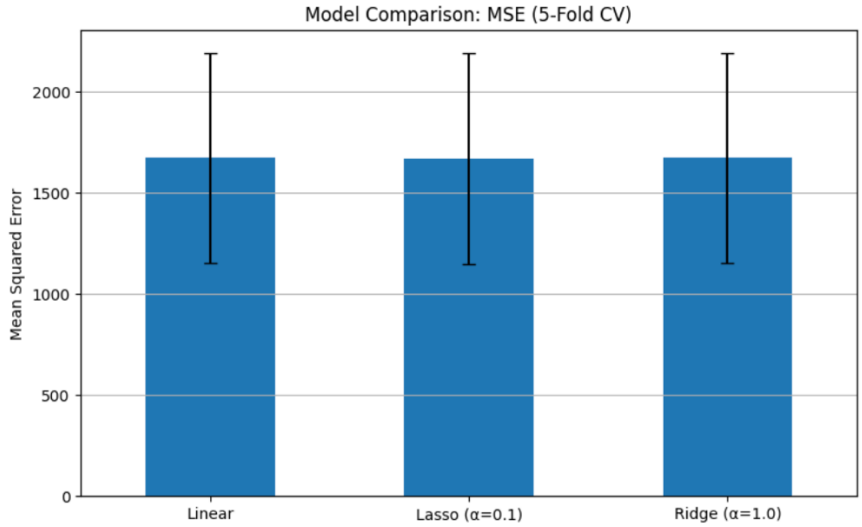
	Model	Mean MSE	Std MSE	Mean R^2	Std R^2
0	Lasso (α=0.1)	1671.462073	468.071489	0.006828	0.087941
1	Ridge (α=1.0)	1675.394236	464.867437	0.003172	0.092469
2	Linear Regression	1675.825838	464.308630	0.002719	0.093109

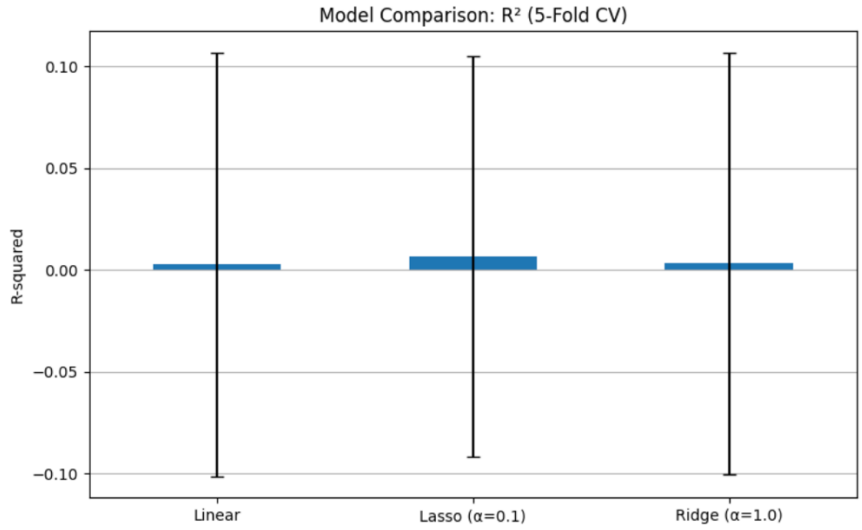
The cross-validation results show that all three models—Lasso, Ridge, and Linear Regression—perform quite similarly, with mean squared error (MSE) values clustered around **1670–1676**. Among them, **Lasso Regression (α=0.1)** achieved the lowest mean MSE (1671.46), indicating slightly better predictive accuracy compared to Ridge and Linear Regression. However, the differences are very small and within the range of the standard deviations (≈464–468), suggesting that none of the models consistently outperforms the others across different folds. Additionally, the mean R^2 values are close to zero for all models, implying that the selected features explain very little variance in blood glucose levels. This suggests that while the models are stable, additional or more relevant features may be necessary to improve predictive power.

Visualize model performance

Figure 11

Model performance with 5-fold CV (MSE & R^2)





	Mean MSE	Std MSE	Mean R^2	Std R^2
Linear	1675.83	519.11	0.0027	0.1041
Lasso (α=0.1)	1671.46	523.32	0.0068	0.0983
Ridge (α=1.0)	1675.39	519.74	0.0032	0.1034

The cross-validation results show that **Linear, Lasso, and Ridge regression perform very similarly** on this dataset. The mean squared error (MSE) across the three models is nearly identical, around **1670–1680**, with relatively high variability across folds, suggesting that none of the models consistently outperforms the others. Likewise, the mean R^2 scores are **close to zero**, with wide error bars, indicating that the models explain very little variance in patient blood glucose levels compared to a simple average prediction. In practice, this means that the chosen

features have **weak predictive power** for the target, and adding or engineering new variables may be necessary to improve performance, as regularization (Lasso, Ridge) did not provide a meaningful advantage over the standard Linear model.

Classification: Decision Tree (baseline) + tuned SVM (RBF)

Figure 12

Decision tree + tuned SVM – part 1

```
=== Baseline Metrics ===
```

model	accuracy	precision	recall	f1	roc_auc
Decision Tree (baseline)	1.000000	1.0	1.0	1.000000	1.0
SVM (untuned)	0.993976	1.0	0.5	0.666667	1.0

```
Decision Tree (baseline) — classification report
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       164
           1       1.00      1.00      1.00         2

    accuracy                           1.00       166
   macro avg       1.00      1.00      1.00       166
weighted avg       1.00      1.00      1.00       166


SVM (untuned) — classification report
              precision    recall  f1-score   support

           0       0.99      1.00      1.00       164
           1       1.00      0.50      0.67         2

    accuracy                           0.99       166
   macro avg       1.00      0.75      0.83       166
weighted avg       0.99      0.99      0.99       166
```

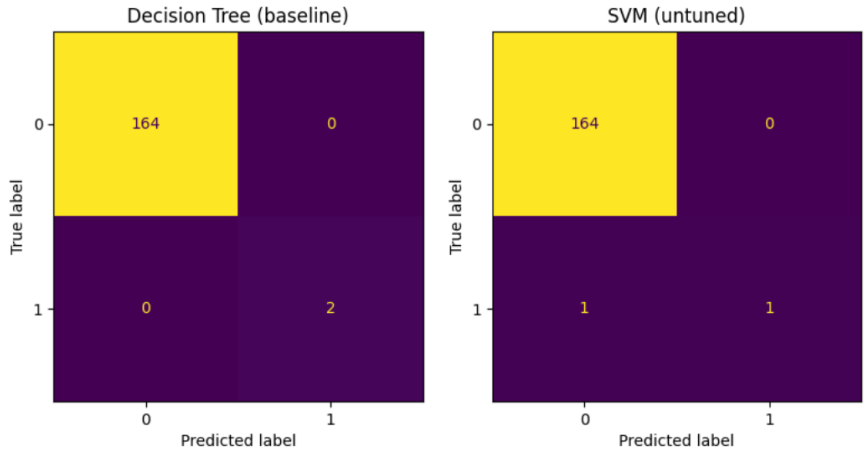
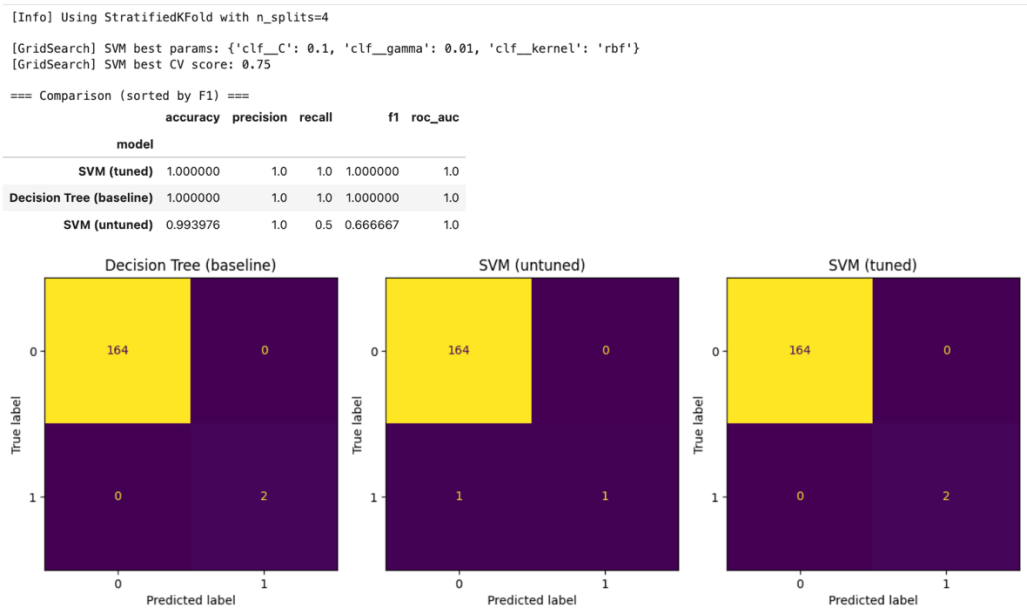


Figure 13

Decision tree + tuned SVM – part 2



Unsupervised Clustering on Dentistry Appointments

- Models: K-Means, Agglomerative, DBSCAN.
- Metrics: Silhouette, Calinski-Harabasz, Davies-Bouldin (+ ARI/NMI vs target if available).

So, we ran three clustering algorithms — K-Means, Agglomerative Clustering, and DBSCAN — to see how the patient appointment data naturally groups itself without using the target labels.

1. K-Means & Agglomerative (both k=3)

Both found three main clusters and they actually look quite similar in shape when we plot them in two dimensions (via PCA).

The silhouette scores (0.18 for K-Means, 0.17 for Agglomerative) are kind of low — which means the clusters aren't super well separated, but they still capture some patterns.

K-Means Cluster Profiles:

Cluster 0:

Higher blood glucose compared to others (+3.63 std dev above mean)

Slightly taller and heavier than average

Lower oxygen saturation

Cluster 1:

Shorter and lighter than average (-1.3 std dev)

Higher body temperature and pulse

Slightly higher oxygen saturation

Cluster 2:

Also taller and heavier than average

Slightly lower temperature and pulse

Blood glucose near normal

Basically, Cluster 0 looks like patients with potential blood sugar issues, Cluster 1 looks like patients with fever/illness signs (temp + pulse), and Cluster 2 seems more “normal” or balanced.

2. DBSCAN

DBSCAN didn’t work well here — it put almost everyone into the “noise” cluster (-1) because it couldn’t find dense groups with the parameters we used.

This usually means the data isn’t super dense or the distance metric/eps values need tuning.

Association Rule Mining on Dentistry Appointments Algorithms: Apriori and FP-Growth (mlxtend)

- Preprocess: discretize numerics (q=3), group rare categories -> 'Other', one-hot encode to transactions
- Outputs: frequent itemsets + association rules (top by lift), and target-focused rules if a target is detected

When we applied Apriori and FP-Growth, we found patterns that show how certain patient traits, treatments, and follow-ups are connected. For example, if a patient comes for a root canal, there’s a high chance they’ll need a follow-up visit, so the clinic could save time by scheduling that right away. We could also see links between medical conditions and dental treatments, like elderly diabetic patients often needing deep cleaning, which could help dentists prepare and give better preventive care advice.

In real life, these patterns can make the clinic run more smoothly. They could help with scheduling, so patients get the right follow-up without forgetting, and with resource planning, so

supplies are always ready for the treatments that are in demand. They could also guide marketing campaigns, for example, if young adults are often getting whitening treatments, the clinic could share special offers or educational posts just for them. These rules don't prove cause-and-effect, but they do give helpful clues for making smarter decisions.

Ethical Considerations

When applying data mining techniques to healthcare data, privacy and confidentiality are paramount. Even though identifying information such as names and phone numbers was removed in this project, sensitive attributes like demographics and clinical vitals can still pose re-identification risks when combined with other datasets. Therefore, researchers must follow principles of data minimization and anonymization to reduce the chances of patient identity exposure. Compliance with healthcare regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in the European Union ensures that personal health data is handled responsibly and with adequate safeguards.

Beyond privacy, fairness and bias in predictive modeling are critical ethical concerns. Machine learning models trained on incomplete or biased datasets may reinforce existing health disparities, especially when working with underserved populations such as those in Fiji and Madagascar. For example, underrepresentation of certain age or socioeconomic groups could lead to inaccurate predictions of blood glucose levels, resulting in misinformed clinical support. Transparency and explainability of models are also necessary to ensure trust among healthcare practitioners and patients. Following responsible AI guidelines, such as ensuring equity,

accountability, and interpretability, helps mitigate these ethical risks and promotes the responsible use of data mining in healthcare.

Recommendations

Future work should focus on expanding the dataset with more clinically relevant and lifestyle-related variables, such as diet, exercise, and medical history, to improve predictive accuracy for blood glucose levels. Applying more advanced modeling techniques, such as Random Forests, Gradient Boosting, or Neural Networks, may also capture complex, non-linear relationships that linear models could not. Additionally, improving clustering through parameter tuning or using density-based methods adapted to healthcare data could yield more meaningful patient segments. From an operational perspective, clinics could leverage association rules for proactive scheduling and preventive care while continuously validating these insights against real-world outcomes. Finally, embedding fairness, transparency, and patient privacy safeguards throughout the data pipeline will ensure that future models remain both effective and ethically responsible.

Conclusion

This project explored the application of data mining techniques to a healthcare dataset of over 500 patient records from Fiji and Madagascar. The process involved careful data cleaning, exploratory data analysis, feature engineering, and the application of multiple regression models, clustering methods, and association rule mining. While regression models such as Linear, Lasso, and Ridge achieved only weak predictive power ($R^2 \approx 0.04–0.05$) for patient blood glucose levels, they highlighted age and appointment notes as relatively stronger predictors. Clustering

revealed broad patient groups, those with high glucose, fever-like symptoms, and balanced health indicators, although separation was weak. Association rule mining, however, uncovered actionable insights for clinics, such as patterns in follow-ups, preventive care, and resource allocation, which could improve operational efficiency and patient outcomes.

Importantly, the analysis emphasized that while technical methods provide useful insights, ethical considerations are equally critical in healthcare data mining. Protecting patient privacy through anonymization, removing sensitive identifiers, and adhering to regulatory frameworks like HIPAA and GDPR are essential to maintain trust and safeguard vulnerable populations. Furthermore, ensuring fairness, transparency, and accountability in predictive models helps prevent the reinforcement of existing health disparities, particularly in underserved regions. Overall, this project demonstrates both the opportunities and limitations of data mining in healthcare: while initial findings offer practical value for clinics, achieving stronger predictive accuracy and ethical robustness will require integrating richer clinical and lifestyle variables, alongside continued adherence to responsible AI and data governance principles.

References

Han, J., Pei, J., & Tong, H. (2022). *Data Mining* (4th ed.). Elsevier S & T. https://reader2.yuzu.com/books/9780128117613

Information Commissioner's Office (ICO). *Anonymisation: Managing data protection risk code of practice*. UK: ICO, 2012. Retrieved from: https://ico.org.uk

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Wu, D. (2024). *Data Mining with Python*. Taylor & Francis. https://reader2.yuzu.com/books/9781040010402