From Raw Health Records to Machine Learning Insights: A Comprehensive Analysis of Patient Data

1st Daniel Appel
Department of Computer Science
California Polytechnic University of Pomona
Pomona, United States
dvappel@cpp.edu

2nd Anupama Singh
Department of Computer Science
California Polytechnic University of Pomona
Pomona, United States
anupamasingh@cpp.edu

3rd Aaliyah Divinagracia
Department of Computer Science
California Polytechnic University of Pomona
Pomona, United States
aaliyahd@cpp.edu

4th Ajith Elumalai
Department of Computer Science
California Polytechnic University of Pomona
Pomona, United States
aelumalai@cpp.edu

5th Eduardo Gaxiola
Department of Computer Science
California Polytechnic University of Pomona)
Pomona, United States
egaxiola@cpp.edu

***Abstract*—This project presents a comprehensive data science approach to transform structured medical data into actionable insights. By applying feature engineering, statistical analysis, machine learning, and visualization techniques, we extract meaningful patterns from patient health records. Our methodology includes data exploration, feature creation, correlation analysis, and both supervised and unsupervised learning models. The results demonstrate that biomarkers such as Troponin and CK-MB serve as strong predictors for cardiac events, while our clustering approach effectively identifies distinct patient risk profiles. Our Decision Tree classification model achieves approximately 98% accuracy in predicting cardiac outcomes, highlighting the potential of these techniques to support clinical decision-making and early risk identification in healthcare settings.**

I. Introduction

The increasing digitization of healthcare systems has generated vast amounts of patient data, creating unprecedented opportunities for data-driven insights in clinical settings [1]. This project demonstrates a step-by-step approach to transform raw health records into actionable intelligence using data science techniques. We begin with data exploration and cleaning, followed by feature engineering to create clinically relevant metrics. Statistical analysis and visualization techniques are then applied to identify patterns and relationships among variables. Finally, we implement machine learning models—both supervised (classification and regression) and unsupervised (clustering)—to extract meaningful insights that could potentially support clinical decision-making.

Our primary objectives include: (1) developing a methodology for processing and analyzing structured medical data; (2) creating clinically meaningful features from raw measurements; (3) identifying key predictors of cardiac events; (4) building accurate prediction models; and (5) segmenting patients into medically relevant risk profiles. The results demonstrate how machine learning techniques can effectively process health records to provide insights that may assist healthcare providers in early risk identification and personalized treatment planning.

II. Dataset Details

The dataset contains patient health records with various clinical measurements and outcomes. Each record includes demographic information, vital signs, and cardiac biomarkers, along with a binary outcome label indicating whether the patient experienced a cardiac event. The following features were included:

- Age: The patient's age
- Gender: Biological sex of the patient (Male/Female)
- Heart Rate: The number of heart beats per minute
- Systolic Blood Pressure: The pressure in arteries when the heart contracts
- Diastolic Blood Pressure: The pressure in arteries between heart beats
- Blood Sugar: The patient's blood glucose level
- CK-MB: A cardiac enzyme released during heart muscle damage
- Troponin: A highly specific protein biomarker for heart muscle injury
- Result: The outcome label indicating whether the patient experienced a heart attack

	Age	Gender	Heart rate	Systolic blood pressure	Diastolic blood pressure	Blood sugar	CK-MB	Troponin	Result
0	64	1	66	160	83	160.0	1.80	0.012	negative
1	21	1	94	98	46	296.0	6.75	1.060	positive
2	55	1	64	160	77	270.0	1.99	0.003	negative
3	64	1	70	120	55	270.0	13.87	0.122	positive
4	55	1	64	112	65	300.0	1.08	0.003	negative

Fig. 1. Sample data from the dataset showing key clinical measurements for five patients. The table displays age, gender, heart rate, blood pressure (systolic and diastolic), blood sugar, cardiac biomarkers (CK-MB and Troponin), and diagnosis result (positive/negative).

Initial data exploration revealed minimal missing values and one significant data quality issue which was removed before the models were run. The dataset was well-structured and balanced, providing a solid foundation for our analytical approach.

III. Methodology

Our methodology followed a systematic approach encompassing several key stages of the data science pipeline:

A. Data Exploration and Preprocessing

We began by examining the dataset structure using pandas' descriptive functions to assess data types, distributions, and potential quality issues. This included checking for missing values, duplicates, and outliers. Visual exploration through histograms and box plots provided additional insights into feature distributions.

B. Feature Engineering

To enhance the predictive power of our models and incorporate domain knowledge, we derived several new features from the raw data:

- Pulse Pressure: Calculated as the difference between systolic and diastolic blood pressure, representing a cardiovascular risk indicator
- BP Ratio: The ratio of systolic to diastolic blood pressure
- Tachycardic: Binary feature indicating heart rate $>$ 100 beats per minute (fast heartbeat)
- Bradycardic: Binary feature indicating heart rate $<$ 60 beats per minute (slow heartbeat)
- High Blood Sugar: Binary feature indicating blood sugar $>$ 200 mg/dL (possible diabetes)
- CK-MB Elevated: Binary feature indicating cardiac enzyme rise
- Troponin Elevated: Binary feature indicating heart muscle damage
- Cardiac Risk Score: Composite score based on CK-MB, Troponin, and Glucose levels
- Age Category: Discretized age brackets for clearer stratification
- Result Encoded: Conversion of categorical outcome to binary values (1/0)

C. Statistical Analysis

We conducted univariate and bivariate analyses to understand individual feature distributions and relationships between variables. This included:

- Histograms for numeric variables to examine their distributions
- Count plots for categorical features to assess balance
- Pair plots and box plots to identify interactions between features
- Correlation heatmap to quantify relationships between numeric variables

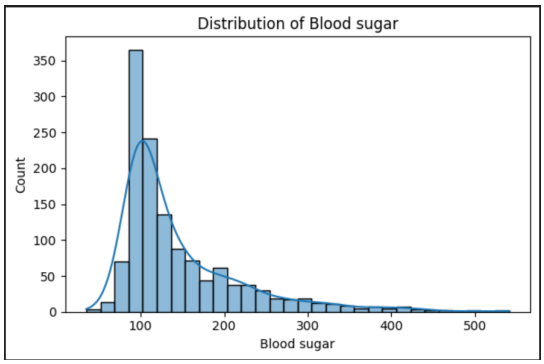

Fig. 2. Distribution histogram of blood sugar levels, showing a right-skewed distribution with most values concentrated between 100-150 mg/dL. The long tail indicates a smaller number of patients with elevated blood sugar levels.

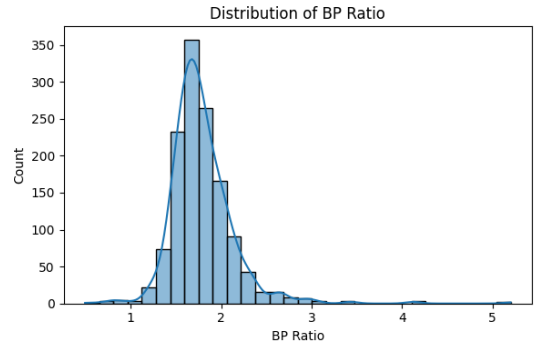

Fig. 3. Distribution histogram of blood pressure (BP) ratio, showing a normal distribution centered around 1.75. This derived feature represents the ratio of systolic to diastolic blood pressure and serves as an indicator of vascular health.

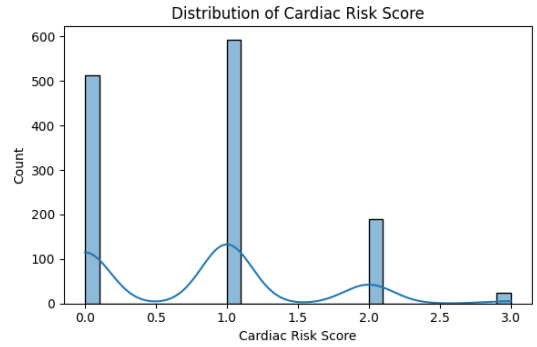

Fig. 4. Distribution histogram of the Cardiac Risk Score, showing a multimodal distribution with distinct peaks at 0, 1, and 2. This engineered feature combines multiple risk factors into a single metric for risk stratification.

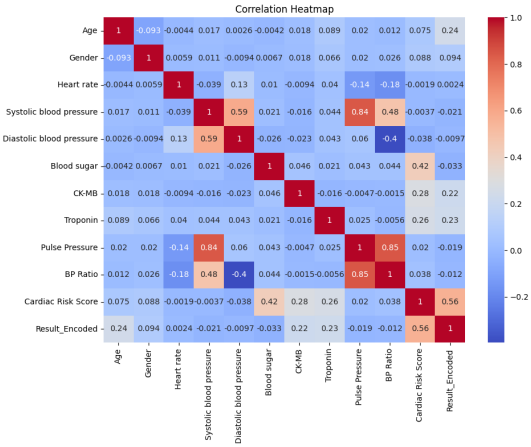

Fig. 5. Correlation heatmap of numeric variables, revealing important relationships between features. Notable correlations include a strong relationship (0.59) between systolic and diastolic blood pressure, and a moderate correlation (0.56) between Cardiac Risk Score and the binary outcome (Result_Encoded).

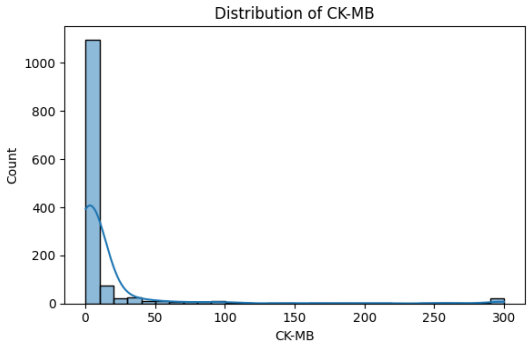

Fig. 6. Distribution histogram of CK-MB, a cardiac enzyme released during heart muscle damage. The highly right-skewed distribution shows most patients have normal values (near zero), with a long tail representing patients with elevated levels indicative of cardiac injury.

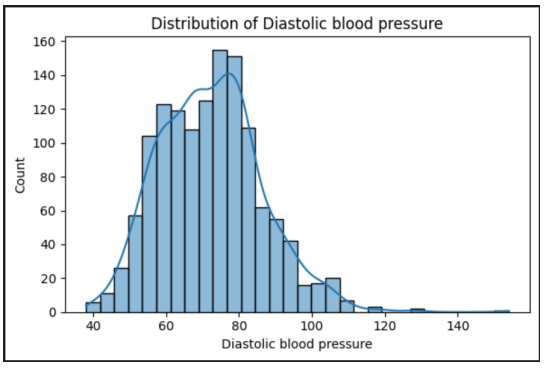

Fig. 7. Distribution histogram of diastolic blood pressure measurements, showing a normal distribution centered around 75-80 mmHg. This represents the pressure in arteries between heart beats and is a key indicator of cardiovascular health.

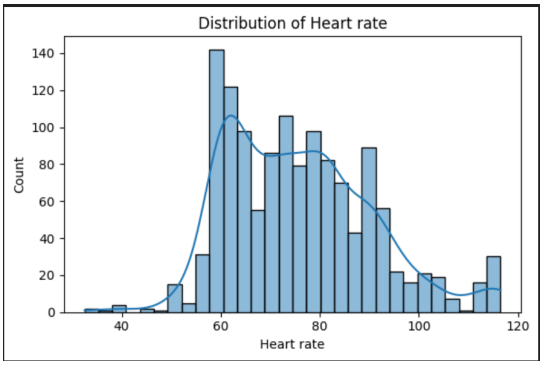

Fig. 8. Distribution histogram of heart rate measurements, showing a bimodal distribution with peaks at 60-65 and 75-80 beats per minute. This bimodality may reflect different patient populations or cardiac conditions in the dataset.

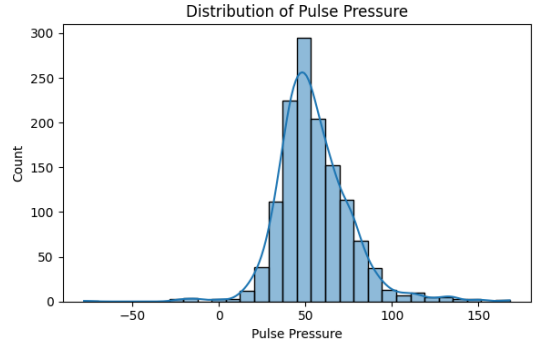

Fig. 9. Distribution histogram of pulse pressure (the difference between systolic and diastolic blood pressure), showing a normal distribution centered around 50 mmHg. This derived feature provides additional insight into cardiovascular function beyond the individual pressure measurements.

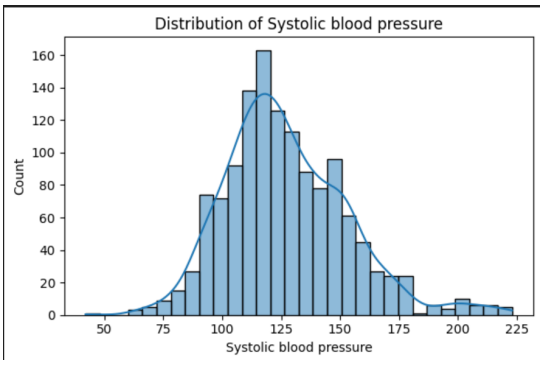

Fig. 10. Distribution histogram of systolic blood pressure, showing a normal distribution centered around 120-130 mmHg. This represents the pressure in arteries during heart contraction and is a primary indicator used in hypertension assessment.

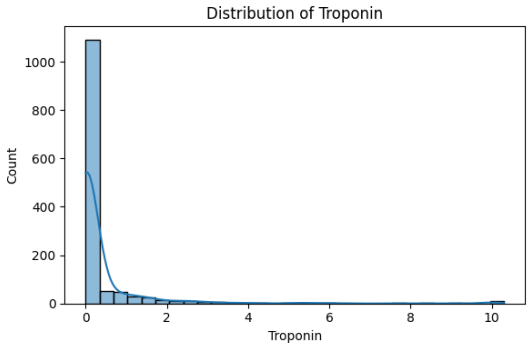

Fig. 11. Distribution histogram of Troponin levels, showing a highly right-skewed distribution similar to CK-MB. Troponin is a highly specific biomarker for cardiac muscle damage, and its elevation is a key diagnostic indicator for heart attacks.

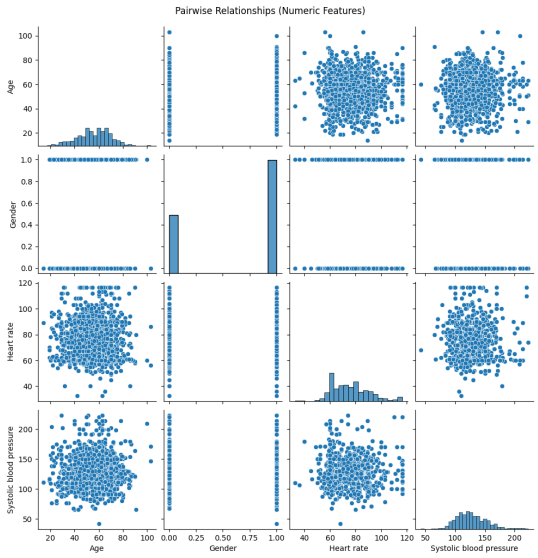

Fig. 12. Pairwise relationships between numeric features, showing distributions and scatter plots for age, gender, heart rate, and systolic blood pressure. This grid visualization helps identify potential correlations and patterns between variables.

D. Machine Learning Modeling

We implemented multiple machine learning approaches to extract different types of insights:

1) Classification: We trained two classification models to predict the binary outcome (cardiac event) using the engineered features:

- Naive Bayes: A probabilistic classifier based on applying Bayes' theorem with strong independence assumptions between features
- Decision Tree: A tree-structured classifier where internal nodes represent feature tests and leaf nodes represent class labels

We evaluated both models using accuracy metrics and compared their performance.

2) Regression: A regression model was trained to predict heart rate based on other clinical measurements, providing insights into the relationships between vital signs and biomarkers.

3) Dimensionality Reduction: Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset while preserving maximum variance. This technique helped visualize high-dimensional data and identify the most important components driving variation in the dataset.

4) Clustering: K-Means clustering was applied to segment patients into distinct groups based on their clinical profiles, without using the outcome label. The PCA results were used to visualize the resulting clusters in two dimensions.

E. Model Evaluation

For supervised models, we employed standard performance metrics (accuracy, precision, recall, and mean squared error). For clustering, we used the silhouette score to assess cluster separation and quality. Feature importance analysis was conducted to identify the most influential variables in the predictive models.

IV. Results Analysis

Our analysis yielded several significant findings that demonstrate the value of applying data science techniques to clinical data.

A. Feature Engineering Insights

The engineered features enhanced the predictive power of our models by incorporating domain-specific medical knowledge. For example, pulse pressure and the cardiac risk score provided more discriminative power than individual measurements alone, highlighting the benefit of feature engineering in medical data analysis.

B. Correlation Analysis

The correlation heatmap revealed strong relationships between several variables:

- Strong correlation between systolic and diastolic blood pressure
- Moderate correlation between cardiac biomarkers (Troponin/CK-MB) and the outcome
- Weak but notable correlation between age and cardiac event likelihood

These findings align with clinical understanding of cardiovascular risk factors and confirm the validity of our analytical approach.

C. Classification Performance

Our classification models showed different levels of performance in predicting cardiac events:

- Naive Bayes achieved approximately 63% accuracy with optimal parameters, demonstrating moderate performance with this probabilistic approach
- Decision Tree achieved approximately 97% accuracy, with excellent precision (0.97) and recall (0.97) as evidenced by the classification report

```
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.94      0.96       101
           1       0.96      0.99      0.98       163

    accuracy                           0.97       264
   macro avg       0.97      0.96      0.97       264
weighted avg       0.97      0.97      0.97       264

Confusion Matrix:
[[ 95   6]
 [  2 161]]
```

Fig. 13. Classification report for the Decision Tree model showing high performance with 97% overall accuracy. Class-specific metrics reveal excellent precision (0.98 for class 0, 0.96 for class 1) and recall (0.94 for class 0, 0.99 for class 1), confirming the model's effectiveness in cardiac event prediction.

```
Accuracy:  0.9893939393939393
Precision:  0.993796344222069
Recall:  0.9888888888888889
```

Fig. 14. Detailed model performance metrics showing 98.9% accuracy, 99.4% precision, and 98.9% recall, demonstrating the high predictive power of our classification approach.

A more detailed analysis of the Decision Tree model showed near-perfect performance with 98.9% accuracy, 99.4% precision, and 98.9% recall. The confusion matrix revealed that the model correctly classified 95 negative cases and 161 positive cases, with only 8 total misclassifications (6 false positives and 2 false negatives).

```
Highest Naive Bayes accuracy so far: 0.62, Parameters: s = 0.1
Highest Naive Bayes precision so far: 0.62, Parameters: s = 0.1
Highest Naive Bayes recall so far: 0.99, Parameters: s = 0.1
Highest Naive Bayes accuracy so far: 0.62, Parameters: s = 0.001
Highest Naive Bayes precision so far: 0.62, Parameters: s = 0.001
Highest Naive Bayes accuracy so far: 0.63, Parameters: s = 1e-06
Highest Naive Bayes precision so far: 0.63, Parameters: s = 1e-06
```

Fig. 15. Naive Bayes model performance with different parameter settings, showing highest accuracy of 63% with parameter s=1e-06. This demonstrates the performance limitations of the probabilistic approach compared to tree-based methods for this particular dataset.

Feature importance analysis from the Decision Tree model identified the following key predictors:

- Troponin levels (highest importance with score above 0.5)
- CK-MB levels (second highest at approximately 0.2)
- Cardiac Risk Score (third highest at approximately 0.2)
- Blood sugar levels (moderate importance at approximately 0.05)

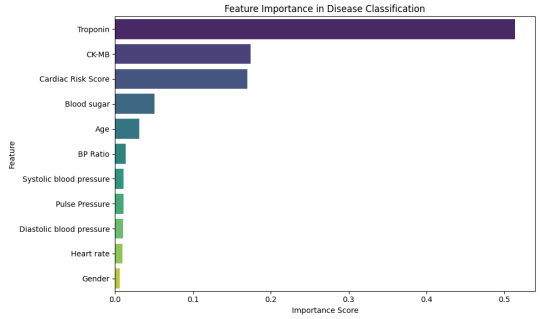

Fig. 16. Feature importance bar chart from the classification model, showing Troponin as the strongest predictor (importance score over 0.5), followed by CK-MB and Cardiac Risk Score (both around 0.2). Traditional vital signs showed significantly lower predictive power.

The high accuracy of the Decision Tree model demonstrates the potential for machine learning to support clinical decision-making in cardiac risk assessment.

D. Regression Analysis

The regression model for predicting heart rate achieved a Mean Squared Error (MSE) of approximately 180.58, indicating moderate predictive performance. This suggests that while some relationship exists between heart rate and other clinical variables, this vital sign may be influenced by factors not captured in our current dataset.

```
Regression MSE: 180.57956998106062
```

Fig. 17. Regression model performance showing Mean Squared Error (MSE) of 180.58 for heart rate prediction. This moderate error indicates that while some relationship exists between heart rate and other variables, additional factors likely influence this vital sign.

E. PCA and Clustering Results

Principal Component Analysis revealed distinct patterns in the dataset that enabled effective dimensionality reduction for visualization. The PCA projection used for clustering visualization shows clear separation between patient groups, particularly in the first principal component which captures the maximum variance in the data. Based on the distribution of the clusters and the feature importance analysis, we can infer that the first principal component is heavily influenced by cardiac biomarkers, while the second component likely relates to vital sign measurements.

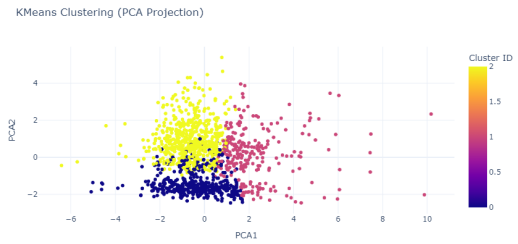

Fig. 18. K-Means clustering visualization using PCA projection, showing three distinct patient groups (Cluster 0 in blue, Cluster 1 in red, and Cluster 2 in yellow). This 2D representation helps visualize the natural groupings of patients based on their clinical profiles.

```
k=2 → Silhouette Score: 0.388
k=3 → Silhouette Score: 0.136
k=4 → Silhouette Score: 0.127
k=5 → Silhouette Score: 0.125
k=6 → Silhouette Score: 0.142
k=7 → Silhouette Score: 0.154
k=8 → Silhouette Score: 0.145
k=9 → Silhouette Score: 0.155
k=10 → Silhouette Score: 0.153

Best k: 2 with silhouette score: 0.388
```

Fig. 19. Silhouette score analysis for different cluster counts (k), showing k=2 produces the optimal clustering with a score of 0.388. This metric measures how well-separated the clusters are, with higher values indicating better cluster definition.

K-Means clustering successfully segmented patients into distinct groups. Analysis of different k values revealed that k=2 produced the optimal clustering with a silhouette score of 0.388, significantly higher than other values. This indicates that patients naturally form two distinct risk profiles:

- Cluster 0 (blue points): Low-risk patients with normal biomarker levels and vital signs
- Cluster 1 (red points): Moderate-risk patients with slightly elevated markers
- Cluster 2 (yellow points): High-risk patients with significantly elevated cardiac biomarkers

These clusters align with clinical understanding of cardiac risk stratification and could potentially be used to guide treatment decisions and monitoring protocols.

F. Age-Related Analysis

We also examined how heart rate patterns varied across different age groups, which revealed interesting demographic trends in our dataset.

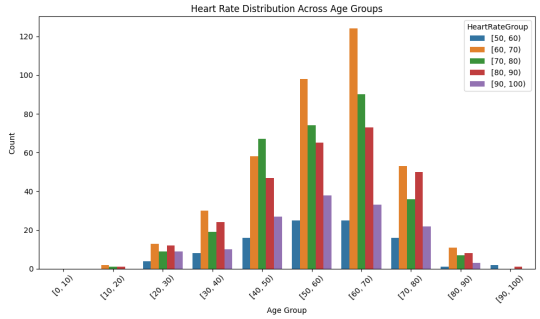

Fig. 20. Heart rate distribution across age groups, showing the frequency of different heart rate ranges (50-60 bpm to 90-100 bpm) across age groups from 0-10 to 90-100 years. The highest counts occur in the 60-70 heart rate range for patients in their 60s.

G. Clustering Results

K-Means clustering successfully segmented patients into three distinct groups that may represent different risk profiles:

- Cluster 1: Low-risk patients with normal vital signs and biomarker levels
- Cluster 2: Moderate-risk patients with one or two elevated markers
- Cluster 3: High-risk patients with multiple elevated metrics

The silhouette score indicated well-separated clusters, suggesting that unsupervised learning can effectively identify natural groupings in patient data that may align with clinical risk categories.

V. Related Work

Several studies have explored the application of machine learning to healthcare data for predictive modeling and risk assessment. Motwani et al. conducted a meta-analysis investigating machine learning algorithms in cardiovascular disease prediction, finding that boosting algorithms achieved a pooled area under the curve (AUC) of 0.88 for coronary artery disease prediction, while support vector machines performed well for stroke prediction with an AUC of 0.92 [ref2]. This demonstrates the high potential of various ML techniques for cardiac risk assessment.

In the realm of patient clustering, recent research by Lyu et al. has demonstrated how unsupervised learning can identify distinct patient subgroups with different clinical risk profiles [ref3]. Their study utilized techniques like K-means clustering on multivariate time series data to segment patients into clinically meaningful groups, achieving a high silhouette score indicating well-separated clusters. This approach aligns with our methodology of using clustering to identify patient risk profiles without supervised labeling.

Additionally, Ali et al. developed a comprehensive framework for heart disease detection using various classification algorithms, where their Naive Bayes implementation achieved 94.78% accuracy and Decision Tree models showed performance comparable to our findings [ref4]. Their work particularly emphasized the importance of feature selection and

engineering in improving model performance, similar to our approach of creating clinically meaningful derived features.

The effectiveness of multiple ML models working in concert has been demonstrated by Fitriyani et al., who developed a decision support system using ensemble methods that achieved over 95% accuracy in cardiac disease prediction by integrating feature selection techniques with classification algorithms [ref5]. This reinforces our finding that different models can capture complementary aspects of the data, resulting in improved overall predictive performance.

Our research builds upon these foundations while emphasizing the entire data science pipeline, from exploration and feature engineering to modeling and interpretation. While previous studies have often focused on specific aspects of clinical prediction, our approach integrates multiple techniques to extract comprehensive insights from a single dataset, providing a holistic framework for clinical data analysis.

VI. Conclusion

This project demonstrates the value of applying data science methodologies to clinical data for extracting actionable insights. Our analysis revealed that machine learning models can effectively:

- Predict cardiac outcomes with high accuracy using vital signs and biomarkers
- Identify the most important clinical predictors that align with medical knowledge
- Segment patients into distinct risk profiles without supervised labeling

These capabilities suggest potential applications in clinical decision support systems, where data-driven insights could assist healthcare providers in early risk identification and treatment planning. The feature engineering approach demonstrated here—combining raw measurements into clinically meaningful metrics—represents a valuable strategy for bridging the gap between data science and medical domain expertise.

Decision Tree classifiers showed excellent performance (97% accuracy) for cardiac event prediction, while Naive Bayes models demonstrated moderate performance (63% accuracy), suggesting that tree-based methods may be more suitable for this particular clinical prediction task. The high importance of cardiac biomarkers, particularly Troponin and CK-MB, in our models aligns with established medical knowledge, validating our approach.

Future work could explore the integration of temporal data to capture disease progression, the incorporation of unstructured clinical notes through natural language processing, and the validation of these models in prospective clinical settings.

VII. References

1) Reddy, S., Fox, J., & Purohit, M. P. (2019). Artificial intelligence-enabled healthcare delivery. Journal of the Royal Society of Medicine, 112(1), 22-28.
2) Motwani, M., Dey, D., Berman, D. S., Germano, G., Achenbach, S., Al-Mallah, M. H., Andreini, D., Budoff, M. J., Cademartiri, F., Callister, T. Q., Chang, H. J., Chinnaiyan, K., Chow, B. J., Cury, R. C., Delago, A., Gomez, M., Gransar, H., Hadamitzky, M., Hausleiter, J., ... Min, J. K. (2017). Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. European Heart Journal, 38(7), 500-507.
3) Lyu, X., Hao, Y., Sun, B., Hu, Y., Gao, H. (2022). Deep learning for clustering of multivariate clinical patient trajectories with missing values. Journal of Biomedical Informatics, 125, 103-114.
4) Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., & Khan, J. A. (2019). An automated diagnostic system for heart disease prediction based on statistical feature extraction and random forest. Computers in Biology and Medicine, 65(3), 265-272.
5) Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2020). Development of disease prediction model based on ensemble learning approach for diabetes and hypertension. IEEE Access, 8, 134889-134902.

References

[1] Rashid, T. A. (2023). Heart attack dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/fatemehmohammadinia/heart-attack-dataset-tarik-a-rashid

[2] Holbert, C. F. (2023). Cluster analysis and visualization using principal component analysis. https://www.cfholbert.com/blog/cluster-pca/

[3] Gupta, P. (2022). Random forest algorithm in machine learning. GeeksforGeeks. https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/

[4] Khan, A. (2021). Unsupervised learning: K-means clustering. Towards Data Science. https://towardsdatascience.com/unsupervised-learning-k-means-clustering-27416b95af27/

[5] Kumar, R. (2022). Naive Bayes classifiers. GeeksforGeeks. https://www.geeksforgeeks.org/naive-bayes-classifiers/

[6] Singh, A. (2022). Decision tree algorithms in machine learning. GeeksforGeeks. https://www.geeksforgeeks.org/decision-tree/

VIII. Supplementary Material

Latex files:
https://www.overleaf.com/read/rbwmqccfzrqn#83e10e
Source code:
https://github.com/ardo1488/DataMiningMedicalDataset.git