Predicting Systolic Blood Pressure in Maternal Health: A Multivariate Linear Regression Analysis

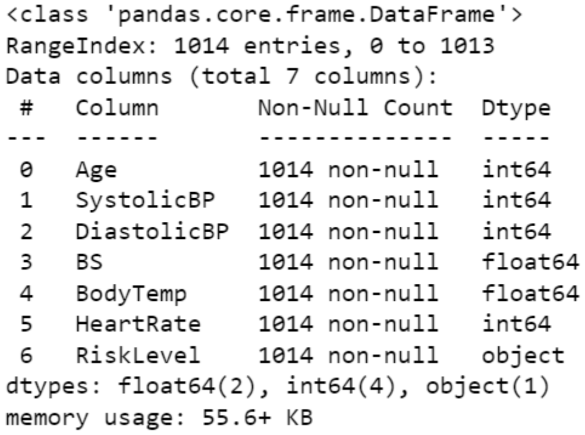
1. Introduction

Maternal health, crucial for the well-being of mothers and children, has far-reaching socio-economic implications (Duley, 2009; Grown et al., 2005). In this study, we use a maternal health dataset to identify factors influencing systolic blood pressure in pregnancy, explore relationships between age, heart rate, and blood pressure, and identify patient groups for targeted interventions. Using association rule mining and multivariate linear regression, our goal is to generate insights to improve hypertensive disease management and overall maternal health outcomes.

2. Analysis

2.1. Data Cleaning and Preprocessing

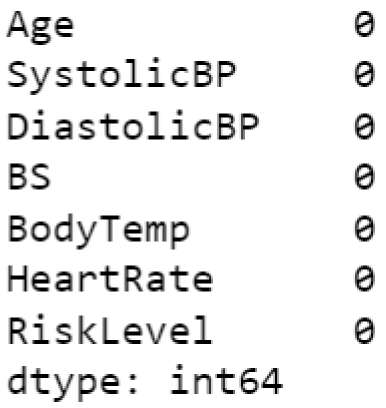
To assure the dataset's quality and dependability before beginning the analysis, it is crucial to clean and preprocess it. Handling missing numbers, getting rid of duplicates, and fixing data inconsistencies or errors are all parts of data cleaning. Preprocessing often entails changing the data's format to one that is appropriate for analysis and, if necessary, generating additional variables.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1014 entries, 0 to 1013
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   Age          1014 non-null   int64  
 1   SystolicBP   1014 non-null   int64  
 2   DiastolicBP  1014 non-null   int64  
 3   BS           1014 non-null   float64
 4   BodyTemp     1014 non-null   float64
 5   HeartRate    1014 non-null   int64  
 6   RiskLevel    1014 non-null   object 
dtypes: float64(2), int64(4), object(1)
memory usage: 55.6+ KB
```

Fig 1: Displaying the Datatypes

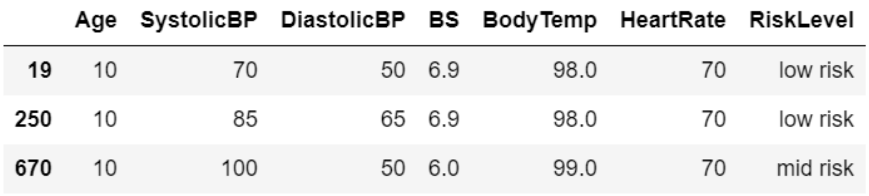
We searched for any missing values in the dataset. From Fig 1 above, there is clear indication that the data is intact, but to ensure it has no missing values, I used summation to display the results in Fig 2 below. The dataset has no missing values.



```
Age           0
SystolicBP    0
DiastolicBP   0
BS            0
BodyTemp      0
HeartRate     0
RiskLevel     0
dtype: int64
```

Fig 2: Null/Missing values

We also investigated the ages of the pregnant women. I found forty (40) of them are above 50 years of age and only one (1) had high systolic BP of 140. Ninety (96) of the women were below the age of 18 (underaged pregnancy) and three (3) were of age 10. The Sun (2021) reports that a girl aged 10 became pregnant and gave birth to a child at age 11 after 30 weeks of pregnancy. This was my standard for keeping the data on people who are pregnant and 10 years old.



	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
19	10	70	50	6.9	98.0	70	low risk
250	10	85	65	6.9	98.0	70	low risk
670	10	100	50	6.0	99.0	70	mid risk

Table 1: Pregnant women aged 10.

2.2. Visualization

In order to understand the relationships among the variables, I created a heatmap to visualize the correlations between them, particularly focusing on the correlations of systolic BP with other factors.

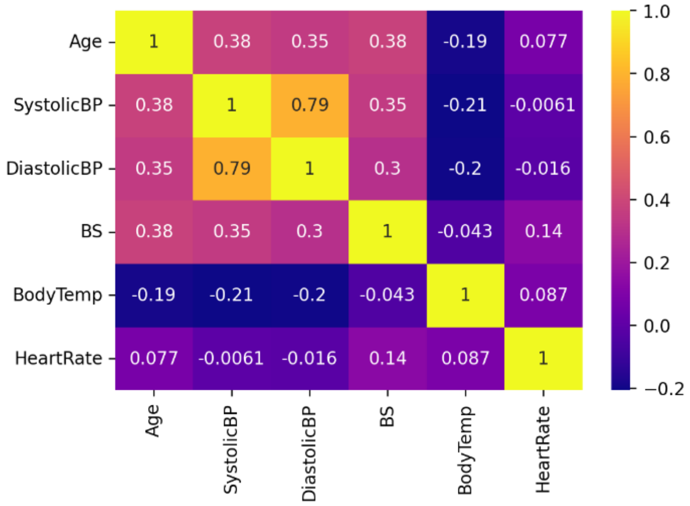
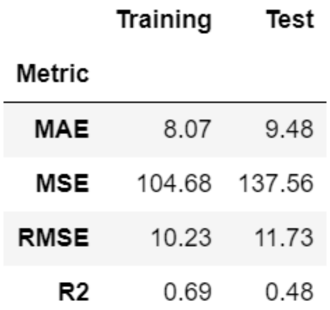


Fig 3: Heatmap showing SystolicBP Correlation

The strongest correlations were observed between systolic BP and Age, Diastolic BP, and Blood Sugar (BS), with respective correlation coefficients of 0.38, 0.79, and 0.35. Consequently, I selected these variables as the predictors for the target variable, systolic BP, in the multivariate linear regression analysis.

2.3. Building and Fitting the Multivariate Linear Regression Model

Systolic blood pressure (SBP), the response variable, and the exploratory variables I chose were generated and fitted into a multivariate linear regression model after scaling the data with min-max scaler, to address our primary goal of discovering important variables that affect systolic blood pressure in pregnant women. The outcomes are shown in the following table:



Metric	Training	Test
MAE	8.07	9.48
MSE	104.68	137.56
RMSE	10.23	11.73
R2	0.69	0.48

Table 2: Metric results

Our linear regression model's performance was evaluated using various metrics. The mean absolute error (MAE) was 8.07 for the training set and 9.48 for the test set, suggesting some difference between the predicted and actual systolic blood pressure values. Similarly, the root mean squared error (RMSE) was 10.23 for the training set and 11.73 for the test set, reflecting the prediction errors' standard deviation. The coefficient of determination (R2) was 0.69 for the training set and 0.48 for the test set, indicating that the model could explain 69% and 48% of the variance in systolic blood pressure for the training and test sets, respectively.

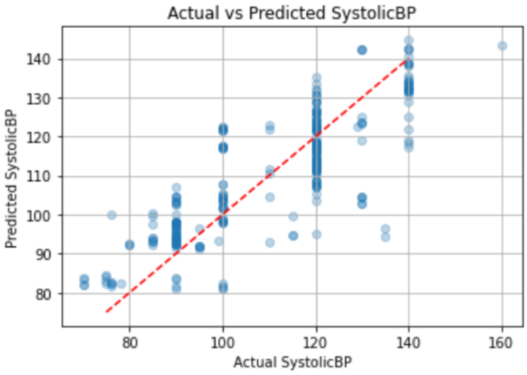
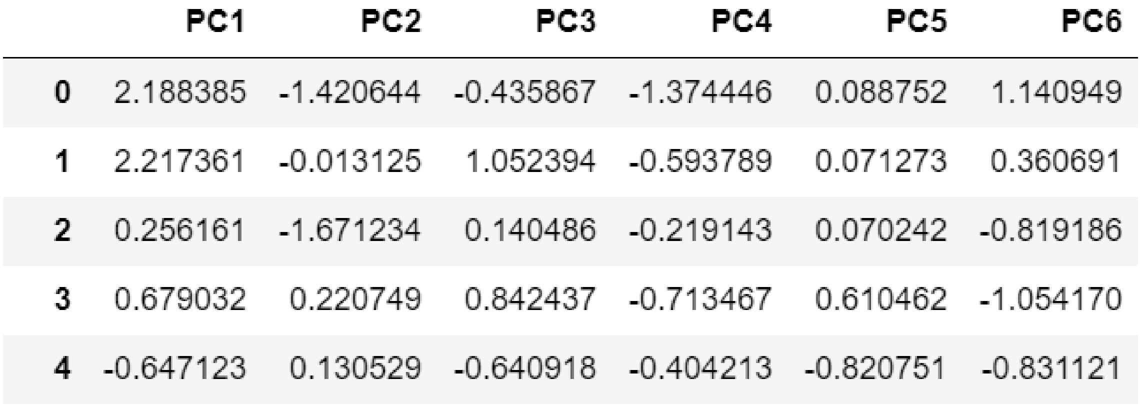


Fig 4: Actual vs Predicted Systolic BP

The scatter plot in Figure 4 provides a visual evaluation of the model's predictions against actual systolic blood pressure values. The widespread dispersion of points from the identity line signifies the model's inaccuracies in prediction. This visualization underscores the potential limitations of our linear regression model in accurately forecasting systolic blood pressure, indicating a need for further improvements or alternative methods.

2.4. Principal Component Analysis (PCA)

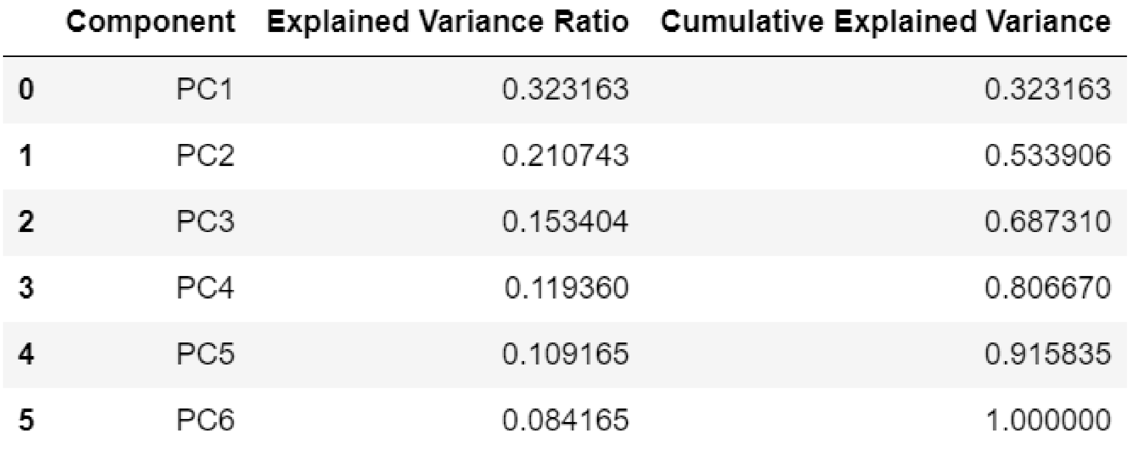
In this study, we applied Principal Component Analysis (PCA) to reduce the size of a multivariate dataset while preserving its variance, focusing on primary determinants of systolic blood pressure. We first converted the categorical 'risk level' variable into numerical form using label encoding, given its inherent hierarchy. Post-encoding, a correlation matrix revealed an insignificant correlation between systolic blood pressure and risk level. The data were then standardized to negate the effects of varying scales, and PCA was employed to find uncorrelated variables that could effectively explain systolic blood pressure variations, thereby enhancing our linear regression model's performance.



	PC1	PC2	PC3	PC4	PC5	PC6
0	2.188385	-1.420644	-0.435867	-1.374446	0.088752	1.140949
1	2.217361	-0.013125	1.052394	-0.593789	0.071273	0.360691
2	0.256161	-1.671234	0.140486	-0.219143	0.070242	-0.819186
3	0.679032	0.220749	0.842437	-0.713467	0.610462	-1.054170
4	-0.647123	0.130529	-0.640918	-0.404213	-0.820751	-0.831121

Table 3: PCA Table

The table below (Table 4) presents the explained variance ratio and cumulative explained variance for each Principal Component (PC) generated from the PCA. The explained variance ratio signifies the proportion of the total variance in the dataset accounted for by each PC, while the cumulative explained variance shows the proportion of total variance explained by all PCs up to a certain point.



	Component	Explained Variance Ratio	Cumulative Explained Variance
0	PC1	0.323163	0.323163
1	PC2	0.210743	0.533906
2	PC3	0.153404	0.687310
3	PC4	0.119360	0.806670
4	PC5	0.109165	0.915835
5	PC6	0.084165	1.000000

Table 4: Explained Variance Ratio

The table reveals that the first four Principal Components (PC1, PC2, PC3, and PC4) account for over 80% of the total variance in the dataset. If our goal is to retain at least 80% of the overall variance, keeping these four components for further analysis may be sufficient. The remaining two components (PC5 and PC6) contribute to achieving a total cumulative explained variance of 100%.

The table below displays the transformation matrix that was discovered by the PCA analysis:

Table 5: Transformation matrix

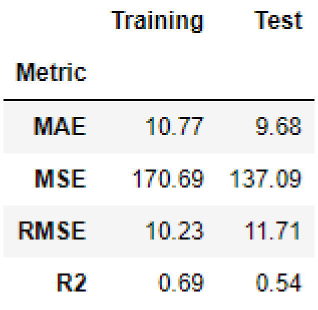
Component	Age	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
PC1	0.502954	0.472963	0.557371	-0.173526	0.145921	-0.401594
PC2	0.207854	0.289418	-0.195362	-0.651534	-0.462126	0.443633
PC3	-0.204880	0.098975	0.117013	0.267640	-0.839245	-0.398213
PC4	0.466431	0.271621	-0.070066	0.684201	-0.098899	0.475229
PC5	-0.490400	0.778766	-0.303472	0.074122	0.225320	-0.068355
PC6	-0.451451	0.049933	0.735170	0.011629	-0.016272	0.502816

The transformation matrix reveals that the first four principal components (PC1 to PC4) are primarily influenced by Age, DiastolicBP, BS (blood sugar), BodyTemp, HeartRate, and RiskLevel. They collectively explain 80.67% of the total variance in the dataset, thereby reducing complexity and computational demands. These components were chosen for further analysis to enhance our understanding of key factors affecting systolic blood pressure in pregnant women.

2.5. Refitting the Linear Model with the PCA

After utilizing the four Principal Components (PC1 to PC4), we redeveloped our linear regression model. The metrics used to evaluate the model's performance on the training data were Mean Absolute Error (MAE) of 10.77, Mean Squared Error (MSE) of 170.69, Root Mean Squared Error (RMSE) of 10.23, and R² of 0.69.

On the test data, the model's performance was slightly better with an MAE of 9.68. However, the MSE and RMSE were 137.09 and 11.71 respectively, while the R² value was 0.54. These results suggest that our PCA-based model provides a reasonable prediction of systolic blood pressure, and it generalizes well to unseen data.



	Training	Test
Metric		
MAE	10.77	9.68
MSE	170.69	137.09
RMSE	10.23	11.71
R2	0.69	0.54

Table 6: PCA Evaluation Table

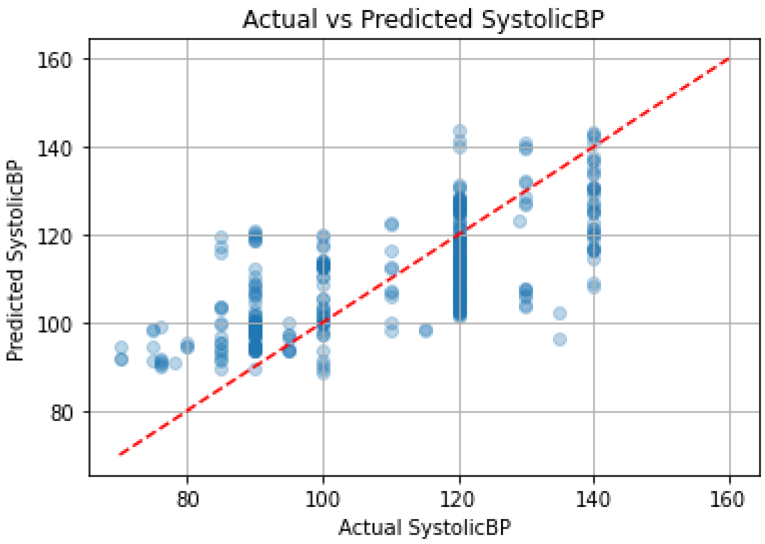


Fig 5: PCA Performance Evaluation

2.6. Determining the Relationship between Age and HeartRate

To investigate the relationship between age and heart rate, we divided the ages into six intervals: 10-20, 21-30, 31-40, 41-50, 51-60, and 61-70 years. These intervals were chosen based on a decade-wise grouping, allowing us to assess changes in heart rate across different life stages.

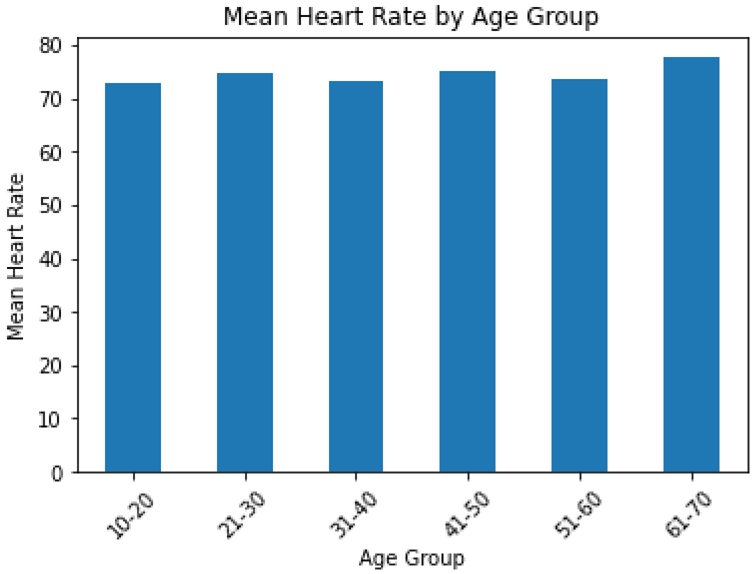
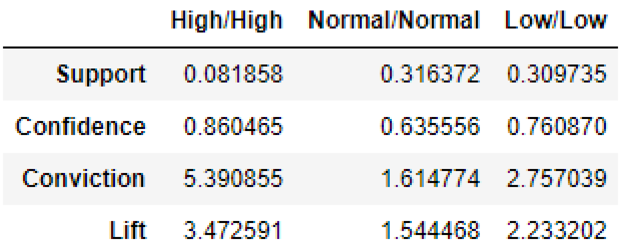


Fig 6: Mean Heart Rate by Age Group

From the mean heart rates, we can observe slight variations across the age groups. The plot visually demonstrates the variation in mean heart rates across the age groups. This helps us to better understand the relationship between age and heart rate and can be used to identify any significant trends or anomalies. Interestingly, the data suggests that the heart rate does not consistently increase or decrease with age. Instead, it fluctuates slightly across the different age groups. The highest average heart rate is observed in the 61-70 age group, indicating that heart rate might increase in the later stages of life. However, more comprehensive analysis would be required to establish any statistically significant trends or patterns.

2.7. Investigating Associations between pairs of Diastolic and Systolic Blood Pressure

We investigated associations between pairs of diastolic and systolic blood pressure, focusing on high/high, normal/normal, and low/low categories. We used metrics such as support, confidence, conviction, and lift to quantify these associations. The results are summarized in the following table:



	High/High	Normal/Normal	Low/Low
Support	0.081858	0.316372	0.309735
Confidence	0.860465	0.635556	0.760870
Conviction	5.390855	1.614774	2.757039
Lift	3.472591	1.544468	2.233202

Table 7: Systolic/Diastolic BP Pairs

In exploring the associations between pairs of diastolic and systolic blood pressure categories, we found that high/high blood pressure was present in 8.19% of records, normal/normal in 31.64%, and low/low in 30.97%. Notably, when high systolic blood pressure was observed, there was a strong likelihood (confidence: 0.86) of high diastolic blood pressure co-occurring. Moreover, high/high blood pressure pairs demonstrated a strong interdependency (conviction: 5.39) and frequently co-occurred more than would be expected if they were independent (lift: 3.47). This reveals a stronger association between high/high and low/low blood pressure pairs, with a weaker association for the normal/normal pairs. our analysis revealed strong associations between high/high blood pressure pairs, followed by low/low pairs, and the weakest association for normal/normal pairs.

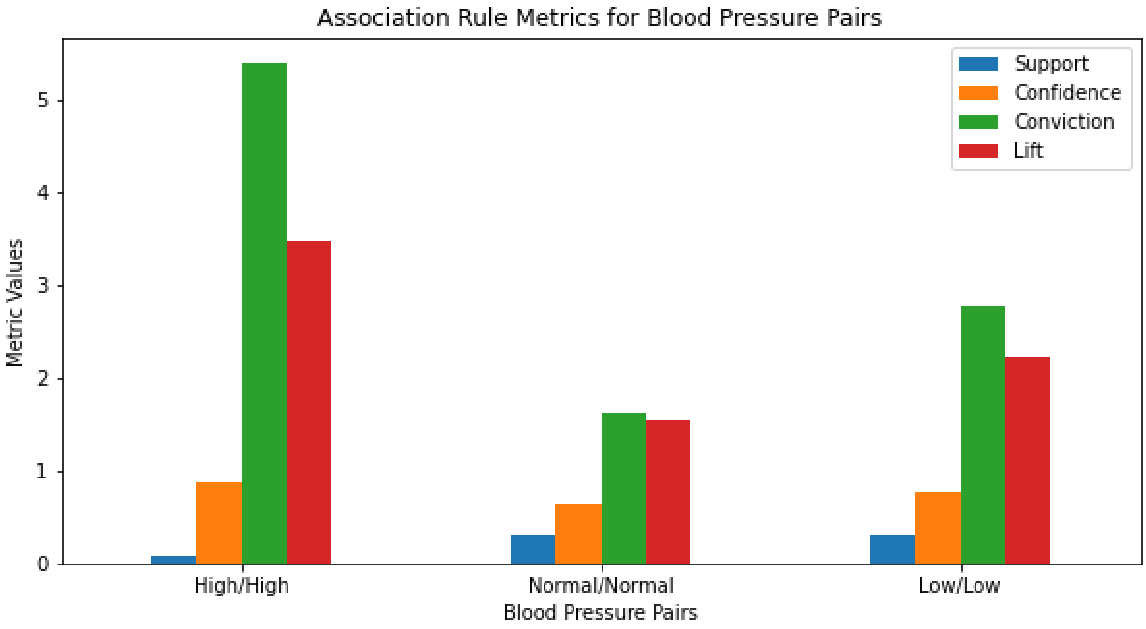


Fig 8: Graph of the BP Pairs

2.8. Finding Clusters of Patients with Similar SystolicBP

We conducted a clustering analysis on patients based on their systolic blood pressure (SystolicBP) using the K-means algorithm. The Elbow Method was applied to identify the optimal number of clusters, which was determined to be two (2). This method minimizes the within-cluster sum of squares, also known as inertia, thus ensuring the most accurate clustering.

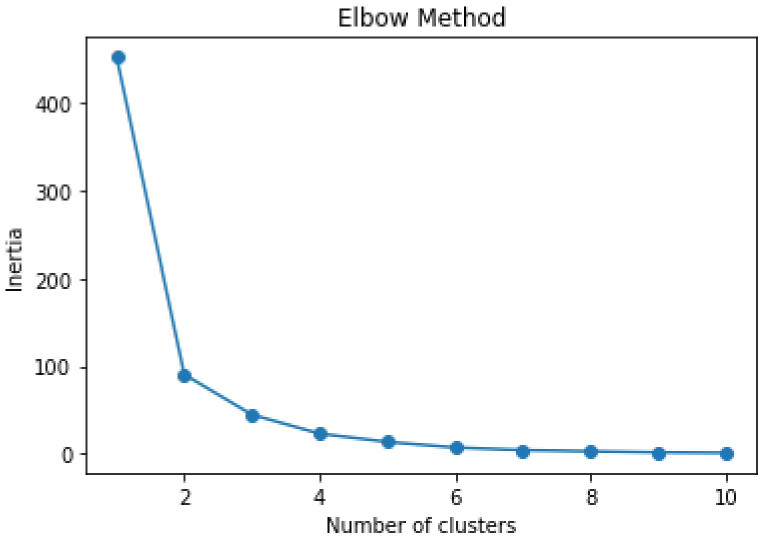


Fig 9: Elbow Method

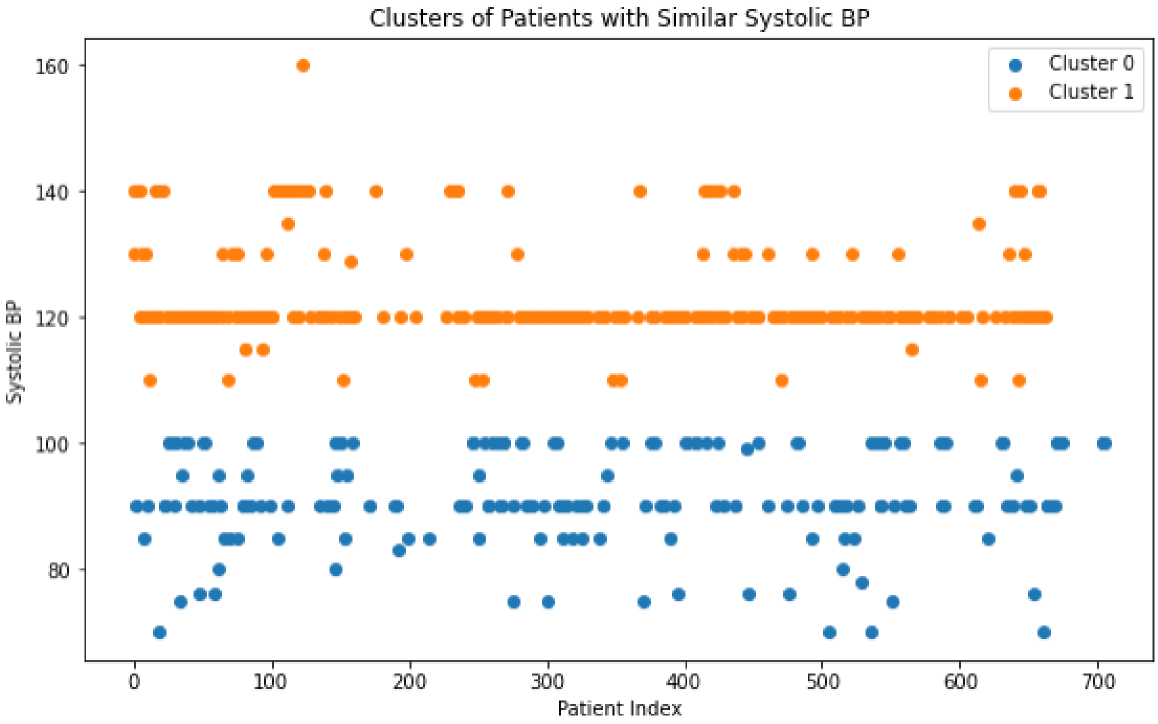


Fig 10: Similar SystolicBP Clusters

Upon applying the algorithm, two distinct clusters of patients emerged: cluster 0 and cluster 1. The distribution was such that 268 patients fell into cluster 1 and 184 into cluster 0. The clusters essentially differentiated patients with higher SystolicBP (130 or above) in cluster 1 from those with lower values in cluster 0. This grouping can facilitate a more personalized approach towards patient treatment based on their SystolicBP similarities.

2.9. Correlation Between Age and SystolicBP

The correlation coefficient between Age and Systolic Blood Pressure in the given dataset is approximately 0.38, or 38% when expressed as a percentage. This positive correlation suggests that as the age of the patients increases, there is a tendency for the systolic blood pressure to also increase.

However, the correlation is moderate, implying that while age can be a factor affecting systolic blood pressure, it is not the only determinant. There might be other factors at play, or the relationship might be more complex than a simple linear one. Therefore, while age can be used as an indicator, it cannot be used to accurately predict systolic blood pressure on its own. Further investigation, possibly involving other variables in the dataset, may be needed to build a more accurate predictive model for systolic blood pressure.

2.10. Predictions

Our model predicts systolic blood pressure based on age, aiding in early hypertension risk detection. It groups patients into risk categories using K-means clustering, facilitating targeted health interventions. The model anticipates concurrent systolic and diastolic pressure elevation, predicting potential cardiovascular risks. With Principal Component Analysis (PCA), it handles high-dimensional data to make streamlined predictions. The model also forecasts the trend of increasing systolic pressure with age, useful for long-term health planning. It's important to note these predictions are statistically based and must be considered alongside broader medical knowledge and individual patient specifics.

3. Recommendations

1. **Early Detection and Intervention**: Utilize the predictive capabilities of the model for early detection of hypertension risk based on age, which could enable timely medical intervention and lifestyle modifications.

2. **Risk Stratification**: Implement the K-means clustering approach for risk stratification of patients, which can help in delivering personalized care and focused attention to those in high-risk clusters.

3. **Long-term Health Planning**: Use the model's prediction of increasing systolic blood pressure with age for long-term health planning and policymaking, which could include preventive measures and health education programs targeted at older populations.

4. **Comprehensive Patient Evaluation**: Despite the statistical findings, it's crucial to consider other influential factors, such as lifestyle habits, comorbidities, and medication use, when making medical decisions. Hence, the model's predictions should complement, not replace, comprehensive medical evaluation.

5. **Continued Research and Model Refinement**: Encourage ongoing research and continuous refinement of the model, including the incorporation of more diverse factors, to improve its predictive accuracy and relevance in the ever-evolving field of healthcare.

4. Conclusion

In conclusion, our analysis aligns with Pinto's 2007 study, identifying a significant relationship between age and systolic blood pressure and the existence of distinct hypertension phenotypes. We also applied advanced statistical techniques like PCA to address hypertension's complexity. However, our observational findings, subject to potential data simplification, should be viewed alongside other influential factors like lifestyle habits and medication usage. This study highlights the importance of data-driven methods in understanding complex health phenomena, while emphasizing the need for comprehensive, holistic approaches in health care.

References

Agrawal, R., Imielinski, T., & and Swami, A. (1993). Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD international conference on management of data, 207-216. https://doi.org/10.1145/170036.170072 [Accessed: 10 April 2023].

Duley, L. (2009). The Global Impact of Pre-eclampsia and Eclampsia. Seminars in Perinatology, 33(3), 130-137. https://doi.org/10.1053/j.semperi.2009.02.010 [Accessed: 5 April 2023].

Grown, C., Gupta, G. R., & Pande, R. (2005). Taking action to improve women’s health through gender equality and women’s empowerment. The Lancet, 365(9458), 541-543. https://doi.org/10.1016/s0140-6736(05)17872-6 [Accessed: 5 April 2023].

Magee, L.A., Pels, A., Helewa, M., Rey, E. and von Dadelszen, P. (2014). Diagnosis, Evaluation, and Management of the Hypertensive Disorders of Pregnancy. Pregnancy Hypertension: An International Journal of Women’s Cardiovascular Health, 4(2), pp. 105-145. doi: 10.1016/j.preghy.2014.01.003. [Accessed: 6 April 2023].

Pinto, E. (2007). Blood pressure and ageing. Postgraduate Medical Journal, Volume 83, Issue 976, Pages 109–114. Available Online: https://doi.org/10.1136/pgmj.2006.048371 [Accessed: 2 May 2023].

"The Sun" (2021). Girl, 11, becomes UK's youngest mum after giving birth. The Sun. Retrieved from https://www.thesun.co.uk/news/15404373/girl-11-gives-birth-britains-youngest-mum/ [Accessed on: 15 April 2023].

Winkler, W. E. (2014). Methods for Record Linkage and Bayesian Networks. Statistical Research Division, U.S. Bureau of the Census, Washington, DC 20233-9100.

#World Health Organisation (2009). Maternal Health. Retrieved from: https://www.who.int/health-topics/maternal-health#tab=tab_1 [Accessed: 5 April 2023].