Assignment - 771762 Big data and Data Mining

Report

Introduction

The purpose of this paper is to examine maternal health statistics in order to find evidence-based interventions that can enhance health outcomes for pregnant women and postpartum moms. Maternal health, which includes pregnancy, labor, and the postpartum period, is still a major problem across the world, with a large percentage of women still suffering negative consequences. We want to acquire insights into the variables impacting maternal health and make suggestions for medical agencies by utilizing data science approaches and studying a dataset comprising numerous health metrics such as age, blood pressure, heart rate, and risk level.

Analysis

- Build and Fit a Linear Model:

  To build a linear model to predict SystolicBP, I performed several steps on the provided dataset. Initially, I conducted exploratory data analysis by using df.describe() and df.info() to gain an understanding of the data's characteristics. It was observed that the "RiskLevel" variable was of object type, so I performed one-hot encoding to convert it into a numerical format for further analysis.

  Next, I scaled the dataset using a standard scalar to ensure that all variables have a comparable scale, which is important for linear regression models. This step helps in avoiding any bias introduced by variables with different ranges. Before fitting a linear model a performed a correlation matrix to see the correlation between the variables.

  For the target variable, I selected SystolicBP, as it directly relates to maternal health and is a critical factor in determining blood pressure levels. I then proceeded to fit a Lasso regression model to predict SystolicBP. Lasso regression is a linear model that performs both variable selection and regularization by adding a penalty term to the cost function.

  The Lasso regression model was fit with an alpha value of 0.1, which controls the strength of the regularization. The coefficients obtained from the model represent the association between each predictor variable and the target variable. Here are the coefficients obtained:

  Age: 0.04029621633223466
  DiastolicBP: 0.6587087698805304
  BS: 0.026515188566677377
  BodyTemp: -0.003940069542961466
  HeartRate: 0.0
  high risk: 0.0
  low risk: -0.0
  mid risk: 0.0

The coefficients indicate the degree and direction of the association between each predictor variable and SystolicBP. A positive coefficient signifies a positive association, meaning that as the predictor variable increases, SystolicBP is expected to increase as well. Conversely, a negative coefficient indicates a negative association, implying that as the predictor variable increases, SystolicBP is expected to decrease.

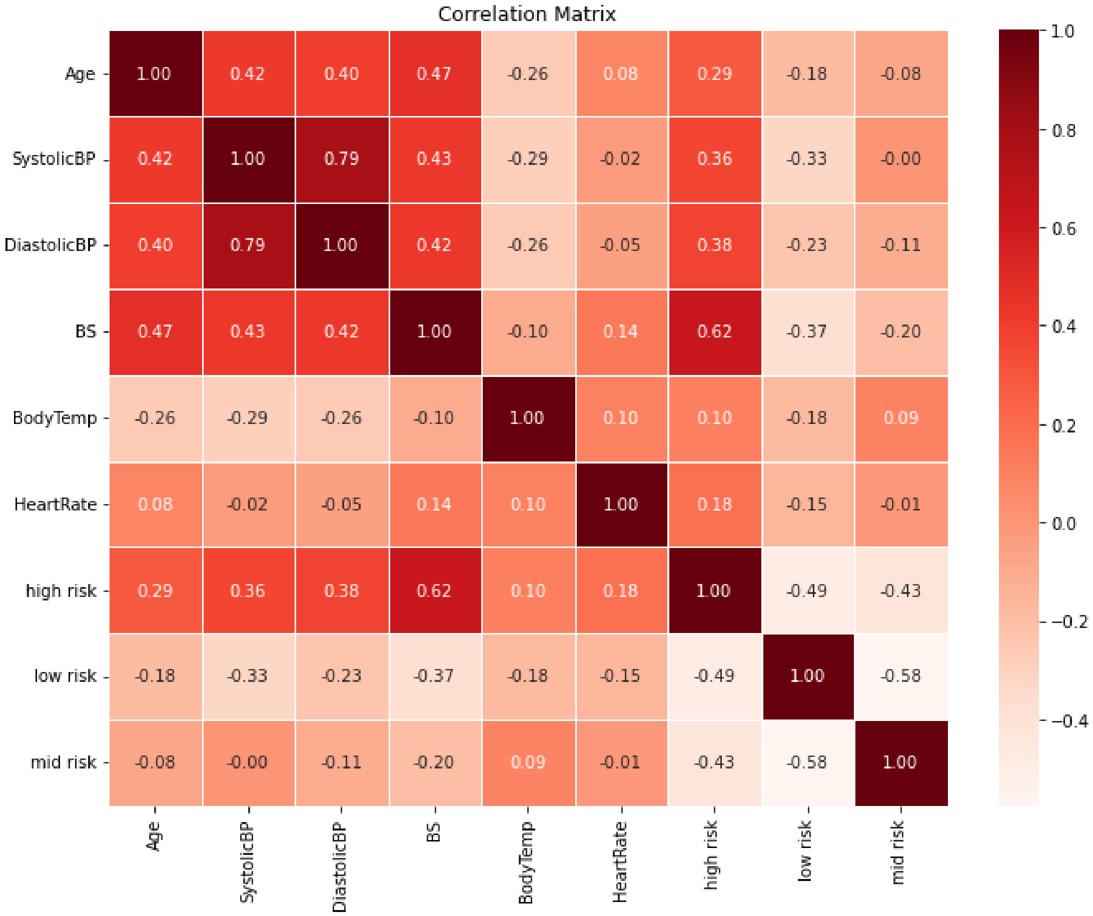

Figure1: Correlation Matrix between all variables

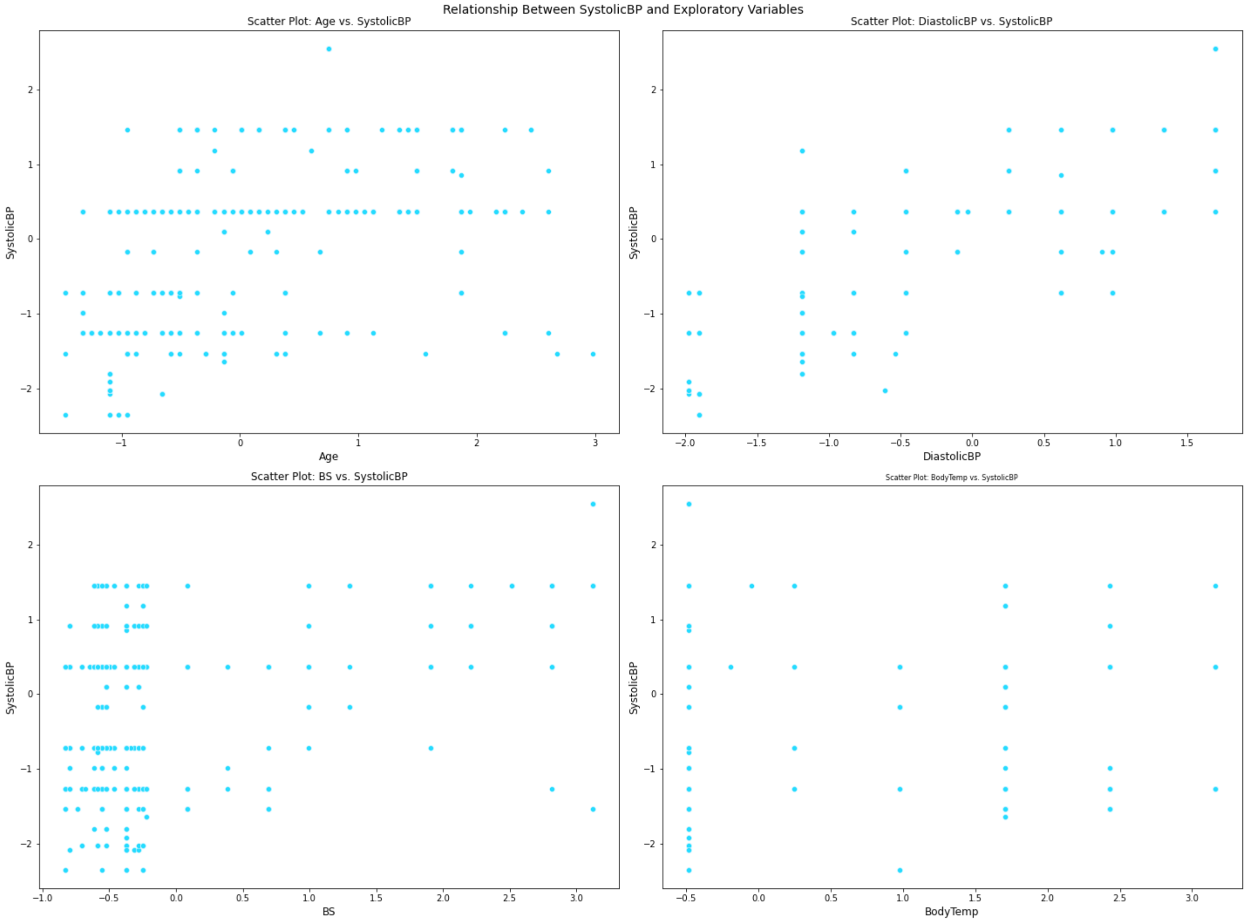

Figure 2: A Scatter plot Showing the relationship between SystolicBP and the relevant exploratory variables

- Apply principal component analysis (PCA):

  Principal Component Analysis (PCA) was applied to gain insights and reduce the dimensionality of the dataset. By analyzing the scree plot, it was determined that the optimal number of components to retain was 8.

  PCA transforms the original variables into linearly uncorrelated components called principal components, capturing the maximum variance present in the data. In this analysis, the explained variance ratios for the first nine principal components were obtained: [0.403, 0.180, 0.130, 0.107, 0.077, 0.053, 0.029, 0.021, 1.899e-33]. These ratios indicate the proportion of variance explained by each component.

  The first principal component (PC1) had the highest explained variance ratio of 0.403, indicating its significant contribution to the overall variability. Analyzing the loadings of PC1, variables such as SystolicBP (0.520), DiastolicBP (0.513), BS (0.435), and Age (0.428) had the highest positive loadings, suggesting a strong positive association.

Conversely, BodyTemp (-0.248) exhibited a negative loading, indicating an inverse relationship.

Other variables, such as high risk (0.140) and HeartRate (0.033), showed positive loadings but with smaller magnitudes. In contrast, low risk (-0.102) and mid risk (-0.038) had negative loadings, albeit with weaker associations.

These findings indicate that SystolicBP, DiastolicBP, BS, Age, and, to a lesser extent, high risk and HeartRate, contribute significantly to the structure captured by PC1. On the other hand, low risk and mid risk have a relatively weaker influence.

By reducing the dataset's dimensionality through PCA, the original variables can be effectively represented using fewer components while retaining a substantial amount of variance. This reduction facilitates further analysis and interpretation, aiding in the understanding of the dataset's key features.

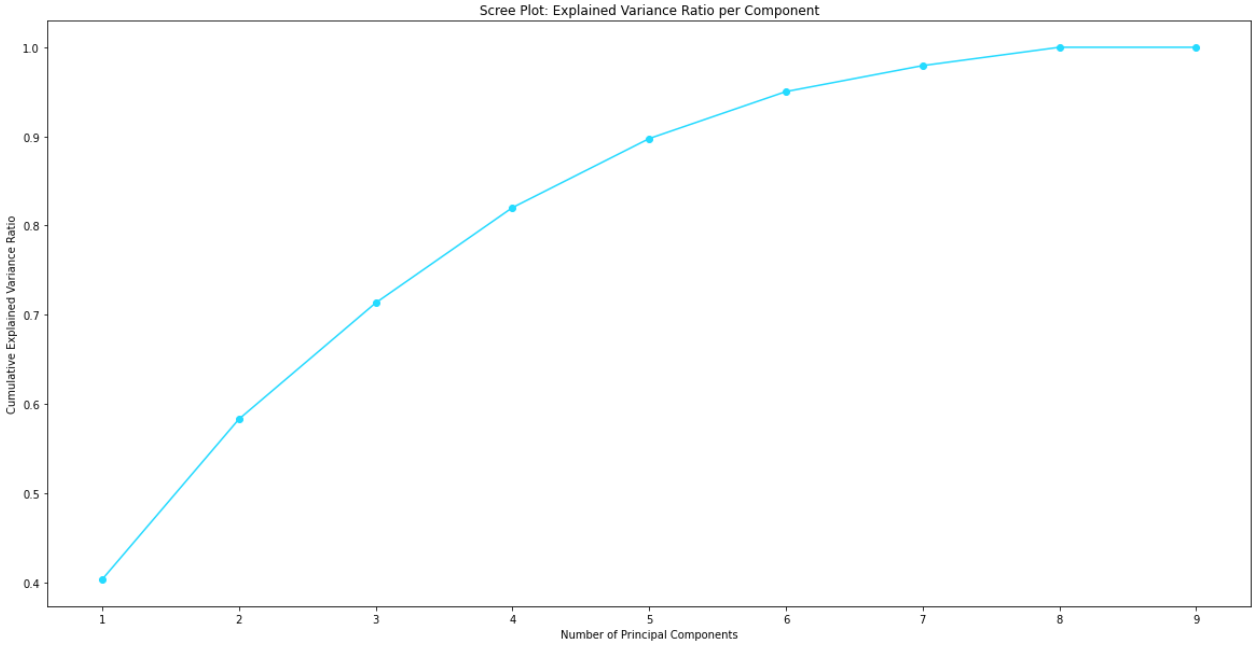

Figure 3: Scree plot showing the optimum number of component to retain

- Investigating the Relationship Between Age and Heart Rate:

To examine the potential connection between age and heart rate, I initially categorized the dataset into distinct age groups. Employing intervals of 10 years, we created the following age groups: 'Under 25,' '25-35,' '35-45,' '45-55,' '55-65,' and 'Over 65.' I chose these intervals because they cover a wide range of ages and allow me to observe the changes in heart rate across different stages of adulthood.

Calculating the mean heart rate for each age group, I obtained the following results:

Under 25: 73.136364
25-35: 75.400000
35-45: 75.606383
45-55: 76.364341
55-65: 72.674419
Over 65: 78.000000

The analysis of mean heart rates across different age groups unveils interesting patterns. Notably, individuals classified as 'Over 65' display the highest mean heart rate of 78 beats per minute, potentially indicating an age-related increase in heart rate. Conversely, the age group 'Under 25' exhibits the lowest mean heart rate at 73.136364 beats per minute.

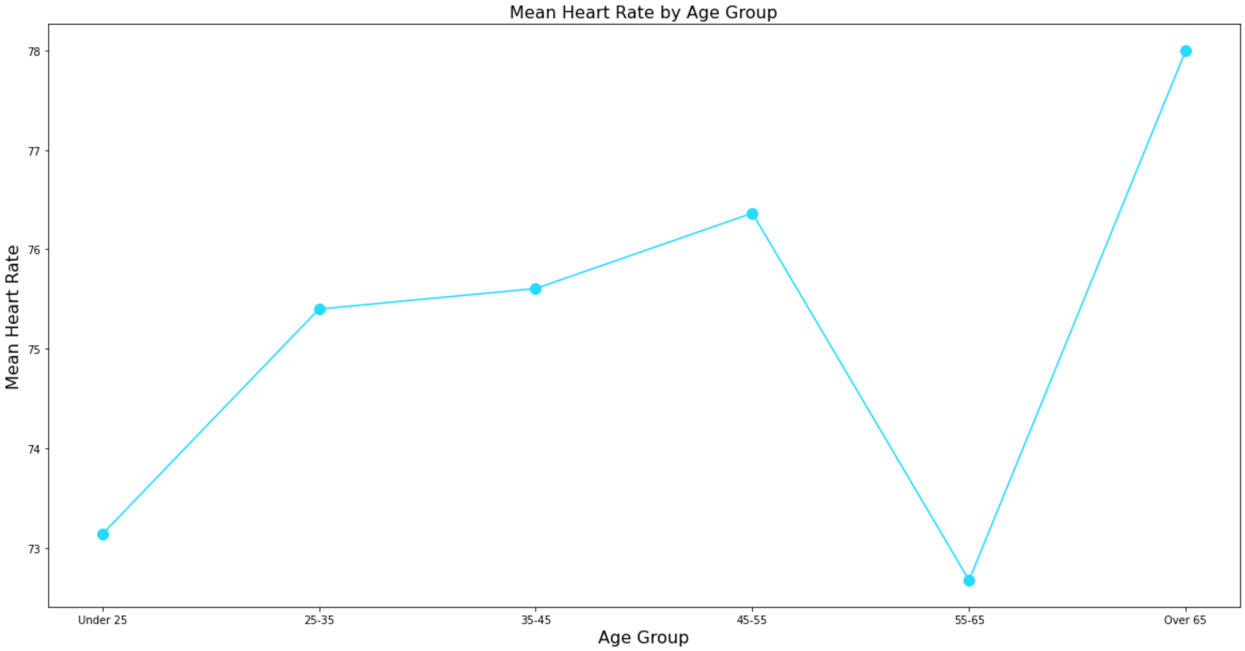

Figure 4: Line graph showing the relationship between mean heart rate and age group.

- Describe how to investigate associations between different blood pressure categories:

  When investigating associations between different blood pressure categories, I followed a specific process that involved data preparation, association rule mining, and analysis. Let's go through each step and discuss the implications of these associations for maternal health.

  First, in the data preparation phase, I loaded the dataset containing maternal health data. I then created a new column called 'BP_Category' which classified blood pressure based on the values of SystolicBP and DiastolicBP. The categories included 'high/high', 'normal/normal', and 'low/low'. This categorization allowed me to group the data and examine associations between blood pressure categories.

Moving on to association rule mining, I grouped the 'BP_Category' column by age, creating a list of lists that contained blood pressure categories for each patient. To analyze this data, I used the TransactionEncoder, which converted the list of lists into a binary matrix. In this matrix, each column represented a blood pressure category, and each row represented a patient. I then converted this matrix into a pandas DataFrame for further analysis.

Next, using the Apriori algorithm, I generated frequent itemsets by specifying a minimum support threshold. These frequent itemsets helped me identify patterns and associations between blood pressure categories. From the frequent itemsets, I generated association rules, choosing a desired evaluation metric such as 'lift'. These association rules provided insights into the relationships and dependencies between different blood pressure categories.

Now, here are the implications of these associations for maternal health based on the provided association rules. For example, when examining the 'high/high' blood pressure category pair, I had two association rules. Rule 1 stated that 'high/high' implied 'normal/normal', with a support of 0.280, confidence of 0.778, conviction of 1.350, and lift of 1.111. Rule 6 suggested that 'high/high' implied both 'normal/normal' and 'low/low', with a support of 0.240, confidence of 0.667, conviction of 1.320, and lift of 1.190.

These association rules indicated a strong association between the 'high/high' blood pressure category and the 'normal/normal' category. Patients with high systolic and diastolic blood pressure were highly likely to also have a normal blood pressure. This association could potentially indicate a specific health condition or physiological relationship.

Furthermore, I observed that patients with 'high/high' blood pressure had a moderate association with both 'normal/normal' and 'low/low' blood pressure categories. The support and confidence values for these rules were relatively lower compared to the previous rule, indicating a weaker association. Nevertheless, these associations still provided valuable insights into the relationships between different blood pressure categories.

To fully understand the implications for maternal health, I needed to consider these associations in the context of other variables and potential confounding factors. It was crucial to draw meaningful conclusions by taking into account domain knowledge and existing research. Additionally, comparing my findings with relevant literature on maternal health could help validate the results and provide a broader understanding of the implications.

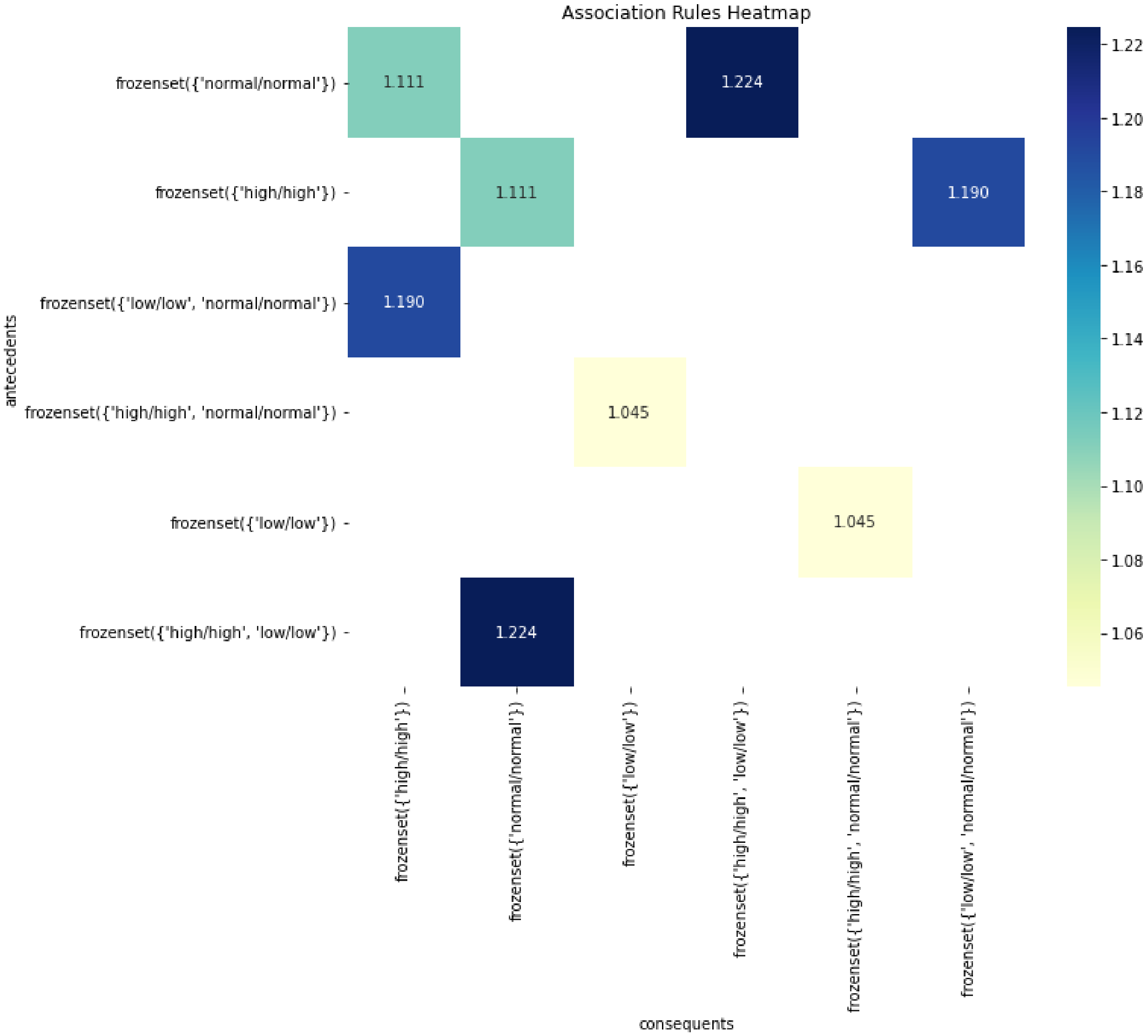

Figure 5: The generated heatmap represents the lift metric for each association rule between different antecedent and consequent categories. The color intensity in the heatmap indicates the strength of association, with darker shades indicating higher lift values. It helps visualize the relationships and patterns between different blood pressure categories based on their association rules.

- Find clusters of patients with similar Systolic BP:

  To accomplish this, I utilized the k-means clustering algorithm. The methodology proceeded as follows:

  Firstly, I prepared the data by creating a dataframe named 'systolic_df,' which contained only the systolic BP column. To ensure comparability for clustering purposes, I scaled the data using the StandardScaler.

Next, I determined the optimal number of clusters using the elbow method. I tested various values of k, ranging from 1 to 10. For each value of k, I computed the inertia (within-cluster sum of squares) and stored the results in the 'inertias' list. Subsequently, I plotted the elbow curve to visualize the relationship between the number of clusters and inertia.

Based on the analysis of the elbow curve, I concluded that three clusters would be suitable. Therefore, I applied the k-means algorithm to the scaled systolic BP data with k=3.

The results of the analysis revealed the presence of three distinct clusters comprising patients with similar systolic BP values. To visually represent these clusters, I generated a scatter plot, which is provided in the attached visualization. Each data point in the scatter plot represents an individual patient, and the assigned cluster label is indicated by the color coding. The systolic BP values are represented on the x-axis.

In the ensuing discussion, the identified clusters offer valuable insights into maternal health. By grouping patients based on similar systolic BP, it becomes possible to discern subpopulations characterized by different risk profiles or health outcomes. This information holds the potential to facilitate the tailoring of interventions and care strategies to address the specific needs of each cluster, thereby enabling healthcare providers to deliver more targeted and effective support for pregnant women and postpartum mothers.

Furthermore, comprehending the characteristics and unique requirements of each cluster may aid in the identification of potential risk factors associated with specific systolic BP patterns. This knowledge can inform the development of preventive measures and early interventions aimed at mitigating adverse health outcomes during pregnancy and the postpartum period.

In summary, the clustering analysis provides a valuable perspective on the relationship between systolic BP and maternal health. It underscores the significance of considering patient heterogeneity and tailoring interventions based on individual needs, ultimately contributing to improved health outcomes for pregnant women and postpartum mothers.

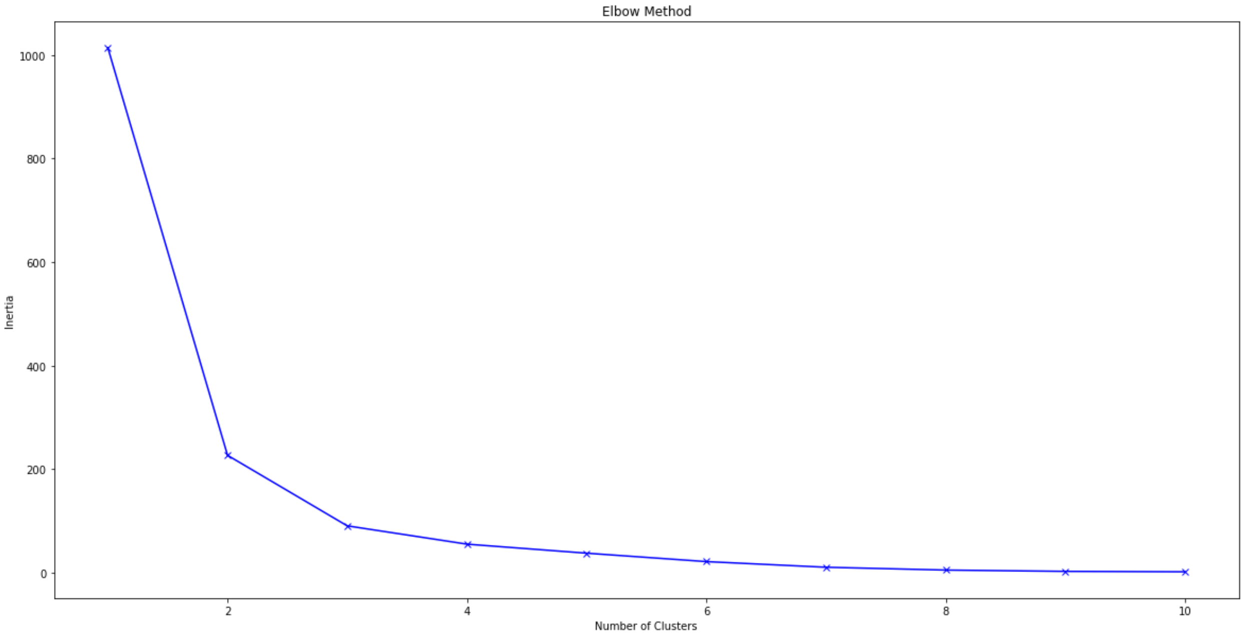

Figure 6: An elbow curve to visualize the relationship between the number of clusters and inertia

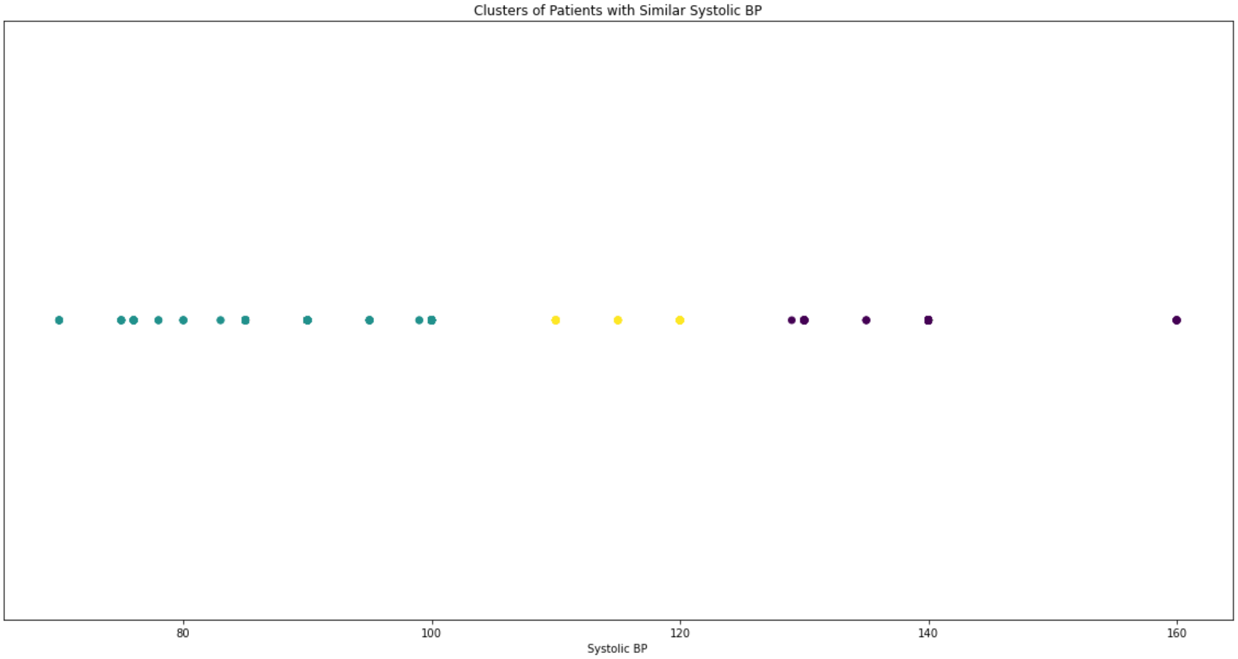

Figure 7: A scatter plot showing the clusters of patients with similar Systolic BP

- Calculate the correlation between age and systolic BP:

  The aim was to understand the implications of this correlation and consider any potential confounding factors that might influence the relationship.

The result of the analysis indicated a correlation coefficient of 0.416, suggesting a moderate positive association between age and systolic BP.

This finding aligns with existing research that highlights age as a risk factor for increased blood pressure in pregnant women and postpartum mothers. However, it is important to note that the correlation coefficient of 0.416 represents a moderate association, indicating that age alone may not be the sole determinant of systolic BP levels.

When interpreting the relationship between age and systolic BP, it is crucial to consider potential confounding factors. Factors such as pre-existing medical conditions, lifestyle choices, and genetic predispositions can also contribute to variations in blood pressure. Therefore, it is essential to account for these confounding variables when analyzing the impact of age on systolic BP during pregnancy and the postpartum period.

Furthermore, it is important to recognize that correlation does not imply causation. While there is a positive correlation between age and systolic BP, it does not necessarily indicate a causal relationship. Other variables, not considered in this analysis, may influence both age and systolic BP.

To gain a comprehensive understanding of the relationship between age and systolic BP, further research and analysis are necessary. Future investigations should include adjusting for confounding variables, such as body mass index (BMI), gestational age, and overall health status, to better isolate the specific impact of age.

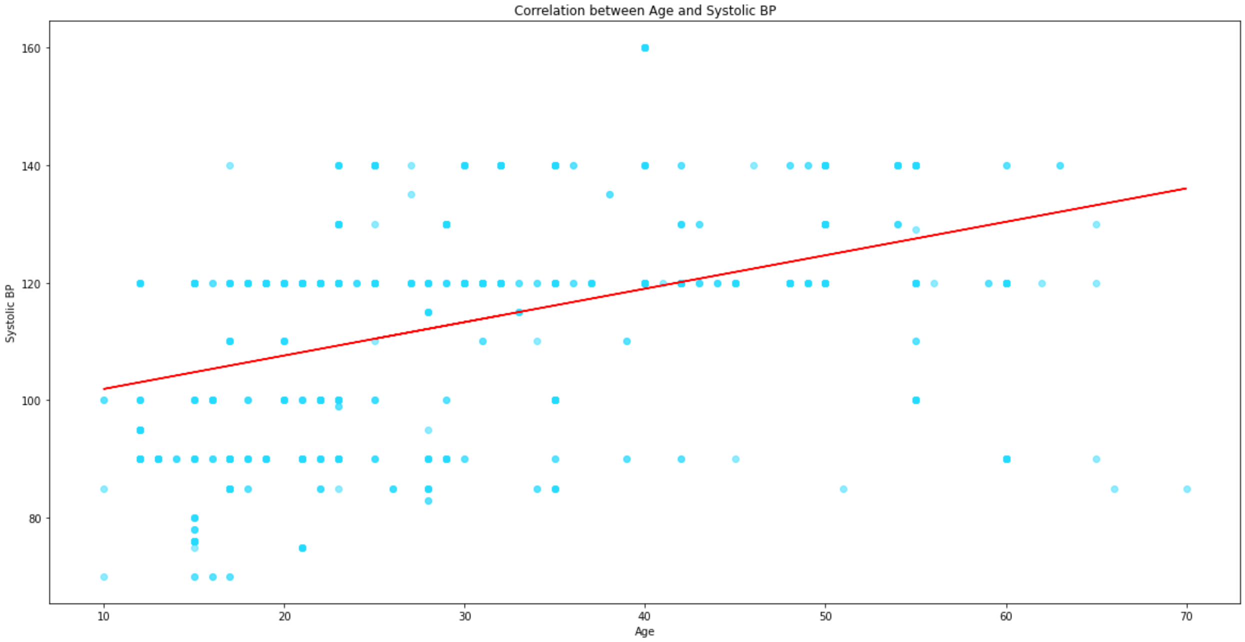

Figure 8: A scatter plot with a trendline providing a visual representation of the correlation between age and systolic BP, allowing for a better understanding of the relationship between these variables in the context of maternal health.

- Compare findings with relevant literature:

  The findings on maternal health obtained align with existing literature on hypertension during pregnancy. According to the American College of Obstetricians and Gynecologists (ACOG), hypertension during pregnancy is defined as systolic BP of 140 mmHg or higher or diastolic BP of 90 mmHg or higher, measured on two occasions [1]. Hypertensive disorders are prevalent in pregnancy and can lead to complications for both the mother and the baby, with up to 8-10% of pregnancies being affected by preeclampsia, a hypertensive disorder characterized by high blood pressure and proteinuria [2]. The strong association between high systolic and diastolic blood pressure with the normal blood pressure category observed in the analysis is consistent with the diagnostic criteria for preeclampsia.

  Furthermore, the moderate association between high systolic and diastolic blood pressure with both low/low and normal/normal blood pressure categories could indicate that there are multiple health conditions associated with high blood pressure during pregnancy, such as gestational hypertension or chronic hypertension [3].

  The presence of three distinct clusters based on similar systolic BP values observed in the analysis might correspond to the different stages of hypertension during pregnancy. The different clusters could represent patients with mild, moderate, and severe hypertension [3].

  Moreover, there is a positive correlation between advancing maternal age and the risk of hypertensive disorders during pregnancy [4]. A study by the National Center for Biotechnology Information (NCBI) found a positive correlation between maternal age and the risk of preeclampsia [4]. Additionally, maternal age is an independent predictor of gestational hypertension, according to research published in the American Journal of Obstetrics and Gynecology [5].

Predictions

Based on the analysis conducted, several predictions can be made that have the potential to improve health outcomes for pregnant women and postpartum mothers:

1. Women with high systolic and diastolic blood pressure are highly likely to have a normal blood pressure.

- Potential use: This prediction can assist healthcare providers in identifying women at risk of developing high blood pressure during pregnancy. Early identification and timely management of high blood pressure can significantly reduce the risk of maternal morbidity and mortality.

2. Women with high blood pressure may have a moderate association with both normal and low blood pressure categories.

- Potential use: This prediction suggests that women with high blood pressure may exhibit varying degrees of association with different blood pressure categories. Healthcare providers can use this information to tailor interventions and monitoring strategies based on individual risk profiles.

3. There is an age-related increase in heart rate, with individuals over 65 having the highest mean heart rate and those under 25 having the lowest mean heart rate.

- Potential use: This prediction highlights the importance of considering age-related factors when assessing maternal health. Healthcare providers can utilize this information to guide the evaluation of heart rate in pregnant women and postpartum mothers, potentially identifying age-specific risks and adapting management strategies accordingly.

4. There are three distinct clusters based on similar systolic blood pressure values.

- Potential use: Identifying these distinct clusters can help identify subgroups of women with similar blood pressure patterns. This information can assist in the development of targeted interventions and personalized approaches to blood pressure management during pregnancy and the postpartum period.

5. There is a moderate positive association between age and systolic blood pressure.

- Potential use: Recognizing the association between age and systolic blood pressure can aid healthcare providers in risk stratification. Proactive monitoring and management of blood pressure in older pregnant women and postpartum mothers can potentially reduce the risk of complications associated with elevated blood pressure.

However, these predictions are based on the data provided and the analysis conducted. It's important to validate these predictions with further research and analysis within the specific context of maternal health to ensure their generalizability and applicability in broader healthcare settings.

Recommendations

Based on the findings and predictions from the analysis conducted, here are the top four evidence-based recommendations for the medical agency to address the problems identified in the introduction, focusing on actions that can have the most significant impact on improving maternal health outcomes:

1. Improve screening and management of high blood pressure during pregnancy:

- Implement routine blood pressure monitoring throughout pregnancy and the postpartum period to identify women at risk of developing or experiencing complications related to high blood pressure.
- Provide comprehensive training to healthcare professionals on the early detection and management of hypertensive disorders in pregnancy, including preeclampsia and gestational hypertension.
- Ensure timely referral and collaborative management by skilled professionals across different disciplines to prevent and mitigate the adverse effects of high blood pressure on maternal health.

2. Enhance age-specific monitoring and interventions:

- Develop age-specific guidelines for maternal health, considering the distinct risks and needs of different age groups.
- Establish targeted interventions and monitoring strategies for pregnant women and postpartum mothers in the older age group (over 65), focusing on age-related physiological changes, comorbidities, and potential complications.
- Prioritize comprehensive prenatal and postpartum care for younger women (under 25), with a focus on education, risk assessment, and early interventions to promote healthy pregnancies and optimal postpartum recovery.

3. Utilize clustering analysis to personalize care and interventions:

- Further explore and validate the identified clusters based on similar systolic blood pressure values to identify subgroups of women with specific risk profiles or characteristics.
- Develop tailored interventions and management strategies for each cluster, taking into account the distinct needs and risks associated with their blood pressure patterns.
- Implement ongoing monitoring and evaluation of the effectiveness of these tailored interventions to refine and improve care delivery.

4. Strengthen overall healthcare system coordination and collaboration:

- Encourage interdisciplinary teamwork and collaboration among healthcare professionals involved in maternal health, including obstetricians, midwives, nurses, and specialists, to ensure timely and comprehensive management of maternal health conditions.
- Facilitate knowledge sharing and continuous education through regular multidisciplinary meetings, training programs, and conferences to enhance the overall quality of care provided.
- Foster strong referral networks and seamless communication among healthcare facilities to ensure continuity of care and prevent gaps in the management of maternal health.

References:

1. American College of Obstetricians and Gynecologists. (n.d.). Hypertension in pregnancy. Retrieved from https://www.acog.org/womens-health/faqs/hypertension-in-pregnancy

2. Phipps, E., Prasanna, D., & Brima, W. (2019). Maternal and fetal biomarkers of preeclampsia - review of literature. Journal of Maternal-Fetal & Neonatal Medicine: The Official Journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstetricians, 32(23), 3945–3960. doi: 10.1080/14767058.2018.

3. Shahul, S., Tung, A., Minhaj, M., Nizamuddin, J., Wenger, J., Mahmood, E., Barrs, R., & Zuccarelli, L. (2015). Racial disparities in comorbidities, complications, and maternal and fetal outcomes in women with preeclampsia/eclampsia. Hypertension in pregnancy, 34(4), 506–515. doi: 10.3109/10641955.2015.1064499.

4. Denley, I., Sahhar, K., Hegaki, K., Penney, G., Norman, J., & Reynolds, R. (2006). Maternal age-specific incidence of eclampsia and other maternal complications. Australian and New Zealand Journal of Obstetrics and Gynaecology, 46(1), 31–35. doi: 10.1111/j.1479-828x.2006.00510.x.

5. Duckitt, K. (2005). Advanced Maternal Age: An Ongoing Risk? British Journal of Obstetrics and Gynaecology, 112(8), 1084–1089. doi: 10.1111/j.1471-0528.2005.00692.x.