Big Data and Data Mining Assignment

Olatunde Ibrahim - 202209928

1.0 Introduction

Maternal Health (MH) has been an important subject of focus in the healthcare industry for decades as it involves the entire well-being of women during the period of pregnancy, childbirth and the postnatal (World Health Organization, 2005). The MH is a key indicator of the overall health of a community because of its impacts on the mother's health, her child and other stakeholders especially in low-income nations (Tiruneh G et. al., 2021). World Health Organization (2020), published that approximately 287,000 women died during pregnancy and after childbirth in 2020 despite the significant progress that has been made in the past two decades. This poses a huge concern and necessitates continuous improvement of maternal health and experience.

In this report, further study was carried out through mining of the Maternal Health Risk Dataset in order to highlight data-led actions to upscale health outcomes, through the analysis of maternal health risk variables such as age, blood sugar, blood pressure, body temperature etc to provide recommendations to reduce maternal related mortality. This report also detailed various analytical operations carried out on the dataset, and careful considerations of maternal health risk indicators with fact-based interpretations.

2.0 Data Evaluation

2.1 Dataset

The Maternal Health Risk Dataset is a sample information from 1014 observations (patients) containing six different indicators for their (women's) health during their maternity period as well as the risk level assigned to each patient based on the indicators. These indicators include Age, Systolic Blood Pressure, Diastolic Blood Pressure, Blood Sugar (BS), Body Temperature (BodyTemp) and the Heart Rate.

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
0	25	130	80	15.0	98.0	86	high risk
1	35	140	90	13.0	98.0	70	high risk
2	29	90	70	8.0	100.0	80	high risk
3	30	140	85	7.0	98.0	70	high risk
4	35	120	60	6.1	98.0	76	low risk
...	...	...	...	...	...	...	...
1009	22	120	60	15.0	98.0	80	high risk
1010	55	120	90	18.0	98.0	60	high risk
1011	35	85	60	19.0	98.0	86	high risk
1012	43	120	90	18.0	98.0	70	high risk
1013	32	120	65	6.0	101.0	76	mid risk

1014 rows × 7 columns

Figure 1: View of dataset showing details of patients' maternal health indicators and assigned risk level.

This dataset offers the basis for exploring the relationship between different health indicators and maternal health outcomes. The analysis of the dataset can provide us the possibility of identifying critical risk factors in maternal mortality and morbidity in order to make meaningful recommendations to upscale the overall maternal health and care, thus averting adverse outcomes.

2.2 Data Cleaning and Assessment

A critical aspect of data cleaning is the observation of missing and null values (Koszalinski R, et al., 2018). In this dataset, it was found that no missing or null was present, hence no need for imputation or the deletion of any record. This further affords the opportunity to use the entire details of the dataset without losing any information. Furthermore, 562 of the 1,014 observations occurred as duplicates, however, this has been considered as mere coincidence and not necessarily accidental error as it is possible for two individuals to have similar key health indicators described in the dataset.

```
<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)
```

Figure 2: View of the information of the dataset showing non-null values.

3.0 Data Analysis

The assessment of the dataset further gives insight to the context of ranges of values for each of the variables. The age observation having a minimum of 10 and a maximum of 70 as shown in table 1. According to Cavazos-Rehg P. et al. (2015), complication risks such as preterm delivery, endometritis, superimposed preeclampsia, chorioamnionitis, fetal distress, poor fetal growth and postpartum haemorrhage have the highest odds among women between the ages of 11-19 and ages above 35 compared to women between ages 25-29.

	count	mean	std	min	25%	50%	75%	max
Age	1014.0	29.871795	13.474386	10.0	19.0	26.0	39.0	70.0
SystolicBP	1014.0	113.198225	18.403913	70.0	100.0	120.0	120.0	160.0
DiastolicBP	1014.0	76.460552	13.885796	49.0	65.0	80.0	90.0	100.0
BS	1014.0	8.725986	3.293532	6.0	6.9	7.5	8.0	19.0
BodyTemp	1014.0	98.665089	1.371384	98.0	98.0	98.0	98.0	103.0
HeartRate	1014.0	74.301775	8.088702	7.0	70.0	76.0	80.0	90.0

Table 1: Description of each variable in the Maternal Health Risk Dataset

While these complications are in part a subset of the overall risk in maternal health assessment, it can be assumed that they form a significant metric for the measurement of our own risk level. While the finding by Cavazos-Rehg P. holds quite true for ages between 25-29 and above 35 with reference to the Maternal Health Risk Dataset, it is quite at variance to evaluation made on ages between 11-19 in our dataset as shown in Figure 3.

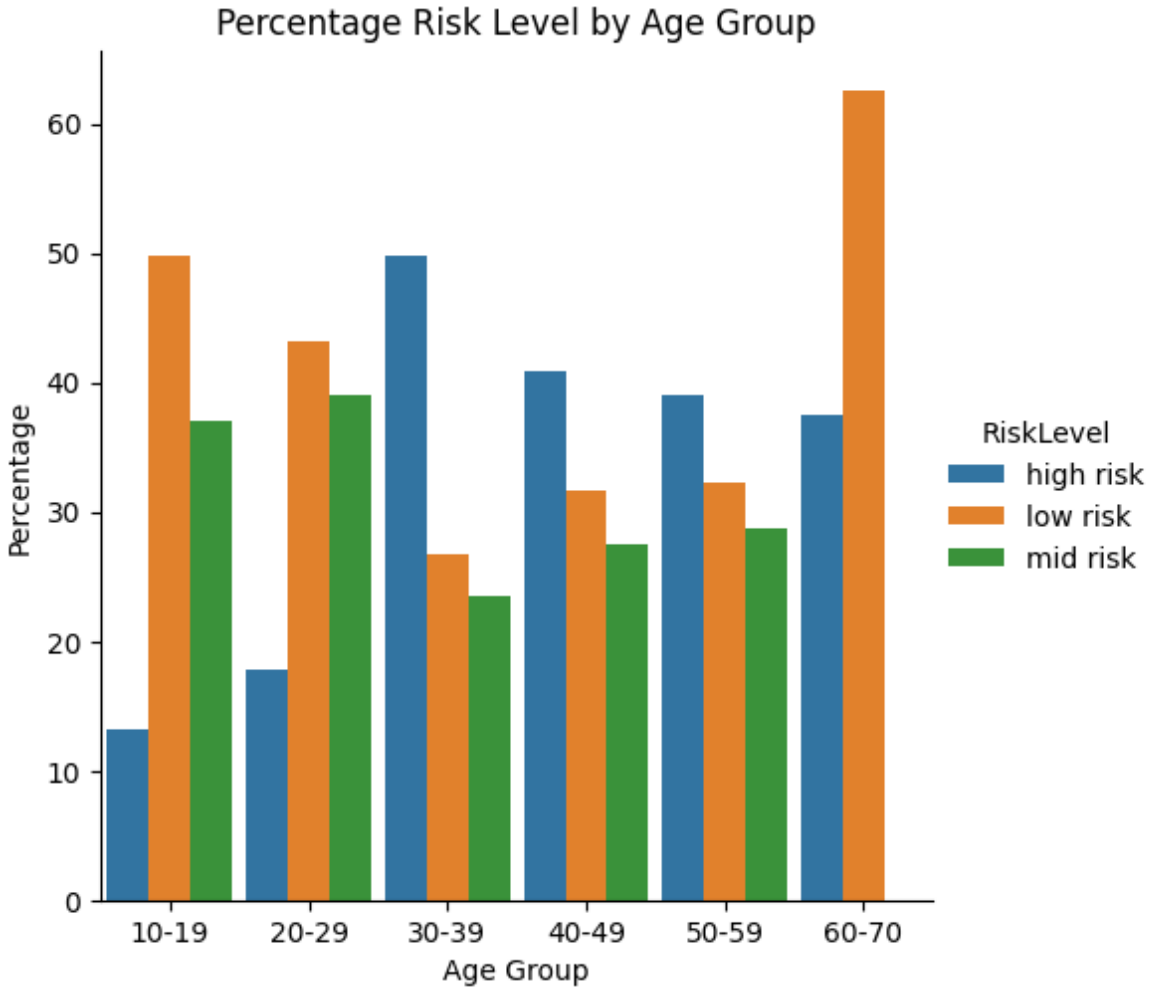

Figure 3: Percentage evaluation of risk levels at chosen age groups.

The above evaluation was done using percentages because each age group does not have equal observations, so it makes sense to evaluate the risk levels of individual age group in percents. High percentage of high-risk exposure are predominant in women above 30years with the largest occurrence of 50 percent in women between ages 30-39. The lowest percentage of high-risk level occurred in women between 10-19 years of age.

3.1 Systolic BP

The Systolic blood pressure is a measure of how much pressure the blood exerts against the artery walls when the heart beats (American Heart Association, 2023). Systolic blood pressure is a crucial indicator of the cardiovascular health of any person and is often used to diagnosis and monitoring of hypertension (Stevens S., 2016). In women, elevated systolic blood pressure throughout the period of pregnancy is a major cause of hypertensive disorders of pregnancy (HDP). This disorder is the second rank cause of maternal mortality across the globe after maternal haemorrhage (Kassebaum J. et al, 2016). According to (Vesna D. et al., 2022) elevated

systolic blood pressure even below the diagnostic threshold for hypertension can be associated with increased risk of preterm delivery. Analysis of the dataset further confirms these ideologies where approximately 95percent of patients with high systolic BP are at high maternal risk. There seems not to be much difference of high maternal risk exposure between patients with normal and low systolic BP. Moreover, it can be hypothesized that patients with low systolic blood pressure have least tendencies for risk exposures as 60percent of the entire low systolic BP category has low risk exposures.

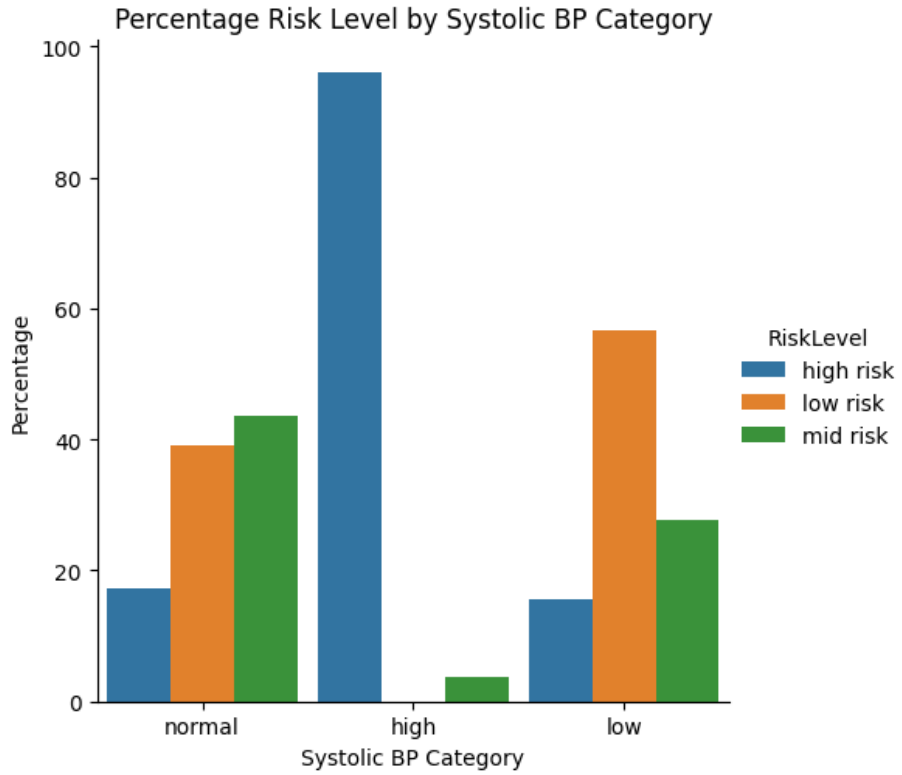

Figure 4: Plot of maternal risk levels with consideration to systolic BP categories. The low SBP category are observations below 110mmHg, the normal systolic BP include 110mmHg but below 140mmHg while high SBP categories are observations from 140mmHg and above.

3.2 Diastolic BP

The Diastolic blood pressure is a measure of how much pressure the blood exerts against the artery walls when the heart is at rest between beats. Gunnarsdottir J. et al (2019), in their research, concluded that increased diastolic blood pressure from early to mid-gestation was directly proportional to risks of preeclampsia and small-for-gestational-age birth (SGA). The results generated through the research also imply that diastolic BP increase around mid-gestation in women and may eventually lead to placental dysfunction disorders. The results are in synch with the outcome of the plot of diastolic blood pressure category and risk level.

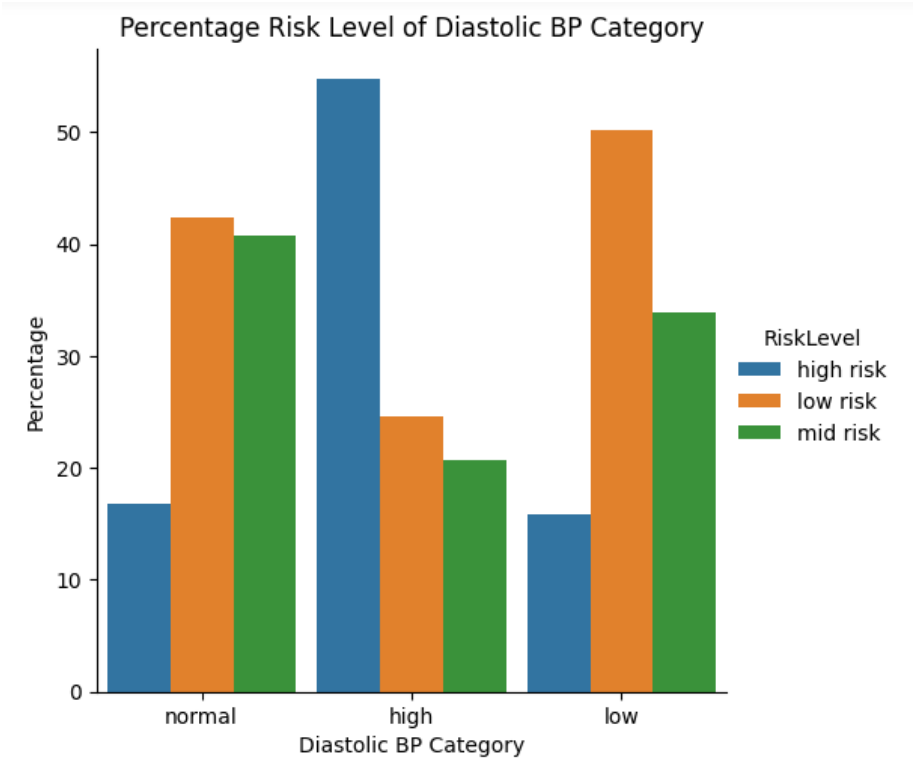

Figure 5: Plot of maternal risk levels with consideration to diastolic BP categories. The low DBP category are observations below 70mmHg, the normal systolic BP include 70mmHg but below 90mmHg while high SBP categories are observations from 90mmHg and above.

3.3 Analysing Relationships between Risk Factors/Variables

Stated in section 1.0, analysis of the relationships between the risk factors will further help to know critical associations of the variables in assessment of maternal health risks. Here, the correlation matrix will first be used to outline these relationships. The correlation matrix shows the level of the linear relationship between pairs of variables in the dataset. The correction matrix of the variables is expressed using heatmap as shown in Figure 6. The matrix represents this relationship between -1 and 1. A correlation of -1 implies a perfect negative correlation, while 1 implies perfect positive correlation. There is no relationship between variables that have correlation of zero. For instance, the correlation between Age and Systolic BP is 0.42, which indicates a moderate positive correlation. This means that as Age increases, Systolic BP tends to increase as well. Although correlation does not necessarily imply causation (Rohrer J, 2018), other factors such lifestyle, gene etc. can influence changes in Systolic BP. However, according to Michael G. et al. (2012) epidemiological surveys revealed a steady increase in Systolic BP with Age. This same linear relationship holds for Diastolic BP having a correction of 0.4 with age, although at a lower rate than Systolic BP; Diastolic BP have tendencies to fall at older ages (Franklin S. et al, 1997). A correlation of 0.79 between Systolic BP and Diastolic BP and implies a strong positive relationship between the two variables, such that as Systolic BP increases, the value of Diastolic BP tends to increase likewise. The correlation of -0.26 between Age and BodyTemp is a weak negative correlation. As a Age increases, BodyTemp decrease slightly.

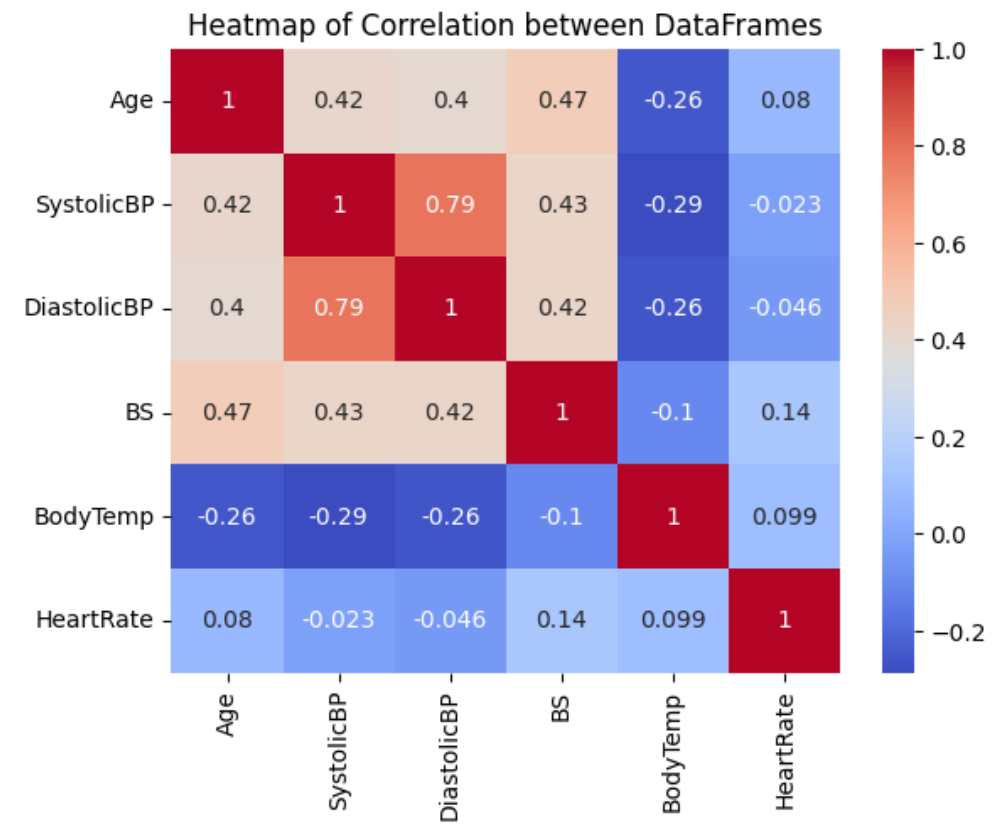

Figure 6: Correlation matrix of variables in the dataset.

3.4 Relationship between Age and Heart rate

The relationship between age and one's heartrate is quite complex and may also depend on other variable such as medication, lifestyle etc. However, heart rate is expected to decrease linearly with aging as a result of aging heart muscles. Komazawa M. et al (2017) concluded that women's heart rate become lower after 40years while men's heart rate tends to keep increasing until the 50's. In this research, the age group as been classified in intervals of 10, starting from the least age represented in the dataset, except for the age group between 60-70 years with interval of 11. This interval of 10 was chosen in order to have equal width in majority of the classes. The classification showed gradual increase of the mean heart rate between ages 10 to 49 and a consequent decrease between 50-59 as in figure 9. This agrees with Komazawa's research. However, the sudden rise in the mean heart rate between after 59years poses question for further examination or question.

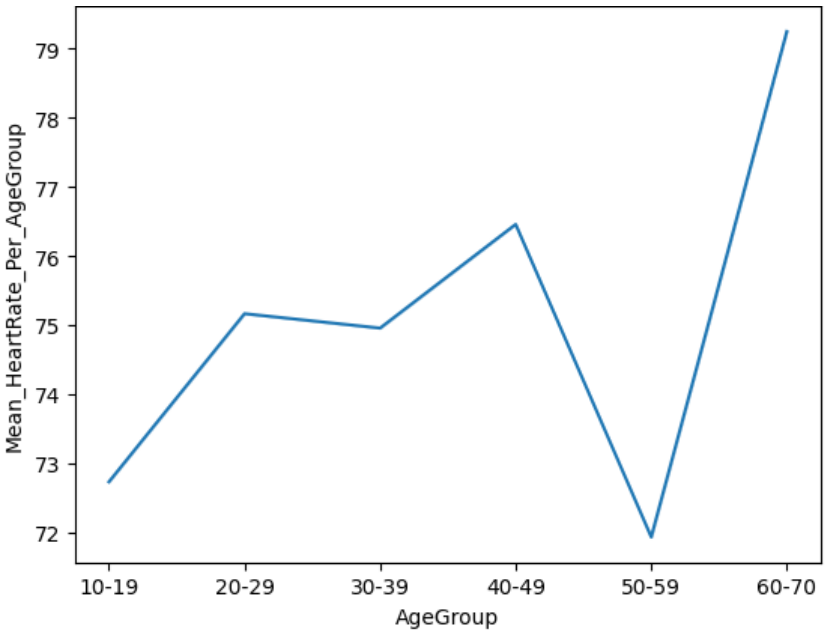

Figure 7: Plot showing Mean heart rate against age group.

3.5 Associations between Systolic BP and Diastolic BP

In order to investigate relationships, association rules mining technique was used. This rule offers us interesting analysis of the connection(s) between variables in our dataset; here we observe only Systolic BP and Diastolic BP. Both blood pressures have been categorized into three levels: High, Normal and Low. Below are the parameters for the categorization.

Blood Pressure	**High (mmHg)**	**Normal (mmHg)**	**Low (mmHg)**
Systolic BP	≥140	$110 \leq X < 140$	<110
Diastolic BP	≥90	$70 \leq X < 90$	<70

Table 2: Boundaries for the categorization of Systolic BP and Diastolic BP

For the high Systolic BP – high Diastolic BP association, the Support, Confidence, Conviction and Lift values being are 0.1154, 0.90, 7.28 and 3.31 respectively. This implies that 11.54% of the transactions in the observations have both high Systolic and high Diastolic BP, where 90% of the observations that contains high systolic BP contains high diastolic BP also. Patients with high systolic BP are 7.28times likely to have high systolic blood pressure than those with low or normal systolic BP. The lift further suggests that both high SBP and high DBP are 3times likely to occur together. Summarily, there exists a strong relationship between high SBP and high DBP. In the same vein, moderate relationship holds between a normal SBP and normal DBP. Just like the high Systolic BP – high Diastolic BP association, there exists a strong relationship between low Systolic BP – low Diastolic BP association.

Measure of Association	**High-High**	**Normal-Normal**	**Low-Low**
Support	0.1154	0.3353	0.2663
Confidence	0.90	0.8153	0.8411
Conviction	7.28	2.53	4.16
Lift	3.31	1.53	2.48

Table 3: Outcome of Association measures between pairs of Systolic and Diastolic BP

3.6 Clustering Systolic BP

In this aspect, patients with similar Systolic blood pressures are categorized Using K-Means clustering. K-means is one of the popular unsupervised machine learning algorithms for data set. In order to achieve the optimal number of clusters, the elbow methos was used. Shown below in Figure 10, the best number of clusters suggested is 3. The cluster outcomes of patients with similar Systolic BP are shown in Figure 11. This was matched against the patients age because age is well correlated with the SBP.

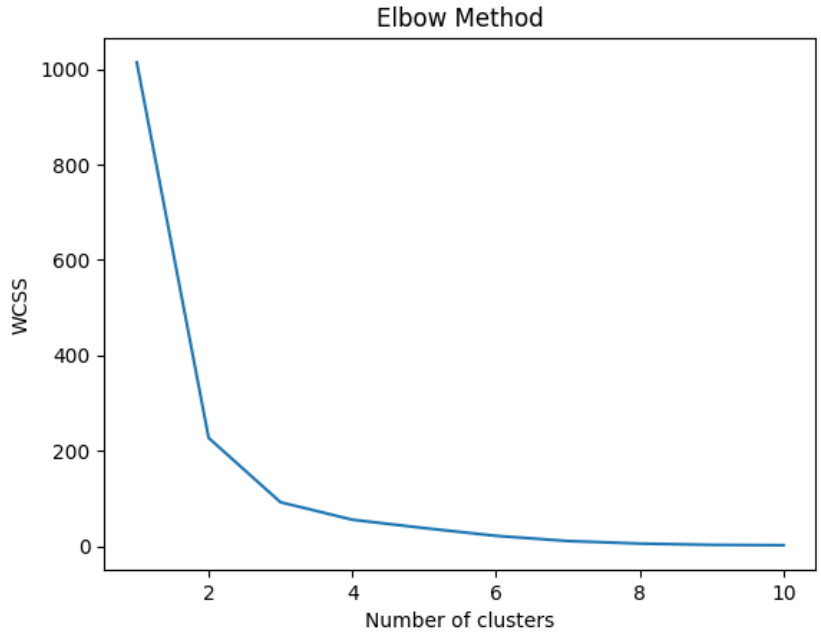

Figure 8. Elbow method used for determining optimal number cluster.

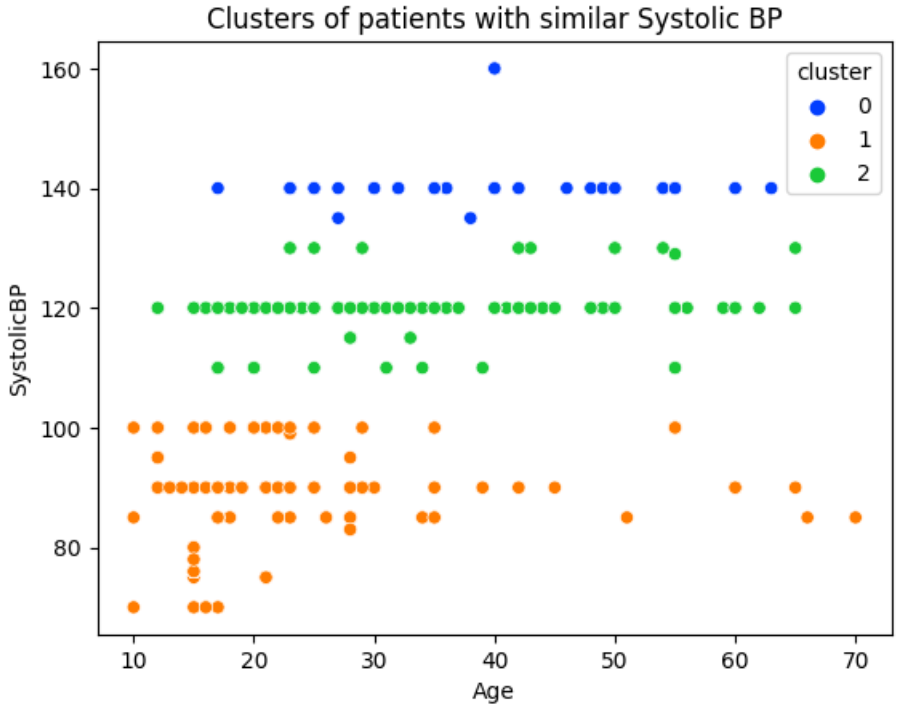

Figure 9. Cluster of Patients with Similar Systolic BP

4.0 Linear Model and Prediction

In the consideration of a linear model that can predict maternal health outcomes, a linear regression model was considered and developed. The explanatory variable – Diastolic BP was tested against the response variable – Systolic BP. The Diastolic BP was chosen because it has strong correlation (0.79) with the response variable such that changes in DBP can significantly affect the SBP. The dataset was split in to training and testing set in ratio 7:3. A linear regression object was used to fit the training set to the model by the fit() method. The model produced a Coefficients of 1.06 and Intercept of 31.77. This implies that a unit increase in the DBP produces 1.06 increase in the SBP.

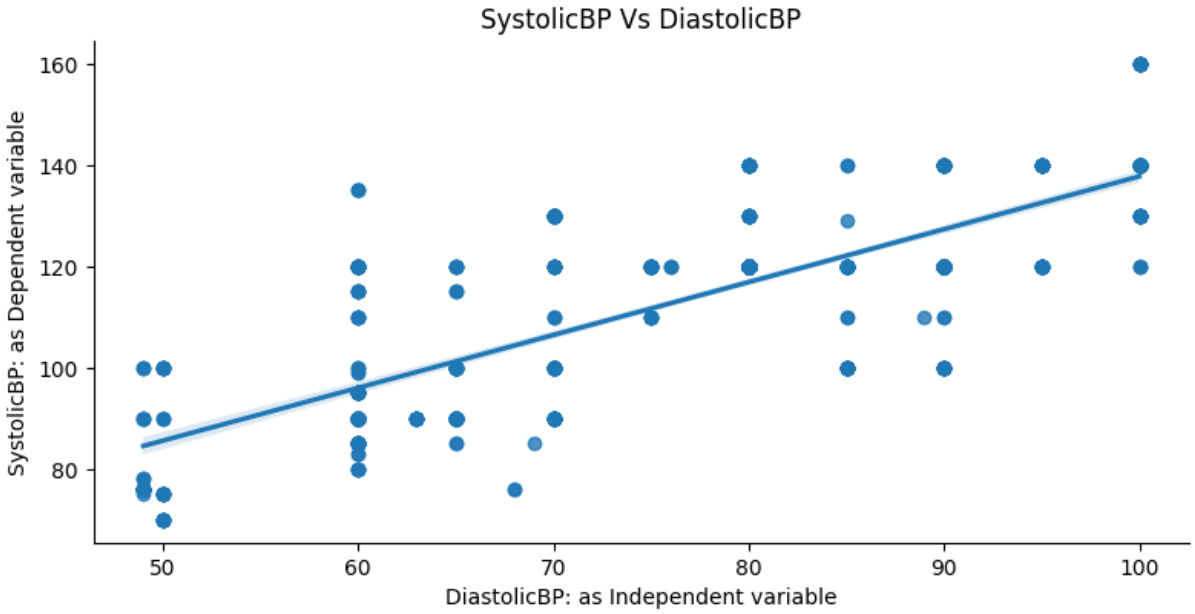

Figure 10. Plot of linear relationship between Systolic BP and Diastolic BP

5.0 Conclusion and Recommendation

Following the different analysis of the risk factors associated with maternal health safety, the major contributing factors to the overall risk levels are elevated Systolic blood pressure, High Diastolic blood pressure, Age and Blood Sugar. While analysis done have shown substantial agreement with previous research, it was also established that blood sugar is well correlated with Systolic and Diastolic BPs. In order to improve the overall maternal health and safety, the following recommendations are suggested.

- Regular blood pressure monitoring: This will afford women, particularly pregnant women the opportunity to detect unusual changes in their systolic and diastolic blood pressure early, in order to seek medical attentions to mitigate or avert any associated risk.
- Maintenance of healthy lifestyle: This includes adoption of healthy eating habits as high sugar content food can increase blood sugar levels, thus impacting their blood pressures. Moderate and regular exercise should be adopted in other to engage the heart muscles.

- Avoidance of Late Marriage/Conception: Early marriage should be encouraged in women because of the increased exposure to risk as the body ages. Analysis made showed that conception in women at ages between 20-29 possess less risk compared to older ages.
- Collaborative care and Maternal Education: Women should be able to access care before, during and after conception. Education about associated risk should also be made accessible to women in order to avert any ignorance.

REFERENCE

American Heart Association, 2023. Understanding Blood Pressure Readings. [Online] Available at: https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings
[Accessed 11 May 2023].

Cavazos-Rehg PA, Krauss MJ, Spitznagel EL, Bommarito K, Madden T, Olsen MA, Subramaniam H, Peipert JF, Bierut LJ. Maternal age and risk of labor and delivery complications. Matern Child Health J. 2015 Jun;19(6):1202-11. doi: 10.1007/s10995-014-1624-7. PMID: 25366100; PMCID: PMC4418963.

Franklin S., Gustin W, Wong D., Larson M. et al., 1997. Hemodynamic patterns of age-related changes in blood pressure: the Framingham Heart Study. Circulation. 96:308.

Gunnarsdottir J, Akhter T, Högberg U, Cnattingius S, Wikström AK. Elevated diastolic blood pressure until mid-gestation is associated with preeclampsia and small-for-gestational-age birth: a population-based register study. BMC Pregnancy Childbirth. 2019 May 28;19(1):186. doi: 10.1186/s12884-019-2319-2. PMID: 31138157; PMCID: PMC6537437.

Kassebaum J, Barber RM, Bhutta ZA, Dandona L, Gething PW, Hay SI, Kinfu Y, Larson HJ, Liang X, Lim SS, et al. Global, regional, and national levels of maternal mortality, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015.Lancet. 2016; 388:1775–1812. doi: 10.1016/S0140-6736(16)31470-2

Komazawa M., Itao K., Lopez G., & Luo Z. (2017). Evaluation of Heart Rate in Daily Life Based on 10 Million Samples Database. Global Journal of Health Science. 9. 105. 10.5539/gjhs.v9n9p105.

Koszalinski R., Tansakul V., Khojandi A., & Li X. (2018). Missing Data, Data Cleansing, and Treatment from a Primary Study: Implications for Predictive Models. CIN: Computers, Informatics, Nursing. 36. 367-371. 10.1097/CIN.0000000000000473.

Michael G., Aaron B., Daniel E. R., Jonathan S. and Hillard K., 2012. Does Blood Pressure Inevitably Rise With Age? : Longitudinal Evidence Among Forager-Horticulturalists. AHA Journals, 60(1), pp. 6-7.

Rohrer J, 2018. Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. Advances in Methods and Practices in Psychological Science. 2018;1(1):27-42. doi:10.1177/2515245917745629

Stevens S., Wood S, Koshiaris C, Law K, Glasziou P, Stevens RJ, McManus RJ. Blood pressure variability and cardiovascular disease: systematic review and meta-analysis. BMJ. 2016 Aug 9;354:i4098. doi: 10.1136/bmj.i4098. PMID: 27511067; PMCID: PMC4979357.

Tiruneh G., Demissie M, Worku A, Berhane Y, 2021. Community's experience and perceptions of maternal health services across the continuum of care in Ethiopia: A qualitative study. PLoS One. 2021 Aug 4;16(8):e0255404. doi: 10.1371/journal.pone.0255404. PMID: 34347800; PMCID: PMC8336848.

Vesna D. Garovic et al., 2022. Hypertension in Pregnancy: Diagnosis, Blood Pressure Goals, and Pharmacotherapy: A Scientific Statement from the American Heart Association. AHA Journals, 79(2), pp. 21-41.

World Health Organization, 2020. Maternal health. [Online]
Available at: https://www.who.int/health-topics/maternal-health#tab=tab_1
[Accessed 05 May 2023].

World Health Organization. World health report 2005: make every mother and child count. Geneva: World health; 2005.