Table of Contents

Report: Analysis of Maternal Health Data

1.0 Introduction

Maternal health is critical to safeguarding the health of pregnant mothers and their unborn children. Our goal as a data scientist collaborating with health professionals is to analyse maternal health data, find evidence-based solutions, and improve health outcomes. In this study, we will look at the dataset, do statistical analysis, and draw conclusions to help guide initiatives to enhance maternal health outcomes.

The World Health Organisation (WHO) defines maternal health as women's health throughout pregnancy, delivery, and the postpartum period. Maternal mortality, morbidity, and obstetric impairment are identified as three essential components of maternal health by the organization (WHO, 2021).

In this study, maternal health data is studied in order to predict systolic blood pressure using related variables to help in understanding the cause of maternal morbidity.

2.0 Background of study

Before we get into the analysis, Maternal mortality is a major issue in the globe, with roughly 800 women dying each day from avoidable causes connected to pregnancy and childbirth (WHO, 2021). Although high-income nations have made tremendous progress in lowering maternal mortality rates, low-income countries continue to struggle with this issue.
Prenatal blood pressure measurement indicates quality. In well-resourced settings, lower hypertension cutoffs worsen pregnancy outcomes. The research examined blood pressure cutoffs, adverse outcomes, and diagnostic test characteristics in low-resource settings. (Bone et al., 2021)

3.0 Methodology

3.1 Dataset

Using the mhs.csv dataset in this research which contains information on pregnant women's demographic and health variables, such as age, blood sugar, diastolic blood pressure, systolic blood pressure and other factors that may impact maternal health. The data has 1014 entries with an average age of 30. The data does not have any null values and has about 564 duplicated rows which were not dropped because there is not distinct identifier for each row. The variable of importance (target variable) in this research is systolic blood pressure, which measures the pressure in the arteries as the heart beats.

Building a linear model to predict systolic blood pressures will involve both features selection and model selection.

3.2 Feature selection

We categorized the data into numerical and categorical to examine their relationship.
This phase uses one hot encoding. We then generated a heatmap to display the relationship and correlation coefficient to quantify the association between the outcome variables.
After the exploratory data analysis (EDA), we picked the variables based on research by (Bhandari 2022) that a correlation value of 0.3 and -0.3 indicate a moderate correlation. The following independent variables were chosen.

Table 1.0: Interpreting a correlation coefficient.

Correlation coefficient	**Correlation strength**	**Correlation type**
-0.7 to -1	Very strong	Negative
-0.5 to -0.7	Strong	Negative
-0.3 to -0.5	Moderate	Negative
0 to -0.3	Weak	Negative
0	None	Zero
0 to 0.3	Weak	Positive
0.3 to 0.5	Moderate	Positive
0.5 to 0.7	Strong	Positive
0.7 to 1	Very strong	Positive

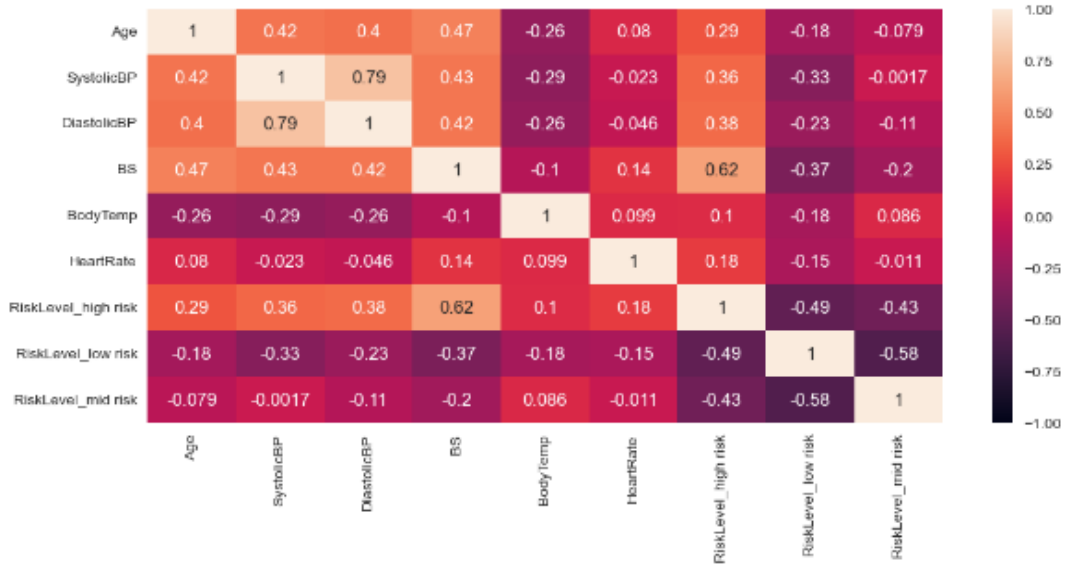


Figure 1.1: Exploratory data analysis to get more insight into the Data.

From the heatmap we selected the various independent variables based on the category of correlation strength as shown in Figure 1.2, Using systolic BP. as the target variable.

INDEPENDENT VARIABLES	CORRELATION VALUE
Age	0.40
Blood Sugar	0.42
Diastolic BP	0.79
RiskLevel_low risk	0.33

Table 1.2: Exploratory variables selected based on the Heatmap.

3.3 Model selection

We adopted the multiple linear regression which enables the inclusion of multiple independent variables in the analysis, allowing me to assess their individual and combined effects on the dependent variable.

3.4 Model building

Building and fitting the model: The data was found to contain a significant number of duplicates. We couldn't drop this duplicate as it may render the rest of the data insignificant as the data were split up into three different categories: training, testing, and validation. The variable for the test

data was predicted using our model, which was also used to train the data we had. After that, we determined the evaluation metrics for the model, which were the mean absolute error, the root mean squared error, and the coefficient of determination. The result reveals the values that were computed for the relevant measures, which were revealed to be 8.55, 10.68, and 0.65 correspondingly. These numbers provide some insight into how well the linear regression model is working.

3.5 Principal Component Analysis (PCA)

PCA is a common data analysis tool in science. Principal component analysis optimizes data re-expression. By using this technique, data patterns and noise should become clearer. (2019, Kurita) To reduce variables and extract the most significant data from the dataset, we employed Principal Component Analysis (PCA). PCA lets us find the most important components, decrease dimensions, and save as much information as possible. Principal component analysis (PCA) will reveal the data's structure and maybe reveal hidden factors that affect maternal health. The dataset's multicollinearity is minimal, however the PCA performs poorly. PCA's data may be inadequate to reflect it. Also, we discovered that using nonlinear regression as the base model, the PCA performed better.

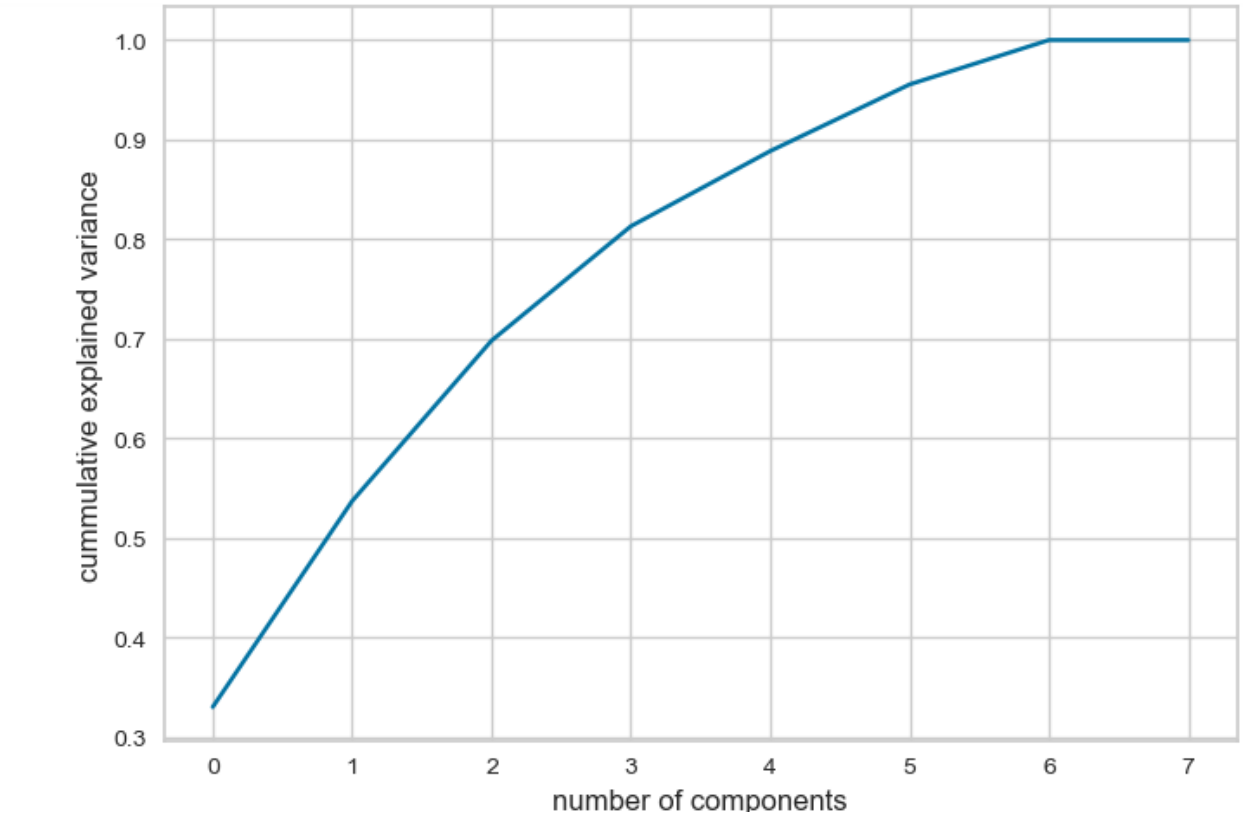


Figure 1.2 Graph of Cumulative explained variance

These lines illustrate the cumulative variance ratio explained as a function of the number of components. It aids in visualizing the quantity of data retained by each additional component.

By looking at the plot, you can determine how much of the data's variance each additional component explains. The plot enables me to determine the number of retained principal components based on the intended level of variance explained. To avoid overfitting or excessive dimensionality, I will typically select the number of components that explain 85% or more variance by identifying the index where the cumulative explained variance ratio crosses or exceeds 0.85 on the components.

There is evidence of the PCA not performing well on the dataset, the number of multi collinearity present in the dataset is not many. The information captured by PCA may be insufficient to effectively represent the data. The results obtained are presented in the table below.

Table 1.3 Evaluation metrics.

Metrics	Linear Regression	PCA
Mean absolute Error	8.55	10.75
Root mean squared error	10.68	13.12
Coefficient of determination	0.65	0.47

3.6 Age Grouping and Heart Rate

We investigated the relationship between age and heart rate by grouping the dataset into specific age intervals. We calculated the mean heart rate for each age group and visualized the results graphically. This analysis helps us understand how heart rate varies across different age groups and provides insights into age-related patterns in heart rate during pregnancy.

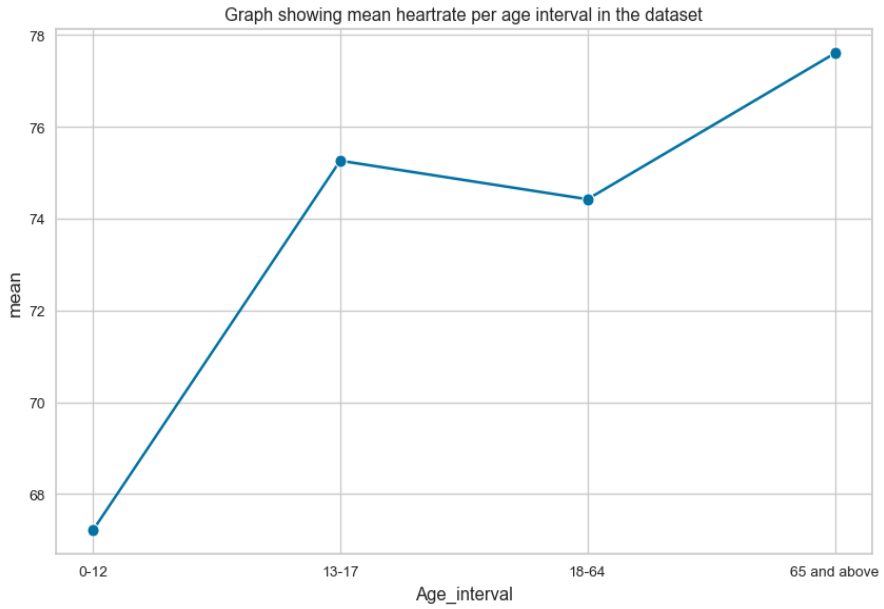


Figure 1.3 Graph showing mean heartrate per age group in the dataset.

The specified age ranges for children are 0–12, adolescents are 13–17, adults are 18–64, and older adult are 65 and beyond. These age ranges, which represent several life stages, are frequently utilised in health research. (Age, 2022)

Our findings indicate that maternal heart rate can be affected by age, with elderly women generally exhibiting greater heart rates.

When it comes to blood pressure, systolic and diastolic readings of at least 140 and 90 are regarded as high blood pressure, respectively, whereas readings between 110 and 140 and 70 and 90 are regarded as normal. Low blood pressure is defined as systolic and diastolic readings under 110.

3.7 Association Analysis of Blood Pressure Categories

An association data mining pattern is an analytical technique that discovers common patterns, relationships, or causal structures from data sets found in many types of databases including relational databases, transactional databases, and other data repositories. This kind of data mining pattern is also known as association data mining (Patel, 2018).

We conducted an association analysis to investigate the associations between different blood pressure categories (high/high, normal/normal, and low/low) for both systolic and diastolic blood pressure. We calculated support, confidence, conviction, and lift measures to quantify the strength of these associations. This analysis allows us to understand the relationships between different blood pressure categories and identify patterns that may guide interventions and treatment strategies.

Antecedents (Diastolic BP)	Consequents (Systolic BP)	support	Confidence	Lift	Convictions
High	High	0.12	0.42	3.31	1.51
Normal	Normal	0.34	0.82	2.48	2.53
Low	Low	0.27	0.84	1.53	4.20

Table 1.5 Table showing the association data mining pattern.

In the context of association analysis, the information provided in the above explains the notions of support, confidence, lift, and conviction. When doing association rule mining, these measurements are often used in order to examine the links that exist between the various variables or characteristics. The following is a rundown of their respective meanings:

Support: The support of an itemset (in this example, diastolic BP and systolic BP) is the percentage of transactions in the dataset that include that itemset. It provides an indication of the item set's occurrence frequency within the dataset. For instance, if a transaction has a support of 12% for both high diastolic and high systolic blood pressure, this indicates that the combination of these two characteristics occurs in 12% of all transactions.

Confidence: The confidence of an association rule (diastolic BP -> systolic BP) counts the percentage of transactions that include both the antecedent and the consequent of the rule (diastolic BP and systolic BP, respectively). It provides a numerical representation of the power of the rule by revealing the frequency with which the consequent is found in dealings that include the antecedent. For instance, a confidence level of 82% for the rule normal diastolic BP -> normal systolic BP implies that normal systolic BP is likewise present in 82% of the transactions in which normal diastolic BP is present.

The lift metric, which accounts for the independence of the two attributes, may assess an association rule's antecedent-consequence link. It achieves this by comparing the rule's actual support to the expected support if the antecedent and consequent were unconnected. A lift value greater than 1 indicates that the antecedent increases the likelihood of future occurrence. For example, a lift of 3.3 indicates that high diastolic blood pressure increases the likelihood of high systolic blood pressure by 3.3 times.

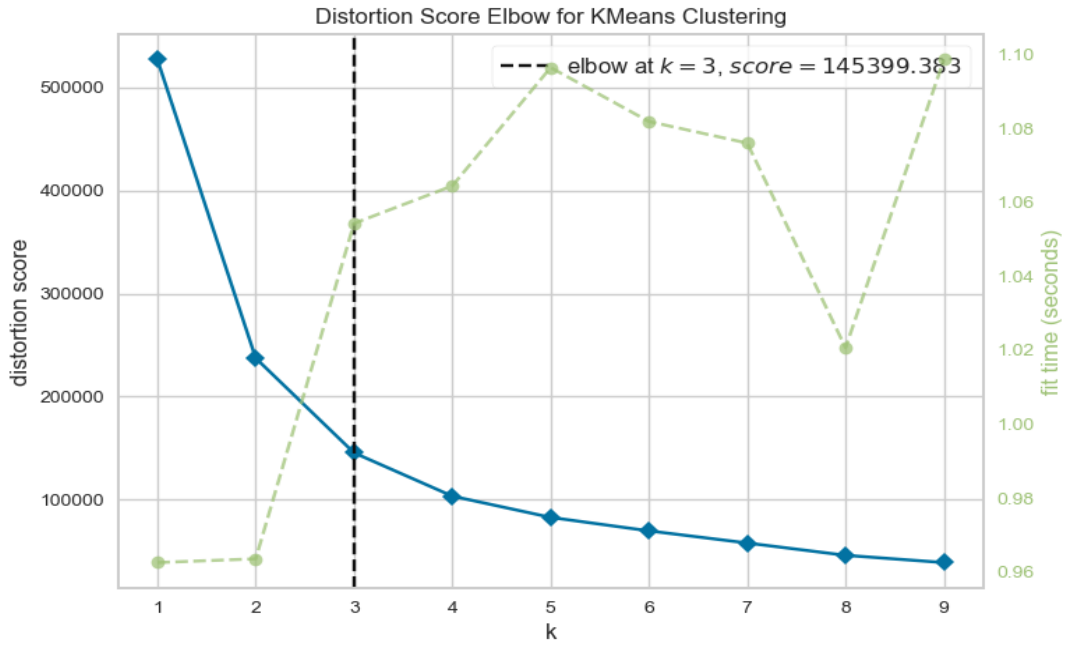
Conviction measures an association rule's antecedent-consequence relationship. Conviction rates. As predicted, it compares this to the frequency of the antecedent without the consequent. A conviction value larger than one implies a positive relationship, implying that the lack of the consequent diminishes the possibility of the antecedent. A belief of 4.2 for low diastolic BP -> low systolic BP suggests that the chance of having one without the other is 4.2 times lower than if the two traits were unrelated. Low diastolic BP -> high systolic BP is a 100% belief.

In general, these measurements provide light on the connections and interdependencies that exist between diastolic blood pressure and systolic blood pressure. They are helpful in quantifying the

degree of connection, and they may be used to find patterns and draw conclusions about the relationship between these two variables in the information that was provided.

3.8 Clustering Analysis

Clustering is a valuable data science technique. It is a method for discovering cluster structure in a data set that is characterized by the highest degree of similarity within the same cluster and the highest degree of dissimilarity between clusters. (K. P. Sinaga & M. -S. Yang, 2020)



- Figure 1.4 Graph showing Distortion Score Elbow for KMeans Clustering.

In order to locate the group of individuals that have systolic blood pressures that are comparable I used K-means clustering and the elbow technique to figure out the number of clusters, and it presented the graph above to show that it found there to be three of them.

In the clustering procedure, the silhouette score was used as a measurement tool to determine how well-defined the clusters are. A silhouette score of 0.52 shows that the clusters produced by the k-means algorithm are well-separated and distinct. This is shown by the fact that the score was calculated.

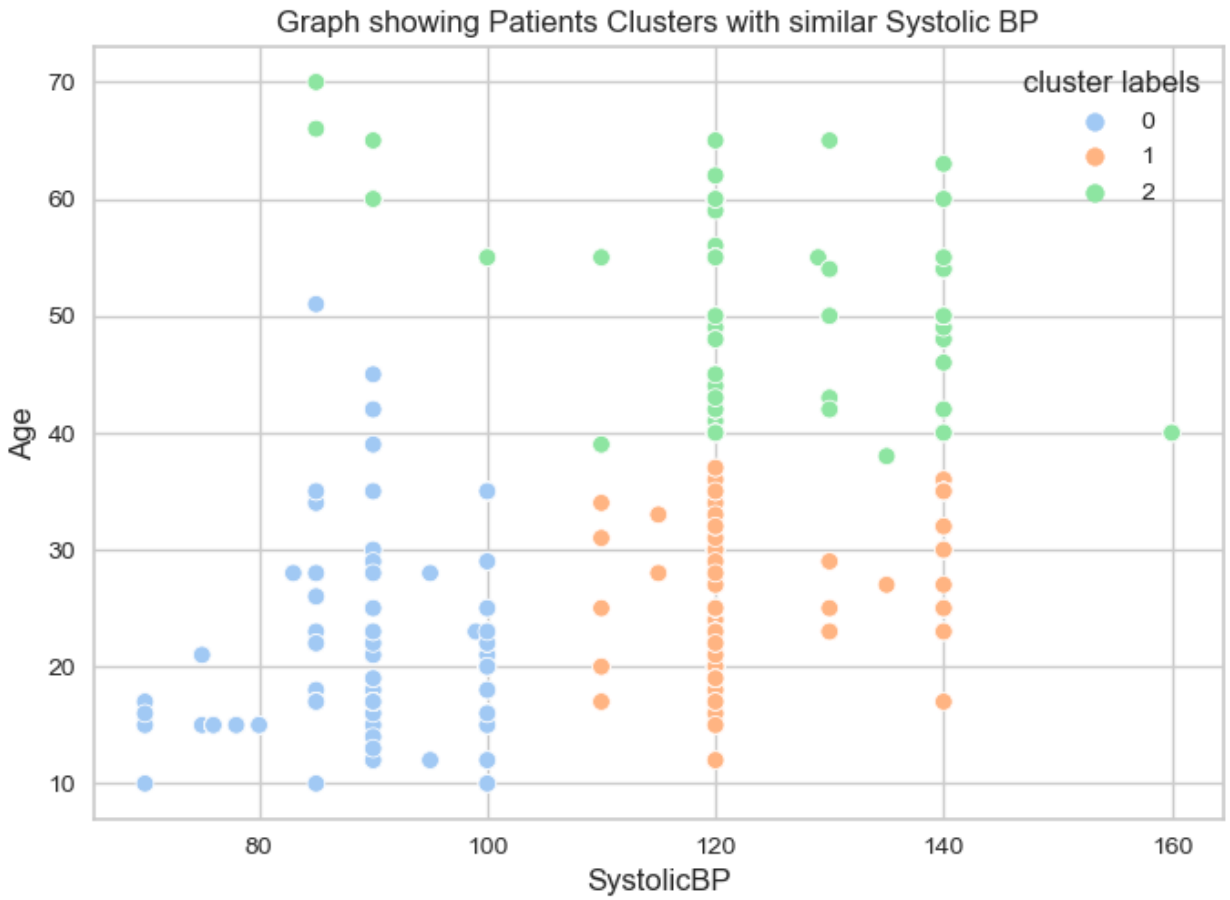


Figure 1.5 Visualization of patient clusters with Similar Systolic BP

3.9 Correlation Analysis

Age and systolic blood pressure were correlated to determine their association. Correlation assesses linear connection strength and direction. Understanding the age-systolic blood pressure connection helps us uncover possible relationships and guides treatments and monitoring techniques for various age groups.

We determined the connection between age and systolic blood pressure, and it turned out to be a positive one albeit a weak one with a value of 0.42. This indicates that there is a correlation between being older and having a greater risk of having higher systolic blood pressure.

4.0 Prediction

The model built was able to predict the systolic blood pressure using the variables obtained from the EDA analysis.

5.0 Conclusion and recommendation

There is a need for comprehensive care solutions for hypertension during pregnancy since high systolic and diastolic blood pressure often coexist.
Based on clustering analysis, tailored therapies should be created to target certain patient groupings with comparable systolic blood pressure characteristics. The following are the recommendation suggested:

- To improve treatment results and reduce problems, customize care regimens to clustering analysis-identified patient groupings.
- Research prenatal weight growth, pre-existing medical disorders, and socioeconomic factors that may contribute to hypertension during pregnancy beyond age and heart rate.
- Encourage prenatal blood pressure and heart rate monitoring to identify and treat hypertension-related problems regardless of mother age.
- Work with healthcare providers and researchers to design and test lifestyle, stress management, and nutrition treatments to reduce pregnancy-related hypertension risk factors.

Bibliography

Anekwe, L. (2020) Ethnic disparities in maternal care. *Bmj,* 368.

Bone et al., (2021) Blood pressure thresholds in pregnancy for identifying maternal and infant risk: A secondary analysis of community-level interventions for pre-eclampsia (CLIP) trial data. *The Lancet Global Health,* 9 (8), e1119-e1128.

Crowe, R. (2022) Factors contributing to maternal health inequalities for women who are not white british in the UK. *British Journal of Midwifery,* 30 (3), 160-171.

Darney, P. D., Nakamura-Pereira, M., Regan, L., Serur, F. & Thapa, K. (2020) Maternal mortality in the United States compared with ethiopia, nepal, brazil, and the United Kingdom: Contrasts in reproductive health policies. *Obstetrics & Gynecology,* 135 (6), 1362-1366.

de Swiet, M. (2000) Maternal mortality: Confidential enquiries into maternal deaths in the United Kingdom. *American Journal of Obstetrics & Gynecology,* 182 (4), 760-766.

K. P. Sinaga & M. -S. Yang. (2020) Unsupervised K-means clustering algorithm.

Knight, M., Kenyon, S., Brocklehurst, P., Neilson, J., Shakespeare, J. & Kurinczuk, J. J. (2017) Saving lives, improving mothers' care: Lessons learned to inform future maternity care from the UK and ireland confidential enquiries into maternal deaths and morbidity 2009-2012.

Knight, M., Nair, M., Tuffnell, D., Kenyon, S., Shakespeare, J., Brocklehurst, P. & Kurinczuk, J. J. (2016) *Saving lives, improving mothers' care: Surveillance of maternal deaths in the UK 2012-14 and lessons learned to inform maternity care from the UK and ireland confidential enquiries into maternal deaths and morbidity 2009-14* Oxuniprint.

Kurita, T. (2019) Principal component analysis (PCA). *Computer Vision: A Reference Guide,* 1-4.

Lewis, G. (2012) Saving mothers' lives: The continuing benefits for maternal health from the United Kingdom (UK) confidential enquires into maternal deaths. *Seminars in perinatology.* Elsevier.

Propper, C., Rigg, J. & Burgess, S. (2007) Child health: Evidence on the roles of family income and maternal mental health from a UK birth cohort. *Health Economics,* 16 (11), 1245-1269.

Say, L. & Raine, R. (2007) A systematic review of inequalities in the use of maternal health care in developing countries: Examining the scale of the problem and the importance of context. *Bulletin of the World Health Organization,* 85 (10), 812-819.

Shakespeare, J. & Knight, M. (2015) Maternal health in pregnancy: Messages from the 2014 UK confidential enquiry into maternal death. *British Journal of General Practice,* 65 (638), 444-445.