Student Name: Ifeanyi Anthony Okpala

Student ID: 202203449

Module: Big Data and Data Mining (771762)

ANALYSING RISK FACTORS FOR MATERNAL HEALTH

1.0 Introduction

Systolic Blood Pressure (BP) is caused when the heart pushes blood out to the rest of the body, while Diastolic BP is the rest between beats (National Health Service, 2019). In a maternal health scenario, a controlled systolic BP can help prevent or manage hypertension and reduce the risk of related health complications (Zhou et al., 2021).

In this report, we will be working with a maternal health dataset of 1014 rows, consisting of the systolic BP, Diastolic BP, Age, Blood Sugar (BS), Body Temperature, Heart Rate, and Risk Level features. Our aim is focused on analysing the mutual relationships between systolic BP and the other features and drawing insights on how to improve maternal health.

2.0 Analysis

In this section, we will begin by investigating and cleaning our data, before analysing our clean dataset for the task at hand.

2.1 Data Cleaning and Exploratory Data Analysis (EDA)

Data cleaning is the process of modifying data that has incorrect, irrelevant, or improperly formatted values for the purpose of analysis (Sisense, 2019). During this process, we observed the data type format and checked for duplicated values. 562 identical rows were detected but not removed because there was no identifier column such as patient id to prove they are duplicated values.

Further investigation was done by using statistical attributes as shown in **Table 1** below.

Table1. Statistical Representation of the Numerical Features in our Dataset. *This is a data cleaning process used to check for any incorrectness or anomaly in our dataset. We can point out a few anomalies such as the minimum Heart Rate and Age in our data set which has a value of 7 and 10, respectively. In addition, Systolic BP and Age have a maximum value of 160 and 70, respectively, which can also be considered irregular.*

	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

From **Figure 1**, we can see a visual relationship between Systolic BP and Heart Rate using a scattered plot.

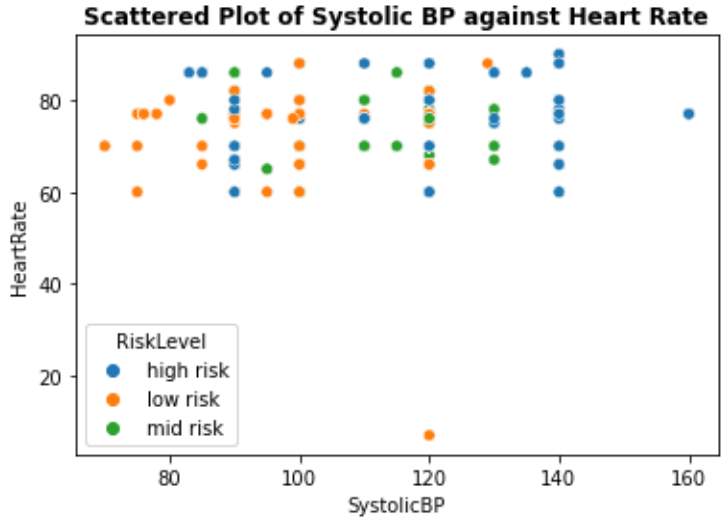


Figure 1. Visual Representation of the Systolic BP against Heart Rate. *This is a scattered plot showing the relationship between Systolic BP and Heart Rate. We included the Risk Level feature as the hue for categorization, which gives us more insights on the data. We can detect few outliers, which has a Heart Rate below 20 and Systolic BP at 160.*

After investigating our data through statistical and visualization methods, the following were considered:

- Retain the irregular ages of 10 and 70, since regardless of age, any woman can become pregnant through artificial reproductive technology (ART) medical assistance such as in vitro fertilization (IVF) (Kay, 2020).
- Retain Systolic BP maximum value of 160, since the hue shows it's under the high-risk level category as shown in figure 1.
- Drop the heart rate of 7 beats per minute (bpm), since the normal Heartbeat for an adult fall within the range of 60 – 100 bpm during resting mode (British Heart Foundation, 2019).

After performing data cleaning, we used a pair plot to have an overview of our cleaned data as shown in **Figure 2**.

Pair-Plot EDA Representation

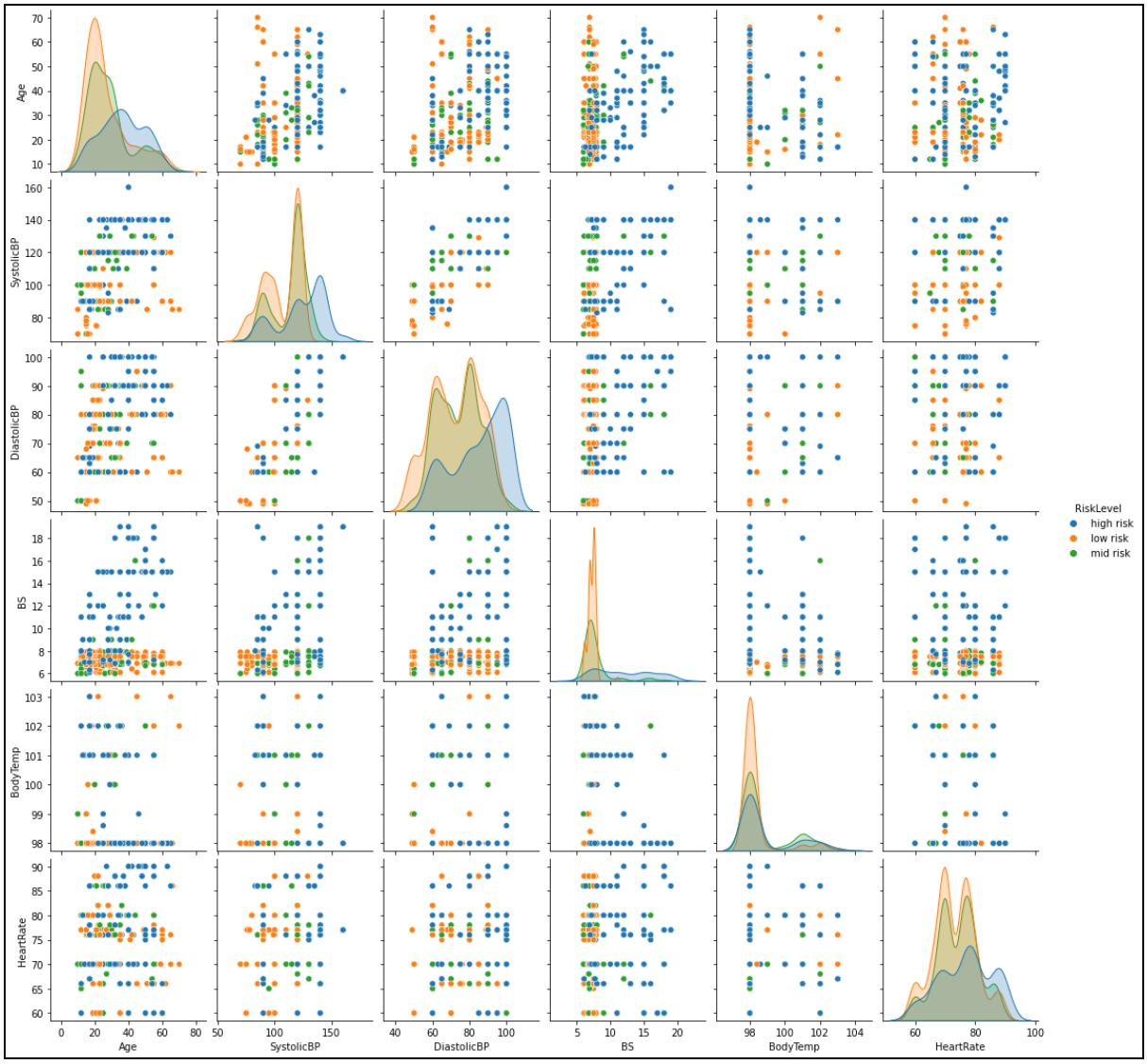


Figure 2. Exploratory Data Analysis (EDA) Representation of Cleaned Data using Pair-Plot. *This shows a visual relationship of the pair combination of our feature variables. In the case of anomalies, not much can be drawn except the case of the high-risk systolic bp which was earlier discussed.*

2.2 Feature Selection and Linear Modelling

At this stage, we will be using different techniques for feature selection to determine the relationship between the response variable, Systolic BP, and other features in our data before then creating a linear model for our analysis. In order to maintain all features, converting Risk Level categorical data into numeric using the get dummies function was done.

A heat map was used to understand the correlation between paired features in our data set as shown in **Figure 3.**

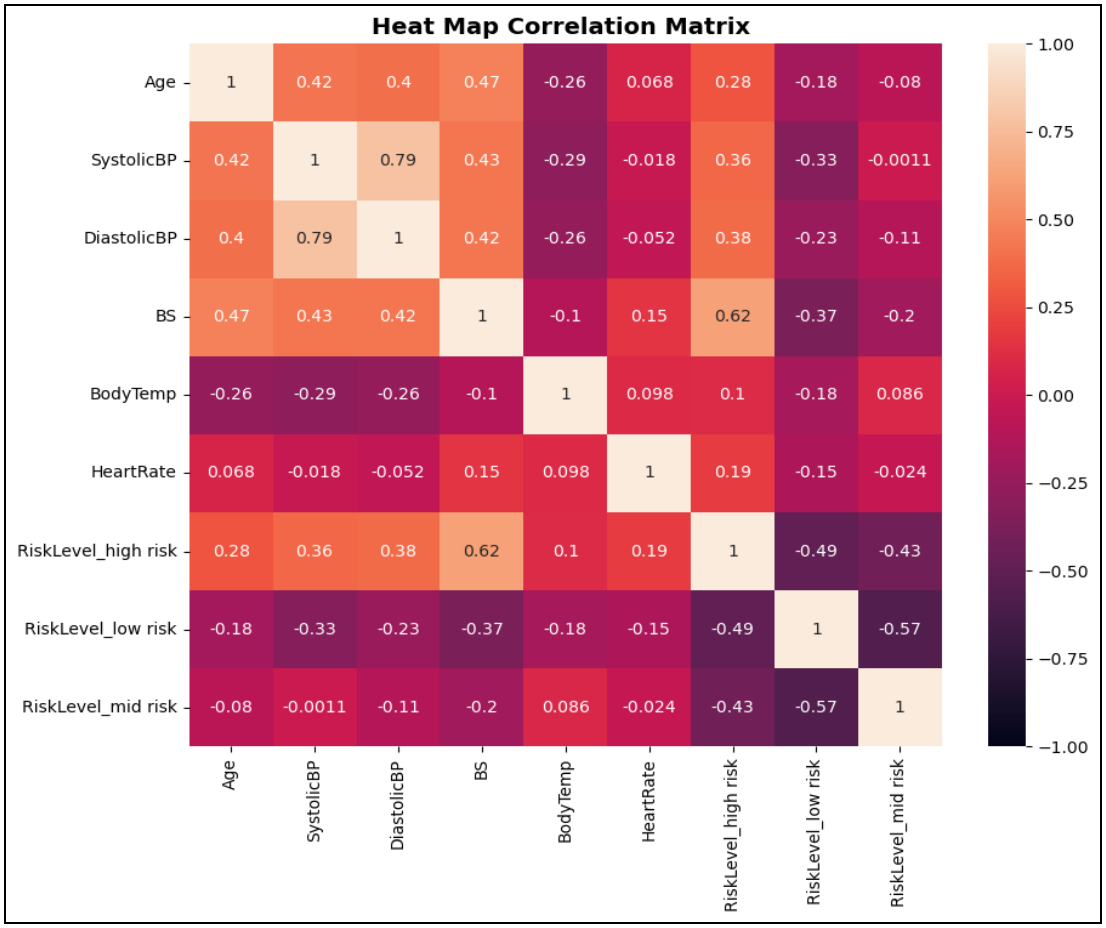


Figure 3. Exploratory Data Analysis (EDA) Representation using Heat Map. *This shows the correlation between each of the feature variables. Considering absolute values, we can see that Diastolic BP, Age, Blood Sugar (BS), High-Risk Level, and Low-Risk Level all have a decent correlation with the response variable.*

Making use of the SelectKBest method for feature selection with f_Regression as the scoring function shown in **Figure 4**, we observed a similar distribution as the latter.

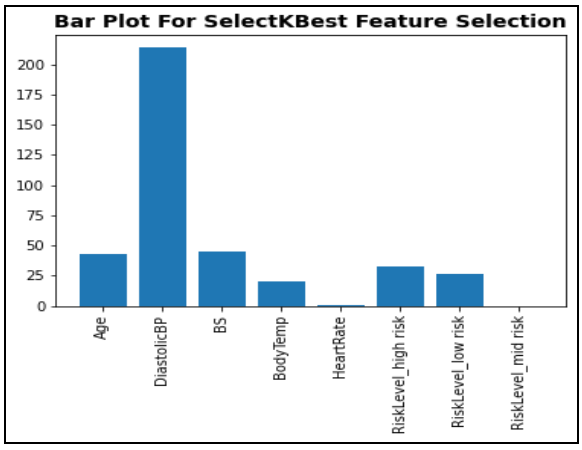


Figure 4. SelectKBest Feature Selection Representation using Bar Plot. *This shows each feature importance to the response variable, and an identical interpretation with the heatmap correlation matrix.*

Forward Feature Selection was also used as a feature selection technique. This process involves iterating through each feature and evaluating the performance. **Figure 5** shows a graphical representation of the best feature combinations at each iteration.

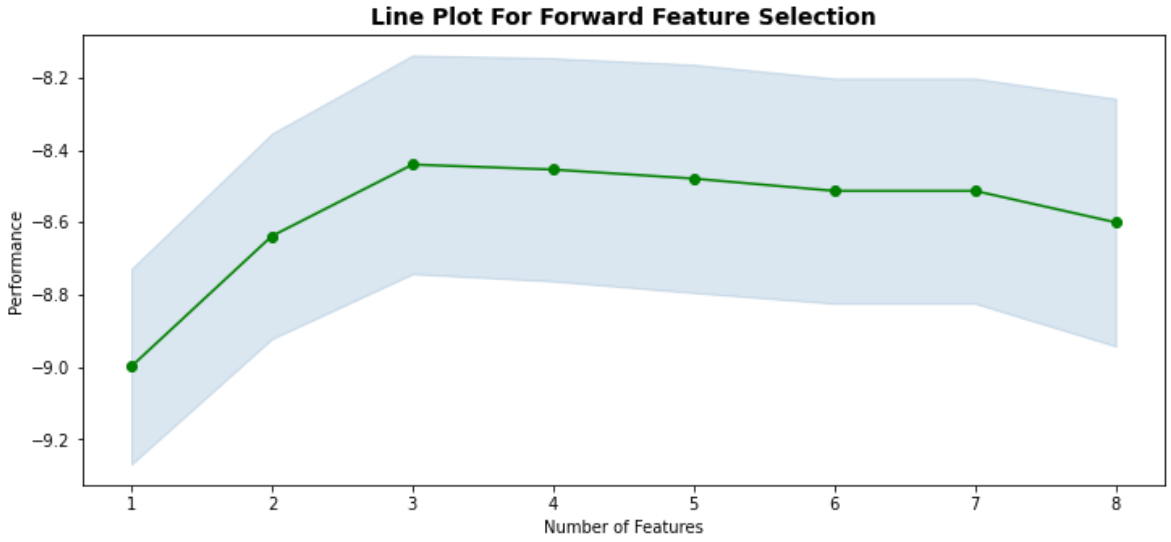


Figure 5. Forward Feature Selection Line Plot. *This shows the best feature iterations beginning with one feature all the way to eight. Three feature combinations performed best.*

The process of creating an instance of our linear regression model was done after normalization and splitting of the dataset for testing and training. The equation to predict the target variable in a linear model constitute the dot product of $\mathbf{W}$ and $\mathbf{X}$, plus b as shown in **Equation 1** (Ng, 2020).

$$f_{\mathbf{w},b}(\mathbf{x}^{(i)}) = \mathbf{w} \cdot \mathbf{x}^{(i)} + b \tag{1}$$

Where $\mathbf{W}$ and $\mathbf{X}$ are vector values of the weight and feature data, respectively, while b is the bias parameter.

The linear regression model was built both using all independent features and then selected features.

2.3 Principal Component Analysis (PCA)

We will be using PCA for dimensionality reduction by transforming the original feature variables into a new set. PCA makes use of both eigenvectors and eigenvalues, The eigenvector with the largest eigenvalues corresponds to a larger variability in the data.

After instantiating our PCA components, we will use the variance ratio to tell us about the variance between the PCA component and the original data. **Figure 6** shows the variance ratio representation with a PCA component of 3.

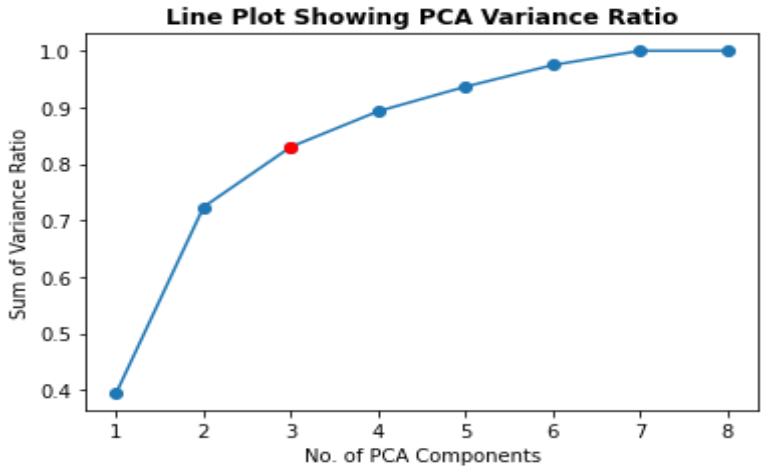


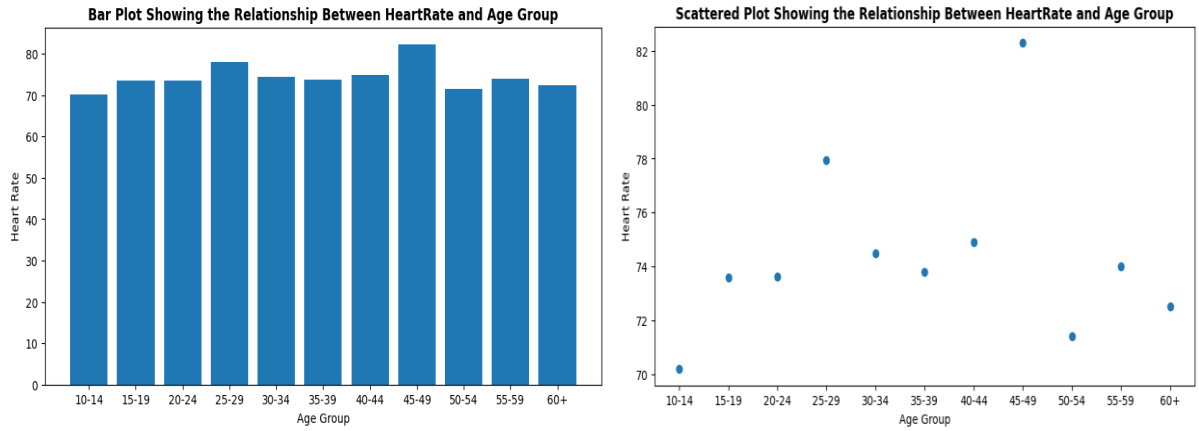
Figure 6. Line Plot Diagram of Variance Ratio with 3 PCA Components. *This plot shows the relationship between the variance ratio and the number of PCA component. With a PCA of 3 components, our variance ratio is 83% with information loss of 17%*

Table 2. Variance Ratio with 3 PCA Component. *This shows the percentage value of the information gotten from the original feature for each component and their summation.*

	PCA Component 1	PCA Component 2	PCA Component 3	Total Sum
Variance Ratio	**39.5%**	**32.9%**	**10.6%**	**83.0%**

2.4 Relationship Between Age and Heart Rate Feature Data

To find the relationship between Age and Heart Rate, we created a data frame consisting of both the age group and the mean Heart Rate per grouping. The age grouping was done using a range of 5. This gives a more informative analysis of the demographic as shown in **Figure 7.**



a) *Bar plot for age group against the mean heart rate, using an age group range of 5 with an interval of 10.*

b) Scattered plot for age group against mean heart rate, using an age group range of 5, with no linear correlation.

Figure 7. Visual Representation Between Age Group and the Mean Heart Rate. *This shows the relationship between age and the mean heart rate. We can see a uniformity between 0 to 70 bpm from the bar plot. The shattered plot tends to be more informative by showing a better distribution.*

From **Figure 7b**, we can see there is no linear relationship between age and heart rate, which means a change in heart rate is not dependent on age. According to Loerup et al. (2019), blood pressure (BP) is what is expected to have a linear correlation with heart rate.

2.5 Association Pattern Mining (APM)

For use to investigate associations between pairs of high-high, normal-normal, and low-low diastolic and systolic blood pressure, a few steps were taken:

- Created new feature groupings of both Systolic and Diastolic BP with their ranges.
- Assigned variable names for high, normal, and low for both Systolic and Diastolic BP.
- Created functions for the support, confidence, conviction, and lift association rules.

Table 3. Association Rules for each Pair. *This table shows all pair combinations of the associated rules, highlighting the diastolic against systolic values.*

APM	HS/HD	NS/ND	LS/LD	HD/HS	ND/NS	LD/LS
support	0.115613	0.333992	0.266798	0.115613	0.333992	0.266798
confidence	0.900000	0.628253	0.784884	0.423913	0.814458	0.841121
conviction	7.272727	1.586887	3.174127	1.512864	2.524383	4.154615
lift	3.300000	1.532028	2.474462	3.300000	1.532028	2.474462

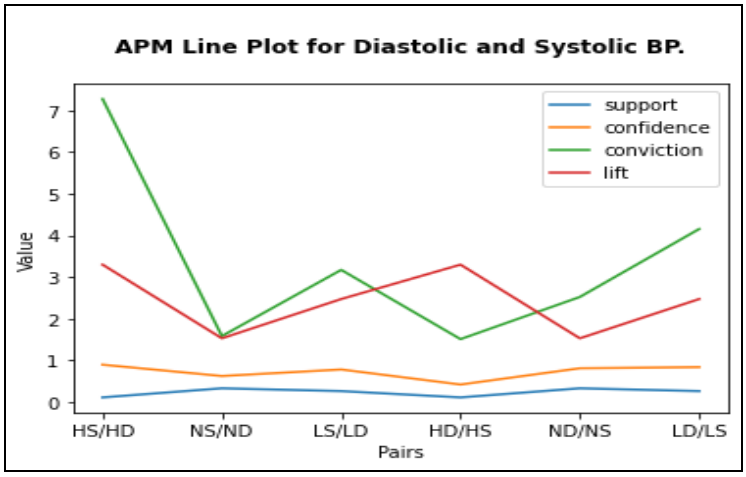


Figure 8. Line Plot Showing the Association Rules for each Pair. *This consist of both systolic against diastolic bp and vice versa, with both having similar pattern for the support and lift, while different in conviction and confidence association rule.*

2.5.1 APM Interpretation for Diastolic against Systolic

From the highlighted section in **Table 3** above, we can observe that the high-paired BP had the lowest occurrence compared to other pairs, with a support value of approximately 12%. Observing the confidence value of the same pair, 42% of patients with high diastolic BP do not have high systolic BP.

For the normal and low pairs, the support values are 33% and 27% respectively. This shows an occurrence of over one-quarter percent of our dataset. The confidence value above 80%, informs a strong relationship between systolic and diastolic BP.

Across all paired categories, the lift and conviction values were above one, which tells us there is usefulness and dependency in each pair.

2.6 Systolic BP Clustering

Since this is an unsupervised learning approach, to achieve our task of clustering patients with similar systolic BP, selecting the number of centroids is a key aspect of this task.

Two methods were used to determine the number of centroids, the Elbow Method, and the Silhouette and Davies Bouldin score comparison method as shown in **Figure 9** and **Table 4** respectively.

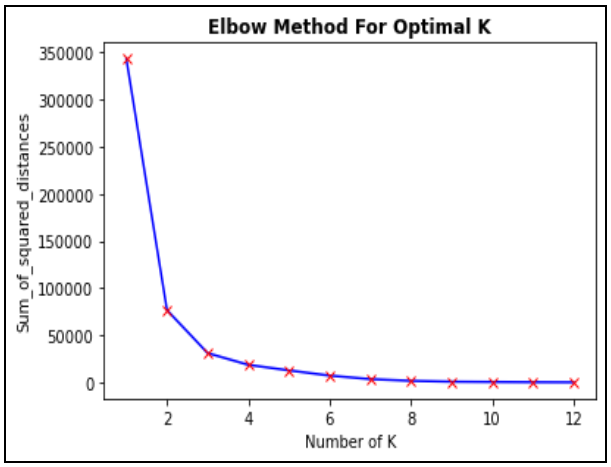


Figure 9. Elbow diagram Representing the number of clusters (K) against distance. *Where K is 3, there is no significant change in distance.*

Table 4. Silhouette and Davies Bouldin Score. *This shows the scores with cluster 2, 3 and 4. A higher Silhouette Score and lower Bouldin Score gives us the optimal cluster number.*

No. of K	silhouette_score	davies_bouldin_score
2	0.758957	0.379113
3	0.801318	0.289553
4	0.787984	0.366633

If the elbow method doesn't show a clear representation of the best cluster, the Silhouette and Davies Bouldin Score can be an alternative.

2.7 Correlation Between Age and Systolic BP

A statistical measure called the Pearson correlation coefficient was used to calculate the correlation coefficient and the p-value between Age and Systolic BP. **Table 5** shows the result.

Table 5. Correlation Coefficient and the P-Value between Age and Systolic BP. *This shows both the Correlation Coefficient and P-Value between Age and Systolic BP using Pearson Statistical Method.*

Method	Correlation Coefficient	P-Value
Pearson Correlation Coefficient	0.417	6.55×10^{-44}

From the Correlation Coefficient, there is a 42% relationship between the features. The P-value at 6.55× 10-44 shows strong evidence against the null hypothesis, interpreting that there is a linear relationship between both features since the P-value is significantly lower than the standard threshold of 0.05.

3.0 Predictions

3.1 Linear Model Prediction

For this model prediction, the results of each feature selection method including PCA were evaluated using statistical measures such as the Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2).

Table 6. Linear Model Prediction. *This includes methods used in feature selection, the number of features, and the evaluated performance.*

Method	No. Features	MAE	MSE	RMSE	R^2
Complete Features	8	8.73	114.74	10.71	0.6907
Select K Best	5	9.08	123.34	11.11	0.6675
Forward Feature Selection	**3**	**8.73**	**114.73**	**10.71**	**0.6920**
Principal Component Analysis	3	9.94	150.51	12.27	0.5942

From **Table 6**, The Forward Feature Selection is the most optimal, using only three features which are **diastolic BP, body temperature, and low-risk level**, with the lowest error across the board and highest R^2 value of 69.2%.

According to Inamdar (2021), age and systolic BP both have a strong correlation. This is also proven true in our previous analysis, although age was not considered a top feature for the best linear model.

The PCA with 3 components had the least performance. This can be attributed to the small number of features resulting in low dimensionality, also information loss from the unselected components can affect performance.

3.2 Clustering Predictions

From the scattered plot in **Figure 10**, we can see a well-separated cluster that shows a mutual relationship to the 3 levels of systolic BP which are low, normal, and high.

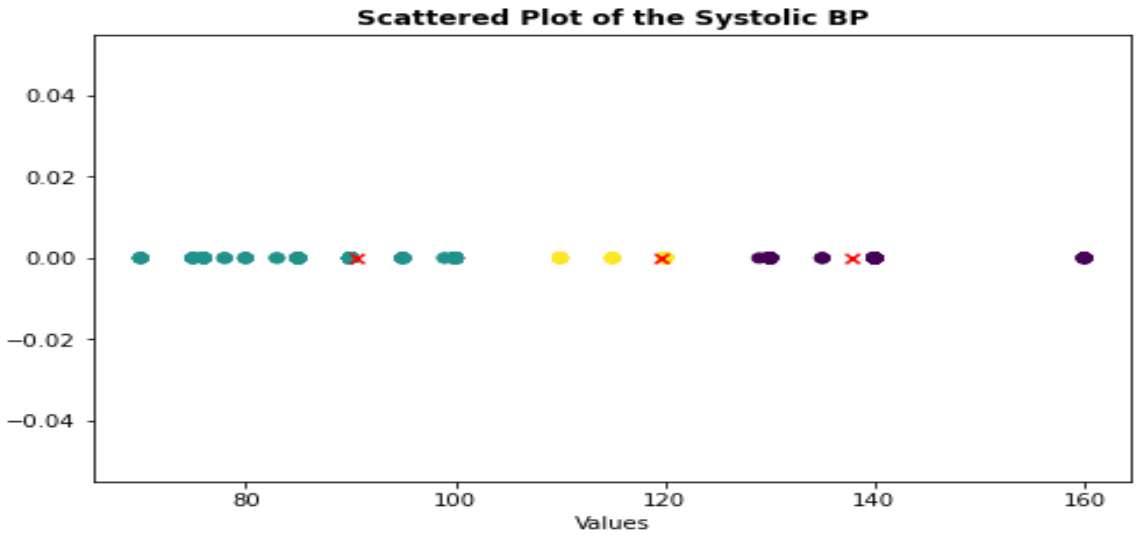


Figure 10. Scattered Plot Representing Clustering of Systolic BP. *This shows both the centroid points and the clusters in different colours, with 3 being the number of clusters.*

4.0 Recommendations

From our analysis, here are some recommendations for medical agencies:

- Prenatal Check-ups: This helps to monitor blood pressure, blood sugar levels, and body temperature which shows to be correlated with systolic bp.
- Healthy diet: This helps to maintain a low-risk level which is an important future in our analysis.
- Stress management: Deep breathing exercises and prenatal yoga can help prevent high blood pressure in maternity health due to stress.
- Patient Clustering: Grouping patients by their BP level can improve promptness to medical attention.

5.0 Conclusion

In conclusion, we successfully analysed the relationship of Systolic Blood Pressure to maternal health, it was found that further research is needed to investigate the features that affect systolic BP. Inamdar (2021), and Zhou et al. (2021) showed that systolic BP has a strong correlation to Age and BS respectively, but these features had no impact on our top-performing linear model.

Reference

British Heart Foundation (2019) *Heart Rate Bhf.org.uk*. British Heart Foundation. Available online: https://www.bhf.org.uk/informationsupport/how-a-healthy-heart-works/your-heart-rate.

Inamdar, A. (2021) Abstract 10023: Association of Systolic and Diastolic Blood Pressure with Age in Rural Indians. *Circulation*, 144(Suppl_1). Available online: https://doi.org/10.1161/circ.144.suppl_1.10023.

Kay, C. (2020) *How Old Is Too Old to Have a Baby? Healthline*. Available online: https://www.healthline.com/health/pregnancy/how-old-is-too-old-to-have-a-baby.

Loerup, L., Pullon, R.M., Birks, J., Fleming, S., Mackillop, L.H., Gerry, S. & Watkinson, P.J. (2019) Trends of blood pressure and heart rate in normal pregnancies: a systematic review and meta-analysis. *BMC Medicine*, 17(1). Available online: https://doi.org/10.1186/s12916-019-1399-1.

National Health Service (2019) *What is blood pressure? NHS*. Available online: https://www.nhs.uk/common-health-questions/lifestyle/what-is-blood-pressure/.

Ng, A. (2019) *Machine Learning | Coursera Coursera*. Available online: https://www.coursera.org/learn/machine-learning.

Sisense (2019) *What is Data Cleaning? | Sisense Sisense*. Sisense. Available online: https://www.sisense.com/glossary/data-cleaning/.

Zhou, Z., Deng, C. & Xiang, X. (2021) Blood glucose related to pregnancy induced hypertension syndrome. *American Journal of Translational Research*, 13(5), 5301–5307. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8205659/ [Accessed 17/April/2022].