MATERNAL HEALTH ANALYSIS AND PREDICTION

1.0 Introduction

The World Health Organization (2019) describes maternal health as women's health during and after pregnancy. In 2020, a woman died every two minutes due to maternity related causes; however, this is avoidable following adequate care by healthcare professionals (World Health Organization, 2020).

Presented with data of 1,014 women with information on their Age, Systolic and Diastolic Blood Pressure, Blood Sugar, Body Temperature, Heart Rate and Risk Level, this report aims to analyze factors that could adversely affect maternal health, with systolic blood pressure as a reference, and provide recommendations to healthcare providers to improve pregnancy outcomes.

2.0 Analysis

This section outlines the steps taken to complete the tasks.

2.1 The Dataset

The dataset contained 1,014 rows and 7 columns – *'Age'*, *'SystolicBP'*, *'DiastolicBP'*, *'BS'* – *Blood Sugar*, *'BodyTemp'*, *'HeartRate'*, *'RiskLevel'* (high, low, and mid risk). All variables apart from the 'RiskLevel' are numerical.

2.2 Exploratory Data Analysis (EDA)

EDA is done to prepare the data for modelling and to gain important insights.

2.2.1 Data Cleaning

The data was examined for null, empty, and duplicated values. No null or empty values were found; however, 562 duplicated values were present. The duplicates were not dropped because individual rows lacked unique identifiers like *'Patient ID'* or *'Patient Name'* to suggest that the values are duplicates. Also, different patients can have similar medical results.

2.2.2 Statistical Overview

The statistical overview showed a minimum heart rate of 7.0; a maximum SystolicBP of 160.0; and a minimum and maximum age of 10 and 70 respectively. This heart rate value was dropped from the dataset as a healthy human's average heart rate is between 60 – 100 which could increase by 10% in pregnant and post-partum women (British Heart Foundation, 2022; Green et al., 2021; Loerup et al., 2019).

As most women have a SystolicBP of between 100 and 120, a maximum value of 160 stands out as an outlier.

Pregnancy at 10 and 70 years, though unlikely is not impossible. Although the average age at Menarche is 12 years, it differs for different women and could start earlier (Canelón and Boland, 2020). In addition, research has shown that the human uterus can support gestation past the age of menopause (Paulson et al., 1997). These values were left in the dataset based on these findings.

Table 2.1: Statistical Overview of the Dataset. *This table shows the key statistical features of the data. The statistical overview can be used to quickly point out anomalies and outliers in a dataset. It was observed that the minimum heart rate is 7.0; the minimum and maximum age are 10 and 70 respectively; and the maximum Systolic BP is 160.0.*

	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

2.2.3 Univariate and Bivariate Analysis

Univariate and bivariate analysis was performed to gain insights into the data.

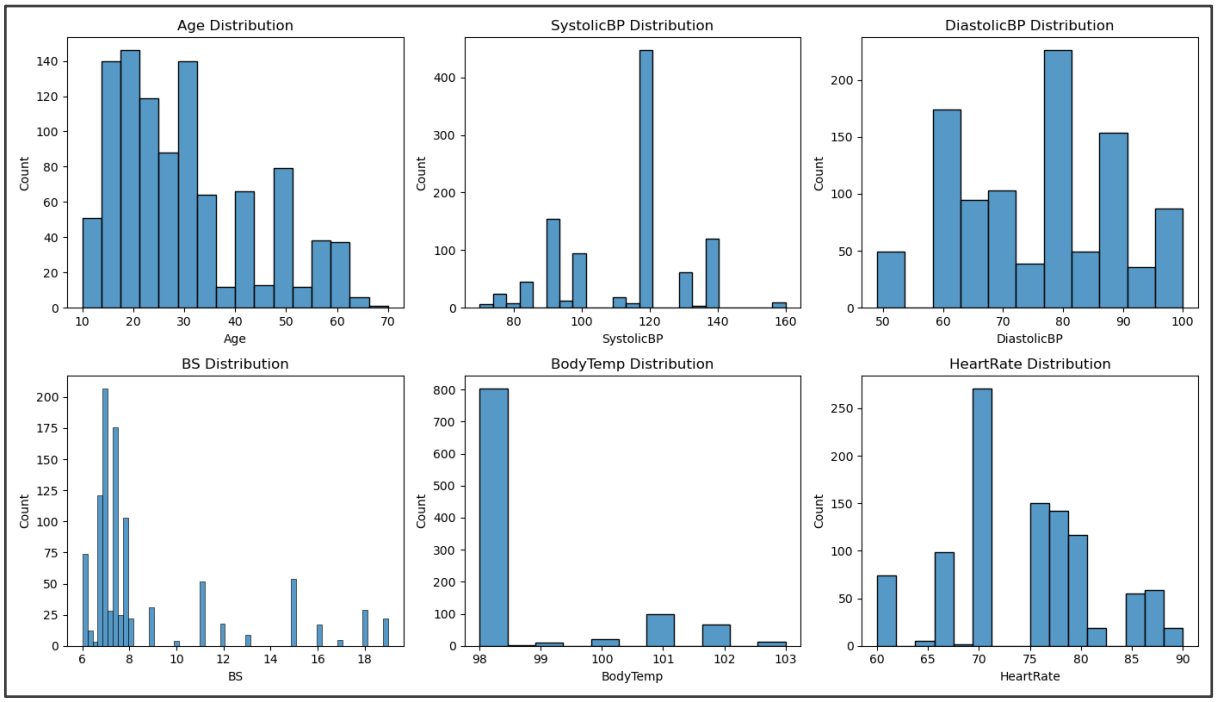

Figure 2.1: Distribution of the numerical variables. This figure shows the univariate analysis of the numerical variables. *It shows that the HeartRate, SystolicBP and DiastolicBP variables are slightly normally distributed, while Age, BS, BodyTemp are right-skewed.*

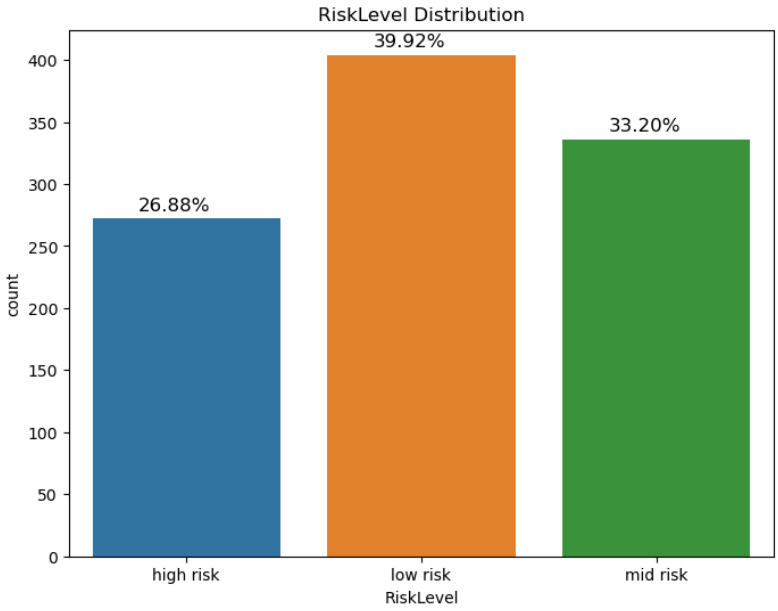

Figure 2.2: Distribution of Risk Levels in the Dataset. *This figure shows the percentage distribution of the risk levels in the data. 26.88% of women are at high risk, 39.92% are at low risk while 33.20% are at mid risk.*

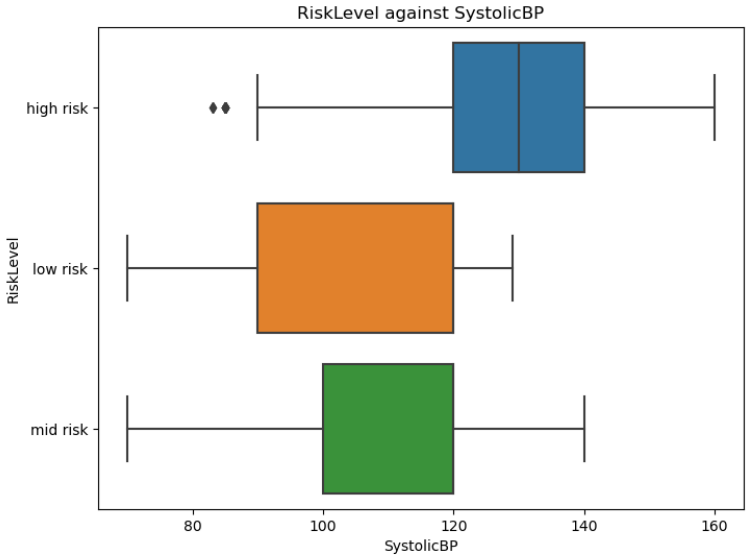

Figure 2.3: Boxplot showing RiskLevel against Systolic BP. *This figure shows the systolic blood pressure ranges for the different risk levels. Women at high risk mostly have a systolic blood pressure of between 120 – 140, low risk is between 90 – 120, while mid risk is between 100 and 120.*

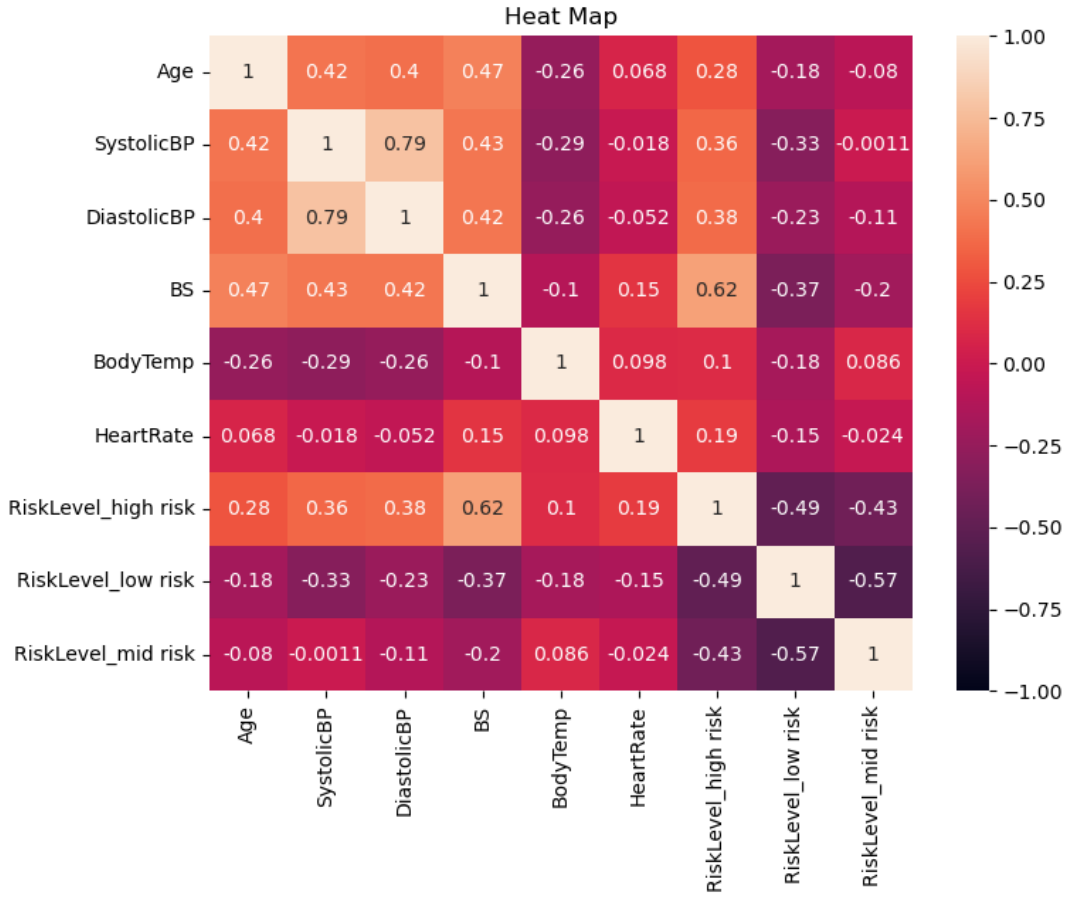

Figure 2.4: Heat Map Showing the Correlation between Variables. This figure shows the correlation between the variables in the dataset. *SystolicBP and DiastolicBP have a high positive correlation which means that as one increases, the other also increases. Age and Blood Sugar have a moderate positive correlation with SystolicBP.*

2.3 Linear Regression Model

A linear regression model was built with SystolicBP as the response variable. The data was split with the 80% training and 20% testing sharing formula and normalized using the StandardScaler technique. Feature selection was performed to select the best variables for the model using the K-best algorithm and the ***'f_regression'*** scoring function. The 'HeartRate' and 'RiskLevel_mid risk' scored the lowest as seen in **Fig 2.5**.

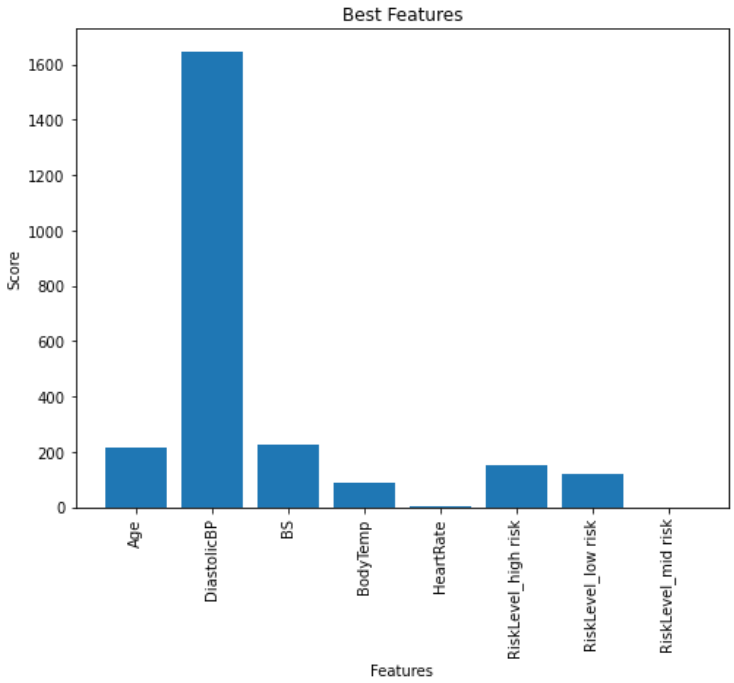

Figure 2.5: Feature Selection using K-best Algorithm. *This figure shows a bar plot of the features and their scores. DiastolicBP had the highest score. HeartRate and RiskLevel_mid risk had the lowest scores. This corroborates information obtained from the Heatmap.*

Table 2.2: Model Coefficients. *The table below shows the coefficients the model assigned to each of the variables. The aim of linear regression is to provide estimates of coefficients that minimize the errors between predicted and actual values. The formular is* ***Y = β0 + β1X1 + β2X2 + ... + βn*Xn + ε*** *where Y is the response variable; β0 is the intercept; β1, β2, βn are the coefficients; X1, X2, Xn are the independent variables; and ε is the error term. Using Bloo dSugar as an example, it can be interpreted that a unit increase in Blood Sugar increases the SystolicBP by 0.60996 assuming other variables remain constant. A negative coefficient causes a reduction in the value of the response variable.*

	Age	DiastolicBP	BS	BodyTemp	RiskLevel_high risk	RiskLevel_low risk
0	1.572805	12.364172	0.60996	-1.78219	-0.470826	-3.096335

The Forward Feature Selection method was also used with the scoring hyperparameter optimized for a lower Mean Absolute Error (MAE) value. **Figure 3.6** shows the MAE improved after three features were added and then started to decline. The three features identified were ***'DiastolicBP'***, ***'BodyTemp'*** and ***'RiskLevel_low risk'***.

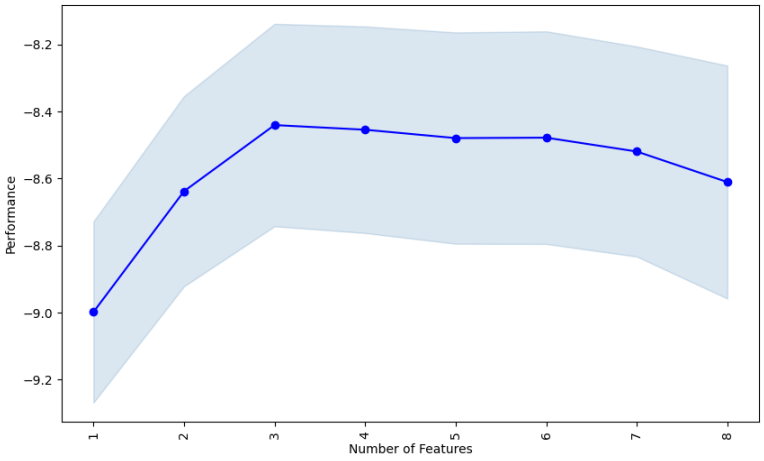

Figure 2.6: Forward Feature Selection Plot. *This figure shows a line plot of the features and their performance for the MAE scoring optimization. The performance (MAE) value reduces up the y_axis hence the negative values.*

2.4 Principal Component Analysis (PCA)

PCA is used for dimensionality reduction for large datasets. It summarizes the important features in the data and loses some information in the process. The Explained Variance Ratio (EVR) is the percentage of variation in the data explained by the principal components, and it informs the optimal number of principal components. The first principal component explains the largest variance, followed by the second principal component and so on. Three principal components were used to rebuild the model.

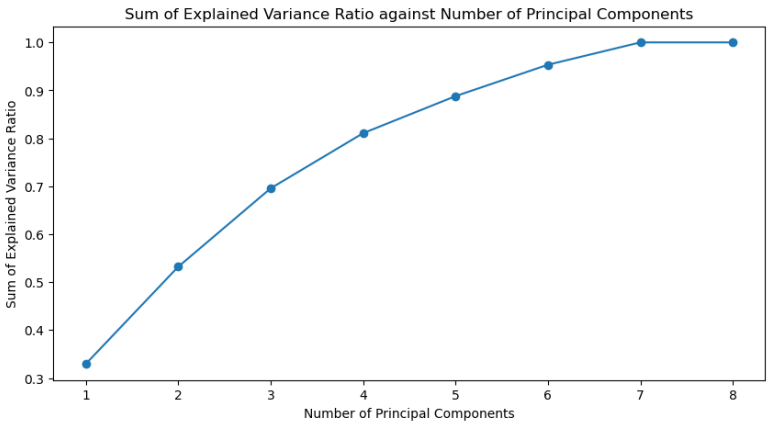

Figure 2.7: Sum of Explained Variance Ratio for the Principal Components. *The line plot shows the summed EVR for each principal component. Three principal components explained about* ***70%*** *of the variation in the data.*

2.5 Relationship between Age and Heart Rate

Age was grouped in bands of five to ensure consistency and aid comparison across different domains. The grouping started from 0 – 14 which falls under the average age of menarche and ends at 60 which is the age most women would have reached menopause (de Kat et al., 2019).

From the data, it is observed that there is no correlation between age and heart rate. This is further confirmed by the heatmap which showed a correlation of **0.06** between the variables. An almost balanced average heart rate range between all age groups was observed as seen in **Figure 2.9**.

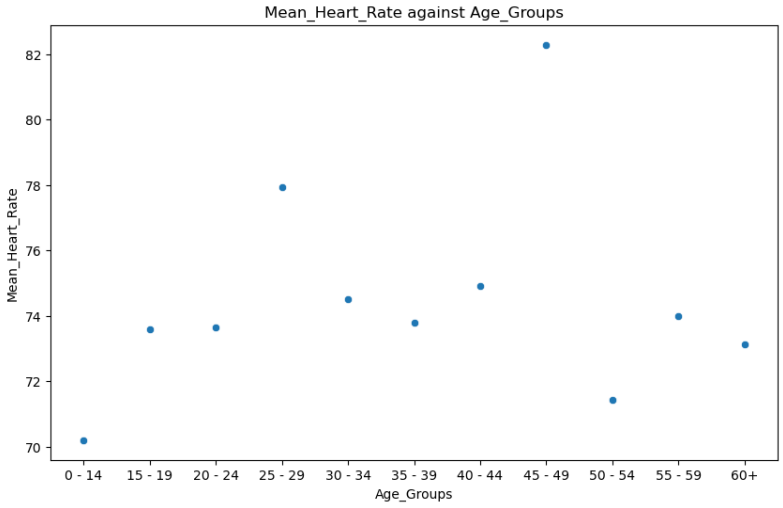

Figure 2.8: Scatterplot Showing Mean HeartRate against Age Groups.
No correlation is observed between mean heart rate and the eleven age groups.

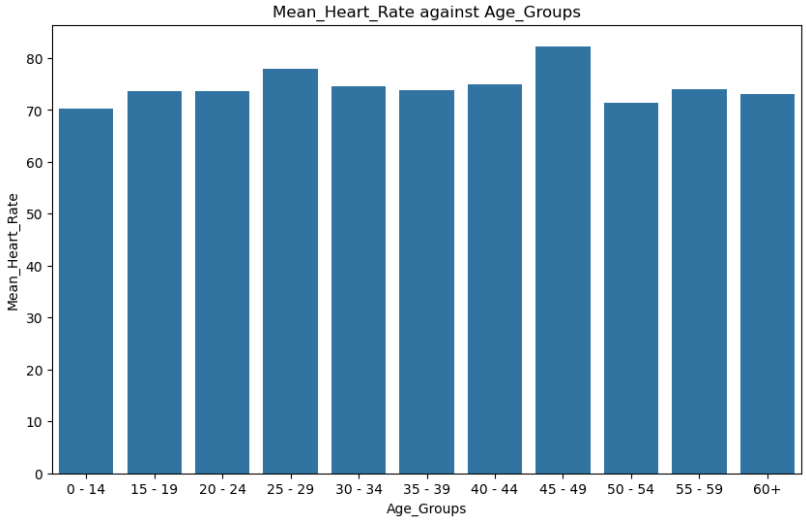

Figure 2.9: Barplot Showing Mean HeartRate against Age Groups.
An almost balanced mean heart rate is observed across the age groups.

2.6 Associations Between Blood Pressure Pairs

Association rules describe the relationship between two or more items in a dataset. They consist of the antecedent – items in the data that act as the condition for the rule, and the consequent – the inferred items based on the presence of the antecedent in the data.

Support, Confidence, Lift, and Conviction are discussed and were calculated using the apriori and association rules algorithms present in the mlxtend frequent patterns library. Two functions were used to group the blood pressure values into *'high'*, *'normal'* and *'low'* based on the specified ranges.

Table 2.3: Association Rule Metrics for Blood Pressure Pairs. *The table below shows the results of the association rule metrics for high/high, low/low, and normal/normal DiastolicBP and SystolicBP readings. It assumes DiastolicBP as the antecedent while SystolicBP is the consequent.*

	Support	Confidence	Lift	Conviction
High Diastolic / High Systolic **(HD/HS)**	0.12	0.42	3.30	1.51
Normal Diastolic / Normal Systolic **(ND/NS)**	0.33	0.81	1.53	2.52
Low Diastolic / Low Systolic **(LD/LS)**	0.27	0.84	2.47	4.15

The **HD/HS** pair occurred in 12% of the dataset (support of 0.12). It had a confidence of 42% meaning that 42% of high diastolic occurrences also had a high systolic occurrence; a strong lift of 3.30 meaning that the presence of high diastolic increases the likelihood of having high systolic by 3.30 times; and a conviction of 1.51 meaning that the relationship between the two variables have a strong dependency. The **ND/NS** and **LD/LS** pairs occurred in 33% and 27% of the data respectively, with high confidence values of 80% and 84% respectively, and strong lift and conviction values.

2.7 Clusters of Patients with Similar SystolicBP

Clustering is an unsupervised learning technique that groups data into clusters based on their similarities to each other, with each cluster being distinct. K-means clustering was used, and the optimal number of clusters obtained by computing the inertia which measures the effectiveness of the clusters. Three clusters were computed with a high silhouette score of **0.8** showing high similarity within clusters and a good inter-cluster separation.

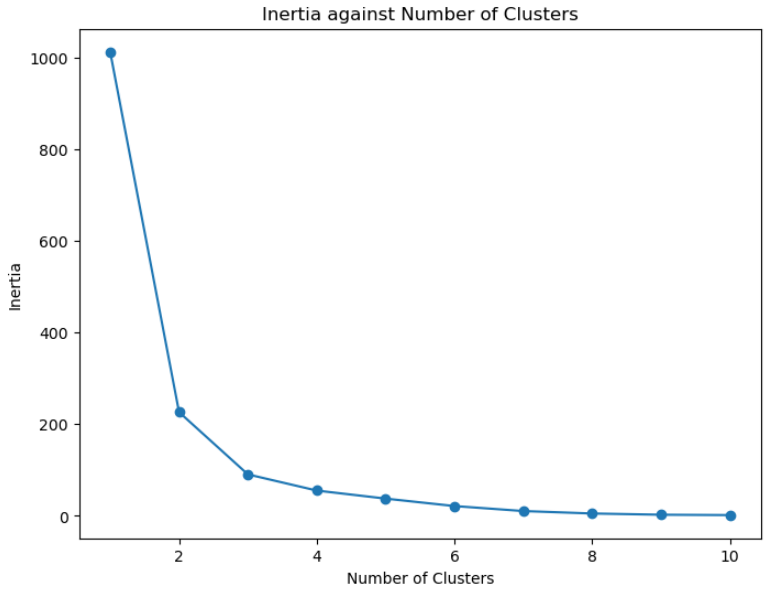

Figure 2.10: Elbow Curve Showing Inertia Against Number of Clusters. *This figure shows the elbow curve which is used to select the optimal number of clusters. 3 is optimal as it specifies the point where the rate of decrease in inertia slows down significantly.*

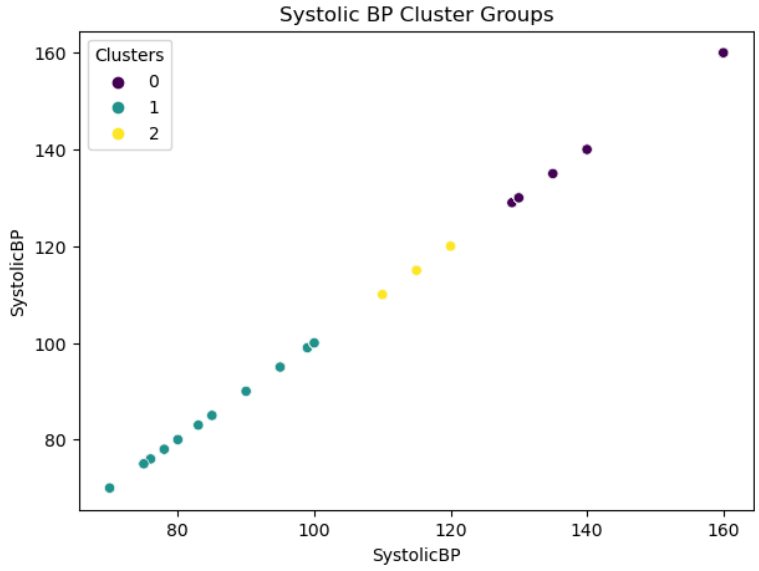

Figure 2.11: Plot of SystolicBP Showing the Clusters. *This figure shows three distinct clusters for SystolicBP from the dataset.* ***Cluster 0*** *showed readings between 129 – 160,* ***Cluster 1*** *showed readings between 70 – 100, and* ***Cluster 2*** *showed readings between 110 – 120.*

2.8 Correlation between Age and SystolicBP

It was observed that age had a moderate positive correlation (**0.42**) with systolic BP. The p-value computed from the regression analysis was **6.56 x 10^-44**.

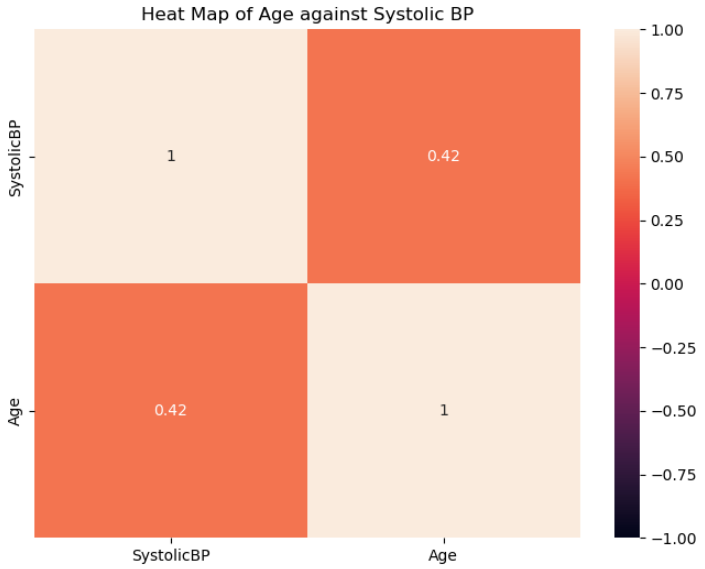

Figure 2.12: Heat Map Showing the Correlation between Age and SystolicBP. *This figure shows a moderate positive correlation of 0.42.*

OLS Regression Results

Dep. Variable:	SystolicBP	R-squared:	0.174
Model:	OLS	Adj. R-squared:	0.173
Method:	Least Squares	F-statistic:	213.0
Date:	Thu, 11 May 2023	Prob (F-statistic):	6.56e-44
Time:	23:43:43	Log-Likelihood:	-4287.0
No. Observations:	1012	AIC:	8578.
Df Residuals:	1010	BIC:	8588.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	96.1280	1.282	74.987	0.000	93.612	98.643
Age	0.5705	0.039	14.593	0.000	0.494	0.647

Omnibus:	24.212	Durbin-Watson:	1.955
Prob(Omnibus):	0.000	Jarque-Bera (JB):	20.412
Skew:	-0.277	Prob(JB):	3.69e-05
Kurtosis:	2.579	Cond. No.	79.9

Figure 2.12: Regression Analysis Showing P-value. *This regression analysis investigates the relationship between Age and Systolic BP. The computed p-value which is very small (6.56e-44) rejects the null hypothesis which assumes no relationship between the variables exist. This points to a definite relationship between age and Systolic BP.*

3.0 Prediction

The models were evaluated using the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Coefficient of Determination (r^2) and Adjusted r^2 metrics.

Table 3.1: Linear Regression Evaluation Metrics. The table below shows the results from the linear regression models built using different feature selection techniques.

Technique	Features Count	MSE	RMSE	MAE	MAPE	r^2	Adjusted r^2
K_best	8	114.76	10.71	8.73	7.96	0.69	0.68
Forward Feature Selection	**3**	**114.24**	**10.69**	**8.59**	**7.83**	**0.69**	**0.69**
PCA	3	116.76	12.72	10.48	7.69	0.56	0.56

The forward feature selection technique produced the best results. PCA performed poorly, which could be due to further reduction in the data dimension resulting in information loss and a lower explained variance. This can be seen in the reduced r^2 value. It could also point to low correlation observed between the independent variables and the response variable.

The K-best algorithm showed DiastolicBP, Age and Blood Sugar as determinants for predicting SystolicBP; while the Forward Feature selection showed DiastolicBP and BodyTemp which was corroborated by Xiong et al., 2020.

A high diastolic blood pressure is closely associated with a high systolic blood pressure as observed from the data. This is a sign of hypertension which is a risk factor for pregnancy complications like pre-eclampsia and accounts for 18% of maternal deaths and 3% of pregnancy complications. Pre-eclampsia is further complicated if a mother is over 40 years (El Abasse, 2020; Hemapriya .L, Bhandiwad and Desai, 2020).

4.0 Recommendations

To ensure successful pregnancy and childbirth outcomes, the following recommendations are provided:

- Regular antenatal visits to monitor blood pressure, blood sugar, heart rate and other vital signs as these reading may change during pregnancy
- Older mothers are at higher risk of developing pre-eclampsia and should be monitored closely
- Regular non-strenuous exercise is advised to regulate heart rate and blood pressure levels
- A balanced diet is advised to maintain blood sugar levels and the avoidance of extremely cold or warm weather which can adversely affect blood pressure.

5.0 Conclusion

This report focused on predicting systolic blood pressure and determining variables that affect it. The relationship between independent variables in the data was also analysed to derive insights for improved maternal health outcomes. The best performing model highlighted diastolic blood pressure and body temperature as important factors which has been proven by existing literature. This research can be expanded to classify maternal mortality risk levels based on related risk factors.

Predictions could be improved with the addition of more data and inclusion of variables that show a higher correlation with the response variable. The absence of this negatively affected the PCA results. From the systolic blood pressure predictions, future study can focus on hypertension which has been recognized as an important maternal health risk factor.

BIBLIOGRAPHY

British Heart Foundation (2022). *Your Heart Rate*. [online] Bhf.org.uk. Available at: https://www.bhf.org.uk/informationsupport/how-a-healthy-heart-works/your-heart-rate#NORM [Accessed 9 May 2023].

Canelón, S.P. and Boland, M.R. (2020). A Systematic Literature Review of Factors Affecting the Timing of Menarche: The Potential for Climate Change to Impact Women's Health. *International Journal of Environmental Research and Public Health*, [online] 17(5), p.1703. doi:https://doi.org/10.3390/ijerph17051703.

de Kat, A.C., van der Schouw, Y.T., Eijkemans, M.J.C., Broer, S.L., Verschuren, W.M.M. and Broekmans, F.J.M. (2019). Can Menopause Prediction Be Improved with Multiple AMH Measurements? Results from the Prospective Doetinchem Cohort Study. *The Journal of Clinical Endocrinology & Metabolism*, 104(11), pp.5024–5031. doi:https://doi.org/10.1210/jc.2018-02607.

El Abasse, Z. (2020). Peculiarity of Hypertension and Pregnancy about 544 Cases. *Archives of Cardiovascular Diseases Supplements*, [online] 12(1), p.138. doi:https://doi.org/10.1016/j.acvdsp.2019.09.394.

Green, L.J., Pullon, R., Mackillop, L.H., Gerry, S., Birks, J., Salvi, D., Davidson, S., Loerup, L., Tarassenko, L., Mossop, J., Edwards, C., Gauntlett, R., Harding, K., Chappell, L.C., Knight, M. and Watkinson, P.J. (2021). Postpartum-Specific Vital Sign Reference Ranges. *Obstetrics & Gynecology*, [online] 137(2), pp.295–304. doi:https://doi.org/10.1097/aog.0000000000004239.

Hemapriya .L, Bhandiwad, A. and Desai, N. (2020). Maternal Complications of Hypertension in Pregnancy – A Five Year Study. *Indian Journal of Obstetrics and Gynecology Research*, [online] 5(3), pp.349–352. doi:https://doi.org/10.18231/2394-2754.2018.0080.

Loerup, L., Pullon, R.M., Birks, J., Fleming, S., Mackillop, L.H., Gerry, S. and Watkinson, P.J. (2019). Trends of Blood Pressure and Heart Rate in Normal Pregnancies: A Systematic Review and Meta-analysis. *BMC Medicine*, [online] 17(1). doi:https://doi.org/10.1186/s12916-019-1399-1.

Paulson, R.J., Thornton, M.H., Francis, M.M. and Salvador, H.S. (1997). Successful Pregnancy in a 63-year-old Woman. *Maturitas*, 28(1), pp.96–97. doi:https://doi.org/10.1016/s0378-5122(97)13268-6.

World Health Organization (2019). *Maternal Health*. [online] Who.int. Available at: https://www.who.int/health-topics/maternal-health#tab=tab_1 [Accessed 9 May 2023].

World Health Organization (2020). *World Health Statistics 2020: Monitoring Health for the SDGs, Sustainable Development Goals*. [online] Available at: https://digitalcommons.fiu.edu/srhreports/health/health/28/ [Accessed 9 May 2023].

Xiong, T., Chen, P., Mu, Y., Li, X., Di, B., Li, J., Qu, Y., Tang, J., Liang, J. and Mu, D. (2020). Association between Ambient Temperature and Hypertensive Disorders in Pregnancy in China. *Nature Communications*, [online] 11(1), pp.1–11. doi:https://doi.org/10.1038/s41467-020-16775-8.