**A Maternal Health Dataset Analysis on Potential Risk Factors for Pregnancy Problems.**

**Introduction:**

Pregnancy and delivery-related issues are women's most common cause of death globally (World Health Organization, 2019). Maternal health is a crucial topic for healthcare providers to concentrate on because there is a need to detect potential risk factors for problems during pregnancy.

This report analyses a maternal health dataset to identify potential risk factors for pregnancy problems. It includes physiological measurements like Age, blood pressure, blood sugar, body temperature, heart rate, and risk level. The report aims to answer questions regarding the distribution of variables, relationships between variables, and predicting the risk level during pregnancy.

**Literature Review**

The risk of developing preeclampsia is significantly higher in women with prior hypertension records during pregnancy or maternal illness as reported by Fox et al. in their study published in 2019. Moreover certain parameters like nulliparity status along with an age below forty years and BMI less than thirty five kilograms per square meter can categorize an individual into moderate risk status along with family history and multiple pregnancies or conception delay beyond ten years.

Pregnancy complications leading to maternal mortality are critical issues globally, especially in low-income countries. Risk reduction during pregnancy has been suggested due to predictive modelling using machine learning. Studies have found mixed outcomes that utilised machine learning to forecast issues and choose delivery methods.

Using IADPSG (International Association of Diabetes in Pregnancy Study Group) criteria rather than WHO 1999 criteria, Bhavadharini et al. (2016) discovered a greater prevalence of gestational diabetes mellitus (GDM) in urban regions, indicating the value of early detection and therapy of GDM. Urban areas have greater rates of GDM prevalence, highlighting the need for both urban and rural populations in India to receive early detection and treatment of GDM.

By identifying risk factors and suggesting efficient treatments to improve maternal and foetal health outcomes in response to the high prevalence of maternal fatalities worldwide, the study seeks to contribute to maternal health research.

The following sections of this study will include a review of the database, visualisations and statistical calculations to identify possible risk indicators. We will also develop prediction models using machine learning techniques to estimate a woman's risk level during pregnancy.

**Methodology**

The following methodology was used to build and fit a linear model for a report on the relationship between Systolic BP and exploratory variables:

1. Data collection: Relevant data on Systolic BP, Age, DiastolicBP, and BS were obtained from the chosen source.
2. Data cleaning: The collected data were carefully reviewed to remove any inconsistencies, missing values, or outliers that could affect the model's accuracy (Peng, 2021).
3. Exploratory data analysis: Basic statistics and visualisations were used to explore the relationships between Systolic BP and the exploratory variables (Gelman et al., 2020).
4. Model selection: Based on the exploratory data analysis, a multiple linear regression model was chosen to analyse the relationship between Systolic BP and Age, DiastolicBP, and BS (Muller & Guido, 2017).
5. Model development: The exploratory variables were defined as predictors and a constant term was added to the model. The model was then fit using Ordinary Least Squares (OLS) regression analysis (Wooldridge, 2019).
6. Model evaluation: The R-squared value, coefficient estimates, p-values, and other pertinent statistics were used to assess the fitted model's goodness of fit, accuracy, and statistical significance.
7. Interpretation of results: The last stage was to assess the model analysis findings and make judgements about how Systolic Blood Pressure related to the exploratory variables.

This methodology was used to build and fit a linear model to investigate the relationship between Systolic BP and Age, DiastolicBP, and BS.

The findings from this analysis will be presented in the report to provide insights and recommendations for healthcare professionals.

Exploratory data analysis:

Data analysis and visualisations were conducted to address research objectives and understand risk factors and interactions between variables in a maternal health dataset. The dataset contains information on pregnant women's ages, blood pressure, blood sugar levels, body temperatures, heart rates, and risk factors.

Descriptive Statistics:

Table

Description automatically generated

Fig 1.1: The pandas' DataFrame function data. Describe () is used to obtain a summary of statistical information for each column in the dataset. The dataset contains 1014 entries, each corresponding to a patient and including demographic information like Age and clinical measurements like SystolicBP, DiastolicBP, BS, body temp, Risk Level, and heart rate. The RiskLevel column shows the level of risk associated with each patient as determined by a medical professional. Analysing the summary statistics from data. Describe () can help identify any anomalies or unusual trends in the data and provide valuable insights into each variable's distribution and range of values.

Age VS Systolic Blood Pressure:

Chart, scatter chart

Description automatically generated

Fig 1.2: The relationship between age and systolic blood pressure is depicted via a scatter plot. The graphic suggests that age and systolic blood pressure may be positively correlated.

Diastolic Blood Pressure Distribution:

Chart, bar chart, histogram

Description automatically generated

Fig 1.3: A histogram displaying the dataset's Diastolic Blood Pressure values distribution. The plot has 20 bins; the x-axis represents Diastolic Blood Pressure values, while the y-axis represents the frequency of occurrence of each value. The title of the plot is "Distribution of Diastolic Blood Pressure".

Distribution of Blood Sugar Levels:

Chart, box and whisker chart

Description automatically generated

Fig 1.4: A boxplot showing the distribution of the dataset's Blood Sugar (BS) column. The plot provides information on the minimum and maximum values and any outliers or unusual patterns in the data.

Body Temperature vs Heart Rate:

Chart, scatter chart

Description automatically generated

Fig 1.5: A scatter plot showing the relationship between Body Temperature and Heart Rate. The plot displays a positive correlation between Body Temperature and Heart Rate.

According to Medical News Today (2022), a healthy person should have systolic and diastolic blood pressure between 70 and 90 mmH and 110 and 140 mmHg, respectivelv. In our study, we found that pregnant women's blood pressure averaged within 100 and 140 mmg for systolic values and 60 to 90 mmH for diastolic values. Blood pressure exceeding 140 (systolic) or 90 mmH (diastolic) indicates high blood pressure.

According to the study results, almost all expectant women maintained an average body temperature equivalent to what Naila et al.'s previous investigation found- around 98 degrees Fahrenheit. The data suggests that having such an elevated body temperature does not pose significant risks during pregnancy. Besides this finding, it was also observed by MediLexicon International in their report published in 2023 that majority of pregnant individuals recorded normal adult-level heart rates. But it is important to acknowledge variables such as age and level of activity which can influence heart rates' potential role as a predictor for pregnancy-related issues.

These results offer helpful information for identifying potential risk factors for problems during pregnancy overall. While monitoring and evaluating the health of pregnant patients, healthcare professionals can use these ranges as a guide.

**Interpretation of results**

As our exploratory variables, we used Age, Diastolic Blood Pressure, and Blood Sugar (BS), as the correlation matrix indicates that these variables significantly correlate with Systolic Blood Pressure compared to others.

Chart

Description automatically generated

Fig 1.6: Is a heatmap showing the connections between variables. The variables include Age, SystolicBP, DiastolicBP, BS, BodyTemp, heart rate, and risk level. The heatmap indicates the strength and direction of the correlation between these variables, with warmer colours representing stronger positive correlations and more excellent colours representing stronger negative correlations.

The Mean Absolute Error (MAE) a measure of the magnitude of error in predictions, shows that on average, the model's predictions for systolic bp are off by 8.883 units. In contrast, the Root Mean Squared Error, which measures model accuracy indicates the model is off by 11.096 units. Lastly the as indicated by the R-squared value of 0.636, the model explains 64% of the changes in systolic BP.

A screenshot of a computer

Description automatically generated with low confidence

Table 1.1: summarises the multiple linear regression model using age, diastolic blood pressure, and blood sugar levels as independent variables to forecast systolic blood pressure. The model contains Mean Absolute Errors, Root Mean Squared Error and R-squared.

The model has a reasonably good fit overall with a high R-squared value and low error.

Principal Component Analysis:

Based on the PCA analysis, We performed PCA on a feature matrix with seven features and obtained five principal components.

|  |  |  |
| --- | --- | --- |
| **Number\_of\_components** | **PCA\_Variance\_Ratio** | **Cummulative\_Percentage** |
| 1 | 35.08% | 40.11% |
| 2 | 18.39% | 61.14% |
| 3 | 15.91% | 79.33% |
| 4 | 10.23% | 91.02% |
| 5 | 7.85% | 100.00% |
| **Total** | **87.46%** |  |

Table 1.2: Summary of principal component analysis (PCA) results showing the percentage of variance explained by each number of components and the cumulative percentage of variance explained. The results indicate that using five components explains 87.46% of the total variance.

We considered the desired trade-off between explained variance and dimensionality reduction to determine the optimal number of principal components to retain. A commonly used rule of thumb is to retain principal components that explain a total variance of at least 80% (Smith,2002). In this case, the first three principal components explain a variance of 0.6938 or 69.38%. Therefore, we might consider retaining only the first three principal components for further analysis, which would significantly reduce the dimensionality of the feature matrix from 7 to 3.

Chart, line chart

Description automatically generated

Fig 1.7: An elbow plot generated using the PCA method to determine the optimal number of components to include in the model. The x-axis represents the number of components, and the y-axis represents the explained variance ratio. The plot shows a decreasing trend in the explained variance ratio with an increasing number of components. The elbow joint of the plot indicates the optimal number of components for the model.

Age Distribution Analysis:

To examine the association between age and heart rate, the data were divided into age intervals using the quartiles of the age distribution. The age intervals range from 10–19, 20-29, 30-39,40-49, 50-59, and 60-70. With each interval holding nearly the same amount of observations, this choice of intervals enables a balanced representation of the age groups. Additionally, it offers a helpful way to divide the data into age brackets that cover a wide range of ages while providing adequate detail to examine any age-related variations in heart rate.

Chart, line chart

Description automatically generated

Fig 1.8 shows a line plot representing the relationship between age groups and mean heart rate. The plot displays a positive correlation between Age and mean heart rate, except for the 50-59 age group with the lowest mean heart rate. The plot is titled "Mean Heart Rate by Age Group" and includes marker points and grid lines for better visualization.

Association Rule Mining Apriori Algorithm:

The table below summarises the relationship rule mining between systolic and diastolic blood pressure categories.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Index** | **Category** | **Support** | **Confidence** | **Conviction** | **Lift** |
| 0 | High/High | 0.115385 | 0.9 | 8.846154 | 7.8 |
| 1 | Normal/Normal | 0.335306 | 0.62963 | 1.794675 | 1.877778 |
| 2 | Low/Low | 0.266272 | 0.784884 | 3.410843 | 2.947674 |

Table 1.2: Association rule mining results for a dataset. The table shows support, confidence, conviction, and lift values for High/High, Normal/Normal, and Low/Low. The High/High category has the highest confidence and conviction values, indicating a strong association between the two items in this category. The Normal/Normal category has the highest support value, while the Low/Low category has the highest lift value.

The data suggests that people with high systolic and diastolic blood pressure are more vulnerable to health risks. This is supported by a high support value of 0.115, high confidence level of 0.9, a high conviction level of 8.85, and a high lift value of 7.8. These values fall into the "High/High" category. On the other hand, those with low systolic and diastolic blood pressure are less likely to be affected, with a lower support value of 0.266, high confidence level of 0.785, a high conviction level of 3.41, and a lift value of 2.95, which fall into the "Low/Low" category.

As per the summary provided above individuals experiencing elevated levels for both systolic and diastolic blood pressures run close to nine times higher chances for classification as high risk when compared against their peers not experiencing these issues. Meanwhile patients displaying relatively lower readings for both categories stand three times better chances for classification in the non high risk category. The results imply that receiving a diagnosis of high blood pressure increases the chances of an individuals exposure to danger. However those individuals who experience lower levels may face lesser risks (MediLexicon International, 2023).

Clustering Analysis:

|  |  |
| --- | --- |
| **Cluster** | **Num\_of\_Patients** |
| 1 | 194 |
| 2 | 344 |
| 3 | 476 |

Table 1.2: Summary of the number of patients in each cluster after performing KMeans clustering on the SystolicBP column. The table shows the cluster number and the corresponding number of patients in that cluster.

Chart, histogram

Description automatically generated

Fig 1.2: A histogram showing each cluster's SystolicBP values distribution. The plot includes three histograms, each representing one cluster. The x-axis represents SystolicBP values, while the y-axis represents frequency. It includes a legend to distinguish between the three clusters.

Based on the patient's systolic blood pressure measurements, three clusters have been created using the clustering technique. Cluster 1 comprises 194 individuals, Cluster 2 has 344, and Cluster 3 has 476 patients, all of whom have similar systolic blood pressure values. Identifying patients who could be at risk for specific illnesses and planning treatments are possible using this data.

Age and SystolicBP Correlation:

Chart, treemap chart

Description automatically generated

Fig 1.2: A heatmap displaying the correlation between Age and SystolicBP. The heatmap includes correlation coefficients as annotations and uses the 'cool, warm' colour map to represent the strength of the correlation.

A moderately positive linear association between Age and systolic blood pressure is indicated by their correlation of 0.416. Systolic blood pressure consequently tends to increase with age. Correlation does not imply causation, but the link between age and systolic blood pressure cannot be disregarded.

The results presented in this research suggest this analysis is consistent with the relevant literature on machine learning methods, digital health technology, and big data analytics to enhance maternal health outcomes.

Machine learning algorithms may correctly predict pregnancy outcomes, including as gestational diabetes, preeclampsia, and premature birth, according to research by Islam et al. (2022). Before using them in clinical practise, further excellent study is necessary. Ethical and societal implications should also be considered.

According to Kuwabara et al. (2019), digital health technology can lead to better maternal health outcomes. The authors emphasised that real-time data may be collected and analysed through wearables, mobile health apps, and telemedicine to spot possible health issues and stop them from worsening.

Bogale et al. (2022) used machine learning to predict perinatal death and found the models highly accurate. They suggest that machine learning can identify high-risk pregnancies and reduce mortality rates.

The subsequent predictions might be made in light of the analysis and model created:

1. The model may predict a pregnant woman's risk of developing hypertension based on her age, BMI, and other relevant factors.
2. The model may also predict a woman's likelihood of developing preeclampsia based on the woman's blood pressure, gestational age, and other clinical characteristics.
3. The model can help detect high-risk pregnancies and provide early interventions to prevent bad outcomes for the mother and the foetus.
4. The model can predict the likelihood of gestational diabetes based on age, weight, and medical history.
5. The model may estimate the probability of premature labour based on gestational Age, past pregnancy history, and clinical symptoms.
6. The model can identify pregnant women needing special care, such as those with multiple gestations or pre-existing medical issues.

Although it is possible to make these predictions, they should always be assessed using clinical judgement and additional testing. The model does not intend to take the role of qualified medical personnel in making decisions regarding the health and welfare of expecting mothers and their unborn children.

**Recommendations**

The following recommendations to the medical organisation can be made based on the dataset analysis and predictions made by the machine learning model:

1. Targeted interventions: The medical organisation should consider implementing targeted therapies to lower high blood pressure among pregnant women, given the apparent association between high blood pressure and a higher risk of maternal morbidity. This may involve regular blood pressure readings, dietary adjustments, and physical activity modifications to enhance lifestyle choices and lower blood pressure.
2. Early detection of high-risk pregnancies: The machine learning algorithm can identify high-risk pregnancies early to provide fast and effective care, leading to decreased mortality by ensuring that women with high-risk pregnancies receive the care and support they require throughout their pregnancy.
3. Use of digital health technologies: Digital health solutions, including as wearables and smartphone apps, can help in the early diagnosis of potential health issues by collecting and evaluating maternal health data in real time. Telemedicine can be explored to reach pregnant women in remote locations for care..
4. Continuing research: Continue research into using big data analytics and other cutting-edge technology to effectively detect and appreciate risk concerns and offer more specialised remedies.

References

Bhavadharini, B., Mahalakshmi, M.M., Anjana, R.M., Maheswari, K., Uma, R., Deepa, M., Unnikrishnan, R., Ranjani, H., Pastakia, S.D., Kayal, A., Ninov, L., Malanda, B., Belton, A. & Mohan, V. (2016). Prevalence of Gestational Diabetes Mellitus in urban and rural Tamil Nadu using IADPSG and WHO 1999 criteria (WINGS 6). *Clinical Diabetes and Endocrinology*, 2, 8. Available online: https://doi.org/10.1186/s40842-016-0028-6.

Bogale, D.S., Abuhay, T.M. & Dejene, B.E. (2022). Predicting perinatal mortality based on maternal health status and health insurance service using homogeneous ensemble machine learning methods. *BMC Medical Informatics and Decision Making*, 22(1), 341. Available online: https://doi.org/10.1186/s12911-022-02084-1.

Fox, R., Kitt, J., Leeson, P., Aye, C.Y.L. & Lewandowski, A.J. (2019). Preeclampsia: Risk Factors, Diagnosis, Management, and the Cardiovascular Impact on the Offspring. *Journal of Clinical Medicine*, 8(10), 1625. Available online: https://doi.org/10.3390/jcm8101625.

Gelman, A., Hill, J. & Vehtari, A. (2020). *Regression and Other Stories*. Cambridge University Press.

Islam, M.N., Mustafina, S.N., Mahmud, T. & Khan, N.I. (2022) Machine learning to predict pregnancy outcomes: a systematic review, synthesising framework and future research agenda. *BMC Pregnancy and Childbirth*, 22(1), 348. Available online: https://doi.org/10.1186/s12884-022-04594-2.

Kuwabara, A., Su, S. & Krauss, J. (2020). Utilising Digital Health Technologies for Patient Education in Lifestyle Medicine. *American Journal of Lifestyle Medicine*, 14(2), 137–142. Available online: https://doi.org/10.1177/1559827619892547.

Medical News Today (2017). *Heart rate: What is a regular heart rate?* Available online: https://www.medicalnewstoday.com/articles/235710 [Accessed 07/05/2023].

Müller, A.C. & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O’Reilly Media, Inc.

Naila Baig-Ansari, Salma Halai Badruddin, Rozina Karmaliani, Hillary Harris, Imtiaz Jehan, Omrana Pasha, Nancy Moss, Elizabeth M. McClur, & Robert L. Goldenberg (no date) *Anemia Prevalence and Risk Factors in Pregnant Women in an Urban Area of Pakistan*. Available online: https://doi.org/10.1177/156482650802900207.

Peng, R.D. (no date) R Programming for Data Science.

Smith, L.I. (2002). A tutorial on Principal Components Analysis. Available online: https://ourarchive.otago.ac.nz/handle/10523/7534 [Accessed 06/05/2023].

Wooldridge, J.M. (2015). *Introductory Econometrics: A Modern Approach*. Cengage Learning.

World Health Organization (no date) *Maternal mortality*. Available online: https://www.who.int/news-room/fact-sheets/detail/maternal-mortality [Accessed 07/05/2023].