

**ST3189 MACHINE LEARNING**

**COURSEWORK**

**Student registration NUMBER (SRN): 220458971**

**NAME: SEOW XIN YONG**

**DATE: 03-04-2025**

**TABLE OF CONTENT**

|  |  |  |
| --- | --- | --- |
| **Index no.** | **Content** | **Page No.** |
| 1.0 | Introduction | 2 |
| 2.0 | Literature Review | 3 |
| 3.0 | Methodology  3.1 Datasets and Preprocessing  3.2 Model Suitability  a) Task 1: Unsupervised Learning  b) Task 2: Regression  c) Task 3: Classification | 4-7 |
| 4.0 | Interpretation of Results  4.1 Clustering  4.2 Regression  4.3 Classification | 7-10 |
| 5.0 | Conclusion | 11 |
| 6.0 | Reference | 12 |

1. **Introduction**

Machine Learning is effective in analyzing complex datasets, revealing the underlying patterns and trends of the data, extracting the useful insights and making predictive conclusions with a high level of accuracy. In this research, the maternal health factors and risk level are the main concern of each machine learning task which are Unsupervised Learning, Regression and Classification. According to the World Health Organization (WHO), a maternal death occurred almost every two minutes in 2020, thus highlighting that it is crucial to identify the risk factors accompanied with maternal health condition (World Health Organization: WHO, 2024).

In this research, the research questions for each of the approaches are formed due to the attention put on the maternal health. For Clustering, the research question is how can we segment the pregnant individuals according to the maternal health factors such as maternal age and others, while for Regression, the research question is how do the maternal health factors affect the pregnant individual’s blood sugar. For Classification, the focus is on the ability to predict the health risk level of a pregnant individual based on maternal health factors.

Furthermore, the model’s suitability and compatibility to each of the machine learning tasks are examined from different aspects. One of the most critical aspects of this study is to ensure the effectiveness of each model by evaluating the model performance through metrics such as Mean Squared Error (MSE), F-statistics, and confusion matrix. Additionally, cross-validation is carried out to compare and ensure the model performance is consistent.

In Clustering, we will compare the K-means algorithm results with hierarchical clustering to assess the validity of the segmentation. Meanwhile, the ANOVA test is carried out on linear regression and polynomial regression model to identify the suitability of each model. Likewise, the best-fit-line is plotted to represent a visualization of actual values against the predicted values for the blood sugar level. Furthermore, in the classification task, the random forest model is used to prevent overfitting, also the performance is being reevaluated by comparing to the Variable Importance Plot.

Nonetheless, there are challenges and critical problems that may be faced during the machine learning applications. The most fundamental challenge will be the quality and quantity of datasets, datasets with noises, biased and inconsistent variance may create unstable result. Consequently, an unstable and unreliable result may cause distortion in the contribution to the risk assessment, health diagnosis and optimal treatment to the pregnant individuals.

Conclusively, the study is about the discussion and knowledge on the pregnant individuals from different types of perspectives. From the maternal health factors and maternal health risks during pregnancy period, many meaningful insights can be gained from the differences and similarities among the pregnant individuals. By analysing the most impactful features and exploring the potential predictors variables, better treatment and more accurate diagnosis can be made to the pregnant individuals from the early stage of pregnancy.

**2.0 Literature Review**

Suffering from a health disorder condition in the pregnancy period could lead to demise. The most poignant part is most of the maternal health complications could be prevented with an early intervention. While this tragedy has been brought to attention, many studies are done under the maternal health risk topic to identify the factors that affect the maternal health.

A study, “Cluster Analysis: A New Approach for Identification of Underlying Risk Factors and Demographic Features of First Trimester Pregnancy Women”, (Gárate-Escamilla et al., 2020) investigates the clusters of pregnant individuals that can be formed according to the predictor variables of the thyroid disease. The underlying relationship between the pregnant individuals’ health profile and thyroid disease is explained in the research. In light of this, the attention for the current research is put on the factors such as health risk level, body temperature, body mass index (BMI), and the others.

During the pregnancy period, the placenta will stimulate the hormones that cause disorder in the insulin uptake in the body, called the “insulin resistance”. Also, the pregnant individuals will have a higher food consumption compared to normal individual, these two conditions will result in a loss of control for the glucose amount in the pregnant individuals’ body. It is critical in consistently monitoring the maternal health condition to prevent high blood sugar level.

Another study, “Association between Maternal Blood Glucose Levels during Pregnancy and Birth Outcomes: A Birth Cohort Study”, (Zhao et al., 2023) which reveals the positive relationship between the abnormal blood glucose level and the neonatal birth. This suggests the importance of a stable blood glucose level on the health condition of the infant. Yet, the possible factors that may affect the blood glucose level have not been found, thus there is an interest in the investigation of the factors affect the blood sugar level. Since blood sugar is a critical source of fuel for cells throughout the body, it is crucial to identify the condition to stabilize the blood sugar level.

For classification, a study titled “Deep Hybrid Model for Maternal Health Risk Classification in Pregnancy: Synergy of ANN and Random Forest”, (Togunwa et al., 2023) combines the Artificial Neutral Networks (ANN) and Random Forest to classify the maternal health risk levels. This proves that the hybrid models can improve the accuracy and reliability of the classification result. The focus is seemingly more on the contrast of the hybrid models and single model, instead of the aim on classifying the pregnant individuals under the intention to give a better early diagnosis and optimal treatment. Hence, the intention is being focus in the current study to investigate further on the maternal health risk levels of the [regnant individuals.

In conclusion, machine learning demonstrates its effectiveness and potential on making predictions, finding the underlying relationship between independent and dependant variables in the previous studies. While Clustering approach helps to segment patients, Regression model predicts the health outcome, and Classification technique enhance risk assessment. Furthermore, the current study adopts an interest on the maternal health factors and risks at a granular level. Although there exist challenges on the data quality, missing values and interpretability, machine learning offers useful insights that can further improve the maternal health outcomes.

**3.0 Methodology**

**3.1 Datasets and Preprocessing**

The datasets used for this project were two different datasets with similar features, found from UCI Machine Learning Repository and Mendeley Data. One of each (called the Risks dataset) was collected from different medical organisations of Bangladesh through the Internet of Things (IoT) based on risk monitoring system, while the other (called the Factors dataset) owned by Daffodil International University. The first one was collected under the aim on the maternal health risks, while the other was collected with the focus on the maternal health factors. There were 1013 records for and 1025 records consecutively for both the datasets.

First of all, both the datasets were being investigated and analysed by exploring the data distribution based on the predictor variables, and the descriptive statistical values such as mean, median and mode. The following figures were the histogram of the variable distribution.

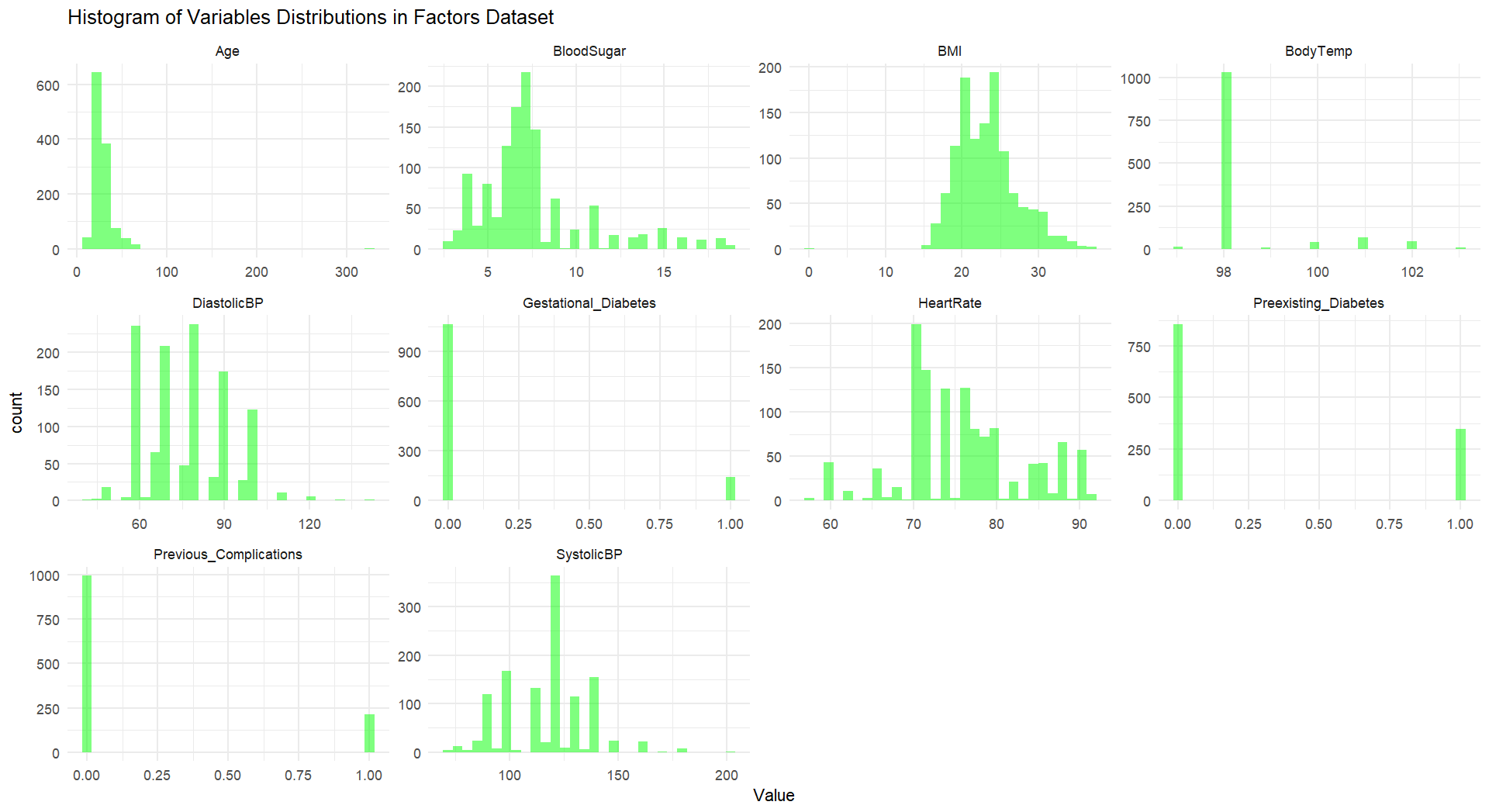
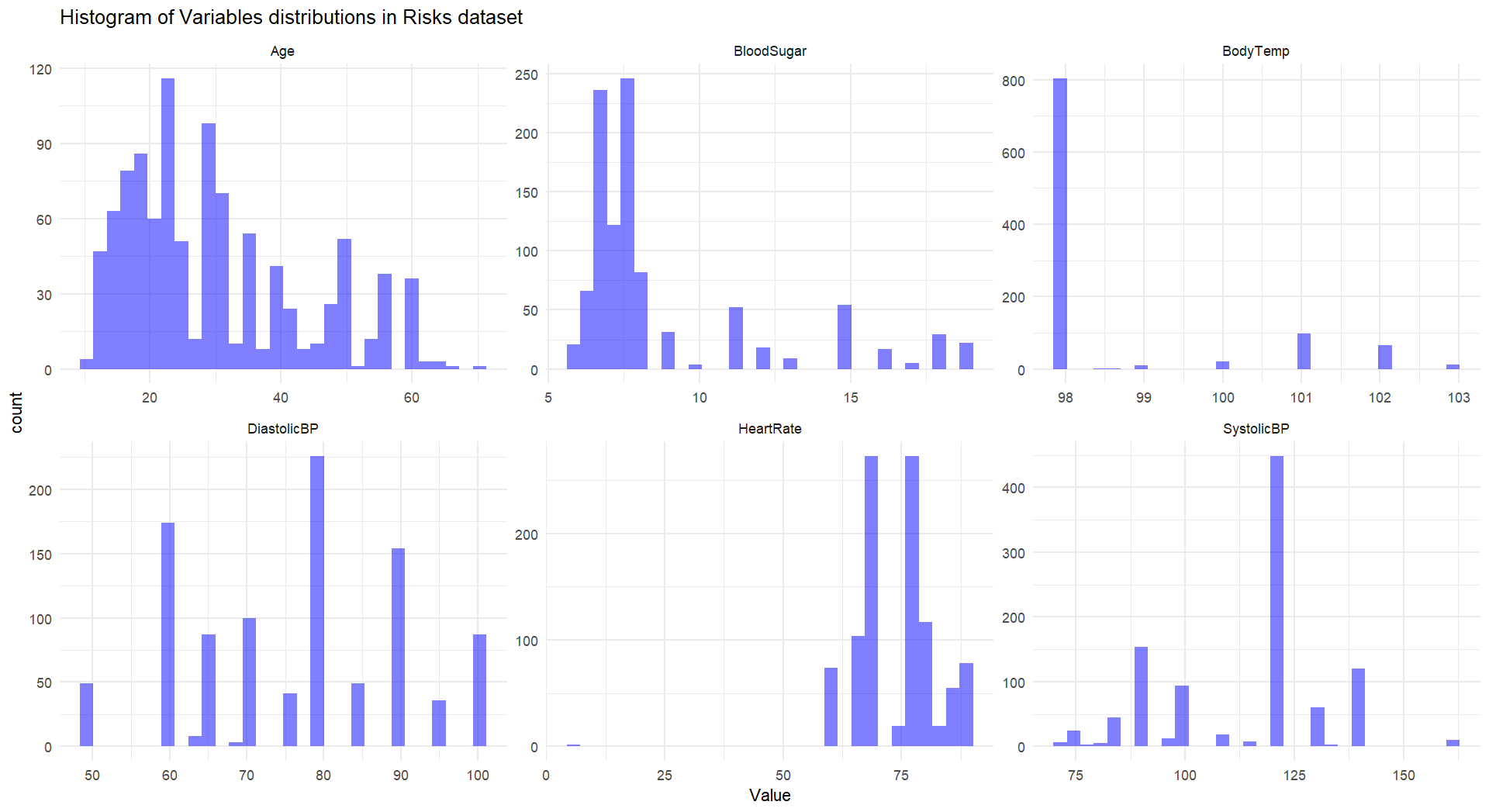


Figure 3.1a: Histogram of variables distribution for both datasets

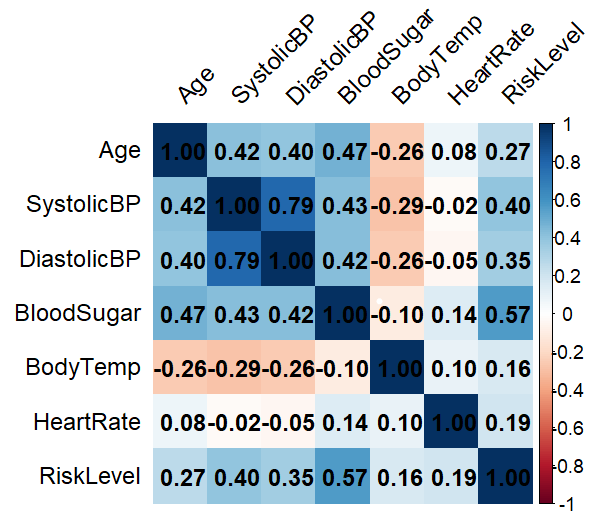
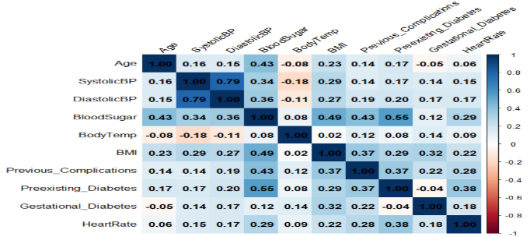
From the figure on the left, the maternal aga records mostly fell within the range of years 15 to 35, while there existed a few peaks beyond 40 years. It showed a slightly right-skewed distribution compared to the blood sugar plot, which had a more right-skewed shape. It was clear to see there was three peaks (60, 80, 90) in the “DiastolicBP” which may suggest the existence of distinct groups. For the right-hand side figure, it showed more peaks in the different histogram, such as “DiasstolicBP”, “Preexisting\_Diabetes”, “Previous\_Complication”, and “SystolicBP”. These suggested the presence of clustered groups among the pregnant individuals. For the records that fell within the higher range in the plot of “BMI”, “HeartRate”, “BloodSugar”, and “BodyTemp”, these were suspected to have an abnormal body health condition of a pregnant individual.

Figure 3.1b: The Correlation matrix plot of variables in both datasets.

The above figures were the correlation matrix of the variables in each dataset. When the correlation values greater than 0.7, the variables would be considered as strongly correlated, weekly correlated when the correlation value less than 0.3. In Risks dataset, the correlation between the features was acceptable, the average of the correlation values was around 0.4, which suggested that there exists a weak relationship between the features. Meanwhile in the Factors dataset, the predictor variables were having a very weak correlation with each other, which can be negligible. Specifically, “BloodSugar” showed a moderately correlation with most of the variables, which indicated that there may have a moderate effect of the other variables on the value of “BloodSugar”.

However, there existed missing values and not applicable (NA) values in some of the feature columns for the Factors dataset. Hence, the rows that contains missing values and NA, were omitted to improve the model accuracy and prevent bias in the analysis. Since the number of rows omitted was 27, which no more than 60 records (5% of the whole data records of the dataset). Thus, 27 rows of records were considered as a number that is safe to be omitted.

Besides, some adjustments and engineering processes were needed on the datasets to enhance the interpretability of the datasets. The feature “RiskLevel” in risks dataset was encoded to a categorical variable to enable the classification to occur. Both datasets have similar features, hence this helped to maintain the unity of the results for three different tasks. Thus, the result was expected to be more representative for the research aim of each different task. In this project, we put the main focus on the mother’s conditions during the pregnancy period.

**3.2 Models Suitability**

**a) Task 1: Unsupervised Learning Model**

**Research question**: How can we segment pregnant individuals according to the maternal health factors such as maternal age, blood pressures, blood sugar, body temperature, heart rate, and maternal health risk level?

The research question was formed due to the intention of figuring out, whether it was possible to separate the pregnant individuals based on the maternal health factors. For unsupervised learning task, clustering was being chosen to fulfil the requirement. Clustering was expected to find out the underlying dimensions within the pregnant individuals who may had the similar characteristics and profile. The convex shape (convex hulls) was chosen for the clustering visualization to represent a clearer boundary of each cluster formed, and to better represent the overlapping region of the clusters if there is any.

K-means algorithm was first being used to separate the data into different clusters, due to its simplicity and productivity. For the k mean values, it was determined by using the Elbow Method at 3. Then, the similarity measure was done to measure the distance between each data to its closest mean, thus forming a centroid like shape around each mean value.

However, to determine the suitability of the clusters formed by K-means algorithm, hierarchical clustering was used to form three different clusters for the comparison to the clusters formed by K-means algorithm. Hierarchical clustering created a tree-like structure, called Dendogram, where it will merge each data points due to the similarity of the points. The figure should be viewed form the bottom to the top, the clusters will be formed by cutting the tree at a threshold point.

**b) Task 2: Regression model**

**Research question**: How do the maternal health factors affect the pregnant individual’s blood sugar (in mmol/L)?

The model of the regression was as follow:

Since the intention was to investigate the relationship between the maternal health factors and the maternal blood sugar level, hence regression model was the most suitable approach as it studied the effect between yet within independent and dependent variables. However, it was uncertain among multilinear and polynomial regression model, whether which is more compatible to the dataset. An ANOVA table was generated to compare the linear model and polynomial model, the polynomial had a lower residual sum of squares (RSS = 3966.9), a very large F-statistic (49.338), and an extremely small p-value (<2.2e-16), compared to the linear model. Thus, the polynomial regression model was chosen.

The reason of the features such as “Age”, “SystolicBP”, “DiastolicBP”, “BodyTemp”, “BMI”, being chosen as a higher degree term, was due to the previous knowledge and experience. As blood sugar level often increases along with age, blood sugar was often proved to have a non-linear relationship with the other features. For example, a high body temperature condition (such as fever), may disrupt the blood sugar level in a non-linear way, while a high BMI condition (such as obesity) may experience abnormal blood sugar level.

To measure the model performance, the train-test split was used here by separating the dataset into two groups, where the size ratio was 8:2, 80% of the data was used for training purpose, the remaining 20% was used as train data. Moreover, the mean squared error (MSE) and R-squared value were calculated and a graph was plot (such as ggplot), to have a better vision on the fitness of the regression model to the dataset.

**c) Task 3: Classification model**

**Research question**: Can we predict the health risk level of a pregnant individual based on maternal health factors?

The classification model was considered as the most suitable approach to reveal the answer to the research question. The research question was being chosen under the intention to have a more complete understanding on the pregnant individuals’ health risk, thus to give a better consultation on the health condition for the pregnant individuals. Along with the intention, the possible health factors that will cause an effect on the maternal health risk level were expected to be revealed.

To classify the pregnant individuals’ health risk level according to their health factors, the risk level of pregnant was modified to be an ordinal variable by assigning the number “1”, ”2”, ”3” to the health risk level, “Low”, “Medium”, “High”, accordingly. Train- test split was used in this model, where the pregnant individuals were separated into two groups with the ratio of the group size as 8:2, 80% of the data were used for training and 20% of them were used for testing the model on the unseen data. This was carried out to measure the model performance by comparing the theoretical and practical condition.

To prevent overfitting and ensure the accuracy of the model, the random forest technique was used to train the dataset. By using random subsets of the data, random forest built and combined multiple decision trees to get a generalized result. Since random forest was actually a combination of multiple decision trees, hence It was expected to be a good approach to prevent overfitting and improve the accuracy of the model.

By comparing the actual and predicted values, the confusion matrix was used to evaluate the model performance in a more completed way. Yet, as a comparison for the model evaluation, the Variable Importance Plot (varImpPlot), was used to visualize the importance of the effect of each feature on the classification result.

**4.0 Interpretation of Results**

**4.1 Clustering**

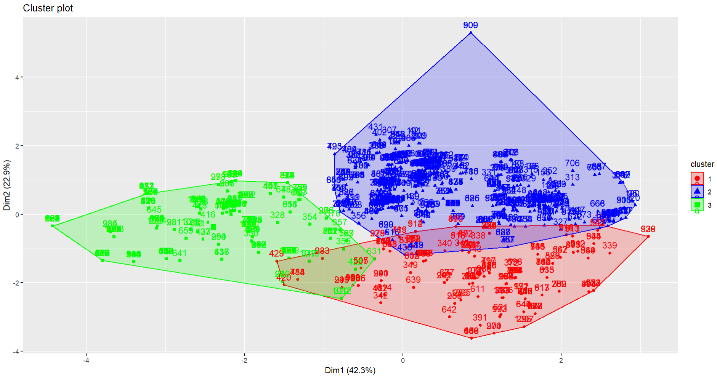
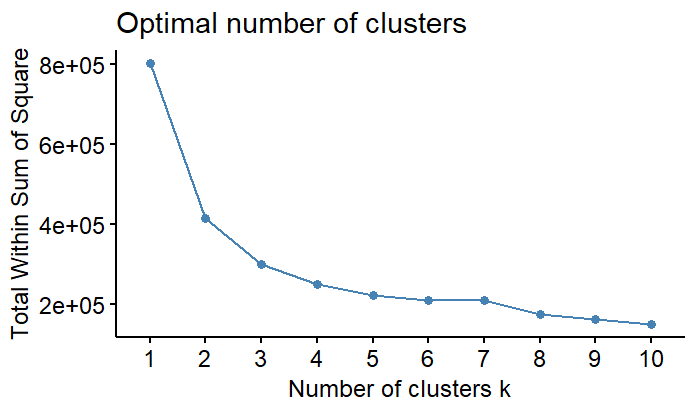


Figure 4.1a: Elbow method Figure 4.1b: K-means cluster plot

The above figures show the elbow method result where it helped to set the k mean value to 3, and the K-means cluster plot. Since dimension 1 explained 42.3% of the variation while dimension 2 explained 22.9% of the variation, a total of 65.2% of variation were explained, which was a significant percentage. From the figure, it was clear that there was no overlapping region between Cluster 2 (Blue), and Cluster 3 (Green), indicating that they were two distinct groups. Yet, both of them overlapped with Cluster 1 (Red), which showed that there was a degree of similarity between them and Cluster 1. Hence, hierarchical clustering was being carried out to ensure the consistency of the result and compare the difference in the cluster structures.

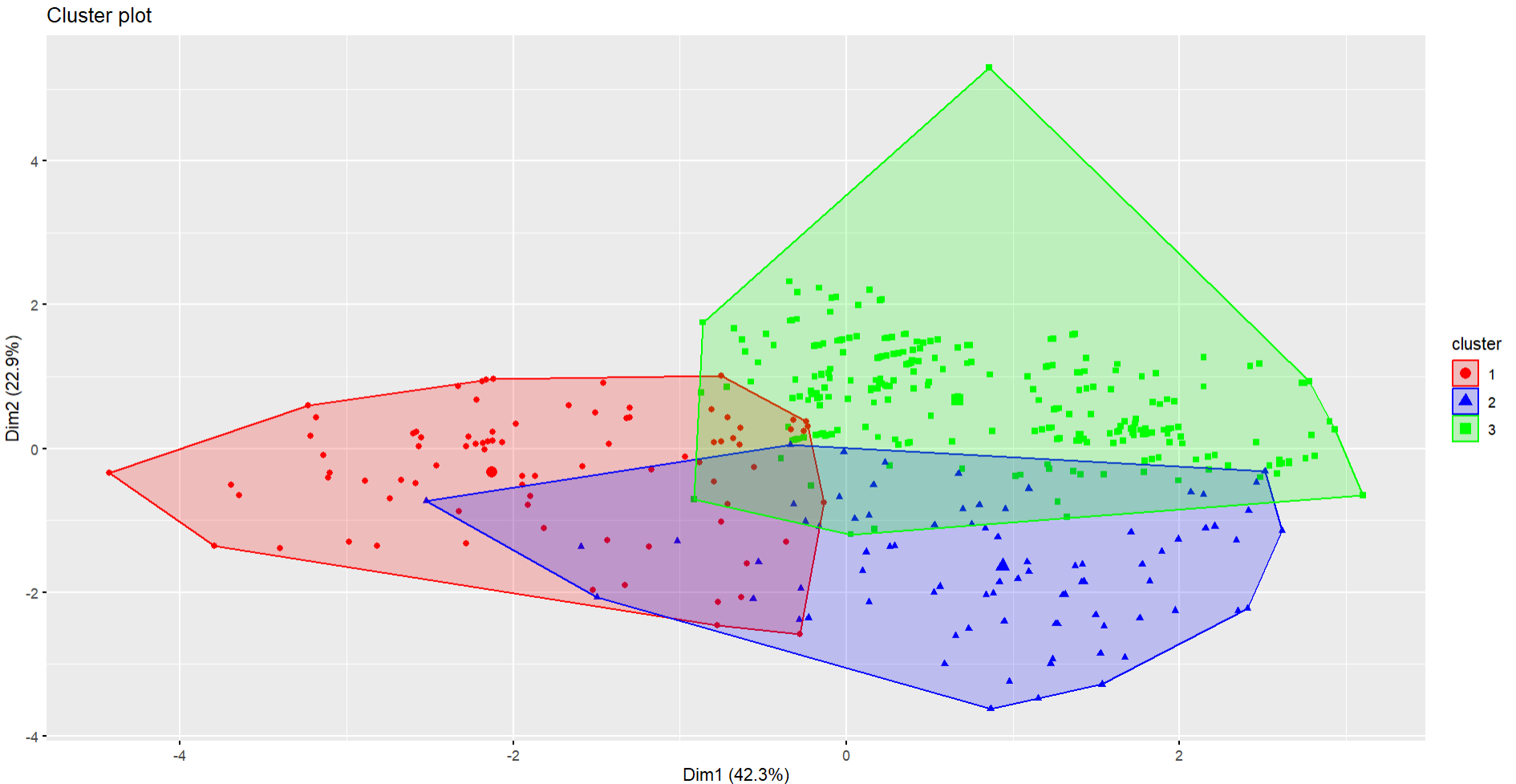
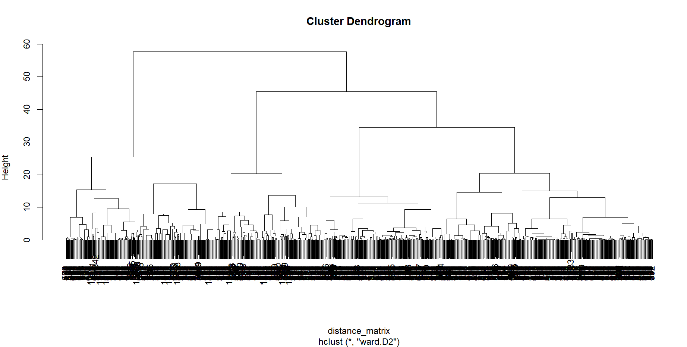


Figure 4.1c: Dendogram Figure 4.1d: Hierarchical clustering plot

Consequently, Hierarchical clustering was carried out, and the number of clusters was determined by using the Dendogram plot. Although the total variation that the dimensions accounted for (65.2%) was the same as K-means clustering, yet the cluster plot represented a different visualization compared to the K-means cluster plot. Three clusters were overlapping each other in a wider region and there was no any space between the clusters, indicating there may be a higher degree of similarity between the clusters. Since it was significantly different from the K-means cluster plot, thus it was certain that a further investigation is needed for the dataset.

**4.2 Regression**

For the polynomial regression model, an ANOVA test was carried out and confirmed that there was a high possibility of the non-linear relationship between the independent and dependant variables. The higher degree terms were determined based on the prior knowledge to blood sugar of the body and previous experience.

First of all, the intercept (6.649), indicated the expected blood sugar level when all the features were at their mean values. Terms with a highly significant p-value were “Age” (p< 2e-16 for first degree), “Previous\_Complications” (p< 0.001), “Preexisting\_Diabetes” (p< 0.001), “Gestational\_Diabetes” (0.0549), “BMI” (p< 0.001), “HighRiskLevel” (p= 0.0306). Specifically to mention that only the first-degree term was significant for both “BodyTemp” and “BMI”, which suggested that there was no non-linear relationship between both predictors and “BloodSugar”, instead of a strong positive linear relationship for both the predictors.

Followed by the insignificant terms, which included “SystolicBP” (p= 0.1190 for first degree), “DiastolicBP” (p= 0.0752 for first degree), and “HeartRate”. A high number in p-value suggested a weak relationship between the predictors and “BloodSugar”. Both “SystolicBp” and “DiastolicBP” had non-significant p values for first and second-degree term, which suggested that they had no significant linear and quadratic effect on “BloodSugar”.

Furthermore, the value of Multiple R-squared (0.6358) referred that the model explained 63.58% of the variance in “BloodSugar”, which is moderately strong explanation power. With a high value in F-statistic (101.2), and an extremely small p-value (< 2.2e-16), it was suggested that the overall model performance was highly statistically significant.

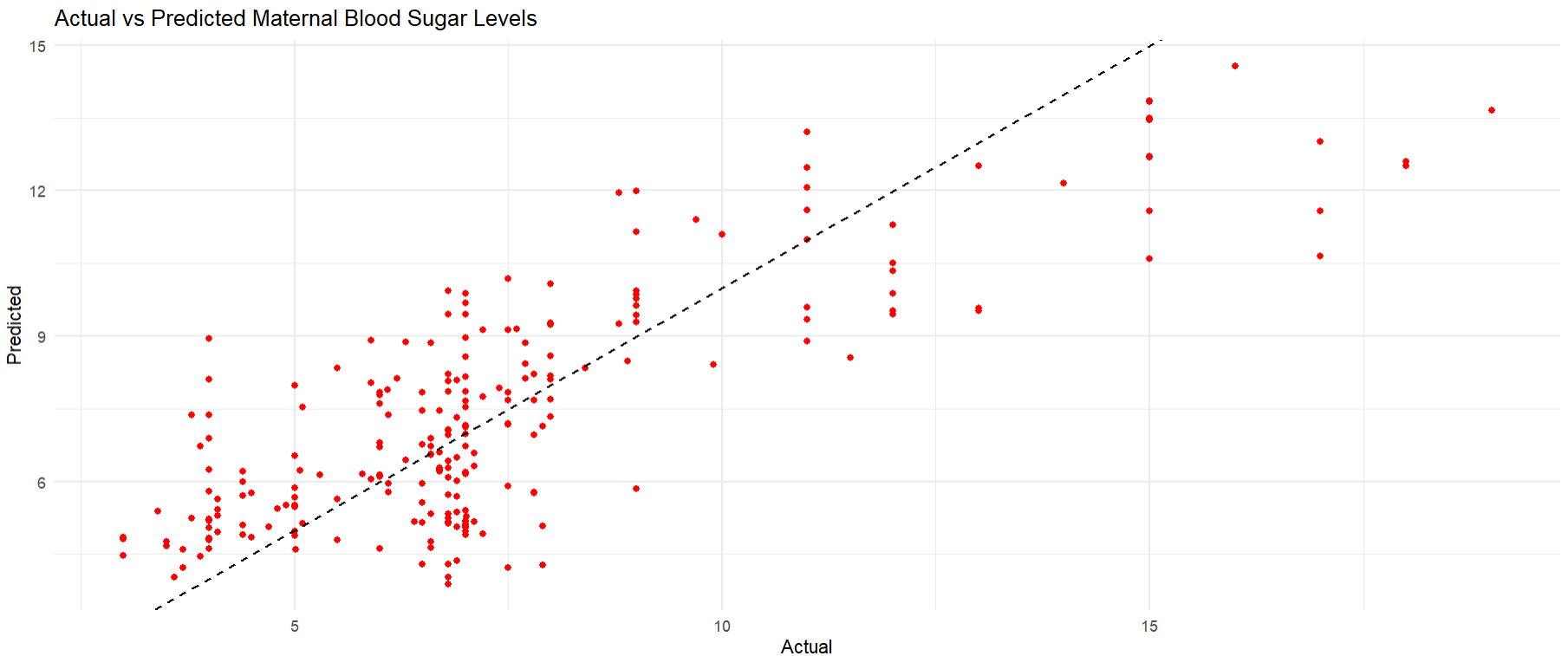


Figure 4.2: Actual VS Predicted Maternal Blood Sugar Levels

The actual and predicted values of blood sugar were being plot to give a better visualization of the model performance and ensure the accuracy of the model. Most of the points scattered more around the best fit line when the blood sugar level was less than 10, yet the points were scattered far from the best fit line when the blood sugar level was greater than 10. This indicated that there may be misestimation for the actual value of blood sugar when it is greater than 10.

**4.3 Classification**

The variable “RiskLevel” was modified and encoded as a categorical variable and the train-test split was being done. From the confusion matrix, the most significant observation was the table of predicted class and actual class of each data. It was clearly that most of the data had the same result for the predicted and actual condition, where Class 1 (70), Class 2 (51), and Class 3 (50). The model performance in overall was at an accuracy of 0.8465, where the p-value for the accuracy showed a very strong statistical significance that the accuracy was valid.

While the Kappa score (0.7668), the measurement on the degree of compatibility between the actual value and predictions, inferred that the model’s classification ability was high. For the McNemar’s Test, p-value (0.9083), which was greater than 0.5, suggested that there was no extreme bias and the misclassification errors were fairly balanced.

In general, the specificity of the model which measured the true negative rate of the result, was high (all above 90%) for each class. However, the sensitivity of the model that measured the true positive rate of the model showed the lowest rate at 76.12% for Class 2, which was lower than the other two classes. Hence, it suggested that the predictions for Class 2 had a higher chance to be wrong. Conclusively, from the confusion matrix, the model was a strong model, but more investigation on the data were needed for the Class 2.

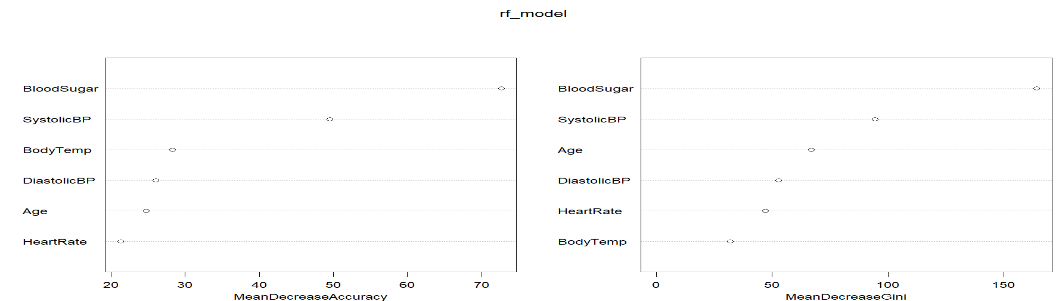


Figure 4.3: The varImpPlot

Yet, the varImpPlot was created to represent the visualization of the importance of each predictor variable on the classification process. The plot on the left showed the importance level of each predictor variable on the model performance, where the importance level increased with the values on the x-axis. While the plot on the right showed the rank of variables on the impurity reduction level. It was clear that “BloodSugar” and “SystolicBP”, were the two significant variables on the classification power.

**5.0 Conclusion**

In general, the datasets used for the three tasks provided moderate level of utility for the study, although its utility was somewhat limited. For both classification and regression models, they were both statistically strong yet there still had room for improvements, while for clustering, the result of the model was moderate. Hence, a conclusion came out that the datasets had to be investigated more according to the compatibility and suitability of the features, even the size of the data records.

First and foremost, it was essential to have a distinct and least correlated features for a clustering model, and often the number of features had to be reduced by removing features which were less contributed to the clustering result. Needless to say, it was crucial to be always certain on the nature of each independent variable, while reducing the variables that may create noise. Specifically, for the K-means algorithm, the k value should be optimised to avoid under or over- clustering and ensure the accuracy of the result of the model. Meanwhile, different distance metrics and different hierarchical linkages should be tested to determine the optimal method to use for the datasets.

Since normally distributed residuals with a constant variance would minimise errors optimally and prevent poor statistical inference, hence it was critical to enhance the distribution of the residuals and homoscedasticity (where there exists a constant value of the variance of residuals). Multicollinearity, which was the relationship between independent variables, should be checked consistently to avoid a high correlation between each term, while the polynomial terms chosen should be reevaluated based on its effect on the model performance. Nonetheless, a clearer scale and proper labelling enhanced the visibility of the best-fit line. Additionally, adding a confidence interval would help to illustrate the range in which the predicted results were likely to fall.

For classification model, the model performance in general was satisfying, however it could be improved by weight balancing the Class 2. To prevent the imbalance accuracy of the classes (Class 2 had the lowest sensitivity among three classes), Class 2 could be assigned a higher weight to its misclassifications while the number of features being adjusted at each split. Thus, the model would put more attention on the Class 2.

Conclusively, the dataset investigation should be done in more thorough manner, and essential dimension reduction step should be included to improve the accuracy and effectiveness of the model. Nevertheless, the actions needed to enhance the model performance should be implemented in every step.

**6.0 Reference**

1. Ahmed, M., Kashem, M. A., Rahman, M., & Khatun, S. (2020). Review and analysis of risk factor of maternal health in remote area using the Internet of Things (IoT). In Lecture notes in electrical engineering (pp. 357–365). <https://doi.org/10.1007/978-981-15-2317-5_30>
2. Togunwa, T. O., Babatunde, A. O., & Abdullah, K. (2023). Deep hybrid model for maternal health risk classification in pregnancy: synergy of ANN and random forest. Frontiers in Artificial Intelligence, 6. <https://doi.org/10.3389/frai.2023.1213436>
3. Zhao, D., Liu, D., Shi, W., Shan, L., Yue, W., Qu, P., Yin, C., & Mi, Y. (2023). Association between Maternal Blood Glucose Levels during Pregnancy and Birth Outcomes: A Birth Cohort Study. International Journal of Environmental Research and Public Health, 20(3), 2102. <https://doi.org/10.3390/ijerph20032102>
4. Mennickent, D., Rodríguez, A., Opazo, M. C., Riedel, C. A., Castro, E., Eriz-Salinas, A., Appel-Rubio, J., Aguayo, C., Damiano, A. E., Guzmán-Gutiérrez, E., & Araya, J. (2023). Machine learning applied in maternal and fetal health: a narrative review focused on pregnancy diseases and complications. Frontiers in Endocrinology, 14. <https://doi.org/10.3389/fendo.2023.1130139>
5. Gárate-Escamilla, A. K., Garza-Padilla, E., Rivera, A. C., Salas-Castro, C., Andrès, E., & Hassani, A. H. E. (2020). Cluster Analysis: a new approach for identification of underlying risk factors and demographic features of first trimester pregnancy women. Journal of Clinical Medicine, 9(7), 2247. <https://doi.org/10.3390/jcm9072247>
6. Deshpande, H. (2021, April 23). Mother’s Significant Feature (MSF) dataset. IEEE DataPort. <https://ieee-dataport.org/open-access/mother%E2%80%99s-significant-feature-msf-dataset?utm_source=PANTHEON_STRIPPED>
7. UCI Machine Learning Repository. (n.d.). <https://archive.ics.uci.edu/dataset/863/maternal%2Bhealth%2Brisk?utm_source=chatgpt.com>
8. Dataset - updated.csv - Mendeley Data. (n.d.). <https://data.mendeley.com/datasets/p5w98dvbbk/1/files/e7b178fd-09ae-4859-b4b3-5675f79f470c>
9. Rahman, M. M., Abe, S. K., Kanda, M., Narita, S., Rahman, M. S., Bilano, V., Ota, E., Gilmour, S., & Shibuya, K. (2015). Maternal body mass index and risk of birth and maternal health outcomes in low‐ and middle‐income countries: a systematic review and meta‐analysis. Obesity Reviews, 16(9), 758–770. <https://doi.org/10.1111/obr.12293>

<https://mhlw-grants.niph.go.jp/system/files/2016/161021/201604001B_upload/201604001B0018.pdf>

1. Managing blood sugar in Pregnancy | InfantRisk Center. (n.d.). <https://www.infantrisk.com/content/managing-blood-sugar-pregnancy>
2. World Health Organization: WHO. (2024, April 26). Maternal mortality. <https://www.who.int/news-room/fact-sheets/detail/maternal-mortality>