**Table of Content**

Introduction …………………………………………………………………………………2

Business understanding………………………………………………………………………2

Project Objectives…………………………………………………………………………….2

General goal …………………………………………………………………………………3

Project Scope……………………………… …………………………………………………3

Methodology………………………………………………………………………………….3

Maternal complications……………………………………………………………………...4

Milestones Achieved…………………………………………………………………………5

Models and machine learning algorithms…………………………………………………..7

Challenges Faced……………………………………………………………………………..8

Data Attributes……………………………………………………………………………......8

Key Insights from Data Analysis…………………………………………………………….8

Descriptive statistics and Data visualisation………………………………………………..8

Heatmap …………………………………………………………………………….….12

Correlation: Pair plot …………………………………………………………………….....14

MinMaxScaler…………………………………………………………………………….…15

Principal Component Analysis (PCA)……………………………………………………...16

Statistical Shapiro-Wilk test……………………………… ……………………………….17

Machine Learning Algorithms by Train and Test………………………………………....18

Machine Learning Modelling by Mean…………………………………………………....19

Fine-Tuning Hyperparameter……………………………………………………………...20

Cross Validation…………………………………………………………………………….21

Model Deployment and Results…………………………………………………………...22

Proposed Solutions for Future Improvements……………………………………………24

Conclusion………………………………………………………………………………….25

Recommendations…………………………………………………………………………26

Citations and Bibliography…………………………………………………………………27

GitHub link…………………………………………………………………………………28

**The Maternal Health Risk Analysis Project**

By Olena Zymek, sba23334

**1. Introduction:**

This capstone project explores the complex interactions between health metrics during pregnancy, concentrating on characteristics like heart rate, blood glucose levels (BS), and diastolic blood pressure (DiastolicBP). Diastolic blood pressure, or lower blood pressure, is a vital indicator of vascular resistance during pregnancy and has a major impact on the health of the mother. Since blood glucose levels (BS) affect the health of both the mother and the fetus, it is crucial to measure them in terms of molar concentration, and predicting potential pregnancy-related complications. Furthermore, one of the basic physiological parameters that is used to assess cardiovascular health is heart rate. In order determine diastolic blood pressure, examine the relationship between blood glucose levels and other variables, investigate the effect of heart rate on pregnancy outcomes, and ultimately predict the risk level during pregnancy, the main goal of this capstone project is to develop predictive models that utilize these critical attributes. This project aims to contribute valuable insights into pregnancy risk factors by identifying complex patterns within these health metrics, opening the door to tailored interventions and well-informed healthcare strategies. This project aims to improve our understanding of maternal health during pregnancy while also emphasising the importance of navigating the complexities of health data. The study also delineated future research opportunities to facilitate the existing initiatives for reducing maternal health risks, revealing the unknown reasons of maternal complications, developing usable and useful clinical decision support systems to be used by the expecting mothers and health professionals, enhancing dataset and its accessibility, and exploring the potentiality of surgical robotic/programming tools. (Islam et al., 2022) This project will provide a state-of-the-art paradigm of maternal healthcare that will aid in clinical decision-making, anticipating pregnancy problems and delivery mode, and medical diagnosis and treatment.

**Business understanding**

A comprehensive understanding of the business context and complexities surrounding maternal health risks enables stakeholders to formulate well-informed strategies, policies, and interventions aimed at enhancing maternal health outcomes and safeguarding the health of mothers and their infants. This holistic approach is essential for tackling the diverse challenges associated with maternal health and fostering lasting enhancements in the delivery and accessibility of maternal healthcare services.

**Project Objectives:**

Based on dataset, reports and other publicly available resources, the goal is deep understanding of which medical conditions carry the greatest risk indicators for pregnancy-related health issues. The project objectives encompass a range of goals, including predictive modelling for estimating diastolic blood pressure, analysing the relationship between blood glucose levels and other variables, understanding the impact of heart rate on pregnancy outcomes, forecasting risk levels during pregnancy, examining feature importance, and proposing health intervention strategies. Additionally, the project aims to provide patient-focused recommendations, perform data visualization and analysis, evaluate model strength, and uphold ethical considerations. These objectives use the mentioned characteristics to develop predictive models, comprehend pregnancy-related health indicators, and offer realistic recommendations for improved outcomes.

**General goal:**

The general goal of this project is to utilize data analysis and predictive modeling techniques to gain insights into maternal health during pregnancy, specifically focusing on diastolic blood pressure (DiastolicBP) and blood glucose levels (BS). By examining the relationships between these variables and other relevant factors, the aim is to identify potential risk factors and develop strategies for mitigating health risks and promoting healthier pregnancies. Additionally, the project seeks to provide personalized health recommendations to expectant mothers based on their individual health metrics. Through data visualization and rigorous model evaluation, the goal is to create robust predictive models that can be effectively used for maternal health assessment and intervention planning.

Maternal Health Risk Analysis is the process of obtaining important information from a dataset pertaining to maternal health. By using data to anticipate and comprehend the possible risks of pregnancy, healthcare providers can better prepare for these risks and customise their interventions to improve the health of both the mother and the fetus.

**Project Scope:**

One innovative project to improve prenatal care is creating a machine learning model to forecast maternal health risks based on patient data. With the use of cutting-edge computational methods, this novel approach analyses a variety of patient data to identify high-risk pregnancies and personalise interventions for better outcomes of both the expecting mother as well as the developing baby.

**Methodology**

*Problem Definition:*

The issue at hand is the requirement for an all-encompassing comprehension and proactive management of the risks to maternal health. Predicting and addressing variables that may result in difficulties during pregnancy, childbirth, and the postpartum phase are part of this. Maternal health is complicated, so it's critical to use sophisticated analytical techniques and tools to identify, anticipate, and control these risks. If maternal health risks are not managed early on, it can cause serious problems for the mother as well as the child. Our goals are to considerably raise the standard of prenatal care, lower the risk of unfavourable health outcomes, and advance maternal healthcare in general by creating an efficient predictive model and intervention plan.

*Boundaries:*

Strict observance of ethical guidelines and data privacy laws, guaranteeing the safe management of patient information.

The development of interpretable machine learning models is prioritized in order to aid healthcare professionals in understanding. Real-time predictions for risk assessment during pregnancy.

*Work Breakdown Structure (WBS):*

This project will be split into smaller, manageable tasks. The Maternal Health Risk Analysis project involved the evaluation of several machine learning algorithms to identify the most effective model for predicting maternal health risk levels. The following algorithms were considered: Logistic Regression (LR), Linear Discriminant Analysis (LDA), k-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Naive Bayes (NB), and Support Vector Machines (SVM). Firstly, dataset will be chosen for structural analysis. In Jupyter Notebook data will be assest, trimmed accordingly, modelled and analysed due to chosen timeframe. Then the actual project will be finished in Word document.

*Timeline Development:*

Exploration and introduction to dataset. (7 days)

Data preparation after EDA and feature engineering to enhance predictive variables. (15 days)

Model Development with implementation of machine learning algorithms and tuning hyperparameters. (6 days)

Model evaluation and using accuracy, precision, and recall metrics. (5 days)

Documentation of code, findings, and methodologies in written report. (7 days)

**Maternal complications**

The common maternal complications responsible for the majority of maternal risks are gestational diabetes, hypertensive disorders (preeclampsia, eclampsia), postpartum hemorrhage and preterm labor.

Here will be investigated how different pregnancy-related health metrics relate to the incidence of maternal complications. In particular, we will look at the relationships between blood glucose levels, blood pressure, and other physiological parameters and gestational diabetes, hypertensive disorders (like preeclampsia and eclampsia), postpartum hemorrhage, and preterm labor. Through the examination of these correlations, we can pinpoint plausible risk factors and create prognostic models to foresee and reduce the probability of these pregnancy-related issues happening. These issues can physically and mentally distress the mother and/or baby, resulting in mild to life-threatening situations. Some of these problems can occur due to past medical conditions such as diabetes, family history of preeclampsia, previous cesarean delivery, previous surgery and due to bad habits.

Figure 1

A diagram of a system

Description automatically generated

(Islam et al., 2022)

**Milestones Achieved:**

1. Understanding Data thorough exploration of maternal health dataset and identification of key features influencing maternal health outcomes. Ensured dataset include a mix of categorical and numerical features. Dataset consists from 1014 rows, 7 columns, from which 562 are duplicates.

That's very big number. Duplicate data takes up unnecessary storage space and slows down calculations at a minimum. At worst, duplicate data can skew analysis results and threaten the integrity of the data set. Including them will essentially lead to the model overfitting this subset of points. But in this case, we will try to analyse this data first to understand the distribution of the data without removing them as in some cases, duplicates may represent valid observations or unique instances that provide valuable information. Our given dataset contain personal identifiers or sensitive health information, it's often advisable to err on the side of caution and refrain from removing duplicates to avoid any potential breach and to mitigate the impact of duplicates on analysis outcomes, such as adjusting analytical methods. Since each entry in the dataset represents information about a unique woman and there may be instances where multiple women have identical data, removing duplicates could inadvertently lead to the loss of valuable information about individual patients. By retaining all entries, even if they appear to be duplicates, the dataset maintains its integrity and ensures that no woman's information is inadvertently excluded from analysis.

Figure 2

A screenshot of a graph

Description automatically generatedA table with numbers and letters

Description automatically generated

Based on the information provided, it's evident that the number of duplicates present in our dataset constitutes approximately half of the total entries.

2. Successful data preprocessing, addressing data quality challenges with creation of relevant features to improve model accuracy. Univariate analysis separately explores the distribution of each variable in a data set. It looks at the range of values, as well as the central tendency of the values. Univariate data analysis does not look at relationships between various variables (like bivariate and multivariate analysis) rather it summarises each variable on its own. Methods to perform univariate analysis will depend on whether the variable is categorical or numerical. For numerical variable, we would explore its shape of distribution (distribution can either be symmetric or skewed) using histogram and density plots. For categorical variables, we would use bar plots to visualise the absolute and proportional frequency distribution. For categorical variables, we'll just checking the frequency distribution of the data using bar plot. Another way to show the relationships between classes or categories of a variable is in a pie or circle chart. In a pie chart, each "slice" represents the proportion of the total phenomenon that is due to each of the classes or groups. (The majority of pregnant women in this dataset appear to be at low risk for health issues. However, since it was decided to keep duplicates – outcome has changed slightly. Pregnant women 0 - Low Risk is updated now to 40.0% from 51.8%, 1 - medium risk is 33.1% compared to 23.5% and 2 - high risk is 26.8% new while 24.8% after removing duplicates.

A pie chart with numbers and a few percentages

Description automatically generatedFigure 3

To learn more and understand why pregnant women are at a different health risk, we will examine the data. Every variable that could have an impact on it will be examined. Let's examine each one separately.)

**Models and machine learning algorithms**

1. ***Model development*** in implementation of machine learning algorithms (e.g., KNN, LDA). Hyperparameter tuning for optimal model performance. Cross-validation helps in detecting and reducing overfitting, a common issue where a model performs well on the training data but poorly on new, unseen data. provides estimates of the model's performance metrics, such as accuracy, precision, recall, and F1-score, which are essential for assessing its effectiveness and reliability. In order to highlight the advantages and disadvantages of the nine machine learning algorithms —

Linear Discriminant Analysis, Decision Tree, Random Forest, Extreme Gradient Boosting, K-Nearest Neighbors (KNN), Gaussian NB, Support Vector Machine,Multi-layer Perceptron Classifier and Logistic Regression — this project evaluates their performance on chosen dataset. Hyperparameter tuning and performance assessment are carried out in three experiments divided into two phases.

2. ***User Interface*** - development of an intuitive and accessible interface. Integration of the predictive model into healthcare systems.

3. Comprehensive ***documentation*** of code and methodologies. Clear and transparent reporting of findings.

4. Different ***libraries*** have been used to perform different tasks and modeling of algorithms. These may include: Pandas, Numpy, Seaborn, Matplotlib, scipy, etc.

**Challenges Faced:**

Data Quality Issues. (Addressing duplicated data and outliers in the maternal health dataset. Iterative data enineering to ensure robustness.)

Model Complexity vs. Interpretability. (Balancing model complexity with the need for interpretability. Fine-tuning models to provide transparent and actionable results.)

Integration into Healthcare Systems. (Overcoming challenges in seamlessly integrating the model. Collaboration with IT professionals for successful implementation.

# **Data Attributes**

1. Age: Represents the age of a woman when she is pregnant, measured in years.
2. SystolicBP: Denotes the upper value of blood pressure in millimeters of mercury (mmHg), a significant attribute during pregnancy.
3. DiastolicBP: Represents the lower value of blood pressure in millimeters of mercury (mmHg), another crucial attribute in maternal health monitoring.
4. BS: Indicates blood glucose levels measured in terms of molar concentration, specifically in millimoles per liter (mmol/L).
5. BodyTemp: Denotes the body temperature of the pregnant woman in Fahrenheit (°F), providing additional insights into the physiological aspects relevant to maternal health.
6. HeartRate: Represents the normal resting heart rate in beats per minute (bpm), a vital parameter for assessing cardiovascular health during pregnancy.
7. RiskLevel (Target Variable): Predicted Risk Intensity Level during pregnancy, taking into account the aforementioned health attributes. This variable serves as the target for our machine learning models.

**Key Insights from Data Analysis:**

*Age and Health Outcomes.* Identified age groups associated with higher maternal health risks. Tailored interventions for specific age demographic Identified age groups associated with higher maternal health risks.

*Lifestyle Factors:* Analysed lifestyle variables contributing to health outcomes. Informed personalised care based on lifestyle patterns.

*Predictive Power of the Model.* Established the predictive model's accuracy and reliability. Validated model outcomes against real-world data.

**Descriptive statistics and Data visualisation**

In the descriptive statistics, we have provided an overview of our dataset using techniques such as examining the first few rows or generating summary statistics.

Figure 4

**A screenshot of a computer

Description automatically generated**

The provided results encompass statistical summaries of numerical features within the dataset, along with their correlation coefficients.

Figure 5

**A table of numbers with different colored numbers

Description automatically generated with medium confidence**

To gain insights into the distribution of our target variable, Risk Level, we need to visualise its frequency or distribution across the dataset.

Figure 6

A graph with blue squares

Description automatically generated

We have used the following mapping to easily convert our target variable, Risk Level, to numerical values:

"low risk": 0, "mid risk": 1, "high risk": 2

Numerical calculations and analysis on the target variable in our dataset are made easier by this mapping.

Figure 7

A graph of a graph showing the number of blue rectangular bars

Description automatically generated with medium confidence

During the course of this project, we gained valuable insights from the correlation analysis, which revealed the presence of multicollinearity within the dataset. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other. In this case, the high positive correlation between SystolicBP and DiastolicBP suggests that they contain redundant or overlapping information regarding blood pressure. One approach is to consider removing one of the highly correlated variables from the model if they are providing redundant information. Alternatively, techniques such as principal component analysis (PCA) will mitigate the effects of multicollinearity in our dataset. (Zach, 2019) Despite that fact that we retained duplicates in comparison to our JN file CA2. The notable correlation observed primarily lies between Systolic and Diastolic Blood Pressure (BP). This relationship extends to age an BS (blood sugar), a factor that can significantly impact maternity conditions, birth outcomes, and overall child health.

A **heatmap** visualisation helps to identify areas with high or low concentrations of data, providing insights into the distribution and relationships within the dataset.

Figure 8

A screenshot of a computer

Description automatically generated

Only two variables exhibit a higher correlation: diastolic and systolic blood pressure (BP) with Blood Sugar. Preeclampsia is a pregnancy complication that we can predict as a high risk at maternity risk levels based on symptoms and evidence found in our dataset. Pregnant mother may experience elevated blood pressure and high urine protein levels if she has preeclampsia. Preeclampsia can cause major, potentially fatal complications for both the mother and the child if it is not treated. (HSE.ie, n.d.) Consequently, we aim to delve deeper into understanding the intricate relationships between the variables to facilitate further analysis and insights.

A graph of a number of different colored boxes

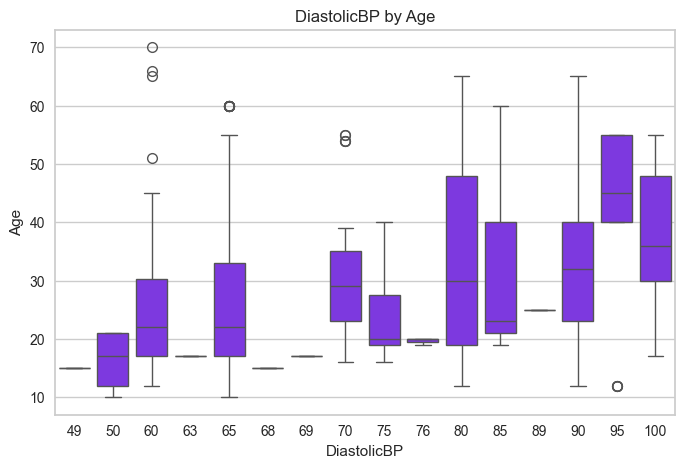
Description automatically generated with medium confidenceHere, we can visually compare the Diastolic and Systolic blood pressure readings with blood sugar levels and age. It's evident that there are numerous outliers present in the data, indicating potential anomalies or extreme values that may need further investigation.

Figure 9

Figure 10

A graph of a number of people

Description automatically generated with medium confidenceA graph with purple and white lines

Description automatically generated  
Observation: It appears that a skewed distribution is caused by outliers in almost all variables. We will just ignore that outlier for the time being as the value appears to be normal in this situation. There is an outlier in that variable whose value deviates too much from the rest of the range. While one option could be to consider dropping these outliers, I'm inclined to retain them for now to gain insights into the nature of the provided dataset.

The line plot (Figure 11) illustrates the distribution of individuals across different age groups. Each point on the plot represents the risk level of individuals within specific age categories. It provides insight into how the risk level varies across different age groups, highlighting potential trends or patterns in maternal health risk.

Figure 11

A graph showing the distribution of individuals across age groups

Description automatically generated

The majority of patient records fall within the age range of 20-29, with the next highest number in the 10-19 age group, followed by 30-39, 40-49, 50-59, and 60-69. The number of records for individuals aged 70 and above is significantly lower, which is expected.

**Correlation: Pair plot**

By plotting each variable against every other variable, the pair plot allows for the identification of any specific patterns or trends that may exist among the variables, aiding in the understanding of their relationships and potential impacts on the target variable.

Figure 12 Pair plot Distribution of the Maternal Health Risks dataset

A screenshot of a data visualization

Description automatically generated

The pair plot visualisation provides a comprehensive overview of the relationships between various health metrics, including diastolic and systolic blood pressure, blood sugar levels, and age. The plot reveals potential patterns, such as correlations or clusters, among these variables, which are valuable insights into maternal health risks.

**MinMaxScaler**

The correlation matrix analysis uncovers a substantial correlation between Blood Sugar (BS) levels and Risk Level, implying a robust association between these two variables within the dataset. This observation suggests that fluctuations or abnormalities in blood sugar levels may serve as significant indicators or predictors of maternal health risks.

Figure 13

A graph with different colored squares

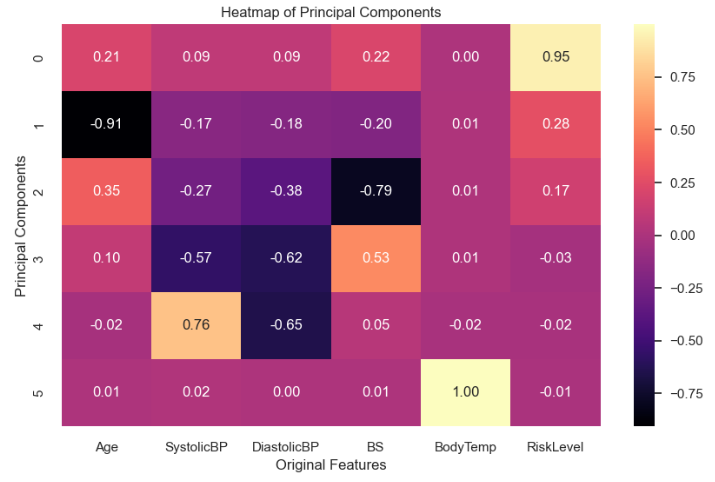
Description automatically generated

Correlation visualisation provides us with valuable evidence into the underlying mechanisms linking blood sugar dynamics to the overall risk profile during pregnancy, potentially informing targeted interventions or monitoring protocols for at-risk individuals.

**Principal Component Analysis (PCA)**

The heatmap generated from PCA results reveals a strong correlation between certain principal components and key variables such as Risk Level, Body Temperature, Systolic Blood Pressure (SBP), and Blood Sugar (BS).

Figure 14



This correlation suggests that these principal components effectively capture the variance associated with these critical health metrics. Specifically, high positive or negative correlations between principal components and these variables indicate that changes in these components correspond to significant variations in Risk Level, Body Temperature, SBP, and BS, highlighting their importance in characterizing the dataset's underlying structure.

**Statistical Shapiro-Wilk test**

The output of the Shapiro-Wilk test for Systolic Blood Pressure, Blood Glucose, and Age indicates that none of these variables are normally distributed, as evidenced by the p-values being less than 0.05. This suggests that these variables do not follow a Gaussian distribution. Instead, they exhibit significant departures from normality, which is ok with this dataset, as we have not treated outliers.

Figure 15

A graph showing a number of blue bars

Description automatically generated with medium confidence

A white background with black text

Description automatically generated

**Machine Learning Algorithms by Train and Test**

It was decided to train 9 different models using both the mean and the training, testing data, on the same dataset, allows us robust evaluation of their performance. This approach helps in comparing the efficacy of each model in handling outliers and duplicates, providing insights into their generalisation capabilities and potential overfitting tendencies. Additionally, by using both mean and training data, any discrepancies or biases introduced by outliers or duplicates can be effectively identified and addressed.

Among the classifiers tested, Random Forest, XGBClassifier and Decision Tree achieved the highest training accuracy (64.36%) and also exhibited the highest test accuracy (64.53%). This indicates that all three models performed the best in terms of both training and generalization on unseen data. Conversely, Logistic Regression showed the lowest performance with a training accuracy of 59.06% and a test accuracy of 57.14%. These results suggest that Random Forest, XGBClassifier or Decision Tree may be the most suitable classifier for this dataset, while Logistic Regression may not be as effective.

Figure 16

**A graph of different colored bars

Description automatically generated**

**Machine Learning Modelling by Mean**

The Maternal Health Risk Analysis project involved the evaluation of several machine learning algorithms to identify the most effective model for predicting maternal health risk levels. The following algorithms were considered: Logistic Regression (LR), Linear Discriminant Analysis (LDA), k-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Naive Bayes (NB), and Support Vector Machines (SVM), Random Forest (RF), MLPClassifier (MLP) and XGBClassifier (XGB). From these results, it can be concluded that XGBoost and Random Forest performed the best among the classifiers tested, with Random Forest having slightly higher accuracy but also a slightly higher standard deviation compared to XGBoost. While some classifiers performed better than others, the overall performance was reasonable across the board, indicating that machine learning models can effectively contribute to maternal health risk analysis.

Figure 17

A graph of a bar chart

Description automatically generated with medium confidence

In conclusion, the Maternal Health Risk Analysis project benefits from the promising predictive capabilities of XGBoost classifier. The final selection should consider the trade-offs between accuracy, interpretability, and computational efficiency, and additional fine-tuning may further optimise the chosen algorithm for deployment in real-world healthcare scenarios.

The robust performance of the Random Forest model in both testing scenarios indicates its potential as a reliable predictor for maternal health risks. By subjecting this model to further refinement through hyperparameter tuning and cross-validation, we can enhance its predictive accuracy and ensure its suitability for real-world deployment. This optimised model can then serve as a valuable tool for stakeholders and medical advisors, empowering them to make informed decisions and interventions aimed at improving maternal health outcomes.

**Fine-Tuning Hyperparameter**

Hyperparameter tuning helps in creating models that generalize well to unseen data. Models with carefully tuned hyperparameters are less likely to overfit or underfit the training data, resulting in better performance on new, unseen data.

Optimising hyperparameters aids in building models with strong data generalisation. Well-tuned hyperparameters reduce the likelihood of overfitting or underfitting the training set, which improves the model's performance on fresh, untested data. By systematically exploring the hyperparameter space and finding the best configuration for each algorithm, we can enhance their predictive power, improve generalisation, and ensure robustness across different datasets and problem domains. This approach will enable us to extract the maximum value from our models and produce more accurate and reliable predictions for maternal health risk analysis. (Guido, Groccia and Conforti, 2022)

Figure 18

A graph of a number of blue bars

Description automatically generated with medium confidence

While grid search helps in systematically exploring the hyperparameter space, cross-validation provides a more reliable estimate of model performance, especially when dealing with limited data or when the dataset is prone to variability.

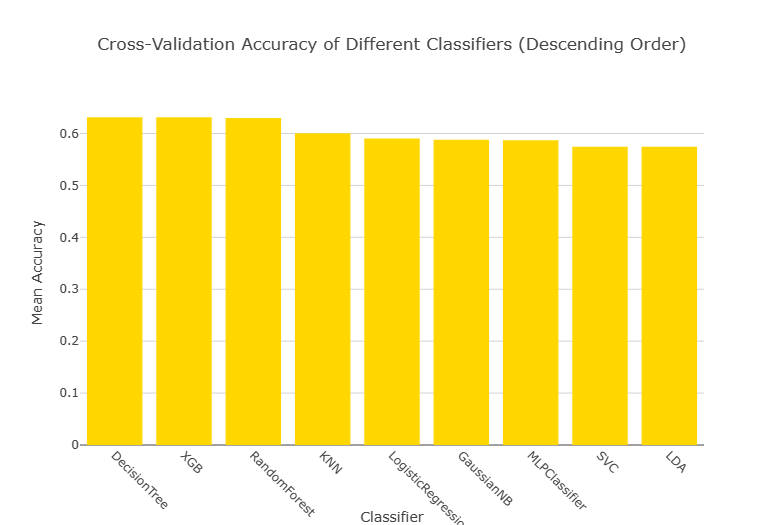
**Cross Validation**

The unexpected performance of the Decision Tree classifier, showing a slightly higher accuracy compared to Random Forest (RF) despite RF's generally superior performance, is indeed noteworthy. This outcome suggests that in this particular dataset, the inherent simplicity and interpretability of decision trees might be capturing the underlying patterns effectively, possibly due to the nature of the features or the relationships within the data.

While Random Forest is often favored for its robustness and ability to handle complex relationships by aggregating multiple decision trees, the specific characteristics of the dataset can sometimes lead to surprising results.

Understanding why certain models outperform others can provide valuable insights into the data and inform future modeling efforts. Additionally, it highlights the importance of not relying solely on mean cross-validation accuracies but also considering individual model performance and potentially exploring ensemble methods to leverage the strengths of different classifiers**.**

Figure 19

****

Decision Tree and XGBoost (XGB) indeed demonstrated the highest mean cross-validation accuracies among all classifiers, achieving approximately 63.13% respectively. These results indicate that both Random Forest and XGBoost are promising choices for modeling this dataset due to their consistently high mean accuracies. Additionally, Random Forest performed admirably well, with a mean cross-validation accuracy of around 83.48%, suggesting its effectiveness as a simpler yet competitive alternative to more complex ensemble methods like Random Forest and XGBoost. Overall, these findings underscore the importance of considering various classifiers and their respective performances to identify the most suitable model for a given dataset.

**Model Deployment and Results**

A strong positive correlation between recall and high risk means that as recall increases (i.e., the model becomes better at correctly identifying high-risk cases), the number of false negatives decreases, leading to a more accurate identification of high-risk instances. High precision for low-risk predictions in the confusion matrix suggests that the model is reliable in correctly identifying low-risk instances, minimizing the number of false positives in its predictions.

This enhances operational efficiency but also ensures timely and accurate identification of high-risk pregnancies, enabling healthcare providers to intervene proactively and potentially prevent adverse maternal outcomes. Organistions can adjust to shifting market conditions and new healthcare challenges by iteratively improving and updating deployed models based on feedback from the real world and changing healthcare trends. This will ultimately improve maternal health outcomes and maximize the impact of machine learning on healthcare delivery.

Figure 20

**A screenshot of a graph

Description automatically generated**

A graph showing different colored lines

Description automatically generatedA graph of a number of different colored lines

Description automatically generated with medium confidenceModel deployment is essential to converting the knowledge gathered from machine learning development into real advances in the provision of maternal healthcare in maternal health projects.

Figure 21 CA3 Result

Figure 22 CA2 Result

A graph showing a line graph

Description automatically generated with medium confidenceFigure 24 Result CA2A graph of a graph

Description automatically generated with medium confidence

Figure 23 Result CA3

Comparing CA2 and CA3, we observe significant differences and improvements in several aspects. In CA3, Decision Tree unexpectedly outperformed Random Forest and XGBoost in mean cross-validation accuracy, indicating a deeper understanding of the dataset and better model selection. This suggests advancements in data exploration, feature engineering, or insights gained from previous analyses. Moreover, CA3 highlights the importance of model deployment in real-world maternal healthcare settings, emphasising the practical implications of machine learning for improving healthcare delivery and patient outcomes. In general, the comparison showcases a progression in understanding, analysis, and application of machine learning techniques in addressing maternal health risks and enhancing healthcare outcomes.

**Proposed Solutions for Future Improvements:**

Collaborate with healthcare institutions or research organizations to acquire additional data and validate the model's predictive capabilities across diverse populations and settings.

Use cutting-edge strategies to reduce prediction bias. Update the model frequently to account for changing healthcare dynamics.

Regular updates to adapt to evolving healthcare practices, patient demographics, and risk factors.

**Conclusion:**

The Maternal Health Risk Analysis project successfully navigated through various phases, achieving milestones and gaining valuable insights. Challenges were addressed with a collaborative and adaptive approach. The key findings and limitations provide a foundation for future enhancements, ensuring continuous improvement in predictive maternal healthcare analytics. The project contributes to the broader goal of leveraging technology for improved healthcare outcomes.

Based on the results obtained from cross-validation and grid search, as well as the evaluation of various classifiers on both training and testing and mean data, we can draw the following conclusions. The mean cross-validation accuracy obtained across different folds is approximately 63.13%, with a standard deviation of 3.67%. This indicates that the models generally perform consistently across different subsets of the data. Tuned models achieved varying levels of accuracy, with XGBoost and Decision Tree models achieving the highest accuracy of approximately 63.13%, followed by Random Forest with an accuracy of 63.00%. Logistic Regression, on the other hand, achieved a lower accuracy of 58.68%. The performance of the tuned classifiers on unseen test data varied. XGBoost and Decision Tree models maintained their high accuracy on the test set, while Logistic Regression exhibited a lower accuracy of 54.19%. In general, the XGBoost and Decision Tree models emerged as the top performers, demonstrating robustness in terms of accuracy across both training and testing datasets. These models are recommended for further consideration due to their superior performance.

Moreover, there is a noteworthy positive correlation between BS and Age, Systolic and Diastolic blood pressure, suggesting caution for expectant mothers with elevated levels of these blood pressure parameters, which leads to hypertension (pre-eclampsia and eclampsia). (HSE.ie, n.d.)

The project provides actionable insights for healthcare professionals in predicting and managing maternal health risks. As for valid suggestions, it would be crucial to implement a system for continuous monitoring of the model's performance in real-world scenarios and update the model as needed, which provides healthcare professionals with insights into feature importance to enhance interpretability and trust in the model's predictions. Also, as recommendation I would suggest collaboration between data scientists and healthcare professionals for better integration of the model into clinical workflows. By incorporating these suggestions, it can further enhance the effectiveness, interpretability, and real-world impact of the Maternal Health Risk Analysis project and the main - it will help to maintain healthy pregnancy and labouring strong children.

**Recommendations:**

*Emphasis on Hypertension Management:*

Implement targeted interventions and regular monitoring protocols to control blood pressure levels, reducing the incidence of hypertensive disorders and associated complications during pregnancy.

*Coordinated Observation:*

Put in place an integrated monitoring system that takes into account several vital signs at once. This all-encompassing method can offer a more thorough comprehension of maternal health.

*Age-Group-Specific Interventions:*

Adapt medical interventions to different age groups. Programs targeted at women who are 25 years of age or older may be helpful in reducing possible health hazards.

*Exploration of Additional Health Metrics:*

Conduct further research to identify novel features or parameters that can provide valuable insights into maternal health status and aid in early risk detection and prevention strategies.

*Additional Heart Rate Research:*

Investigate other features or indicators that might improve heart rate's ability to predict health risks in future studies.

In conclusion, these insights can inform strategic interventions that can lead to better outcomes for maternal healthcare.

**Citations and Bibliography:**

1. Seaborn.countplot. [Online]. Available at: https://seaborn.pydata.org/generated/seaborn.countplot.html [Accessed: 30 November 2023].

2. Gupta, B. K., 2023. Scaler.com. [Online] Available at: <https://www.scaler.com/topics/machine-learning/missing-value-in-machine-learning/> [Accessed 7 December 2023].

3. Scikit Learn, 2023. scikit-learn.org. [Online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html> [Accessed 8 December 2023].

4.  COLORMAP reference. [Online]. Available at: <https://matplotlib.org/stable/gallery/color/colormap_reference.html>

[Accessed: 11 December 2023].

5. Real Python. (2023). Python enumerate(): Simplify loops that need counters. Real Python, Real Python. [Online]. Available at: <https://realpython.com/python-enumerate/>

[Accessed: 11 December 2023].

6. Elflein, J. (2023). Birth rate 10-14 year-old girls U.S. 1991-2021. Statista. [Online]. Available at: <https://www.statista.com/statistics/410744/birth-rate-for-us-girls/>

[Accessed: 14 December 2023].

7. Branham, R.L. (1990) Sparse matrices, SpringerLink. Available at: https://link.springer.com/chapter/10.1007/978-1-4612-3362-6\_3 (Accessed: 31 March 2024).

8. Kumar, A. (2023). Classification problems in Machine Learning: Examples. Analytics Yogi. [Online]. Available at: https://vitalflux.com/classification-problems-real-world-examples/ [Accessed: 31 March 2024].

9. note.nkmk.me. (2024). pandas: Get summary statistics for each column with describe() | note.nkmk.me. [online] Available at: https://note.nkmk.me/en/python-pandas-describe/ [Accessed 13 Apr. 2024].

10. Stack Overflow. (n.d.). How to plot only the lower triangle of a seaborn heatmap? [online] Available at: https://stackoverflow.com/questions/57414771/how-to-plot-only-the-lower-triangle-of-a-seaborn-heatmap [Accessed 23 Apr. 2024].

11. Bonnar, J., MacGillivray, I. and Symonds, E.M. (2012). Pregnancy Hypertension. Springer.

12. Ashworth, D.C., Maule, S.P., Stewart, F., Nathan, H.L., Shennan, A.H. and Chappell, L.C. (2020). Setting and techniques for monitoring blood pressure during pregnancy. Cochrane Database of Systematic Reviews. doi:https://doi.org/10.1002/14651858.cd012739.pub2.

13. docs.dataprep.ai. (n.d.). Installation — DataPrep 0.4.0 documentation. [online] Available at: https://docs.dataprep.ai/installation.html [Accessed 27 Apr. 2024].

14. World Health Organization (2023). Maternal Mortality. [online] World Health Organization. Available at: https://www.who.int/news-room/fact-sheets/detail/maternal-mortality.

15. HSE.ie. (n.d.). Pre-eclampsia. [online] Available at: https://www2.hse.ie/conditions/pre-eclampsia/.

16. Zach (2019). A Guide to Multicollinearity & VIF in Regression. [online] Statology. Available at: https://www.statology.org/multicollinearity-regression/.

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

18. Guido, R., Groccia, M.C. and Conforti, D. (2022). A hyper-parameter tuning approach for cost-sensitive support vector machine classifiers. Soft Computing. doi:https://doi.org/10.1007/s00500-022-06768-8.

19. HSE.ie. (n.d.). Pre-eclampsia. [online] Available at: https://www2.hse.ie/conditions/pre-eclampsia/.

**GitHub link:**

<https://github.com/AlesUA23/CA3_Strat_Think_sba23334>