KZ Khumalo 222248893

ABSTRACT

Maternal health models are important tools in the global effort to identify and reduce potential risks during pregnancy, with the aim of lowering maternal mortality rates. My main objective is to contribute to ongoing research in this domain by developing a model that can assist medical professionals in identifying high-risk pregnancies early, allowing for timely intervention and improved outcomes.

I began the research by preprocessing my data to ensure its clean to avoid any errors and inaccurate predictions.

Pre-processing:

The dataset contained no null values and had consistent data types across all columns. However, a significant number of duplicate rows were identified, many of which involved entire rows with identical values across columns. These duplicates were removed to avoid compromising the model's accuracy.

Initially, I removed the 'Body Temperature' feature because I assumed it provided minimal predictive value. However, upon observing a drop in performance—particularly in the model's ability to correctly classify high-risk cases—I added it back into the study, leading to improved high-risk predictions.

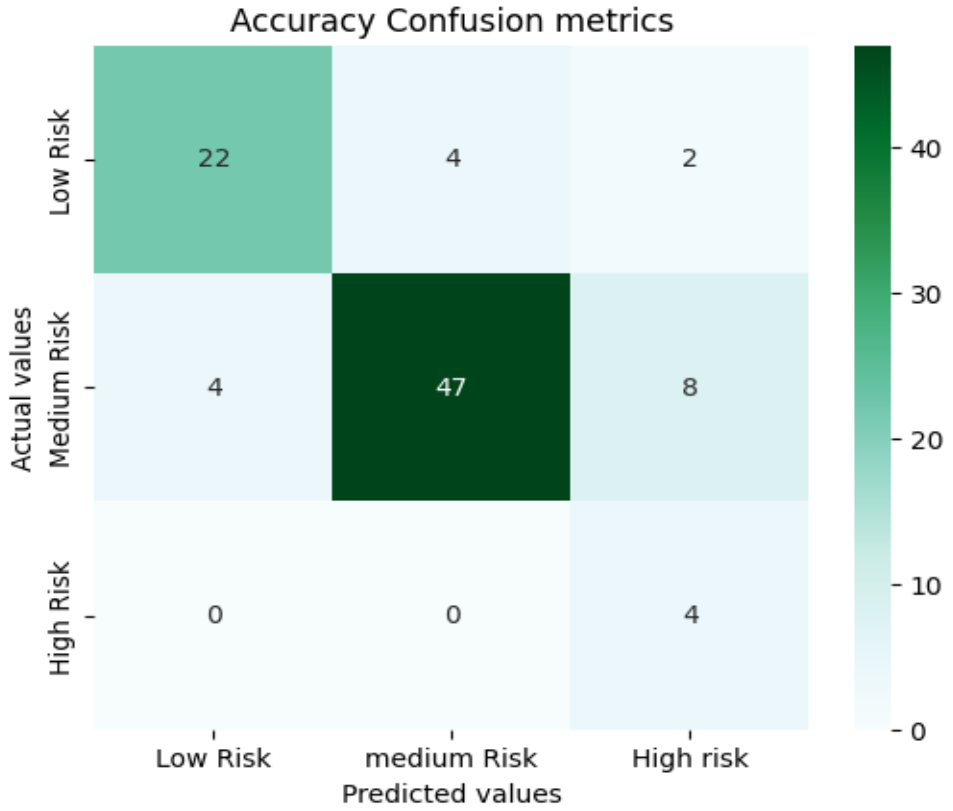
Models used to conduct my study:

- **Decision Tree**: Initially considered for its simplicity and tolerance for less clean data but ultimately dismissed due to poor performance and reduced accuracy.
- **Random Forest**: Chosen for its ensemble structure, which gave better results. It helped to avoid overfitting which is a common problem with decision trees. This model produced an accuracy of approximately **78%**, correctly identifying most medium-risk cases but struggling with high-risk classification.
- **Gradient Boosting Classifier**: added to address the weaknesses of the Random Forest and improve performance through sequential learning (corrects previous mistakes).
- **Stacking Classifier**: I used it to combine predictions from both Random Forest and Gradient Boosting. Logistic Regression was used as the final estimator, which improved overall accuracy to **80.2%**.

  I decided to combine all these models to harvest off their unique advantages and to help reduce their weaknesses thus improving the overall accuracy of the model.

Evaluation through a **confusion matrix** showed that while medium-risk predictions were most accurate, there were still a few misclassifications—specifically, **two low-risk individuals being classified as high-risk**, raising curiosity about possible data inconsistency or imbalance because it's too high of a spectrum.

Attempts to fine-tune the model's performance using different train_test_split ratios revealed that a 20% test set consistently provided the highest accuracy.



KEY TAKEAWAYS

Through exploration I confirmed that feature selection, model stacking, and attention to false positives—especially in critical categories like high-risk pregnancies—are essential to developing a clinically valuable prediction system.

I was satisfied with the 78% accuracy until I studied the confusion that revealed that the main category, we are aiming to predict (high risks) was actually predicted less accurately than the others.

This project highlights the potential of machine learning in supporting maternal healthcare and underscores the importance of careful preprocessing, model comparison, and iterative tuning to achieve reliable and actionable outcomes.