III METHODOLOGY

This section sheds a light on research methodology ,along with data collection , preprocessing , model development and it's evaluation technique which were used for analysis and risk factor prediction.

A DATA COLLECTION :

We used the maternal health risk factors dataset [28] for this study, which we got from the UCI machine learning public repository. The majority of the 1014 rows and 7 columns in this dataset reflect pregnant women who pose minimal health risks. Of the 1014 observations, 406 (40%) were classified as low risk, 336 (33.1%) as medium risk, and 272 (26.8%) as high risk pregnant women. After that, we balanced the dataset using categorical encoding and hypertuned it, yielding a total of 1218 observations. The following provides a full description of each dataset attribute:

Age: Represents the age of pregnant women in years.

Systolic BP: Denotes the maximum blood pressure in millimeters of mercury, a crucial parameter during pregnancy.

Diastolic BP: Indicates the lower blood pressure measurement in millimeters of mercury, another vital consideration during pregnancy.

BS: Refers to the amount of glucose in the blood, measured in mmol/L.

Heart rate: Represents the normal heart rate in beats per minute.

Risk Level: Signifies the intensity level of risk prediction during pregnancy, dependent on the preceding attributes.

In this context, the target variable is the risk level, while the remaining features serve as predictor variables.

RISKS N THEIR LEVEL FIGURE

B DATA PREPROCESSING :

The initial phase in enhancing the predictive model involves ensuring that the data is meticulously prepared for analysis. The effectiveness of the model is substantially heightened when the data undergoes appropriate transformations.The dataset exhibited a commendable absence of missing values, reflecting a high degree of data completeness.. But when we looked at the kind of pregnancies in our dataset, we noticed that only 26.8% of them were considered high-risk.This imbalance might affect our analysis, so we need to be careful when drawing conclusions about risk factors.

SCREEN SHOT OF IMBALANCE DATA SET

\*\*Categorical Encoding\*\*\*

In the preprocessing phase, we conducted categorical encoding on the 'RiskLevel' variable, converting it into numerical values. This transformation is crucial for machine learning algorithms, which typically require numerical inputs for optimal performance. By ensuring compatibility and effectiveness in subsequent analyses, this encoding step facilitates seamless integration of the 'RiskLevel' variable into our machine learning models, contributing to improve their overall accuracy and predictive capabilities.

\*\*\*Feature Selection\*\*\*

In the process of feature selection, a strategic decision was made to eliminate the 'DiastolicBP' column from our dataset. Feature selection is a crucial preprocessing technique in machine learning aimed at optimizing model performance. By carefully choosing relevant features and discarding less impactful ones, we can streamline our data, reducing complexity and potentially enhancing the predictive capabilities of our models. This approach is geared towards improving model efficiency and accuracy by focusing on the most informative variables, a fundamental practice in machine learning engineerin

\*\*\* Handling Class Imbalance\*\*\*

Addressing class imbalance is pivotal for an unbiased model. The Synthetic Minority Over-sampling Technique (SMOTE) was employed to rectify the class imbalance in the target variable 'RiskLevel'. SMOTE adeptly generates synthetic samples for the minority class, guaranteeing a balanced representation within the dataset.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Model Development \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Train-Test Split:

The dataset was split into two parts for machine learning purposes: 80% for training the model and 20% for testing its performance. This allows the model to learn from most of the data and then evaluate its effectiveness on unseen data, a standard practice in machine learning.

TABLE OF TEST TRAIN GIVEN BY SIR

Model Development :

Several machine learning methods, including Random Forest, Support Vector Classification (SVC), and XGBoost, were initially considered. Following a thorough evaluation, the Random Forest Classifier emerged as the top performer among these methods.

Random Forest Classifier:

The Random Forest Classifier, our chosen predictive model, excels as an ensemble learning method by aggregating insights from multiple decision trees, thereby enhancing predictive accuracy and addressing overfitting concerns.

Hyperparameter Tuning:

Optimizing the Random Forest Classifier involved the implementation of Grid Search, an advanced hyperparameter optimization technique. This approach systematically explored various hyperparameter configurations, with a focus on key parameters like the number of estimators (trees), maximum tree depth, minimum samples for node splitting, and minimum samples for a leaf. The goal was to fine-tune these parameters for the optimal model performance.

Model Training:

Following hyperparameter tuning, the Random Forest Classifier underwent intensive training using the preprocessed dataset. The optimal hyperparameter configuration derived from the Grid Search process was applied, ensuring the model is finely tuned for superior performance. This well-trained classifier forms the cornerstone for subsequent evaluation and testing stages, showcasing its efficacy in handling real-world data with precision and reliability.

\*\*\*\*\*\*\*\*\*\*\*\*\* Model Evaluation \*\*\*\*\*\*\*\*\*

Performance Metrics:

Evaluation of the model's performance was conducted using fundamental classification metrics, namely accuracy, precision, and recall. Accuracy gauges overall correctness, precision assesses the relevance of positive predictions, and recall evaluates the model's capability to capture positive instances.

Cross-Validation:

To fortify the model evaluation process and minimize the risk of overfitting, a robust 5-fold cross-validation strategy was employed during the hyperparameter tuning phase. This involved partitioning the dataset into five subsets, training the model on four subsets, and evaluating on the fifth. The process was iterated five times, and the results were averaged to provide a comprehensive assessment.

Results Reporting:

The conclusive reporting encompassed the model's accuracy, precision, recall, and an intricate classification report. The classification report served to offer nuanced insights into the model's performance across distinct classes, providing a holistic evaluation perspective.

Results And Discussion:

The table below provides the accuracy percentages for various machine learning algorithms used in predicting maternal health risks along with their Precision, Recall and F1 Score.

TABLEEE

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*CONCLUSION\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

This research endeavors to forecast maternal health risks through the application of diverse machine learning algorithms. Our study reveals that the random forest with hyperparameterization has exhibited remarkable efficacy compared to alternative models.

Through comprehensive comparison with existing methodologies in the literature, our proposed model stands out with an exceptional accuracy rate of 94.26%.

These findings underscore the potential of machine learning in enhancing prenatal care by enabling early identification and mitigation of maternal health complications. The success of our model signifies a significant stride toward personalized healthcare interventions tailored to expectant mothers. By leveraging advanced computational techniques, healthcare practitioners can potentially preemptively identify and mitigate risks, thereby enhancing prenatal care outcomes and maternal well-being.

\*\*\*\*\*\*\*\*\*\*\*\*