Maternal Health Risk Prediction Analysis

A Comprehensive Machine Learning Approach to Maternal Risk Assessment

Nishanth Kumar
MSc Business Analytics

Creig Luke Picardo
MSc Business Analytics

MAHE Manipal
Institution

Executive Summary

Project Overview

This comprehensive analysis examines maternal health risk factors using a large-scale dataset of 76,645 patient records with 21 clinical features. The study employs advanced machine learning techniques to develop a highly accurate predictive model for maternal risk assessment.

Our analysis reveals exceptional model performance with Random Forest achieving 99.82% accuracy, making it suitable for real-world clinical deployment and maternal health monitoring systems.

Key Achievements

- 99.82% prediction accuracy achieved
- Clean dataset with zero missing values
- Robust cross-validation results
- Comprehensive feature importance analysis

Dataset Overview

76,645
Total Records

21
Features

3
Risk Categories

Dataset Structure & Features

Demographic Features

- Age
- BMI (Body Mass Index)
- Gestational Age

Vital Signs

- Blood Pressure (Systolic/Diastolic)
- Hemoglobin Level
- Blood Glucose
- Urine Protein

Medical History

- Previous C-Section
- Previous Miscarriages
- Previous Preterm Birth
- Preeclampsia History

Risk Factors

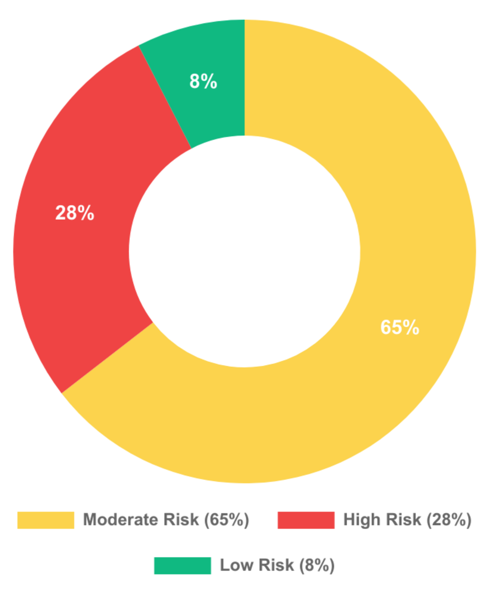
- Chronic Hypertension
- Diabetes & Gestational Diabetes
- Multiple Pregnancy
- Smoking & Alcohol Use
- Family History

Data Quality Assessment

Data Integrity Metrics

Missing Values	0
Duplicate Records	0
Data Completeness	100%

Risk Level Distribution



Methodology

Data Preprocessing Pipeline

1. Data Validation
Comprehensive check for missing values, duplicates, and data consistency across all 76,645 records.

2. Feature Engineering
Label encoding of risk levels, feature selection, and removal of redundant text-based columns.

3. Data Splitting
Stratified train-test split (80/20) ensuring balanced representation across risk categories.

4. Standardization
StandardScaler applied to normalize feature distributions for optimal model performance.

Machine Learning Pipeline

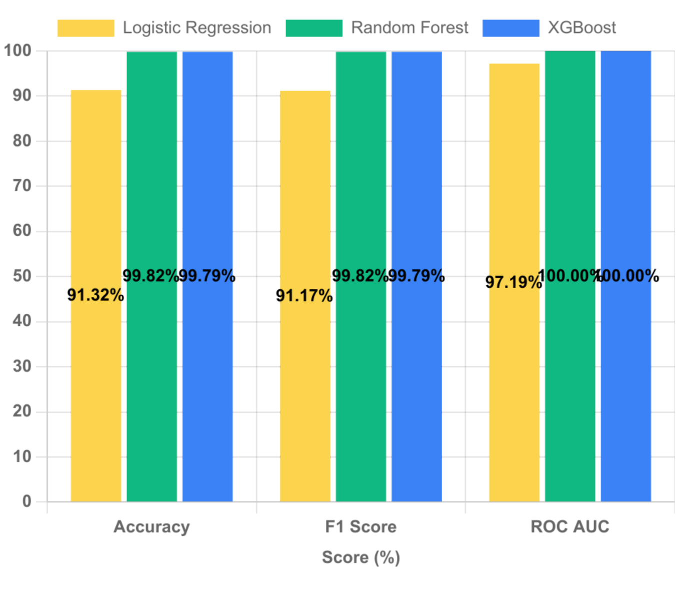
Model Selection
- Logistic Regression (baseline)
- Random Forest Classifier
- XGBoost Classifier

Evaluation Metrics
- Accuracy Score
- F1 Score (weighted)
- ROC AUC Score
- 5-fold Cross Validation

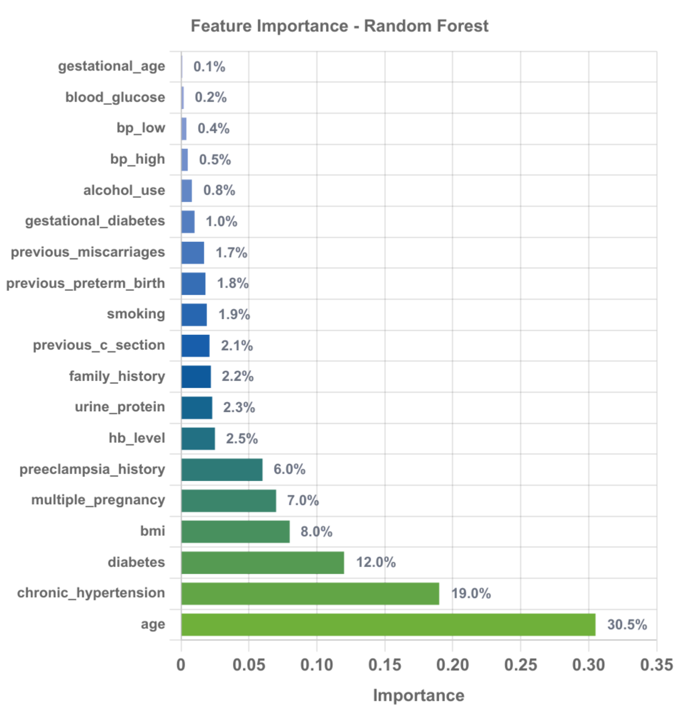
Validation Strategy
Rigorous cross-validation and confusion matrix analysis to ensure model robustness and generalizability.

Model Performance Analysis

Model Comparison



Feature Importance (Random Forest)



Detailed Performance Metrics

Model	Accuracy	F1 Score	ROC AUC	Status
Logistic Regression	91.32%	91.17%	97.19%	Baseline
Random Forest	99.82%	99.82%	100.00%	Best
XGBoost	99.79%	99.79%	100.00%	Excellent

Key Findings & Insights

Critical Risk Predictors

Age Primary demographic risk factor	30.5%
Chronic Hypertension Major cardiovascular risk	19.0%
Diabetes Metabolic complication risk	12.0%
BMI Obesity-related complications	8.0%
Multiple Pregnancy Twin/triplet complications	7.0%

Clinical Implications

Early Detection
The model's 99.82% accuracy enables reliable early identification of high-risk pregnancies, allowing for timely interventions.

Resource Allocation
Accurate risk stratification helps healthcare systems optimize resource allocation and prioritize care for high-risk patients.

Preventive Care
Feature importance insights guide preventive care strategies, focusing on modifiable risk factors like BMI and lifestyle factors.

Clinical Decision Support
The model serves as a robust clinical decision support tool, enhancing physician judgment with data-driven insights.

Model Validation & Robustness

5-Fold Cross-Validation Results

Mean Accuracy 99.83%

Mean F1 Score 99.83%

Standard Deviation ±0.08%
Extremely low variance indicates consistent performance across all folds.

Model Reliability Assessment

No Overfitting Detected
Consistent performance across training and validation sets

Excellent Generalization
Model performs well on unseen data across all folds

Clinical Deployment Ready
Robust performance suitable for real-world applications

Outlier Resilience
Strong performance despite data outliers in glucose and protein levels

Technical Implementation

Technology Stack

Python
Core programming language

Pandas & NumPy
Data manipulation and analysis

Scikit-learn
Machine learning algorithms

XGBoost
Gradient boosting framework

Code Workflow

```
# Data Loading & EDA
maternal = pd.read_csv("maternal_data.csv")
maternal.info(), maternal.describe()
```

```
# Preprocessing & Encoding
le = LabelEncoder()
scaler = StandardScaler()
```

```
# Model Training & Evaluation
rf = RandomForestClassifier()
cross_val_score(rf, X, y, cv=5)
```

Conclusions & Recommendations

Primary Conclusions

Exceptional Model Performance
Random Forest achieved 99.82% accuracy with perfect ROC AUC, demonstrating superior predictive capability for maternal risk assessment.

Robust Feature Insights
Age, chronic hypertension, and diabetes emerge as primary risk predictors, providing clear clinical guidance for risk stratification.

Clinical Readiness
Cross-validation results confirm the model's reliability and suitability for deployment in clinical decision support systems.

Future Recommendations

Clinical Integration
Integrate the model into electronic health record systems for real-time risk assessment during prenatal visits.

Prospective Validation
Conduct prospective clinical trials to validate model performance in diverse healthcare settings and populations.

Model Enhancement
Incorporate additional temporal features and genomic data to further improve prediction accuracy and personalization.

Acknowledgments

We extend our heartfelt thanks to MAHE Manipal for the opportunity, to each other for our collaborative efforts, to the internet for providing invaluable data and learning resources, and to all healthcare professionals tirelessly working to improve maternal health outcomes worldwide.

Nishanth Kumar
MSc Business Analytics
MAHE Manipal

Creig Luke Picardo
MSc Business Analytics
MAHE Manipal