



Early Detection of Preeclampsia using Machine Learning Algorithm.

by

Fazlay Rabby (201016001)
Fabliha Mahazabin Ahana (201016002)
Nusrat Jahan Lammy (201016004)
Julian Thoma Gomes (201016011)

A project report submitted in partial fulfillment of the requirements for the degree
of Bachelor of Science in Electrical and Electronic Engineering

Supervisor: Abul Barkat Mollah Sayeed Ud Doulah.,
Associate Professor, Dept of EEE, ULAB.

Department of Electrical and Electronic Engineering
University of Liberal Arts Bangladesh
688 Beribadh Road
Mohammadpur, Dhaka -1207,
Bangladesh.

5 May, 2024

DECLARATION

This Final year Capstone Project report is submitted to the department of Electrical and Electronic Engineering, University of Liberal Arts Bangladesh in partial fulfillment of the requirements for the degree of Bachelor of Science. So, we hereby declare that this report is based on the work done by us. Materials of work found by other researchers are mentioned by reference. This project report, neither in whole, nor in part, has been previously submitted for any degree.

1 Name:

Signature
:

Date:

2 Name:

.

Signature
:

Date:

3 Name:

.

Signature
:

Date:

4 Name:

.

Signature
:

Date:

SUPERVISOR’S CERTIFICATION / APPROVAL

The Final year Capstone Project report entitled “Early Detection of Preeclampsia Using Machine Learning Algorithm” is submitted to the department of Electrical and Electronic Engineering, University of Liberal Arts Bangladesh, Dhaka in partial fulfillment of the requirements for the degree of Bachelor of Science.

Dr. ABM Sayeed Ud Doulah

Supervisor

Associate Professor
Department of Electrical and Electronic
Engineering

University of Liberal Arts Bangladesh

Signature

Date:

Dr. Mirza Rasheduzzaman

Acting Head of Department

Associate Professor
Department of Electrical and Electronic
Engineering

University of Liberal Arts Bangladesh

Signature

Date:

ACKNOWLEDGEMENT

First and foremost, we would like to thank our supervisor Dr. ABM Sayeed Ud Doulah, Associate Professor and Coordinator of Electrical and Electronic Engineering at the University of Liberal Arts Bangladesh (ULAB), for setting up the completion of this Final Year Capstone Project for us as well as for his excellent guidance, encouragement, patience and kindness. We also want to thank Dr. Mirza Rasheduzzaman, Acting Head and Associate Professor of Electrical and Electronic Engineering, and our department's esteemed faculty members, in particular, Dr. Md. Shahjalal, Dr. Litton Hossain, are to be thanked for their guidance, co-operation and lectures during the study. We also want to express our gratitude to our family and friends, who supported and encouraged us emotionally during the endeavor. Our motivation and focus have been maintained in large part by their unfailing support. Finally, we would like to thank everyone who took part in the research and gave us their insightful opinions and input, without whose help this project would not have been feasible. We appreciate your help and support with this initiative.

Abstract

Preeclampsia is a serious pregnancy complication that poses significant risks to both maternal and fetal health worldwide. Early detection and intervention are crucial to prevent adverse outcomes and improve health outcomes for mothers and infants. This study aims to develop advanced machine learning models to predict preeclampsia risk using data from online datasets and local hospital records. The study employs various machine learning techniques, including k-nearest neighbors (KNN), logistic regression (LR), and support vector machines (SVM), to analyze medical history, vital signs, and clinical records for predictive modeling. Data preprocessing involves cleaning, transforming, and organizing the dataset for compatibility with the machine learning algorithm. The models are trained using techniques such as K-fold cross-validation and leave-one-out cross-validation to improve generalizability and robustness. Performance evaluation criteria, including F1-score, recall, accuracy, and precision, are used to assess the models' effectiveness in predicting preeclampsia. Results show that the KNN, LR, and SVM models demonstrate promising accuracy and effectiveness in early detection, providing healthcare professionals with valuable tools for vigilant monitoring and prompt intervention. These findings hold potential for improving prenatal care, particularly in low- and middle-income countries where resources may be limited. This study contributes to the broader goal of enhancing maternal and fetal health by developing reliable and precise early detection models for preeclampsia.

Keywords- preeclampsia, k-nearest neighbors (KNN), logistic regression (LR), support vector machines (SVM), K-fold cross-validation, leave-one-out cross-validation, F1-score.

Table of Contents

CHAPTER	TITLE	PAGE
	Title Page	i
	Declaration	ii
	Supervisor's Approval	iii
	Acknowledgement	iv
	Abstract	v
	Table of Contents	vi
	List of Tables	vii
	List of Figures	viii
	List of Abbreviations	ix
1	Introduction	1
	1.1 Background and Motivation	3
	1.2 Objective	4
2	Literature Review	6
3	Methodology	26
	3.1 Data Acquisition	26
	3.2 Data Processing, Labeling and Selection	29
	3.2.1 Data Labeling	30
	3.2.2 Data Processing and Selection	31
	3.3 Data Validation	33
	3.3.1 Univariate Analysis	35
	3.3.2 Bivariate Analysis	38
	3.4 Feature Selection	43
	3.5 Standardization	45
	3.6 Cross Validation	47
	3.6.1 K Fold Cross-validation	48
	3.6.2 Leave One Out Cross-validation	48
	3.7 Classification	49
	3.7.1 K-Nearest Neighbor	49
	3.7.2 Logistic Regression	50
	3.7.3 Support Vector Machine	50
	3.7.4 Main Model	51
	3.8 Performance Evaluation	54
4	Results and Discussion	55
	4.1 Confusion Matrix	56
	4.2 ROC Curve Comparison	60
5	Future work and Limitations	63
	5.1 Limitations	63
	5.2 Future work	63
	5.3 Conclusion	64
6	CO-PO mapping	66
	References	68
	APPENDIX A	73
	APPENDIX B	80

List of Tables

Table 1: Performance of ML models

List of Figures

Fig. 1.1. Signs and symptoms of preeclampsia.

Fig. 1.2. A dysfunctional placenta causing maternal and fetal morbidity through altered angiogenic and antiangiogenic factors.

Fig. 3.1. Gestational hypertension

Fig. 3.2 Target distribution

Fig. 3.3 Box plot (a) with outliers and (b) without outliers.

Fig. 3.4 Density Curve

Fig. 3.5 KDE plot

Fig. 3.6 Bar plot

Fig. 3.7 Correlation matrix

Fig 3.8. Online data histogram

Fig. 3.9. p values from Chi-Square Test.

Fig. 3.10. Ranges of values from the online dataset

Fig. 3.11. Ranges of standardized values from the online dataset

Fig. 3.12. Ranges of standardized values from the offline dataset

Fig. 3.13. Flow chart of the methodology of detection of preeclampsia for the online data set.

Fig. 3.14. Flow chart of the methodology of detection of preeclampsia for the local data set.

Fig. 4.1. Confusion matrix for the KNN model.

Fig. 4.2. Confusion matrix for the LR model.

Fig. 4.3. Confusion matrix for the SVM model.

Fig. 4.4. ROC curve comparison for three different machine learning models.

List of Abbreviations

Placental Dysfunction-Related Disorders (PDDs)

Machine Learning (ML)

World Health Organization (WHO)

Preeclampsia (PE)

Blood Pressure (BP)

Blood Urea Nitrogen (BUN)

Creatinine Ratio (UPCR)

Aspartate Transaminase (AST)

Alanine Transaminase (ALT)

Systolic Blood Pressure (SBP)

Diastolic Blood Pressure (DBP)

Albumin-To-Creatinine Ratio (AC Ratio)

Protein/Creatinine Ratio (PC Ratio)

Comma Separated Values (CSV)

Exploratory Data Analysis (EDA)

Interquartile Range (IQR)

Least Squares Regression Line (LSRL)

Kernel Density Estimate (KDE)

Variance Inflation Factor (VIF)

Leave-One-Out Cross-Validation (LOOCV)

Mean Squared Error (MSE)

K-Nearest Neighbor (KNN)

Logistic Regression (LR)

Support Vector Machine (SVM)

True Positive (TP)

True Negative (TN)

False Positive (FP)

False Negative (FN)
False Negative Rate (FNR)
False Positive Rate (FPR)
True Positive Rate (TPR)
False Positive Rate (FPR)
Receiver Operating Characteristic (ROC)
Area Under The Curve (AUC)
Placental Dysfunction-Related Disorders (Pdds)
Machine Learning (ML)
World Health Organization (WHO)
Preeclampsia (PE)
Blood Pressure (BP)
Blood Urea Nitrogen (BUN)
Creatinine Ratio (UPCR)
Aspartate Transaminase (AST)
Alanine Transaminase (ALT)
Systolic Blood Pressure (SBP)
Diastolic Blood Pressure (DBP)
Albumin-To-Creatinine Ratio (AC Ratio)
Protein/Creatinine Ratio (PC Ratio)
Comma Separated Values (CSV)
Exploratory Data Analysis (EDA)
Interquartile Range (IQR)
Least Squares Regression Line (LSRL)
Kernel Density Estimate (KDE)
Variance Inflation Factor (VIF)
Leave-One-Out Cross-Validation (LOOCV)
Mean Squared Error (MSE)

K-Nearest Neighbor (KNN)

Logistic Regression (LR)

Support Vector Machine (SVM)

True Positive (TP)

True Negative (TN)

False Positive (FP)

False Negative (FN)

False Negative Rate (FNR)

False Positive Rate (FPR)

True Positive Rate (TPR)

False Positive Rate (FPR)

Receiver Operating Characteristic (ROC)

Area Under The Curve (AUC)

Chapter 1: Introduction

Preeclampsia, a grave pregnancy complication, poses significant threats to both maternal and fetal well-being on a global scale. Placental dysfunction-related disorders (PDDs), such as preeclampsia and intrauterine growth restriction, require that a referral choice be made quickly. Preeclampsia is the world's most common cause of maternal death and morbidity [1]. It is a hazardous medical condition that can develop about halfway through gestation (after 20 weeks) and is associated with substantial mortality and morbidity for the mother, the fetus, and the newborn [2]. The consequences of untreated or inadequately managed preeclampsia can be devastating, resulting in adverse outcomes for both mothers and infants. Each year, this condition claims the lives of hundreds of thousands of fetuses and newborns, alongside a significant number of mothers, highlighting the urgent need for effective prevention and management strategies. The disorders associated with placental dysfunction (PDDs) are conditions that call for prompt referral decision-making. There are two relevant examples of such conditions; Preeclampsia and intrauterine growth restriction. where the primary cause of maternal mortality and morbidity globally is preeclampsia [3]. By the third trimester of pregnancy, proteinuria and gestational hypertension starts to increase frequently [5]. Due to the condition's impact on mothers and limitations on fetal development, low birth weights increase the chance of neonatal mortality [6]. The symptoms of preeclampsia include headaches, blurred vision, increased blood pressure, urine containing protein, and stress on the mother's heart and other organs [7]. Additionally, this condition adversely affects the maternal blood supply, amplifying its potential for serious complications and underscoring the urgency of effective detection and management strategies.

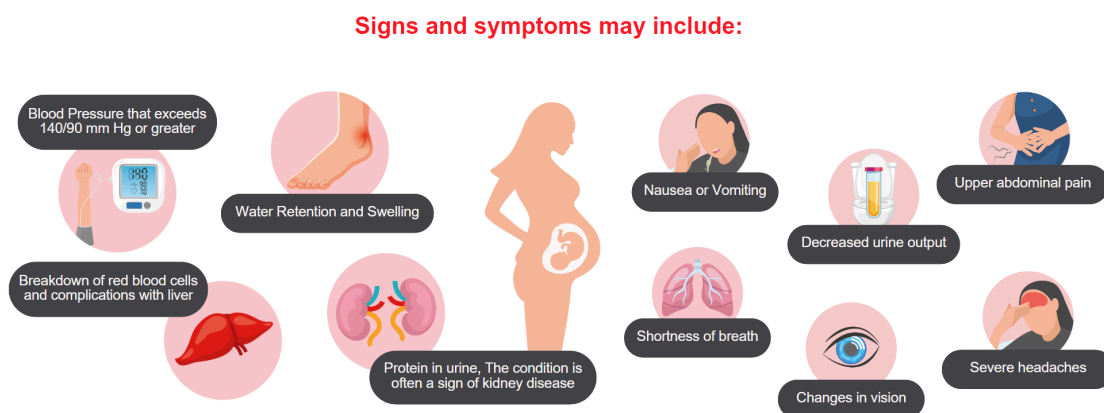


Fig. 1.1. Signs and symptoms of preeclampsia.

Preeclampsia is characterized by a constellation of symptoms, including high blood pressure (hypertension), fluid retention (edema), and an elevated protein level in the urine

(proteinuria) [10]. However, these indicators may not be readily apparent at the onset of pregnancy, making the early detection of preeclampsia particularly challenging. Nevertheless, the urgency of early detection cannot be overstated, as it allows for vigilant monitoring and prompt intervention, thereby mitigating the risks associated with this condition. Importantly, women with a history of preeclampsia in previous pregnancies face elevated risks in subsequent pregnancies, underscoring the critical need for effective screening and intervention strategies. It is imperative to develop precise detection models and algorithms meticulously engineered to identify individuals at risk of preeclampsia at the earliest possible juncture. By leveraging advancements in medical technology and data analytics, our primary mission is to enhance prenatal care by developing reliable and accurate prediction models for preeclampsia, ultimately improving outcomes for mothers and infants worldwide.

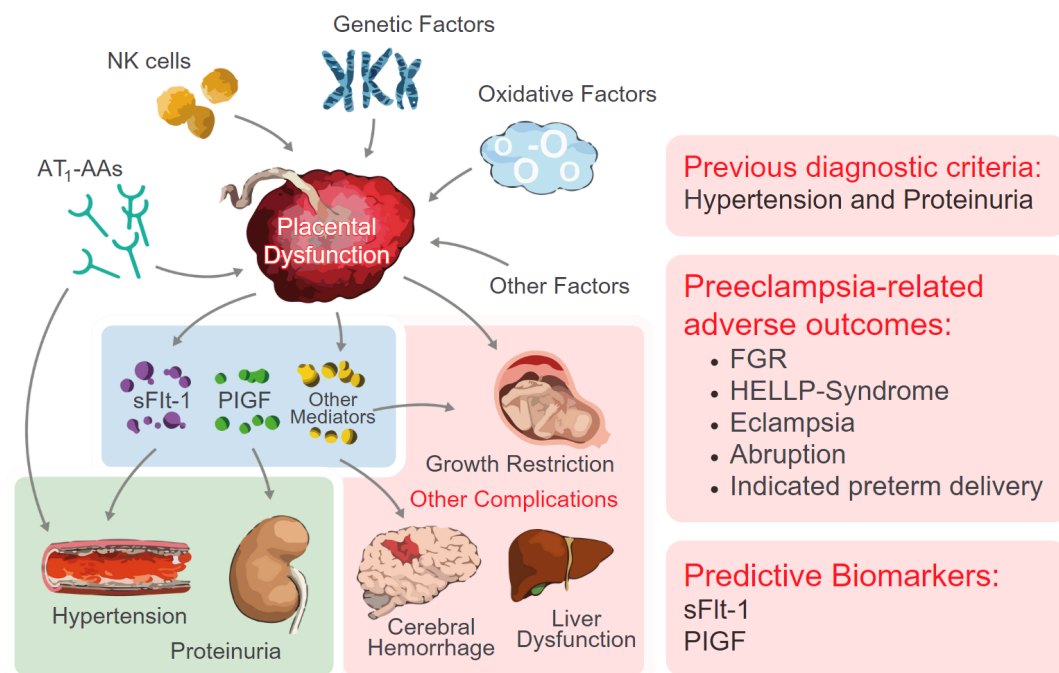


Fig. 1.2. A dysfunctional placenta causing maternal and fetal morbidity through altered angiogenic and antiangiogenic factors.

Preeclampsia is a condition characterized by a dysfunctional placenta, which leads to maternal and fetal symptoms stemming from an imbalance of angiogenic and antiangiogenic factors. While the exact causes of the initial placental disturbance are not entirely known, it is evident that factors such as genetics, immunology, and the environment play a role. Three key insights into the clinical outcomes of preeclampsia are:

Diagnostic Criteria: Traditional diagnostic criteria occur at the end of a lengthy pathophysiological process and represent just one of several forms of end-organ damage.

Maternal and Fetal Outcomes: The key drivers of maternal and fetal morbidity and mortality are the adverse effects and outcomes, which are downstream effects of placental dysfunction. Predicting these complications is crucial.

Angiogenic and Antiangiogenic Factors: These factors are central to the pathophysiological cascade and can improve the prediction and diagnosis of placenta-related complications.

As a result of the pressing need to enhance early detection and management of preeclampsia, considerable research efforts have been dedicated to the development of effective prediction techniques [11]. A reliable prediction algorithm holds the potential to streamline monitoring processes for individuals not predisposed to preeclampsia while enhancing surveillance for those at high risk. Despite prior investigations into clinical aspects and biomarkers for prediction [12–13], few studies have exhibited clinically adequate qualities, highlighting the ongoing challenges in this field. One promising avenue that has garnered significant attention is the application of machine learning (ML) techniques. ML methodologies offer the capability to uncover intricate relationships between disparate data items, which may be challenging to identify using conventional techniques [14, 15]. Given the vast and complex nature of medical data, ML emerges as a promising tool for disease diagnosis and clinical outcome forecasting. Numerous ML algorithms have been successfully implemented in clinical settings, demonstrating superior accuracy in illness prediction compared to traditional approaches [16,17].

1.1 Background and motivation

Preeclampsia is indeed a widespread and severe pregnancy complication that poses significant risks to maternal and fetal health globally. It is characterized by elevated blood pressure, fetal growth impairment, and organ damage, particularly affecting the liver and kidneys. The condition typically emerges after the 20th week of pregnancy and is a leading cause of maternal and fetal mortality worldwide.

The incidence of preeclampsia is increasing, and it affects approximately 3-5% of pregnant women. The consequences can be devastating, with an estimated 500,000 fetal and newborn deaths and around 46,000 maternal deaths attributed to preeclampsia each year. These fatalities disproportionately occur in low- and middle-income countries, highlighting the urgent need for effective detection and intervention strategies. Characterized by its insidious onset, typically after the 20th week of pregnancy, this condition unveils a troubling

constellation of symptoms, notably elevated blood pressure, impaired fetal growth leading to low birth weight, and the insidious damage to vital organs, with a particular emphasis on the liver and kidneys. [4]

According to the World Health Organization (WHO), the tragic reality persists that over 800 women lose their lives every day due to preventable causes related to pregnancy across the globe [8]. Among these causes, preeclampsia (PE) emerges as a significant contributor to maternal and fetal morbidity and mortality [9]. This complex disorder poses inherent dangers that can lead to severe medical complications for both the mother and the baby, presenting a formidable challenge to healthcare providers worldwide. The study seeks to utilize hospital electronic medical records to develop machine learning models that can anticipate the onset of preeclampsia, with the goal of enhancing early detection and intervention. Early identification of individuals at risk of preeclampsia is crucial, as it allows healthcare professionals to closely monitor affected patients and take prompt action to prevent life-threatening complications for both mothers and their unborn babies. By evaluating and comparing the efficacy of various machine learning models, the research aims to improve clinical decision-making and patient care, particularly in low- and middle-income countries where resources may be limited.

1.2 Objective

The objective of this project is to develop advanced machine learning algorithms for the early detection of preeclampsia with significantly improved accuracy compared to existing methods. The project will leverage public datasets and clinical reports from various stages of pregnancy to identify pregnant individuals at the highest risk of developing preeclampsia. By utilizing these enhanced detection models, healthcare professionals will gain actionable insights to help prevent or mitigate the severity of preeclampsia, potentially reducing adverse outcomes for both mothers and their unborn babies.

Furthermore, the project will explore the feasibility of developing machine learning-based algorithms capable of accurately predicting preeclampsia within the first trimester of pregnancy. This proactive approach will enable earlier intervention and more personalized care, ultimately aiming to improve maternal and fetal health outcomes. The project will also assess the generalizability and scalability of the developed models across different

populations and healthcare settings, particularly in low- and middle-income countries where preeclampsia poses a significant health burden.

Chapter 2: Literature Review

The purpose of this chapter is to provide an overview of the existing research and findings related to the detection of preeclampsia using machine learning algorithms. This review will serve as a foundation for understanding the current state of the field and identifying the gaps that our project aims to fill. The literature review provides a comprehensive overview of recent advancements in the early detection and prediction of preeclampsia, a critical condition during pregnancy. Researchers worldwide are leveraging machine learning techniques, advanced statistical methods, and novel biomarkers to develop predictive models that can identify individuals at risk of developing preeclampsia. These efforts aim to revolutionize prenatal care by enabling early intervention, reducing maternal and neonatal mortality and morbidity. Key themes emerge from the reviewed papers, including the use of machine learning algorithms such as Random Forest, Logistic Regression, and Support Vector Machine, along with deep learning models, to analyze various datasets encompassing clinical, genetic, and biochemical information. These models demonstrate impressive prediction accuracy, surpassing traditional methods and offering new insights into the complex interplay of risk factors associated with preeclampsia. Furthermore, the integration of novel biomarkers, genetic data, and advanced imaging techniques into predictive models holds promise for enhancing the accuracy of early detection. Studies emphasize the importance of multi-parametric algorithms that combine maternal characteristics with biophysical and biochemical markers to improve predictive power. However, the review also acknowledges challenges such as limited sample sizes, dataset heterogeneity, and the need for standardized diagnostic criteria. Despite these limitations, the findings underscore the transformative potential of machine learning and advanced analytical techniques in prenatal care, highlighting the urgent need for further research and collaboration to refine predictive models and optimize clinical outcomes for mothers and infants worldwide.

Preliminary Research

The systematic literature review titled “Advancing Preeclampsia Prediction with Machine Learning” emphasizes the potential of machine learning algorithms in the early detection of preeclampsia, a critical condition during pregnancy. The study underscores that machine learning surpasses other methods in predicting pregnancy outcomes, with a focus on features such as maternal demographics, medical history, current medical conditions, and laboratory data [11].

A variety of algorithms, including Random Forest, Logistic Regression, and Support Vector Machine, have been employed to analyze these features. Among these, machine learning techniques, particularly Stochastic Gradient Boosting, have demonstrated remarkable prediction accuracy, reaching as high as 97.3%. This highlights machine learning's dominance in the field [11].

Despite these advancements, the review calls for continued research to identify the most effective algorithm for preeclampsia diagnosis. It notes that while machine learning is currently the leading method for predicting pregnancy outcomes, the quest to find the optimal algorithm for prenatal diagnosis is ongoing. Interestingly, the review does not mention any specific limitations within the provided contexts, suggesting a positive outlook on the use of machine learning in this domain [11].

The study "Preeclampsia Predictor with Machine Learning: A Comprehensive and Bias-Free Machine Learning Pipeline" presents a novel approach to predicting preeclampsia in a nulliparous study cohort. By constructing bias-free classifiers that incorporate known risk factors for severe outcomes, the research aims to enhance the accuracy and reliability of preeclampsia prediction. The machine learning techniques employed in this study are designed to build a model that is free from biases that could skew the results [12].

Screening models play a crucial role in this research, as they are used to predict the risk of preeclampsia through various tests, including the analysis of biomarkers and ultrasound measurements. These models take into account significant features such as BMI, blood pressure, and specific analytes, which are highly indicative of the potential for developing preeclampsia [12].

One of the limitations noted in the study is the limited number of cases involving existing medical conditions, which could affect the generalizability of the findings. Despite this, the screening models show promise in predicting the risk of preeclampsia using biomarkers and ultrasound data, marking a step forward in the early detection and management of this pregnancy-related condition. The study's findings underscore the importance of further research to refine these models and improve prenatal care outcomes [12].

The research presented in "Prediction of Preeclampsia from Clinical and Genetic Risk Factors in Early and Late Pregnancy Using Machine Learning and Polygenic Risk Scores" delves into the intricate interplay between clinical and genetic factors in predicting preeclampsia. The study harnesses a range of machine learning models, including RandomOversampler, Linear

Regression, and XGBoosted ensemble models, to analyze these factors. Feature selection and data exclusion based on the stage of pregnancy are critical steps in refining the model's accuracy [13].

Key predictors identified include blood pressure, renal disease, and uric acid values, with patients exhibiting high systolic blood pressure polygenic risk scores (SBP PRS) showing correspondingly higher blood pressure levels. Sensitivity analysis within a White population cohort revealed similar patterns, reinforcing the predictors' validity [13].

However, the study acknowledges limitations, particularly the potential for overfitting, which necessitates a cautious approach to model interpretation. The use of SBP PRS derived from a White population also raises questions about the model's applicability across diverse groups. Despite these challenges, the study's findings affirm the efficacy of using combined clinical and genetic data to predict preeclampsia risk, with strong predictors like blood pressure, renal disease history, and uric acid levels leading the way in early detection efforts [13]

The research paper "Preterm Preeclampsia Risk Modeling: Examining Hemodynamic, Biochemical, and Biophysical Markers Prior to Pregnancy" presents a pioneering approach to assessing the risk of preterm preeclampsia using machine learning. The study harnesses a wealth of hemodynamic, biophysical, and biochemical data to construct models that can predict preterm preeclampsia risk before pregnancy occurs [14].

Central to the study's methodology is the use of Support Vector Machine (SVM) techniques, which have been optimized to deliver high performance. The features considered by these models include maternal cardiovascular hemodynamics, biochemical biomarkers, and biophysical measurements. Impressively, the SVM model achieved a Receiver Operating Characteristic Area Under the Curve (ROC AUC) of up to 0.88 and detection rates as high as 0.70 [14].

The models demonstrated their ability to leverage biophysical and hemodynamic biomarkers effectively, with six out of seven SVM model assessments showing true positive rates of 0.4 or higher. Additionally, the models took into account maternal BMI and various combinations of biophysical data. A feature reduction analysis was conducted, identifying 16 hemodynamic features that capture the differences associated with preterm preeclampsia [14].

Despite the impressive performance, there were instances of misclassifications in multiclass models, highlighting areas for further refinement. The study also points to the relatively

unexplored opportunity of collecting biomarker data pre-pregnancy, which could have significant clinical implications for maternal care [14].

Overall, the study underscores the potential of machine learning in identifying pre-pregnancy biomarkers for preterm preeclampsia risk. The high performance of SVM models in detecting this risk suggests that data from pre-gestational timepoints could be invaluable in improving outcomes for maternal health [14].

The paper "A Comprehensive and Bias-Free Machine Learning Approach for Risk Prediction of Preeclampsia with Severe Features in a Nulliparous Study Cohort" explores the application of machine learning models to predict the onset of preeclampsia in nulliparous pregnant women. The study stands out for its bias-corrected models, which achieved an Area Under the Receiver Operating Characteristic curve (AUROC) between 0.72 and 0.77, indicating a strong predictive capability [15].

Key to the study's approach is the use of placental analytes as top features for early preeclampsia screening. The research utilized ensemble methods, particularly Random Forest (RF) and XGBoost, which are known for their robustness and accuracy in predictive modeling. The development of the pipeline was carried out using Python 3, showcasing the practical application of programming in medical research [15].

The prediction models' performance was impressive, with RF models performing the best, achieving an AUC ranging from 0.64 to 0.83. However, the study did encounter limitations, such as a limited number of cases involving existing medical conditions in the placental analytes sub-study. Additionally, the focus of the study was narrowed to comparing patients with severe preeclampsia (sPEE) and normal pregnancy history (NPH) [15].

Another limitation was the data source; the nuMoM2b study data limits the analysis to the nulliparous US population, which may affect the generalizability of the findings. Despite these constraints, the early screening models showed promise in predicting preeclampsia risk within the nulliparous study cohort [15].

The study underscores the importance of placental analytes and multiple tests in making accurate predictions. The strong predictive power of ensemble methods like RF and XGBoost highlights the potential of machine learning in revolutionizing prenatal care by enabling early detection and intervention for conditions like preeclampsia [15].

The paper "An Appropriate Feature Selection Technique for Use on Socio-Demographic Predictor Variables to Enable Early Detection of Preeclampsia: A Review of Literature" investigates the potential of socio-demographic predictors in the early detection of preeclampsia, a significant concern for maternal health. The study conducts a thorough review of journal articles to pinpoint both independent and intervening variables that are closely linked to preeclampsia [16].

In an effort to refine the predictive process, the research utilizes advanced machine learning algorithms such as the Imperialist Competitive Algorithm and SPD algorithm for effective clustering. Additionally, the Particle Swarm Optimization algorithm is adopted to streamline the feature selection process, particularly for deep learning applications. The study also incorporates correlation-based feature selection and the C5.0 algorithm to monitor pregnancy risks more accurately [16].

A key contribution of this paper is the proposed model that amalgamates preeclampsia risk factors with socio-demographic features, aiming to boost the accuracy and performance in the early detection of this condition. This approach addresses a gap in previous studies where the influence of socio-demographics on prediction accuracy was not fully explored [16].

While the paper does not explicitly outline specific limitations, it suggests that a limited focus on socio-demographics in past research could impact the effectiveness of current predictive models. The proposed model, therefore, represents a significant step forward, enhancing the precision of early preeclampsia detection and potentially improving maternal health outcomes [16].

The review paper "Prediction of Preeclampsia Using Machine Learning and Deep Learning Models" provides a comprehensive analysis of the methodologies employed from 2018 to 2022 for predicting preeclampsia. It offers a critical examination of the data types, algorithms, and the inherent challenges and opportunities associated with preeclampsia prediction [17].

The paper discusses a variety of machine learning (ML) and deep learning (DL) methods, including Random Forest, Support Vector Machine (SVM), and Artificial Neural Networks (ANN). These methods have been applied to clinical data to predict preeclampsia with notable accuracy. The study highlights the effectiveness of combining maternal factors and antenatal laboratory data, particularly using UPCR and UA Doppler screening tests, to enhance prediction accuracy [17].

Furthermore, the review acknowledges the importance of data diversity and representation in training ML models, noting that the size of the dataset can significantly affect issues of overfitting and underfitting. It also points out the challenges posed by imbalanced datasets, which are often a result of the low prevalence of the disease [17].

Despite the high accuracy achieved by the reviewed methods, the paper identifies limitations such as the restricted data sizes and reliance on single-center data sources. The data sources reviewed were varied, encompassing clinical data, questionnaires, and laboratory test data. The studies incorporated a range of factors, including maternal characteristics, biophysical parameters, and Placental Growth Factor (PIGF), to construct their predictive models. In essence, the review underscores the potential of ML and DL in the realm of preeclampsia prediction while also highlighting the need for larger, more diverse datasets to build more robust and generalizable models [17].

The study "Preeclampsia Susceptibility Assessment Based on Deep Learning Modeling and Single Nucleotide Polymorphism Analysis" explores the frontier of preeclampsia diagnosis by integrating deep learning models with single nucleotide polymorphism (SNP) analysis. This innovative approach focuses on the early diagnosis of preeclampsia by examining the genetic pathways, particularly the interleukin-13 and interleukin-4 genes, which have been identified as differentiators between preeclampsia and normal pregnancy [18].

The research utilizes raw data from the GSE149440 microarray dataset, applying it to sophisticated models such as the multilayer perceptron and PPI graph convolutional neural networks. These models are designed to capture the complex genetic interactions that contribute to preeclampsia susceptibility. The study also includes practical genotyping tests for interleukin-13 gene polymorphisms using the PCR method, further solidifying the link between gene expression levels and the condition [18].

One of the key findings is the significant difference in genotype distribution and allelic frequencies between cases and controls, highlighting the potential of these genetic markers in preeclampsia diagnostics. However, the study is careful to note its limitations, including its focus on a single pathway related to systemic inflammation and the exclusion of samples with preterm birth conditions, which may affect the breadth of its applicability [18].

Overall, the combined use of SNP tests and deep learning models presents a promising future for preeclampsia diagnostics, offering a more nuanced understanding of the genetic factors at play and paving the way for more personalized approaches to maternal healthcare [18].

The study "Detecting Preeclampsia with a Multiple Protein Serum Test: Assay and Algorithm Development" marks a significant advancement in the early detection of preeclampsia. The researchers have developed a test that utilizes a panel of eight proteins and one clinical biomarker, each representing different physiological pathways involved in preeclampsia, to form a binary classification system [19].

In the development of the algorithm, nine biomarkers were identified and adjusted for gestational age dependency, ensuring the algorithm's consistent performance across different stages of pregnancy. The authors, who are employees or consultants of Progenity, contributed their expertise to the algorithm development, biomarker discovery, and statistical analysis [19].

The algorithm's training and optimization were carried out through multivariable logistic regression modeling. The sandwich immunoassay technique, which involves antigen-specific antibodies and detection antibodies, was employed to measure the biomarkers accurately [19].

The algorithm's performance was noteworthy, achieving 90.6% sensitivity and 78.4% specificity in detecting preeclampsia. This high level of sensitivity is particularly crucial for the ratio of free Placental Growth Factor (PlGF) to total PlGF, which is vital for the assay's accuracy. However, the complexity of diagnosing preeclampsia, especially in the presence of chronic hypertension, presents a challenge [19].

The assay developed through this study offers a sensitivity of 90.6% and a negative predictive value (NPV) of 96.9% for preeclampsia, making it a valuable tool for differentiating patients who may need more aggressive medical management. This development holds the potential to significantly aid in the early detection and treatment of preeclampsia, improving outcomes for both mothers and babies [19].

Clinicians rely on a combination of clinical, biophysical, and biochemical tests to identify early signs of this condition. Among the clinical tests, the isometric hand-grip exercise test offers a non-invasive means to predict vascular reactivity, although its practicality in routine screening is debated [20].

Biophysical tests, such as the assessment of platelet angiotensin II receptors and platelet calcium sensitivity, provide insights into the vascular changes that may herald the onset of preeclampsia. These tests, while promising, are part of an ongoing discussion about their predictive reliability [20].

Biochemical markers have also been at the forefront of research, with fibronectin levels and ED 1 fibronectin showing potential as early indicators of preeclampsia. However, the sensitivity and predictive values of these tests are subjects of contention among researchers [20].

The complexity of certain tests, like the angiotensin sensitivity test, poses a challenge due to their intricate procedures and time requirements, which may not be feasible for widespread screening. Moreover, the lack of comprehensive data on other potential markers, such as thrombin antithrombin III and protein C levels, limits their current utility in predictive models [20].

Despite these challenges, increased fibronectin levels have emerged as a robust predictor of preeclampsia, suggesting that a focus on this biomarker could improve early detection strategies. The ongoing debate over the ideal screening test reflects the need to balance various factors, including accuracy, cost, and accessibility [20].

Researchers have developed a Machine Learning (ML) model to predict pre-eclampsia and intrauterine growth restriction in pregnant women. This innovative model utilizes Doppler measures, sFlt-1, and PlGF values to make its predictions, employing a data transformation and method adaptation approach tailored for multi-label learning. The features such as Doppler measures and the sFlt-1/PlGF ratio are identified as crucial for the model's accuracy [21].

The study explored various classifiers for the ML model, including Decision Tree, Extra Tree, Random Forest, and K-Nearest Neighbors. Additionally, Gaussian Naive Bayes, Support Vector Machine, and K Neighbors Classifier were also evaluated. Among these, the Extra Tree classifiers stood out, achieving the best AUC ROC metric of 0.87, indicating a high level of predictive performance [21].

The development of this ML model underscores the importance of sharing health data and programming code to enhance research reproducibility. However, the researchers caution that the limited number of samples in the dataset might overestimate the model's performance. They highlight the challenges faced due to the lack of large publicly available datasets for researchers and emphasize the need for external validation to confirm the predictive model's performance [21].

In the realm of women's healthcare, Machine Learning (ML) is paving the way for groundbreaking advancements, particularly in predicting complications during pregnancy. A

systematic review has shed light on the efficacy of ML in foreseeing perinatal complications, with a focus on algorithms like Random Forest, Logistic Regression, Neural Networks, and Support Vector Machines, among others. These models are adept at analyzing electronic medical records to identify potential risks. However, the studies vary in their baselines, inputs, and the complications they examine, making direct comparisons challenging. Despite the promise shown by these ML applications, the field faces hurdles such as a scarcity of multicenter studies and small sample sizes, which impede their clinical application. The review calls for future prospective studies to validate these models further. It also highlights the transformative potential of AI in enhancing diagnostic accuracy and reducing healthcare costs, advocating for more research into ML applications that can lead to multicenter clinical solutions and ultimately, a reduction in pregnancy-related complications. This endeavor not only aims to advance medical knowledge but also to foster a collaborative spirit by promoting the sharing of health data and programming code for the greater good of research reproducibility and clinical application [22].

Recent advancements in machine-learning models have shown promising results in the medical field, particularly in predicting the delivery timeline for patients diagnosed with early onset preeclampsia (eoPE). A study has demonstrated that these models can accurately forecast the likelihood of delivery within 7 days following an eoPE diagnosis. The use of a Support Vector Machine (SVM) model, enhanced by an evolutionary feature selection process, has been pivotal in providing predictive information [23].

The study employed various statistical methods for univariate comparisons, including the t-test, Mann-Whitney U-test, chi-square test, and Fisher's exact test, to validate the effectiveness of the SVM model. This model, alongside Models D2 and HA2, has been instrumental in predicting potential complications and aiding healthcare professionals in making optimized clinical decisions. The high negative predictive value (NPV) of these models is particularly beneficial in identifying candidates suitable for expectant management, thus avoiding unnecessary interventions [23].

However, the study also acknowledges certain limitations, such as the small sample size and the absence of external validation for the models. It highlights the non-interchangeability of values of soluble fms-like tyrosine kinase-1 (sFlt-1) and placental growth factor (PlGF) between different laboratories, which could affect the generalizability of the results. Additionally, the lack of sonographers trained in feto-maternal Doppler studies within the emergency department poses a challenge in the widespread application of these findings [23].

Despite these constraints, the SVM model's ability to predict the need for delivery within 7 days post-eoPE diagnosis remains a significant step forward. Models D2 and HA2 not only forecast maternal and perinatal outcomes but also support the selection process for expectant management, potentially improving patient care and resource allocation. Nevertheless, the study cautions that the results may not be universally applicable across different patient subsets, indicating the need for further research and validation in diverse clinical settings [23].

Researchers have developed a model for predicting the risk of preeclampsia (PE) during early pregnancy using a combination of biomarkers. The model incorporates several factors, including PTX3, sFlt-1, ADAM12, parity, and gestational age. Notably, the genetic risk factor FLT1 rs4769613 TC significantly enhances the accuracy of the prediction [24].

The study employed a 6PLEX assay, which combines serum biomarker data with genetic information, resulting in improved preeclampsia risk prediction. Machine learning strategies were utilized to create the PE prediction models. Specifically, genotyping of placental FLT1 rs4769613 T/C played a crucial role in refining the model [24].

The developed PE prediction model demonstrated high sensitivity, accurately identifying cases that later developed preeclampsia. The model achieved an overall accuracy of 88% with a specificity of 80%. Incorporating placental genotype data further increased the prediction accuracy [24].

However, the study acknowledges limitations. Mid-pregnancy serum biomarker data alone may not be sufficient for reliable PE prediction models. The sample size was moderate, and real-life pregnancy variability was lacking. Additionally, the study population had a narrow demographic origin [24].

Despite these challenges, the developed model enables first-trimester PE screening with both high sensitivity and specificity. Further research is needed to identify specific maternal serum biomarkers during the second trimester [24].

Researchers have developed a model for predicting the risk of preeclampsia (PE) during early pregnancy using a combination of biomarkers. The model incorporates several factors, including PTX3, sFlt-1, ADAM12, parity, and gestational age. Notably, the genetic risk factor FLT1 rs4769613 TC significantly enhances the accuracy of the prediction [25].

The study employed a 6PLEX assay, which combines serum biomarker data with genetic information, resulting in improved preeclampsia risk prediction. Machine learning strategies were utilized to create the PE prediction models. Specifically, genotyping of placental FLT1 rs4769613 T/C played a crucial role in refining the model [25].

The developed PE prediction model demonstrated high sensitivity, accurately identifying cases that later developed preeclampsia. The model achieved an overall accuracy of 88% with a specificity of 80%. Incorporating placental genotype data further increased the prediction accuracy [25].

However, the study acknowledges limitations. Mid-pregnancy serum biomarker data alone may not be sufficient for reliable PE prediction models. The sample size was moderate, and real-life pregnancy variability was lacking. Additionally, the study population had a narrow demographic origin [25].

Despite these challenges, the developed model enables first-trimester PE screening with both high sensitivity and specificity. Further research is needed to identify specific maternal serum biomarkers during the second trimester [25].

Preeclampsia research is increasingly focusing on identifying early predictors and developing effective treatment strategies. Among the various methods being explored, uterine artery Doppler and serum biomarkers such as vitamin D, the sFLT1 PlGF ratio, and soluble endoglin (sEng), along with T-lymphocytes, are showing promising results. Researchers are using multiparametric algorithms that combine maternal characteristics with biophysical and biochemical markers to improve the accuracy of preeclampsia prediction [26].

The use of uterine artery Doppler, in conjunction with biomarkers like vitamin D, has been particularly encouraging. These methods have been effective in identifying women at risk for preeclampsia, which is crucial for early intervention. Known risk factors for the condition include a history of preeclampsia, chronic hypertension, and multifetal pregnancy. Interestingly, low levels of placental growth factor (PlGF) have been found to correlate with preterm delivery, regardless of whether the mother has been diagnosed with preeclampsia [26].

Despite these advances, there are no clear recommendations for the clinical use of these early predictors. The superiority of individual markers or combinations of markers remains uncertain. Consequently, efforts are being directed towards a deeper understanding of the

pathology of preeclampsia, with the aim of preventing and treating the condition more effectively [26].

Screening methods that include uterine artery Doppler and serum biomarkers are showing potential for early detection. However, large-scale trials are necessary to evaluate the effectiveness of these screening strategies and interventions. Such trials will help determine the best approaches for predicting and managing preeclampsia, ultimately improving outcomes for both mothers and their babies [26].

The systematic review on autonomic dysfunction in preeclampsia has shed light on its potential prognostic value. The review summarized various studies that examined autonomic nervous system (ANS) dysfunction in women with preeclampsia (PE). These studies involved data extraction that included assessments of autonomic function, catecholamine plasma concentration, among other factors. Non-invasive methods were employed to assess autonomic cardiovascular control, which is crucial for understanding the impact of PE on the body's regulatory systems [27].

However, the review also highlighted several challenges in this area of research. There were contradicting results from studies on autonomic nervous activity in preeclampsia, which points to the complexity of the condition and its effects on the body. A notable issue was the inconsistency in the definition of preeclampsia among different studies, which can lead to difficulties in comparing and synthesizing results. Moreover, some studies lacked clear diagnostic criteria, and there were variations in the exclusion criteria for preeclampsia patients, such as the presence of various medical conditions, chronic diseases, and antihypertensive treatment [27].

Despite these challenges, ANS testing has emerged as a promising tool that may predict the risk of PE early in pregnancy. This could be pivotal in the early identification and management of preeclampsia, potentially improving outcomes for both mothers and babies. The review underscores the need for standardized definitions and criteria in preeclampsia research to advance the understanding and treatment of this complex condition [27].

In a retrospective cohort study conducted in China, researchers have made significant strides in the development of a prediction model for preeclampsia using machine learning methods. The study utilized prenatal screening data to train various machine learning algorithms, including Deep Neural Networks (DNN), Logistic Regression (LR), Support Vector

Machines (SVM), Decision Trees (DT), and Random Forests (RF), with the aim of creating effective predictive tools [28].

The Random Forest model emerged as the standout performer, demonstrating the highest accuracy rate in predicting the risk of preeclampsia. This model was particularly adept at identifying important predictive features for preeclampsia risk, which is a crucial step in the early detection and management of the condition [28].

However, the study did face certain limitations. The data imbalance and the fact that it was a single-center study may affect the generalizability of the findings. Additionally, while the Random Forest model showed promising results, the models have not yet been validated using external datasets, which is an important step in confirming their reliability and applicability in different populations [28].

Despite these challenges, the study represents a significant advancement in the use of machine learning for medical predictions. The integration of maternal history and laboratory data into the models has the potential to enhance clinical decision-making and improve outcomes for pregnant women at risk of preeclampsia. As the research progresses, further validation and refinement of these models will be essential to ensure their effectiveness in diverse clinical settings [28].

The scoping review on potential biomarkers for late-onset and term preeclampsia has delved into the complexities of predicting this condition. The exploration of biomarkers is crucial, as it highlights the challenges in developing reliable predictive models. Researchers are leveraging machine learning algorithms and artificial intelligence to enhance these models, utilizing various methods such as protein measurement with ELISA, RT-PCR for mRNA and microRNA studies, and NMR spectroscopy and LC-MS for metabolomics studies [29].

The review emphasizes that while multivariable models that combine maternal factors with biochemical markers show promise, no single molecular biomarker currently offers sufficient clinical sensitivity or specificity for screening. This underscores the need for more effective biomarkers and further research into novel candidates that could improve prediction rates [29].

Identifying critical factors for candidate markers, such as defining subtypes and optimal testing conditions, is essential. However, the review also points out the poor reproducibility of metabolic markers due to external factors and the lack of heterogeneity and quality assessment in clinical performance evaluation. Additionally, the exclusion of studies with

biomarkers tested after disease diagnosis presents a limitation in the current understanding [29].

The review concludes that collaboration with clinicians and consistent subtype definitions are recommended to advance the prediction of late-onset preeclampsia. As the field moves forward, it is clear that a concerted effort is required to identify and validate new biomarkers that can be used in conjunction with existing risk factors to enhance the detection and management of preeclampsia [29].

The systematic review and meta-analysis on prediction models for Intrauterine Growth Restriction (IUGR) using Artificial Intelligence (AI) and Machine Learning (ML) have provided valuable insights into the capabilities of these technologies. AI and ML models, particularly those utilizing fetal heart rate variability as an input variable, have shown high accuracy in predicting IUGR. The study utilized a variety of ML models, with the Random Forest-Support Vector Machine (RF-SVM) model standing out for its high accuracy rate [30].

The review followed the PRISMA checklist to ensure quality assessment and employed a thorough search strategy across medical databases to gather relevant data. The findings suggest that AI/ML models have high diagnostic performance in predicting IUGR, which is a significant step forward in prenatal care [30].

However, the study also points out the need for uniform diagnostic criteria before these models can be clinically implemented. Algorithmic improvement and refinement are necessary to ensure the models are ready for clinical use. While AI and ML show great potential in predicting IUGR, further validation is required to confirm their effectiveness [30].

The integration of AI with noninvasive methods is highlighted as a way to enhance diagnostic tools, and future research is encouraged to focus on the transparency and efficacy of AI/ML models. This will be crucial in moving towards more reliable and accessible prenatal diagnostic methods for IUGR, ultimately improving outcomes for at-risk pregnancies [30].

In the quest to predict early- and late-onset preeclampsia in its preclinical stage, researchers have turned their focus to placenta-specific extracellular miRNA profiling. This innovative approach has led to the identification of miRNAs that differentiate between early-onset preeclampsia (ePE) and late-onset preeclampsia (lPE) with remarkable specificity. By analyzing placenta-specific miRNAs, scientists aim to detect preeclampsia early, improving outcomes for mothers and babies [31].

The research methodology includes small RNA deep sequencing and real-time quantitative PCR analysis, which are employed to profile miRNAs in placental tissue and blood plasma. Notably, the Let-7 family of miRNAs has shown differential expression in ePE and lPE, suggesting distinct pathophysiological processes in these conditions. Other crucial miRNAs, such as miR-20a-5p, miR-451a, miR-22-5p, and miR-30e-5p, have also been identified as key regulators of trophoblast functions and signaling pathways [31].

The discovery of these miRNAs has led to the development of a universal model that predicts the onset of preeclampsia with high accuracy. While the research did not mention specific limitations, it is understood that further studies are needed to validate these findings and integrate them into clinical practice. The differences in miRNA expression not only reflect the variations in the pathogenesis of early and late PE but also offer a new avenue for developing targeted interventions [31].

This groundbreaking work opens up new possibilities for prenatal care, offering hope for a future where preeclampsia can be predicted and managed more effectively, ensuring the health and safety of both mothers and their unborn children [31].

The accuracy of first-trimester screening for preeclampsia has been a subject of extensive research, with a focus on integrating maternal factors and biomarkers. This multiparametric screening approach has shown to significantly improve detection rates for early-onset preeclampsia. However, implementing such comprehensive screening in low to middle-income settings presents challenges due to the need for substantial initial and ongoing investment in infrastructure, training, and care pathways [32].

The Fetal Medicine Foundation (FMF) algorithm has been adapted for preeclampsia screening in various populations, including Asian demographics, and has been found to enhance the effectiveness of these screenings. Despite the promise shown by multiparametric screening tests, which combine patient characteristics with serum biomarkers and ultrasound, their implementation as a standard first-trimester screening method has been limited [32].

While these screening methods are more effective in detecting high-risk cases, the cost-effectiveness of multiparametric screening for preeclampsia is not yet fully established. The FMF algorithm, in particular, has shown improvements in screening across different populations. As research continues, future studies should address the transparency and efficacy of these models to ensure they can be reliably integrated into clinical practice [32].

In a prospective study examining the predictive performance of machine learning-based methods for preeclampsia, a range of models were evaluated for their accuracy in predicting different subtypes of the condition. The study utilized decision tree, Naïve Bayes, Support Vector Machine, and Random Forest models, incorporating maternal characteristics, mean arterial pressure, serum biomarkers, and the uterine artery pulsatility index as variables [33].

The findings revealed that early-onset preeclampsia is best predicted by decision tree and Support Vector Machine models. Meanwhile, Naïve Bayes and Random Forest models showed the highest performance across all types of preeclampsia. However, the predictive performance was modest for moderate and severe types of the condition, indicating that these models have limitations when it comes to predicting more complex cases of preeclampsia [33].

Additionally, the study observed a high cesarean delivery rate and lower birth weight in the preeclampsia group, suggesting that these machine learning models could be particularly useful for the early prediction of preeclampsia, potentially allowing for better management of the condition and improved outcomes for both mothers and infants. Despite the challenges, the use of machine learning in this context shows promise and could lead to more personalized and effective prenatal care [33].

In the realm of prenatal care, a significant advancement has been made with the development of a predictive model for preeclampsia that synergizes four biomarkers: sFlt-1, PlGF, NT-proBNP, and uric acid. This model stands out for its substantial improvement in the accuracy of predicting preterm preeclampsia [34].

To construct this model, serum samples were meticulously collected to measure the levels of PlGF, sFlt-1, NT-proBNP, and uric acid. A machine-learning model, equipped with six predictors, was then employed to forecast the likelihood of preeclampsia. The robustness of the model was assessed using statistical tools such as SPSS, ANOVA, and Chi-squared tests [34].

The findings were telling: elevated levels of sFlt-1, the sFlt-1/PlGF ratio, NT-proBNP, and uric acid were consistently observed in cases of preeclampsia. The predictive model, which combines the sFlt-1/PlGF ratio with NT-proBNP and uric acid, showed improved outcomes, indicating its potential as a reliable tool for early detection. Notably, increased NT-proBNP levels were also found in women with hypertensive disorders of pregnancy, further validating the model's relevance [34].

However, the study acknowledges the need for more data from randomized trials to validate these findings for clinical practice. It's also important to note that the cut-off values for the sFlt-1/PlGF ratio are specific to Elecsys immunoassays, which may limit their generalizability [34].

In summary, the machine-learning model that combines the sFlt-1/PlGF ratio with NT-proBNP and uric acid offers a more nuanced prediction of preeclampsia, particularly for preterm cases, compared to using the ratio alone. This model represents a promising step forward in the early identification and management of preeclampsia, potentially leading to better maternal and fetal health outcomes [34].

In a groundbreaking study, researchers have developed an imbalance-aware deep neural network, known as a Cost-Sensitive Deep Neural Network (CSDNN), designed to enhance the early prediction of preeclampsia. This model is particularly attuned to the nuances of diverse minority populations, employing focal loss and weighted cross-entropy loss functions to refine its predictive capabilities [35].

The CSDNN model has achieved superior Area Under the Curve (AUC) results, outperforming state-of-the-art techniques in predicting preeclampsia. By leveraging clinical databases, the model has demonstrated a robust predictive power for preeclampsia among minority groups, a critical step forward in addressing healthcare disparities [35].

The study delved into various network architectures and hyperparameter optimization algorithms, including Bayesian optimization, Hyperband, and random search, to optimize the model's performance. Despite facing challenges such as sparse data, racial disparities, and imbalanced datasets, the CSDNN model with the focal loss function emerged as a superior tool for preeclampsia prediction [35].

However, the researchers acknowledge that false-negative predictions of preeclampsia can lead to high maternal morbidity and mortality, underscoring the need for reliable outpatient management tools for preeclamptic patients. The study also highlights the importance of addressing racial disparities in health outcomes and calls for future research to extend these models to both early and late-onset preeclampsia, aiming to improve the model's applicability and accuracy further [35].

The systematic review protocol for the use of artificial intelligence (AI) and machine learning (ML) methods in early pregnancy preeclampsia (PE) screening represents a structured approach to evaluating the effectiveness of these technologies. The review focuses on AI and

ML methods specifically tailored for PE risk prediction during the early stages of pregnancy. It involves the extraction of data from various studies that detail PE risk factors and the AI/ML methods used to predict them [36].

The protocol includes the inclusion of original research studies that apply AI and ML methods to PE risk prediction. An Excel spreadsheet is utilized to systematically extract data from the included studies, ensuring a thorough and organized analysis. A meta-analysis is then conducted, contingent upon the quality and availability of the data, to synthesize the findings and draw conclusions about the efficacy of AI/ML methods in this context [36].

The review aims to identify the AI/ML methods that are most effective for PE risk prediction during early pregnancy by examining study details, risk factors, and outcomes. While the protocol does not specify any limitations or provide conclusions in the available contexts, it sets the groundwork for a comprehensive assessment of AI/ML applications in prenatal care, with the potential to significantly impact the screening and management of preeclampsia [36].

The pursuit of early detection and management of preeclampsia (PE) has led to the development of various screening methods and biomarkers. These tools are designed to identify the condition in its nascent stages, with a focus on maternal risk factors, uterine artery Doppler, and mean arterial pressure as key indicators. A logistic regression model that includes glycerol, 3-hydroxyisovalerate, arginine, and uterine artery pulsatility index has been utilized to enhance the accuracy of these screenings [37].

The combination of these methods and biomarkers has been shown to detect preterm PE with approximately 75% accuracy while maintaining a 10% false-positive rate. This level of precision is significant, as it represents a substantial improvement over previous detection rates. The use of uterine artery Doppler and mean arterial pressure, in particular, has proven to be effective screening tools [37].

However, the performance of these screenings can vary due to methodological differences, and criteria based solely on maternal characteristics may not capture all cases. Despite these challenges, an effective screening for preterm PE can be achieved, with a combination of maternal risk factors, uterine artery Doppler, mean arterial pressure, and maternal serum pregnancy-associated plasma protein-A (PAPP-A) and placental growth factor (PlGF) identifying about three-quarters of preterm PE cases for a false-positive rate of 10% [37].

This integrated approach to screening signifies a leap forward in the proactive management of preeclampsia, offering a beacon of hope for expectant mothers and the healthcare professionals dedicated to their care [37].

In a prospective study, the predictive performance of various machine learning models for preeclampsia was assessed, revealing that models such as Random Forest (RF), Decision Tree (DT), Naïve Bayes (NB), and Support Vector Machine (SVM) exhibit high predictive performance. These models were applied to first-trimester data, including clinical and paraclinical parameters, to predict preeclampsia subtypes with accuracies ranging notably from moderate to high [38].

The study incorporated serum levels of PP-13 in the analysis, which, along with other factors, influenced the predictive performance. While the models showed promising results, their performance was limited when it came to moderate and severe types of preeclampsia. Additionally, certain patient conditions and incomplete records were excluded from the study, which could impact the generalizability of the findings [38].

Despite these limitations, the study highlights the potential of machine learning models in the early prediction of preeclampsia, with DT and NB models showing high accuracy for predicting preeclampsia subtypes, and SVM and RF models excelling in predicting early-onset preeclampsia. This research underscores the importance of machine learning in enhancing prenatal care and the management of preeclampsia [38].

The study on improving preeclampsia risk prediction has made significant strides by utilizing electronic medical record (EMR) data combined with machine learning techniques. A variety of feature selection methods, including LASSO, univariate analysis, XGBoost, and random forest, were employed to construct predictive models that assess the risk of preeclampsia at different stages of pregnancy [39].

These models underwent rigorous validation on independent datasets, demonstrating high accuracy in predicting preeclampsia. They have also been instrumental in identifying both known and potential risk factors associated with the condition. The success of these models has led to the development of personalized precision therapeutic strategies for patients deemed at risk, marking a notable advancement in prenatal care [39].

However, the study faced challenges such as missing data from patients who received care at other facilities and the exclusion of detailed physician notes and over-the-counter medications

from the analysis. Additionally, the reliance on ICD codes alone may not provide a comprehensive identification of preeclampsia patients [39].

Despite these limitations, the use of machine learning has enhanced the accuracy of preeclampsia prediction beyond what standard clinical practice can achieve. The study has even identified new risk factors, such as complete blood count, for preeclampsia. The predictive models have proven effective across different pregnancy periods, establishing a new benchmark for early detection and management of preeclampsia, ultimately aiming to improve maternal and fetal health outcomes [39].

The study on machine-learning-based prediction of pre-eclampsia has brought to light the significant role of first-trimester maternal characteristics and biomarkers in forecasting the condition. Machine learning techniques, particularly non-linear classifiers and neural networks, have been employed to analyze these factors, demonstrating that race information is a critical component for accurate prediction, especially for non-white populations [40].

The research indicates that raw biomarker data can effectively predict preterm preeclampsia without the need for normalization, simplifying the screening process. The inclusion of race in the prediction algorithms has been emphasized as essential for maintaining accuracy. These machine-learning methods have been shown to enhance the prediction of pre-eclampsia, offering an advantage over traditional logistic regression by not requiring population-based normalization [40].

However, the study acknowledges certain limitations, such as the lack of testing of the prediction algorithms in diverse populations, which could introduce bias. Despite these challenges, the study's strengths, including large populations, standardized protocols, and the measurement of PlGF (placental growth factor), provide a solid foundation for the findings. The study underscores the potential of machine learning in improving pre-eclampsia screening and highlights the need for further research to validate these algorithms across various populations [40].

Chapter 3: Methodology

Preeclampsia, a hypertensive disorder that occurs during pregnancy, poses significant risks to both maternal and fetal health. Timely detection and intervention are crucial to mitigate adverse outcomes. In recent years, machine learning (ML) algorithms have emerged as powerful tools for medical diagnosis and risk prediction. This study focuses on leveraging ML techniques to enhance early detection of preeclampsia. The significance of early detection lies in several key aspects:

- **Maternal Health:** Preeclampsia can lead to severe complications such as eclampsia, organ failure, and maternal mortality. Detecting warning signs early allows for timely medical intervention, reducing the impact on maternal health.
- **Fetal Well-Being:** Preeclampsia affects fetal growth, leading to low birth weight, preterm birth, and developmental issues. Identifying high-risk pregnancies early enables targeted monitoring and management.
- **Challenges in Clinical Diagnosis:** Clinical diagnosis relies on traditional risk factors (e.g., hypertension, proteinuria), which may not manifest until later stages. ML algorithms can integrate diverse features and patterns to provide earlier predictions.

Our research aims to achieve the following objectives:

1. **Feature Extraction:** Extract relevant features from clinical data, including vital signs, laboratory results, and patient history.
2. **Model Development:** Train and evaluate ML models (such as logistic regression, random forests, and neural networks) to predict preeclampsia risk.
3. **Validation and Generalization:** Validate model performance using independent datasets and explore generalizability across diverse patient populations.

By enhancing early detection, our study contributes to improved maternal and fetal outcomes. We anticipate that ML-based approaches will complement existing clinical practices, aiding healthcare providers in making informed decisions.

3.1 Data Acquisition

An extensive dataset, obtainable via the Figshare database, has been made available by Yonsei University Healthcare Centre, which encompasses Severance Hospital and Gangnam

Severance Hospital in Seoul, Korea, and provides data from 2005 to 2017 [22]. This dataset encompasses information on both preeclampsia and normal patients, totaling approximately 42,965 patient records. The systematic collection of data occurred from the commencement of the second trimester until reaching a gestational age of 34 weeks. The dataset includes demographic details such as age, blood type, blood pressure (BP), height, weight, and gestational age, along with various biochemical parameters. Features within the dataset cover blood urea nitrogen (BUN), serum creatinine, spot urine protein to creatinine ratio (UPCR), urine albumin to creatinine ratio, hemoglobin, fasting blood glucose, serum albumin, uric acid, total bilirubin, aspartate transaminase (AST), alanine transaminase (ALT), total cholesterol, triglycerides, high-density lipoprotein cholesterol, and low-density lipoprotein cholesterol. The primary research goal was to assess the likelihood of preeclampsia development, defined as the emergence of new hypertension diagnosed with at least two separate measurements, each taken at least two hours apart, with a diastolic BP of 90 mm Hg or a systolic blood pressure of 140 mm Hg. Additional diagnostic criteria included clinically significant proteinuria, indicated by a 24-hour urine protein level of ≥ 300 mg, a platelet count of less than 100,000/ μ L, a creatinine level exceeding 1.1 mg/dL, serum transaminase levels twice the normal range, or the manifestation of cerebral symptoms or pulmonary edema after 34 weeks of gestation [23].

In the process of acquiring knowledge, we often find ourselves conducting research in new areas. This was the case when we initiated a project to gather data from various hospitals, both private and government-owned. This task led us through numerous facilities, introduced us to a multitude of medical professionals, and presented us with unexpected challenges and successes. In this account, we will detail our experiences, the initial rejections faced, the lessons learned, and the alternative methods employed to gather the required data. With a clear objective and a list of establishments, we began our research in the healthcare sector, eager to uncover valuable information. Hospitals, both private and government-run, became our primary research sites as we sought to gain insights into their operations.

As we approached each hospital, we were often met with initial rejection. Convincing medical professionals to share their knowledge and data proved to be a significant challenge. However, we remained undeterred and persisted in our efforts. Each rejection only served to strengthen our resolve. Our initial visits to hospitals were not completely for naught. Despite the setbacks, we decided to visit hospitals repeatedly, aiming to gain the trust and cooperation of the medical fraternity. These visits not only allowed us to familiarize ourselves with the

healthcare system but also provided invaluable insights into the routines, procedures, and protocols followed by medical professionals.

The acquisition of data often requires the use of alternative methods. In our search for information, we discovered different ways to collect data that deviated from standard procedures. We devised strategies that allowed us to gather the necessary data without strictly adhering to established protocols. These alternative methods opened doors to previously inaccessible information and granted us a unique perspective on the world of hospitals. After numerous rejections and relentless perseverance, we finally managed to collect the desired data from a hospital. This accomplishment was not without its challenges. We spent several days immersing ourselves in the hospital's environment, observing, interacting, and documenting every detail. The experience was enlightening, offering a glimpse into the intricacies of medical operations and the dedication of healthcare professionals. We learned a lot about the practical field and gained plenty of insights, such as:

- Most private hospitals do not want to share their patients' data; they maintain confidentiality; only a select few agree to it.
- Only a few hospitals have a specified ward/department for preeclampsia and eclampsia.
- Government hospitals do not build any database or record any data, hence do not have an actual stored dataset.
- A few professors of our country in this specific field personally collect data or indulge in this type of research/work, but not all of them are cooperative.
- Many doctors of government hospitals wanted some sort of commission/ benefits for agreeing to help us collect data.
- All pregnant women are given blood tests and urine tests
- A few tests were specifically given to patients with more serious conditions, so we could not collect all the data we wanted.

The task of collecting data from multiple hospitals was indeed a significant undertaking. It tested our resilience, required us to adapt, and allowed us to discover alternative means of acquiring information. Through rejection, persistence, and alternative methods, we gained a deeper understanding of the healthcare landscape. This account serves as a testament to the lengths one can go in the pursuit of knowledge, and the rewards that await those who dare to venture beyond conventional boundaries.

3.2 Data Processing, Labeling and Selection

Data processing is a crucial step in ML as it involves cleaning and transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and contains many errors. Data processing helps to eliminate such issues and makes the data suitable for an ML model. It involves data cleaning, data integration, data transformation, and data reduction.

- **Data Cleaning:** This involves handling missing and noisy data. Missing values can be filled in with a certain value or estimated using a method like mean, median, etc. Noisy data (random error or variance in a measured variable) can be smoothed by applying techniques like binning, regression, etc.
- **Data Integration:** This is the process of combining data from different sources and providing users with a unified view of these data.
- **Data Transformation:** This involves normalizing the data within a specific range (like -1.0 to 1.0), or changing the data to improve its accuracy.
- **Data Reduction:** This involves reducing the volume of data, making it easier to store and process.

Data Labeling: In supervised learning, models learn from labeled data, i.e., data paired with its corresponding output label. Data labeling involves assigning labels to the data points. For example, in a binary classification problem, each data point is labeled as 0 (negative class) or 1 (positive class). The quality and accuracy of labels directly impact the performance of the ML model. Incorrect labels can lead to incorrect predictions by the model.

Data Selection: This involves selecting the most relevant data for your ML model from the entire dataset. Not all data is useful and relevant to the problem you are trying to solve. Irrelevant or less important data can negatively impact model performance. Techniques like correlation matrices, feature importance from tree-based models, or methods like Recursive Feature Elimination can be used to select the most relevant features.

Data processing, labeling, and selection are critical steps in the ML pipeline. They ensure that the data is clean, relevant, and in the right format, which in turn helps in building accurate and efficient ML models. Remember, "Garbage in, garbage out". If the data going into the model isn't good, the predictions coming out won't be good either. Hence, spending time on these steps is crucial for any ML project.

3.2.1 Data Labeling

The data gathered from medical records was initially non-standardized. Many characteristics linked to medical indicators are included in the dataset. The field experts and references were followed in conducting this investigation [25]. This analysis comprised 42965 pregnant women. 6% of all pregnancies, or 645 cases of PE, were discovered in this data. The following pregnancies met the inclusion criteria: (1) those with no history of chronic hypertension or hypertension before 20 weeks of gestation; (2) those with full information on all 26 characteristics; and (3) those with deliveries at the Yonsei University Healthcare Centre. The terms "Threshold_feature1" and "Threshold_feature2" refer to thresholds that are defined for SBP and DBP features, respectively. Only in the event that particular criteria are satisfied—namely, when both SBP and DBP are below their corresponding thresholds—is the 'Label' column changed to '0'. Features like Urine_AC, Urine_PC, ALT, AST, BUN, PLT, and CR_SE have additional criteria assigned to them. In case any of the given requirements are not fulfilled, the 'Label' column is changed to '0'. Examining different threshold combinations for distinct attributes is required to meet these objectives. The entire procedure includes preprocessing the data and assigning labels based on predetermined thresholds to make the categorization of preeclampsia easier. By taking into consideration various criteria that are developed from medical indicators found in the dataset, the current technique assists in labeling the data to improve the accuracy of the diagnosis process. However, the use of this technique allows for a more efficient and accurate classification of patients.

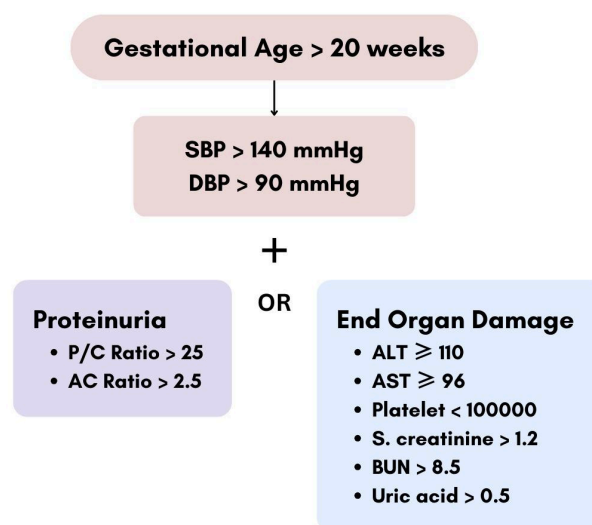


Fig. 3.1. Gestational hypertension

Gestational hypertension (PE) was defined as a gestational age greater than 20 weeks (systolic BP (SBP) at 140 mm Hg and/or diastolic BP (DBP) at 90 mm Hg, two times with at least 4 hours between them after 20 weeks of gestation in a previously normotensive woman), associated with proteinuria. Therefore, it is possible to diagnose preeclampsia. Afterwards, in order to validate the occurrence of preeclampsia, proteinuria and end organ damage features must be investigated. If the urinary albumin-to-creatinine ratio (AC ratio) is greater than 2.5 or the protein/creatinine ratio (PC ratio) is greater than 25, preeclampsia must be confirmed. The last characteristics of organ damage include uric acid >0.5 , serum creatinine >1.2 , BUN >8 , ALT >110 , AST >96 , or uric acid >1.2 , which should be obtained in addition to platelet counts $<100,000$. This investigative approach adhered to the guidance of field experts and references [25]. These standards are crucial for validating the preeclampsia diagnosis and setting it apart from other illnesses. In order to treat and care for pregnant patients with suspected preeclampsia, it is imperative to evaluate these indicators.

3.2.2 Data Processing and Selection

Data Collection

We started our project by collecting data, which is a crucial first step in any data analysis or machine learning project. We gathered data from a variety of sources, including online databases and raw data from hospitals. The type of data we collected depended on the problem we were trying to solve. The key was to identify the most relevant sources that would provide the necessary data for our specific problem.

Data Preparation

After collecting the data, we had to prepare it for analysis. This stage involved transforming the data into a format suitable for further analysis or modeling. The data we collected online was initially in Excel and txt formats. We converted these formats into CSV (Comma Separated Values), a widely used format that can be easily imported into many data analysis tools and programming languages like Python and R.

Data Cleaning

Once we had our data in a suitable format, we moved on to the data cleaning stage. Real-world data is often messy and contains errors, outliers, and missing values that could negatively impact the results of our analysis. Therefore, we cleaned the data by handling these issues appropriately.

Handling Missing Values

One common issue we encountered during data cleaning was missing values. There are several strategies to handle missing values, and the choice of strategy often depends on the nature of the data and the specific problem at hand. We considered various strategies such as:

- **Deleting Rows:** This strategy involved deleting the rows that contained missing values. We used this approach when the number of rows with missing values was relatively small and unlikely to impact the overall analysis.
- **Imputation:** This strategy involved filling in the missing values based on other data. This could be the mean, median, or mode of the column, or it could be a more complex method like using a machine learning algorithm to predict the missing values based on other data.
- **Ignoring:** In some cases, it might be appropriate to simply ignore the missing values, especially if the data analysis method we were using could handle missing values appropriately.

By following these steps, we were able to create a clean, high-quality dataset that we could use to generate reliable and valid insights from our data analysis or machine learning model. Remember, the goal of these steps is to prevent data loss and errors, which is why they are so crucial in any data analysis or machine learning project.

Data Selection

After we had cleaned and prepared our data, the next step in our process was data selection. This is a crucial stage as it involves choosing the most relevant data for our analysis from the entire dataset. Not all data is useful and relevant to the problem we were trying to solve. Irrelevant or less important data can negatively impact model performance.

In our case, we were particularly interested in blood reports and urine reports. These tests were the most common and provided a wealth of information that was relevant to our analysis. Blood tests can reveal a lot about a person's health, including information about their organs' function, the presence of any diseases, and the effectiveness of treatments they might be receiving. Similarly, urine tests can provide information about kidney function, hydration, and the presence of infection among other things.

By focusing on these reports, we were able to narrow down our data to the most relevant and useful information. This not only made our analysis more manageable but also more accurate

and effective. We were able to avoid the noise and confusion that can come from dealing with irrelevant data and could focus our efforts on the data that truly mattered for our project. This is why data selection is such a critical step in any data analysis or machine learning project. It allows us to make the most of our data and ensures that our results are based on the best possible information.

3.3 Data Validation

Data Validation is the process of checking that data meets certain requirements by comparing it to a set of rules that have already been defined. This procedure involves performing a series of checks known as check routines. Data validation ensures that the data is clean, accurate, and usable. Only validated data should be imported, saved, or used; otherwise, programs may stop working, results may be erroneous or other potentially disastrous problems may arise. Now, let's discuss the importance of data validation:

- **Accuracy:** Data validation provides accuracy, cleanness, and completeness to the dataset by eliminating data errors from any project to ensure that the data is not corrupted. Inaccurate and incomplete data may lead the end-users to lose trust in data.
- **Efficiency:** Data validation is an essential part of the ETL process (Extract, Transform, and Load), which involves moving the source database to the target data warehouse. Performing data validation is required to enhance the value of the data warehouse and the information stored there.
- **Reliability:** Data validation is crucial for maintaining the accuracy and reliability of data within a system. By applying validation rules, you can ensure that the data entered or imported meets specific quality standards, reducing inaccuracies, inconsistencies, and errors in the data.
- **Decision Making:** You risk making decisions based on inaccurate, unrepresentative data without validating data. Validating the accuracy, clarity, and specificity of data is necessary to fix any project problems.

Data validation is a critical step in data handling tasks to create data that is consistent, accurate, and complete to prevent data loss and errors. It allows users to check that the data they are dealing with is valid by means of end-to-end testing such as testing for data accuracy, data completeness, and data quality. It is considered to be the foundation for efficient data management since it facilitates analytics based on meaningful and valid datasets. Exploratory data analysis was used for this project

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a critical step in the data analysis pipeline. It involves analyzing datasets to summarize their main characteristics, often using data visualization methods[43][44]. EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them[43].

Here are the key steps involved in EDA[34]:

1. **Dataset Overview and Descriptive Statistics:** This involves understanding the number of observations, number and types of features, overall missing rate, and percentage of duplicate observations[44].
2. **Feature Assessment and Visualization:** This step involves visualizing each field in the raw dataset with summary statistics. It allows you to assess the relationship between each variable in the dataset and the target variable[43].
3. **Data Quality Evaluation:** This step involves checking for data quality issues such as missing values, outliers, and inconsistent data entries[44].

EDA involves specific statistical functions and techniques, including[43]:

1. **Univariate visualization:** This involves visualizing each field in the raw dataset, along with summary statistics[43].
2. **Bivariate visualizations:** These allow you to assess the relationship between each variable in the dataset and the target variable[43].

EDA is important because it helps data scientists ensure the results they produce are valid and applicable to any desired business outcomes and goals[43]. It also helps stakeholders confirm they are asking the right questions[43]. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling[43].

3.3.1 Univariate Analysis

Univariate analysis is a type of statistical analysis that is applied to data that consists of observations on only a single characteristic or attribute. The term "univariate" comes from "uni" meaning "one", indicating that the analysis is performed on one variable[41]. The primary purpose of univariate analysis is to describe the data and find patterns within it[41]. It doesn't deal with causes or relationships, unlike regression[41]. One way to perform

univariate analysis is Charts. Charts can be used to visualize the distribution of values for a certain variable. We used boxplot, and density curve.

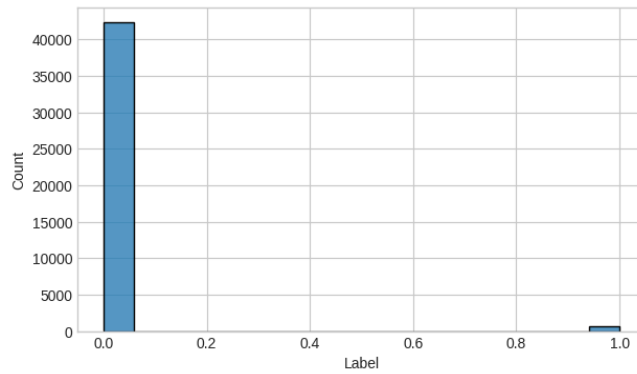


Fig. 3.2. Target distribution

The dataset we have consists of information from both preeclampsia patients and normal pregnant women. Upon reviewing the data, we observe a significant imbalance, with data from 42,320 normal pregnant women and only 645 preeclampsia patients. This imbalance presents challenges for model development, as the vast majority of the data represents the normal population, while the minority class—preeclampsia patients—is underrepresented. When working with imbalanced datasets, it is crucial to address the disproportionate representation to avoid biases in model training.

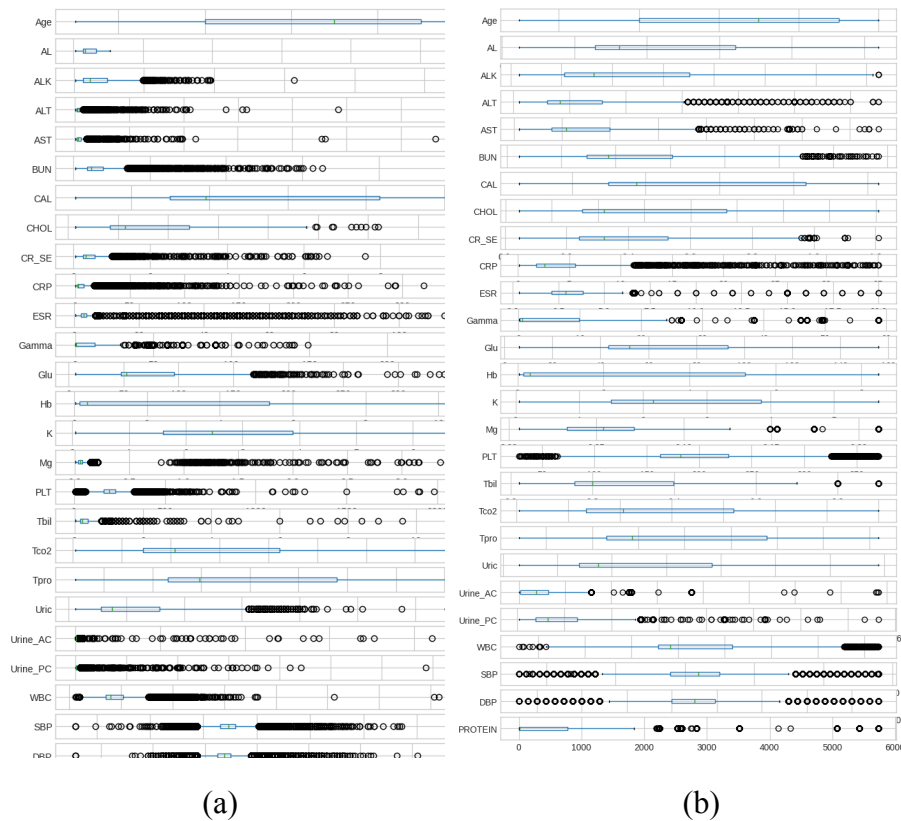


Fig. 3.3. Box plot (a) with outliers and (b) without outliers.

The figure above is a box plot, which is a graphical representation of data that provides a summary of the distribution's central tendency, spread, and variability. It displays the median, quartiles, and potential outliers within the data. The box represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3). The line inside the box shows the median, while the whiskers extend from the box to the smallest and largest values within 1.5 times the IQR from Q1 and Q3, respectively. Points beyond these whiskers are considered potential outliers. Outliers are data points that deviate significantly from the majority of observations within a dataset. They are identified in a box plot as points that lie outside the whiskers. Outliers can arise from measurement errors, data entry errors, or natural variations within the data.

It is important to exclude outliers in the machine learning process because they can distort the model's training data and lead to biased or unreliable predictions. Outliers can disproportionately influence statistical measures and violate assumptions about the data's distribution, which can impact the model's accuracy and generalizability. By removing outliers, we can improve the quality of the data and help ensure that the model learns patterns that are representative of the majority of observations.

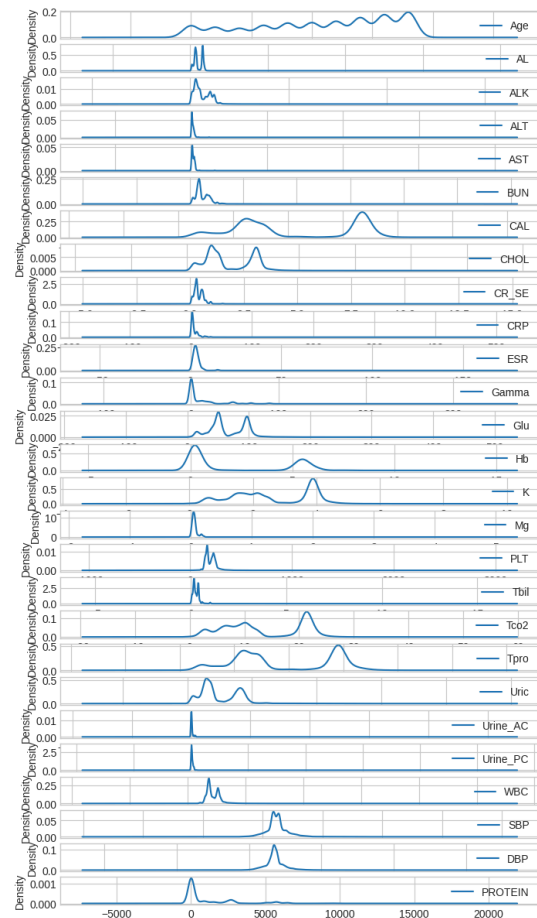


Fig. 3.4. Density Curve

The figure displayed above depicts a density curve, which serves as a graphical representation helping us comprehend the distribution of a univariate dataset. This curve is smooth and provides an estimate of the probability density associated with the data. On observing the vertical axis, we can discern that it represents the probability density, signifying the probability of encountering a specific value. Conversely, the horizontal axis corresponds to the values of the variable under investigation. It is constructed in a manner that ensures the total area beneath it equals 1, thereby encapsulating the overall probability distribution. Consequently, the height of the curve at any given point indicates the relative likelihood of encountering a value around that particular point.

By analyzing the shape of the density curve, we can glean insights into the data distribution. For instance, a symmetric curve suggests a normal or bell-shaped distribution, while a skewed curve suggests a deviation from symmetry. The curve's peak denotes the mode, representing the value with the highest density. Density curves are invaluable for comprehending the overall data pattern, offering insights into central tendency, variability, and skewness. Moreover, they facilitate easy comparisons between different datasets or groups by overlaying multiple density curves on the same plot.

3.3.2 Bivariate Analysis

Bivariate Analysis is a statistical method that helps you investigate the relationship between two variables. It's one of the simplest forms of quantitative (statistical) analysis[42][45]. The purpose of bivariate analysis is to determine the empirical relationship between two variables, often denoted as X and Y[42][45]. Here's how it works:

- **Bivariate Regression:** This is a statistical technique used to investigate how variation in one or more variables predicts or explains variation in another variable. Bivariate regression aims to identify the equation representing the optimal line that defines the relationship between two variables based on a particular data set.
- **Simple Linear Regression:** This is a statistical method used to model the linear relationship between an independent variable and a dependent variable. It assumes a linear relationship between the variables and is sensitive to outliers.
- **Least Squares Regression Line (LSRL):** The LSRL is a method in simple linear regression for modeling the linear relationship between two variables, and it serves as a tool for making predictions based on new values of the independent variable.

Bivariate analysis can be helpful in testing simple hypotheses of association. It can help determine to what extent it becomes easier to know and predict a value for one variable (possibly a dependent variable) if we know the value of the other variable (possibly the independent variable). Bivariate analysis can be contrasted with univariate analysis in which only one variable is analyzed.

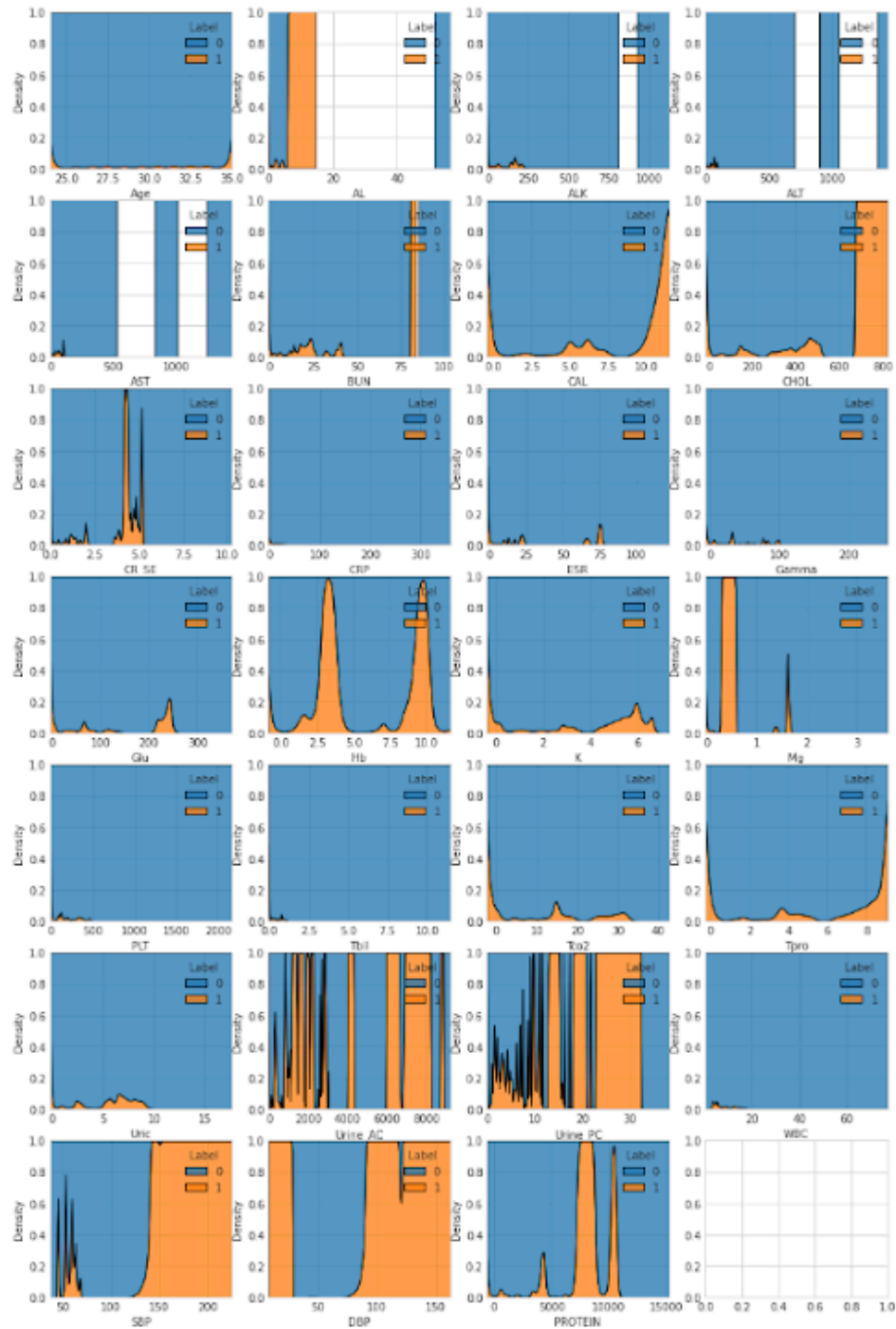


Fig. 3.5. KDE plot

The above image represents a kernel density estimate (KDE) plot, a visualization technique used in bivariate analysis. The plot displays the estimated probability density of the data

points in a two-dimensional space. Contours or a color gradient indicate the density of data points, with darker or thicker lines representing higher density areas. By examining the KDE plot, we can identify patterns, trends, and clusters in the relationship between the two variables being analyzed. KDE plots are especially useful for visualizing the joint distribution of continuous variables, providing a comprehensive understanding of their association. In the context of the above image, the KDE plot helps us gain insights into the relationship between the two variables, showcasing density patterns and highlighting areas of interest within the data.

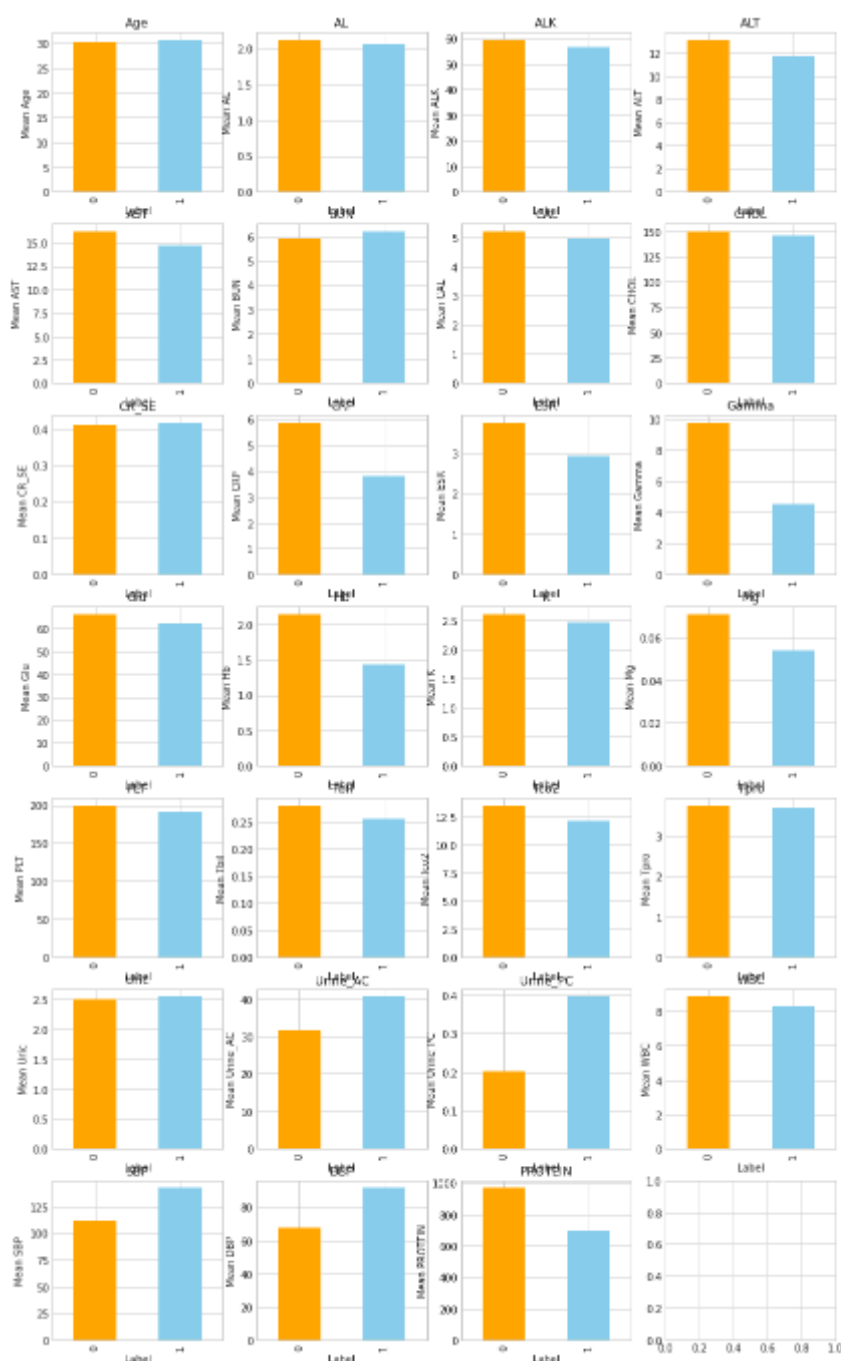


Fig. 3.6. Bar plot

The above image represents a bar plot used in bivariate analysis. The plot showcases the distribution and frequency of categories for two categorical variables. The horizontal axis represents one variable, while the height of the bars represents the frequency or proportion of the corresponding categories. In the context of the above image, the bar plot allows for a visual comparison of category distributions between the two variables being analyzed. It helps identify which categories are more or less prevalent in each variable and provides insights into their relationship. By examining the bar plot, we can easily interpret the patterns and trends in the bivariate relationship. It offers a clear and intuitive visualization of the categorical data, facilitating a better understanding of the variables' association.

In summary, the bar plot in the above image serves as a visual tool for bivariate analysis. It enables us to compare and analyze the distribution of categories between two categorical variables, providing valuable insights into their relationship.

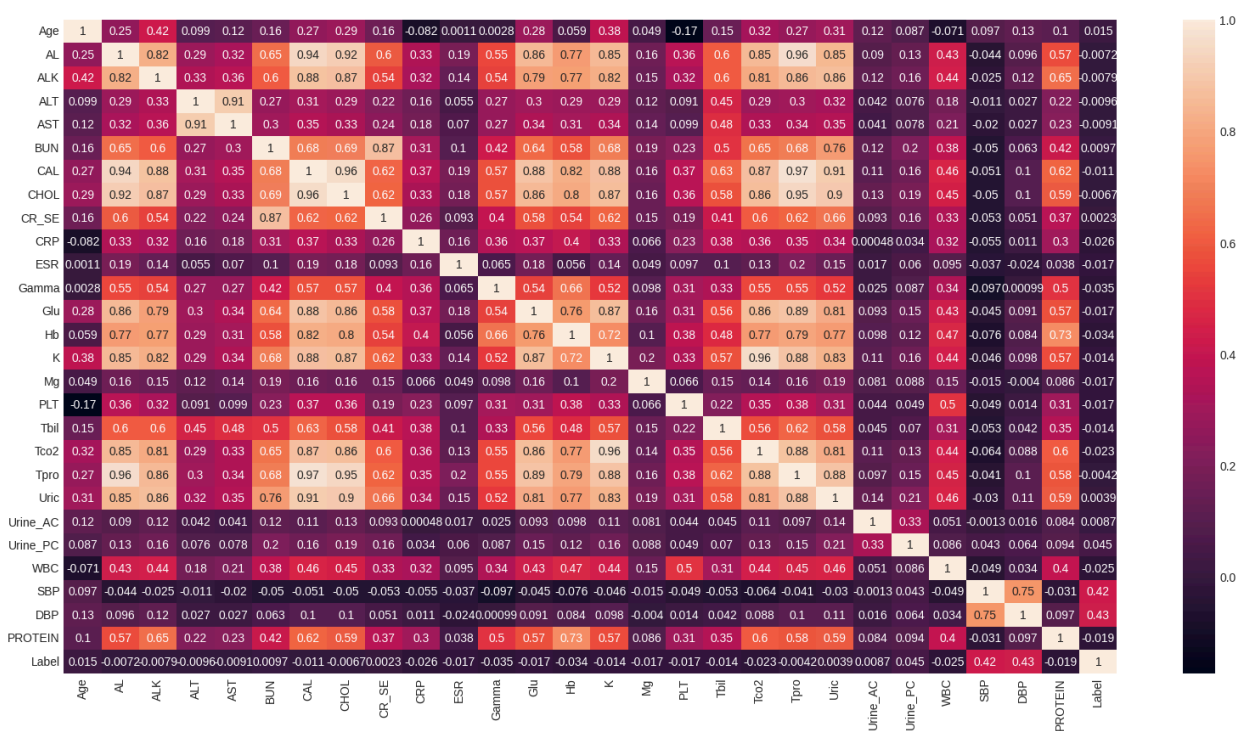


Fig. 3.7. Correlation matrix

The above image represents a correlation matrix, which displays the correlation coefficients between multiple variables in a dataset. The matrix allows for a visual representation of the strength and direction of the linear relationships between pairs of variables. In the context of the above image, the correlation matrix provides insights into the interdependencies among the variables being analyzed. By examining the cells in the matrix, we can identify the degree of correlation between each pair of variables. Positive correlations indicate that the variables tend to increase or decrease together, while negative correlations suggest an inverse

relationship. The correlation matrix is a valuable tool for understanding the connections between variables and identifying potential patterns or associations. It aids in feature selection, identifying redundant variables, and gaining insights into the underlying structure of the data. It is important to note that correlation does not imply causation. While a high correlation coefficient suggests a strong relationship between variables, it does not establish a causal link. Further analysis and domain knowledge are necessary to determine causal relationships.

To summarize, the correlation matrix presented in the above image provides a visual representation of the relationships between variables. It allows for the identification of correlations, aiding in the exploration and understanding of the interdependencies within the dataset.

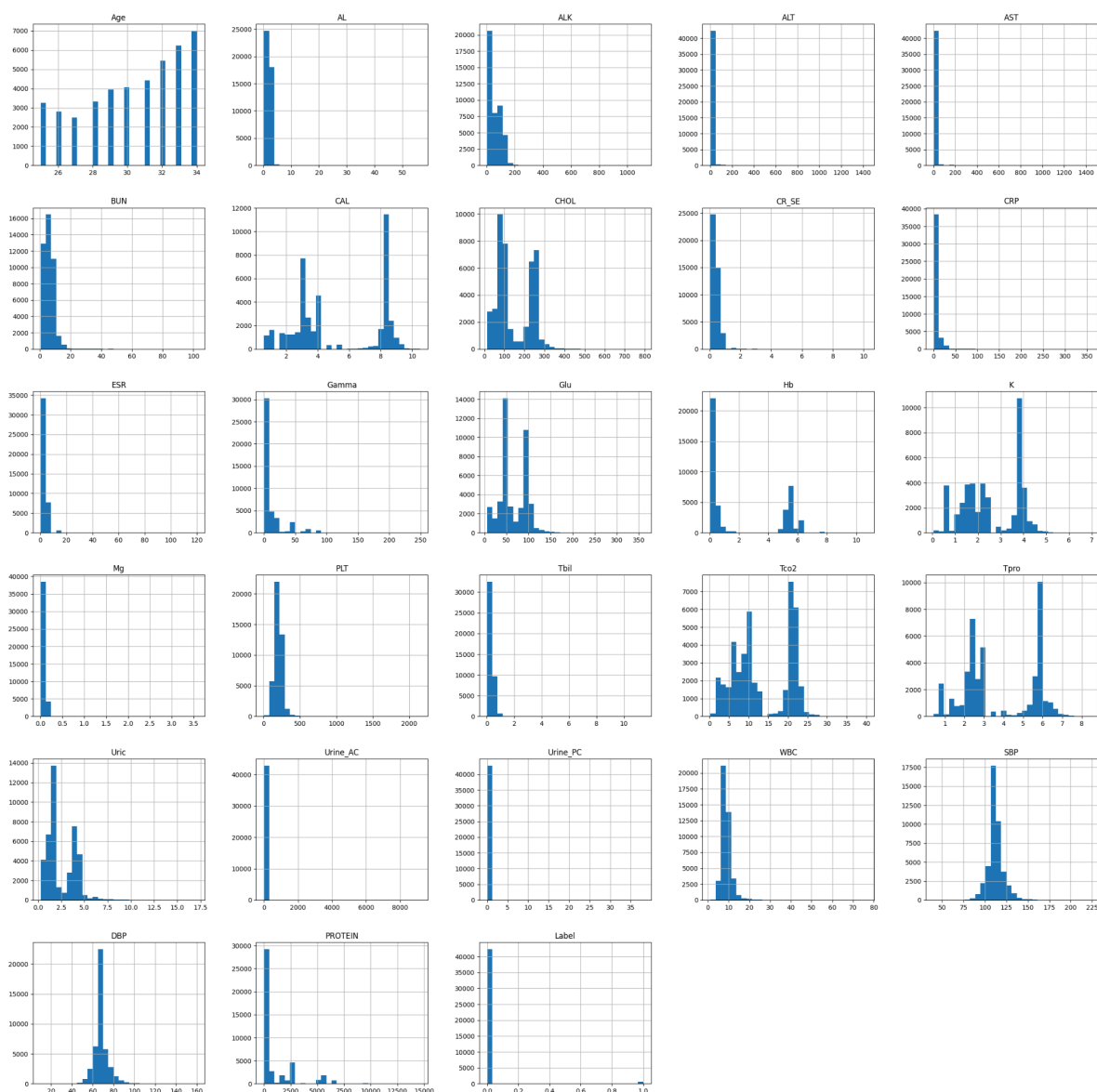


Fig 3.8. Online data histogram

The image you've uploaded appears to be a collection of 20 different graphs, each representing various types of data through bar and line graphs. The graphs are a form of data visualization used to represent information. Each graph is labeled with variables as per our dataset. The scales on the x-axes tell us if the patient is healthy or has preeclampsia, and y-axes tell us the amount of preeclampsia or healthy patients.

3.4 Feature Selection

When we have a large number of features, it's not always clear which ones are the most informative for our model. This is where feature selection methods like SelectKBest, Chi-Square test, and VIF test come in handy.

SelectKBest

SelectKBest is a method used for feature selection that removes all but the highest scoring features. It uses a scoring function to rate each feature and then selects the 'k' best scoring features. The scoring function could be anything from chi-squared for non-negative features to `f_classif` for F-value between label/feature for classification tasks and `f_regression` for F-value between label/feature for regression tasks. By selecting the 'k' best features, we found the least significant features like `CR_SE`, `Uric`, `Tpro`, `ALK`, `AL`, `CHOL`, `Urine_AC`.

Chi-Square Test

The Chi-Square test is a statistical test used to determine if there is a significant association between two categorical variables. In the context of feature selection, it's used to test the independence of features and labels. If the output (p-value) is less than a threshold (typically 0.05), it suggests that the feature is dependent on the label and should be kept. It's important to note that the Chi-Square test assumes that the variables are categorical and that the data meets certain criteria (such as expected frequencies of the categories). The Chi-Square test helped us identify and retain features that had a significant association with the output variable. Features that were independent of the output variable (i.e., had a high p-value in the Chi-Square test) such as `CR_SE`, `AL`, `Tbil`, `Trpo`, `Uric` were considered less informative.

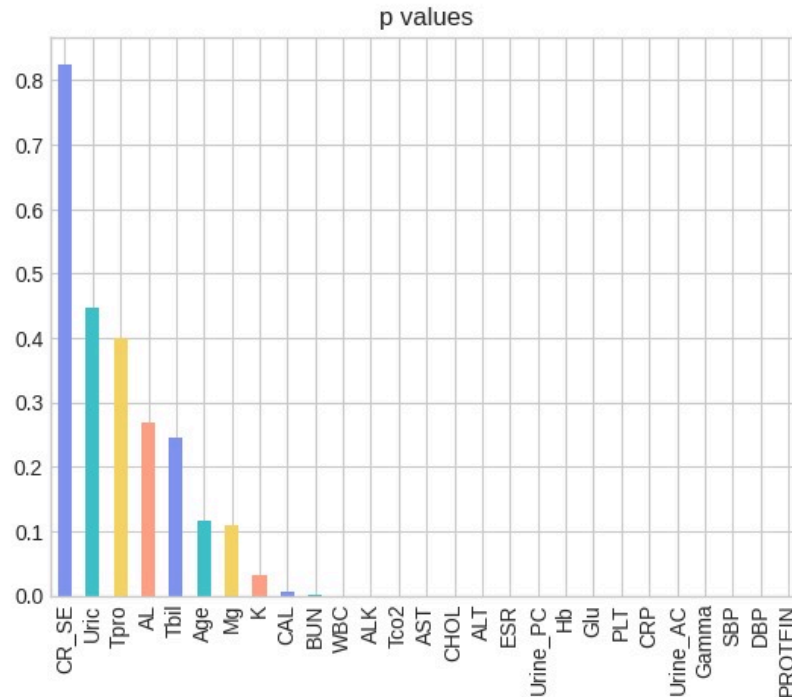


Fig. 3.9. p values from Chi-Square Test.

Variance Inflation Factor (VIF) Test

The Variance Inflation Factor (VIF) is a measure of multicollinearity in a set of multiple regression variables. Multicollinearity is when two or more independent variables in a regression are highly related to each other. High VIF (typically a $VIF > 5$) indicates high correlation and multicollinearity. In terms of feature selection, features with a high VIF are usually removed as they do not provide additional information that isn't already captured by other features. The VIF test was particularly useful for detecting multicollinearity, which is when one feature can be linearly predicted from others. Features with a high VIF such as Tpro, AL, CHOL, CR_SE didn't provide additional information that wasn't already captured by other features.

By employing these methods, SelectKBest, Chi-Square test, and VIF test, we were able to identify and remove similar less important features like AL, CHOL, Tpro, Uric, CR_SE. This improved the performance and interpretability of our machine learning models by reducing noise, multicollinearity, and overfitting. Eliminating these less significant features helps focus the model on the most impactful variables, leading to more efficient and accurate predictions.

3.5 Standardization

Data Standardization is the process of transforming raw data into a uniform format or structure, ensuring consistency and conformity to predefined rules. This process is crucial because it allows different systems to share and efficiently use data.

In Python, data standardization is often done using the 'StandardScaler' function from the 'sklearn.preprocessing' module. In our code:

- We first import the necessary modules.
- We create some example data in a NumPy array.
- We create a 'StandardScaler' object.
- We use the 'fit_transform' method of the 'StandardScaler' object to standardize the data. This method first fits the scaler to the data (i.e., computes the mean and standard deviation), and then it transforms the data using these parameters (i.e., subtracts the mean and divides by the standard deviation).
- Finally, we print the standardized data.

The result is that the data is transformed such that each feature (i.e., column) has a mean of 0 and a standard deviation of 1. This is important for many machine learning algorithms that are sensitive to the scale of the input features. By standardizing the data, we ensure that all features are on a similar scale, which can improve the performance of our machine learning models.

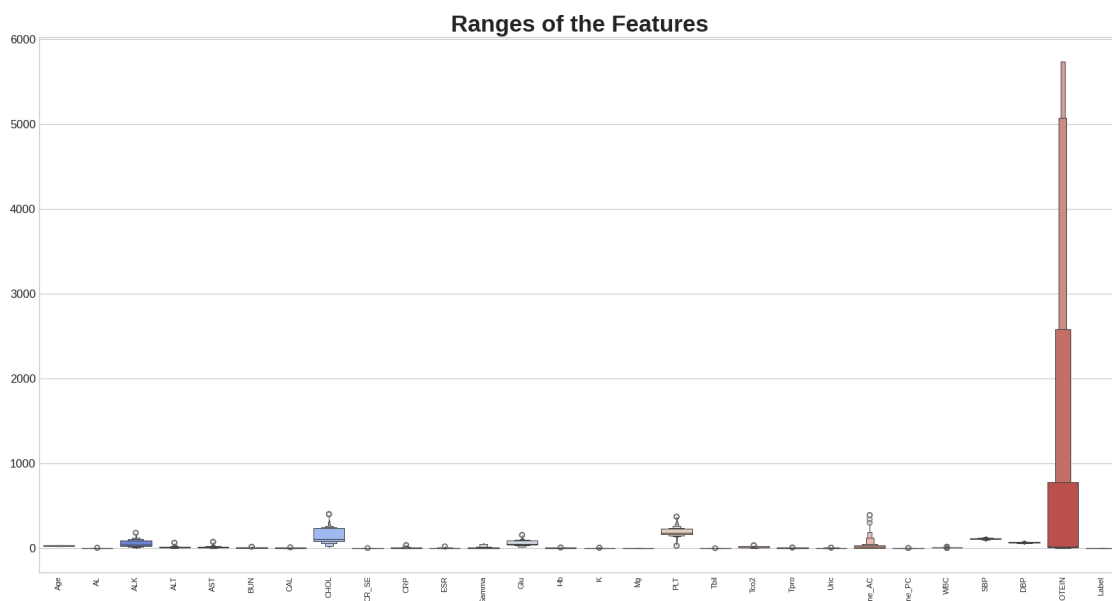


Fig. 3.10. Ranges of values from the online dataset

Figure 3.9 visualizes the range of values for different features in a dataset. The plot uses color to differentiate between the ranges of various features. The blue vertical bar represents a feature with a smaller range of values, as it is plotted lower on the y-axis. In contrast, the red vertical bar indicates a feature with a significantly higher range of values, extending almost to the top of the graph. The x-axis of the plot contains various symbols and texts representing different features, while the y-axis shows the numerical range of these features, ranging from 0 to 6000. Each vertical bar corresponds to a feature and has a small horizontal line at its top end, indicating the maximum value within that feature's range. The plot suggests that most features have very small ranges, as indicated by short lines close to zero on the y-axis, except for the feature represented by the red bar, which has a much larger range.

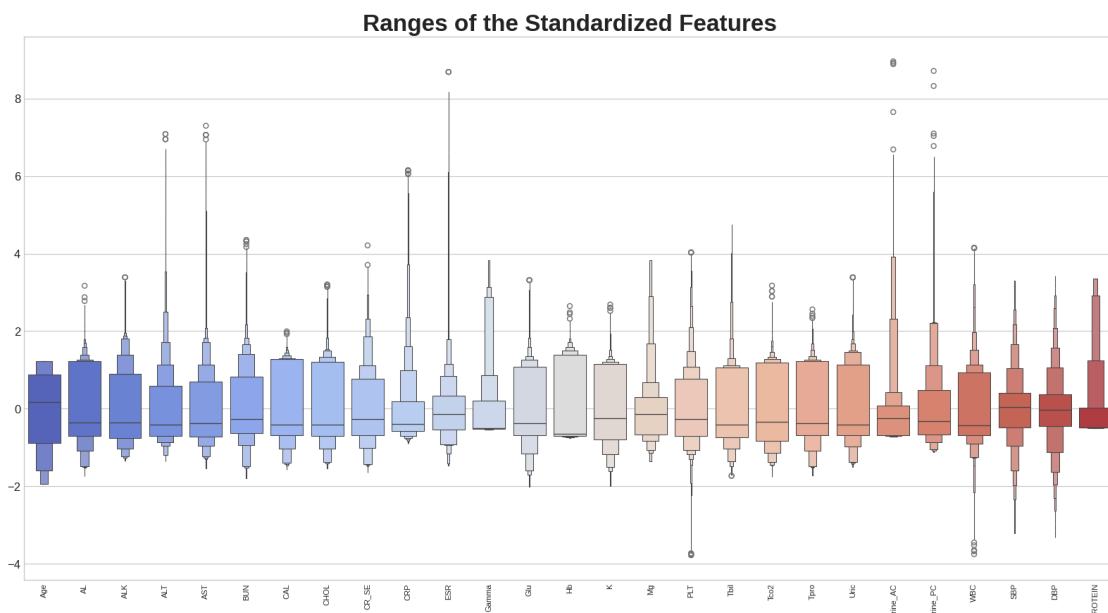


Fig. 3.11. Ranges of standardized values from the online dataset

Figure 3.10 visualizes the distribution of standardized feature values in a dataset. The graph uses different colors to represent distinct groups of features. There are three groups: blue, gray, and red. Each group's color indicates a set of features that have been standardized, with outliers shown as individual points above and below the main boxplot. The x-axis of the graph contains labels for different features, while the y-axis represents standardized values ranging from -4 to 8. The boxplots show the median, quartiles, and potential outliers for each feature. The median is indicated by a line within each box, and its position can suggest whether the distribution of values is skewed.

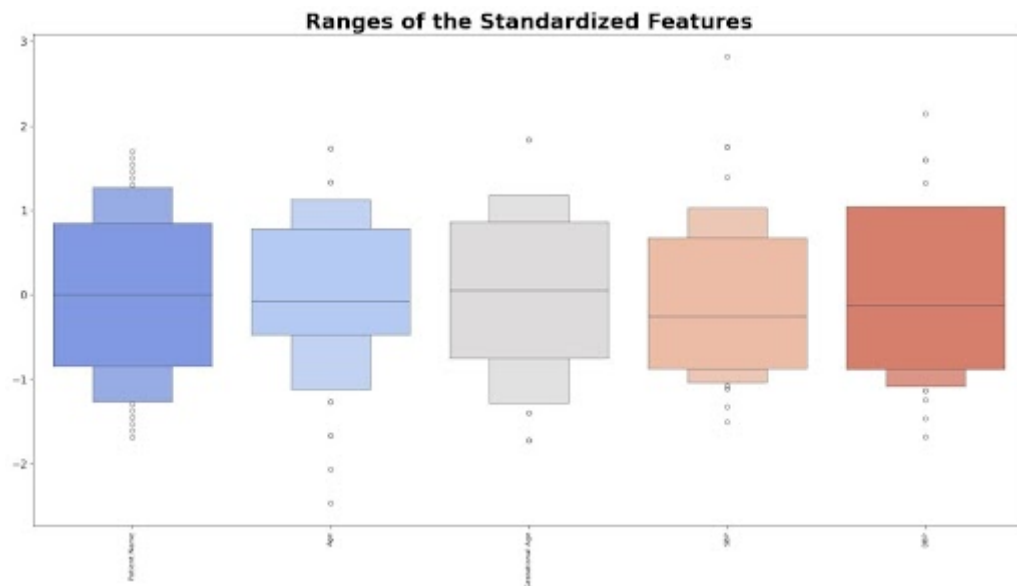


Fig. 3.12. Ranges of standardized values from the offline dataset

The image you've uploaded is a box plot graph titled "Ranges of the Standardized Features," which visualizes the distribution of standardized feature values in a dataset. The graph uses different colors to represent distinct groups of features. There are three groups: blue, grey, and red. Each group's color indicates a set of features that have been standardized, with outliers shown as individual points above and below the main boxplot. The x-axis of the graph contains labels for different features, while the y-axis represents standardized values ranging from -4 to 8. The boxplots show the median, quartiles, and potential outliers for each feature. The median is indicated by a line within each box, and its position can suggest whether the distribution of values is skewed.

This type of visualization is useful for comparing the distributions of multiple features after standardization, which typically involves scaling the features to have a mean of 0 and a standard deviation of 1. It helps in identifying features with extreme values or outliers that may affect the performance of machine learning models.

3.6 Cross Validation

Cross-validation is a technique used to assess the performance of machine learning models. It involves dividing the dataset into 'k' subsets, training the model on 'k-1' subsets, and validating the model on the remaining subset. This process is repeated 'k' times, each time with a different subset used for validation, to provide a more robust measure of model performance.

3.6.1 K fold Cross-Validation

We used K-fold cross-validation to handle the imbalanced data from the public dataset and mitigate overfitting. In K-fold cross-validation, the data is divided into K groups or "folds." The model is trained on K-1 folds and validated on the remaining fold, repeating this process K times so each fold serves as the validation set once. This approach is important because it provides a comprehensive assessment of the model's performance across different subsets of the data, improving its robustness and reliability.

By systematically varying the training and validation sets, K-fold cross-validation helps the model generalize better and perform more consistently on unseen data. This method is especially useful for handling imbalanced datasets, as it ensures that each fold is a representative sample of the entire dataset, reducing the risk of biased predictions.

3.6.2 Leave One Out Cross-Validation

Leave-One-Out Cross-Validation (LOOCV) is a specific form of cross-validation, an important technique in machine learning for assessing the performance of a model[46]. Here's how it works:

1. **Splitting the Dataset:** In LOOCV, the dataset is split into two parts: a training set and a testing set[46]. The training set includes all the data except for one observation[46]. This is where the method gets its name "leave-one-out" cross-validation[46].
2. **Model Building:** The model is built using only the data from the training set[46].
3. **Prediction and Evaluation:** The model is used to predict the response value of the one observation left out of the model and the Mean Squared Error (MSE) is calculated[46].
4. **Iteration:** This process is repeated n times (where n is the total number of observations in the dataset), leaving out a different observation from the training set each time[46].
5. **Average Error:** The test MSE is calculated as the average of all the MSEs[46].

The advantage of LOOCV is that it uses almost all the data for training, and hence can provide a less biased measure of model effectiveness[46]. However, it can be computationally expensive for large datasets, as the model needs to be retrained for each observation[46]. LOOCV is a useful technique for model validation as it provides a robust estimate of the model's performance[46]. It's particularly useful when you have a small dataset, as it allows you to make the most of your data[46]. This method has been used for the public dataset that we collected.

3.7 Classification

Many machine learning algorithms, prediction model establishment algorithms, and challenges to evaluate fetal health and predict and diagnose obstetric diseases exist for classification [26]. This study used k-nearest neighbor (KNN), logistic regression (LR), and support vector machine (SVM) models as four different ML techniques to address the classification problem and diagnose preeclampsia. The aggregation of the output from different models can improve the precision of the prediction. K-fold cross validation has been used for online data while leave-one-out method has been used for offline data.

3.7.1 K-Nearest Neighbor

This section describes the implementation of the K-Nearest Neighbors (KNN) classification algorithm. KNN is a non-parametric supervised learning method that classifies data points based on the similarity to their k nearest neighbors in the training data. The core principle is that data points are more likely to belong to the same class as their closest neighbors.

Model Training and Hyperparameter Tuning:

The KNN model was implemented using the scikit-learn library in Python. To optimize the model's performance, a GridSearchCV approach was employed. This technique involves defining a grid of hyperparameter values and evaluating the model's performance on each combination using cross-validation. The GridSearchCV object was configured with 10-fold cross-validation (cv=10) and utilized all available cores (n_jobs=-1) to expedite the search process. The model was evaluated using the F1-weighted scoring metric (scoring='f1_weighted'), which considers both precision and recall for imbalanced class distributions.

Model Fitting and Evaluation

Following hyperparameter tuning, a KNN model was instantiated with the optimal parameters and fit on the training data. The model's performance was then evaluated on the unseen test data. The accuracy score was calculated using the accuracy_score function from scikit-learn, and a classification report was generated using the classification_report function. The classification report provides detailed metrics such as precision, recall, F1-score, and support for each class.

3.7.2 Logistic Regression

Logistic Regression is a widely used classification algorithm that models the relationship between features and a binary class label using a sigmoid function. It is suitable for problems with well-separated classes and interpretable feature coefficients. A Logistic Regression model was initially fitted on the training data. The model's performance was evaluated using the `accuracy_score` function from `scikit-learn`, calculating both the training accuracy and test accuracy. On the other hand Hyperparameter tuning is crucial for optimizing the performance of machine learning models. In this study, a `RandomizedSearchCV` approach was employed to efficiently explore a defined hyperparameter space. `RandomizedSearchCV` iteratively samples random hyperparameter combinations from a provided grid and evaluates the model's performance on each combination using cross-validation. This approach is computationally less expensive compared to `GridSearchCV`, which exhaustively evaluates all possible combinations. A new Logistic Regression model is then instantiated with the optimal hyperparameters and refit on the training data for a final evaluation. The training and test accuracies of the best model are calculated, along with a classification report using the `classification_report` function.

This approach allows for a more robust evaluation of the Logistic Regression model's performance on unseen data and provides insights into the model's ability to classify data points correctly for each class.

3.7.3 Support Vector Machine

Support Vector Machine (SVM) is a powerful machine learning algorithm used for classification, regression, and outlier detection. It's particularly popular due to its ability to handle high dimensional data, and its versatility in modeling diverse sources of data. Here's how it works:

- **Hyperplane:** SVM works by finding a hyperplane that best separates the data into classes. In a two-dimensional space, this hyperplane is a line. But in higher dimensions, it could be a plane or a hyperplane.
- **Support Vectors:** Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, SVM maximizes the margin of the classifier.

- **Margin:** The margin is defined as the distance between the separating hyperplane (decision boundary) and the nearest data point from either class². The goal of an SVM is to find the optimal hyperplane that maximizes this margin².
- **Kernel Trick:** Real-world data is often not linearly separable. That is, you can't draw a straight line (or hyperplane) to separate the classes. The kernel trick is used to transform the input data into a higher-dimensional space where a hyperplane can be used to separate the data⁵.
- **Optimization:** SVMs are trained by solving a type of optimization problem that maximizes the margin². This involves some complex math, including concepts like Lagrange multipliers².

In Python, the `sklearn` library provides an implementation of SVMs. The `svm.SVC` function can be used for classification tasks.

SVMs are a powerful tool in machine learning. They can handle complex, high-dimensional data. However, they can be sensitive to the choice of the kernel parameters and the regularization term, which control the complexity of the decision boundary.

3.7.4 Classification

We began the process by acquiring data from an online dataset, which included various features that could be utilized in our machine learning model to assess the risk of preeclampsia. After obtaining the data, we proceeded with data preprocessing, which involved cleaning, transforming, and organizing the dataset to ensure it was high quality and compatible with the machine learning algorithm. Once the data was preprocessed, we conducted feature selection to identify the most relevant features that significantly contributed to the prediction of preeclampsia. This step helped streamline the model by reducing complexity and improving its accuracy.

Following feature selection, we performed data standardization and data validation to ensure the data was consistent and suitable for model training. We then split the data into a 70/30 ratio for training and testing, respectively. This separation allowed us to train the model on a larger portion of the data while retaining a subset for later testing.

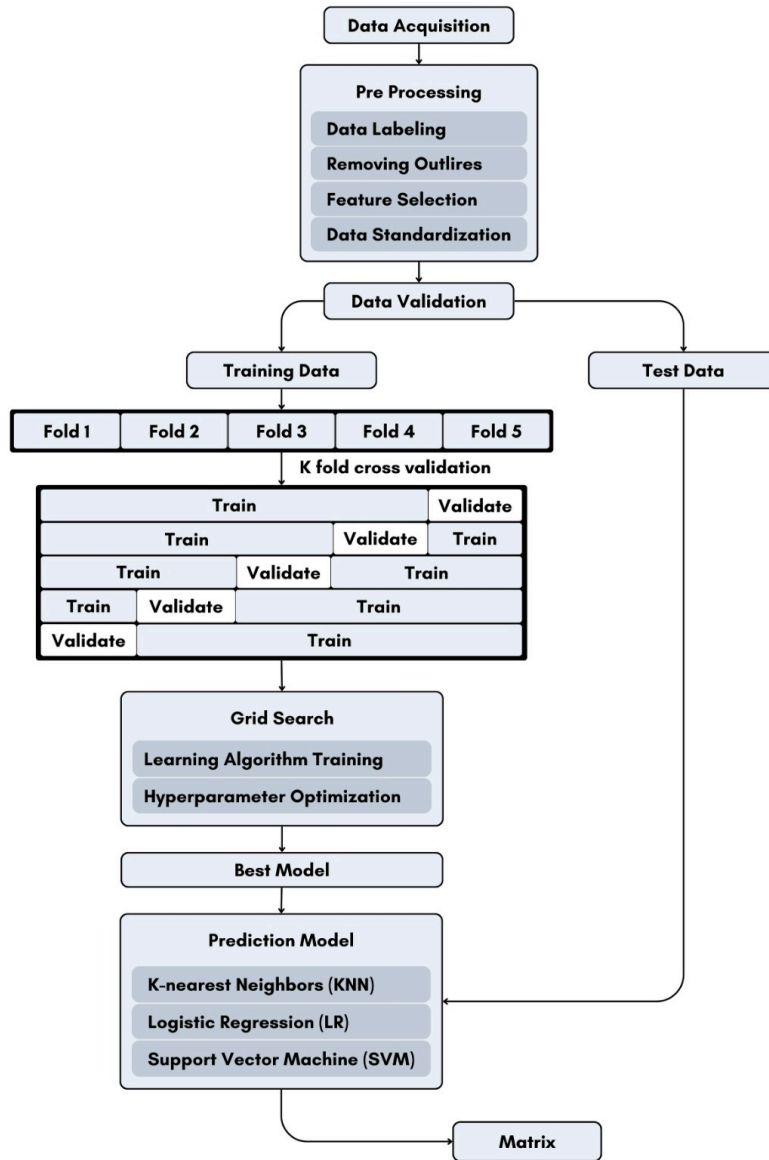


Fig. 3.13. Flow chart of the methodology of detection of preeclampsia for the online data set.

Additionally, we implemented K-fold cross-validation on the training data to enhance the model's generalizability and robustness. In the model training phase, the machine learning algorithm analyzed the data to identify patterns and relationships, creating a predictive model for preeclampsia risk. We then evaluated the model's performance by testing its predictions against a separate set of data with known outcomes. The evaluation results allowed us to analyze the accuracy and effectiveness of the early detection system, informing us of the model's potential utility in real-world clinical settings.

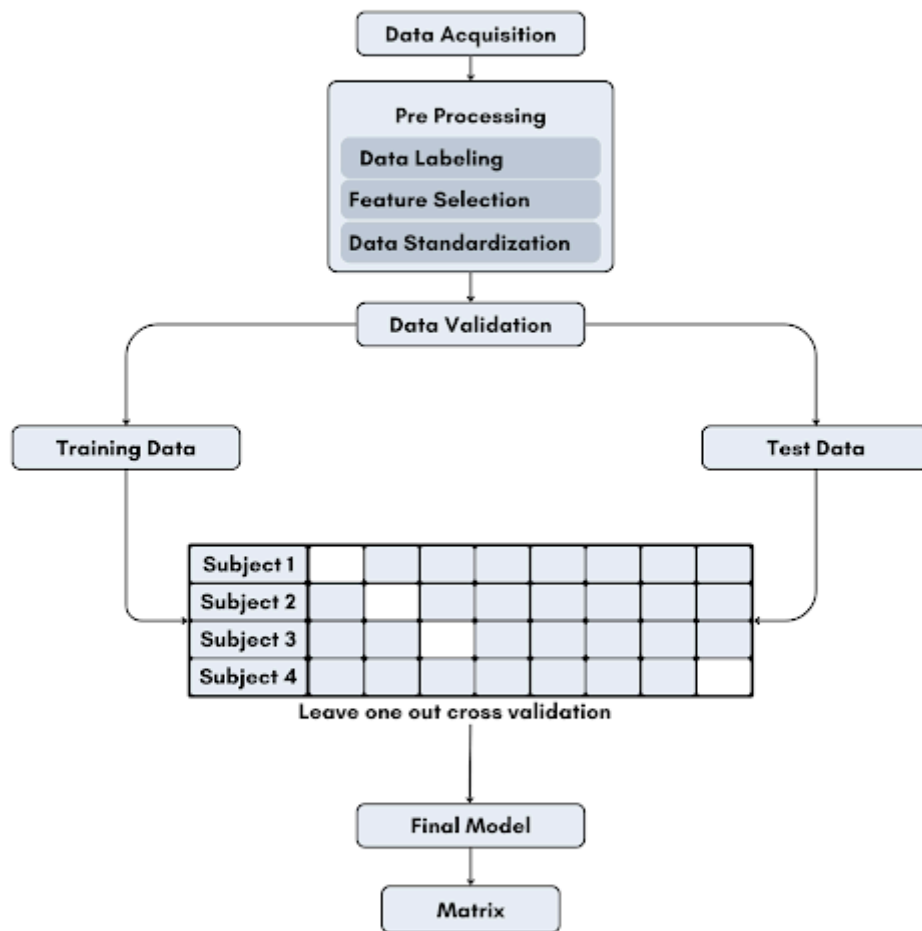


Fig. 3.14. Flow chart of the methodology of detection of preeclampsia for the local data set.

For our second model, we began by acquiring patient data from local hospitals, which included factors such as gastrocolic age and vital signs. Once the data was obtained, we initiated the preprocessing phase, where we cleaned, transformed, and organized the data to ensure its quality and compatibility with the machine learning algorithm. This step was essential to prepare the data for further analysis. After preprocessing, we proceeded with feature selection. Given the limited number of features in our dataset, we opted to use all of them for the model. This decision allowed us to maximize the use of the available data and potentially improve the model's performance.

Since we had a smaller dataset, we chose leave-one-out cross-validation to train the model. This approach involved leaving out one data point at a time for testing while using the rest for training, allowing us to make the most of the limited data we had. This method helped us assess the model's generalizability and robustness more accurately. After that, we used Logistic Regression to train our model. Once the model was trained, we evaluated its performance and effectiveness in predicting preeclampsia by testing it on the data that was excluded earlier in the process. This evaluation enabled us to analyze the model's accuracy

and the effectiveness of the early detection system based on the outcomes. Through this process, we gained insights into how well our model could predict preeclampsia and its potential utility in clinical practice.

3.8 Performance Evaluation

Several performance criteria were assessed in order to verify the effectiveness of the suggested technique for the identification of preeclampsia episodes. To find the best configuration in this model performance we have split the data. For this research train-validation-test splits is 75% training data, 12.5% validation, and 12.5% test data. In the performance evaluation positive (P) denotes an epoch that is predicted to be preeclampsia; negative (N) denotes an epoch that is not preeclampsia; true positive (TP) denotes an epoch that is positive and predicted to be positive; true negative (TN) denotes an epoch that is negative and predicted to be negative; false positive (FP) denotes an observation that is positive but predicted to be negative; and false negative (FN) defines an epoch as negative but predicted to be positive [24]. With the use of these factors, the results were computed, assessing F1-score, recall, accuracy, sensitivity, specificity, and precision. The following formulas apply to the performance metrics:

Sensitivity / True Positive Rate / Recall

$$sensitivity = \frac{TP}{TP + FN}$$

Sensitivity tells us what proportion of the positive class got correctly classified.

False Negative Rate

$$FNR = \frac{FN}{TP + FN}$$

False Negative Rate (FNR) tells us what proportion of the positive class got incorrectly classified by the classifier.

Specificity / True Negative Rate

$$specificity = \frac{TN}{TN + FP}$$

Specificity tells us what proportion of the negative class got correctly classified.

False Positive Rate

$$FPR = \frac{FP}{TN + FP} = 1 - Specificity$$

FPR tells us what proportion of the negative class got incorrectly classified by the classifier.

Precision:

$$Precision = \frac{TP}{TP + FP}$$

Recall:

$$Recall = \frac{TP}{TP + FN}$$

F1 Score:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Chapter 4: Results and Discussion

This chapter elucidates the empirical findings obtained from the rigorous analysis conducted as part of this research. The results are articulated in an unambiguous manner, buttressed by pertinent data, figures, and tables. Each result is expounded in relation to the research objectives and the extant body of literature. The discussion segment provides an in-depth interpretation and extrapolation of these results. It aligns the findings with the theoretical foundations and the research hypotheses posited at the commencement of the study. This segment also concedes the limitations encountered during the study and proffers potential avenues for future research. Collectively, the results and discussion chapter constitutes the fulcrum of the research, offering invaluable insights and contributing to the wider scholarly dialogue in the field.

Table 1: Performance of ML models

Dataset Type	ML Model	Precision	Recall	F1-score	Sensitivity	Specificity
Online	KNN	0.96	0.83	0.89	0.62	0.554
	LR	0.93	0.88	0.91	0.7268	0.73773
	SVM	0.97	0.94	0.95	0.92000	0.96769
Offline	LR	0.91	0.92	0.91	0.866666	0.958

The table shows the performance of three machine learning models (KNN, LR, SVM) for early detection of pre-eclampsia for online and LR for offline datasets. The performance is measured using five metrics: precision, recall, F1-score, sensitivity, and specificity.

Precision: measures the proportion of positive predictions that were actually correct. A high precision means that the model is good at filtering out false positives. In the context of pre-eclampsia detection, a high precision would mean that most of the patients flagged by the model as having pre-eclampsia actually have the condition.

Recall: measures the proportion of actual positive cases that were correctly identified by the model. A high recall means that the model is good at catching most of the true positives. In the context of pre-eclampsia detection, a high recall would mean that the model is able to identify most of the patients who actually have the condition.

F1-score: is a harmonic mean between precision and recall. It's a way to balance between the two metrics and get a single measure of the model's performance.

Sensitivity: is also known as the True Positive Rate (TPR). It measures the proportion of actual positive cases that were correctly identified by the model. In the context of pre-eclampsia detection, a high sensitivity would mean that the model is able to identify most of the patients who actually have the condition.

Specificity: is also known as the True Negative Rate (TNR). It measures the proportion of actual negative cases that were correctly identified by the model. In the context of pre-eclampsia detection, a high specificity would mean that the model is good at correctly identifying patients who do not have pre-eclampsia.

Looking at the table, we can see that all the models have a high precision (>0.91) for both online and offline datasets. This means that the models are good at filtering out false positives. The SVM model has the highest overall performance across all metrics for the online dataset, with a precision of 0.97, recall of 0.94, and F1-score of 0.95. The LR model has the highest overall performance across all metrics for the offline dataset, with a precision of 0.91, recall of 0.92, and F1-score of 0.91.

It's important to note that the best model for this task depends on the specific priorities. If it's more important to avoid false positives (e.g., to reduce unnecessary medical interventions), then a model with higher precision might be preferable. On the other hand, if it's more important to catch as many true positives as possible (e.g., to avoid missing cases of pre-eclampsia), then a model with higher recall might be preferable.

Overall, the results show that all three machine learning models have the potential to be useful for early detection of pre-eclampsia. The choice of the best model will depend on the specific priorities of the application.

4.1 Confusion Matrix

The confusion matrix is a visual tool used to evaluate the performance of a classification model. Here's an explanation of its components:

- **True Positives (TP):** The top left quadrant, red in color, with the label "Normal" for both actual and predicted values, indicating correct predictions when the actual class is positive.
- **False Negatives (FN):** The top right quadrant, blue in color, labeled "Normal" for actual values and "Suspect" for predicted values, indicating instances where the model incorrectly predicted the negative class when it was actually positive.

- **False Positives (FP):** The bottom left quadrant, also blue, labeled "Suspect" for actual values and "Normal" for predicted values, showing instances where the model incorrectly predicted the positive class when it was actually negative.
- **True Negatives (TN):** The bottom right quadrant, red in color, with both labels as "Suspect," indicating correct predictions when the actual class is negative.

The color difference, with red indicating correct predictions and blue indicating incorrect predictions, helps to quickly identify the model's accuracy in predicting each class. The values in each quadrant of the confusion matrix 8 represent the number of observations that fall into each category based on the model's predictions and the actual values.. This matrix is particularly useful for visualizing the performance of the SVM model in classifying data into 'Normal' or 'Suspect' categories.

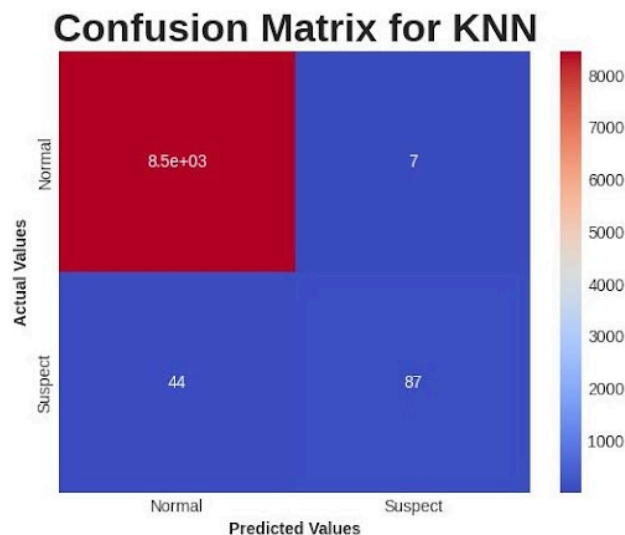


Fig. 4.1. Confusion matrix for the KNN model.

The confusion matrix shows the performance of the KNN model on a dataset of pre-eclampsia cases.

True Positives (TP): 8500. These are the cases where the model correctly predicted that a patient has pre-eclampsia.

False Negatives (FN): 7. These are the cases where the model incorrectly predicted that a patient does not have pre-eclampsia, but they actually do.

False Positives (FP): 44. These are the cases where the model incorrectly predicted that a patient has pre-eclampsia, but they actually do not.

True Negatives (TN): 87. These are the cases where the model correctly predicted that a patient does not have pre-eclampsia.

Based on the confusion matrix, we have calculated the precision, recall, f1-score for the KNN model's performance on this dataset. These metrics all indicate that the KNN model is performing very well on this dataset. It has a very high precision, which means that it is very good at filtering out false positives. It also has a very high recall, which means that it is very good at catching true positives. The F1-score is a harmonic mean of precision and recall, and it is also very high, which means that the model is balanced in its performance.

Confusion Matrix for Best Logistic Regression Model

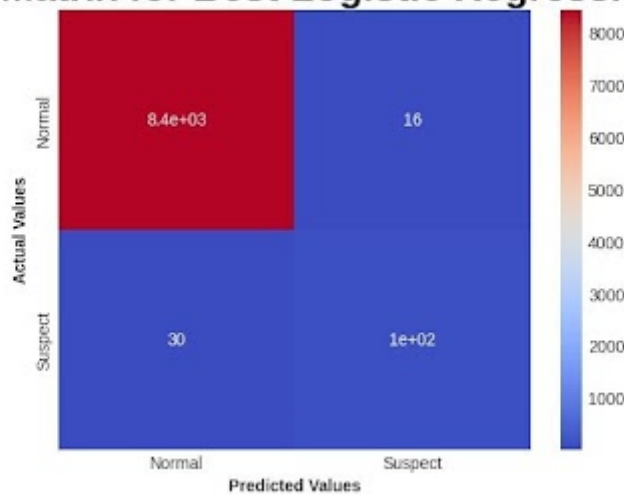


Fig. 4.2. Confusion matrix for the LR model.

The confusion matrix shows the performance of the LR model on a dataset of pre-eclampsia cases.

True Positives (TP): 8400. These are the cases where the model correctly predicted that a patient has pre-eclampsia.

False Negatives (FN): 16. These are the cases where the model incorrectly predicted that a patient does not have pre-eclampsia, but they actually do.

False Positives (FP): 30. These are the cases where the model incorrectly predicted that a patient has pre-eclampsia, but they actually do not.

True Negatives (TN): 100. These are the cases where the model correctly predicted that a patient does not have pre-eclampsia.

Based on the confusion matrix, we have calculated the precision, recall, f1-score for the LR model's performance on this dataset. These metrics all indicate that the LR model is performing very well on this dataset. It has a very high precision, which means that it is very good at filtering out false positives. It also has a very high recall, which means that it is very good at catching true positives. The F1-score is a harmonic mean of precision and recall, and it is also very high, which means that the model is balanced in its performance.

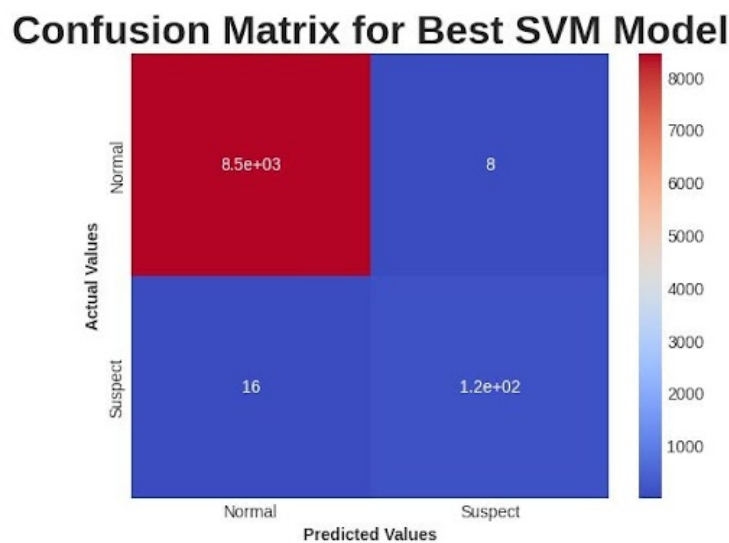


Fig. 4.3. Confusion matrix for the SVM model.

The confusion matrix shows the performance of the SVM model on a dataset of pre-eclampsia cases.

True Positives (TP): 8500. These are the cases where the model correctly predicted that a patient has pre-eclampsia.

False Negatives (FN): 8. These are the cases where the model incorrectly predicted that a patient does not have pre-eclampsia, but they actually do.

False Positives (FP): 16. These are the cases where the model incorrectly predicted that a patient has pre-eclampsia, but they actually do not.

True Negatives (TN): 120. These are the cases where the model correctly predicted that a patient does not have pre-eclampsia.

Based on the confusion matrix, we have calculated the precision, recall, f1-score for the SVM model's performance on this dataset. These metrics all indicate that the SVM model is

performing very well on this dataset. It has a very high precision, which means that it is very good at filtering out false positives. It also has a very high recall, which means that it is very good at catching true positives. The F1-score is a harmonic mean of precision and recall, and it is also very high, which means that the model is balanced in its performance.

4.2 ROC Curve Comparison

The Receiver Operating Characteristic (ROC) curve for three different machine learning models: K-Nearest Neighbors (KNN), Logistic Regression, and Support Vector Classifier (SVC). The ROC curve is a useful tool for evaluating the performance of binary classification models. It plots the True Positive Rate (TPR) on the y-axis against the False Positive Rate (FPR) on the x-axis. The TPR is the proportion of positive cases that were correctly identified by the model, while the FPR is the proportion of negative cases that were incorrectly classified as positive.

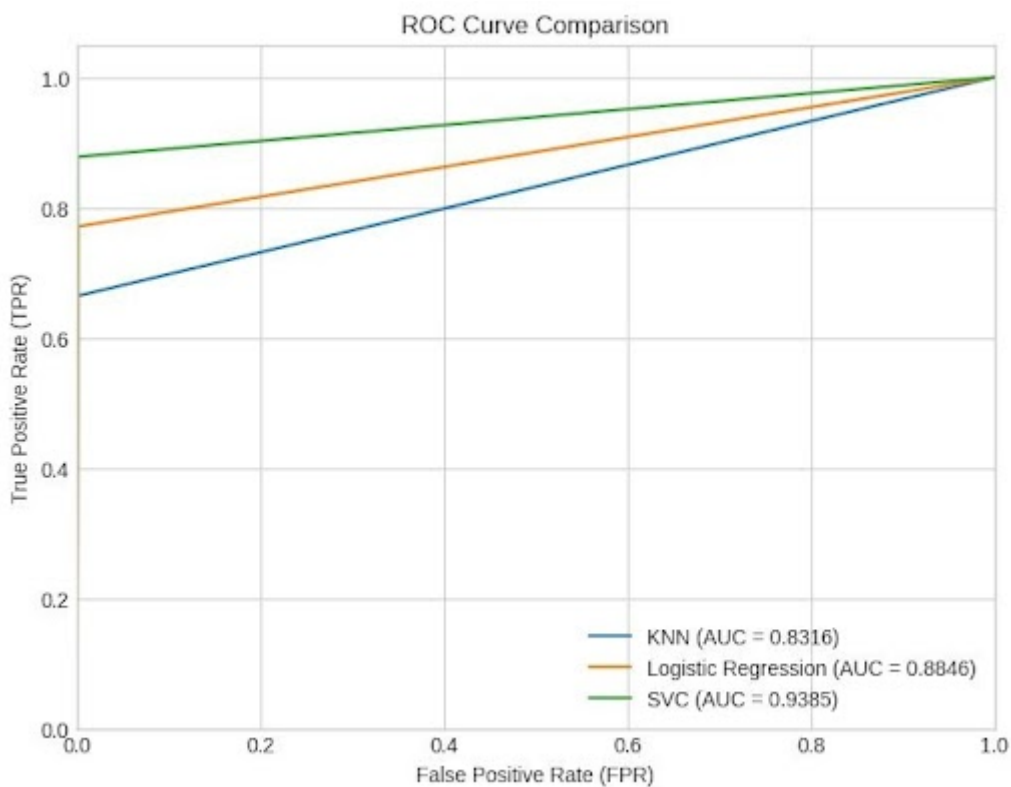


Fig. 4.4. ROC curve comparison for three different machine learning models.

The ROC curve for a perfect classifier would be a straight line in the upper left corner of the graph. This means that the model would be able to correctly classify all positive cases (TPR = 1) without any false positive classifications (FPR = 0). As the model's performance worsens, the ROC curve will deviate from this ideal line and move down and to the right.

In the ROC curve the SVC model has the highest Area Under the Curve (AUC) score, which is indicated by the largest enclosed area beneath its curve. This means that the SVC model has the best overall performance in classifying the positive and negative cases. The KNN model has the second-highest AUC score, followed by the Logistic Regression model.

Interpreting the AUC scores (0.8316 for KNN, 0.8846 for Logistic Regression, and 0.9385 for SVC) in your ROC AUC curve suggests the following:

- **SVC Model Performance:** The Support Vector Classifier (SVC) model, with an Area Under the Curve (AUC) of 0.9385, exhibits the highest classification ability among the three models being compared. This high AUC value signifies that the SVC model is exceptionally successful in distinguishing between positive and negative cases, which is a critical aspect of model performance. In simpler terms, the SVC model is the most accurate at correctly classifying positive instances, while also minimizing the number of false positive classifications. This balance between sensitivity (true positive rate) and specificity (true negative rate) is crucial in many applications, especially those where the cost of misclassification is high. The performance of the SVC model can be influenced by several factors, including the choice of kernel, the regularization parameter C , and the gamma parameter in the case of the RBF kernel. Therefore, careful tuning of these parameters using techniques like grid search or cross-validation can lead to improved model performance. Additionally, preprocessing steps such as feature scaling are particularly important for SVC as it is a distance-based algorithm. Lastly, while the AUC is a comprehensive performance metric, it's also important to consider other metrics such as precision, recall, F1 score, and confusion matrix depending on the specific problem and business context. These additional metrics can provide a more nuanced understanding of the model's performance. For instance, precision and recall are particularly useful when dealing with imbalanced datasets. Therefore, a comprehensive evaluation of the model should consider these various aspects.
- **Logistic Regression Performance:** The Logistic Regression model's performance is evaluated using the Area Under the Curve (AUC) metric, which in this case is 0.8846. This value, while lower than that of the Support Vector Classifier (SVC), is still considered good. The AUC value indicates that the model's predictive power is better than random guessing, meaning it can effectively distinguish between positive and negative cases. Compared to the K-Nearest Neighbors (KNN) model, the Logistic Regression model achieves a higher AUC. This suggests that it has a better balance between true positive rates (sensitivity) and false positive rates (1-specificity), which is

crucial in many applications to avoid costly misclassifications. The higher AUC also implies that the Logistic Regression model has a higher likelihood of ranking a randomly chosen positive instance higher than a randomly chosen negative one. It's worth noting that the performance of the Logistic Regression model can be influenced by several factors, including the selection of features, the handling of categorical variables, and the treatment of missing values. Therefore, careful preprocessing of the data and feature engineering can lead to improved model performance. Additionally, hyperparameter tuning, such as adjusting the regularization strength, can also help optimize the model's performance. Finally, it's important to remember that while AUC is a useful metric, it's not the only one to consider when evaluating model performance. Other metrics, such as precision, recall, and F1 score, should also be considered depending on the specific problem and business context.

- **KNN Model Performance:** The K-Nearest Neighbors (KNN) model's performance, as indicated by the Area Under the Curve (AUC) of 0.8316, is the lowest among the three models being compared. An AUC above 0.5 suggests that the model's predictions are better than random chance, but the lower AUC for the KNN model indicates a weaker ability to distinguish between positive and negative instances compared to the Support Vector Classifier (SVC) and Logistic Regression models. The suboptimal performance of the KNN model could be attributed to the selection of the distance metric or the number of neighbors used for classification. The distance metric in KNN determines how distance is calculated between instances, and choosing an inappropriate metric could lead to poor model performance. Similarly, the number of neighbors in KNN affects the decision boundary of the classifier. If too few or too many neighbors are chosen, the model may overfit or underfit the data, respectively. Therefore, it's crucial to fine-tune these parameters during the model building process to ensure optimal performance. Techniques such as grid search or cross-validation can be used to find the optimal values for these parameters. Additionally, preprocessing steps like feature scaling can also significantly impact the performance of KNN, as it is a distance-based algorithm. Hence, ensuring that the data is appropriately preprocessed before model training is another important consideration.

Chapter 5: Future work and Limitations

5.1 Limitations

Our research on the early detection of preeclampsia using machine learning faced substantial limitations due to challenges in collecting a comprehensive dataset from public hospitals in Bangladesh. The stringent rules and regulations of these hospitals created significant barriers to accessing data, as they required prolonged negotiations and bureaucratic procedures to gain necessary permissions. The lengthy and time-consuming process of securing ethical clearance from Institutional Review Boards further complicated data acquisition efforts. Many local public and private hospitals demonstrated a lack of focus on research and were hesitant to provide data, often discouraging our attempts. This limited our ability to obtain a comprehensive and diverse dataset necessary for building robust and generalizable predictive models. As a workaround, we relied on existing connections and permissions from hospital staff and doctors to manually collect a smaller set of data directly from patients, but this process was resource-intensive and constrained by logistical challenges.

The data we managed to gather was limited in both size and scope, leading to inconsistencies and reduced feature sets that may have impacted the accuracy and reliability of our machine learning models. These constraints also affected our ability to generalize the findings to a wider population, thereby potentially limiting the broader applicability of our models.

Overall, these limitations underscore the need for stronger collaborations with public healthcare institutions in future research endeavors. Establishing streamlined data acquisition protocols and fostering a culture of research within hospitals can facilitate the collection of more extensive and representative datasets. Such improvements would enhance the development of effective early detection models for preeclampsia and improve outcomes for pregnant individuals in Bangladesh and similar contexts worldwide.

5.2 Future work

In the future, we aim to address the limitations we faced during our project by collaborating with medical researchers who share similar interests. These partnerships will facilitate easier access to larger datasets, which will enable us to develop more comprehensive and reliable predictive models. We are currently in the process of collaborating with Dr. Fariha Haseen, Head of the Division of Reproductive and Child Health in the Department of Public Health & Informatics at Bangabandhu Sheikh Mujib Medical University, to acquire a substantial

amount of data for our project. Once we secure the data, we will employ advanced machine learning algorithms to create a predictive model for the early detection of preeclampsia in the first trimester. We will evaluate and compare various algorithms such as logistic regression, support vector machines, random forests, and deep neural networks to determine the most accurate and reliable model.

After thoroughly testing and confirming the effectiveness of our machine learning model, we will work closely with healthcare professionals and experts to integrate the model into real clinical settings. This collaboration will involve testing the model within existing healthcare systems and hospitals to ensure its practicality and efficiency. Additionally, we will develop user-friendly interfaces and decision support systems to aid healthcare providers in making precise and timely decisions regarding the diagnosis and management of preeclampsia. Through these efforts, we hope to improve patient care and outcomes while providing valuable tools for healthcare practitioners in the early detection and management of preeclampsia.

5.3 Conclusion

Our research on the early detection of preeclampsia using machine learning has shown significant promise in improving prenatal care and health outcomes for mothers and infants worldwide. We tackled the challenge of working with imbalanced datasets and successfully employed a variety of machine learning models to predict preeclampsia risk. Here we have created two machine learning models, one for a public dataset we have obtained from online and one for the data we have collected from a local hospital. We faced the challenge of working with imbalanced datasets but successfully employed a variety of machine learning models to predict preeclampsia risk, creating one model using a public dataset sourced online and another using data collected from a local hospital. In our first model, the combination of machine learning techniques such as k-nearest neighbor (KNN), logistic regression (LR), and support vector machine (SVM), along with the integration of advanced algorithms, helped improve model accuracy and reliability. For our local dataset, we employed leave-one-out cross-validation methods to rigorously test the model's performance in predicting preeclampsia with a high degree of confidence. Our careful data preprocessing, feature selection, and performance assessment strategies allowed us to gain meaningful insights into the patterns and relationships within the data, leading to more precise and effective prediction models. One of the key achievements of our research was the utilization of online and offline datasets to develop and validate our models. Despite facing limitations in obtaining data from

local hospitals due to ethical and logistical challenges, we managed to overcome these obstacles and utilize the data available to us effectively. Our findings hold great potential for informing clinical practice by providing healthcare professionals with reliable tools for early detection and prediction of preeclampsia. By identifying individuals at risk early in pregnancy, healthcare providers can closely monitor these patients and take timely actions to mitigate the risks associated with this serious condition. This can lead to better management strategies, reduced maternal and neonatal mortality and morbidity, and improved overall prenatal care.

The study highlights the urgent need for effective and cost-efficient early detection methods, especially in low- and middle-income countries where preeclampsia has a disproportionately high impact. Our research contributes to the broader objective of improving global maternal and fetal health by developing robust machine learning models that can be integrated into clinical practice. Moving forward, further research and collaboration with medical institutions and experts are essential to refine and optimize our models. Additionally, ongoing efforts to expand the dataset and improve data quality will enhance the generalizability and applicability of our findings. By continuing to invest in this area, we aim to develop cutting-edge solutions that will revolutionize prenatal care and save lives around the world.

Chapter 6: CO-PO mapping

Course Outcome (CO)	Program Outcome (PO)	Knowledge Profile (K)	Complex Engineering Problems, Solutions, and Activities (P&A)	Our Project
CO1	PO1	K1	P1	This course outcome likely relates to fundamental knowledge or skills. Our project (P1) aligns with it by using ML to analyze blood samples for early detection of preeclampsia, which requires a strong foundation in relevant medical and machine learning knowledge (K1).
CO2	PO2	K2	P2, A2	CO2 may involve more advanced concepts or skills. Our project (P2, A2) contributes to this outcome by implementing complex ML algorithms and addressing challenging issues in early preeclampsia detection. It requires knowledge (K2) and advanced analytical techniques (A2).
CO3	PO3	K3	P1, P3	CO3 could focus on problem-solving. Our project (P1, P3) aligns with this by addressing the critical issue of preeclampsia detection, involving problem-solving skills and knowledge of relevant medical data (K3).
CO4	PO4	K4	P1, P4, A5	CO4 may relate to interdisciplinary skills and teamwork. Our project (P1, P4, A5) aligns with this by involving multiple aspects such as data collection, algorithm development (K4), and teamwork (A5).

CO5	PO5	K3, K4, K5, K6, K8	P1	CO5 seems to involve a broader knowledge base. Our project (P1) aligns with this by requiring a combination of various knowledge profiles (K3, K4, K5, K6, K8) to effectively address the complexities of preeclampsia detection
CO6	PO6	K6	P1, P6, A2	CO6 might focus on specialization in a particular area. Our project (P1, P6, and A2) aligns with this by specializing in early preeclampsia detection (K6) and implementing advanced ML techniques (A2).
CO7	PO7	K7	P7, A2	CO7 could involve ethical and societal aspects. Our project (P7, A2) aligns with this by considering ethical implications (K7) and addressing societal challenges (A2) in healthcare.
CO8	PO8	K3, K4, K5, K6, K8	P1	CO8 appears to encompass a comprehensive understanding of multiple areas. Our project (P1) aligns with this by leveraging a combination of knowledge profiles (K3, K4, K5, K6, and K8) to tackle the complex problem of early preeclampsia detection.

References

- [1] Fabjan-Vodusek, V.; Kumer, K.; Osredkar, J.; Verdenik, I.; Gersak, K.; Premru-Srsen, T. Correlation between Uterine Artery Doppler and the sFlt-1/PlGF Ratio in Different Phenotypes of Placental Dysfunction. *Hypertens. Pregnancy* 2019, 38, 32–40.
- [2] Alrowaili, M.M.; Zakari, N.M.A.; Hamadi, H.Y.; Moawed, S. Management of Gestational Hypertension Disorders in Saudi Arabia by Primary Care Nurses. *Saudi Crit. Care J.* 2020, 4, 103.
- [3] Vesna Fabjan-Vodušek, K. Kumer, Joško Osredkar, I. Verdenik, Ksenija Geršak, and T. Premru-Srsen, “Correlation between uterine artery Doppler and the sFlt-1/PlGF ratio in different phenotypes of placental dysfunction,” *Hypertension in Pregnancy*, vol. 38, no. 1, pp. 32–40, Nov. 2018.
- [4] Karrar and P. L. Hong, “Preeclampsia,” *PubMed*, Feb. 13, 2023.
- [5] J. M. Roberts and H. S. Gammill, “Preeclampsia,” *Hypertension*, vol. 46, no. 6, pp. 1243–1249, Dec. 2005, doi:<https://doi.org/10.1161/01.hyp.0000188408.49896.c5>.
- [6] S. M. de C. Pelicia, S. M. W. Fekete, J. E. Corrente, and L. M. S. de S. Rugolo, “Impact of early-onset preeclampsia on feeding tolerance and growth of very low birth weight infants during hospitalization,” *Revista Paulista de Pediatria*, vol. 41, 2023, doi: <https://doi.org/10.1590/1984-0462/2023/41/2021203>.
- [7] E. A. Rokotyanskaya et al., “Technologies for Prediction of Preeclampsia,” *Sovremennye Tekhnologii V Meditsine*, vol. 12, no. 5, pp. 78–84, 2021, doi: <https://doi.org/10.17691/stm2020.12.5.09>.
- [8] A. Bertini, R. Salas, S. Chabert, L. Sobrevia, and F. Pardo, “Using Machine Learning to Predict Complications in Pregnancy: A Systematic Review,” *Front Bioeng Biotechnol*, vol. 9, no. January, pp. 1–16, 2022, doi: 10.3389/fbioe.2021.780389.
- [9] I. C. Udenze, A. P. Arikawe, C. C. Makwe, and O. F. Olowoselu, “A prospective cohort study on the clinical utility of second trimester mean arterial blood pressure in the prediction of late-onset preeclampsia among Nigerian women,” *Niger J Clin Pract*, vol. 20, no. 6, pp. 741–745, 2017, doi: 10.4103/1119-3077.208948.
- [10] I. Bellos, V. Pergialiotis, D. Loutradis, A. Papapanagiotou, and G. Daskalakis, “The role of hemoglobin degradation pathway in preeclampsia: A systematic review and meta-analysis,” *Placenta*, vol. 92, pp. 9–16, 2020, doi: 10.1016/J.PLACENTA.2020.01.014.

- [11]R. T.-A. Rahman, M.-M. Lakulu, and I.-Y. Panessai, “Advancing Preeclampsia Prediction with Machine Learning: a Comprehensive Systematic Literature Review,” Original Research Paper International Journal of Intelligent Systems and Applications in Engineering IJISAE, vol. 11, no. 3, pp. 13–23, Jul. 2023, Accessed: Apr. 14, 2024. [Online]. Available: <https://ijisae.org/index.php/IJISAE/article/view/3138>
- [12]Y. C. Lin et al., “Preeclampsia Predictor with Machine Learning: A Comprehensive and Bias-Free Machine Learning Pipeline,” Jun. 2022, doi: <https://doi.org/10.1101/2022.06.08.22276107>.
- [13]V. P. Kovacheva, B. W. Eberhard, R. Y. Cohen, M. Maher, R. Saxena, and K. J. Gray, “Prediction of Preeclampsia from Clinical and Genetic Risk Factors in Early and Late Pregnancy Using Machine Learning and Polygenic Risk Scores,” Feb. 2023, doi: <https://doi.org/10.1101/2023.02.03.23285385>.
- [14]B. C. Loftness, I. Bernstein, C. A. McBride, N. Cheney, E. W. McGinnis, and R. S. McGinnis, “Preterm Preeclampsia Risk Modeling: Examining Hemodynamic, Biochemical, and Biophysical Markers Prior to Pregnancy,” Mar. 2023, doi: <https://doi.org/10.1101/2023.02.28.23286590>.
- [15]Y. Lin et al., “A Comprehensive and Bias-Free Machine Learning Approach for Risk Prediction of Preeclampsia with Severe Features in a Nulliparous Study Cohort,” Apr. 2023, doi: <https://doi.org/10.21203/rs.3.rs-2635419/v1>.
- [16]A. A. Jamwa, M. Mvurya, A. Luvanda, and P. Kimetto, “An Appropriate Feature Selection Technique for Use on Socio-Demographic Predictor Variables to Enable Early Detection of Preeclampsia: A Review of Literature,” Computer Engineering and Intelligent Systems, vol. 13, no. 4, Aug. 2022, doi: <https://doi.org/10.7176/ceis/13-4-02>.
- [17]S. S. Aljameel et al., “Prediction of Preeclampsia Using Machine Learning and Deep Learning Models: A Review,” Big Data and Cognitive Computing, vol. 7, no. 1, p. 32, Feb. 2023, doi: <https://doi.org/10.3390/bdcc7010032>.
- [18]A. Saadaty, S. Parhoudeh, K. Khashei Varnamkhasti, M. Moghanibashi, and S. Naeimi, “Preeclampsia Susceptibility Assessment Based on Deep Learning Modeling and Single Nucleotide Polymorphism Analysis,” Biomedicines, vol. 11, no. 5, p. 1257, Apr. 2023, doi: <https://doi.org/10.3390/biomedicines11051257>.
- [19]A. R. Mazloom et al., “Detecting preeclampsia with a multiple protein serum test: Assay and algorithm development,” Jan. 2023, doi: <https://doi.org/10.21203/rs.3.rs-2196679/v1>.

- [20]G. A. Dekker and B. M. Sibai, “Early detection of preeclampsia,” *American Journal of Obstetrics and Gynecology*, vol. 165, no. 1, pp. 160–172, Jul. 1991, doi: [https://doi.org/10.1016/0002-9378\(91\)90245-m](https://doi.org/10.1016/0002-9378(91)90245-m).
- [21]L. Gómez-Jemes, A. M. Oprescu, Á. Chimenea-Toscano, L. García-Díaz, and M. del C. Romero-Tertero, “Machine Learning to Predict Pre-Eclampsia and Intrauterine Growth Restriction in Pregnant Women,” *Electronics*, vol. 11, no. 19, p. 3240, Oct. 2022, doi: <https://doi.org/10.3390/electronics11193240>.
- [22]A. Bertini, R. Salas, S. Chabert, L. Sobrevia, and F. Pardo, “Using Machine Learning to Predict Complications in Pregnancy: A Systematic Review,” *Frontiers in Bioengineering and Biotechnology*, vol. 9, Jan. 2022, doi: <https://doi.org/10.3389/fbioe.2021.780389>.
- [23]C. Villalain, I. Herraiz, P. Domínguez-Del Olmo, P. Angulo, J. L. Ayala, and A. Galindo, “Prediction of Delivery Within 7 Days After Diagnosis of Early Onset Preeclampsia Using Machine-Learning Models,” *Frontiers in Cardiovascular Medicine*, vol. 9, Jul. 2022, doi: <https://doi.org/10.3389/fcvm.2022.910701>.
- [24]K. Ratnik et al., “Novel Early Pregnancy Multimarker Screening Test for Preeclampsia Risk Prediction,” *Frontiers in Cardiovascular Medicine*, vol. 9, Jul. 2022, doi: <https://doi.org/10.3389/fcvm.2022.932480>.
- [25]D. Zheng et al., “Comparison of machine learning and logistic regression as predictive models for adverse maternal and neonatal outcomes of preeclampsia: A retrospective study,” *Frontiers in Cardiovascular Medicine*, vol. 9, p. 959649, Oct. 2022, doi: <https://doi.org/10.3389/fcvm.2022.959649>.
- [26]R. A. Abbas, L. Ghulmiyyah, E. Hobeika, I. M. Usta, F. Mirza, and A. H. Nassar, “Preeclampsia: A Review of Early Predictors,” *Maternal-Fetal Medicine*, vol. 3, no. 3, pp. 197–202, Feb. 2021, doi: <https://doi.org/10.1097/fm9.0000000000000088>.
- [27]D. Yousif et al., “Autonomic Dysfunction in Preeclampsia: A Systematic Review,” *Frontiers in Neurology*, vol. 10, Aug. 2019, doi: <https://doi.org/10.3389/fneur.2019.00816>.
- [28]M. Liu et al., “Development of a prediction model on preeclampsia using machine learning-based method: a retrospective cohort study in China,” *Frontiers in Physiology*, vol. 13, Aug. 2022, doi: <https://doi.org/10.3389/fphys.2022.896969>.
- [29]L. Han, O. J. Holland, F. D. S. Costa, and A. V. Perkins, “Potential Biomarkers for late-onset and Term preeclampsia: A Scoping Review,” 2023.

- [30]R. Rescinito, M. Ratti, A. B. Payedimarri, and M. Panella, “Prediction Models for Intrauterine Growth Restriction Using Artificial Intelligence and Machine Learning: A Systematic Review and Meta-Analysis,” *Healthcare*, vol. 11, no. 11, p. 1617, Jun. 2023, doi: <https://doi.org/10.3390/healthcare11111617>.
- [31]A. V. Timofeeva, I. S. Fedorov, Y. V. Sukhova, T. Y. Ivanets, and G. T. Sukhikh, “Prediction of Early- and Late-Onset Pre-Eclampsia in the Preclinical Stage via Placenta-Specific Extracellular miRNA Profiling,” *International Journal of Molecular Sciences*, vol. 24, no. 9, p. 8006, Apr. 2023, doi: <https://doi.org/10.3390/ijms24098006>.
- [32]S. L. Malone, R. H. Yahya, and S. C. Kane, “Reviewing Accuracy of First Trimester Screening for Preeclampsia Using Maternal Factors and Biomarkers,” *International Journal of Women’s Health*, vol. Volume 14, pp. 1371–1384, Sep. 2022, doi: <https://doi.org/10.2147/ijwh.s283239>.
- [33]A.-S. Melinte-Popescu, I.-A. Vasilache, D. Socolov, and M. Melinte-Popescu, “Predictive Performance of Machine Learning-Based Methods for the Prediction of Preeclampsia—A Prospective Study,” *Journal of Clinical Medicine*, vol. 12, no. 2, p. 418, Jan. 2023, doi: <https://doi.org/10.3390/jcm12020418>.
- [34]C. Garrido-Giménez et al., “Predictive Model for Preeclampsia Combining sFlt-1, PlGF, NT-proBNP, and Uric Acid as Biomarkers,” *Journal of Clinical Medicine*, vol. 12, no. 2, p. 431, Jan. 2023, doi: <https://doi.org/10.3390/jcm12020431>.
- [35]R. Bennett, Z. D. Mulla, P. Parikh, A. Hauspurg, and T. Razzaghi, “An imbalance-aware deep neural network for early prediction of preeclampsia,” *PLOS ONE*, vol. 17, no. 4, p. e0266042, Apr. 2022, doi: <https://doi.org/10.1371/journal.pone.0266042>.
- [36]P. L. Hedley, C. M. Hagen, C. Wilstrup, and M. Christiansen, “The use of artificial intelligence and machine learning methods in early pregnancy pre-eclampsia screening: A systematic review protocol,” *PLOS ONE*, vol. 18, no. 4, p. e0272465, Apr. 2023, doi: <https://doi.org/10.1371/journal.pone.0272465>.
- [37]N. O’Gorman, K. H. Nicolaides, and L. C. Poon, “The Use of Ultrasound and other Markers for Early Detection of Preeclampsia,” *Women’s Health*, vol. 12, no. 2, pp. 199–207, Mar. 2016, doi: <https://doi.org/10.2217/whe.15.95>.
- [38]A. Melinte, I.-A. Vasilache, D. Socolov, and M. Melinte, “Predictive Performance of Machine Learning- Based Methods for The Prediction of Preeclampsia- A Prospective Study,” Nov. 2022, doi: <https://doi.org/10.20944/preprints202211.0539.v1>.

- [39]S. Li et al., “Improving preeclampsia risk prediction by modeling pregnancy trajectories from routinely collected electronic medical record data,” *npj Digital Medicine*, vol. 5, no. 1, Jun. 2022, doi: <https://doi.org/10.1038/s41746-022-00612-x>.
- [40]Z. Ansbacher-Feldman, A. Syngelaki, H. Meiri, R. Cirkin, K. H. Nicolaides, and Y. Louzoun, “Machine-learning-based prediction of pre-eclampsia using first-trimester maternal characteristics and biomarkers,” *Ultrasound in Obstetrics & Gynecology*, vol. 60, no. 6, pp. 739–745, Dec. 2022, doi: <https://doi.org/10.1002/uog.26105>.
- [41]Zach, “What is Univariate Analysis? (Definition & Example),” *Statology*, Feb. 25, 2021. <https://www.statology.org/univariate-analysis/>
- [42]S. Glen, “Univariate Analysis: Definition, Examples,” *Statistics How To*, Dec. 10, 2014. <https://www.statisticshowto.com/univariate/>
- [43]IBM, “What Is Exploratory Data Analysis? | IBM,” *www.ibm.com*, 2020. <https://www.ibm.com/topics/exploratory-data-analysis>
- [44]M. Santos, “A Data Scientist’s Essential Guide to Exploratory Data Analysis,” *Medium*, Jul. 11, 2023. <https://towardsdatascience.com/a-data-scientists-essential-guide-to-exploratory-data-analysis-25637eee0cf6>
- [45]Q. Collaborators, “Bivariate Analysis: What is it, Types + Examples,” *QuestionPro*, Feb. 15, 2023. <https://www.questionpro.com/blog/bivariate-analysis/>
- [46]Zach, “A Quick Intro to Leave-One-Out Cross-Validation (LOOCV),” *Statology*, Nov. 03, 2020. <https://www.statology.org/leave-one-out-cross-validation/>

Appendix A

Source Code

```
1. Import Library
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import preprocessing
import random
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.model_selection import RandomizedSearchCV
from sklearn.metrics import accuracy_score, classification_report
from sklearn.metrics import confusion_matrix

# for Machine learning
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import StackingClassifier
from sklearn.metrics import accuracy_score, classification_report, roc_auc_score,
confusion_matrix
from sklearn.model_selection import train_test_split, cross_val_score, GridSearchCV,
ShuffleSplit
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import f_classif
from sklearn.model_selection import learning_curve
import pickle

# To ignore unwanted warnings
import warnings
warnings.filterwarnings('ignore')

# for styling
plt.style.use('seaborn-whitegrid')
2. Load and split data
from google.colab import drive
drive.mount('/content/drive')
# Load your CSV data into a DataFrame, replace 'your_data.csv' with your actual file
path
df = pd.read_csv('/content/drive/MyDrive/Capstone/output_dataset.csv')
df.head(5)
3.Exploratory data analysis
Dataset Overview
# Label encoding
# (1=in care patient), (0=out care patient)
```

```

df['Label'] = df.Label.replace({"Preeclampsia":1, 'Normal':0})

# get all the features
features = [feat for feat in df.columns if feat != 'Label']
print(f"The dataset contains {df.shape[0]} rows and {df.shape[1]} columns")

num_features = [feat for feat in features if df[feat].dtype != object]
cat_features = [feat for feat in features if df[feat].dtype == object]

print(f"Total number of features : {len(features)}")
print(f"Number of numerical features : {len(num_features)}")
print(f"Number of categorical features : {len(cat_features)}\n")
df.hist(bins=30, figsize=(30, 30))
Missing values
# Show the percentage of missing values in each feature

df.isna().mean().to_frame(name='Missing %')
Unique values
# To count number unique values in each feature

df.nunique().to_frame(name='# of unique values')
3.1 Univariate analysis
3.1.1 Target

name : Label

There are two classes in the target

0 : Normal patient 1 : Preeclampsia patient
# Target distribution

plt.figure(figsize=(7,4))
fig = sns.histplot(x='Label', data=df)
# Check no. of data points for each class

df.Label.value_counts().to_frame(name='Count')
corellation
plt.figure(figsize=(20, 10))
sns.heatmap(df.corr(), annot=True)
plt.show()
# Describe statistics for each numerical features

df[num_features].describe()
# Boxplot for each numerical feature
fig, axes = plt.subplots(len(num_features), 1, figsize=(8, 15))

# Only loop through the number of available axes
for i, c in enumerate(num_features[:len(axes)]):
    f = df[[c]].boxplot(ax=axes[i], vert=False)
# Kernel Density Estimation plot for each numerical feature
fig, axes = plt.subplots(len(num_features), 1, figsize=(8, 15))

```

```

# Only loop through the number of available axes
for i, c in enumerate(num_features[:len(axes)]):
    f = df[[c]].plot(kind='kde', ax=axes[i])
from scipy import stats

# Data noise using z score method

z_scores = stats.zscore(df.select_dtypes(include=["float64", "int64"]),
nan_policy='omit')

z_scores_df = pd.DataFrame(z_scores, columns=df.select_dtypes(include=["float64",
"int64"]).columns)

outliers_filtered_df = df[(z_scores_df < 3).all(axis=1) & (z_scores_df >
-3).all(axis=1)]

print('Original dataset shape:', df.shape)
print('Filtered dataset shape:', outliers_filtered_df.shape)
# Boxplot for each numerical feature
fig, axes = plt.subplots(len(num_features), 1, figsize=(8, 15))

# Only loop through the number of available axes
for i, c in enumerate(num_features[:len(axes)]):
    f = outliers_filtered_df[[c]].boxplot(ax=axes[i], vert=False)

Feature Selection

Feature selection should be performed to find the best set of features that allow the
creation of optimized models of the studied phenomena.

# Feature selection with K-Best
X = df.drop(['Label'], axis=1)
Y = df['Label']

bestfeatures = SelectKBest(score_func=f_classif, k='all')

fit = bestfeatures.fit(X, Y)
dfscores = pd.DataFrame(fit.scores_)
SelectKBest uses statistical tests such as the chi-square test, ANOVA F test, or mutual
information score to evaluate and rank features based on their correlation with the
target variable. It then selects the top K features with the most important importance
score to be included in the final feature subset.
# Concatenate two dataframes for better understanding
featureScores = pd.concat([dfcolumns, dfscores], axis=1)
featureScores.columns = ['Feature', 'Score']

# Sort features based on importance scores
featureScores = featureScores.sort_values(by='Score', ascending=False)
print(featureScores)
# Visualize the importance scores
fig, ax = plt.subplots(figsize=(15, 10))

```

```

plot = sns.barplot(data=featureScores, x='Score', y='Feature', palette='coolwarm',
linewidth=0.5, saturation=2,
orient='h')

```

```

plt.title('Feature Importance',
         fontsize=24,
         fontweight='bold')
# Finding the least correlated features with fetal health
low_score_features = featureScores[featureScores['Score'] < 15]
print(low_score_features)
# Dropping the features that are below the threshold
features_to_drop = ['CR_SE', 'Uric', 'Tpro', 'CHOL']
data = df.drop(features_to_drop, axis=1)
Feature Scaling

```

Feature Scaling is a technique for standardizing individual features in data into a fixed range. It is performed to handle highly variable quantities or values or units.

If feature scaling is not done, a machine learning algorithm will tend to weigh larger values as higher values and treat smaller values as lower values, regardless of the unit of the values.

```

# Ranges of the features
plt.figure(figsize=(20, 10))
sns.boxenplot(data=df, palette='coolwarm')

plt.title('Ranges of the Features',
         fontsize=24,
         fontweight='bold')
plt.xticks(rotation=90, fontsize=8)
plt.yticks(fontsize=12)
# Splitting data into features (X) and outcomes (y)
X = data.drop(["Label"], axis=1)
y = data["Label"]
col_names = list(X.columns)
s_scaler = preprocessing.StandardScaler()

```

```

X_df = s_scaler.fit_transform(X)
X_df = pd.DataFrame(X_df, columns=col_names)
# Description of the standardized features
X_df.describe().T
# Plot the standardized features
plt.figure(figsize=(20, 10))
sns.boxenplot(data=X_df, palette='coolwarm')

```

```

plt.title('Ranges of the Standardized Features',
         fontsize=24,
         fontweight='bold')
plt.xticks(rotation=90, fontsize=8)
plt.yticks(fontsize=12)
# Split the dataset into training and test set

```

```

X_train, X_test, y_train, y_test = train_test_split(X_df, y, test_size=0.2,
random_state=42)
X_train.shape, X_test.shape, y_train.shape, y_test.shape
# K-Nearest-Neighbors Classification
knn = KNeighborsClassifier()
# Parameter tuning with GridSearchCV
param_knn = {
    'n_neighbors': list(range(1, 31)),
    'p': (1, 2),
    'weights': ('uniform', 'distance'),
    'metric': ('minkowski', 'chebyshev')}
# Fitting the model for GridSearchCV
grid_search_knn = GridSearchCV(knn, param_knn, cv=10, n_jobs=-1, scoring='f1_weighted')
grid_search_knn.fit(X_train, y_train)
# Best parameters for KNN
grid_search_knn.best_params_
# Fitting KNN with the best parameters
knn = KNeighborsClassifier(n_neighbors=4, metric='minkowski', p=1, weights='distance')
knn.fit(X_train, y_train)
# Prediction and accuracy
y_pred = knn.predict(X_test)
accuracy_score(y_test, y_pred) * 100
# Show 20 predicted and actual values - KNN
pd.DataFrame({'Prediction': y_pred[:20], 'Actual': y_test[:20]})
# KNN Classification report (precision, recall, f1-score, support, accuracy)
print(classification_report(y_test, y_pred))
# Confusion matrix for KNN
ax = plt.subplot()
sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, ax=ax, cmap='coolwarm')

ax.set_xlabel('Predicted Values',
               fontweight='bold')
ax.set_ylabel('Actual Values',
               fontweight='bold')
ax.set_title('Confusion Matrix for KNN',
              fontsize=24,
              fontweight='bold')
ax.xaxis.set_ticklabels(['Normal', 'Suspect'])
logreg = LogisticRegression()

logreg.fit(X_train, y_train)

print("Train accuracy : ", accuracy_score(y_train, logreg.predict(X_train)))
print("Test accuracy : ", accuracy_score(y_test, logreg.predict(X_test)))
# Hyperparameters
param_grid = {'C': [0.001, 0.01, 0.1, 1, 10, 100, 1000],
               'penalty': ['l1', 'l2', 'elasticnet', 'none'],
               'fit_intercept': [True, False],
               'max_iter': [100, 200, 300],
               'class_weight': [{0: 1, 1: 1}, {0: 1, 1: 2}, {0: 1, 1: 4}, {0: 1, 1: 5}]
              }

```

```

# Random search for best hyperparameters
search = RandomizedSearchCV(LogisticRegression(random_state=1),
                             param_grid,
                             scoring='f1_weighted',
                             cv=3,
                             verbose=1,
                             random_state=1,
                             n_iter=30)

search.fit(X_train, y_train)

# Best parameters for Logistic regression classifier
search.best_params_
# Retrain with best model

best_logreg = search.best_estimator_

best_logreg.fit(X_train, y_train)
print("Best train accuracy : ", accuracy_score(y_train, best_logreg.predict(X_train)))
print("Best test accuracy : ", accuracy_score(y_test, best_logreg.predict(X_test)))
print(classification_report(y_test, best_logreg.predict(X_test)))
# Confusion Matrix for Best Model
best_logreg_pred = best_logreg.predict(X_test)
confusion_matrix_lr = confusion_matrix(y_test, best_logreg_pred)

ax = plt.subplot()
sns.heatmap(confusion_matrix_lr, annot=True, ax=ax, cmap='coolwarm')

ax.set_xlabel('Predicted Values', fontweight='bold')
ax.set_ylabel('Actual Values', fontweight='bold')
ax.set_title('Confusion Matrix for Best Logistic Regression Model', fontsize=24,
fontweight='bold')
ax.xaxis.set_ticklabels(['Normal', 'Suspect'])
ax.yaxis.set_ticklabels(['Normal', 'Suspect'])

# Add a colorbar
plt.colorbar(ax=ax)
plt.show()
svc = SVC(random_state=1)

svc.fit(X_train, y_train)

print("Train accuracy : ", accuracy_score(y_train, svc.predict(X_train)))
print("Test accuracy : ", accuracy_score(y_test, svc.predict(X_test)))
# Hyperparameters
param_grid = {'C': [0.1, 1, 10, 100, 1000],
              'gamma': [1, 0.1, 0.01, 0.001, 0.0001],
              'kernel': ['linear', 'rbf', 'poly'],
              'degree': [0, 1, 2, 3, 4, 5, 6]
              }

```

```

# Random search for best hyperparameters
search = RandomizedSearchCV(SVC(random_state=1),
                             param_grid,
                             scoring='f1_weighted',
                             cv=3,
                             verbose=1,
                             random_state=1,
                             n_iter=30)

search.fit(X_train, y_train)

# Best parameters for Support vector classifier
search.best_params_
# Retrain with best model

best_svc = search.best_estimator_

best_svc.fit(X_train, y_train)
print(classification_report(y_test, best_svc.predict(X_test)))
# Confusion Matrix for Best Model
best_svc_pred = best_svc.predict(X_test)
confusion_matrix_svc = confusion_matrix(y_test, best_svc_pred)

ax = plt.subplot()
sns.heatmap(confusion_matrix_svc, annot=True, ax=ax, cmap='coolwarm')

ax.set_xlabel('Predicted Values', fontweight='bold')
ax.set_ylabel('Actual Values', fontweight='bold')
ax.set_title('Confusion Matrix for Best SVM Model', fontsize=24, fontweight='bold')
ax.xaxis.set_ticklabels(['Normal', 'Suspect'])
ax.yaxis.set_ticklabels(['Normal', 'Suspect'])

# Add a colorbar
plt.colorbar(ax=ax)
plt.show()

print(classification_report(y_test, best_svc.predict(X_test)))
import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve, auc

# Assuming you have fit the KNN, Logistic Regression, and SVC models (knn, best_logreg,
best_svc)
# and have the true labels (y_test) for the test set

# KNN ROC AUC
knn_fpr, knn_tpr, _ = roc_curve(y_test, knn.predict(X_test))
knn_auc = auc(knn_fpr, knn_tpr)

# Logistic Regression ROC AUC

```

```

logreg_fpr, logreg_tpr, _ = roc_curve(y_test, best_logreg.predict(X_test))
logreg_auc = auc(logreg_fpr, logreg_tpr)

# SVC ROC AUC
svc_fpr, svc_tpr, _ = roc_curve(y_test, best_svc.predict(X_test))
svc_auc = auc(svc_fpr, svc_tpr)

# Plot ROC curves
plt.figure(figsize=(8, 6))
plt.plot(knn_fpr, knn_tpr, label='KNN (AUC = %0.4f)' % knn_auc)
plt.plot(logreg_fpr, logreg_tpr, label='Logistic Regression (AUC = %0.4f)' %
logreg_auc)
plt.plot(svc_fpr, svc_tpr, label='SVC (AUC = %0.4f)' % svc_auc)
plt.title('ROC Curve Comparison')
plt.xlabel('False Positive Rate (FPR)')
plt.ylabel('True Positive Rate (TPR)')
plt.legend(loc='lower right')
plt.grid(True)
plt.xlim(0, 1)
plt.ylim(0, 1.05)
plt.show()

```

Appendix B

Source Code

```

# We import the libraries
import numpy as np
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import LeaveOneOut
from sklearn.metrics import classification_report

from google.colab import drive
drive.mount('/content/drive')
# Load your CSV data into a DataFrame, replace 'your_data.csv' with your actual file
path
df = pd.read_csv('/content/drive/MyDrive/Capstone/output_dataset2.csv')
# Splitting data into features (X) and outcomes (y)
X = df.drop(["Label"], axis=1)
y = df["Label"]
from sklearn import preprocessing
col_names = list(X.columns)
s_scaler = preprocessing.StandardScaler()

X_df = s_scaler.fit_transform(X)
X_df = pd.DataFrame(X_df, columns=col_names)
# Description of the standardized features
X_df.describe().T
import matplotlib.pyplot as plt

```

```

import seaborn as sns

# Plot the standardized features
plt.figure(figsize=(20, 10))
sns.boxenplot(data=X_df, palette='coolwarm')

plt.title('Ranges of the Standardized Features',
          fontsize=24,
          fontweight='bold')
plt.xticks(rotation=90, fontsize=8)
plt.yticks(fontsize=12)
from sklearn.model_selection import train_test_split
# Split the dataset into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X_df, y, test_size=0.2,
                                                    random_state=42)
# Define the model
model = LogisticRegression()

# Implement the leave one out cross-validation
loo = LeaveOneOut()

# Train the model
oof = np.zeros((len(X_train), 2))
predictions = np.zeros((len(X_test), 2))

for i, (train_index, val_index) in enumerate(loo.split(X_train)):
    X_train_fold, X_val_fold = X_train.iloc[train_index], X_train.iloc[val_index]
    y_train_fold, y_val_fold = y_train.iloc[train_index], y_train.iloc[val_index]

    model.fit(X_train_fold, y_train_fold)

    oof[val_index] = model.predict_proba(X_val_fold)
    predictions += model.predict_proba(X_test) / loo.get_n_splits(X_train)

# Evaluate the model using the classification report
print(classification_report(y_train, np.argmax(oof, axis=1)))

```