2022-07-01

Analysis Of Factors Affecting Maternal Health Using Data Mining Techniques

Prajina Edayath
University of Texas at El Paso

ANALYSIS OF FACTORS AFFECTING MATERNAL HEALTH USING DATA MINING TECHNIQUES

PRAJINA EDAYATH

Master's Program in Industrial Engineering

APPROVED:

Sreenath Chalil Madathil, Ph.D.

Amit J Lopes, Ph.D.

Palvi Aggarwal, Ph.D.

Sergio Alberto Luna Fong, Ph.D.

Stephen L. Crites, Jr., Ph.D.
Dean of the Graduate School

ANALYSIS OF FACTORS AFFECTING MATERNAL HEALTH USING DATA MINING TECHNIQUES

by

PRAJINA EDAYATH, B. Tech

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Industrial, Manufacturing, and Systems Engineering

THE UNIVERSITY OF TEXAS AT EL PASO

August 2022

ACKNOWLEDGMENTS

I would like to express my sincere thanks to my adviser Dr. Sreenath Chalil Madathil for the help and support throughout my thesis work. I extend special thanks to Dr. Alfred Myrtede and Anindita Nath who helped me to complete the thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

According to the Center for Disease Control's (CDC) report on maternal health [1], [2], the United States has the highest maternal mortality and morbidity rates among the developed countries. Recent Pregnancy Mortality Surveillance System (PMSS) reports that the maternal mortality ratio in the US is 17.3 deaths per 100,000 live births. Figure 1 shows the trends in maternal mortality from 1987 to 2017 [1]. Figure 2 shows the trends in Severe Maternal Morbidity from blood transfusion during the time period 1993 to 2014 [2]. Moreover, the US ranks 56th in the maternal death world ranking by the World Health Organization (WHO). This study intended to figure out the major features that adversely affect maternal health.

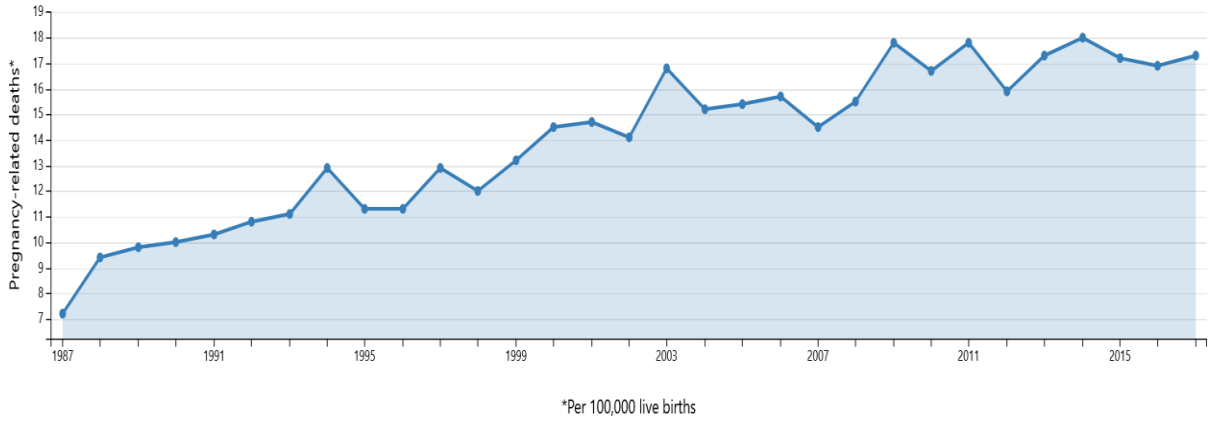

Figure 1 Trends in pregnancy-related deaths

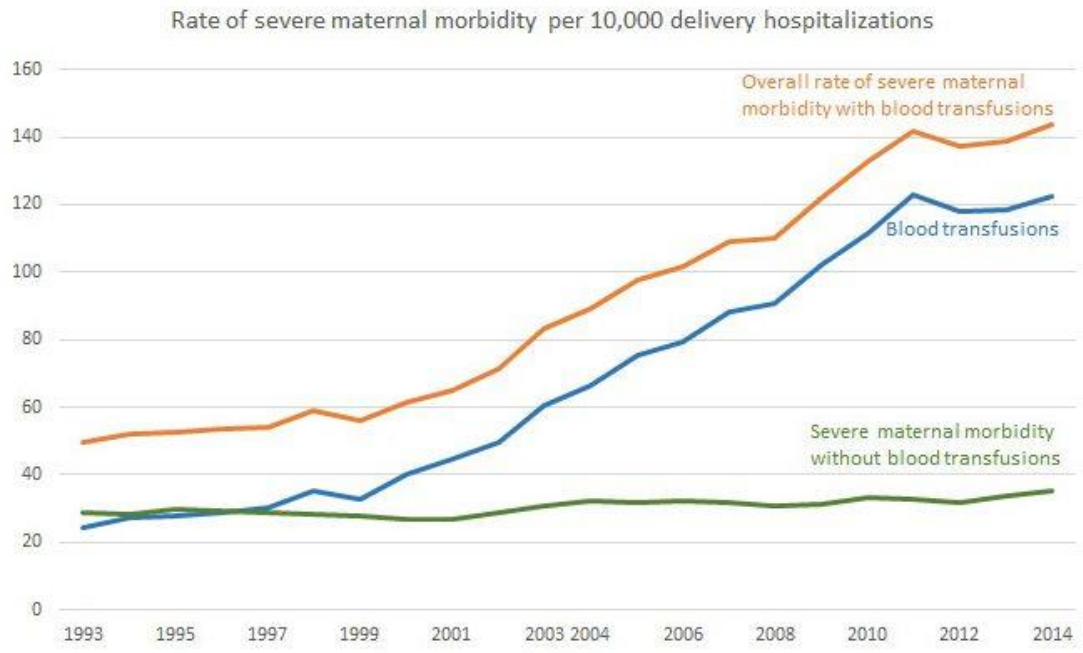

Figure 2 Trends in Severe maternal morbidity

Maternal health's duration includes the time during pregnancy, labor and childbirth, and until the end of the postpartum period, i.e., 1 year after giving birth [3]. All the health concerns in this period can be considered maternal morbidity. CDC defined maternal mortality as the death of a woman while pregnant or one year after the termination of pregnancy. Maternal morbidity or Severe Maternal Morbidity (SMM) is defined as, unexpected outcomes of labor and delivery that result in significant short or long-term consequences on a woman's health [2]. Since 1999, maternal morbidity has been identified from the hospital discharge data and the International Classification of Disease, Tenth Revision, Clinical Modification (ICD10CM) code. Before 1999, the ICD9CM code was used to identify maternal morbidity. The WHO publishes the ICD9CM and the ICD10 codes to identify the diseases and health problems.

Data mining concepts have been used in healthcare and other fields to identify the hidden trends and patterns in complex data since the middle of the 1990s **[4]**. It has the power to analyze data and contribute to decision-making. This study aims to develop and compare various prediction

models using data mining techniques and determine the best-performing model. Furthermore, this research aims to find out the significant features that have a negative impact on maternal outcomes.

Research questions

Data mining applications in healthcare include improving healthcare processes, reducing overall costs, and enhancing efficiency. Following are the two main questions addressed by this research.

- How to identify significant clinical, and sociodemographic factors that impact maternal health in the US?
- How to identify efficient data analytic techniques to predict adverse maternal outcomes.

Research significance

Reports [1][2] show that the maternal healthcare system needs a lot of improvement, since the rate of maternal mortality and SMM shows an increasing trend. According to WHO, timely intervention by healthcare providers can prevent the majority of maternal mortality cases [3]. Moreover, several data mining techniques have been used in healthcare systems for the past years to analyze the data and predict the outcome, identify the risk factors, and help in decision-making. In the maternal healthcare system, such advanced techniques can help providers to figure out the factors causing undesired maternal outcomes. These providers can determine appropriate care pathways to a patient who has conditions that can lead to adverse outcomes.

This study uses three data mining classifiers, viz., logistic regression, Naïve Bayes, and random forest, to determine significant factors that contribute to abnormal pregnancy based on the data from AllofUs followed by a discussion on the performance of these techniques to classify the data.

Research uniqueness

The data used for this research is from National Institute of Health's (NIH) AllofUs database which contains anonymous patient records from the US. It has the patient's data from different race and ethnic backgrounds, different socio-economic and sociodemographic backgrounds, different clinical conditions, and different age groups.

Unlike relying on hospital data, data from NIH's All of Us research workbench is used for this research. This data contains patient records from all over the US. The hospital data has limitations such as the kind of population or health conditions. Usage of All of Us provides data that includes the different populations with different conditions from different regions of the country.

For this research, the factors considered for predicting maternal outcome, visit type, procedure, and the lab are unique considering the other research mainly focused on demographic and socio-economic factors.

CHAPTER 2: LITERATURE REVIEW

The first part of this chapter summarizes various literature that discuss factors impacting maternal health such as sociodemographic, socioeconomic, and clinical factors. The latter part of this chapter summarizes several data analytical techniques used in this research.

Maternal Health

Pregnancy-related death is any death occurring within one year of pregnancy due to a pregnancy-related complication, a sequence of events triggered by the pregnancy, or detrimental physiological effects of pregnancy on an unrelated condition [5][6]. Maternal mortality specifically describes deaths occurring within 42 days (about 1 and a half months) of pregnancy [6][7]. Maternal mortality has doubled in the past 30 years with estimates as high as 26.4 deaths per 100.000 live births in 2015 [8]–[11] . The leading causes of maternal mortality include cardiovascular conditions, infections (including sepsis), hemorrhage, cardiomyopathy, hypertensive disorders of pregnancy, and thrombotic pulmonary and other embolisms; each condition contributes to 9 -15 percent of deaths [9] [12]. For every maternal death, there are 100 cases of severe maternal morbidity (SMM), such as peripartum hysterectomy, hemorrhage, pulmonary embolism, and septic shock, affecting an additional 100,000 women [13][14]. Maternal mortality and SMM cost billions of dollars each year, with preeclampsia alone costing above $1 billion (about $3 per person in the US) to treat annually [14]. Maternal mortality is also associated with increased infant and child mortality, loss of income, reduced social mobility, and cycles of poverty that impact both families and society [14][15].

Demographic factors on MM and SMM: Women of color and women of low socioeconomic status are more likely to suffer mortality and SMM. Black women are 3 to 4 times more likely to experience maternal mortality than white women; representing the largest disparity among all the

conventional population perinatal health measures [13]. Pregnancy-related deaths are elevated among Native Americans/Native Alaskans, Asians/Pacific Islanders, and for certain subgroups of Latina women, including Puerto Ricans, as well [16]–[18]. Women of color also experience higher rates of SMM, an unexpected and potentially fatal outcome of labor and delivery that result in significant short or long-term consequences on a woman's health, and higher case fatality rates, even without higher prevalence [7], [10], [19]–[21].

System-level factors on quality of care: An estimated 45 – 60 percent of all maternal deaths, SMM, and near-misses are preventable with timely and appropriate care [21]–[24]. Mortality and SMM from several specific conditions, such as hemorrhage and preeclampsia, have much higher rates of preventability; in some cases, as high as 93 percent [13], [20], [21]. These findings suggest opportunities to intervene at the point of care to reduce mortality and SMM and improve overall maternal care [13][22]. However, there is little research specifically examining quality of care issues such as inadequate teamwork, delays, and poor coordination that contribute to these adverse outcomes and disparities, particularly in the context of the national effort to help improve maternal care [22][24].

Thirty-five states in. the U.S. and major cities such as Baltimore, New York City, Philadelphia, and Washington D.C. have initiated or currently creating maternal mortality review committees (MMRC) to review pregnancy-related deaths, assess their preventability, and develop recommendations to improve the health of pregnant and postpartum patients [24]. These MMRC can apply an equity lens to these investigations and examine SMM in a limited capacity, as they are much more prevalent. Several cases of SMM are caused by transfusion of four or more units of blood and ICU admission [24]. Prior research and commentary implore the need for local and systems-level assessments, the reports from the nine MMRCs noted that while surveillance (such

as vital statistics) highlights trends and disparities, there exists a need for smaller scale efforts to assess the preventability and causes of deaths and identify opportunities to improve care. Social determinants of health research have identified specific influences on pregnant women's health, including access to and utilization of prenatal care, reliable transportation, and healthy food [24][25]. However, the high percentage (40 – 60 percent) of preventable deaths, SMM cases, and near-misses suggests opportunities for improvement at the point of care [22][23]. While patient factors, including prevalence of co-morbidities and substance abuse, and provider factors, including failure to identify progressing severity and unnecessary use of medical interventions, contribute to maternal death and SMM, systems factors, including lack of available ICU beds, delayed diagnosis, inadequate record of blood loss, and poor coordination among clinical services, also contribute to adverse maternal outcomes. These systems factors were cited as a cause in approximately 25 percent of the preventable deaths [23]–[26]. However, 25 percent represents a conservative estimate as the relationship between the provider and system of care is interrelated and bad systems, including poorly designed and integrated electronic health records (EHRs), negatively impact clinicians' ability to provide effective care (REF). An in-depth examination of these systems factors, the structures and processes of care within the clinical system, are critical for understanding sociotechnical challenges in maternal care, their contributions to healthcare disparities[27], and improving the quality of care [23][24].

Data mining

Data mining is the process of classifying and identifying patterns and trends by sorting large data sets for predictions and to get an insight into the data. It is a concept that started in the middle of the 1990s [4]. It is helpful to reduce costs, increase revenue and increase the efficiency of the

healthcare system. Its application in healthcare includes the identification and classification of the high-risk population (Race or Ethnicity), procedures, visit type, etc. [28].

Cavazoz-Rhg et al. studied the relationship between maternal age and the risk of maternal morbidity and delivery complications as part of the US Healthcare Cost and Utilization Project. They used hospital billing information from the United States Nationwide Inpatient Sample (NIS) to collect data and classified maternal conditions using ICD-9-CM codes. They applied logistic regression on the data and found that younger age (11-18) has a higher risk to develop complications compared to the age group 25-29. Furthermore, the age group >35 is more likely to have conditions such as preterm delivery, hypertension, and preeclampsia. [29]. Frolich et al. in their study used logistic regression to find out the relation between the distance between a patient's residence and hospital. They found that longer residential distance from the hospital was a significant factor in maternal death [30]. Leonard et al. studied the relation between maternal characteristics and cesarean section delivery that leads to Severe Maternal Morbidity (SMM). They conducted a multivariate logistic regression analysis to study the impact of characteristics such as advanced maternal age, pre-pregnancy-obesity, pre-pregnancy-comorbidity, and cesarean delivery with SMM. ICD-9-CM codes were used to identify SMM. They found advanced maternal age and pre-pregnancy obesity and some procedures as the predictors of SMM [31]. Azimi et al. used six data mining classifiers to predict the factors impacting infections after surgeries [32].

Table 1 summarizes literature reviews, editorial/reports reviewed for this study. The papers published after 1990 are considered for review.

Table 1: Literature discussing maternal healthcare

Citation	**Type of Review**	**Duration**	**Number of papers reviewed**	**Objective**	**Outcome**
[24]	Literature review	1982-1993	72	The role of nurses in preventive measures to reduce ectopic pregnancy is discussed.	The nurses along with other healthcare workers can contribute in preventive measures to reduce ectopic pregnancy
[22]	Literature review	1977-1992	97	Examine the evidence behind the assumptions behind cesarean delivery	Cesarean delivery is highly beneficial to both infants and mothers
[4]	Literature review	1963-2010	110	Reviewing different data mining techniques used in healthcare	Points out the advantages and problems of datamining in healthcare
[12]	Literature review	1990-2013	29	Reviewing the global and the programs within the US on Maternal mortality and morbidity	Women's overall health, nutrition, access to care and socioeconomic issues should be taken care of to address maternal mortality and morbidity
[10]	Literature review	2005-2014	86	Review the evidence on maternal health disparities and the impact of Affordable Care Act on these disparities.	The Affordable Care Act could reduce the disparities, but it cannot prevent disparities.
[28]	Survey	1996-2015	35	Gather the data to show the importance of datamining in healthcare	Figured out the advantages and challenges of datamining in healthcare
[9]	Literature review	2003-2018	33	Identify the actions taken to improve maternal outcomes in California	Identified and discussing the four-step model helped to improve maternal outcomes in California

[27]	Report/Editorial			Address maternal healthcare and health plans	
[7]	Literature review	1980-2017	46	Review studies and reports on maternal mortality and morbidity in the USA	Improvement of quality of care of hospitals and good coordination improved maternal outcomes and reduced disparities
[13]	Literature review	1999-2018	39	Reduce maternal mortality and morbidity using patient safety tools.	A standardized approach to the conditions causing maternal mortality and morbidity will increase the positive outcome. Addressing racial disparities can reduce adverse outcomes.
[5]	Literature review	1997-2019	54	Analyze ERAS (Enhanced Recovery After Surgery) as a tool to reduce maternal mortality and morbidity	ERAS is capable of reducing the negative maternal outcome after cesarean
[23]	Literature review	2000-2018	55	Analyze the differences in risk factors associated with maternal mortality in racial and geographic populations	Proposed literature and theory-based framework to address social determinants of maternal health
[25]	Proposal	2000-2018	46	Address disparity in maternal mortality and identify ways to reduce it	Suggested eight steps to improve quality of care and reduce disparity

Table 2 shows the list of papers that used statistical methods for data analysis. The type of data used are demographic, laboratory data (Lab), procedure, visit type, and conditions. The demographic data includes factors such as age, race/ethnicity, income, gender, income, and education. The laboratory data consists of different lab works such as iogonadotropin.beta subunit (pregnancy test) [Presence] in Urine, Choriogonadotropin [Mass/volume] in Serum or Plasma, Choriogonadotropin (pregnancy test) [Presence] in Urine, Choriogonadotropin.beta subunit [Units/volume] in Serum or Plasma, Choriogonadotropin [Units/volume] in Serum or Plasma, Choriogonadotropin (pregnancy test) [Presence] in Serum or Plasma. Several types of hospital visits during the maternity period are included in the visit type data. The hospital visits such as inpatient visits, outpatient visits, emergency room visits, and laboratory visits are listed in visit type data. Procedure data are the clinical procedures done during pregnancy and the postpartum period. Excision of the fallopian tube and surgical removal of ectopic pregnancy, repair of current obstetric laceration of rectum and sphincter ani, aspiration curettage of the uterus for termination of pregnancy, repair of obstetric laceration of bladder and urethra, pregnancy detection examination are some of the procedures included in procedure data. The conditions occurring during pregnancy and the postpartum period are considered in the condition data. Different types of conditions include the data such as pregnancy, post-term pregnancy, the disorder of pregnancy, and delivery normal.

Table 2: Literature discussing statistical analysis in healthcare

	Type of Data						Methodology	Objective	Outcomes
Citation	**Demographics**	**Lab**	**Procedure**	**Visit type**	**Conditions**	**Number of patients**			
[14]	✓				✓	139	Bivariate Analysis	Review of California-Pregnancy Associated Mortality Review	CA-PAMR found additional maternal deaths and more accuracy in case findings. It could contribute to quality improvement.
[15]					✓	64330	Logistic correlation analysis	Factors causing c-section delivery	
[6]	✓				✓	7025	Chi-square and Fisher exact test	Figure out the relation between hypertensive disorder and SMM	There is a strong relation between hypertensive disorder and SMM
[16]	✓				✓	122	Chi-square test	Determining the causes of spontaneous abortion	Age, race and marital status are found aa the common reasons
[18]	✓				✓	588,232	Logistic regression	Determine the effect of socioeconomic factors on ethnicity causing SMM	Found significant relation between Race/ethnicity and SMM
[17]	✓				✓	1030350	Multivariate analysis	Examining racial disparities in maternal outcomes among four ethnic groups	African American are found to have worst maternal outcome considering

									socioeconomic, sociodemographic, and clinical factors
[19]	✓					76912	Condition-specific and Multivariate analysis	Examining the factors causing SMM and black and white disparity in maternal morbidity	Found racial/ethnic disparity in maternal morbidity
[20]	✓					62588	Multivariate analysis	Analyzing the patterns of care and management of ectopic pregnancy find if there is disparity	Found racial and economic disparities in the management of ectopic pregnancy
[30]	✓				✓	296	Univariate and Multivariate analysis	Analyze the maternal mortality trends in 10 years between 189 and 1998	Racial disparity is a common trend in New York city
[29]	✓				✓	7810762	Logistic Regression	Determining the relationship between maternal age and Maternal morbidity	Found significant correlation between maternal age and Labor and delivery complications
[30]	✓				✓	835	Chi-square and Fisher exact test and Logistic Regression	Analysis of maternal mortality at the University of Alabama from 1990 to 2010	Found a relationship between residential and hospital location
[31]	✓					3,556,206	Multivariate Logistic Regression	To find the relationship between cesarean delivery and maternal characteristics	Found that maternal age and obesity can result in cesarean delivery
[32]					✓	208	Logistic Regression, Naïve Bayes,	Develop models to predict post-operative infections	Predicted the features that can cause infections after surgery

							Decision Tree, ANN, SVM		
This research	✓	✓	✓	✓	✓	✓	Logistic Regression, Naïve Bayes, Random Forest	Develop models to predict factors affecting maternal health	Predicted factors affecting maternal health

CHAPTER 3: METHODOLOGY

Data mining is a powerful technique that has been using in healthcare field to analyze the data [4]. It can explore the data for predictive analysis (supervised learning) and for descriptive analysis (unsupervised learning). In predictive data mining, the data records are classified into different groups according to the target or the dependent attribute. Data mining classifiers predict the target attribute for each independent attribute by using the groups already created. Descriptive analysis clusters the data to find the similarities or relationships and reveals the hidden patterns in the data. The aim of this study is to analyze the data and find the most crucial factors causing the abnormal maternal outcome. This research uses predictive analysis or supervised learning techniques such as Logistic Regression, Naïve Bayes, and Random Forest for the classification of the data and outcome prediction. The dependent target variable is maternal outcome ('TARGET'), which has two categories, 'Normal' and 'Abnormal'. The independent variables are Age, Race-Ethnicity, Procedure, Lab, Visit type, and Conditions. The data mining techniques used to predict the adverse maternal outcome are Logistic Regression, Naïve Bayes, and Random Forest. Figure 3 shows the steps used to analyze the data and build a prediction model.

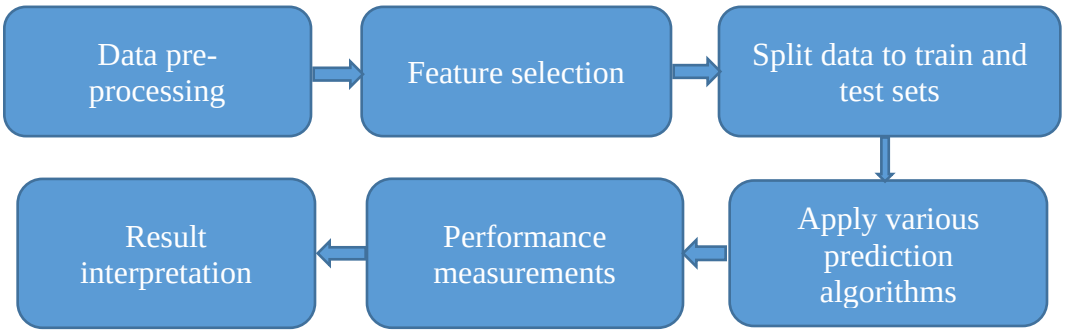

Figure 3: Steps for data analysis

Python 3.0 is used for the whole analysis and the packages used are sklearn and statsmodels.

Data

NIH's All of US researcher workbench has patient data from the United States. They utilize surveys, Electronic Health Records (EHR), bio samples, physical measurements, and data from wearables like Fitbit as the data sources. This research used the data from the NIH All of Us which contains anonymous patient records from the US. The cohort included in this research has characteristics such as 'female', 'pregnant', and 'gestation'. The cohort has patients with condition termed as "only pregnant". The second cohort has patients with conditions related to Abnormal Pregnancy. The cohort subset for normal pregnancy contains all patients in "Only pregnant" cohort after removing patients conditions related to "abnormal pregnancy". AllOfUs Data allows the use of Python and R using Jupyter notebook and this research used Python for data analysis.

The concept set used to build the cohort is given in Table 3.

Table 3: Concept set

Demographics	Gender identity-Female
	Race: Black or African American
	Race: Asian
	Race: White
	Ethnicity: Hispanic or Latino
	Ethnicity: Not Hispanic or Latino
Conditions	Post term pregnancy OR pregnant
	Antepartum Condition or complication
	Post term pregnancy, delivered, with or without mention of antepartum condition
	Parent complication of pregnancy
	Childbirth and/or the puerperium
	Parent complication occurring during pregnancy

The cohort data can include complete records for the patients. However, this research require data during the maternal period only. Hence, a sorted patient data helped to create the final dataset that include Date of birth (DOB), Person ID (PID), Race, Ethnicity, and Gender. The

dataset created during the maternal period based on the information of gestation start, condition start, and pregnancy period include records on patients' conditions, labs, procedures, visit types, and drug information.

One concept set in AllOfUs contains survey questions related to health insurance, tobacco and alcohol use, disability, and history of complications of pregnancy. The next step is to add these observation values related to pregnancy, maternity, and delivery from this survey concept set. The concept id used to obtain abnormal conditions are post-term pregnancy, antepartum condition, or complication, complication of pregnancy, childbirth, and/or the puerperium, and complications occurring during pregnancy.

The same steps are repeated for the cohort 'Only pregnant' to get the data for Age, Race-Ethnicity, Lab, Visit type, Procedure, and Other conditions.

The 'Normal Pregnancy' cohort is built by deleting the 'Abnormal Pregnancy' data from the cohort 'Only Pregnant'. The data used for classification and prediction of adverse maternal outcomes are generated by concatenating the Normal Pregnancy and Abnormal Pregnancy cohorts.

The data used in this study is highly imbalanced, i.e., there are 18691unique patient records with 18683 abnormal outcome and 614 normal outcomes. Since the data is imbalanced heavily to the abnormal outcome, such data may not provide a good prediction model.

Dummy variables

The variables race-ethnicity, procedure, lab types, visit type, and conditions are categorical variables. Thus, dummy variables are created for all unique categories. These are artificial variables that take the values 0 or 1, in which 0 indicates the sample (patient) does not belong to that category and 1 means it belongs to that category. For example, for the factor Race/Ethnicity

there are eight categories (White-Hispanic or Latino, Black or African American-Not Hispanic or Latino, White-Not Hispanic or Latino, White-Not Hispanic or Latino, Asian-Hispanic or Latino, Asian-Not Hispanic or Latino, Black or African American-Hispanic or Latino, Another Single Population-Hispanic or Latino). Each of these categories takes values 0 or 1. The race/ethnicity to which the patient belongs will take the variable 1 while all the others will have the value 0. In the dataset, age is a continuous variable. After creating the dummy using the method 'get_dummies' variables there are 256 features.

Feature selection

Feature selection is a technique used to select the most prominent features to create a better predictive model. It is important when dealing with a dataset that contains several features. These types of high-dimensional datasets can end up with a few problems while fitting the model such as longer training time for the model, and occasional overfitting of the model.

Three classes of feature selection methods are Filter Methods, Wrapper Methods, and Embedded Methods [32]. The filter method is based on the uniqueness of evaluating data and selecting the subset. It uses the exact assessment criterion that includes the exact information. Here, the variable selection is based on the ranking technique or the scores of each feature. The ranking will exclude irrelevant features before the classification. In the wrapper method, the model evaluates the interaction between the variables and selects the best combinations that provide the best prediction. The embedded method is an iterative method and performs feature selection as part of the training process. It eliminates the irrelevant feature by adding a penalty to the objective.

Since the number of features is large for this data, dimensionality reduction is necessary to predict the exact features those will adversely affect maternal health. There are various feature

selection methods used to reduce the irrelevant features, such as Recursive Feature elimination (RFE), and Principal Component Analysis (PCA). This research uses RFE which is a wrapper type feature selection method. The RFE works by recursively eliminating the attributes and building a model on the remaining attributes. Each feature is given a rank according to its importance and eliminates the least important feature at each iteration. The model selects the most significant features once it reaches any specified stopping criteria. The RFE uses the accuracy of the model to identify the features that can contribute to predicting the model.

Feature selection models

The packages used for processing the data are 'sklearn' and 'statsmodels'. The data obtained after feature selection split into subsets training and testing data in a ratio 60:40. The function 'train_test_split' in sklearn is used to split the data.

Logistic regression

Logistic regression is a statistical analysis method for classification of the data and predicting the outcome of the event by analyzing the data. Logistic regression works best in classification problems and to estimate the probability. It is used when the target variable is categorical. Even though the output variable is binary multinomial logistic regression extend the application to scenarios where there are more than two outcomes[33].

Naïve Bayes

Naïve Bayes datamining method is a probabilistic machine learning method used to predict the probability of a feature or event that can affect the outcome. It is a classifier based on Bayes theorem Eq (1).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad Eq.(01)$$

Naïve Bayes classifier is suitable for large data set. It works on the assumption that each sample is independent and does not interact each other. The main advantage of Naïve Bayes is that it is computationally fast when dealing with high dimensional data set [34].

Random forest

Random forest is a machine learning method that constructs multiple decision trees for prediction and classification. It constructs several decision trees while training the data. These decision trees are trained parallelly on different subset on the features which is called bootstrapping method. Random Forest classifier combines the individual decision trees for the outcome prediction. Compared to the other classifiers it provides more accuracy without model overfitting [34].

Performance measurements

This research used a confusion matrix to evaluate the performance of feature selection model. When evaluating a classification model there will be four outcomes since there are two classes. For instance, if a factor can be a reason for adverse maternal outcome and it is predicted to be positive (1) it is called true positive (TP) and if it is predicted to be negative it is false positive (FP). Similarly, if a factor cannot be a reason for adverse maternal outcome and if it is predicted to be negative it is called true negative (TN) and if it is predicted to be positive it is called false negative (FN). Table 4 shows the four outcomes that construct a to-by-two matrix called a confusion matrix.

Table 4: Confusion matrix

		True Classes	
		Negative	Positive
Predicted class	Negative	TN	FN
	Positive	FP	TP

This confusion matrix helps to calculate Accuracy, Sensitivity, Specificity and Precision. Accuracy (Eq 02) indicates the percentage of data points causing adverse effects on maternal health and predicted correctly.

$$Accuracy = \frac{TN + TP}{TN + FN + FP + TP} \qquad Eq.(02)$$

Sensitivity (Eq. 03) is the proportion of adversely affecting data points that are predicted correctly.

$$Sensitivity = \frac{TP}{TP + FN} \qquad Eq.(03)$$

Specificity (Eq. 04) shows the percentage of data points that do not cause adverse outcome and are correctly predicted negative.

$$Specificity = \frac{TN}{TN + FP} \qquad Eq.(04)$$

Precision (Eq. 05) shows the percentage of data points predicted to cause an adverse outcome and causes adverse outcome.

$$Precision = \frac{TP}{TP + FP} \qquad Eq.(05)$$

Internal validation

Internal validation is an unavoidable step in the development of a prediction model. It calculates the reproducibility of the model and adjusts the model for overfitting. In this research, for cross-validation, the data was split into training and testing data in a 60:40 ratio. Each model is repeated 6 times with different random seeds ranging from 5 to 10 for the reproducibility of the resulting features.

Data summary

The data contains a total number of 18697 unique patients that has 614 patients with normal outcome and 18689 abnormal outcomes. After applying the concept set to the unique patients, total number of records for these patients is 209571 with 3195 records for patients with normal outcome and 206376 records for patients with abnormal outcomes. Table 5 shows the classification of total data.

Table 5: Data Classification

	Unique Patients	Total number of records
Total	18697	209571
Normal	614	3195
Abnormal	18689	206376

For this study, the factors considered for predicting the normal and abnormal maternal outcomes are Age, Race or Ethnicity, Pregnancy conditions (Other conditions), Lab, Visit type, and Procedures.

Age while pregnant

Age is a key factor determining the overall health of pregnant women. Since the age used to build the cohort is the current age (age at the time of running the code), may not be relevant for the analysis. Hence, the updated data includes a new calculated column with the age at the time of pregnancy by subtracting the pregnancy start date and the date of birth. The histogram (Figure 4) of age distribution shows ages while pregnant ranging from 17 to 69 years.

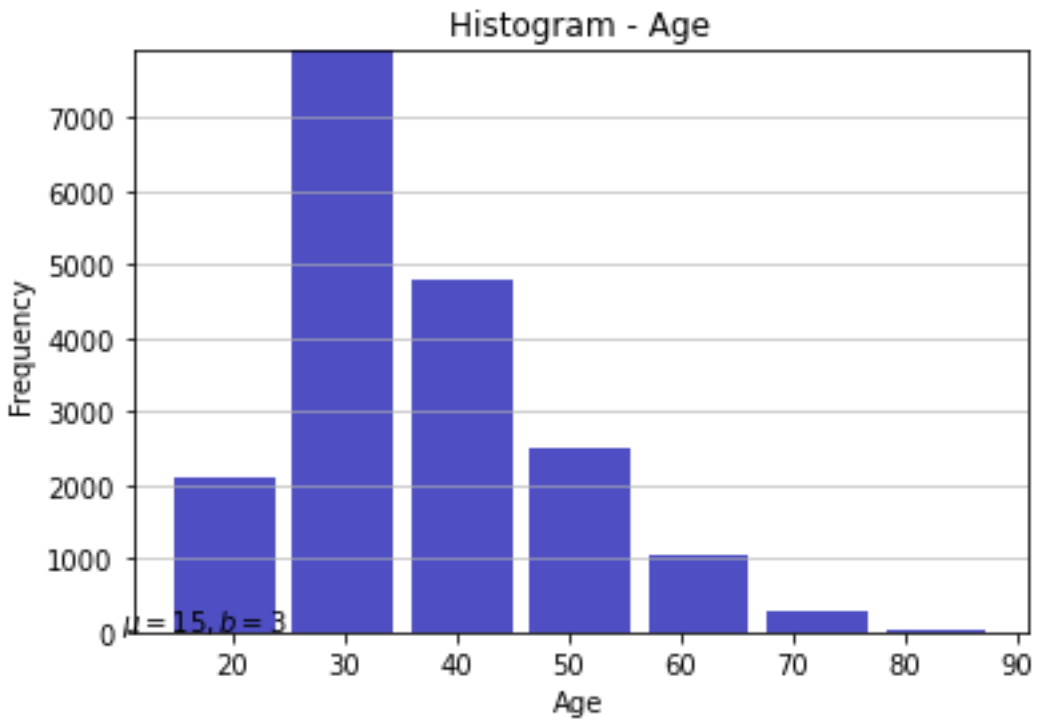

Figure 4: Histogram of age at the time of pregnancy

Race and Ethnicity

The CDC report indicates that the race and ethnicity of women play a major role in determining the maternal health outcome. It shows maternal mortality and morbidity among minority women are high compared to non-Hispanic White women. In this study, there are eight sub-categories for race and ethnicity such as White: Non-Hispanic or Latino, White: Hispanic or Latino, Black or African American: Non-Hispanic or Latino, Black or African American: Hispanic or Latino, Asian: Non-Hispanic or Latino, Asian: Hispanic or Latino, Another population: Hispanic or Latino, more than one population: Hispanic or Latino. Figure 5 shows the distribution of race and ethnicity in normal and abnormal data.

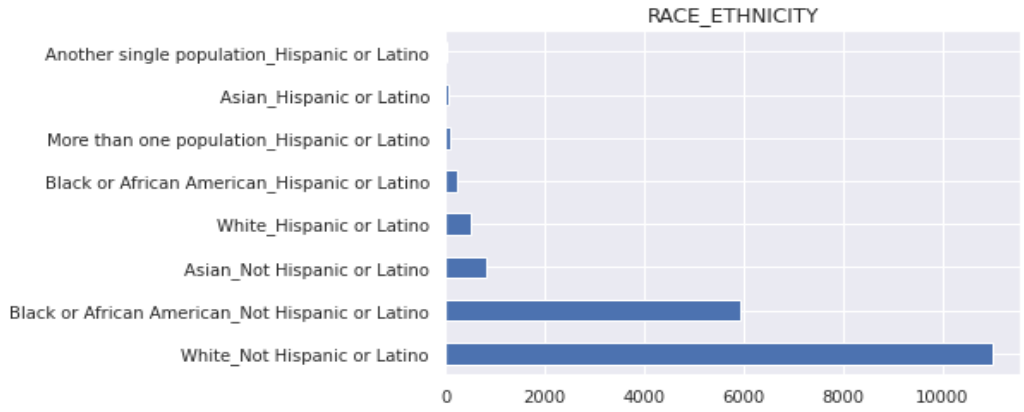

Figure 5: Race-Ethnicity distribution

Lab

Figure 6 shows the top twenty-one different lab works obtained from the data (Figure 6). The type of lab work performed can be an indication for adverse maternal health outcomes. Lab work is one of the features that is not present in any previously reviewed literature.

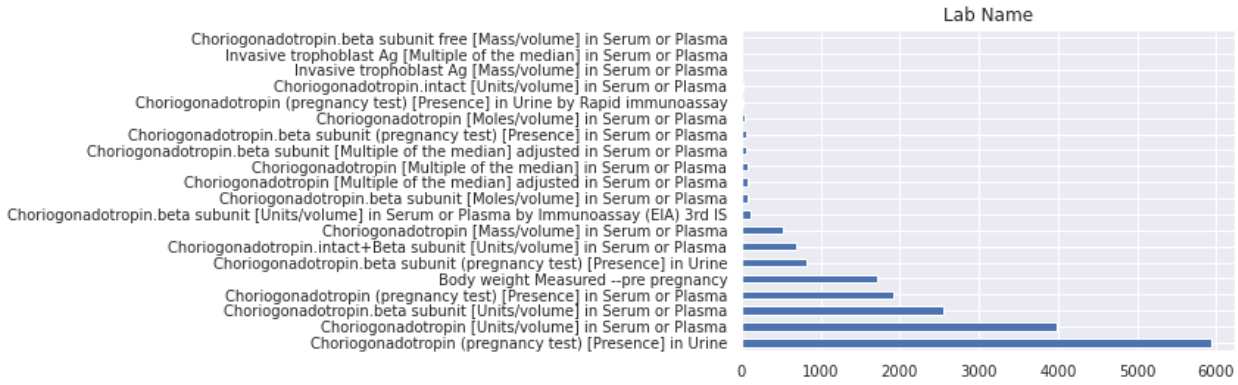

Figure 6: Lab work distribution

Visit type

A patient's visit type can determine the current condition and emergency situation of the patient. It is one of the least considered factors in previous research. Figure 7 shows the top twenty-two distinct types of hospital visits present in the data.

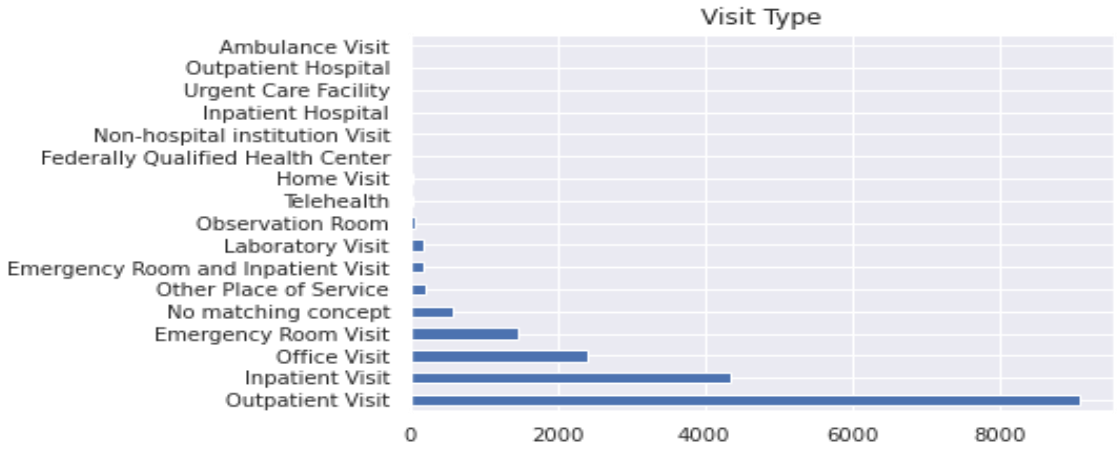

Figure 7: Visit type distribution

Pregnancy conditions

Figure 8 depcits the frequency of the top twenty pregnancy conditions present in the final data. The occurrences of any of these conditions can determine whether the woman is going to have a normal or abnormal pregnancy outcome. Identifying these condition can avoid many abnormal pregnancy outcomes.

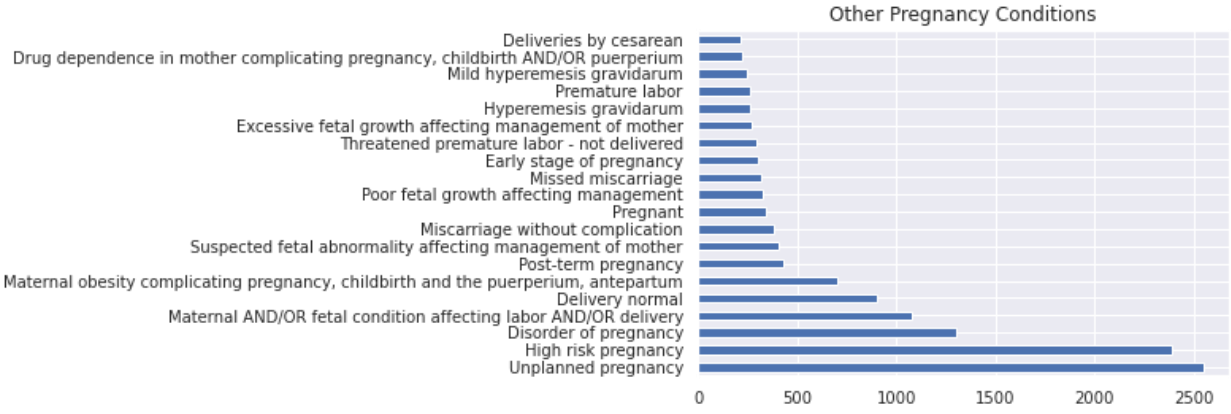

Figure 8: Pregnancy conditions distribution

Procedure

Procedure data are the clinical procedures done during pregnancy and the postpartum period. Procedures during the maternity period can reveal some of the risk factors associated with

pregnancy. It is a good predictive feature for the adverse maternal outcome. Some of the procedures included in procedure data include excision of the fallopian tube and surgical removal of ectopic pregnancy, repair of current obstetric laceration of rectum and sphincter ani, aspiration curettage of uterus for termination of pregnancy, repair of obstetric laceration of bladder and urethra, and pregnancy detection examination. Figure 9 summarizes the top 15 procedures obtained from the data.

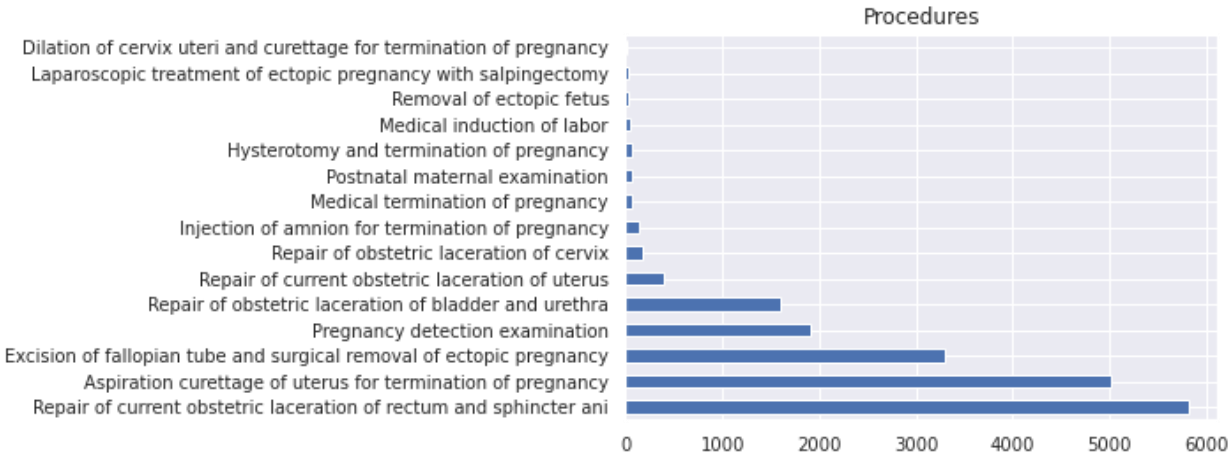

Figure 9: Procedure graph

CHAPTER 4: RESULTS

There are 474 features observed after creating the dummy variables. After applying Recursive Feature Elimination (RFE) there are 236 features found to be relevant. Since running the classifier with the whole data made the running environment (Jupyter notebook) dead, a sample count of 3195 records were used to classify the data. (The data used for this study is highly imbalanced data, 206376 records are abnormal and less than 3195 are normal. So, to run the classifiers without error, a sample count is used). The results obtained for different classifiers are explained below:

Logistic regression

In this study out of 236 features obtained after feature elimination in, 15 features are found to be statistically significant ($P<0.05$) considering 6 repetitions by changing the random seeds. These 15 features are repeating at least 3 times in every run. Table 6 shows the predicted features.

Table 6: Predicted features (Logistic regression)

Pregnancy conditions	Primigravida
	Hyperemesis gravidarum
	Pregnant
	Uterine size for dates discrepancy
	Poor fetal growth affecting management
	Urinary tract infection in pregnancy
Visit type	No matching concept
	Emergency Room and Inpatient Visit
	Other Place of Service
Lab name	Choriogonadotropin [Multiple of the median] adjusted in Serum or Plasma
	Choriogonadotropin [Mass/volume] in Serum or Plasma
	Choriogonadotropin (pregnancy test) [Presence] in Serum or Plasma
Race-Ethnicity	Black or African American Hispanic or Latino

The confusion matrix and the performance measures are shown below in Table 7 and the performance measures are shown in Table 8.

Table 7: Confusion matrix (Logistic regression)

		True class	
		Negative	Positive
Predicted Class	Negative	824	448
	Positive	519	765

Table 8: Performance measurements (Logistic regression)

Accuracy	Sensitivity	Specificity	Precision
0.621	0.647	0.595	0.60

Naïve Bayes

There are 28 features that are significant among 236 features (Feature ranking > 0.001). The Naïve Bayes classifier repeated 6 times by changing the random seeds from 5 to 10. Each of these features repeated at least 3 times in every run. Those are listed in Table 9. The attribute 'permutation_importance' in 'sklearn' is used for feature ranking of each feature. The ranks indicate how much the model depends on the feature [35].

Table 9: Predicted features (Naïve Bayes)

Pregnancy conditions	Advanced maternal age gravida
	Drug dependence in mother complicating pregnancy, childbirth AND/OR puerperium
	Elderly primigravida with antenatal problem
	Hyperemesis gravidarum
	Maternal AND/OR fetal condition affecting labor AND/OR delivery
	Maternal obesity complicating pregnancy, childbirth, and the puerperium, antepartum
	Normal birth
	Poor fetal growth affecting management
	Pregnancy with abortive outcome
	Pregnant
	Primigravida
	Suspected fetal abnormality affecting management of mother
	Uterine size for dates discrepancy
Lab name	Choriogonadotropin [Mass/volume] in Serum or Plasma
	Choriogonadotropin [Multiple of the median] adjusted in Serum or Plasma

	Choriogonadotropin (pregnancy test) [Presence] in Urine
	Choriogonadotropin [Moles/volume] in Serum or Plasma
Race-Ethnicity	Black or African American Hispanic or Latino
	White Hispanic or Latino
Visit type	Emergency Room Visit
	Inpatient Visit
	Laboratory Visit
	No matching concept
	Other Place of Service
	Outpatient Visit
	Telehealth

The confusion matrix and the performance measures for Naïve Bayes are shown in the tables below (Table 10), (Table 11). The highest accuracy obtained for one of 6 instances in Naïve Bayes is 62.08%.

Table 10: Confusion matrix (Naïve Bayes)

		True class	
		Negative	Positive
Predicted Class	Negative	815	464
	Positive	513	785

Table 11: Performance measurements (Naïve Bayes)

Accuracy	Sensitivity	Specificity	Precision
0.620	0.637	0.604	0.63

Random forest

Random forest classifier predicted 50 features as significant among the 236 features obtaind after feature elimination. These features repeated at least 3 times in every run. The attribute 'feature_importances' in sklearn is used to find the feature ranking of each feature. It measures the mean and standard deviation of accumulation of impurity decrease [35]. Table 12 shows the list of features predicted by the random forest classifier.

Table 12: Predicted features (Random forest)

Pregnancy conditions	Drug dependence in mother complicating pregnancy, childbirth AND/OR puerperium
	Pregnant
	Abnormality of organs AND/OR soft tissues of pelvis affecting pregnancy
	Excessive fetal growth affecting management of mother
	Hyperemesis gravidarum
	Urinary tract infection in pregnancy
	Poor fetal growth affecting management
	Missed miscarriage
	Delivery normal
	Fetal condition affecting obstetrical care of mother
	Threatened premature labor - not delivered
	Advanced maternal age gravida
	Premature labor
	Elderly primigravida
	Suspected fetal abnormality affecting management of mother
	Early stage of pregnancy
	Mild hyperemesis gravidarum
	Primigravida
	Maternal obesity complicating pregnancy, childbirth and the puerperium, antepartum
	Maternal AND/OR fetal condition affecting labor AND/OR delivery
	Spotting per vagina in pregnancy
	False labor before 37 completed weeks of gestation
	High risk pregnancy due to history of preterm labor
Visit type	Emergency Room Visit
	Inpatient Visit
	Other Place of Service
	Emergency Room and Inpatient Visit
	No matching concept
	Outpatient Visit
	Office Visit
Race-Ethnicity	White Hispanic or Latino
	White Not Hispanic or Latino
	Black or African American Not Hispanic or Latino
Procedure	Repair of current obstetric laceration of uterus
	Repair of obstetric laceration of bladder and urethra
	Excision of fallopian tube and surgical removal of ectopic pregnancy
	Repair of current obstetric laceration of rectum and sphincter ani
Lab name	Body weight Measured --pre pregnancy
	Choriogonadotropin (pregnancy test) [Presence] in Serum or Plasma
	Choriogonadotropin (pregnancy test) [Presence] in Urine
	Choriogonadotropin [Mass/volume] in Serum or Plasma
	Choriogonadotropin [Multiple of the median] adjusted in Serum or Plasma

	Choriogonadotropin [Units/volume] in Serum or Plasma
	Choriogonadotropin.beta subunit (pregnancy test) [Presence] in Urine
	Choriogonadotropin.beta subunit [Multiple of the median] adjusted in Serum or Plasma
	Choriogonadotropin.beta subunit [Units/volume] in Serum or Plasma by Immunoassay (EIA) 3rd IS
	Choriogonadotropin.beta subunit [Units/volume] in Serum or Plasma
	Choriogonadotropin.intact+Beta subunit [Units/volume] in Serum or Plasma

The table below (Table 13), (Table 14) shows the confusion matrix and the performance measures of Random Forest in one instance of 6 repetitions

Table 13: Confusion matrix (Random forest)

Predicted Class	True class	
	Negative	Positive
Negative	728	544
Positive	432	852

Table 14: Performance measurements (Random forest)

Accuracy	Sensitivity	Specificity	Precision
0.623	0.572	0.663	0.66

Discussion

The AllofUs data used for this study are highly imbalanced. There are two sets of data, one with abnormal results and the other with normal results. There are 18689 unique abnormal records and 614 unique normal records. To classify the data and predict the outcome, we sampled the majority class to match with the count from minority class. The three data mining techniques are applied to this All of Us data. To make the prediction more sensible, each classifier is repeated 6 times by changing the random seed from 5 to 10.

The result obtained from three classifiers indicates that the factors selected to predict the maternal outcome are relevant. Table 15 shows the list of most significant features.

All three classifiers identified the 12 common features as significant features to predict maternal outcome (Table 15).

Table 15: Significant features

Features	Logistic Regression	Naïve Bayes	Random Forest
Choriogonadotropin [Mass/volume] in Serum or Plasma	✓	✓	✓
Choriogonadotropin [Multiple of the median] adjusted in Serum or Plasma	✓	✓	✓
Primigravida	✓	✓	✓
Hyperemesis gravidarum	✓	✓	✓
Pregnant	✓	✓	✓
Uterine size for dates discrepancy	✓	✓	
White Hispanic or Latino	✓	✓	✓
No matching concept	✓	✓	✓
Choriogonadotropin (pregnancy test) [Presence] in Serum or Plasma	✓		✓
Poor fetal growth affecting management	✓	✓	✓
Black or African American Hispanic or Latino	✓	✓	
Emergency Room Visit		✓	✓
Emergency Room and Inpatient Visit	✓		✓
Other Place of Service	✓	✓	✓
Choriogonadotropin (pregnancy test) [Presence] in Urine		✓	✓
Advanced maternal age gravida		✓	✓
Drug dependence in mother complicating pregnancy, childbirth AND/OR puerperium		✓	✓
Maternal AND/OR fetal condition affecting labor AND/OR delivery		✓	✓
Maternal obesity complicating pregnancy, childbirth and the puerperium, antepartum		✓	✓
Suspected fetal abnormality affecting management of mother		✓	✓
Repair of current obstetric laceration of rectum and sphincter ani		✓	✓
Inpatient Visit		✓	✓
Laboratory Visit		✓	✓
Outpatient Visit		✓	✓

The features listed in Table 15 can be considered as the prominent features for predicting maternal mortality and morbidity. It indicates that using Logistic regression, Naïve Bayes or Random Forest can identify the factors that adversely affect the maternal outcome. Identifying

the risk factors can help the providers to determine the treatments and avoid future risks. Furthermore, it is helpful for the patients to understand their conditions to be prepared for treatment for the specific condition. Figure 10 shows the accuracy plot for the three classifiers.

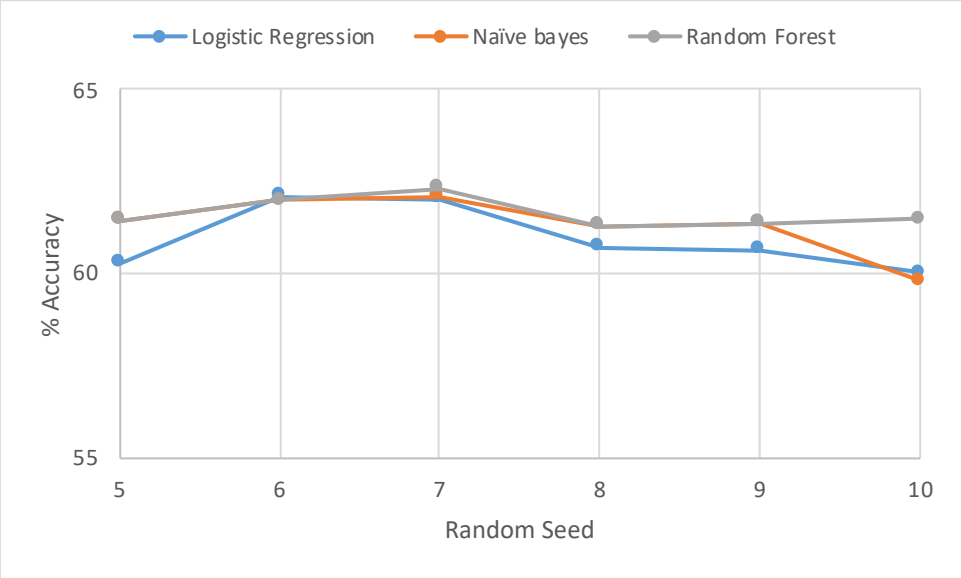

Figure 10: Accuracy plot

CHAPTER 5: CONCLUSION

Maternal health is one of the biggest challenges in the healthcare system of the US. The reports from CDC does not show a good trend in maternal mortality and morbidity. This study is aimed at finding out the factors that lead to adverse maternal outcomes using data mining approaches. This is helpful for the providers to provide better treatment and for the patients to understand the possible risk factors.

For this study, the data used are from NIH's AllOfUs research work bench, which gathers millions of patient records living within the US. Two patient cohorts are built from the All of Us data. One contains the data of all pregnant women available, and the other is data of pregnant women who had abnormal conditions or outcomes. The criteria applied to build the cohort are Female, Race and Ethnicity, Age, Pregnancy, Pregnancy conditions, Childbirth, and puerperium. Data mining approaches in healthcare are great methods to predict the features that can increase the probability of maternal mortality or morbidity. In this study three data mining techniques, Logistic regression, Naïve Bayes, Random Forest, are used. The prediction models developed by all three classifiers performed similar. Considering the features contributing to adverse maternal outcome, all three classifiers identified same 12 features.

LIMITATIONS

This research only focused on classifying abnormal and normal outcomes. However, several other specific outcomes such as mortality and morbidity require further analysis. The research was limited due to the inability to run programs on the AllOfUs servers. Further, analysis to improve the algorithms and identify relevant features are part of the future work. Several features such as lab work, and measures are treated as categorical variables. However, these features also

have values associated with it such as the value of blood pressure, sodium, and other minerals that can provide valuable insights in predicting the maternal outcome.

Future work

The accuracy of the model obtained from three of the datamining techniques used are not the best. Advanced machine learning techniques such as KNN, SVM, etc. will provide more accurate models with better prediction of the features. Furthermore, to address the imbalanced data usage of oversampling technique such as SMOTE will improve the model. Another area to improve in this research is the consideration of frequency of occurrence. For example, a single patient (PID) can have multiple visits, procedure, or lab work for the same reason. So, it can also be a predictor for adverse maternal outcome.

REFERENCES

[1] N. C. for C. D. P. and H. P. Division of Reproductive Health, "Pregnancy Mortality Surveillance System," *Centre for disease control*, 2018. https://www.cdc.gov/reproductivehealth/maternal-mortality/pregnancy-mortality-surveillance-system.htm (accessed Feb. 01, 2022).

[2] N. C. for C. D. P. and H. P. Division of Reproductive Health, "Severe Maternal Morbidity in the United States," *CDC*, 2021. https://www.cdc.gov/reproductivehealth/maternalinfanthealth/severematernalmorbidity.html.

[3] "Maternal health," *World Health Organization*. https://www.who.int/health-topics/maternal-health#tab=tab_1 (accessed Oct. 02, 2022).

[4] I. Yoo *et al.*, "Data mining in healthcare and biomedicine: A survey of the literature," *J. Med. Syst.*, vol. 36, no. 4, 2012, doi: 10.1007/s10916-011-9710-5.

[5] A. F. Peahl, R. Smith, T. R. B. Johnson, D. M. Morgan, and M. D. Pearlman, "Better late than never: why obstetricians must implement enhanced recovery after cesarean," *Am. J. Obstet. Gynecol.*, vol. 221, no. 2, 2019, doi: 10.1016/j.ajog.2019.04.030.

[6] J. Hitti, L. Sienas, S. Walker, T. J. Benedetti, and T. Easterling, "Contribution of hypertension to severe maternal morbidity," *Am. J. Obstet. Gynecol.*, vol. 219, no. 4, 2018, doi: 10.1016/j.ajog.2018.07.002.

[7] P. C. Wong and P. Kitsantas, "A review of maternal mortality and quality of care in the USA," *Journal of Maternal-Fetal and Neonatal Medicine*, vol. 33, no. 19. 2020, doi: 10.1080/14767058.2019.1571032.

[8] J. Hitti, L. Sienas, S. Walker, T. J. Benedetti, and T. Easterling, "Contribution of Hypertension to Severe Maternal Morbidity," *Obstet. Anesth. Dig.*, vol. 39, no. 3, 2019, doi:

10.1097/01.aoa.0000575224.15575.1b.

[9] E. K. Main, C. Markow, and J. Gould, “Addressing maternal mortality and morbidity in California through public-private partnerships,” *Health Aff.*, vol. 37, no. 9, 2018, doi: 10.1377/hlthaff.2018.0463.

[10] P. Mehta, “Addressing reproductive health disparities as a healthcare management priority: Pursuing equity in the era of the Affordable Care Act,” *Current Opinion in Obstetrics and Gynecology*, vol. 26, no. 6. 2014, doi: 10.1097/GCO.0000000000000119.

[11] G. K.D., K. L.M., L. J. D.C., H. J., J. J., and P. J., “Age and racial/ethnic differences in maternal, fetal, and placental conditions in laboring patients,” *Am. J. Obstet. Gynecol.*, vol. 188, no. 6, 2003.

[12] J. E. Edwards and J. C. Hanke, “An update on maternal mortality and morbidity in the United States,” *Nurs. Womens. Health*, vol. 17, no. 5, 2013, doi: 10.1111/1751-486X.12061.

[13] C. Baptiste and M. E. D’Alton, “Applying Patient Safety to Reduce Maternal Mortality,” *Obstetrics and Gynecology Clinics of North America*, vol. 46, no. 2. 2019, doi: 10.1016/j.ogc.2019.01.016.

[14] C. Mitchell, E. Lawton, C. Morton, C. McCain, S. Holtby, and E. Main, “California pregnancy-associated mortality review: Mixed methods approach for improved case identification, cause of death analyses and translation of findings,” *Maternal and Child Health Journal*, vol. 18, no. 3. 2014, doi: 10.1007/s10995-013-1267-0.

[15] J. L. Vázquez-Calzada, “Cesarean childbirth in Puerto Rico: the facts.,” *P. R. Health Sci. J.*, vol. 16, no. 4, 1997.

[16] S. M. Berman, H. T. Mackay, D. A. Grimes, and N. J. Binkin, “Deaths From Spontaneous Abortion in the United States,” *JAMA J. Am. Med. Assoc.*, vol. 253, no. 21, 1985, doi: 10.1001/jama.1985.03350450091028.

[17] J. J. Shen, C. Tymkow, and N. MacMullen, “Disparities in maternal outcomes among four ethnic populations,” *Ethn. Dis.*, vol. 15, no. 3, 2005.

[18] R. E. Howland *et al.*, “Determinants of Severe Maternal Morbidity and Its Racial/Ethnic Disparities in New York City, 2008–2012,” *Matern. Child Health J.*, vol. 23, no. 3, 2019, doi: 10.1007/s10995-018-2682-z.

[19] D. Rosenberg, S. E. Geller, L. Studee, and S. M. Cox, “Disparities in mortality among high risk pregnant women in Illinois: A population based study,” *Ann. Epidemiol.*, vol. 16, no. 1, 2006, doi: 10.1016/j.annepidem.2005.04.007.

[20] J. Y. Hsu *et al.*, “Disparities in the management of ectopic pregnancy,” *Am. J. Obstet. Gynecol.*, vol. 217, no. 1, 2017, doi: 10.1016/j.ajog.2017.03.001.

[21] V. Sundaram, K. L. Liu, and F. Laraque, “Disparity in maternal mortality in New York City.,” *J. Am. Med. Womens. Assoc.*, vol. 60, no. 1, 2005.

[22] E. L. Shearer, “Cesarean section: Medical benefits and costs,” *Soc. Sci. Med.*, vol. 37, no. 10, 1993, doi: 10.1016/0277-9536(93)90334-Z.

[23] M. R. Kramer *et al.*, “Changing the conversation: applying a health equity framework to maternal mortality reviews,” *Am. J. Obstet. Gynecol.*, vol. 221, no. 6, 2019, doi: 10.1016/j.ajog.2019.08.057.

[24] J. Bernstein, “Ectopic Pregnancy: A Nursing Approach to Excess Risk Among Minority Women,” *J. Obstet. Gynecol. Neonatal Nurs.*, vol. 24, no. 9, 1995, doi: 10.1111/j.1552-6909.1995.tb02564.x.

[25] E. A. HOWELL and Z. N. AHMED, “Eight steps for narrowing the maternal health disparity gap,” *Contemp. Ob. Gyn.*, vol. 64, no. 1, 2019.

[26] D. Noell *et al.*, "Ectopic Pregnancy Mortality—Florida, 2009–2010," *Obstet. Gynecol. Surv.*, vol. 67, no. 12, 2012, doi: 10.1097/01.ogx.0000425646.95865.e7.

[27] J. Burns, "Employer Groups Pushing For Improving Maternal Health," *Manag. Care*, vol. 27, no. 8, 2018.

[28] M. H. Tekieh and B. Raahemi, "Importance of data mining in healthcare: A survey," 2015, doi: 10.1145/2808797.2809367.

[29] P. A. Cavazos-Rehg *et al.*, "Maternal Age and Risk of Labor and Delivery Complications," *Matern. Child Health J.*, vol. 19, no. 6, 2015, doi: 10.1007/s10995-014-1624-7.

[30] M. A. Frölich, C. Banks, A. Brooks, A. Sellers, R. Swain, and L. Cooper, "Why do pregnant women die? A review of maternal deaths from 1990 to 2010 at the University of Alabama at Birmingham," *Anesth. Analg.*, vol. 119, no. 5, 2014, doi: 10.1213/ANE.0000000000000457.

[31] S. A. Leonard, E. K. Main, and S. L. Carmichael, "The contribution of maternal characteristics and cesarean delivery to an increasing trend of severe maternal morbidity," *BMC Pregnancy Childbirth*, vol. 19, no. 1, 2019, doi: 10.1186/s12884-018-2169-3.

[32] K. Azimi, M. D. Honaker, S. Chalil Madathil, and M. T. Khasawneh, "Post-Operative Infection Prediction and Risk Factor Analysis in Colorectal Surgery Using Data Mining Techniques: A Pilot Study," *Surg. Infect. (Larchmt).*, vol. 21, no. 9, 2020, doi: 10.1089/sur.2019.138.

[33] A. Subasi, *Practical Machine Learning for Data Analysis Using Python*. 2020.

[34] S. Misra and H. Li, "Noninvasive fracture characterization based on the classification of sonic wave travel times," in *Machine Learning for Subsurface Characterization*, 2019.

[35] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in {P}ython," *J. Mach. Learn. Res.*, vol. 12, pp.

2825–2830, 2011.

VITA

Prajina Edayath was born in India. She acquired her bachelor's degree in Electrical and Electronics Engineering from Cochin University of Science and Technology in 2011. She entered for master's in Industrial Engineering in the University of Texas at El Paso in 2020. She presented papers in IISE annual conference in 2021 and 2022.

Contact Information: pedayath@miners.utep.edu