Impact of Medication Use on Pregnancy Outcomes

Ashritha Dandam

Submitted to the faculty of the School of Informatics in

Partial

fulfillment of the requirements

for the degree of

Master of Science in Health Informatics, Indiana University

APRIL 2019

Accepted by the Faculty of Indiana University, in

partial fulfilment of the requirements for the degree of

Master of Science in Health Informatics

Master's Project

Committee

Dr. Josette Jones, Ph. D.

Director, Health Informatics, SOIC, IUPUI

Faculty Advisor Name, Degree Title

Dr. Sara Quinney, Pharm.D, Ph.D.

Assistant professor, Department of Obstetrics and Gynecology, IU school of medicine

Committee Member Name, Degree, Title

Table of Contents

- LIST OF FIGURES……...6
- TABLES……....8
- Abbreviations……9
- ACKNOWLEDGEMENTS……..10
- ABSTRACT……..11
- INTRODUCTION……12
- LITERATURE REVIEW……... 23
- SIGNIFICANCE……28
- Aim……28
- APPROACH TO SOLUTION……...28
- Research question……...28
- Possible solution……28
- Objective……29
- DATA……..29
- Data Description……29
- METHODOLOGY……32
- Study duration……32
- Inclusion Criteria……32
- Exclusion Criteria……... 33
- Data collection ……33

- Data loading ……..34
- Data preparation ……...34
- Data preprocessing ……40
- Feature Extraction ……46
- Data Analysis ……53
- RESULTS …….....64
- DISCUSSION……72
- LIMITATIONS …….....74
- CONCLUSION …….....74
- Future directions ……....74
- References ……..75
- Poster ……81
- CV ……..82

LIST OF FIGURES

Figure1: Preterm birth for live and stillbirth……23
Figure 2: Patients Medication chart dataset…….32
Figure 3: Data access from user defined directory……..33
Figure 4: Conversion of data from wide to long format……..34
Figure 5: Dictionary and mapping ……..37
Figure 6: Dictionary for Drug name and ATC code……38
Figure 7 : Bubble chart commonly used drug class……41
Figure8: Bar chart for the most commonly used drug class with frequency count…….. 42
Figure 9: SMOTE technique for pregnancy outcome……43
Figure 10: SMOTE technique for preterm birth dataset…….44
Figure 11: SMOTE technique for small for gestational age dataset……45
Figure 12: Recursive Feature Elimination live/stillBirth…….47
Figure 13: LASSO Feature Extraction live/stillBirth……..48
Figure 14: Chi-square Feature Extraction live/stillBirth…….48
Figure 15: RFE with Cross-validation feature selection live/stillBirth……...49
Figure 16: Features extracted by RFE live/stillBirth…….. 50
Figure 17: Features extracted by LASSO preterm birth …….51
Figure 18: Features extracted by RFE with cross-validation preterm birth…….52
Figure 19: Features extracted by LASSO for RFE with cross-validation……...52

Figure 20: Features extracted by LASSO for RFE with cross-validation……...53

Figure 21: Model for analysis…….54

Figure 22: Logistic model accuracy before SMOTE……..57

Figure 23: Confusion matrix logistic regression before SMOTE……...57

Figure 24: Random Forest regression after SMOTE……..58

Figure 25: Compare of different models for pregnancy outcome ……..58

Figure 26: Compare of different models for preterm birth …….61

Figure 27: Comparison of ROC for preterm birth Dataset……..62

Figure28: Drug count vs small for gestational age ……..65

Figure 29: Area under curve for logistic regression…….66

Figure 30: Logistic Regression Model Accuracy…….67

Figure 31: Performance measure of logistic regression……...68

Figure 32:Drug count vs preterm birth…….68

Figure 33: Performance measure of random forest……..69

Figure 34: Comparison of area under curve for preterm birth …….70

Figure 35: Comparison of Various algorithms for small for gestational age……....70

Figure36:Performance measure of small for gestational age……...71

Tables:

Table1: FDA classification of drugs used in pregnancy……21

Table2: Teratogenic mechanism with drug use……...22

Table3: Patients Chart Variable Dataset……...30

Table4: Patients Pregnancy Outcome Dataset……..31

Table5: Individual Drug list……35

Table6: Classification of ATC codes……37

Table7: Mapping drug names with ATC code and group by column ID……..39

Table8: Final dataset after merging……40

Table9: Dataset with drug count……...40

Table10: One Hot encoding technique……45

Table11: Final dataset for pregnancy outcome……56

Table12: Final dataset preterm birth ……59

Table13: Final dataset preterm birth after one hot encoding ……60

Table14: Final dataset small for gestational age …… 63

Table15: Final dataset small for gestational age after one hot encoding ……64

Table16: Evaluation of final models for different datasets……72

Abbreviations:

EHR- Electronic Health Records

ATC-Anatomical Therapeutic Classification of drugs

nuMoM2b - Nulliparous Pregnancy Outcomes Study Monitoring Mothers-to-be

FDA-US Food and Drug administration

HTN- Hypertension

SQL- Structured Query Language

LR- Logistic Regression

RFE- Recursive Feature Elimination/extraction

LASSO- least absolute shrinkage and selection operator

SVM- Support vector machine

ACKNOWLEDGEMENT

I would like to thank my project advisor, Dr. Sara Quinney, for her patient guidance, and timely suggestions throughout the project from IU school of medicine, Department of Obstetrics and Gynecology at Indiana University Purdue University Indianapolis. The door to prof Dr. Josette Jones was always open whenever I had a question about my project.

They allowed me my project to be own work and steering me in the right direction whenever I needed

I would also like to thank the experts who were involved in this project [David Guise, Jarod Watt] with their passionate participation and input the project was conducted successfully.

I extend my hearty thanks and well-wishes to all my colleagues, friends and seniors for their co-operation and help.

Finally, I would like to express our special indebtedness to my family who stood by me always and whose continuous encouragement and unconditional support was an unremitting source of inspiration for this work.

Abstract:

It is estimated that over half of women receive a prescription or nonprescription medication during pregnancy. Though some medications are known to cause birth defects or adverse pregnancy outcomes, which may lead to under prescribing for pregnant women and women of childbearing age. The number of drugs taken by pregnant women is increasing which may lead to increased birth defects. One cause of adverse pregnancy outcomes is adverse drug reactions or drug interactions and no proper alert system in prescribing the combinations of drugs to pregnant women. The aim of the study is to evaluate the impact of medication use on adverse pregnancy outcomes, such as stillbirth, preterm birth, and small for gestational age at delivery. This study aims to identify drugs that lead to adverse pregnancy outcomes and helps in further research to identify the root cause. The data used for this analysis is from the Numom2B study, a prospective cross-sectional study of 10,000 nulliparous women. For this analysis, we are focusing on 326 variables including medications used in pregnancy, demographic variables, and outcomes of life vs. stillbirth, small for gestational age, and preterm birth. The dependent variables are the pregnancy outcomes and independent variables are the medications prescribed. Data cleaning was performed to normalize drug names to third level hierarchy ATC codes. Machine learning models like logistic, random forest and CART were performed to predict the outcomes with an accuracy of 93.5%, 86.7%, and 62% respectively. We found that the number of drugs taken by women does not impact on the pregnancy outcomes. Specific classes of drugs, including anti-hypertensive, anti-diabetic, antidepressants were correlated with adverse pregnancy outcomes.

Introduction:

Medications should be used with caution in women of childbearing age who are pregnant or are contemplating pregnancy. (Viral M, 2019). Though the topical medications are considered safe than the parental agents, or the oral medications, their safety usage must be measured carefully while taking these medications. The information available for the medication use in pregnancy is limited, not always assisted by the FDA pregnancy category system. Now a days most of the drugs prescribed by the pregnant women are being harmful to the evolving fetus. Consequently, during the pregnancy time of the women, women with chronic diseases that involve detail and specific medications frequently reduce or stop taking the drugs that is required for their own health, possibly leading to harm to the fetus than drug itself. Pregnant women who have been exposed to antidepressants or other type of drugs like anxiolytics before or during the sixteenth week of pregnancy, those women have shown three-fold increased risk for preeclampsia when compared to other women who have not taken these antidepressants and anxiolytic medication during their 16^{th} week of pregnancy(Bernard & Forest, 2019). Results also prove that women who has completely stopped these antidepressants or anxiolytics during their pregnancy time have benefitted from reduced preeclampsia risk. In a study by Raffi and Cohen, the estimation of the prescription drug usage involving minerals and vitamins during their first trimester is ranging from around 33 percent to 69 percent in all the well-developed countries (Raffi & Cohen, 2019). In United states, the average number of prescription and over the counter drugs has shown an increase from 2.5 in the yeah 1976-1978 to around 4.2 in the year 2006-2008. Until now, the

quality and quantity of the information about the risk of fetus of the medication usage in the pregnant women continues to remain poor. In a review about 172 medications which has been approved by the US FDA between the years 2000 and 2010 has found that for around 97.7 percent of women, teratogenic risk in their pregnancy was not determined and for around 73.3% there was no data available regarding the risks occurring during their pregnancy.

The adverse birth outcomes include the preterm births which is medically defined as baby who is less than 37 weeks of gestation, low birth weight, baby less than 5.5 pounds. These adverse pregnancy outcomes can lead to complex rates of infection and illness for the newborn babies, health problems and as well as long term neurological problems.

The rate of obesity and overweight are increasing in the pregnant women now a days in all developed countries. These women are more prone to high risk of adverse pregnancy outcomes. The obese women are at an increased risk of gestational diabetes, pre-eclampsia and pregnancy induced hypertension. Therefore, active preventative strategies are immediately required (Chaturica, 2010). Several methods for analyzing the risk of adverse pregnancy outcomes, been the source of many publications and debates. The problem is the conflation of time at risk with preterm birth at birth of the newborn baby. Three different approaches in measuring the risk regarding the preterm birth, each addressing a separate etiological or prognostic question that means the risks are associated with cumulative, prospective and instantaneous, suggesting the suitable denominators for each (Kramer & Zhang, 2014).

Low birth weight and preterm birth are the most pertinent elements of newborn infants' existence, both in established and developing countries. There are several risk factors identified in literature both for the preterm birth and low birth weight. Infections involving the

genitourinary tract infections with several genetic and biological factors are the etiological factors for preterm and low birth weight deliveries. Though there is an evidence that subclinical infection sites which is distant from the Genito-urinary tract infection can also be considered as a significant cause for the preterm and low birth weight deliveries. Another possible risk factor for the preterm and low birth weight reported by many authors is the Maternal Periodontal status. Though this review has found reliable association among preterm birth/low birth weight and periodontitis, this should be treated carefully until the heterogeneity sources can be explained (Jagan & Amrutha, 2012). There are several risk factors associated with these preterm birth and numerous pathways are ensuing in labor. The 4 important factors which leads to the preterm labor are namely the maternal/fetal stress, intrauterine infection, excessive uterine stretch and decidual hemorrhage. Other factors leading to preterm birth or low birth weight involves Uteroplacental vascular insufficiency, when the inflammatory response is exaggerated, cervical insufficiency, hormonal factors and genetic predisposition.

The following conditions constitute for the risk of small for gestational age and increase the risk for developing the fetal growth restriction. Few conditions constituting these risk are when there is no proper nutrition during the pregnancy, birth abnormalities or chromosomal defects, when the mother weight is less than 100 pounds, PIH which means pregnancy induced hypertension, abnormalities of placenta and umbilical cord, multiple pregnancies, gestational diabetes in mom and oligohydramnios which means very low levels of amniotic fluid. There are also long-term complications associated with this risk. Present studies say that the intrauterine growth restriction rises the possibility of the problems through the adulthood involving cardiovascular disease, obesity, diabetes and high blood pressure. Small for gestational age newborn babies are prone to

have weakened immune system which rises the risk of developing the infections in hospitals. Regardless of size, small for gestational age newborn babies generally act and look alike the normal sized babies with those of similar gestational age. Few of the small for gestational age newborns appear thin and have very less muscle fat and mass and few of the babies have wizened facies which means having sunken facial features.

There are association among the 14 medical exposures and stillbirth were observed by means of deliveries in 1984 in about ten California counties, around 332 cases were stillbirths and they observed infants' deaths just within 24 hours of birth (Pastore & Hertz, 1999). The randomly selected live births were served as controls and were regularly matched by the county and the maternal age. With the help of important statistical data, questionnaire and relative hazards, modelling was achieved with alteration and adjustment for the possible confounders. The utmost predominant exposures were acetaminophen and ultrasounds. During the first two gestational months, prescription pain medications was powerfully linked with the stillbirths due to the inherited irregularities. 1st and 2nd trimester usage of prescription pain or medication of migraine was certainly linked with all the stillbirths. The fertility drugs were very positively linked with the stillbirths in total and stillbirths involving the complications of the placentas, membranes and the umbilical cord. There was no association found with aspirin, fever, diagnostic X-rays, amniocentesis, reliable with preceding studies. This report is between few studies of definite causes of stillbirth and the medical experiences by the gestational time window.

NuMom2b has started in 2010, it is a study which studies all the pregnant women who will be delivering for the first time- women delivering for the first time are called nulliparous women. This is a prospective cohort study estimating the fundamental interrelated mechanisms of

numerous adverse pregnancy outcomes which are common, and which cannot be predicted in the women who is having little or no pregnancy history at all, to assist guide their treatment. NuMom2b study helps in addressing a critical group of at-risk pregnant women who are presently understudied and signify about 40 percent of US births every year. The results and outcomes of this study will assist in informing the health care providers and their respective patients who are considering pregnancy or pregnant and aids in supporting further research to enhance the outcomes and care in this group of women. A National heart institute has funded a substudy of about 3,600 nuMoM2b women participants, is investigating the relationship among the sleep disorders through pregnancy and its adverse consequences. This nuMoM2B study is enrolling and joining ethnically, racially and geographically various pregnant women over about eight clinical research sites and 12 subsites all over the country. These pregnant women participate in various tests in order to recognize the potential mechanisms of the adverse pregnancy outcomes and analytical, predictive aspects for the consequences they are facing at 4 points during their pregnancy. The main aim of this Nulliparous pregnancy Outcomes study is to regulate the maternal characteristics which involve physiologic and genetic response to pregnancy and environmental features that helps in predicting the adverse pregnancy outcomes (Haas & Parker, 2015). The nulliparous women in their 1st trimester of pregnancy was taken into an observatory cohort study. The women participating in this study were seen at 3 study visits during their pregnancy and at their deliver time. The data was collected during the in-clinic interviews, clinical measurements and take-home survey questionnaires, chart abstractions and ultrasound studies. Maternal biospecimens like urine, plasma, serum and cervicovaginal fluid during the antepartum study visits and delivery specimens like cord blood, placenta and

umbilical cord were collected together, processed and finally stored. The initial goal of this study was defined as pregnancy ending at <37+0 gestation weeks. The key study hypothesis of this study comprises the pregnancy outcomes of preeclampsia, spontaneous preterm birth and fetal growth restriction. This study assists the investigators in planning the future projects. 40 percent of women in US have never given birth. Women have complications with their pregnancy, but there is no information available from a previous pregnancy to detect the problems that have caused it. The nulliparous women are enrolled early in pregnancy and has undergone research assessments around 4 times during their pregnancy. Women are asked to participate in sub studies gathering information regarding sleep patterns, breathing, quality and other areas relating to birth outcomes. Therefore, the only goal of this study is to recognize the women in this group developing problem with their pregnancy and using this information to enhance the health of the pregnant women and their babies in future. This study helps in focusing the problems associated with pregnancy like high blood pressure, very small babies and babies who are born too early.

During pregnancy, it is stated that more than 50% of women receive prescription and nonprescription drugs. The use of drugs or medications is increased among pregnant women from 2010 to 2017. In the world, 2 to 3 % of birth defects are associated with drug-related issues. Usage of prescription drugs among pregnant women is prevalent nowadays. According to statistics, 44-99% of pregnant women were prescribed with medications. The use of prescription drugs in pregnancy is associated with the unfamiliar risk-benefit condition. Drugs prescribed for better healthcare of mother can also mislead or cause risk to the fetus. Till date only a few drugs i.e., less than 30 drugs are known to cause defects in pregnant women. Most of the clinical trials studies are less focused on determining the effect of drugs on pregnant women rather inclined

towards the general population. Very few drugs outlined the adverse effects on pregnant women. The pharmacological effect of drugs in the second and third trimester of pregnancy occurs which can 3 be detected and maintained. The recent incident about the H1N1 that affected pregnancy women demands detailed research about the studies in therapeutic level. The pregnancy is considered one of the distinct and special conditions which are influencing the drug absorption, drug distribution and drug elimination due to the physiological alterations that occur during pregnancy. Some pregnant women have medical conditions that require ongoing treatment and episodic treatments. The United States preterm birth rose from 9.57% to 9.85% during 2014-2016(Martin,2018) Infants born to preterm are vulnerable to many complications including respiratory distress syndrome, injury to intestines and compromised immune system. Advanced maternal age of 40 years was associated with preterm birth (Monet,2018). The preterm birth is considered as a significant cause of death for children under age 5 according to WHO. The spontaneous reason for the preterm births may be due to the rupture of membranes, and few are due to the induction of labor for non-medical and medical reasons. According to Paulo Cesar, 2012, the urogenital infections in pregnant women are the cause of preterm birth. The study is a perspective cross-section with 10000 sample size and 326 variables. The dependent variables are the pregnancy outcomes and independent variables are the medications prescribed. The study was conducted to analyze the pregnancy outcome related to medication use. The inclusion criteria for this Numom2b study include pregnant women whose previous pregnancy has not lost more than 20 weeks and patients age of 15 years and older. The exclusion criteria for this study are pregnant women less than 13 years and patients who are unable to provide informed consent forms. Women using medications during pregnancy are prone to adverse outcomes. 50% of the

women use the medication before the pregnancy and 97% use during pregnancy with a mean of 4 drugs per women. The medications comprise anti-hypertensives, anti-diabetics, vitamins, minerals, etc. The percentage of medication usage is highest between the age groups for 25-32 years. The pregnancy outcomes of many drugs have not been studied effectively and have an undetermined risk of preterm birth. Upon reviewing the patient history of medications, it can reduce the risk associated with the preterm birth. The physicians in the obstetrics department must face a challenge in prescribing medication when the prevalence rate is higher than 9%. The main goal of this study is to analyze and visualize the medication usage in pregnancy outcomes and deliver a healthier approach for pregnant women about the drugs they use in pregnancy. The rate of early preterm birth, < 32 weeks of gestation has not changed since 1990, accounting for 2% of all births but 54% of infant mortality in the United States (David M. HAAS, 2015). The lack of knowledge and information about the root cause of the issue is considered as the major limitation in solving the issue. The older history of pregnancy records would act as a risk manager in assessing the women in their next pregnancy, but 40% of the women lack in nulliparous information. According to Hass (2015)," The identification of women at increased risk for preterm birth has been traditionally directed towards various epidemiologic, clinical, and environmental risk factors". The triggers for the preterm birth include the previous history of similar complication, increased cervical-vaginal fetal levels and shortened the cervical length. In order to overcome this and to identify the impact of drugs on the pregnancy adverse outcomes, the study is designed to act as a risk assessor for all healthcare providers. In summary, the consequence of this study examined the risk of live birth, stillbirth, preterm birth and small of gestational age for drugs and used in pregnancy and developing a clinical decision support

system that contains information on safety and security rating of drugs in pregnancy and to progress clinical outcomes. Clinical decision support system advances medication safety and decreases inappropriate prescribing of medications as it has automation method while ordering which is a key process in health care.

According to the UK General Practice Research Database data, among 81975 pregnancy women, the percentage of women prescribed at least one medication is 65% from which 1 in every 164 had received US Food and Drug Administration (FDA) category X (Box 1) drug in early pregnancy. In the United States and Canada, the percentage of women taking category D and X was 4.8 to 5.2% and 3.9 to 4.6% respectively. The most commonly prescribed class of drugs before the delivery are analgesics, anxiolytics, antidepressants, anti-asthmatic and antibiotics. According to a recent Iris study with a sample size of 23989 pregnancy women, 39.2% of the women were prescribed with at least one mediation. Below figure represents the category of drugs in pregnancy as per US FDA.

Box 1. US Food and Drug Administration (FDA) classification system for drugs used in pregnancy[35]

Category of drug in pregnancy	Explanation
A	Adequate and well-controlled studies have failed to demonstrate a risk to the fetus in the first trimester of pregnancy (and there is no evidence of risk in later trimesters).
B	Animal reproduction studies have failed to demonstrate a risk to the fetus and there are no adequate and well-controlled studies in pregnant women.
C	Animal reproduction studies have shown an adverse effect on the fetus and there are no adequate and well-controlled studies in humans, but potential benefits may warrant use of the drug in pregnant women despite potential risks.
D	There is positive evidence of human fetal risk based on adverse reaction data from investigational or marketing experience or studies in humans, but potential benefits may warrant use of the drug in pregnant women despite potential risks.
X	Studies in animals or humans have demonstrated fetal abnormalities and/or there is positive evidence of human fetal risk based on adverse reaction data from investigational or marketing experience, and the risks involved in use of the drug in pregnant women clearly outweigh potential benefits.

Table1: FDA classification of drugs used in pregnancy

The above table tells us about the US food and drug administration classification system of drugs used in the pregnancy. The table contains the A, B, C, D, X category of drug in pregnancy and its respected explanation.(Table 1).

The prevalence of malformations related to congenital is reported to be 2 to 3% from over al childbirths from which 1% are due to prescription drug usage. As per research in England and Wales in the year 2010, among 723000 live births, 10 in 200 were having drug-related issues.

Box 3. Teratogenic mechanisms associated with medication use[15]

- Folate antagonism, e.g. anti-epileptic drugs, methotrexate, trimethoprim, sulfasalazine and metformin
- Endocrine disruption, e.g. diethylstilbestrol, fertility hormones, oral contraceptives
- Neural crest cell disruption, e.g. retinoids and bosentan
- Oxidative stress, e.g. thalidomide, anti-epileptic drugs, class III anti-arrhythmics, iron supplements
- Vascular disruption, e.g. misoprostol, aspirin, ergotamine, pseudoephedrine
- Specific receptor or enzyme-mediated teratogenesis, e.g. ACE (angiotensin-converting-enzyme) inhibitors, angiotensin II receptor blockers, statins, non-steroidal inflammatory drugs

Table2: Teratogenic mechanism with drug use

The table describes the teratogenic mechanisms which are associated with medication use during the pregnancy taken by the women (Table 2).

Stillbirth is defined as the fetus death after 20 gestational weeks in pregnant women, the trends of live birth a stillbirth are increasing over the years. As per surveys, it is reported that 6.05 stillbirths per 1000 live births from 2006 to 2012. Preterm birth and pregnancy outcome play a key role as a health indicator for the newborn baby.

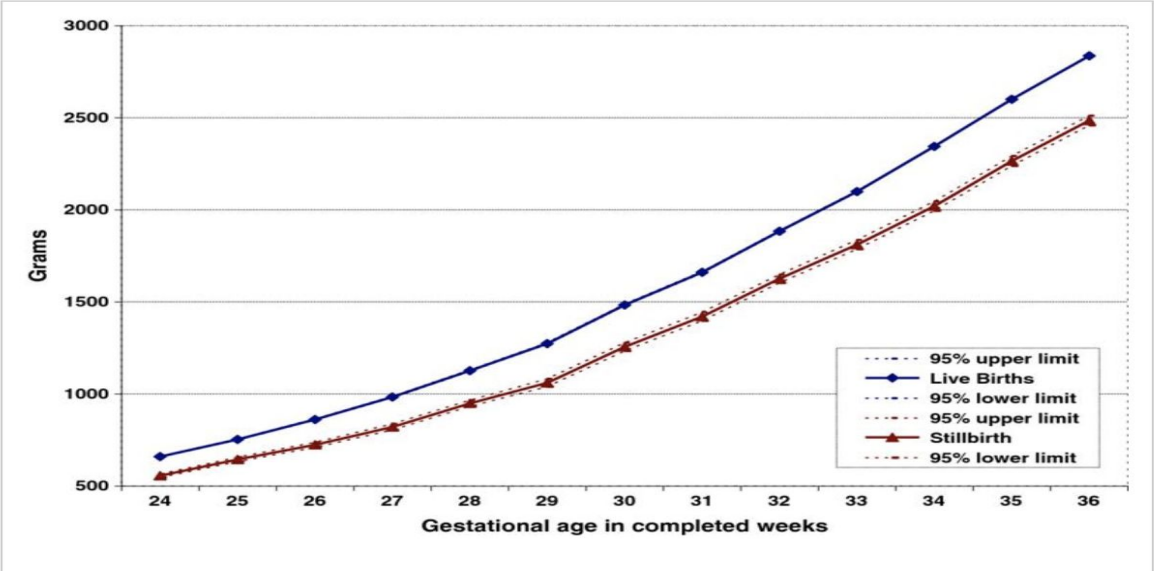


Figure1: Preterm birth for live and stillbirth

The figure shows the outcomes of live births and stillbirth from 24 to 36 weeks of pregnancy(Figure 1).

LITERATURE REVIEW:

The literature review discusses the relationship between antidepressants and preterm births. One of the significant concerns of the healthcare professionals is the preterm birth says the Author Krista F. Huybrechts et.al (). According to this article, increased usage of antidepressants leads to increased risk of depression in pregnant women. The author has done an extensive search using websites like PubMed, Medline and PsycINFO to find out the correlation between antidepressants and pregnancy outcomes. Here the data is collected and extracted independently by researchers whose relative and absolute risks are calculated. The summary measures of the effect were calculated by the random-effects model. The statistical significance results have proven that there is some association between the antidepressants and the pregnancy outcomes. As per the research, one in ten born babies are affected with preterm birth which is one of the

major concerns for the physicians and it accounts for about 15 million per year. The medication usage and complications occurring during pregnancy are considered as major reasons for preterm birth. The author Florian Knock mentions that diabetes mellitus is correlated with long term complications and fetal-maternal morbidities. One of the greatest challenges considered here is preventing the preterm birth as it is affecting the length of incidence and gestation period of the babies. The risk factor identified for the preterm birth is the poor glycemic control, therefore by controlling or regulating the glucose levels, the preterm births can be prevented.

Nutritional and Antimicrobial Interventions to Prevent Preterm Birth: An Overview of Randomized Controlled Trials

This study is to examine and analyze the efficiency of the interventions which are used in preventing the preterm birth associated with the infectious disorders occurring during the pregnancy. The articles are collected from the Cochrane library. The preterm delivery rate and prematurity are considered as primary and secondary outcomes based on the systematic reviews which are done on antibiotic and nutritional interventions in pregnancy. The inclusion criteria don't involve general interventions. Data of the preterm rate is obtained from each of the paper reviewed. Information like the number of members participating, a number of the outcomes obtained, and interventions involved are noted and carefully reviewed. 18 papers were involved which are containing the interventions made on antimicrobial and nutritional agents on the outcome of preterm birth. According to the results obtained from the article, it shows that the iron supplements reduces or prevents the risk of preterm birth and other elements like magnesium, fish oil, zinc have also shown decrease risk of the preterm birth. Metronidazole

which is an antimicrobial agent has also shown a decreased risk of the preterm births in women suffering from the urinal and vaginal infections. Although calcium supplements have promising results in several kinds of literature, it didn't show any impact or association with the preterm birth. With the help of this study, it is understood that there are few drugs reducing the risk of preterm birth in pregnancy outcome and calcium having no effect on PTB.

New perspectives for the effective treatment of preterm labor.

According to the author Keirse MJ, one of the serious issues that must be addressed immediately is the preterm birth of the babies. The author feels that these interventions can be reduced by deliver proper care to pregnant women during her pregnancy and with some proper interventions. This paper mainly deals with tocolytic agents which are mainly useful in postponing the delivery of the pregnant women so that it helps in preventing the preterm births. There are several things to be investigated says the author regarding the preterm births that is identifying the authentic cause of PTB and if identified the cause, trying to prevent the preterm labor and delivering proper care in order to prevent the PTB outcome. It is significant to examine the reasons associated with preterm births and their causing factors. The reasons or factors may be due to the consumption of antibiotics, corticosteroids and some of the nutritional supplements. With the help of identifying and understanding the cause for PTB, it aids in saving and delivering a healthier life to the preterm babies.

Calcium supplementation in the Nulliparous women for prevention of preterm birth

This article written by Caroline A. Crowther states that calcium which is considered as one of the important nutritional supplements reduces the risk of preterm birth in babies. The supplements of

calcium also decrease the risk of pre-eclampsia in pregnant women. This study is considered a randomized controlled double-blinded study and prospectively multicentered. The results of the methodology are much more reliable as it is a highly advanced one. Calcium supplements are usually prescribed for nulliparous pregnant women. 1.8g of the daily dose is used until the delivery. The risk of PTB and relative risk are 0.44 which is considered as the significant p-value. The article highlights that it is a multicenter study. The patients selected in this article are randomized and are from different centers. First-time pregnant women called the nulliparous women are taken for this study. This study is found to limit the generalization if the results. In this article, the author concludes that calcium supplements can reduce the adverse risk of preeclampsia and PTB in nulliparous pregnant women.

Use of azathioprine and corticosteroids during pregnancy and birth outcome in women diagnosed with inflammatory bowel disease.

This article is written by Anne Veie who mentions that the use of medication during pregnancy is one of the major concerns and these permeable drugs pass through the placenta. The interesting concern considered during pregnancy is medication safety, but there is no proper research. Congenital anomalies and other complication of the pregnancy outcomes are led due to the increased use of medications for inflammatory bowel disease. One of the drugs prescribed for the inflammatory bowel disease is Azathioprine but there is no proper evidence that pregnant women exposing to this drug will result in an increased risk of the PTB. The pregnancy outcomes in this study are categorized into 3 main groups namely live birth, stillbirth and preterm birth. The pregnant women with inflammatory bowel disease are prescribed with the drug called

azathioprine showing the amplified risk of the PTB. The article concludes that discounting or reducing the usage of the azathioprine is recommended for pregnant women with inflammatory bowel disease during pregnancy.

Combination antiretroviral use and preterm birth. The Journal of infectious diseases

This study involves the pediatric HIV/AIDS cohort study which is evaluating the use of antiretroviral use in pregnancy and its association with the risk of the outcomes. The pregnant women who have been exposed to the antiretroviral agents during her pregnancy period are at a high risk of PTB and spontaneous PTB which is occurring due to the rupture of the membranes. The findings in this study suggest that the antiretroviral agents used in the early stage of pregnancy are associated with the amplified risk of the prematurity. Thus, it states that increased risk of the PTB is seen with the usages of antiretroviral agents before pregnancy.

Risk of preterm birth following late pregnancy exposure to NSAIDs or COX-2 inhibitors

This article is written by the author Berard who describes that pregnant women are more prone to selective cox-2 inhibitors and non-steroidal anti-inflammatory drugs for regulating the disease flares which occur during the pregnancy. All women having singleton live birth between Jan 1998 and Dec 2009 were included in this Quebec pregnancy cohort study. Around 156,531 pregnancies met the inclusion criteria. Among these, around 391 pregnancies have been exposed to COX-2 inhibitors and NSAIDs. Crude adjusted the odds ratio with almost 95% of confidence intervals which is initially calculated for identifying the risk of the prematurity. The article

concludes that analysis exposed to C0X-2 inhibitors and NSAIDs is associated with 1.65-fold of risk of PTB.

Significance:

The main aim of the study is to identify the impact of prescription drugs on the pregnancy outcomes especially conditions like small for gestational age, stillbirth, live birth, and Preterm birth. The main purpose of the project is to tell with confidence that a patient pregnancy outcome is associated with defects or non- defect.

The aim of the project:

• To identify the drugs or medications leading to stillbirth or live birth

• To identify the small for gestational age of the baby based on the medication's intake

Solution Approach:

Research question: The main of the study is to find the impact of medication use on pregnancy outcomes?

Possible solution: The solution to the research question can be identified by building machine learning algorithms to predict what class of drugs contribute to the pregnancy outcome.

Null hypothesis: There is no association between drug use and pregnancy outcomes

Alternative hypothesis: An association of drug use and pregnancy outcome among the patients.

Objectives: The objective of the study is to identify the impact of mediations on pregnancy outcomes. From the results of this study, the outcome can be useful to pregnant women and physicians as an evidence-based approach in avoiding the drugs that are associated with risk of pregnancy outcome.

Data: The study was a prospective cross-sectional to find out the impact of medications on pregnancy outcomes. All the data was collected from multivariate hospital with the questionnaire method. The study consists of three datasets collected from the Numom2B study. The data is real-world data and it is challenging to handle the data, multiple drugs are prescribed to the same patient ID. Patient data were collected from three datasets namely chart variables, drugs in pregnancy label and pregnancy outcomes.

Dataset description:

Chart Variables:

The chart variables dataset consists of information from various sources namely hospital records, delivery records, previous pregnancy, newborn namely:

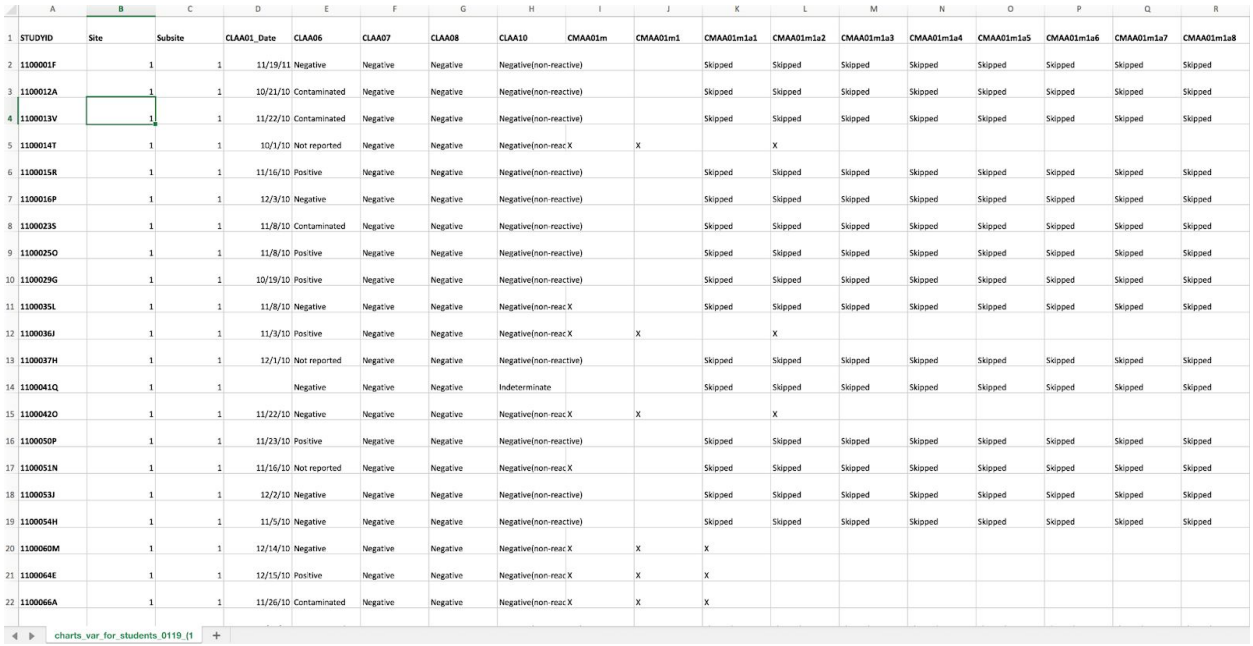
Fetal Anatomy

Newborn Outcomes

Maternal Hypertensive Disorders

Childbirth, Delivery, and Postnatal Experience

Cause of death etc.,



STUDYID	Site	Subsite	CLAA01_Date	CLAA06	CLAA07	CLAA08	CLAA10	CMAA01m	CMAA01m1	CMAA01m1a1	CMAA01m1a2	CMAA01m1a3	CMAA01m1a4	CMAA01m1a5	CMAA01m1a6	CMAA01m1a7	CMAA01m1a8
1100001F	1	1	11/19/11	Negative	Negative	Negative	Negative(non-reactive)			Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped
1100012A	1	1	10/21/10	Contaminated	Negative	Negative	Negative(non-reactive)			Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped
1100013V	1	1	11/22/10	Contaminated	Negative	Negative	Negative(non-reactive)			Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped
1100014T	1	1	10/1/10	Not reported	Negative	Negative	Negative(non-reac	X	X		X						
1100015R	1	1	11/16/10	Positive	Negative	Negative	Negative(non-reactive)			Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped
1100016P	1	1	12/3/10	Negative	Negative	Negative	Negative(non-reactive)			Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped
1100023S	1	1	11/8/10	Contaminated	Negative	Negative	Negative(non-reactive)			Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped
1100025O	1	1	11/8/10	Positive	Negative	Negative	Negative(non-reactive)			Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped
1100029G	1	1	10/19/10	Positive	Negative	Negative	Negative(non-reactive)			Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped
1100035L	1	1	11/8/10	Negative	Negative	Negative	Negative(non-reac	X		Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped
1100036J	1	1	11/3/10	Positive	Negative	Negative	Negative(non-reac	X	X		X						
1100037H	1	1	12/1/10	Not reported	Negative	Negative	Negative(non-reactive)			Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped
1100041Q	1	1		Negative	Negative	Negative	Indeterminate			Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped
1100042O	1	1	11/22/10	Negative	Negative	Negative	Negative(non-reac	X	X		X						
1100050P	1	1	11/23/10	Positive	Negative	Negative	Negative(non-reactive)			Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped
1100051N	1	1	11/16/10	Not reported	Negative	Negative	Negative(non-reac	X		Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped
1100053J	1	1	12/2/10	Negative	Negative	Negative	Negative(non-reactive)			Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped
1100054H	1	1	11/5/10	Negative	Negative	Negative	Negative(non-reactive)			Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped	Skipped
1100060M	1	1	12/14/10	Negative	Negative	Negative	Negative(non-reac	X	X	X							
1100064E	1	1	12/15/10	Positive	Negative	Negative	Negative(non-reac	X	X	X							
1100066A	1	1	11/26/10	Contaminated	Negative	Negative	Negative(non-reac	X	X	X							

charts_var_for_students_0119_(1

Table 3: Patients Chart Variable Dataset

The above table is the patient chart variable dataset containing study ID, hospital and delivery records, fetal anatomy, newborn outcomes etc (Table 3).

Pregnancy outcome:

The dataset consists of information regarding Study ID and pregnancy outcomes namely baby is stillbirth or live birth, preterm birth, Pre-gestational diabetes, chronic hypertension, and Preeclampsia/Gestational HTN. For the study, three outcomes are analyzed namely Pregnancy Outcome Live birth, stillbirth, preterm birth and small for gestational age.

Still Birth: Still birth is defined as the death of the baby after the 20th week of pregnancy

Preterm birth: It is defined as the measurement of pregnancy age taken from women's final period of the menstrual cycle. Normal pregnancy ranges from 38 to 42 weeks, whereas infants born 42 weeks later are considered as post mature.

Small for gestational age: It generally defines the baby weight; in other words, it illustrates about the weight of baby compared to normal weight as per the gestational weeks.

1	Enrollment Study ID Numbe	Pregnancy end date based o	Pregnancy outcome based o	Gestational age (completed	Type of livebirth/stillbirth b	Pre-gestational diabetes bas	Diabetes based on CMAE03	Chronic hypertension based	Preeclampsia/Gestational H	Preeclampsia/Gestational HTN (worst) using adapted ACOG 2013 criteria
2	StudyID	pENDDATE_CA_INT	pOUTCOME	GAwksEND	TYPE_CA_A09	PreGestDM	oDM	ChronHTN	PEgHTN	acog_PEgHTN
3	5100211I	11	Live birth	41	iFTLB	No	No pre-gestational DM or GI	No	No hypertensive disorder	No hypertensive disorder
4	6200344U	2	Live birth	40	iFTLB	No	Gestational diabetes	No	Mild preeclampsia	Mild preeclampsia
5	1100365N	-6	Live birth	39	iFTLB	No	No pre-gestational DM or GI	No	No hypertensive disorder	No hypertensive disorder
6	8100233U	0	Live birth	40	sFTLB	Yes	Pre-gestational diabetes	Yes	No hypertensive disorder	No hypertensive disorder
7	5100473B	7	Live birth	41	iFTLB	No	No pre-gestational DM or GI	Yes	No hypertensive disorder	No hypertensive disorder
8	5101612I	-6	Live birth	39	iFTLB	Yes	Pre-gestational diabetes	No	New onset intrapartum/post	New onset intrapartum/postpartum HTN
9	8101093F	8	Live birth	41	iFTLB	No	No pre-gestational DM or GI	No	No hypertensive disorder	No hypertensive disorder
10	7201828I	4	Live birth	40	iFTLB	No	No pre-gestational DM or GI	No	No hypertensive disorder	No hypertensive disorder
11	5100394U	9	Live birth	41	sFTLB	No	No pre-gestational DM or GI	No	No hypertensive disorder	No hypertensive disorder
12	7200866F	9	Live birth	41	sFTLB	No	No pre-gestational DM or GI	No	No hypertensive disorder	No hypertensive disorder
13	5100513P	6	Live birth	40	iFTLB	No	No pre-gestational DM or GI	No	New onset antepartum HTN	New onset antepartum HTN

Table 4: Patients Pregnancy Outcome Dataset

The above table is the patient's pregnancy outcomes dataset containing columns like study ID, Type of birth (live birth or stillbirth), Birth weight, small for gestational age, preeclampsia associated etc (Table 4).

Drugs in pregnancy: The dataset consist of information regarding study ID and corresponding medication taking pattern. The dataset consists of drugs in combination and explains patient medication consumption pattern.

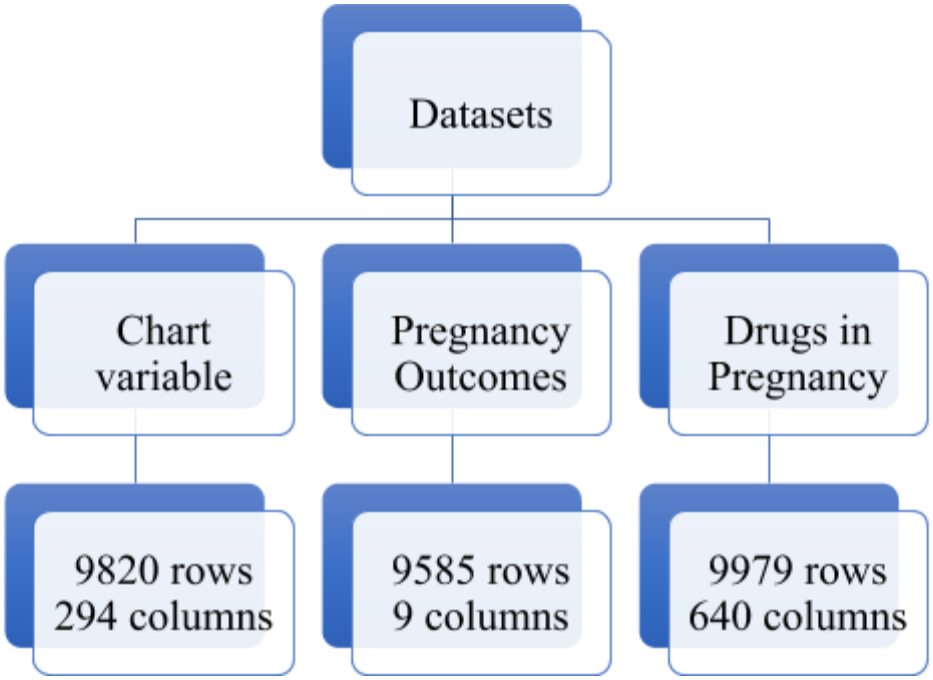


Figure 2: Patients Medication chart dataset

A final dataset is prepared from the three datasets by extracting the variables as per project needs. The dataset contains the chart variable containing 9820 rows and 294 columns, pregnancy outcomes table containing 9585 rows and 9 columns, drugs in pregnancy containing 9979 rows and 640 columns (Figure 2).

Methodology:

Study duration:

The study began in December 2019 to April 2019 for a period of five months

Inclusion criteria:

The inclusion criteria of the study include:

- Pregnancy women greater than 15 years of old
- Pregnancy women with outcomes associated with weight, birth and gestational weeks

Exclusion Criteria:

- Pregnancy women less than 15 years of age
- Vaccines and vitamin medications are removed from the final dataset

Hypothesis:

Null Hypothesis: The drugs have no impact on the adverse pregnancy outcomes

Alternative Hypothesis: Medications have an impact on the pregnancy outcomes namely still/live birth, small for preterm birth and preterm birth.

The Python packages like Matplotlib, NumPy operations and tableau were used for data visualization, including bar graphs, line charts, and bubble charts, to evaluate prescription patterns.

Data collection: All the datasets are collected from the Numom2B study. A special directory is created to access the datasets. A Secure access to the directory by user ID and password is created to access the data as it contains sensitive information. Datasets are gathering from multiple resources and prepared.

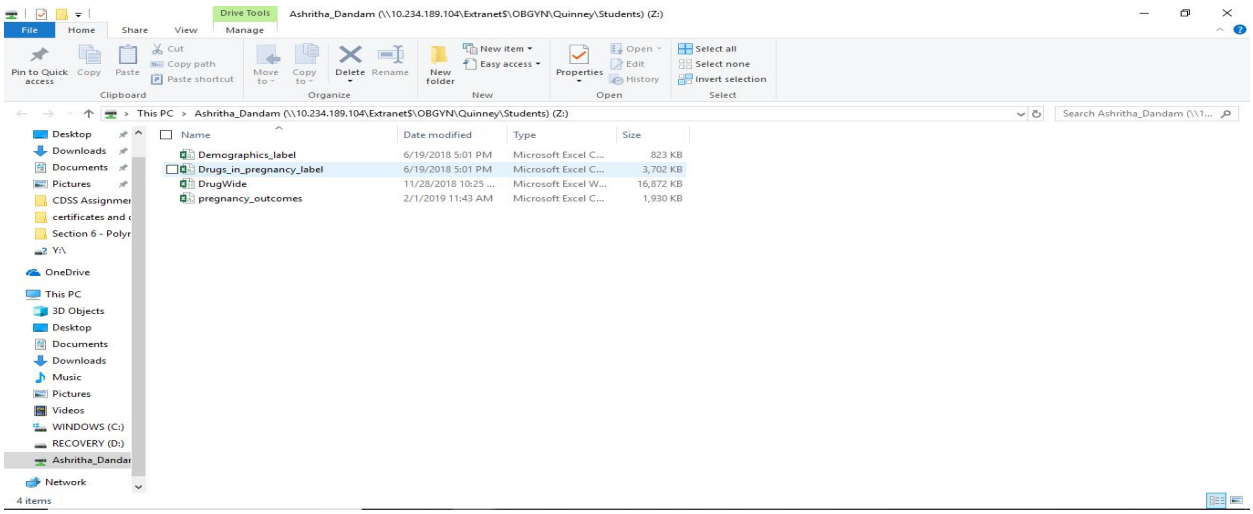


Figure 3: Data access from user defined directory

The above figure is the directory created for me from where I can access all the datasets for the project analysis (Figure 3).

Preparing the data: Data preparation consists of multiple steps namely data cleaning, data preprocessing and exploratory data analysis.

Data loading: All the datasets are loaded in SQL and performed data manipulation techniques line Insert, Update, delete, and update in SQL.

In Python tool library packages are installed and datasets are loaded.

Data preparation: The study consists of three different datasets; each dataset is prepared separately in python. Drugs in pregnancy dataset consist of drug names in column with combination such as(acetaminophen/dichloralphenazone/isometheptene).

Steps in converted Data from wide to long format:

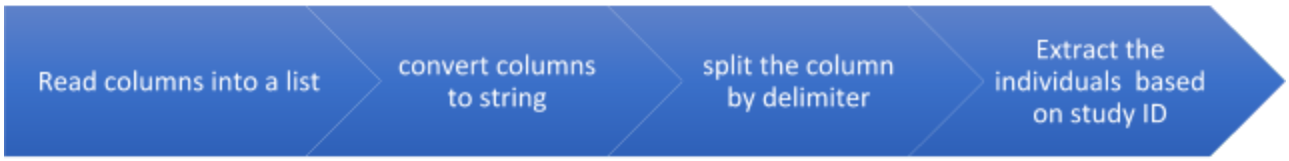


Figure 4: Conversion of data from wide to long format

Steps in converted Data from wide to long format: Drug data was converted from wide format to long format by reading columns into a list, converting columns to string, splitting the columns by delimiters and finally extracting the individual drugs from the combination drugs based on the study ID of the patients (Figure 4).

The dataset is loaded in the python and columns are read into a list. All the columns are converted into a string and sliced by a delimiter and taken into a list. All the drugs are converted into a unique drug name based on the study ID. The combination of drugs in columns are converted into individual drugs and aggregated based on the Study ID. The total number of drugs in the dataset is 602.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	StudyID	PTB	MIFEPRISTONE	CALAMINE	KETOCONAZOLE	SALICYLIC ACID	TETRAHYDROZOLI	VITAMIN B-1	DIAZEPAM	HYDROXYCHLORO	LACTASE	NILOTINIB	ARMODAFINIL	NIACIN	PAROXETINE	EPHEDRINE	CODEINE	ETHO
2	1100001F	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	1100012A	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	1100013V	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	1100014T	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6	1100015R	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
7	1100016P	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
8	1100023S	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
9	1100025O	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	1100029G	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	1100035L	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
12	1100036J	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	1100037H	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	1100041Q	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
15	1100042O	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	1100050P	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
17	1100051N	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18	1100053J	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19	1100060M	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Table 5: Individual Drug list

The above table is the individual drug list which is separated from the combination of drugs taken by the patients and the individual drugs are arranged based on study ID. 602 is the total number of drugs in the table (Table 5).

Preparing dictionary: A dictionary with ATC hierarchy is build corresponding to individual drug names for mapping the drug names. Data about the drugs and corresponding ATC was downloaded from bioportal and a dictionary as built based on one to one, one to many and many to one relationship.

ATC is the representation of medication name with a unique ID. ATC is defined as Anatomical Therapeutic Classification of drugs based on the organs which drug act and the composition of drug with the therapeutical, pharmacological and chemical nature. ATC was published in 1976 and maintained by the World Health Organization (WHO). According to the ATC drugs are classified into five levels of hierarchies.

1st Level	Anatomical main group	Musculo-skeletal system	M
2nd Level	Therapeutical subgroup	Anti-inflammatory and Antirheumatic products	M01
3rd Level	Pharmacological subgroup	Anti-inflammatory and Antirheumatic products, non-steroids	M01A

4th Level	Chemical subgroup	Propionic acid derivatives	M01AE
5th Level	Chemical Substance	ibuprofen	M01AE01

Table6: Classification of ATC codes

The above table explains the classification of the ATC codes, its levels, groups and its codes respectively (Table 6).

Figure 5: Dictionary and mapping

The above figure explains that the dictionary is created, validation is done by mapping the individual drug names with the ATC codes. Then reverse mapping is done to the ATC 3rd level Hierarchy which means the drugs are mapped to its class of drugs. Therefore, the drug count is reduced from 601 to 326 after mapping it with its class of drugs i.e. ATC 3rd level Hierarchy (Figure 5).

```
{ drug name :  atc3 ,
 'mifepristone': 'G03XB',
 'ketoconazole': 'D01AC',
 'salicylic acid': 'D01AE',
 'diazepam': 'N05BA',
 'hydroxychloroquine': 'P01BA',
 'nilotinib': 'L01XE',
 'armodafinil': 'N06BA',
 'niacin': 'C04AC',
 'paroxetine': 'N06AB',
 'ephedrine': 'A08AA',
 'codeine': 'N02AA',
 'ethosuximide': 'N03AD',
 'tacrolimus': 'D11AH',
 'sertraline': 'N06AB',
 'cetirizine': 'R06AE',
 'tetracycline': 'A01AB',
 'metoclopramide': 'A03FA',
 'celecoxib': 'L01XX'
```

Figure 6: Dictionary for Drug name and ATC code

The above figure shows us the drug name with its respective ATC 3rd level hierarchy code (Figure 6).

Mapping: Mapping the drug names to corresponding ATC 5th level codes and followed a bottom-up approach in mapping ATC 5th level to ATC 3rd level along with the description of the ATC class. A final dataset with ATC 3rd level hierarchy is built, and all the common columns are aggregated. Upon mapping the drugs to third level hierarchy the number of drugs decreased from 604 to 326.

	A01AA	A01AB	A01AC	A01AD	A02A	A02BA	A02BB	A02BC	A02BD	A02BX	...	V06A	V06B	V06D	V06DC	V07AC	V07AY	V09GX	study id	PTB
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1100001F	2.0
1	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1100012A	2.0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1100013V	2.0
3	0	1	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	1100014T	2.0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1100015R	2.0
5	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1100016P	1.0
6	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	0	1100023S	2.0
7	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1100025O	2.0
8	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1100029G	2.0
9	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1100035L	1.0
10	0	1	0	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	1100036J	2.0
11	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1100037H	2.0
12	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	1100041Q	2.0
13	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1100042O	1.0
14	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1100050P	2.0
15	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1100051N	2.0
16	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1100053J	2.0
17	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1100060M	2.0
18	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1100064E	2.0
19	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1100066A	2.0
20	0	0	0	0	0	0	0	0	1	0		0	0	0	0	0	0	0	1100069R	2.0

Table7: Mapping drug names with ATC code and group by column ID

The above table shows the mapping of the drug names with the respective ATC code and grouping it with help of the column ID (Table 7).

Merging of datasets: Based on the primary key (Study ID) all the datasets were inner joined and merged to prepare a final dataset for the analysis.

Data validation: Validation of dataset by checking the formats or type of columns. Prior to the join function all the datasets were transformed into a proper format and structure for successful merging.

ID	Pregnancy outcome	Gestational age	birth weight	Unnamed: 0_y	A01AA	A01AB	A01AC	A01AD	...	V04CH	V06A	V06B	V06D	V06DC	V07AC	V07AY	V09GX	drug count	PT
5100211I	Live birth	41	>= 10th %ile	4422	0	0	0	0	...	0	0	0	0	0	0	0	0	1	2
3200344U	Live birth	40	>= 10th %ile	6226	0	0	0	0	...	0	0	0	0	0	0	0	0	11	2
1100365N	Live birth	39	>= 10th %ile	147	0	0	0	0	...	0	0	0	0	0	0	0	0	3	2
3100233U	Live birth	40	>= 10th %ile	7606	0	0	0	0	...	0	0	0	0	0	0	0	0	8	2
			>=																

Table8: Final dataset after merging

The above table is the final dataset after merging, it contains the study ID, Drugs converted to ATC 3rd level Hierarchy and pregnancy outcomes namely pregnancy outcomes (live birth vs stillbirth), preterm birth (Gestational age) and small for gestational age (Birth weight) (Table 8).

Data cleaning: The dataset is cleaned by removing the nulls from final dataset. The nulls contribute a negligible amount from the total dataset. A total of 2% nulls were identified for the data and removed from the final dataset.

Data preprocessing: Exploratory data analysis is performed to demonstrate the hidden pattern in the data. A dataset is prepared with the total number of drugs a patient consumed and plotted against different pregnancy outcomes.

A01AA	A01AB	A01AC	A01AD	A02A	A02BA	A02BB	A02BC	A02BD	A02BX	...	V04CH	V06A	V06B	V06D	V06DC	V07AC	V07AY	V09GX	study id	drug count
0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1100001F	3
0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1100012A	3
0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1100013V	1
0	1	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	1100014T	8
0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1100015R	9

Table9: Dataset with drug count taken by the pregnant women

The above table contains the dataset showing the count of the drugs taken by the pregnant women during their pregnancy time (Table 9).

Data Visualization: Python packages Matplot and tableau are used for visualization., Bar graph, line chart, bubble chart is visualized.

Bubble chart:

A bubble chart is plotted for most used class of drugs. Drugs related to hypertension, infections, cardiovascular and alimentary are most commonly used when compare to others.

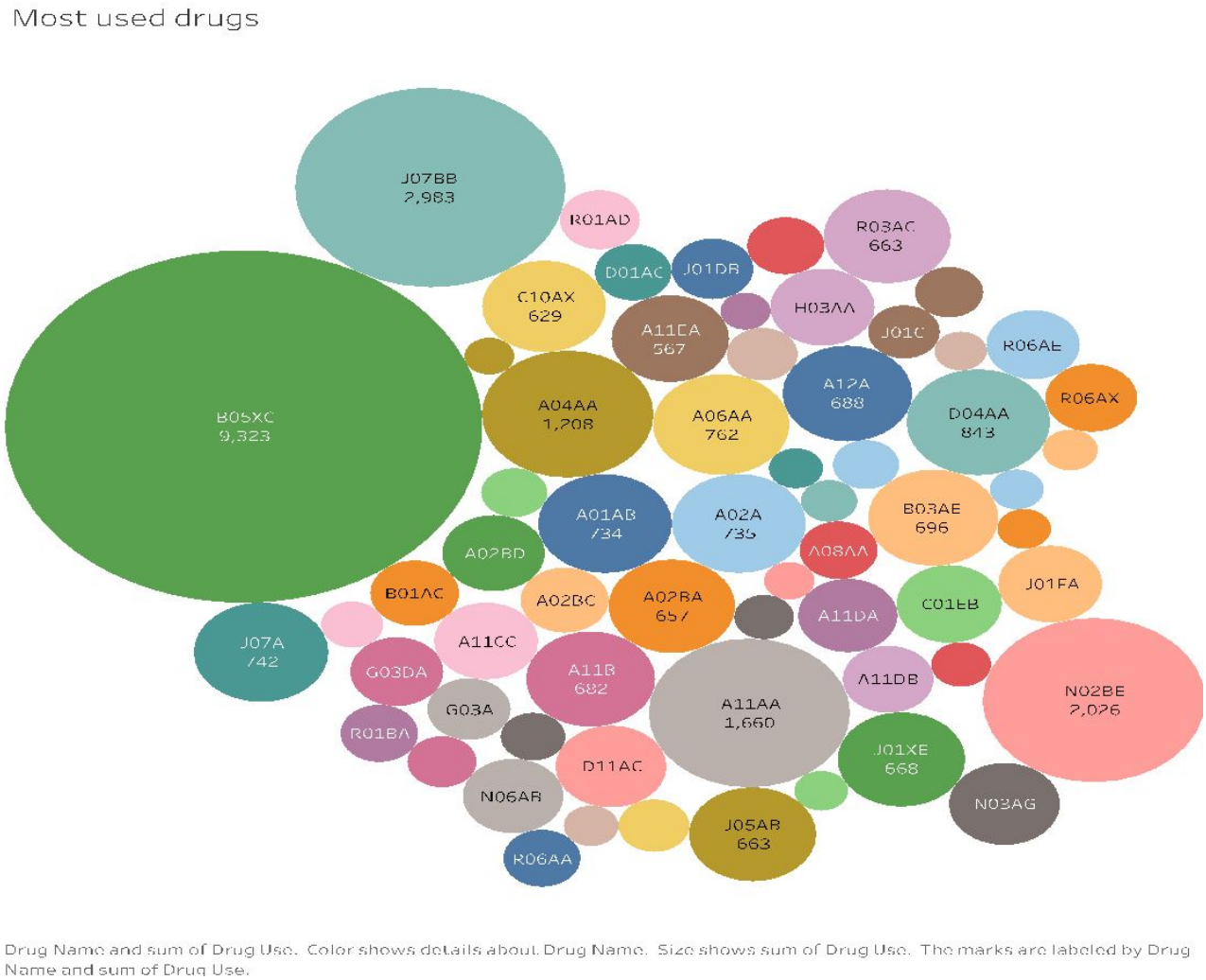


Figure 7: Bubble chart for the most commonly used drug class

Bubble chart: A bubble chart is plotted for most used class of drugs (Figure 7). Drugs related to hypertension, infections, cardiovascular and alimentary are most commonly used when compare to others. Individual drugs with highest use were prenatal vitamins (ATC BO5XC), influenza vaccines (J07BB), neomycin and oxycodone are most consumed more by the pregnancy women.

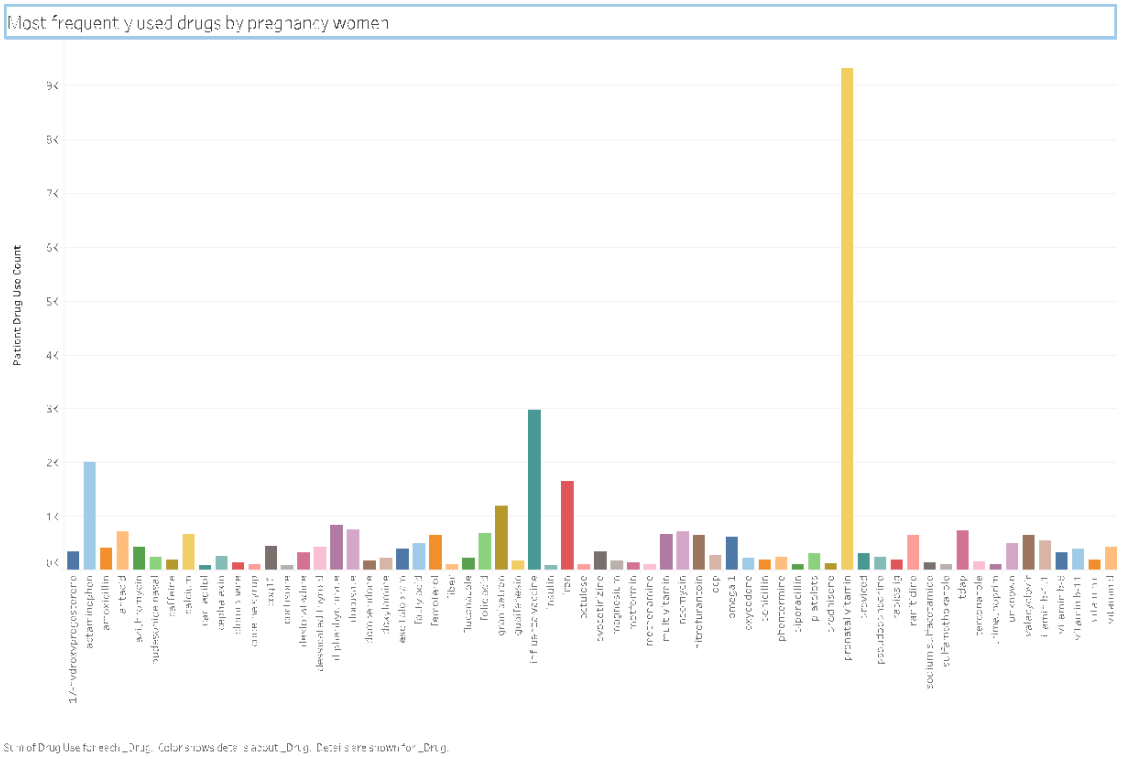


Figure 8: Bar chart for the most commonly used drug class with frequency count

The above figure is the bar graph showing the most commonly used class of drugs. From the above bar graph, it is interpreted that prenatal vitamins, influenza, vaccines, neomycin and oxycodone are most consumed more by the pregnancy women (Figure 8).

Data Imbalance: Data imbalance is defined as unequal distribution of one class against the other. In machine learning imbalance data is the biggest challenge in training the dataset as it

may lead to overfitting. In this study, three datasets were prepared for each outcome and upon investigation it is found that one class count is more than the other. For example, In the dataset with pregnancy outcome as dependent variable, the live birth count is 9398, stillbirth count is 36, proportion of both is 261:1. Building a machine learning model by training this dataset leads to overfitting. To avoid such flaws in the study SMOTE technique is applied to balance the imbalance data. In SMOTE up-sampling is performed for the datasets to increase the minority level class to majority level class. The total dataset size is increased to 18796.

SMOTE for still/live birth:

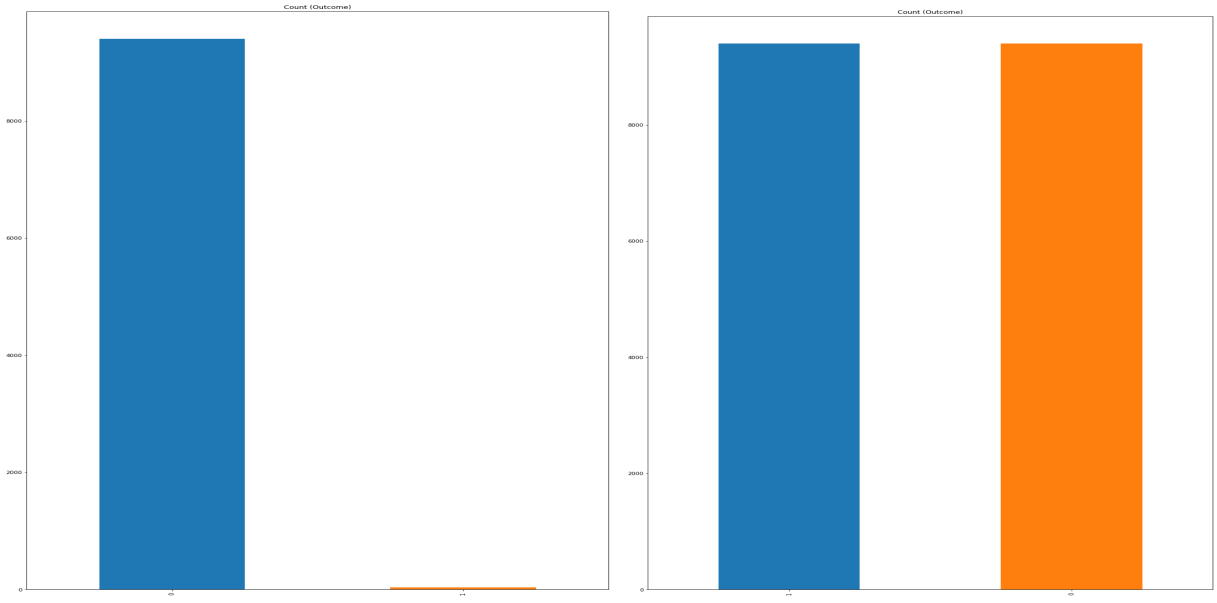


Figure 9: SMOTE technique for Pregnancy outcome (Live birth vs Stillbirth)

Smote technique applied for both live birth and stillbirth samples. After applying the smote technique, it balances both the datasets, i.e. stillbirth data is replicating to balance with live birth (Figure 9).

SMOTE for preterm birth:

The preterm birth dataset is divided into PRETERM AND FULL TERM based on the gestational weeks. If the outcome is greater than >37 the variables are converted into full term and the remaining into preterm. In the dataset with preterm birth outcome as dependent variable, the full term is 8632, preterm count is 802, proportion of both is 10.76: 1. Upon applying SMOTE algorithm the total dataset size is increased to 17264.

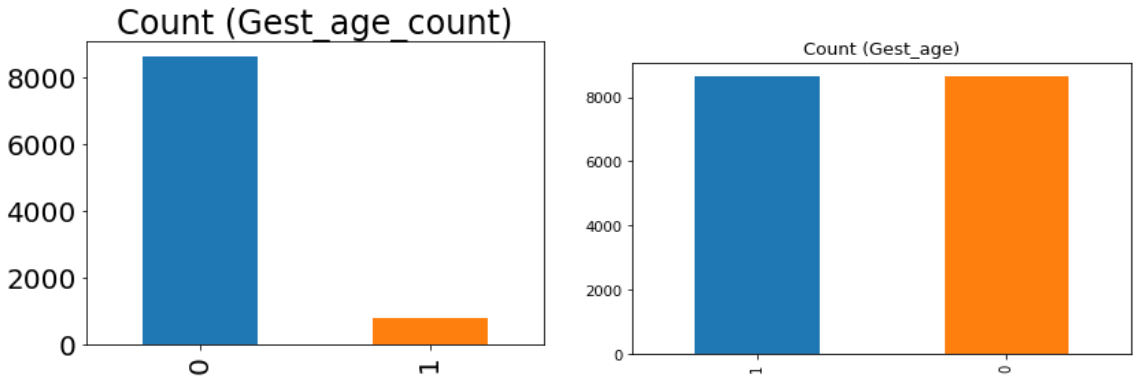


Figure 10: SMOTE technique for preterm birth dataset

The above figure shows the original dataset to the left and to the right it shows the balanced dataset after applying the Smote technique (Figure 10).

SMOTE for Small for gestational age:

In the dataset with small for gestational age outcome as dependent variable, the 5th % ile, 5th to 10th %ile and 10th % ile are divided in different proportions. Upon applying SMOTE algorithm, the total dataset size is increased to 17898.

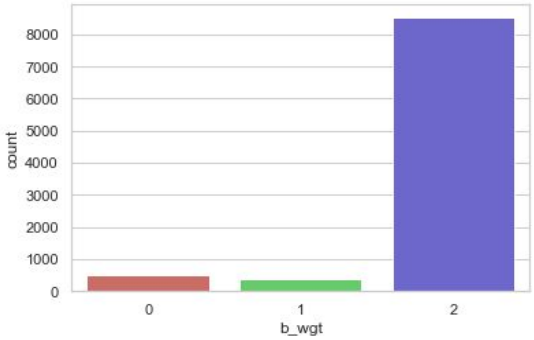


Figure 11: SMOTE technique for small for gestational age dataset

The above figure shows the smote technique used for small for gestational age as we can see the imbalances in the dataset. The three columns in the above figure refers to 5th % ile, 5th to 10th %ile and 10th % ile divided in different proportions (Figure 11).

One Hot encoding:

The process of converting categorical variables into numerical variables is known as One Hot encoding. In the datasets, all the outcome variables are categorical variables namely still birth, live birth, small for gestational age. One Hot encoding helps the conversion of categorical to numeric and improves the prediction of model. This conversion is completed by using label encoder.

```
In [83]: # Create a label encoder object
         from sklearn import preprocessing
         from sklearn.preprocessing import LabelEncoder
         final1["Outcome"] = final1["Outcome"].astype('category')
         final1.dtypes
         final1["Outcome"] = final1["Outcome"].cat.codes
         final1.head()
```

Out[83]:

	Outcome	A01AA	A01AB	A01AC	A01AD	A02A	A02BA	A02BB	A02BC	A02BD	...	V03AB	V04CF	V04CH	V06A	V06B	V06D	V06DC	V07AC	V07AY	V(
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	

Table10: One Hot encoding technique

The above table shows the one hot encoding techniques which means the process of converting categorical to numerical variables. For example, Live birth (categorical variable) is converted to 1(numerical variable) for analysis (Table 10).

Feature extraction:

Feature extraction is the important technique in machine learning, informative and key features that enhance model accuracy can be extracted by this method. Feature extraction helps in reducing overfitting. There are many techniques, but for this study Recursive Feature Elimination, LASSO, Random forest are used for feature selection.

Recursive Feature Elimination: It eliminates the attributes or variables and built a model on the remaining variables. A model accuracy method is used to denote the target variable. The important features are named as true and eliminated features are named as false. The number of features can be either self-defined or a loop is defined to predict the optimum number of features.

LASSO: It is also known as L1 regularization method. For the lost function penalty is added and all the weak features are forced to become zero coefficient. All the non-zeros are the best features and used for model building.

Random Forest: Random forest consists of inbuilt decision tree algorithm with 4-12 trees, each tree is built over randomly distributed variable. Every tree does not see all the features and thus

overfitting can be avoided by such techniques. Each tree is a Boolean based on feature combination. These yields generally good predictive features and with easy interoperability.

Feature extraction for Live/Stillbirth Dataset:

Recursive Feature Elimination: 30 features were extracted

```
#no of features
nof_list=np.arange(1,326)
high_score=0
#Variable to store the optimum features
nof=0
score_list =[]
for n in range(len(nof_list)):
    X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.3, random_state = 0)
    model = LinearRegression()
    rfe = RFE(model,nof_list[n])
    X_train_rfe = rfe.fit_transform(X_train,y_train)
    X_test_rfe = rfe.transform(X_test)
    model.fit(X_train_rfe,y_train)
    score = model.score(X_test_rfe,y_test)
    score_list.append(score)
    if(score>high_score):
        high_score = score
        nof = nof_list[n]
print("Optimum number of features: %d" %nof)
print("Score with %d features: %f" % (nof, high_score))
```

```
Optimum number of features: 30
Score with 30 features: 0.092798
```

Figure 12: Recursive Feature Elimination live/stillBirth

The above figure shows the recursive feature elimination for the pregnancy outcome (Live birth vs stillbirth). It shows about 30 number of optimum features, with a score of 0.092798 (Figure 12)

LASSO: Lasso picked 199 variables and eliminated the other 127 variables

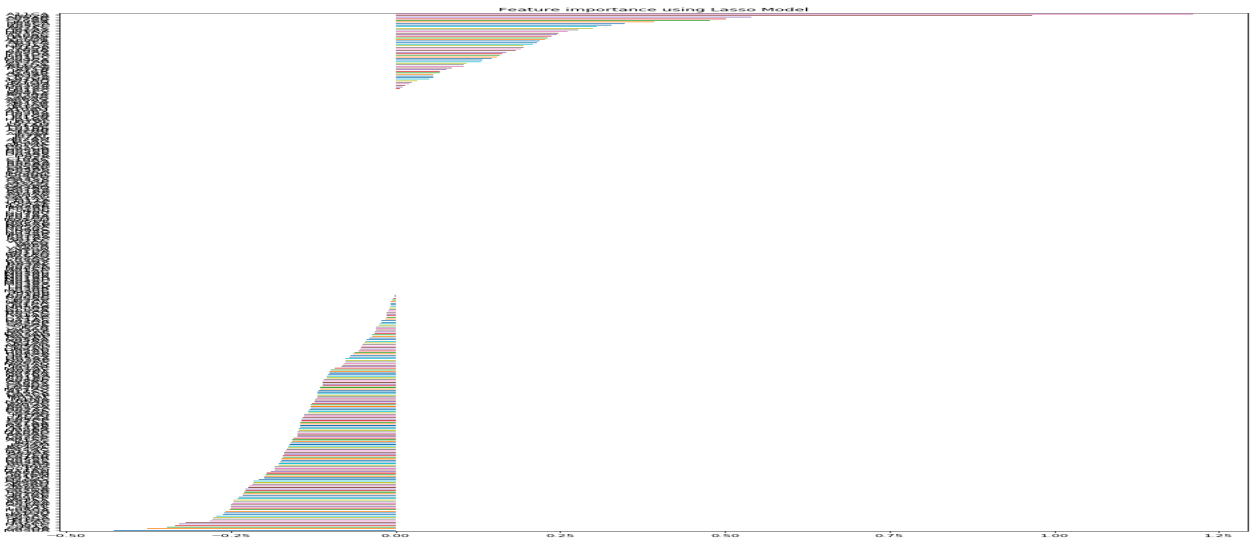


Figure 13: LASSO Feature Extraction live/stillBirth

The above figure shows the LASSO feature extraction for pregnancy outcome (Live birth vs stillbirth) where it helped in picking up 199 variables and eliminating 127 variables. The above variables which are on right corner are having impact on the outcomes and variables which are on left down corner have no impact on the outcomes, so it is eliminating those variables (Figure 13).

Chi-square: chi-square picked 150 variables out of 326 variables

```
In [32]: # chi-square
         from sklearn.feature_selection import SelectKBest
         from sklearn.feature_selection import chi2
         # Select two features with highest chi-squared statistics
         chi2_selector = SelectKBest(chi2, k=150)
         X_kbest = chi2_selector.fit_transform(X, y)
         # Show results
         print('Original number of features:', X.shape[1])
         print('Reduced number of features:', X_kbest.shape[1])

         Original number of features: 326
         Reduced number of features: 150
```

Figure 14: Chi-square Feature Extraction live/stillBirth

The above figure is the Chi-Square method for the pregnancy outcome. This model picked up 150 variables showing impact on the outcomes and eliminated the remaining variables which are showing no impact (Figure 14).

RFE with cross-validation:

In the dataset with pregnancy outcome stillbirth, Recursive Feature elimination is used for feature extraction. Cross validation with Recursive feature elimination is performed for feature selection. The number of features in the final dataset are 327 but the model picked 195 features.

Reason for picking RFE with cross validation: Among various feature selection models RFE with Cross-validation is picked as the cross-validation avoids overfitting of the features and the model built by these features predicts best accuracy.

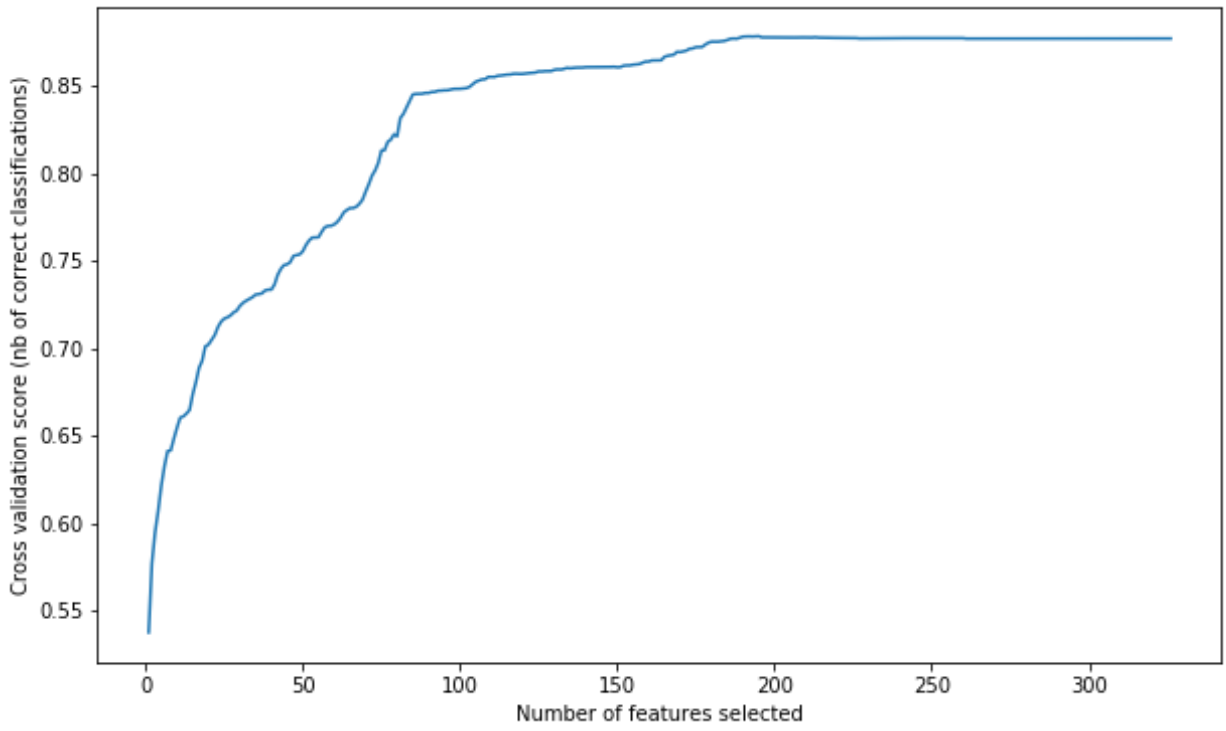


Figure 15: RFE with Cross-validation feature selection live/still Birth

The above figure is the recursive feature elimination with cross validation for live birth vs stillbirth. This method has picked up 195 features out of 327 and has shown the best accuracy among all the other models (Figure 15).

```
Optimal number of features: 195
Selected features: ['A01AB', 'A01AC', 'A02A', 'A02BB', 'A02BC', 'A02BD', 'A03AA', 'A03AB', 'A03AX', 'A03FA', 'A04A',
'A04AA', 'A04AD', 'A05AA', 'A06AA', 'A06AB', 'A06AC', 'A06AD', 'A06AX', 'A07AA', 'A07DA', 'A07EA', 'A07EC', 'A07XA',
'A08AA', 'A10AC', 'A10BA', 'A10BB', 'A10BD', 'A10BH', 'A11AA', 'A11B', 'A11CA', 'A11CC', 'A11DA', 'A11DB', 'A11EA', '
A12A', 'A12CC', 'A12CD', 'A16AX', 'B01AB', 'B01AC', 'B02AA', 'B03AE', 'B05XA', 'B05XB', 'B05XC', 'C01BB', 'C01CA', 'C
01EB', 'C02AB', 'C02DB', 'C03AA', 'C03CA', 'C03DB', 'C04AC', 'C05AA', 'C05AE', 'C05AX', 'C07AA', 'C07AB', 'C07AG', 'C
08CA', 'C08DA', 'C09AA', 'C09CA', 'C10AB', 'C10AX', 'D01AC', 'D01AE', 'D03AA', 'D06AX', 'D06BB', 'D07AA', 'D07AB', 'D
07AC', 'D07AD', 'D08AG', 'D10AD', 'D10AE', 'D10AF', 'D10AX', 'D11AH', 'D11AX', 'G02AC', 'G02CB', 'G02CC', 'G03A', 'G0
3AA', 'G03AC', 'G03CA', 'G03CX', 'G03DA', 'G03GA', 'G03GB', 'G04BX', 'G04CA', 'H02AB', 'H03BB', 'H04AA', 'J01C', 'J01
CA', 'J01CG', 'J01CR', 'J01DB', 'J01DC', 'J01DD', 'J01EA', 'J01EC', 'J01FA', 'J01MA', 'J01XE', 'J01XX', 'J04AC', 'J05
AB', 'J05AF', 'J05AH', 'J05AR', 'J06BB', 'J07A', 'J07AG', 'J07AJ', 'J07AM', 'J07AP', 'J07BB', 'J07BC', 'J07BD', 'J07B
M', 'L01CX', 'L02AE', 'L02BG', 'L03AX', 'L04AB', 'L04AD', 'L04AX', 'M01AC', 'M01AX', 'M03BA', 'M03BX', 'N01AH', 'N02A
', 'N02AC', 'N02AE', 'N02AF', 'N02AX', 'N02BA', 'N02BE', 'N02CA', 'N02CC', 'N03AB', 'N03AD', 'N03AE', 'N03AF', 'N03AX
', 'N05AB', 'N05AH', 'N05AX', 'N05BA', 'N05BB', 'N05CA', 'N05CC', 'N05CF', 'N05CH', 'N06AA', 'N06AB', 'N06AX', 'N06BA
', 'N06BC', 'N07BA', 'N07XX', 'P01AX', 'P01BA', 'P01BB', 'P03AC', 'R01AC', 'R01AD', 'R01BA', 'R03AC', 'R03DC', 'R03DX
', 'R05', 'R05CA', 'R05DA', 'R05DB', 'R05X', 'R06AA', 'R06AB', 'R06AE', 'R06AX', 'S01L', 'V01AA', 'V03AB', 'V04CF', '
V06DC']
```

Figure 16: Features extracted by RFE live/still Birth

The above figure shows the 195 features which are extracted by recursive feature elimination for pregnancy outcome. The about 195 features are the 3rd level hierarchy ATC codes for the drugs used by the pregnant women (Figure 16).

Feature extraction for preterm birth Dataset:

LASSO: 294 features from 326 were identified as top features

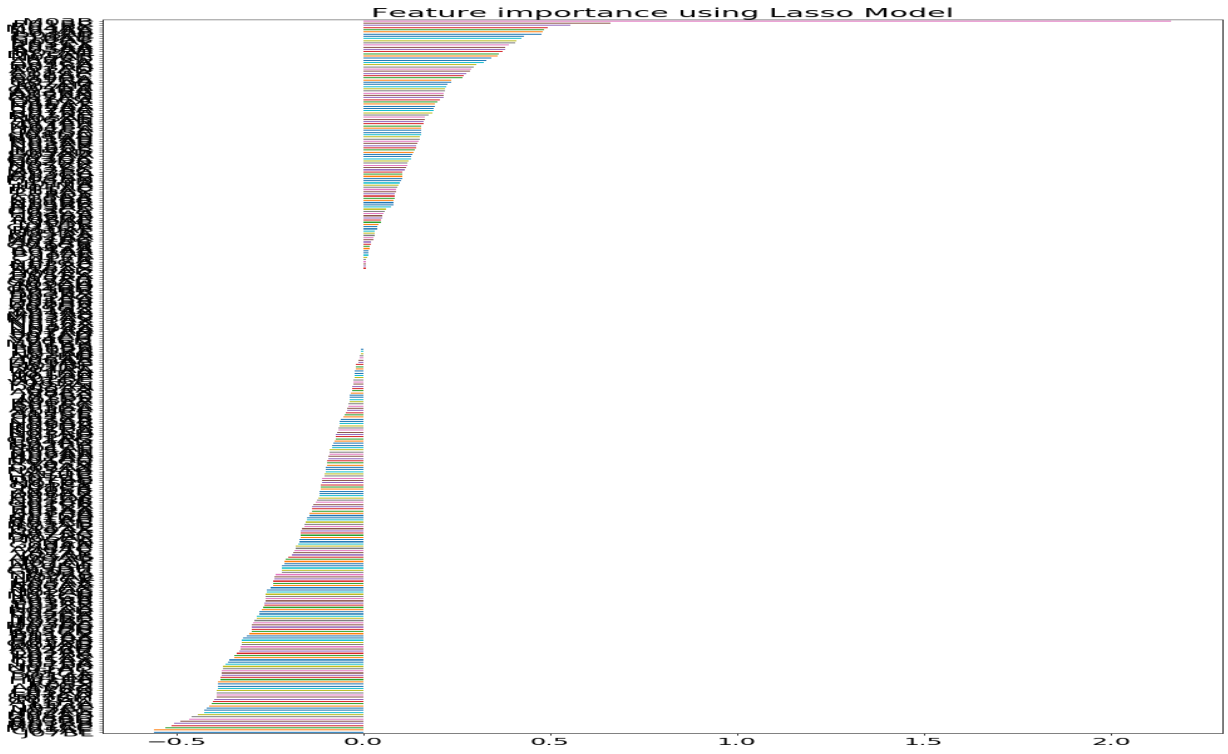


Figure 17: Features extracted by LASSO Preterm birth

The above figure shows the LASSO feature extraction for preterm birth where it helped in picking up 294 variables and eliminating 32 variables. The above variables which are on right corner are having impact on the outcomes and variables which are on left down corner have no impact on the outcomes, so it is eliminating those variables (Figure 17).

RFE with cross-validation: 224 Features were picked from 326

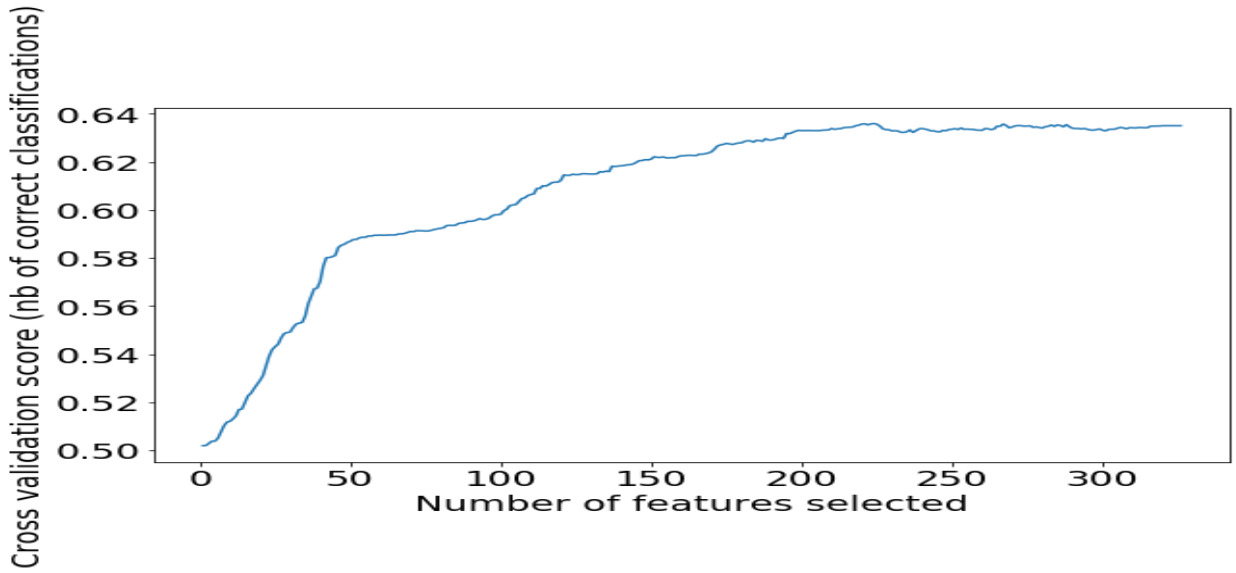


Figure 18: Features extracted by RFE with cross-validation preterm birth

The above figure is the features extracted by the Recursive feature elimination with cross validation for the preterm birth. This model has picked up 224 features from 326 and eliminating the rest (Figure 18).

```
Index(['A01AB', 'A01AC', 'A02A', 'A02BA', 'A02BC', 'A02BD', 'A03FA', 'A04AA',
       'A05AA', 'A06AA', 'A06AB', 'A06AD', 'A07EA', 'A08AA', 'A10AC', 'A10BA',
       'A10BB', 'A11AA', 'A11B', 'A11CC', 'A11DA', 'A11DB', 'A11EA', 'A11GA',
       'A12A', 'A12CC', 'B01AB', 'B01AC', 'B03AE', 'B05XC', 'C01EB', 'C02AB',
       'C07AG', 'C08CA', 'C10AX', 'D01AC', 'D04AA', 'D10AF', 'D11AC', 'G01AG',
       'G02CC', 'G03A', 'G03DA', 'G03GA', 'G03GB', 'H02AB', 'H03AA', 'J01C',
       'J01CA', 'J01DB', 'J01DD', 'J01EA', 'J01FA', 'J01XE', 'J01XX', 'J05AB',
       'J06BB', 'J07A', 'J07AJ', 'J07BB', 'L01CX', 'M03BX', 'N02AA', 'N02BE',
       'N02CC', 'N03AG', 'N03AX', 'N05AB', 'N05BA', 'N06AB', 'N06AX', 'N06BA',
       'N06BC', 'R01AD', 'R01BA', 'R03AC', 'R03DC', 'R05CA', 'R05DA', 'R06AA',
       'R06AE', 'R06AX', 'V06DC'],
      dtype='object')
```

Figure 19: Features extracted by LASSO for RFE with cross-validation

The above figure shows the features extracted by Lasso for RFE with cross validation. The index shows the 3rd level hierarchy ATC codes of the drugs which are used for building the model (Figure 19).

Reason for picking LASSO Regression: Among various feature selection models LASSO is picked for feature selection, LASSO converts all the non-important features to zeros and also the machine learning model built by these features predicts best accuracy.

Feature extraction for small for gestational age Dataset:

RFE with cross-validation:

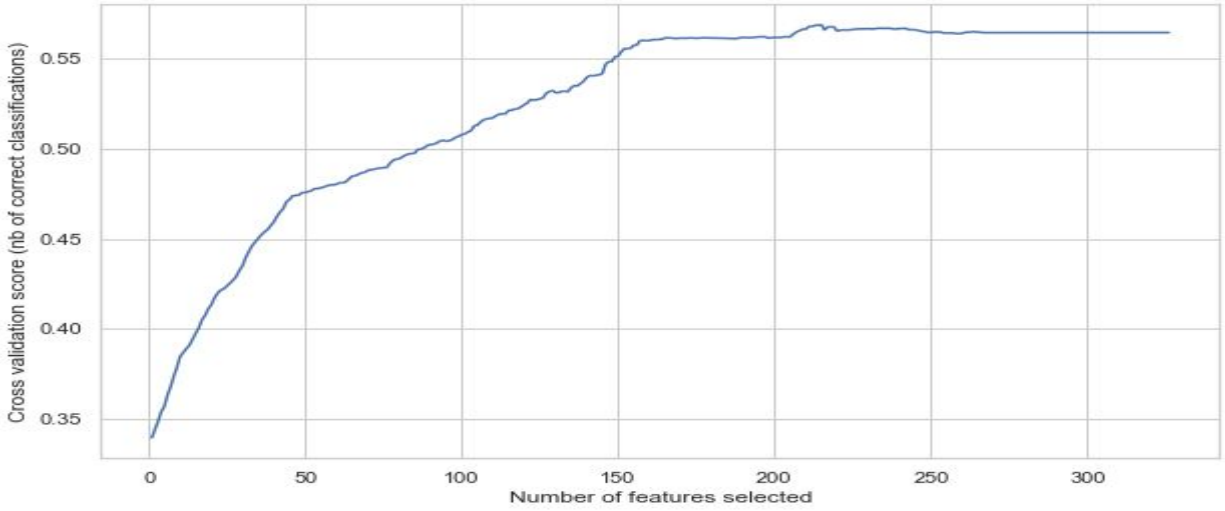


Figure 20: Features extracted by LASSO for RFE with cross-validation

In Birth Weight, Recursive Feature Elimination with cross validation model selects 215 features among which Anilides, anti-infectives and Nitrofuran derivatives are the top medications in the features (Figure 20).

Data Analysis: As the dataset has three outcomes, two machine learning algorithms namely logistic and random forest are performed for individual outcomes.

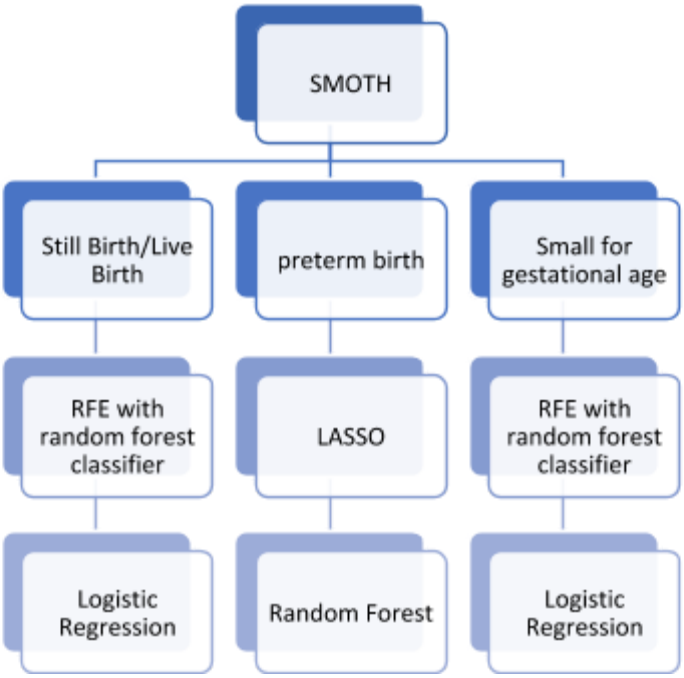


Figure 21: Model for analysis

The above figure describes the feature extraction methods, and models used for analysis for showing the relationship between dependent and independent factors (Figure 21).

Machine learning algorithms in detail:

Logistic regression:

It is a machine learning classification algorithm used for predicting the probability of binary outcome. Logistic Regression is a generalized Linear Regression in which the output is not weighted sum of inputs directly but passed through a function that can map any real value between 0 and 1.

Random Forest:

Random forest comes under supervised learning algorithm used for classification and regression. It has tree inbuilt function which makes the model robust and more flexible. A decision tree is made upon randomly selected samples which has in turn predictions and chooses the best.

Decision Tree Algorithm:

A decision tree is a flowchart-like tree structure where an internal node represents feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning. This flowchart-like structure helps you in decision making. It's visualization like a flowchart diagram which easily mimics the human level thinking. That is why decision trees are easy to understand and interpret.

Support Vector Model:

Support vector machines is a supervised machine learning algorithm that produces high accuracy compared to other regression models. SVM is most commonly used for face detection, classification of mails and recognition of handwritten pattern. The SVM works as a classifier and regression by separating points using a hyperplane with biggest amount of margin.

Machine learning models for Live/Stillbirth pregnancy outcome:

Dataset: The dataset for pregnancy outcome consists of 9434 rows × 327 columns. The columns are ATC 3rd level codes and the data within the columns indicated the drug count. The Outcome is binary variable with live birth as 0 and still birth as 1.

Dependent variable: Pregnancy outcome (live/stillbirth)

Independent variable: Drug with ATC codes

	Outcome	A01AA	A01AB	A01AC	A01AD	A02A	A02BA	A02BB	A02BC	A02BD	...	V03AB	V04CF	V04CH	V06A	V06B	V06D	V06DC	V07AC	V07AY
0	Live birth	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
1	Live birth	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
2	Live birth	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
3	Live birth	0	0	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0
4	Live birth	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
5	Live birth	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
6	Live birth	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
7	Live birth	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
8	Live birth	0	0	0	0	0	0	0	2	0	...	0	0	0	0	0	0	0	0	0
9	Live birth	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0
10	Live birth	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0
11	Live birth	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
12	Live birth	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
13	Stillbirth	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
14	Live birth	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
15	Live birth	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
16	Live birth	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
17	Live birth	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
18	Live birth	0	1	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0

Table 11: Final dataset for pregnancy outcome

The above table shows the final dataset for the pregnancy outcome (Live birth vs stillbirth) (Table 11).

Logistic regression:

In the study, still birth is termed as 0 and live birth is termed as 1. The steps of logistic regression include: splitting of test and train data, initiate the model, evaluate the model using confusion matrix and plotting the Area under curve.

The Pregnancy outcome is considered as dependent variable and the drugs or medications columns considered as independent variable.

```
model = XGBClassifier()
model.fit(X_train[['A01AA']], y_train)
y_pred = model.predict(X_test[['A01AA']])

accuracy = accuracy_score(y_test, y_pred)
print("Accuracy: %.2f%%" % (accuracy * 100.0))
```

```
Accuracy: 99.74%
```

Figure 22: Logistic model accuracy before SMOTE

The above figure shows the logistic model accuracy before showing the Smote technique. The accuracy shown is 99.74% (Figure 22).

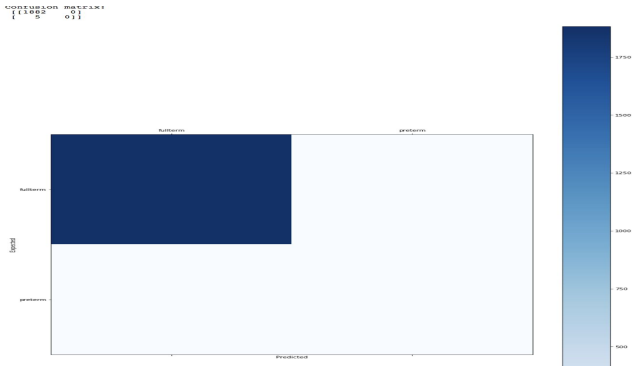


Figure 23: Confusion matrix logistic regression before SMOTE

The above figure shows the confusion matrix logistic regression before using smote techniques. It is not predicting the correct results for live birth and stillbirth (Figure 23).

Random forest:

Random forest is performed as an alternative machine learning model for the pregnancy outcome dataset

```
: from sklearn.ensemble import RandomForestClassifier

#Create a Gaussian Classifier
clf=RandomForestClassifier(n_estimators=100)

#Train the model using the training sets y_pred=clf.predict(X_test)
clf.fit(X_train,y_train)

# prediction on test set
y_pred=clf.predict(X_test)

#Import scikit-learn metrics module for accuracy calculation
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.8462206776715899
```

Figure 24: Random Forest regression after SMOTE

The above figure shows the random forest regression model after using the SMOTE technique. The accuracy shown is 0. 842206 (Figure 24).

Different models were performed to check for the best predicting algorithm. Comparison of different models' accuracy for the outcome live/still birth and compare the final algorithm for the further analysis.

```
t[67]:
```

	MLA Name	MLA Train Accuracy	MLA Test Accuracy	MLA Precission	MLA Recall	MLA AUC
2	ExtraTreesClassifier	0.9226	0.9151	0.854113	1.000000	0.915573
4	RandomForestClassifier	0.9225	0.9149	0.853801	1.000000	0.915362
5	GaussianProcessClassifier	0.9195	0.9127	0.850692	1.000000	0.913246
1	BaggingClassifier	0.9223	0.9115	0.848837	1.000000	0.911976
17	DecisionTreeClassifier	0.9226	0.9108	0.847913	1.000000	0.911342
13	KNeighborsClassifier	0.9078	0.8961	0.846067	0.967038	0.896553

Figure 25: Compare of different models for pregnancy outcome

The above figure compares the different models for the pregnancy outcomes (Live birth vs stillbirth). The table contains the contains the model classifiers, accuracy percentages, precision, recall and AUC values (Figure 25).

Machine learning models for preterm birth outcome:

Dataset: The dataset for preterm birth outcome consists of 9434 rows × 327 columns. The columns are ATC 3rd level codes and the data within the columns indicated the drug count. The Outcome is continuing variable with preterm birth and converted to binary variables as preterm and full-term birth 0 and 1 respectively.

Dependent variable: preterm birth outcome (preterm/full-term)

Independent variable: Drug with ATC codes

	Gest_age	A01AA	A01AB	A01AC	A01AD	A02A	A02BA	A02BB	A02BC	A02BD	...	V03AB	V04CF	V04CH	V06A	V06B	V06D	V06DC	V07AC	V07AY
0	41	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
1	40	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
2	39	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
3	40	0	0	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0
4	41	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
5	39	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
6	41	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
7	40	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
8	41	0	0	0	0	0	0	0	2	0	...	0	0	0	0	0	0	0	0	0
9	41	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0
10	40	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0

Table12: Final dataset preterm birth

The above table is the final dataset for the preterm birth which is ready for model building (Table 12).

In this study, first the preterm birth data is converted into full term and preterm upon 37-week parameter. One hot encoding is used for converting categorical into numerical variables. The preterm birth is considered as dependent variable and the drugs or medications columns considered as independent variable.

	Gest_age	A01AA	A01AB	A01AC	A01AD	A02A	A02BA	A02BB	A02BC	A02BD	...	V03AB	V04CF	V04CH	V06A	V06B	V06D	V06DC	V07AC	V07AY
0	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
5	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
7	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0	2	0	...	0	0	0	0	0	0	0	0	0
9	1	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0
10	1	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0

Table 13: Final dataset preterm birth after one hot encoding

The above table is the final data set for the preterm birth after one hot encoding which helps in converting the categorical variable to numerical variable. For example, Full term birth termed as 1 and preterm birth termed as 0 etc (Table 13).

After feature extraction, the dataset is finalized with the features and different machine learning algorithms are performed to finalize the best algorithm. Below are models with the accuracy precision, recall and Area under curve.

	MLA Name	MLA Train Accuracy	MLA Test Accuracy	MLA Precission	MLA Recall	MLA AUC
2	ExtraTreesClassifier	0.8679	0.8397	0.853224	0.820833	0.839684
4	RandomForestClassifier	0.8671	0.8369	0.844045	0.826852	0.836895
1	BaggingClassifier	0.8665	0.8246	0.824318	0.825463	0.824605
5	GaussianProcessClassifier	0.8410	0.8093	0.828179	0.781019	0.809340
17	DecisionTreeClassifier	0.8679	0.8068	0.798649	0.820833	0.806752
15	NuSVC	0.8177	0.7773	0.810140	0.725000	0.777389
13	KNeighborsClassifier	0.7950	0.7614	0.726544	0.838889	0.761281
3	GradientBoostingClassifier	0.6484	0.6344	0.715556	0.447222	0.634557
0	AdaBoostClassifier	0.6165	0.6126	0.659477	0.467130	0.612739
16	LinearSVC	0.6124	0.6054	0.649836	0.458796	0.605558
14	SVC	0.6081	0.6015	0.712355	0.341667	0.601724
8	RidgeClassifierCV	0.6124	0.6006	0.641192	0.458333	0.600688
6	LogisticRegressionCV	0.6172	0.6006	0.633415	0.479167	0.600669
11	BernoulliNB	0.6045	0.5955	0.638926	0.440741	0.595602
12	GaussianNB	0.5937	0.5860	0.659266	0.357407	0.586171
10	Perceptron	0.5762	0.5711	0.578867	0.525000	0.571173
9	SGDClassifier	0.5758	0.5621	0.546265	0.737963	0.561931
7	PassiveAggressiveClassifier	0.5354	0.5424	0.756233	0.126389	0.542786

Figure 26: Compare of different models for preterm birth

The above figure is the comparison of different models used for the preterm birth. The table contains the columns namely Classifier models, accuracy percentages, precision, recall, AUC values etc (Figure 26).

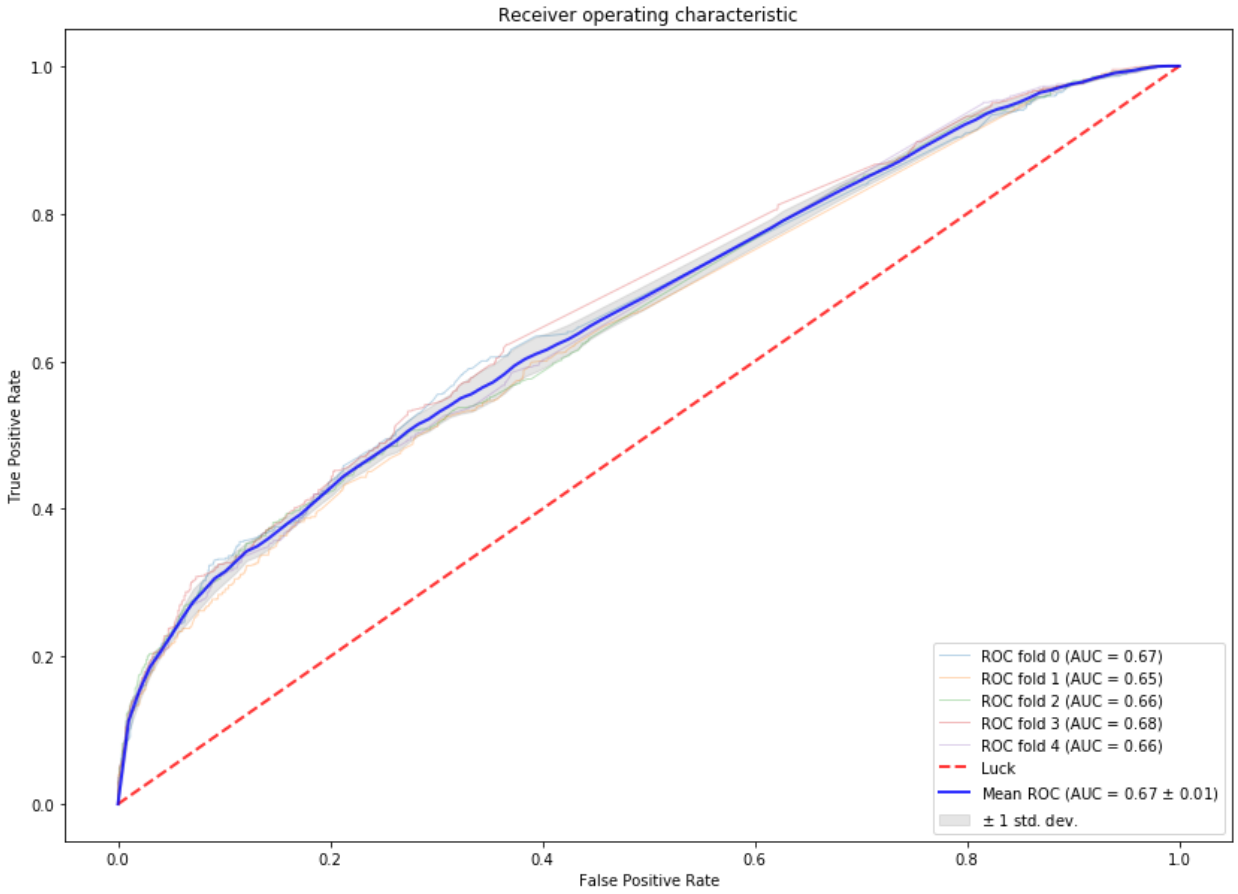


Figure 27: Comparison of ROC for preterm birth Dataset

The above figure shows the comparison of the ROC curve for the preterm birth dataset (Figure 27).

Various models namely SVM, Decision Tree, Logistic regression, Random forest etc., were performed to evaluate the best model.

Machine learning models for small for gestational age outcome:

Dataset: The dataset for small for gestational age outcome consists of 9434 rows × 327 columns. The columns are ATC 3rd level codes and the data within the columns indicated the drug count.

The Outcome is multivariate variable with < 5th %ile as 0, 5th to <10th %ile as 1 and >= 10th %ile as 2.

Dependent variable: small for gestational age outcome (<5, 5 to 10 and >10)

Independent variable: Drug with ATC codes

	b_wgt	A01AA	A01AB	A01AC	A01AD	A02A	A02BA	A02BB	A02BC	A02BD	...	V03AB	V04CF	V04CH	V06A	V06B	V06D	V06DC	V07AC	V07AY
0	>= 10th %ile	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
1	>= 10th %ile	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
2	>= 10th %ile	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
3	>= 10th %ile	0	0	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0
4	>= 10th %ile	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
	>=																			

Table14: Final dataset small for gestational age

The above figure is the final dataset for small for gestational age used for model building (Table 14).

In this study, the small for gestational age dataset consists of three classes namely 5th % ile 5th to 10th % ile and 10th %ile. As there are more than 2 classes in a multiclass/ multinomial regression is performed. One hot encoding is used for converting categorical into numerical variables. The small for gestational age is considered as dependent variable and the drugs or medications columns considered as independent

variable.

	b_wgt	A01AA	A01AB	A01AC	A01AD	A02A	A02BA	A02BB	A02BC	A02BD	...	V03AB	V04CF	V04CH	V06A	V06B	V06D	V06DC	V07AC	V07AY
0	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
1	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
2	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
3	2	0	0	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0
4	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
5	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
6	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
7	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
8	2	0	0	0	0	0	0	0	2	0	...	0	0	0	0	0	0	0	0	0
9	2	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0
10	2	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0
11	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
12	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
13	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
14	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
15	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
16	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
17	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0

Table15: Final dataset birth weight after one hot encoding

The above table is the final dataset for the birth weight after one hot encoding which helps in converting the characteristic variable to numerical variable. Assigning numerical values to 5%, 5-10% and greater than 10%(Table 15).

Results:

Pregnancy outcomes:

Python packages Matplot and tableau are used for visualization., Bar graph, line chart, bubble chart is visualized.

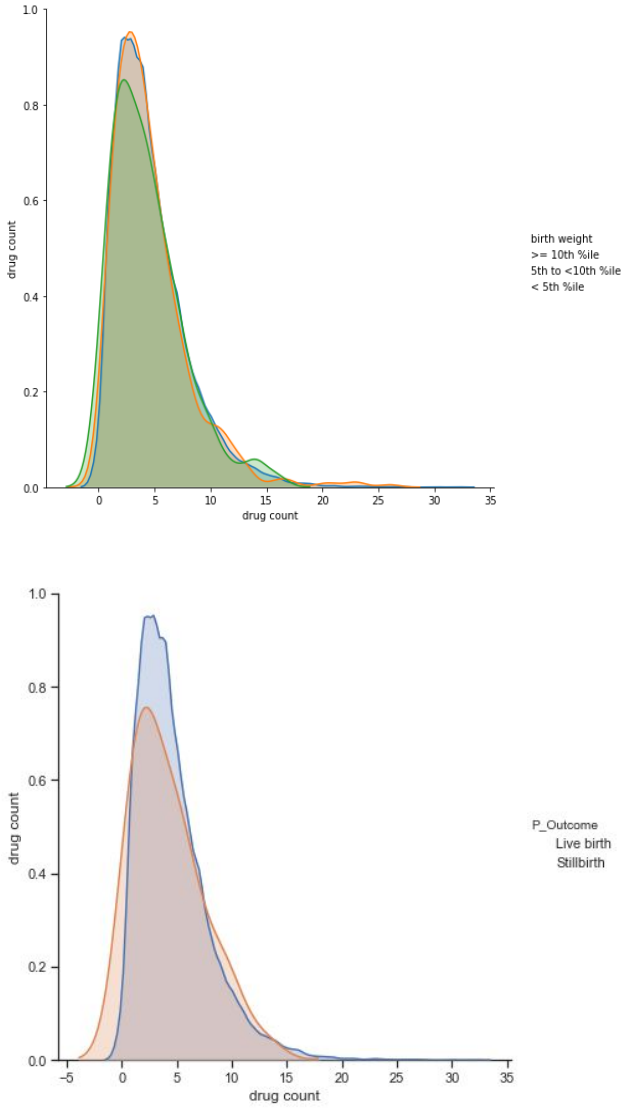


Figure 28: Drug count against small for gestational age and pregnancy outcome

The above figure shows us the drug count against the small for gestational age whose birth weight is greater or equal to 10th percentile, 5^{th} to 10^{th} percentile and less than 5^{th} percentile and drug count for the pregnancy outcome involving live birth and stillbirth(Figure 28).

The prescription pattern i.e., a number of drugs intake the impact on pregnancy outcomes were displayed by using Tableau and Python. The number of drugs (drug count) is taken on the X-axis

and pregnancy probability is plotted on the Y-axis. From the figures, it is estimated that patients who take drugs from 4 to 10 are having highest still and live birth outcomes.

Results of pregnancy outcomes (Live birth vs Stillbirth)

Pregnancy outcome Live/stillbirth:

Logistic regression is performed, and the ROC curve is plotted. The ROC curve is plotted by sensitivity against the specificity, sensitivity is the probability of predicting a true positive to be a positive and specificity is the probability of predicting a true negative to be a positive.

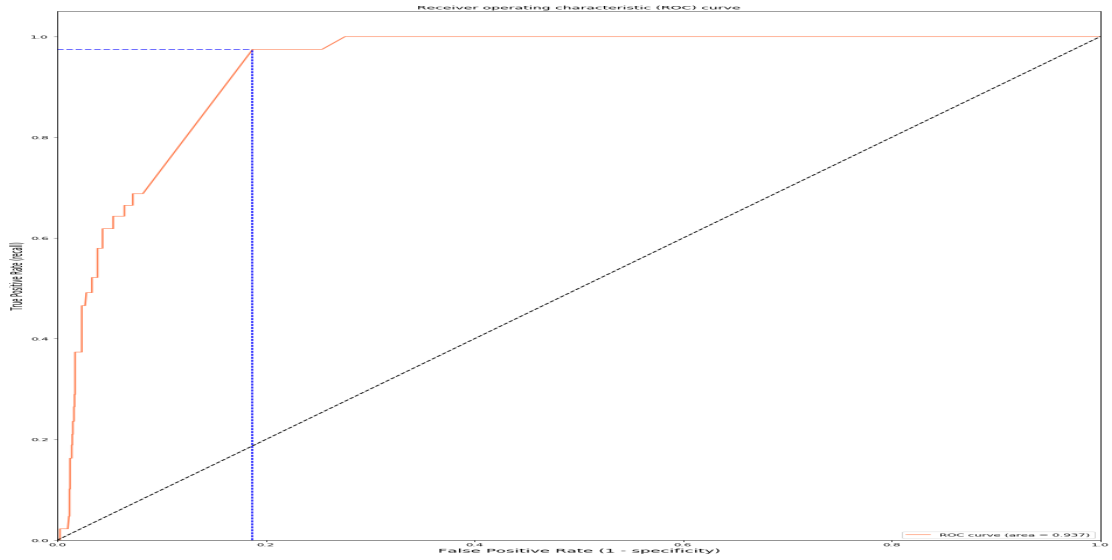


Figure 29: Area under curve for logistic regression

The above figure shows the area under curve for the logistic regression model, this curve on left hand corner indicates that model predicts more accurate and 0.3 log loss shows the good performance of the model (Figure 29).

Area Under Curve- ROC curve which goes closer to the top left-hand corner of the plot indicates that model predicts more accurate. The log loss indicates as the loss function in the logistic regression. The log loss of 0.3 indicates the good performance of the logistic regression model.

```
anged to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)
/anaconda3/lib/python3.6/site-packages/sklearn/linear_model/logistic.py:433: Future
anged to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)

K-fold cross-validation results:
LogisticRegression average accuracy: 0.878 (+/-0.005)
LogisticRegression average log_loss: 0.320 (+/-0.015)
LogisticRegression average auc: 0.933 (+/-0.006)

/anaconda3/lib/python3.6/site-packages/sklearn/linear_model/logistic.py:433: Future
```

Figure 30: Logistic Regression Model Accuracy

The above figure shows the logistic regression model accuracy. It shows 0.933 percent accuracy(Figure 30).

Precision: It is the ratio of true positive to the total predicted positive observations.

$$\text{Precision} = TP/TP+FP$$

Recall: Also known as sensitivity i.e., ratio of predicting real positive observations to all the observations in the actual class.

$$\text{Recall} = TP/TP+FN$$

F-1 Score: It is the weighted average of Precision and Recall. It takes positive and negative into account and it is more important than accuracy.

$$\text{F1 Score} = 2*(\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

	precision	recall	f1-score	support
0	0.96	0.80	0.87	1862
1	0.83	0.97	0.90	1898
micro avg	0.89	0.89	0.89	3760
macro avg	0.90	0.88	0.88	3760
weighted avg	0.90	0.89	0.88	3760

Figure 31: Performance measure of logistic regression

The above figure shows the performance measure of the logistic regression. The table contains the precision values, recall values, f-1 score and support values (Figure 31).

Results of preterm birth:

Pregnancy outcome preterm birth:

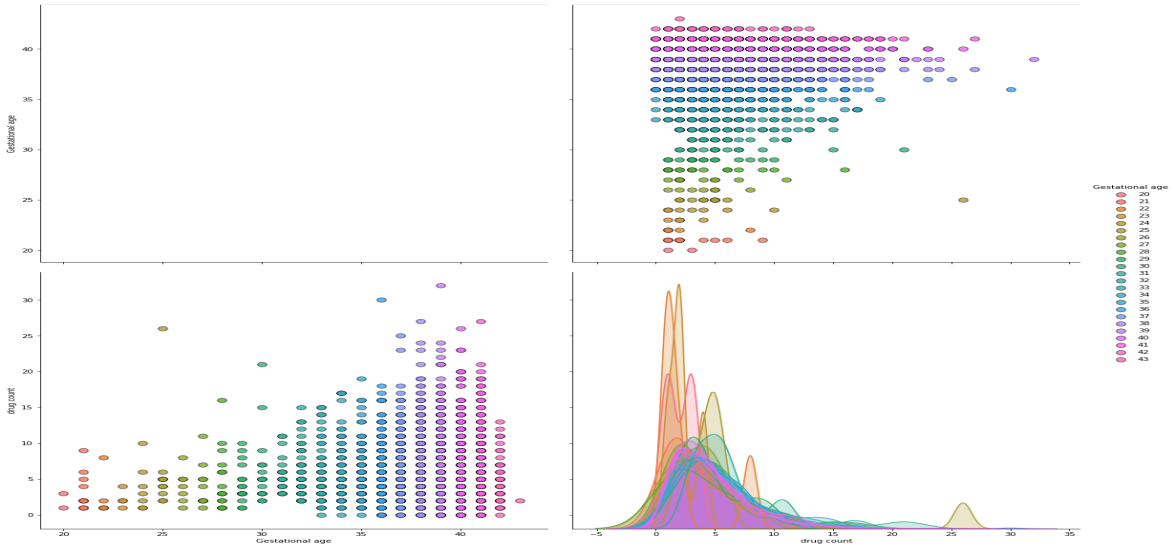


Figure 32: Drug count against preterm birth

The above figure shows the drug count against the preterm birth taking by the pregnant women during their pregnancy. The graphs show the highest number of drugs are taken during 35-40 preterm birth weeks (Figure 32).

In the preterm birth the number of drugs at preterm birth are displayed against the drug count. It has been observed that the greatest number of drugs are consumed between 35 and 40 preterm birth weeks.

	MLA Name	MLA Train Accuracy	MLA Test Accuracy	MLA Precission	MLA Recall	MLA AUC
2	ExtraTreesClassifier	0.8679	0.8397	0.853224	0.820833	0.839684
4	RandomForestClassifier	0.8671	0.8369	0.844045	0.826852	0.836895
1	BaggingClassifier	0.8665	0.8246	0.824318	0.825463	0.824605
5	GaussianProcessClassifier	0.8410	0.8093	0.828179	0.781019	0.809340
17	DecisionTreeClassifier	0.8679	0.8068	0.798649	0.820833	0.806752
15	NuSVC	0.8177	0.7773	0.810140	0.725000	0.777389
13	KNeighborsClassifier	0.7950	0.7614	0.726544	0.838889	0.761281
3	GradientBoostingClassifier	0.6484	0.6344	0.715556	0.447222	0.634557
0	AdaBoostClassifier	0.6165	0.6126	0.659477	0.467130	0.612739
16	LinearSVC	0.6124	0.6054	0.649836	0.458796	0.605558
14	SVC	0.6081	0.6015	0.712355	0.341667	0.601724
8	RidgeClassifierCV	0.6124	0.6006	0.641192	0.458333	0.600688
6	LogisticRegressionCV	0.6172	0.6006	0.633415	0.479167	0.600669
11	BernoulliNB	0.6045	0.5955	0.638926	0.440741	0.595602
12	GaussianNB	0.5937	0.5860	0.659266	0.357407	0.586171
10	Perceptron	0.5762	0.5711	0.578867	0.525000	0.571173
9	SGDClassifier	0.5758	0.5621	0.546265	0.737963	0.561931
7	PassiveAggressiveClassifier	0.5354	0.5424	0.756233	0.126389	0.542786

Figure 33: Performance measure of random forest

The above figure shows the performance measure of random forest model for the preterm birth. The above figure contains the classifier model name, its accuracy, precision, recall values and AUC values for preterm birth. Random forest shows the highest accuracy among all the other models (Figure 33).

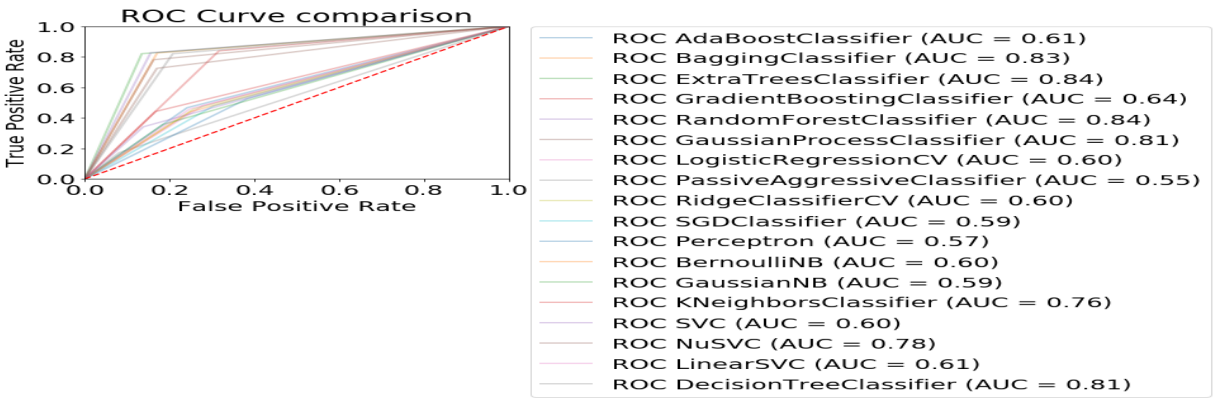


Figure 34: Comparison of area under curve for preterm birth

The above figure shows the comparison of the area under curve for the preterm birth outcome (Figure 34).

Results of small for gestational age outcome:

Pregnancy outcome small for gestational age:

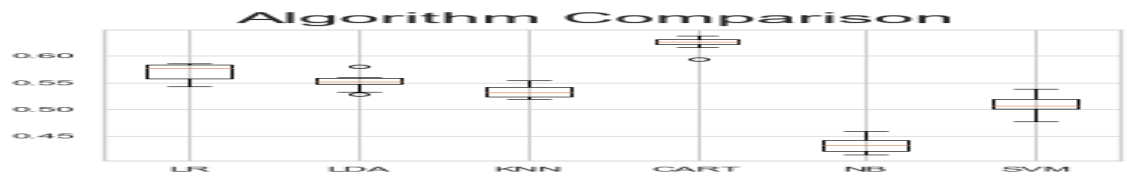


Figure 35: Comparison of Various algorithms for small for gestational age

The above figure shows the comparison of various algorithm models which are used for small for gestational age. Among all the models used CART shows the highest accuracy and it is selected for building the predictive model (Figure 35).

```
0.6153631284916201
[[675 532  28]
 [195 937  30]
 [241 351 591]]
              precision    recall  f1-score   support

           0       0.61      0.55      0.58      1235
           1       0.51      0.81      0.63      1162
           2       0.91      0.50      0.65      1183

   micro avg       0.62      0.62      0.62      3580
   macro avg       0.68      0.62      0.62      3580
weighted avg       0.68      0.62      0.62      3580
```

Figure 36: Performance measures for small for gestational age

The above figure shows the performance measures for small for gestational age. The figure contains the precision values, recall and f-1 score values and support values for the small for gestational age outcome (Figure 36).

Dataset	Feature selection	Model algorithm	Reason
Pregnancy outcome (live/still)	RFE with cross-validation (193 features)	Logistic Regression	More accuracy
preterm birth (preterm/full-term)	LASSO (125 features)	Random forest	Robust model with more precision and recall

small for gestational age	RFE with cross-validation (215 features)	Multinomial regression (CART)	High accuracy and performance measures

Table16: Evaluation of final models for different datasets

The above described following feature selection methods were used for the feature selection, logistic regression, random forest and multinomial regression (CART) is used respectively for pregnancy outcome, preterm birth and small for gestational age (Table 16).

Discussion:

Upon various machine learning algorithms approach, algorithms with the highest accuracy are further considered for building the models. In features extraction process for live/stillbirth outcomes, from fig 17-193 extracted drugs are considered for further model building as they interpret with the highest accuracy. The RFE with cross-validation eliminates the overfitting of the features by dividing the data into test and train. In gestation age, dataset LASSO regression is the choice of algorithm for feature extraction as per fig 19 125 features are extracted. For the dataset with small for gestational age, Recursive Feature Elimination with cross validation algorithm is finalized as it extracts 215 drugs out of 326. In machine learning algorithms for pregnancy outcomes, logistic regression predicts with the highest accuracy of 93.3% as shown in fig 27 when compared to random forest regression 89%. The precision recall and F-1 scores for the logistic regression model are 96%, 97%, and 90% respectively. Upon comparison of various models for preterm birth dataset Random forest classifier is considered as the best algorithm with

an accuracy of 86.7% followed by extra trees classifier 86.79%. The precision and recall for the random forest model are 84% and 82% respectively which denotes the true positive and true negative prediction of our model. Even though the extra tree has more accuracy random forest is chosen due to its robust technique. Figure 37 indicated the various ROC curves of different algorithms plotted against sensitivity and specificity. The third dataset with the outcome of small for gestational age s multinomial regression is performed for the three classes. Figure 35 compares different machine learning models and their accuracy in a box plot diagram. CART predicts more accuracy of 62% when compared to logistic, K-Nearest Neighbor and Support Vector Machine. The precision recall and F-1 score for the CART model are 61%, 81%, and 95% respectively. In Live/Stillbirth, Recursive Feature Elimination with cross-validation Anilides, anti-infectives and Antihistamines are the top medications in the feature extraction. In small for gestational age, Recursive Feature Elimination with cross-validation model predicts features Anilides, anti-infectives, and Nitrofuran derivatives are the top medications. In preterm birth, least absolute shrinkage and selection operator model predicts Anilides, anti-infectives, and Iron in other combinations are the top medications. Mediations of antiepileptic, Anti-Bacterial, Analgesics and Antipyretics, Lipid modifying agents, Stomatological preparations and Cardiovascular drugs have an impact on pregnancy outcomes. Pregnancy women drug intake pattern varies from 4 to 10 with the highest still and live birth outcomes. Among the population who takes 4 to 10 drugs live birth and small for gestational age of >=10% are more prevalent than others. From the above bar graph, in figure 16 it is interpreted that prenatal vitamins, influenza vaccines, neomycin, and oxycodone are most consumed more by the pregnant women.

Limitations:

Small dataset size is the biggest challenge for the study. Due to the small sample, size data is not distributed and biased towards one outcome. The imbalance is data enables all the models to overfit and created a bias in the prediction. Techniques to make the data balance took more time in the analysis part. The time span of the study is less than 4 months which is the greatest challenge to put forward and execute the study plan. Lack of documented data and dictionary for ATC codes during mapping frenzied more time in building a dictionary.

Conclusion:

As per the model most common class of drugs that cause pregnancy outcomes are anti-infective, nervous system drugs benzodiazepines, antidepressants, antineoplastic agents, etc., The number of drug intake is on average of 2 to five in stillbirth and live birth. Research identifies the medications and their impact on the various pregnancy outcomes. Live/Stillbirth logistic model accuracy predicts 93.5% accurate. Multinomial regression approach for small for gestational age yields the best accuracy for CART regression i.e., 62%. Random Forest regression model predicts 86.7% accuracy for preterm birth outcome whether the medications yield preterm/full-term birth.

Future direction: The results are displayed in the drug class level within ATC codes, in future research can be made on the specific drugs within the class to identify the medication which has more impact on the pregnancy outcome. The model can be deployed in the Electronic Health Record or in a mobile device to create an alert system for the physician in suggesting medication for the patient and acts as evidence-based medicine. Further research by the addition of new

variables namely demographics like age, socioeconomic factors namely income, education, associated health conditions would make the model more realistic in the real world.

References:

- Bérard, A., Sheehy, O., Girard, S., Zhao, J. P., & Bernatsky, S. (2018). Risk of preterm birth following late pregnancy exposure to NSAIDs or COX-2 inhibitors. Pain, 159(5), 948-955.
- Bernard, N., Forest, J. C., Tarabulsy, G. M., Bujold, E., Bouvier, D., & Giguère, Y. (2019). Use of antidepressants and anxiolytics in early pregnancy and the risk of preeclampsia and gestational hypertension: a prospective study. BMC pregnancy and childbirth, 19(1), 146. doi:10.1186/s12884-019-2285-8
- Chaturica Athukorala, Alice R Rumbold, Kristyn J Willson and Caroline A Crowther, The risk of adverse pregnancy outcomes in women who are overweight or obese, BMC Pregnancy and Childbirth201010:56, https://doi.org/10.1186/1471-2393-10-56
- Crowther, C. A., Hiller, J. E., Pridmore, B., Bryce, R., Duggan, P., Hague, W. M., & Robinson, J. S. (1999). Calcium Supplementation In Nulliparous Women For The Prevention Of Pregnancy‐Induced Hypertension, Preeclampsia And Preterm Birth: An Australian Randomized Trial. Australian and New Zealand journal of obstetrics and gynaecology, 39(1), 12-18.
- Dr. Jagan Kumar Baskaradoss,1, * Amrita Geevarghese,1 and Abdullah Al Farraj Al Dosari2 Causes of Adverse Pregnancy Outcomes and the Role of Maternal Periodontal

Status – A Review of the Literature, Open Dent J. 2012; 6: 79–84. Published online 2012 May 9. doi: 10.2174/1874210601206010079

- https://americanpregnancy.org/pregnancy-complications/fetal-growth-restriction/
- https://www.nichd.nih.gov/research/supported/nuMoM2b
- Fitton, C. A., Steiner, M. F., Aucott, L., Pell, J. P., Mackay, D. F., Fleming, M., & McLay, J. S. (2017). In-utero exposure to antihypertensive medication and neonatal and child health outcomes: a systematic review. Journal of hypertension, 35(11), 2123.
- Fuchs, F., Monet, B., Ducruet, T., Chaillet, N., & Audibert, F. (2018). Effect of maternal age on the risk of preterm birth: A large cohort study. PloS one, 13(1), e0191002.
- Giraldo, P. C., Araújo, E. D., Junior, J. E., Amaral, R. L. G. D., Passos, M. R., & Gonçalves, A. K. (2012). The prevalence of urogenital infections in pregnant women experiencing preterm and full-term labor. Infectious diseases in obstetrics and gynecology, 2012.
- Grote, N. K., Bridge, J. A., Gavin, A. R., Melville, J. L., Iyengar, S., & Katon, W. J. (2010). A meta-analysis of depression during pregnancy and the risk of preterm birth, low birth weight, and intrauterine growth restriction. Archives of general psychiatry, 67(10), 1012- 1024.
- Gyamfi-Bannerman, C., Thom, E. A., Blackwell, S. C., Tita, A. T., Reddy, U. M., Saade, G. R., ... & Chien, E. K. (2016). Antenatal betamethasone for women at risk for late preterm delivery. New England Journal of Medicine, 374(14), 1311-1320.
- Haas DM1, Parker CB2, Wing DA3, Parry S4, Grobman WA5, Mercer BM6, A description of the methods of the Nulliparous Pregnancy Outcomes Study: monitoring

mothers-to-be (nuMoM2b) Am J Obstet Gynecol. 2015 Apr;212(4): 539.e1-539.e24. doi: 10.1016/j.ajog.2015.01.019. Epub 2015 Jan 31.

- Hendrick, V., Smith, L. M., Suri, R., Hwang, S., Haynes, D., & Altshuler, L. (2003). Birth outcomes after prenatal exposure to antidepressant medication. American journal of obstetrics and gynecology, 188(3), 812-815.
- Hernandez, R. K., Werler, M. M., Romitti, P., Sun, L., Anderka, M., & Study, N. B. D. P. (2012). Nonsteroidal antiinflammatory drug use among women and the risk of birth defects. American journal of obstetrics and gynecology, 206(3), 228-e1.
- Huybrechts, K. F., Sanghani, R. S., Avorn, J., & Urato, A. C. (2014). Preterm birth and antidepressant medication use during pregnancy: a systematic review and meta-analysis. PLoS One, 9(3), e92778.
- Iams, J. D., Romero, R., Culhane, J. F., & Goldenberg, R. L. (2008). Primary, secondary, and tertiary interventions to reduce the morbidity and mortality of preterm birth. The lancet, 371(9607), 164-175.
- Karakosta, P., Alegakis, D., Georgiou, V., Roumeliotaki, T., Fthenou, E., Vassilaki, M., ... & Chatzi, L. (2012). Thyroid dysfunction and autoantibodies in early pregnancy are associated with increased risk of gestational diabetes and adverse birth outcomes. The Journal of Clinical Endocrinology & Metabolism, 97(12), 4464-4472.
- Keirse, M. J. (1995). New perspectives for the effective treatment of preterm labor. American journal of obstetrics and gynecology, 173(2), 618-628.

- Köck, K., Köck, F., Klein, K., Bancher-Todesca, D., & Helmer, H. (2010). Diabetes mellitus and the risk of preterm birth with regard to the risk of spontaneous preterm birth. The Journal of Maternal-Fetal & Neonatal Medicine, 23(9), 1004-1008.
- Kramer MS, Zhang X, Platt RW, Analyzing risks of adverse pregnancy outcomes, Am J Epidemiol. 2014 Feb 1;179(3):361-7. doi: 10.1093/aje/kwt285. Epub 2013 Nov 27
- Martin, J. A. (2018, June 13). National Center for Health Statistics. Retrieved from https://www.cdc.gov/nchs/products/databriefs/db312.htm
- Pastore LM1, Hertz-Picciotto I, Beaumont JJ, Risk of stillbirth from medications, illnesses and medical procedures, Paediatr Perinat Epidemiol. 1999 Oct;13(4):421-30
- Peltoniemi, O. M., Kari, M. A., Tammela, O., Lehtonen, L., Marttila, R., Halmesmäki, E., ... & Hallman, M. (2007). Randomized trial of a single repeat dose of prenatal betamethasone treatment in imminent preterm birth. Pediatrics, 119(2), 290-298.
- Plauborg, A. V., Hansen, A. V., & Garne, E. (2016). Use of azathioprine and corticosteroids during pregnancy and birth outcome in women diagnosed with inflammatory bowel disease. Birth Defects Research Part A: Clinical and Molecular Teratology, 106(6), 494-499.
- Premkumar, A., Baer, R. J., Jelliffe-Pawlowski, L. L., & Norton, M. E. (2018). Hypertensive Disorders of Pregnancy and Preterm Birth Rates among Black Women. American journal of perinatology.
- Raffi ER[1], Nonacs R[2], Cohen LS[2] , Safety of Psychotropic Medications During Pregnancy, Clin Perinatol. 2019 Jun;46(2):215-234. doi: 10.1016/j.clp.2019.02.004. Epub 2019 Mar 28.

- Shim, L., Eslick, G. D., Simring, A. A., Murray, H., & Weltman, M. D. (2011). The effects of azathioprine on birth outcomes in women with inflammatory bowel disease (IBD). Journal of Crohn's and Colitis, 5(3), 234-238.
- Verma, I., Avasthi, K., & Berry, V. (2014). Urogenital Infections as a risk factor for preterm labor: A hospital-based case–control study. The Journal of Obstetrics and Gynecology of India, 64(4), 274-278.
- Viral M. Patel MD Robert A. Schwartz MD, MPH, DSc (Hon), FRCP Edin, Topical antibiotics in pregnancy: a review of safety profiles, 25 April 2019 https://doi.org/10.1111/dth.12951
- Villar, J., Gulmezoglu, A. M., & de Onis, M. (1998). Nutritional and antimicrobial interventions to prevent preterm birth: an overview of randomized controlled trials. Obstetrical & gynecological survey, 53(9), 575-585.
- Wagura, P., Wasunna, A., Laving, A., & Wamalwa, D. (2018). Prevalence and factors associated with preterm birth at kenyatta national hospital. BMC pregnancy and childbirth, 18(1), 107.
- Watts, D. H., Williams, P. L., Kacanek, D., Griner, R., Rich, K., Hazra, R., ... & Pediatric HIV/AIDS Cohort Study. (2012). Combination antiretroviral use and preterm birth. The Journal of infectious diseases, 207(4), 612-621.
- Xie, R. H., Guo, Y., Krewski, D., Mattison, D., Walker, M. C., Nerenberg, K., & Wen, S. W. (2014). Beta‐Blockers increase the risk of being born small for gestational age or of being institutionalised during infancy. BJOG: An International Journal of Obstetrics & Gynaecology, 121(9), 1090-1096.

Poster:

SCHOOL OF INFORMATICS AND COMPUTING
INDIANA UNIVERSITY
IUPUI

Impact of Medication Use on Pregnancy Outcomes

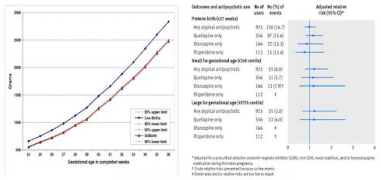
Ashritha Dandam, Josette Jones PhD. RN, Dr. Sara Kay Quinney, PhD

Introduction

- The use of medications among pregnancy women is increased from 2010 to 2017.
- Most of the clinical trials studies are less focused on determining the effect of medications on pregnant women rather inclined towards the general population.

Background

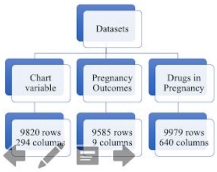
- As per the research, one in ten born babies are affected with preterm birth which is one of the major concerns for the physicians and it accounts for about 15 million per year.
- The main perinatal outcomes assessed within the unmatched cohort, event rates were higher among antipsychotic users than non-users for preterm birth (14.8% v 10.3%) and birth weight >97th centile (3.7% v 2.6%).
- According to the study by Berard women's exposed to C0X-2 inhibitors and NSAIDs during pregnancy is associated with 1.65-fold of risk of PTB.
- According to the Substance Abuse and Mental Health Services Administration's National Survey on Drug Use and Health, 5.4 percent of pregnant women between ages 18-44 had used OTC drugs during their first trimester, 4.8 percent in their second trimester, and 2.4 percent in the last trimester of pregnancy.



Objectives

- The main aim of the study is to identify the impact of prescription drugs on the pregnancy outcomes especially conditions like birth weight, stillbirth, live birth, and gestational age.
- The purpose of the project is to tell with confidence that a patient pregnancy outcome is associated with defects or non- defect.

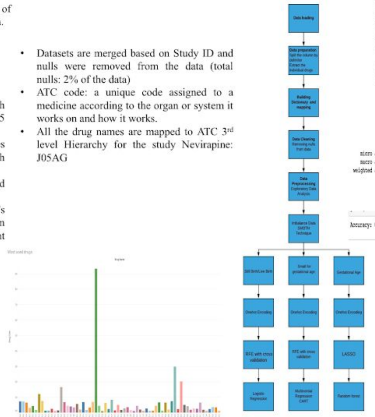
Methodology



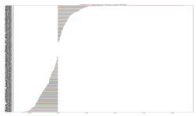
Inclusion criteria:
Pregnancy women with outcomes associated with weight, birth and gestational weeks

Exclusion Criteria:
Vaccines and vitamin medications are removed from the final dataset

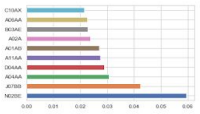
- Datasets are merged based on Study ID and nulls were removed from the data (total nulls: 2% of the data)
- ATC code: a unique code assigned to a medicine according to the organ or system it works on and how it works.
- All the drug names are mapped to ATC 3rd level Hierarchy for the study Nevirapine: J05AG



Results

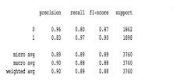


Gestational Age: LASSO(Optimal number of features: 125)

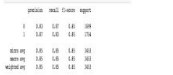


Top 10 Features for Gestational age

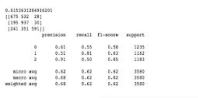
Logistic Regression(Live/Still Birth)

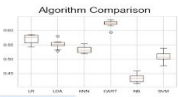


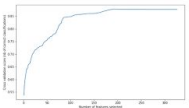
Random Forest(Gestational Age)



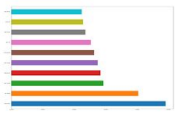
Multinomial Regression(Birth Weight)



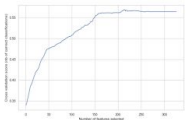




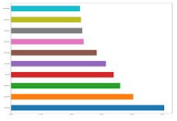
Live Birth/Still Birth: RFE(Optimal number of features: 193)



Top 10 Features for Live/still birth



Birth Weight: RFE (Optimal number of features: 215)



Top 10 Features for Birth weight

Discussion

- Mediations of antiepileptic, Anti-Bacterial, Analgesics and Antipyretics, Lipid modifying agents, Stomatological preparations and Cardiovascular drugs have impact on pregnancy outcomes
- Live/Still birth **logistic model accuracy** predicts 93.5% accurate
- Multinomial regression approach for birth weight yields best accuracy for CART regression i.e., 62%
- Random Forest regression model predicts 85% accuracy for gestational age outcome whether the medications yield preterm/full term birth
- Pregnancy women drug intake pattern varies from 4 to 10 with highest still and live birth outcomes

Conclusion

- Research identifies the medications and its impact on the various pregnancy outcomes
- Development of a clinical decision support system that contains information on safety and secure rating of drugs in pregnancy and to progress clinical outcomes
- Further research on the individual drugs with ATC 5 level hierarchy can be performed

References

- Bérard, A., Sheehy, O., Girard, S., Zhao, J. P., & Bernatsky, S. (2018). Risk of preterm birth following late pregnancy exposure to NSAIDs or COX-2 inhibitors. Pain, 159(5), 948-955.

CV:

ASHRITHA DANDAM

714K Blake St, Indianapolis, IN 46202
Phone: +1 (315) 527 0143 E-mail: ashritha27596@gmail.com

Education:

Master of Science in Health Informatics(Pursuing) Aug2017-May2019
Indiana University Purdue University Indianapolis- GPA: 3.87/4.0

Bachelor of Pharmacy June2013-June2017
Jawaharlal Nehru Technological University- GPA: 3.8/4.0

Academic Projects:

CAPSTONE: Impact of Polypharmacy on Neonatal

Tools: SQL server, R, Python and Tableau

- Planned, designed, and implemented application database code objects, such as stored procedures and views
- Build and maintain SQL scripts, indexes, and complex queries for data analysis and extraction
- Developed Supervised Repressor Model and analysed data to obtain important features, devised data using train_test_split, evaluated R2 to calculate coefficient of determination
- Statistical Analysis like chi-square, Wilcoxon-Mann-Whitney test, ANOVA, REG, Linear Regression performed using Python.
- Visualizing the data using graphs and tables in Tableau, Power BI while creating infographics to better understand the reports for Non-Technical background team and reporting the results through documentation and presentations(ppt).

Analysis of Health Factors Effects on Health Outcomes in the United States

- Extraction, transformation and loading (ETL) of data by using SQL
- Designed, developed, and maintained complex **SQL queries for data analysis** and data extraction as per project requirements
- Created **joins and sub-queries** for complex queries involving multiple tables.
- Data Modeling (logical modeling), **Entity relationship diagram** (ERD) for the tables
- Scripts to increase departmental efficiency and automate repeatable tasks.
- Mapping of data from ZIPCODE to FIPCODE
- Python to make the unstructured data to structure data by **One Hot encoding** for categorical variables
- Scaling features- **iScale function** to reduce multicollinearity between the variables
- Data analysis by performing predictive analysis

Python: Machine learning technique: Recursive feature elimination, Linear regression

- **Performance measure:** R^2, MSE, and RMSE
- Data visualization – **Tableau, D3.js** for summary reports using statistical tools, charts, and graphs
- Python libraries: **Pandas, Numpy, Pylab, SKLearn, Seaborn, and Matplolib** for data analysis

Comparative analysis of various risk factors for Cervical Cancer

- Data collected from UCI Repository and loading of data into R environment
- Data analysis by performing predictive model using R and Python
- Python Packages: Pandas, Matplot, Plotly, scikits-learn, seaborn
- R Packages: ggplot2, dplyr, caret
- R: Boruta for Selection of Variables
- Python: Recursive feature elimination, Logistic regression
- Developed end reports for the predictive model using R, and, Python.
- Tableau, R: Data Visualization

Interoperability between CPOE and LIS Systems

- Extracted data from CDA document in an XML format
- Create XSL to extract required elements from XML
- Generated HL7 message from the XML file
- Deploying of OpenELIS
- Postman to post the HL7-ORU message into OpenELIS
- Documentation and validate the system by checking error logs.

Health Care Education by using Virtual Reality

Designed and developed, Scope statement, Work Break Down schedule, Project work schedule, Communication Plan, Risk Management and budget plan.

Experiences:

Data Analyst Intern, Interprofessional Practice and education center **May2018-July 2018**

Analyze and visualize the **student's responses and their improvements about the courses**

Data Mining and Modeling: Collected, cleansed and provided modeling and analyses of structured and unstructured data

- Efficiently translated research questions into queries and transform the hypothesis into analyzable data
- Analyses the responses and feedbacks provided by the students by using **Python and SQL**.
- Visualized the responses by Sankey diagrams, Origin- Destination matrices, Heat maps, stacked bar plots, time line and violin plots using **Tableau**.

Data Analyst I, Dr. Reddy's laboratories **Jan 2017- May 2017**

- Created summary, customized, and enhanced reports for Clinical database utilizing SQL
- Extracted and analyzed clinical database utilizing SQL and R programming
- Manipulate and build datasets from given data to support analyses through use of efficient SQL code
- Writing complex SQL Queries, Stored Procedures, Triggers, Views, Cursors, Joins, Constraints, DDL, DML and User Defined Functions to implement the business logic and also created clustered and non-clustered indexes
- Perform data cleaning, manipulation and exploratory analyses using Python
- Performed data pre-processing to clean and eliminate outliers in the data and conducted data exploration for detection of correlation, trends and patterns
- Performed ad hoc analysis of data sources for all external and internal customers.
- Data modeling techniques namely logistic regression, classification trees(CART), K-nearest neighbors, SVM
- Conducted independent statistical analysis, descriptive analysis, hypothesis testing, logistic regression, time series analysis, longitudinal data analysis, ANOVA and sampling techniques
- Performance analysis procedures include confusion matrices, and calculation of accuracy, precision, recall, and f1-score. In addition, ROC curves and K-folds cross-validation were performed on the logistic regression model.
- Visualized the data using Origin-Destination matrices, Heatmap, stacked bar plots and timeline using **Tableau**

Statistical Analyst Intern, Aurobindo Pharma Ltd **July 2016-Dec 2016**

Molecular docking guided antiglucansucrase activity of streptococcal mutans by tartaric acid, Gallic acid and naringin from grape, amla and orange respectively and their correlation with invitro anti-streptococcal mutans and invitro anti-glucansucrase activity

- Data collection by using KAP questionnaire and samples from the patients
- Dumping of data into SQL database and extraction using MYSQL Workbench.
- Data cleaning by replacing missing values with mean in variables, Data processing and Data reconciliation
- SAS Base, SAS Macro and SQL Procedures for analysis of Data
- Statistical analysis by performing chi-square, Fisher extract test, descriptive statistics, REG and Linear Regression.
- Graphical representation using tableau

Technical Skills:

- Knowledge in SQL, R, Python and analysis tools like SAS
- Effective in operating various software tools like Oxygen XML editor, Postman, OpenMRS and OpenELIS
- Knowledge and experience with HL7 based interfaces
- Experience and knowledge in Machine learning techniques like Rainforest, Boruta, Linear Regression, Logistic regression using Python, R and SAS.
- Knowledge of Extract, Transform and Load (ETL) frameworks
- Excellent analytical and communication skills
- Outstanding organizational and problem-solving aptitude
- Proficiency in MS Office, Word, Excel & PowerPoint
- Knowledge of MIPS, Value-Based Payment Models and HEDIS
- Knowledge of various medical terms and different medical ontologies MEDDRA, ICD9, ICD10, CPT, ATC, LOINC and SNOMED.
- Proven ability to perform multi-tasking and accomplishing assigned task on time even under pressure
- Potential to share knowledge, communicate efficiently and effectively with all team members

Poster Presentations:

- Poster presentation on **Potential Epidemic transmission of Zika Virus** at G. Pulla Reddy college of pharmacy for One-day national seminar on "Innovations in Pharmaceutical Research-2016 and poster presentations"
- Poster presentation on **An Informative Review on Emerging Prospective of Sphingosomes** at Recent Advancements in Nanoscale particles and colloids in pharmaceutical sciences at Anurag Group of Institutions.
- Poster presentation on **TACE** at Strike Out Cancer Conference held at Ravindra Bharathi, Hyderabad on 29 February 2016.
- Poster presentation on **Artificial Sweeteners- Boon or Bane** in Two Day National symposium held at CPS, IST, and JNTUH.

Awards and others

- Actively participated in the event PRASTHUTHI of **"**AAKRUTHI 2016**",** a National Level Technical Symposium, conducted by the department of PHARMACEUTICS, JNTUH University.
- Paper presentation on **Brain Gate Technology** in Samavarthan 2014 at Bharat Institute of Tech.
- Seminar on **Pathophysiology of Gingivitis and role of Streptococcal Mutants** conducted at Guru Nanak Institute of Technology
- Poster presentation on **Lipodermal Drug Delivery** in Pharmsamhitha-2k16, a National level student Pharma fest held during March 2016 at bharat institute of technology.
- **1500** performances all over the world in dance and **LIMCA BOOK OF RECORD** for Bharatanatyam.

Review article

- Published a review article on **Artificial Sweeteners – Boon or Bane** in **International Journal of Pharmacy.**