A picture containing text, tree

Description automatically generatedA black rectangle with a black background

Description automatically generated with low confidenceLogo, company name

Description automatically generated

DAEN 690

Project Report

Sri Kaushik Reddy Gillela(G01353172)

Srinidh Vobbilisetty(G01352978)

Sai Praneeth Reddy Kethiri(G01355703)

Manideep Kamtala (G01162066)

Shivanvita Jasti(G01354220)

Manish Kovvuri(G01349265)

Spring 2023

**RISK PREDICTION FOR COMPLICATIONS DURING PREGNANCY**

**This page intentionally left blank**

Contents

**Table of Contents**

**ABSTRACT……………………………………………………………………………………………………………………………………………………… 3**

**SECTION 1: PROBLEM DEFINITION………………………………………………………………………………………………………………… 5**

1.1 BACKGROUND……………………………………………………………………………………………………………………………………….. 5

1.2 PROBLEM SPACE……………………………………………………………………………………………………………………………………. 6

1.3 RESEARCH……………………………………………………………………………………………………………………………………………. 8

1.4 SOLUTION SPACE…………………………………………………………………………………………………………………………………. 10

1.5 PROJECT OBJECTIVES……………………………………………………………………………………………………………………………. 12

1.6 PRIMARY USER STORIES ………………………………………………………………………………………………………………………. 13

1.7 PRODUCT VISION…………………………………………………………………………………………………………………………………. 13

1.7.1 SCENARIO #1………………………………………………………………………………………………………………………………………… 13

1.7.2 SCENARIO #2……………………………………………………………………………………………………………………………………….. 13

**SECTION 2: DATASETS…………………………………………………………………………………………………………………………………. 14**

2.1 OVERVIEW……………………………………………………………………………………………………………………………………………. 14

2.2 FIELD DESCRIPTIONS……………………………………………………………………………………………………………………………. 18

2.3 DATA CONTEXT……………………………………………………………………………………………………………………………………. 23

2.4 DATA CONDITIONING………………………………………………………………………………………………………………………….. 23

2.5 DATA QUALITY ASSESSMENT……………………………………………………………………………………………………………….. 24

2.6 OTHER DATA SOURCES………………………………………………………………………………………………………………………… 26

2.7 STORAGE MEDIUM………………………………………………………………………………………………………………………………. 26

2.8 STORAGE SECURITY……………………………………………………………………………………………………………………………… 26

2.9 STORAGE COSTS…………………………………………………………………………………………………………………………………… 26

**SECTION 3: ALGORITHMS & ANALYSIS / ML MODEL EXPLORATION & SELECTION……………………………………….. 26**

3.1 SOLUTION APPROACH………………………………………………………………………………………………………………………….. 26

3.1.1 SYSTEMS ARCHITECTURE……………………………………………………………………………………………………………………… 26

3.1.2 SYSTEMS SECURITY………………………………………………………………………………………………………………………………. 27

3.1.3 ALGORITHMS & ANALYSIS……………………………………………………………………………………………………………………. 27

3.2 MACHINE LEARNING……………………………………………………………………………………………………………………………. 28

3.2.1 MODEL EXPLORATION…………………………………………………………………………………………………………………………. 29

3.2.2 MODEL SELECTION………………………………………………………………………………………………………………………………. 29

**SECTION 4: VISUALIZATIONS / ML MODEL TRAINING, EVALUATION, & VALIDATION………………………….………… 29**

4.1 OVERVIEW……………………………………………………………………………………………………………………………………………. 29

4.2 VISUALIZATIONS…………………………………………………………………………………………………………………………………… 30

4.3 MACHINE LEARNING……………………………………………………………………………………………………………………………. 34

4.3.1 MODEL TRAINING………………………………………………………………………………………………………………………………… 34

4.3.2 MODEL EVALUATION……………………………………………………………………………………………………………………………. 34

4.4 NATURAL LANGUAGE PROCESSING (NLP) ……………………………………………………………………………………………….. 35

4.5 VISIALIZING NLP………………………………………………………………………………………………………………………………..…….. 35

4.6 RESULTS OF NLP………………………………………………………………………………………………………………………………..……. 38

**SECTION 5: FINDINGS………………………………………………………………………………………………………………………………….. 38**

**SECTION 6: SUMMARY………………………………………………………………………………………………………………………………… 39**

**SECTION 7: FUTURE WORK………………………………………………………………………………………………………………………….. 39**

**APPENDIX A: GLOSSARY………………………………………………………………………………………………………………………………. 41**

**APPENDIX B: GITHUB REPOSITORY………………………………………………………………………………………………………………. 42**

**APPENDIX C: RISKS………………………………………………………………………………………………………………………………………. 43**

**APPENDIX D: AGILE DEVELOPMENT…………………………………………………………………………………………………………….. 44**

SCRUM METHODOLOGY…………………………………………………………………………………………………………………………………… 44

SPRINT 1 ANALYSIS…………………………………………………………………………………………………………………………………………… 45

SPRINT 2 ANALYIS……………………………………………………………………………………………………………………………………………… 45

SPRINT 3 ANALYSIS…………………………………………………………………………………………………………………………………………… 45

SPRINT 4 ANALYSIS…………………………………………………………………………………………………………………………………………… 45

SPRINT 5 ANALYSIS…………………………………………………………………………………………………………………………………………… 45

**WORKS CITED……………………………………………………………………………………………………………………………………………… 47**

Abstract

To prevent complications, it is crucial to assess the level of risk associated with pregnancy as early as possible. This is so that the causal factors that are responsible for the risky nature of the pregnancy can be addressed in a timely manner, with a minimal chance of complications, and lesser medical costs incurred. The main concern this project intends to address is labeling new and ongoing pregnancies by their risk level (ranging from Low to High). More specifically, the project aims at tackling this issue by developing a machine learning model that can classify ongoing pregnancies based on their level of risk, and by analyzing the data consisting of the biological, demographic, and clinical parameters associated with previous and completed pregnancies. The machine learning model will be fed said data and trained in a way that lets it classify ongoing pregnancies based on the information it has on completed pregnancies. The data for this project has been provided by Metronomic, Inc and it has since been preprocessed to make it suitable for machine learning. It was also sampled and replicated synthetically to generate more records to ensure better accuracy. Our results are promising.

Report

# Problem Definition

## Background [1]

Pregnancy is a complex and dynamic process that involves a variety of physiological changes and health risks. Preterm labor, gestational diabetes, and preeclampsia are all substantial dangers to mother and fetal health and are a major source of concern for healthcare providers.

To assess the likelihood of difficulties during pregnancy, prenatal care has traditionally relied on demographic and medical history information. This strategy, however, has limitations in correctly forecasting and managing the risk of negative effects. Medical history data, for example, may not accurately reflect the current state of maternal health or the risk of complications, and demographic data, such as age and race, may not effectively reflect the underlying biological and environmental factors that contribute to the risk of complications.

A picture containing bar chart

Description automatically generated

Figure 1.1: Hospital Discharge Statistics

To address these constraints, there is a rising interest in constructing comprehensive and accurate risk prediction models for pregnant problems. To provide a more accurate estimate of the risk of unfavorable outcomes, these models use statistical algorithms to consider many risk factors, such as demographic information, medical history, lifestyle and environmental factors, and biomarkers. These models can provide individualized recommendations for prenatal care and management depending on the evaluated risk of problems by merging numerous data sources and considering the unique characteristics of each individual pregnancy.

The use of risk prediction models has the potential to enhance maternal and fetal outcomes by allowing healthcare practitioners to identify women who are at high risk of difficulties early in pregnancy and take appropriate measures to lower the probability of negative outcomes. This tailored approach to prenatal treatment can result in better outcomes for mothers and kids, such as lower rates of bad outcomes, improved maternal and fetal health, and lower healthcare expenditures.

Shape, logo, company name

Description automatically generated

Figure 1.2: Types of Pregnancy Complications

## Problem Space

Although a good many of pregnancies in the world are uncomplicated, there are a few pregnancies which complications do occur and there are many women and their families who experience the pain resulting from these complications. These complications can occur in any of the three trimesters during pregnancy and those that occur during the second and third trimesters are considered to be higher risk than those that occur early-on in the pregnancy.

Gestational Diabetes [4]: A form of high blood sugar that affects pregnant women.

Diagram

Description automatically generated

Figure 1.3: Gestational Diabetes statistics in the US

Pre-eclampsia [5]: A dangerous pregnancy complication characterized by high blood pressure. It typically develops after 20 weeks of pregnancy and can affect both the mother and the baby.

Map

Description automatically generated

Figure 1.4: The significance of Preeclampsia

C-Section [6]: The condition where the pregnancy ends in a C-section, when it could have been avoided by changing a few factors during the pregnancy.

Chart, line chart

Description automatically generated

Figure 1.5: C-section statistics

Pre-term labor [7]: When a baby is born too early, before 37 weeks of pregnancy have been completed.

Chart, bar chart

Description automatically generated

Figure 1.6: Premature labor statistics

High/low weight in the baby: A condition where the baby develops to a weight that is either lower or higher than its ideal weight.

There are many risks associated with trying to mitigate the complications that occur in the later stages of pregnancy. Complications can range anywhere from Gestational Diabetes, Pre-Eclampsia, Pre-term delivery, possibility of a C-Section and the baby not growing to full term. Trying to tackle these problems in the later stages of pregnancy is associated with a multitude of risks including complications to the health of the mother or the baby and also the incurrence of higher costs. It is considered to be much safer and cost efficient to tackle these risks in the initial stages of the pregnancy itself. There arises the problem statement of this project: developing a machine learning model that is capable of accurately predicting the probability of occurrence of complications in newer patients by discovering trends and patterns in the historic and clinical data from older patients' medical records and then implementing these trends and patterns.

## Research

**Trends in Pregnancy and Childbirth Complications in the U.S** [2]

According to a study by Blue Cross Blue Shield, the Health of America, which used data from over 1.8 million pregnancy episodes between 2014 and 2018, the incidence of pregnancy complications has been increasing in recent years. Some of the key findings of this study are:

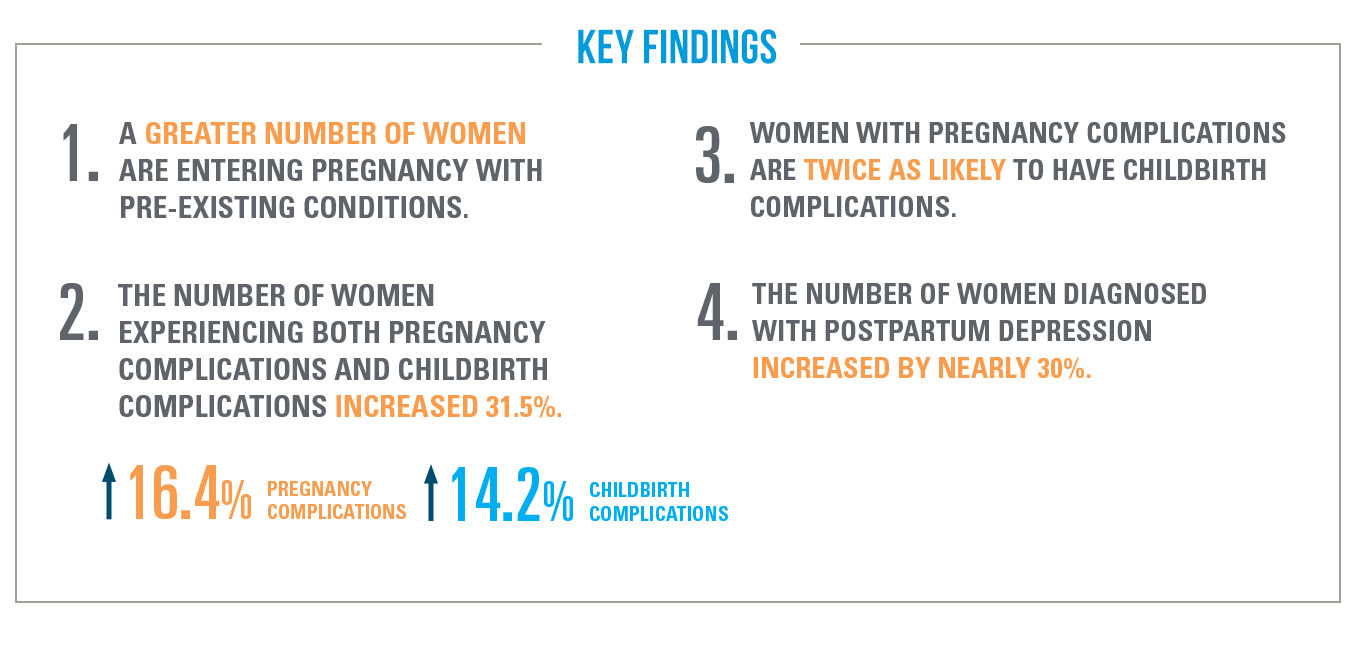


Figure 1.7: Findings on pregnancy complications

Pre-Existing Conditions

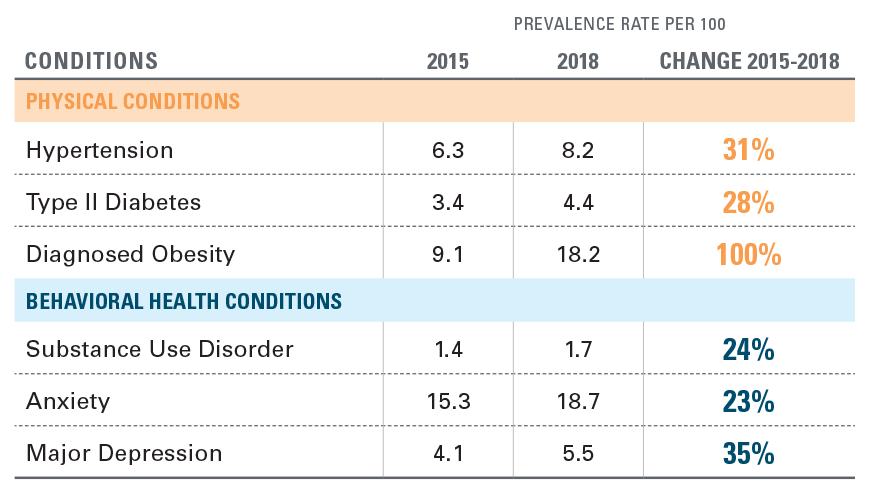
The study shows that there has been a significant increase in the predominance of chronic physical and behavioral health conditions, in which diagnosed obesity and major depression experiencing the biggest rises. These pre-existing conditions can majorly influence the pregnancy and childbirth complications rate. 

Figure 1.8: Rise in rates of pre-existing conditions

How increase in pregnancy complication is affecting the childbirth complications.

Even though 80% of pregnancies and births are healthy, problems are becoming more common. The rates of pregnancy difficulties increased by more than 16% between 2014 and 2018, while the rates of delivery complications increased by more than 14%. Both types of problems affected about seven out of every 1,000 pregnant women, a nearly 31% rise from 2014. There were significantly more pregnancy complications among elder women (ages 34-44).

Their study has shown how that gestational diabetes and pre-eclampsia driving increase in pregnant complication rates.

Rates of childbirth complications among women with/without pregnancy complications in 2018.

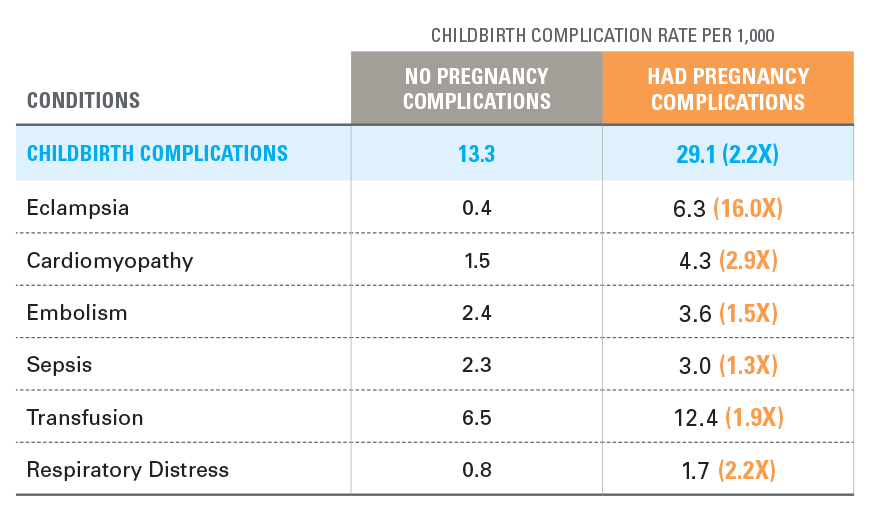


Figure 1.9: Childbirth complication rates per 1000

We can see that the Childbirth Complications are twice with women who had some kind of Pregnancy Complication.

Postpartum depression

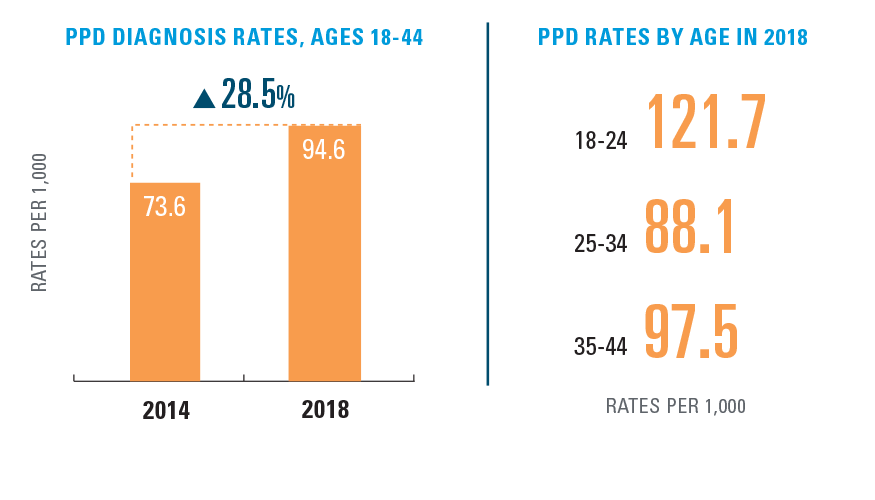
Postpartum depression is a common and serious condition affecting women after childbirth. In recent years, the result of a confluence of biological, psychological.

Figure 1.10: Statistics of postpartum Depression

**Maternal mortality** [3]

According to the World Health Organization (WHO), the maternal mortality rate has remained unacceptably high, with an estimated 830 women dying from preventable causes related to pregnancy and childbirth every day. This trend is seen across all regions of the world, with the highest rates occurring in low-income countries.

WHO attributes the rise in maternal morbidity and mortality to a lack of access to quality prenatal care and inadequate resources and support for new mothers. To address this crisis, WHO is working to improve access to quality maternal health care and reduce disparities in maternal health outcomes between different populations.

Some of key facts mentioned in the article are

* 94% of all maternal deaths occur in low and lower middle-income countries.
* Young adolescents (ages 10-14) face a higher risk of complications and death as a result of pregnancy than other women

## Solution Space

We have conducted extensive research on the pregnancy complications that we plan to predict in this project and have listed out the top most major contributing factors that lead to the occurrence of each complication. The details are as follows:

Gestational Diabetes:

1. Having pre-diabetes / Family History of Type 2 diabetes
2. Obesity
3. Polycystic Ovarian Syndrome (PCOS)
4. Having had gestational diabetes during a previous pregnancy
5. High Blood Pressure
6. Higher Maternal Age

Pre-Eclampsia:

1. Being diabetic/ pre-diabetic
2. Obesity
3. Having had pre-eclampsia in a previous pregnancy
4. The mother being older than 35 - 40 years
5. Multiple pregnancy (twins or more)
6. Being pregnant for the first time
7. Chronic Hypertension

Pre-term labor:

1. Other complications in the pregnancy (GD or pre-eclampsia)
2. Multiple pregnancy (twins or more)
3. Chronic conditions such as diabetes or Hypertension
4. Tobacco and alcohol use / Substance abuse
5. Prior preterm birth
6. Stress

Unnecessary C-Section:

1. Fear of childbirth
2. Scheduling Convenience
3. Higher Maternal age
4. Previous C-section
5. High blood pressure
6. Presence of pregnancy complications
7. Multiple pregnancy (Twins or more)

High/ low weight of the baby:

1. Number of babies
2. Gestational age of the baby
3. Mother’s diet during the pregnancy
4. Age of the mother – less than 20 and greater than 35 have a higher chance of LBW baby
5. Mother’s medical conditions
   1. Anemia, excess blood pressure – underweight baby
   2. Type 1/2 diabetes, Gestational Diabetes – overweight baby

We plan on analyzing these factors further and applying the appropriate machine learning methods/ predictive algorithms to each of the complication. The models will be designed in such a way that the major causal factors discussed above will be used to predict the complication that they are responsible for causing. Using this method, the chances of occurrences of the complications can be predicted to a certain degree of accuracy.

For this project, the historic and clinical data (both the causal factors and the complications) from previous pregnancies will be gathered and analyzed and fed to machine learning models, in order to train the models to expect the occurrence of a certain complication, if all or most of its causal factors are also present in a mother. In this way, when the data of expecting mothers is fed to the machine learning models, they can predict the occurrence of a complication based on the causal factors that are present in the mother.

**1.4.1 Understanding Similar Projects**

In the process of gaining a better understanding of our project and problem statement, we have looked into similar previous projects, projects in the same domain and the projects that have similar problem statements as ours does. Below are the projects that we have found to be the most relevant to out project:

1. Development of a Predictive Risk Assessment Tool for Pregnancy Complications [9]
2. Machine Learning-based Early Warning System for Pregnancy Complications [8]
3. Predictive Modeling of Pregnancy Complications using Electronic Health Records (EHRs) [11]
4. Automated Risk Prediction for Pre-eclampsia [10]

We have extensively investigated these projects and found the methods and technologies used in each of them to find solutions to their respective problem statements. The technologies that have been used most across all the four projects are as follows:

1. Supervised learning
2. Gradient boosted decision trees
3. Machine Learning
4. Deep Learning
5. AUC (area under the curve)
6. Cost sensitive hybrid model (CSHM)
7. Elastic net algorithm
8. Gradient boosting algorithm
9. Cross-fold validation
10. Predictive modeling

We plan on drawing inspiration from these methods and applying similar techniques to our data as well, just to verify the compatibility of our data with machine learning methods. We also plan on testing several types of machine learning models for each of our complications, and finalize the best fitting model through a series of cross validation methods. This way, the highest possible accuracy can be achieved in terms of predicting the probability of occurrence of pregnancy complications.

## Project Objectives

The main aim of this project is to be able to predict the occurrences of complications that might develop in the later stages of the pregnancy. The details are as mentioned below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Complication** | **Predict** | | **Predict by** |
| Gestational Diabetes | | Chances of occurrence | Week 12 |
| Pre-Eclampsia | | Chances of occurrence and risk factors | Second trimester |
| Pre-Term | | Chances of occurrence and risk factors | End of second trimester |
| Unnecessary C-Section | | Chances of normal delivery | Before term |
| High/ Low weight gain | | Ideal weight gain and baby weight | - |

By the end of this project, we wish to understand the reasons why certain complications occur in a pregnancy and what factors majorly impact the development of complications. By predicting the odds of a complication developing in the later stages of a pregnancy, the necessary treatment and medical intervention can be provided to the mother in the early stages of her pregnancy itself, thereby minimizing the risks and the associated costs with the treatment.

We wish to develop a machine learning model for each of the complications listed above in such a manner that depending on the parameters of the pregnancy, the probability of development of complications in the pregnancy can be accurately estimated.

The results of these findings will be extremely beneficial to the pregnant women out there, both in terms of the risk factor in their pregnancy and a financial standpoint. When identified in the early stages of a pregnancy, say the first trimester, the treatment of complications can be simpler, which makes it easier on both the mother and the baby, while also being cost effective.

## Primary User Stories

In the current world of obstetrics, many pregnancy complications are not being detected until it is too late or too complicated to get them treated. This is increasing the risks and morbidity rate associated with pregnancies, leading to an overall unfavorable outcome.

With this project, we wish to address this fact and come up with a technical solution where the chances of occurrence of complications in the later stages of a pregnancy can be estimated in the earlier stages itself.

We are planning on developing an appropriate machine learning model for each of the complications that the project aims to predict, since no two complications can be predicted using the same causal factors, and hence the algorithm has to be trained in a different manner for each of the complications.

## Product Vision

### Scenario #1

Mothers and their families:

The ones who would be greatly benefited by the employment of this solution, and the ones that were kept in mind while devising the solution, are the mothers and their families. The risk factors and the treatment costs of treating pregnancy complications increase proportionally with the increase in the gestational age of the pregnancy. Since this solution focuses on predicting the chances of development of complications during the early stages of the pregnancy, the necessary steps or changes can be taken to lessen the chances of complications early on in the pregnancy, thereby avoiding complex and expensive medical intervention during the later stages of the pregnancy. This makes for an overall safer pregnancy and childbirth for both the mother and the baby.

### Scenario #2

Healthcare providers:

There are multiple advantages that the Healthcare industry can gain by using this solution. Treating the complications in a pregnancy before it is too late or too complicated is much safer and beneficial to the Healthcare providers. This would aid in keeping the overall mortality rate of the healthcare practice at a minimum, while also earning them positive remarks throughout.

# Datasets

## Overview

The dataset that has been used for the purpose of this project has been provided by Metronomic Inc., and it comprises of the clinical and medical historic data of previous pregnancies. There are over 60 descriptive parameters related to each pregnancy that have been gathered in this dataset. The parameters that have been captured in this dataset range from demographic factors of the mothers to their biological factors to their family histories and the clinical records during their pregnancies. There are multiple data types across the dataset, that have to be formatted and converted accordingly in order to make the data feasible for machine learning algorithms.

The datatypes and normal ranges of the data parameters are as mentioned below:

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Possible Data Type** | **Normal Values/ Ranges** |
| Marital Status | String | Married/ Single/ Divorced etc. |
| Ethnicity | String | Asian/ American etc. |
| Maternal age | Numeric | 20 – 40 |
| Age at first child | Numeric | N/A |
| Weight | Numeric (lbs) | 170 – 180 |
| BMI | Numeric | 18.5—24.9 |
| Maternal weight gain | Numeric (lbs) | 25 – 35 |
| Method of conception | String | Medications, IVF etc. |
| Gravida | Numeric | N/A |
| Para | Numeric | N/A |
| Maternal work | String | Engineer, Lawyer etc. |
| Maternal employment status | String | Employed, Unemployed etc. |
| Income | Numeric | N/A |
| Domestic abuse | Boolean | Yes/ No |
| Number of siblings prior | Numeric | N/A |
| Past history of DM | Boolean | Yes/ No |
| Past history of GDM | Boolean | Yes/ No |
| Past history of hypertension | Boolean | Yes/ No |
| Past history of renal disesase | Boolean | Yes/ No |
| Past history of Autoimmune disease | Boolean | Yes/ No |
| Past history of C-section | Boolean | Yes/ No |
| Past history of abruption | Boolean | Yes/ No |
| Past history of pre-eclampsia | Boolean | Yes/ No |
| Past history of preterm delivery | Boolean | Yes/ No |
| Past history premature rupture of membranes | Boolean | Yes/ No |
| Past history of lacerations (tears) | Boolean | Yes/ No |
| Past usage of forceps | Boolean | Yes/ No |
| Past usage of suction | Boolean | Yes/ No |
| Past history of NICU admission | Boolean | Yes/ No |
| Past history of IUGR | Boolean | Yes/ No |
| Family histories | Boolean | Yes/ No |
| Prenatal medications | Boolean | Yes/ No |
| Crohn’s | Boolean | Yes/ No |
| Ulcerative colitis | Boolean | Yes/ No |
| Post partum depression | Boolean | Yes/ No |
| Multiple pregnancy (twins or more) | Boolean | Yes/ No |
| Place of delivery (urban, rural, academic setting) | String | Urban, Rural etc. |
| Smoking | Boolean | Yes/ No |
| Alcohol consumption | Boolean | Yes/ No |
| Drug use | Boolean | Yes/ No |
| Blood Pressure measurements | Numeric (mm Hg) | Less than 120/80 |
| Sex of previous baby | String | M, F |
| Birth weight of previous baby | Numeric (lbs) | 5 – 9 |
| Drug history | Boolean | Yes/ No |
| Breast feeding outcomes | Numeric(oz) | 25-35 oz / 24 hours |
| Fundal height measurements | Numeric (cm) | 18-22 cm (at 20 weeks of gestation) |
| Prior history of circlage | Boolean | Yes, No |
| Prior history of maternal trauma (motor vehicle accident) | Boolean | Yes, No |
| Maternal vaccination status (MMR, T-Bac) | Boolean | Yes, No |
| Hemoglobin | Numeric (g/dL) | 11-14 in the first trimester 10.5-13.5 in the second trimester |
| 10-12 in the third trimester |
| White blood cell count | Numeric (cells/mcL) | 5,000-15,000 |
| platelet count | Numeric (cells/mcL) | 150,000-450,000 |
| GTT | Numeric (mg/dL) | Fasting blood glucose less than 92 |
| 1-hour post-load glucose less than 180 |
| 2-hour post-load glucose less than 153 |
| STD tests | String | Positive, Negative |
| Thyroid levels | Numeric (mU/L) | 0.2-<2.5 in the first trimester of pregnancy |
| 0.3-3 in the remaining trimesters |
| Blood type | String | A, B, AB, O |
| RH status | String | Positive, Negative |
| Group B strep | String | Positive, Negative |
| Urine studies (proteinuria) | Numeric (mg/dL) | Less than 150 (or less than 0.15g/L) |
| Total Cholestrol | Numeric (mg/dL) | Less than 200 |
| HDL | Numeric (mg/dL) | 40-60 |
| LDL | Numeric (mg/dL) | Less than 130 |
| Triglycerides | Numeric (mg/dL) | Less than 150 |
| Adiponectin | Numeric (mg/dL) | 5-30 |
| E-selectin | Numeric (mg/dL) | 10-50 |
| CRP | Numeric (mg/dL) | Less than 10 |
| GGT | Numeric (U/L) | Less than 40 |
| t-PA | Numeric (μg/l) | 5 – 40 |
| Pregnancy associated plasma protein-A (PAPP-A) | Numeric (MOM) | More than or equal to 0.5 |
| Progestorone levels | Numeric (ng/mL) | Pregnancy 1st trimester: 11.2 to 90.0 |
| Pregnancy 2nd trimester: 25.6 to 89.4 |
| Pregnancy 3rd trimester: 48 to 150 to 300 or more |
| Placental growth factor (PIGF) | Numeric (pg/ml) | At 30 weeks gestation: 141 |
| At term: 23 |
| Free beta HCG | Numeric (MOM) | 0.5 – 2 |
| Fetal Fibronectin | Numeric (ng/ml) | 10 – 12 weeks: 4000 |
| 22 weeks: 50 |
| 36 – 37 weeks: undetectable |
| Estriol level | Numeric (ng/ml) | Pregnancy 1st trimester: <= 2.5 |
| Pregnancy 2nd trimester: <= 9.6 |
| Pregnancy 3rd trimester: <= 14.6 |
| 20 week scan (fetal anomaly scan) | String | Normal, Abnormal etc. |
| Organs, growth, fetal presentation, asymmetric growth | String | Normal, Abnormal etc. |
| Uterine artery pulsatility index (Ut-PI) | Numeric | 0.3 – 1.3 |
| Nuchal Translucency | Numeric (mm) | Less than 3mm |
| Quad screen data | Numeric | AFP: 15 – 200 ng/ml |
| Estriol: 0.5 – 2.5 mg/ml |
| hCG: 50,000 – 300,000 mIU/mL |
| Inhibin-A: 50 – 1500 pg/mL |
| Sequential screens | Numeric | N/A |
| Mean Arterial Pressure (MAP) | Numeric (mm Hg) | 70-100 |
| Cervical length and consistency | Numeric (mm) | generally greater than 25 mm |
| Fetal weight and head circumference | Numeric (lbs) | 5 – 9 |
| Amniotic fluid measurements | Numeric (cm) | 5 – 25 |
| Sex of the fetus | String | M, F |
| Placentation (grade and previa) | String | grade 0 (immature) – grade 3 (mature) |
| Umbilical cord details | String | N/A |
| Cord insertion (marginal, central) | String | Marginal, Central etc. |

The initial plan in understanding the data is to analyze each of the complications in the previous pregnancy data, look into the values of the parameters for each pregnancy, and see which of the parameters (contributing factors) are out of their normal ranges according to the analysis done in the above table. This will aid in the formulation of a system where the contributing factors can be linked to the respective complication that they play a major role in causing. This analysis can be seen in the table below:

Table

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

Theerclagers that have been marked as ‘required’ play a major role in the onset of complications during a pregnancy.

## Field Descriptions

The fields in the dataset and their respective datatypes can be described as follows:

1. **Marital Status** (String) : This field describes the marital status of the pregnant woman. It is of the type string and can take the values ‘Married’, ‘Single’, ‘Divorced’ etc.

2. **Ethnicity** (String): This field indicates the ethnicity of the pregnant woman. It is of the type string and takes the value of the ethnic category the mother falls into.

3. **Maternal Age** (Numeric): It indicates the age of the woman at the time of pregnancy, and is of numerical format.

4. **Age at first child** (Numeric): It indicates the age of the woman when she had her first child. It is of numerical format.

5. **Weight** (Numeric / lbs): It indicates the weight of the pregnant woman and the field is of numeric format.

6. **BMI** (Numeric): It denotes the Body Mass Index of the mother and the field is of numeric format.

7. **Maternal Weight gain** (Numeric / lbs): This field indicates the total weight that the woman has gained during her pregnancy. The field is of numeric format.

8. **Method of conception** (String): It indicates the method in which the pregnancy has been conceived. The field is of String format and can take the values ‘IVF’, ‘Donor’ etc.

9. **Gravida** (Numeric): This indicates the number of pregnancies a woman has had previously. The field is of numeric format.

10. **Para** (Numeric): This indicates the number of previous pregnancies of the woman that have crossed 20 weeks of gestation and is of numeric format.

11. **Maternal Work** (String): This indicates the line of work the mother is in, and the field is of String format.

12. **Maternal Employment Status** (String): This indicates the employment status of the woman at the time of her pregnancy (Employed/ Unemployed) and the field is of String format.

13. **Income** (Numeric): This field is of numeric data format and gives information on the income of the mother.

14. **Domestic Abuse** (Boolean): This field indicates if the woman is encountering any form of Domestic Abuse. The field is of Boolean type – as it can only take two values – yes and no.

15. **Number of siblings prior** (Numeric): This field is of numeric format and indicates how many children the woman has birthed prior to her current pregnancy.

16. **Past History** **of DM** (Boolean): This field indicates if the woman has a prior history of DM. The field is of Boolean type – as it can only take two values – yes and no.

17. **Past History of GDM** (Boolean): This field indicates if the woman has a prior history of GDM. The field is of Boolean type – as it can only take two values – yes and no.

18. **Past History of Hypertension** (Boolean): This field indicates if the woman has a prior history of Hypertension. The field is of Boolean type – as it can only take two values – yes and no.

19. **Past History of Renal Disease** (Boolean): This field indicates if the woman has a prior history of Renal Disease. The field is of Boolean type – as it can only take two values – yes and no.

20. **Past History of Autoimmune Disease** (Boolean): This field indicates if the woman has a prior history of Autoimmune Disease. The field is of Boolean type – as it can only take two values – yes and no.

21. **Past History of C-Section** (Boolean): This field indicates if the woman has undergone a C-section for any of her previous pregnancies. The field is of Boolean type – as it can only take two values – yes and no.

22. **Past History of Abruption** (Boolean): This field indicates if the woman has a prior history of Abruption. The field is of Boolean type – as it can only take two values – yes and no.

23. **Past History of Pre-Eclampsia** (Boolean): This field indicates if the woman has a prior history of Pre-Eclampsia in any of her previous pregnancies. The field is of Boolean type – as it can only take two values – yes and no.

24. **Past History of Pre-Term delivery** (Boolean): This field indicates if the woman has a prior history of Pre-Term delivery in any of her previous pregnancies. The field is of Boolean type – as it can only take two values – yes and no.

25. **Past History of Premature rupture of membranes** (Boolean): This field indicates if the woman has a prior history of Premature rupture of membranes in any of her previous pregnancies. The field is of Boolean type – as it can only take two values – yes and no.

25. **Past history of lacerations** (Boolean): This field indicates if the woman has a prior history of lacerations in any of her previous pregnancies. The field is of Boolean type – as it can only take two values – yes and no.

26. **Past usage of Forceps** (Boolean): This field indicates if forceps have been used in any of her previous pregnancies. The field is of Boolean type – as it can only take two values – yes and no.

27. **Past usage of suction** (Boolean): This field indicates if suction has been used in any of her previous pregnancies. The field is of Boolean type – as it can only take two values – yes and no.

28. **Past history of NICU Admissions** (Boolean): This field indicates if any of the mother’s previous children have been admitted to the NICU. The field is of Boolean type – as it can only take two values – yes and no.

29. **Past History of IUGR** (Boolean): This field indicates if the woman has a prior history of IUGR. The field is of Boolean type – as it can only take two values – yes and no.

30. **Family Histories:** This field can take multiple data types based on the conditions and diseases of the family that the field indicates.

31. **Prenatal Medications** (Boolean): This field is indicative of if the mother has taken any medications before and during the pregnancy, before the birth. The field is of Boolean type – as it can only take two values – yes and no.

32. **Crohn’s** (Boolean): This field indicated whether or not the mother suffers from Crohn’s disease. The field is of Boolean type – as it can only take two values – yes and no.

33. **Ulcerative Colitis** (Boolean): This field indicated whether or not the mother suffers from Ulcerative colitis. The field is of Boolean type – as it can only take two values – yes and no.

34. **Post-Partum depression** (Boolean): This field indicates if the woman has a prior history of Post-Partum depression in any of her previous pregnancies. The field is of Boolean type – as it can only take two values – yes and no.

35. **Multiple pregnancy** (Boolean): This field indicates if the current pregnancy is a multiple pregnancy or not (twins or more). The field is of Boolean type – as it can only take two values – yes and no.

36. **Place of delivery** (String): This field describes the place where the delivery has taken place. It is of the type string and can take the values ‘Urban, ‘Rural’ etc.

37. **Smoking** (Boolean): This field indicates if the mother has the habit of smoking or not. The field is of Boolean type – as it can only take two values – yes and no.

38. **Alcohol Consumption** (Boolean): This field indicates if the mother has the habit of Alcohol Consumption or not. The field is of Boolean type – as it can only take two values – yes and no.

39. **Drug Use** (Boolean): This field indicates if the mother has the habit of drug use or not. The field is of Boolean type – as it can only take two values – yes and no.

40. **Blood Pressure measurements** (Numeric / mm Hg): This field denotes the blood pressure of the woman during the pregnancy and the field is of numeric type.

41. **Sex of the previous baby** (String): This field is indicative of the gender of the previous children that the woman has had. The field if of type String.

42. **Drug history** (Boolean): This indicates any previous usage of drugs by the pregnant mother. The field is of Boolean type – as it can only take two values – yes and no.

43. **Breastfeeding Outcomes** (Numeric / oz): This field measures the volume of breastmilk that the mother is generating and is given by a Numeric data type.

44. **Fundal height measurements** (Numeric / cm): This field indicates the height of the fundus in the woman and is of numeric data format.

45. **Prior history of erclage** (Boolean): This field indicates if the woman has a prior history of erclage in any of her previous pregnancies. The field is of Boolean type – as it can only take two values – yes and no.

46. **Prior history of maternal trauma** (Boolean): This field indicates if the woman has experienced trauma during any of her previous pregnancies. The field is of Boolean type – as it can only take two values – yes and no.

47. **Maternal Vaccination Status** (Boolean): This field indicates if the mother has been vaccinated against common viruses and bacteria. The field is of Boolean type – as it can only take two values – yes and no.

48. **Hemoglobin** (Numeric / g/dL): This field indicates the hemoglobin levels in the mother’s blood during her pregnancy. It is of numeric datatype.

49. **White blood cell count** (Numeric / cells/mcL): This field measures the count of white blood cells of the mother and is of numeric datatype.

50. **Platelet cell count** (Numeric / cells/mcL): This field measures the count of platelets of the mother and is of numeric datatype.

51. **GTT** (Numeric / mg/dL): This field measures the glucose levels in the mother’s blood during pregnancy and is of numeric datatype.

52. **STD tests** (String): This field indicates the results of any STD tests that have been performed on the mother. It is of string datatype and the results can take on the values of ‘Positive’ and ‘Negative’.

53. **Thyroid levels** (Numeric / mU/L): This field measures the thyroid levels of the mother during pregnancy and is of numeric datatype.

54. **Blood Type** (String): This field specifies the blood type of the mother. It can have the values ‘A’, ‘B’, ‘AB’ and ‘O’ and is of the type String.

55. **RH status** (String): This field specifies the RH type of the mother. It can have the values ‘Positive’ and ‘Negative’ and is of the type String.

56. **Group B Strep**: This field specifies if the mother has contracted Group B strep and is of type String, as it can have the values ‘Positive’ and ‘Negative’.

57. **Urine studies** (Numeric / mg/dL): This field indicates the protein levels in the mother’s urine and is of Numeric data type.

58. **Total Cholestrol** (Numeric / mg/dL): This field indicates the cholesterol levels of the mother. It is of numeric datatype.

59. **HDL** (Numeric / mg/dL): This field measures the HDL levels in the mother and is of numeric datatype.

60. **LDL** (Numeric / mg/dL): This field measures the LDL levels in the mother and is of numeric datatype.

61. **Triglycerides** (Numeric / md/dL): This field measures the level of Triglycerides in the mother and is of numeric datatype.

62. **Adiponectin** (Numeric / md/dL): This field measures the level of Adiponectin in the mother and is of numeric datatype.

63. **E-selectin** (Numeric / mg/dL): This field measures the level of E-selectin in the mother and is of numeric datatype.

64. **CRP** (Numeric / mg/dL): This field measures the CRP levels in the mother and is of numeric datatype.

65. **GGT** (Numeric / U/L): This field measures the GGT levels in the mother and is of numeric datatype.

66. **t-PA** (Numeric / μg/l): This field measures the t-PA levels in the mother and is of numeric datatype.

67. **Pregnancy associated plasma protein-A (PAPP-A)** (Numeric / MOM): This field measures the level of Pregnancy associated plasma protein-A in the mother and is of numeric datatype.

68. **Progesterone levels** (Numeric / ng/dL): This field measures the level of Progesterone in the mother and is of numeric datatype.

69. **Placental growth factor (PIGF)** (Numeric / pg/ml): This field measures the rise of PIGF in the mother and is of numeric datatype.

70. **Free beta HCG** (Numeric / MOM): This field measures the level of Free beta HCG in the mother and is of numeric datatype.

71. **Fetal Fibronectin** (Numeric / ng/ml): This field measures the levels of Fetal Fibronectin in the mother and is of numeric datatype.

72. **Estriol level** (Numeric / ng/ml): This field measures the levels of Estriol in the mother and is of numeric datatype.

73. **20 week scan** (fetal anomaly scan) (String): This field specifies the results of the scanning conducted at 20 weeks of gestation and is of the type String. It can take the values ‘Normal’, ‘Abnormal’ etc.

74. **Organs, growth, fetal presentation, asymmetric growth** (String): It specifies the conditions of the organs of the fetus and is of the type String. It can take the values ‘Normal’, ‘Abnormal’ etc.

75. **Uterine artery pulsatility index (Ut-PI)** (Numeric): This field measures the level of Ut-PI in the mother and is of numeric datatype.

76. **Nuchal Translucency** (Numeric / mm): It is a measure of the amount of fluid behind the fetus’s neck in the first trimester of pregnancy. The field is of numeric type.

77. **Quad screen data** (Numeric): It consists of the following factors: AFP, Estriol, hCG and Inhibin. All of these factors are of numeric datatype.

78. **Sequential screens** (Numeric): This field is indicative of several blood tests and ultrasounds in the first trimester of the pregnancy. The results are of numeric datatype.

79. **Mean Arterial Pressure** (MAP) (Numeric / mm Hg): It is a measure of the Mean Arterial Pressure in the mother during the pregnancy and is of numeric datatype.

80. **Cervical length and consistency** (Numeric / mm): It is a measure of the Cervical length and consistency in the mother during the pregnancy and is of numeric datatype.

81. **Fetal weight and head circumference** (Numeric / lbs): It is a measure of the Fetal weight and head circumference of the baby and is of numeric datatype.

82. **Amniotic fluid measurements** (Numeric / cm): It is the sum of the measurements of the deepest vertical pockets of amniotic fluid found in each of the four abdominal quadrants. It is measured in a numeric datatype.

83. **Sex of the fetus** (String): This mentions the gender of the fetus and the field if of the type String.

84. **Placentation** (grade and previa) (String): This field describes the grade of contact between the fetus and the mother and ranges from grade 0 – grade 3. The field is of type String.

85. **Umbilical cord details** (Numeric / cm): This field specifies the length of the umbilical cord throughout the course of the pregnancy and is measured by a numeric datatype.

86. **Cord insertion** (String): This field provides information about the way in which the cord has been inserted and is of the datatype String. It takes the values ‘Marginal’, ‘Central’ etc.

## Data Context

Since the primary goal of the project is to predict the occurrences of complications during a pregnancy in the early stages of the pregnancy itself, analysis has to be made on data which has been taken by documenting the parameters and complications of previous pregnancies. This way, the relation between the contributing factors and the pregnancy complications that they are responsible for causing can be examined and interpreted. These relations can further be used on data pertaining to current and on-going pregnancies, and the presence of certain causal factors can be seen as a sign of possibility that their respective complication may occur in the later stages of the pregnancy.

For this project, the data has been gathered by documenting the details of previous pregnancies and the parameters in the data include a wide variety of information regarding the demographic, biological, historical and clinical data of the pregnant women. The data also has been curated from mothers across the world, so that multitudes of possibilities of complications occurring during a pregnancy can be analyzed.

## Data Conditioning

Since the dataset that is being used for this project is fairly lesser in size, and all the three datasets that have been provided by Metronomic,Inc. have been appended into a single dataset, there is no need currently to have an elaborate storage mechanism and make any alterations to the storage medium that is in place. The data and the files generated throughout the course of the project will be stored in a private github repository and based on further developments in the project, if the size of the data increases significantly with Synthetic Data Generation techniques, further data conditioning steps will be explored on the go.

**2.4.1 Data Cleaning and Preprocessing**

In order to make the data viable for machine learning algorithms, several preprocessing steps have been performed on it. The details of which are as follows:

1. Calculating number of previous unsuccessful pregnancies:

Isolated the gravida and para values from ‘previousPregnancy’ column and calculate number of previous unsuccessful pregnancies by calculating ‘gravida – para’.

1. Converging the doctor’s advice notes and clinical observations from the secondary datasets into the first dataset.
2. Created a new column for ‘Age at pregnancy’ by calculating ‘EDD (estimated date of delivery – Date of Birth)
3. Based on the conditions in the past medical history, and based on the causal factors of the complications we have researched in section 2.1, assigned possible pregnancy complications to each patient’s record.
4. Found all the unique parameters in the ‘past medical history’ column and created as many unique columns in the dataset. For these columns, assigned 1 if a patient has that condition, 0 if not.

past\_medical\_history column in raw data:

Graphical user interface, text, application, email

Description automatically generated

Figure 2.1: ‘past\_medical\_history’ column in raw data

processed past\_medical\_history column:

Graphical user interface, text, application

Description automatically generated

Figure 2.2: ‘past\_medical\_history’ column after preprocessing

## Data Quality Assessment

We have received three datasets from Metronomic,Inc. for the purpose of this project. The attributes that are available and defined in these datasets are as follows:

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Parameter** | **Data Type** |
| Primary dataset | Patient Id | Integer |
| Date of birth | Date |
| Medical condition | String |
| Co\_morbidity | String |
| Ethnicity | String |
| Risk\_factor | String |
| Imp | Date |
| Cycle length | Integer |
| EDD (Estimated date of delivery) | Date |
| Number of babies | Integer |
| Other checks | String |
| Infection history | String |
| Past medical history | String |
| Past pregnancy | String |
| Medications | String |
| Secondary datasets | Doctor’s advice notes | String |
| Clinical observations | String |

The primary dataset contains general and demographic information on 97 patients and the secondary dataset contains information regarding the appointments that the patients have had throughout the course of their pregnancies. The columns doctor’s advice notes and clinical observations have been converged into the primary dataset with respect to the patient id column.

The quality assessment of the data is as below:

* Completeness – There are null values present in few columns (past\_medical\_history , infection\_history , other checks)
* Consistency – The data is consistent across all three datasets and within each dataset.
* Uniqueness – All the entities within a dataset are unique.
* Integrity – ‘Doctor’s notes’ and ‘clinical observations’ from the secondary datasets must be converged into the primary dataset.
* Conformity – The data conforms to the right standards.
* Accuracy – The data is accurate and unambiguous.

## Other Data Sources

We have searched for and found one dataset on Kaggle that is relevant to our project. The fields in the dataset are as below: [12]

* Age
* Systolic BP
* Diastolic BP
* Blood Sugar level
* Body temperature
* Heart Rate
* Risk Level

This data can be used to analyze the blood pressure, blood sugar levels, heart rate and body temperatures of the pregnant women to estimate the risk level of the pregnancy (ranging from low risk to high risk). The number of columns (parameters) is limited in this dataset and there is no information on predicting the occurrence of specific complications, just the overall risk level of the pregnancy can be estimated using this data.

## Storage Medium

The primary storage medium that has been chosen for the storage of the data files for this project is github file repository. A private repository has been created on github [13] that is only accessible to the project team and the dataset and the code files generated by this project can be securely stored in the private repository without the risk of data leakage.

## Storage Security

Since the chosen storage medium for the data set in this project is a private github repository with access privileges granted only to the project team, the dataset is stored in a secure manner without the risk of the data being leaked. The data stored in Github private repositories will never be shared with third parties and will not be accessed by Github staff. Hence, the only people who will be accessing the project related data is the team working on the project.

## Storage Costs

The chosen storage medium for the data set in this project is a private github repository, which is offered for free by Microsoft corp and can be used completely free of cost to store the dataset used for this project. Hence, there are no costs associated to the storage of the data in this project as of yet. Further assessments will be made once the size of the dataset becomes known and the github private repository is no longer a viable storage medium due to the size of the dataset, and further evaluations will be made on the storage costs.

# Algorithms & Analysis / ML Model Exploration & Selection

## Solution Approach

### **Systems Architecture**

The basic system architecture and process overflow that has been planned for this project is as follows:

Problem Statement Definition > Data Collection > Data Cleaning and Preprocessing > Exploratory Data Analysis > NLP > Synthetic Data Generation > Machine Learning Algorithms > Model Validation

It can be depicted in the form of a flow chart as follows:

Diagram

Description automatically generated

Figure 3.1: Process overflow in the project

### **Systems Security**

The dataset being secure is the only security risk we are anticipating for this projects, but since the chosen storage medium for the data set in this project is a private github repository with access privileges granted only to the project team, the dataset is stored in a secure manner without the risk of the data being leaked.

We will explore security concerns in the project as the need arises as the project progresses further.

### **Algorithms & Analysis**

We have conducted research on similar projects, which are projects in the same domain or projects with a similar problem statement. The algorithms that have been used extensively across all these projects are as listed below:

**Similar Projects:**

1. Development of a Predictive Risk Assessment Tool for Pregnancy Complications
2. Machine Learning-based Early Warning System for Pregnancy Complications
3. Predictive Modeling of Pregnancy Complications using Electronic Health Records (EHRs)
4. Automated Risk Prediction for Pre-eclampsia

**Algorithms used in similar projects:**

1. Supervised learning
2. Gradient boosted decision trees
3. Machine Learning
4. Deep Learning
5. AUC (area under the curve)
6. Cost sensitive hybrid model (CSHM)
7. Elastic net algorithm
8. Gradient boosting algorithm
9. Cross-fold validation
10. Predictive modeling

Based on the understanding we have about our dataset, the primary target variable that we must focus on is ‘risk factor’, which classifies each pregnancy based on its level of risk (low, medium, high). Since this is a classification problem as opposed to a prediction problem, the focus of this project will be on applying as many classification algorithms as possible to the data and validating the results of each algorithm. The most appropriate algorithm that provides the most accurate results for this data will be narrowed down through several model validation techniques.

Some of the classification algorithms that are the frontrunners are as below:

* Logistic Regression
* Random Forests
* Naïve Bayes
* K-Nearest neighbors
* Decision tree
* Support Vector Machines

The results of these algorithms can be validated through the methods mentioned below:

* ROC AUC
* Precision-Recall validation
* Accuracy metrics and plots

## Machine Learning

Once the dataset gets finalized, we plan on performing the below steps to decide on the machine learning model that delivers the best possible solution to the problem statement:

Evaluating the data based on the following parameters:

* Gathering a clear understanding of the project goal
* Analyzing the level of processing required for the dataset
* Finding the level of linearity of the data
* Narrowing down the number of features and parameters

Choosing one of the following, based on the linearity of the data and the data-types:

* Unsupervised Machine Learning Algorithms
* Supervised Machine Learning Algorithms
* Semi-supervised Machine Learning Algorithms
* Reinforcement Machine Learning Algorithms

The following machine learning algorithms have been performed so far on the dataset with ‘risk factor’ as the target variable, and their accuracies are as follows:

|  |  |
| --- | --- |
| **Algorithm** | **Accuracy** |
| Decision Trees | 76.9% |
| Random Forests | 81.2% |
| Gaussian NB | 60.8% |
| Multinomial NB | 77.8% |
| SVM | 84% |

### **Model Exploration**

Every machine model that is applied to the data gets tested by the following metrics:

* Train/test split and validation
* Accuracy (correct predictions / all predictions)
* Precision (true positives / (true positives + false positives))
* Recall (true positives / (true positives + false negatives))
* Regression Metrics
* Mean Squared Error (MSE)
* R2 coefficient
* Validation curves (underfitting / overfitting)

### **Model Selection**

Based on the results obtained for the above-mentioned validation methods, the outputs for all the algorithms and machine learning models will be compared and the best possible model would be selected for the dataset, that best answers and explains the problem statement.

# Visualizations / ML Model Training, Evaluation, & Validation

## Overview

This project is primarily a machine learning project, with the main goal being classifying the risk factors of future pregnancies by analyzing previous pregnancies. Thousands of new records have been created for the dataset from the existing 97 records using ‘Synthetic Data Vault’ library, to have adequate data to attain accurate results from the machine learning algorithms.

A few preliminary visualizations have been made on the preprocessed dataset to better understand the relationship between the variables.

## Visualizations

Since the target variable is the risk factor of the pregnancy, a few visualizations have been made to interpret its relation with the other variables in the dataset.

Chart, bar chart

Description automatically generated

Figure 4.1: Ethnicity vs Risk Factor

The above barplot suggests that the risk factor in general is higher in women who belong to Hispanic, Latino, Black and African ethnicities. It also indicates that ethnicity is a significant detrimental factor in determining the risk level of a pregnancy.

Chart, box and whisker chart

Description automatically generated

Figure 4.2: Age at Pregnancy vs Risk Factor

The above boxplot depicts the effect that the age of the woman at the time of their pregnancy has on the level of risk of the pregnancy. It can be seen that the women with higher maternal age are relatively prone to having higher risk associated with their pregnancies.

**4.2.1 Probabilities of occurrence of complications**

Identification of Causal factors: Based on the preprocessing performed on the past medical column (section 2.4.1), the causal factors that are responsible for the onset of each of the pregnancy complications have been identified and are as listed below:

**Gestation Diabetes:**

* Obesity
* Diabetes
* High Blood Pressure
* Higher Maternal Age

**Pre-Eclampsia:**

* Obesity
* Diabetes
* High Blood Pressure
* Higher Maternal Age
* Heart Disease

**Unnecessary C-Section:**

* Higher Maternal Age
* Previous Gynecological Surgery
* High Blood Pressure

**Abnormal weight gain of the baby:**

* High Blood Pressure
* Diabetes
* Higher Maternal Age

**Pre-term labor:**

* Diabetes
* Obesity
* High Blood Pressure
* Use of tobacco
* Consumption of Alcohol
* Use of street drugs

Based on the above causal factors, five new columns have been created for each of the 97 patients in the original dataset. The ‘past medical history’ column has been analyzed for each patient and flags have been assigned for each of the complication – 1, if the patient has a probability of developing the complication during the course of their pregnancy and 0 – if the patient has no probability of developing the complication. For example, in the below picture, the first patient has diabetes listed in their past medical history. Since diabetes is one of the causal factors of gestational diabetes, 1 has been assigned to the gestational diabetes column for that patient.

Table

Description automatically generated

## The probabilities of occurrence of complications has also been calculated for each of the patients based on the number of causal factors listed in their past medical history. For example, since the first patient only has diabetes and not the other four conditions that are responsible for causing gestational diabetes, it has been determined that the patient has a 1 in 5 chance (20%) of developing gestational diabetes during their pregnancy. The percentages can be visualized in the below table and plots.

Table

Description automatically generated

Figure 4.3: Statistics of probability of occurrence of Unnecessary C-section

Figure 4.4: Statistics of probability of occurrence of Abnormal baby weight

Figure 4.5: Statistics of probability of occurrence of pre-term birth

Figure 4.6: Statistics of probability of occurrence of Gestational Diabetes

Figure 4.7: Statistics of probability of occurrence of Pre-Eclampsia

## 4.3 Machine Learning

### **4.3.1 Model Training**

The total 1097 records in the dataset have been split into 75% training set and 25% testing set. The 75% of the data has been used to train the machine learning models, with the ‘risk\_factor’ column being the target variable and the remaining variables in the dataset being predictor variables. The models have been trained to associate the values of the predictor variables to the level of risk of the pregnancies.

For the purpose of this project, since the target variable is a categorical variable as opposed to a numeric one, this is a classification problem, which has been tackled by the following machine learning algorithms:

* Decision Trees
* Random Forests
* Gaussian Naïve Bayes
* Multinomial Naïve Bayes
* Support Vector Machine (SVM)

### **4.4.2 Model Evaluation**

The k-fold cross validation method has been used to determine the accuracies of the above-mentioned models, and they have yielded the following results:

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Accuracy (without synthetic data generation)** | **Accuracy (with synthetic data generation using SDV)** | **Accuracy (with synthetic data generation using CTGAN)** |
| Decision Trees | 76.9% | 89.7% | 45% |
| Random Forests | 81.2% | 89.1% | 35.9% |
| Gaussian NB | 60.8% | 51% | 37.2% |
| Multinomial NB | 77.8% | 90.2% | 46.1% |
| SVM | 84% | 89.4% | 40.3% |

Although the models and validations performed on the raw data (97) did not incur any errors, there were a few warnings associated with lesser data than required.

Hence thousands of new records have been synthetically generated by sampling and replicating the already existing 97 records using the ‘sdv’ and ‘ctgan’ libraries in python and a bigger dataset has been created.

When evaluated using k-fold cross validation, the dataset generated by sdv library has yielded better accuracy scores with machine learning, as can be seen in the above table.

To decide on the number of records to generate to ensure optimal accuracy, we have generated more synthetic records and noted the accuracies for each, as can be seen in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | **SDV (1097 records)** | **SDV (2097 records)** | **SDV (3097 records)** | **SDV (5097 records)** |
| Decision Trees | 89.7% | 90.5% | 90.7% | 90.6% |
| Random Forests | 89.1% | 90% | 90.7% | 89.9% |
| Gaussian NB | 51% | 58% | 49% | 47% |
| Multinomial NB | 90.2% | 90.2% | 90.4% | 90.3% |
| SVM | 89.4% | 90.3% | 90.5% | 90.4% |

3000 new records is the optimal number of records for this dataset to obtain the best accuracies. Decision Trees Algorithm is the one that gave the best accuracy at 90.9% at 3097 records.

**4.4 Natural Language Processing (NLP)**

For the machine learning analysis that has been performed until now, just the primary dataset has been considered and has been synthetically generated and analyzed. In addition to the primary dataset, Metronomic Inc has also provided two additional datasets containing data pertaining to the appointments that the patients have had throughout the course of their pregnancies. The significant columns in these datasets have been identified as ‘Appointment reasons’, ‘Clinical Observations’ and ‘Doctor’s Advice Notes’. Since these columns are entirely textual, they could not be analyzed through traditional machine learning methods.

For the purpose of analysis, these columns have been merged into the primary dataset with respect to the patient Id and Natural Language Processing (NLP) has been used to analyze these three columns. In order to make the data viable for NLP, a few preprocessing steps have been performed on the same as follows:

* Entire text has been converted into lowercase
* The columns have been converted into tokens
* Lemmatization: converting multiple forms of the same word into one word for the sake of being recognized as one by NLP (eg: maintain, maintenance, maintaining etc. will be converted into one form of the word maintain, and will be counted as such by NLP)
* Stopwords have been removed (eg: words such as patient, her, the, for, she etc. that are not significant to the analysis)
* Infrequent words have been removed

**4.5 Visualizing NLP**

The results of the NLP have been visualized with the aid of word plots and frequency distribution plots as seen below:

**Word Plots:**

Appointment reasons:

Text

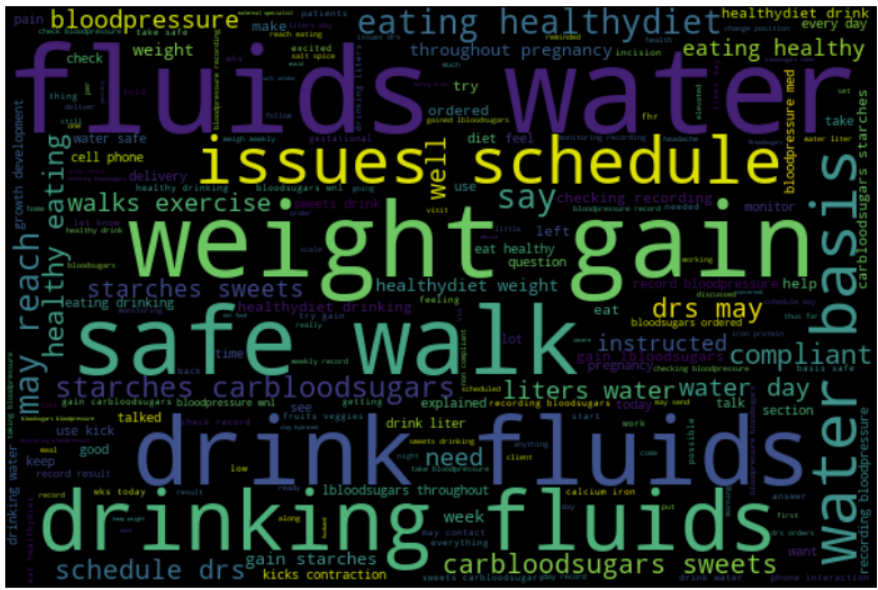
Description automatically generated

Clinical Observations:

Text

Description automatically generated

Doctor’s Advice Notes:



**Frequency Distribution plots:**

Appointment reasons:

Chart, bar chart

Description automatically generated

Clinical Observations:

Chart, bar chart

Description automatically generated

Doctor’s Advice Notes:

Chart, funnel chart

Description automatically generated

**4.6 Results of NLP**

**Appointment Reasons:**

As observed by the frequency distribution plots and wordplots, the most important and frequently observed words across the records of all the patients in their appointment reasons are as follows:

* Glucose Monitoring
* Weight
* Blood Pressure
* Fetal Doppler
* Weight Management

**Clinical Observations:**

As observed by the frequency distribution plots and wordplots, the most important and frequently observedwords across the records of all the patients in their clinical observations are as follows:

* Blood Sugars
* Blood Pressure
* Weight

**Doctor’s Advice Notes:**

As observed by the frequency distribution plots and wordplots, the most important and frequently observed words across the records of all the patients in their Doctor’s Advice Notes are as follows:

* Water
* Weight gain
* Blood Sugars
* Blood Pressure
* Eating Healthy
* Exercise

# Findings

The primary and most significant findings of this project are as below:

The probabilities of occurences of the five major pregnancy complications have been determined among the 97 patients. Based on our research, we have been able to offer speculations as to the probabilities of the patients developing complications during their pregnancy.

Since the original data just had 97 records, which is not adequate for an accurate machine learning algorithm, we have looked into synthetic data generation methods to multiply and expand on the existing data.

Multiple synthetic data generation synthesizers have been looked into – among which Fast\_ML provided us with the highest quality score. Within Fast\_ML, Synthetic Data Vault (SDV) method has generated data which provided us with the highest accuracy in machine learning.

Since the target variable ‘risk factor’ is a categorical column, focus has been put on classification algorithmns to classify the risk factor of new pregnancies. Among all the tested classification algorithmns, highest accuracy has been observed at 3000 new records with SVM – linear.

Results of Natural Language Processing: in order to identify the major concerns throughout the course of the pregnancy – the correlation between the results of NLP performed on all three textual columns has been observed, and the common problems faced by majority of the patients are: High Blood Pressure, Blood Sugars and Abnormal weight gain.

# Summary

This project has set out to analyze the data on completed pregnancies and use the trends and patterns identified in the data to determine the risk level of new and ongoing pregnancies. A secondary goal of the project is also to identify the issues that the patients have faced throughout the course of their pregnancies, by analyzing the data on the appointments they have had during their pregnancies.

The results of the project are promising and a machine learning model has been developed, which is capable of analyzing existing data and classifying new pregnancies based on their risk factors. It could be extremely beneficial to interpret the results of the project and incorporate the findings into new pregnancies, such that if a pregnancy gets classified as being risky, necessary and appropriate measures can be taken in order to mitigate the risk-causing factors as early on in the pregnancy as possible, in order to prevent complications further down the line. It is also advisable to conduct this study during the early stages of the pregnancy, as treating the complications in the later stages is significantly riskier and costlier than in the early stages.

# Future Work

The results and the scope of this project have been slightly limited by the lack of comprehensive data on the completed pregnancies. Should there be data on more number of patients available in the future, without the need of having to generate data synthetically, the results in classifying new pregnancies would be more accurate. The data in the future could also include more demographic, biological, historical and clinical factors pertaining to the completed pregnancies. With the increase in the number of predictor variables in a machine learning model, the sensitivity of the target variable also increases – which indicates that with the presence of higher number of predictor variables (describing factors), the new pregnancies can be classified with increased accuracy as compared to the current results.

Appendix

Appendix A: Glossary

|  |  |
| --- | --- |
| Term | Definition |
| Natural Language Processing (NLP) | The study of making computers understand, interpret human language. It is the study of teaching computers to comprehend human language as it is spoken and written, and to respond in a way that is understandable to people. |
| Synthetic Data Generation | The process of creating artificial data that mimics real-world data, but is not actually collected from the real world. It involves using mathematical algorithms and models to generate data points that follow the statistical patterns and properties of the original data. |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Appendix B: GitHub Repository

**Overview**

All the files that have been utilized in the project and the files that have been generated in the process have been uploaded to a github private repository.

**GitHub Repository Link**

<https://github.com/TeamBlitz-Daen/Blitz>

**GitHub Repository Contents**

1. Original datasets

2. Datasets generated during analysis and preprocessing

3. Code files

Appendix C: Risks

The risks that we have faced throughout the course of the project have been listed in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Risk Name** | **Description** | **Probability** | **Impact** | **Mitigation Steps** |
| Diverse Data | Multiple data types might not get analyzed in single model. | High | Medium | Converting categorical data into numeric by some type of encoding |
| Inadequate fields | Not enough fields for machine learning to accurately predict the risk factor. | High | Medium | Generating as many new fields as possible (eg: LMP – date of birth = age at pregnancy) |
| Insufficient records | There are under a hundred records in the dataset. | High | High | Synthetic Data Generation techniques to expand on the existing data |

**Mitigation Status of risks:**

**Diverse Data** – Mitigated: Categorical data columns have been converted into numerical format using one hot encoding and have been incorporated into machine learning algorithms.

**Inadequate Fields** – Mitigated: As many new fields as can be generated using the existing data have been generated. Example: The past medical history column has been used to generate 28 new columns in the data – one for each of the conditions mentioned in the column. A new column has been generated for age of the patient at the time of their pregnancy by calculating LMP – date of birth.

**Insufficient Records** – Mitigated: Adequate number of records have been synthetically generated using Synthetic Data Generation methods.

Appendix D: Agile Development

**Scrum Methodology**

This project has been managed by employing the scrum/ agile methodology. The entire project has been divided into five sprints, with a duration of 3 weeks each. The main steps performed in each Sprint have been segregated and listed in the flowchart below:

Diagram

Description automatically generated

**Sprint 1 Analysis**

The main focus of the first sprint was for the team to gain as much domain knowledge as possible, so as to facilitate a comprehensive understanding of the problem statement and the requirements of the client. Several projects in the same domain and projects with similar problem statements have been researched in this Sprint, in order to gain information on how similar problems have been solved before, and analyze if any of these steps can be applied to solve the problem statement in this project.

**Sprint 2 Analyis**

The main steps that have been performed in this sprint are conducting research on the five major pregnancy complications and the causal factors that are responsible for the occurrence of these complications. We have also decided on a storage medium for our project related files in this Sprint – and we have decided on using a Github Private Repository.

**Sprint 3 Analysis**

This sprint has been dedicated to preprocessing the data in order to make it feasible for the Machine Learning Algorithms to obtain the highest accuracy possible. Since the original dataset provided by the client had just under a hundred records, it was necessary for the team to generate synthetic data by replicating and multiplying the existing data – in order to have adequate data to obtain satisfactory results during the analysis. These steps have been performed in this Sprint as well. Several classification algorithms have been applied to the synthetically generated data, and they have been evaluated in the subsequent sprint.

**Sprint 4 Analysis**

This Sprint has been dedicated to performing visualizations on the data to ascertain the relationship between the predictor variables in the data and the target variable. The machine learning models that have been applied to the data in Sprint 3 have been evaluated to determine the best model for this data.

**Sprint 5 Analysis**

This Sprint has been about improving aspects of our presentation with each iteration of rehearsals, and preparing for the final showcase presentation.

References

Works Cited

# Bibliography

|  |  |
| --- | --- |
| [1] | "Using Machine Learning to Predict Complications in Pregnancy: A Systematic Review," Frontiers, 10 December 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fbioe.2021.780389/full. [Accessed 1 February 2023]. |
| [2] | "Trends in Pregnancy and Childbirth Complications in the U.S.," Blue Cross Blue Shield , [Online]. Available: https://www.bcbs.com/the-health-of-america/reports/trends-in-pregnancy-and-childbirth-complications-in-the-us. [Accessed 1 February 2023]. |
| [3] | " Maternal mortality," World Health Organization , [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/maternal-mortality. [Accessed 1 February 2023]. |

[4] Healio. (2021, August 17). ‘Time to get serious’: Rate of gestational diabetes increases in US. Retrieved from https://www.healio.com/news/primary-care/20210817/time-to-get-serious-rate-of-gestational-diabetes-increases-in-us

[5] GestVision. (n.d.). Preeclampsia in the United States & Worldwide. Retrieved from https://www.gestvision.com/about-preeclampsia

[6] NaProTechnology. (n.d.). Prevention of Preterm Birth. Retrieved from https://naprotechnology.com/prevention/

[7] statista. (n.d.). Percentage of all live births by Cesarean delivery in the United States from 1997 to 2020. Retrieved from https://www.statista.com/statistics/184079/us-hospital-stays-with-cesarean-section-procedures-since-1997/

[8] DIHI. (n.d.). Development of a Maternal Early Warning System Using Machine Learning. Retrieved from https://dihi.org/project/development-of-a-maternal-early-warning-system-using-machine-learning/

[9] frontiers. (2022, January 19). Using Machine Learning to Predict Complications in Pregnancy: A Systematic Review. Retrieved from https://www.frontiersin.org/articles/10.3389/fbioe.2021.780389/full

[10] nature. (2017, November 27). Electronic Health Record Driven Prediction for Gestational Diabetes Mellitus in Early Pregnancy. Retrieved from https://www.nature.com/articles/s41598-017-16665-y

[11] ScienceDirect. (2020, May). Early prediction of preeclampsia via machine learning. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S2589933320300306

[12] Kaggle. (n.d.). Maternal Health Risk Data. Retrieved from https://www.kaggle.com/datasets/csafrit2/maternal-health-risk-data