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. The results are promising.

Probabilities of occurrence of complications:

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.

	past_medical_history	gestational diabetes	pre-eclampsia	pre-term labor	abnormal weight	unnecessary c-section
2	diabetes	1	0	0	0	0
3	Covid,obesity,drug_allergy_list,advanced_maternal_age	1	1	0	1	0
4	obesity,diabetes,hepatitis_liver_disease,history_of_abnormal_pap	1	1	0	0	0
5	diabetes,allergies,psychiatric,drug_allergy_list	1	0	0	0	0
6	obesity	1	1	0	0	0
7	obesity,hypertension,history_of_abnormal_pap	1	1	1	1	1
8	obesity	1	1	0	0	0
9	other_past_medical_history	1	0	0	0	0
10	obesity,diabetes,hypertension	1	1	1	1	1
11	obesity,hypertension	1	1	1	1	1

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.

	past_medical_history	gestational diabetes	pre-eclampsia	pre-term labor	abnormal weight	unnecessary c-section
2	diabetes	0.2	0.2	0.2	0	0.33
3	Covid,obesity,drug_allergy_list,advanced_maternal_age	0.4	0.4	0	0.33	0.33
4	obesity,diabetes,hepatitis_liver_disease,history_of_abnormal_pap	0.6	0.4	0.2	0	0.33
5	diabetes,allergies,psychiatric,drug_allergy_list	0.2	0.2	0.2	0	0.33
6	obesity	0.2	0.2	0	0	0
7	obesity,hypertension,history_of_abnormal_pap	0.6	0.4	0.2	0.33	0.33
8	obesity	0.2	0.2	0	0	0
9	other_past_medical_history	0	0	0	0	0
10	obesity,diabetes,hypertension	0.6	0.6	0.4	0.33	0.66
11	obesity,hypertension	0.4	0.4	0.2	0.33	0.33

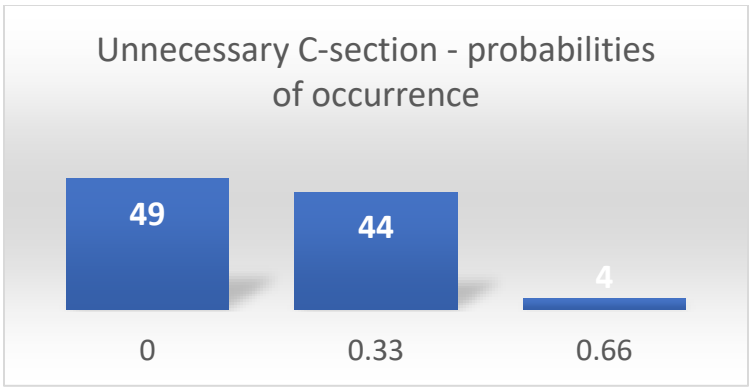

Figure 4.3: Statistics of probability of occurrence of Unnecessary C-section

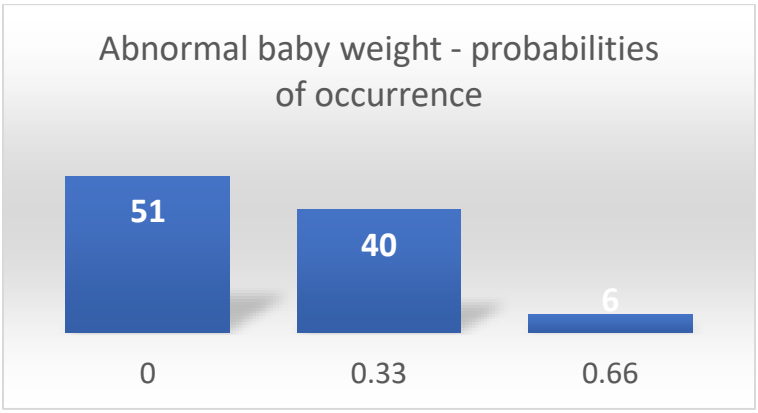

Figure 4.4: Statistics of probability of occurrence of Abnormal baby weight

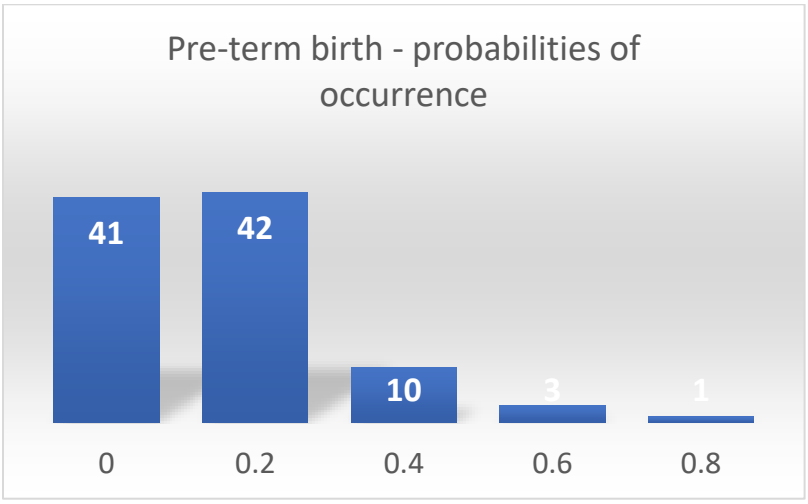

Figure 4.5: Statistics of probability of occurrence of pre-term birth

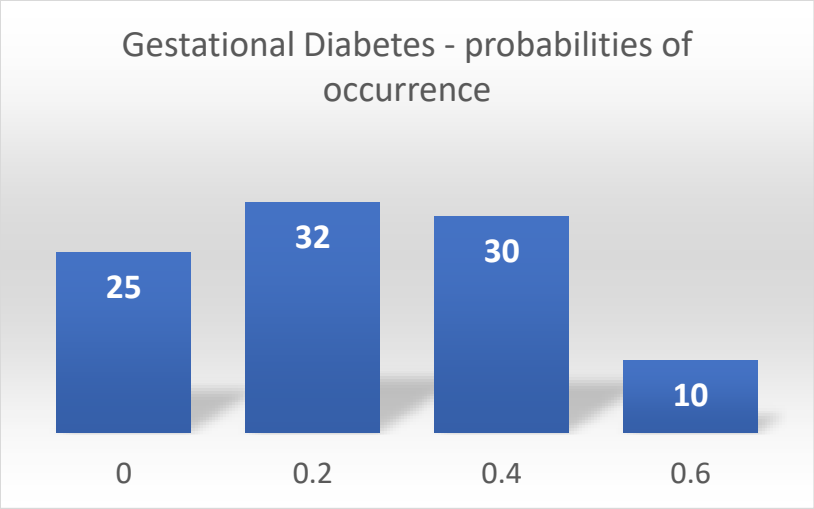

Figure 4.6: Statistics of probability of occurrence of Gestational Diabetes

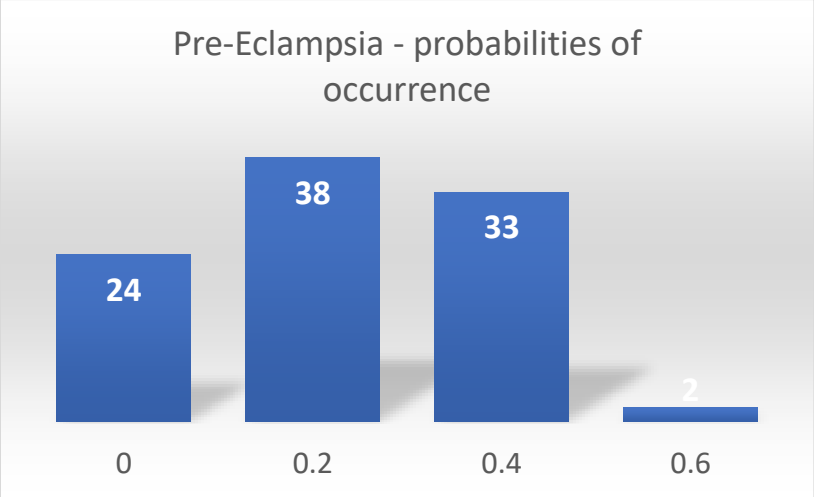

Figure 4.7: Statistics of probability of occurrence of Pre-Eclampsia

Machine Learning and Model Evaluation:

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.

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)
- 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%

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.

Findings:

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

The probabilities of occurrences 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.

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.