



Project Report

TOSOH PRE-NATAL RISK ASSESSMENT ENGINE

Under the guidance of

Dr. Santanu Mondal and Dr. Arnab Majumdar

Olivia Chatterjee

olivia@olivia.net.in

13th March 2025




Table of Contents

1	INTRODUCTION	2
2	PROJECT TASKS.....	2
3	PROJECT SOLUTION STEPS	2
4	SOLUTION DETAILS AND RESULTS.....	3
5	CONCLUSION	8

1 Introduction

This report provides a comprehensive overview of the project "Pre-Natal Risk Assessment Engine." The project aims to analyse the risk for open neural tube defects (ONTD), Down Syndrome, trisomy 18, and Smith-Lemli-Opitz Syndrome (SLOS) using modern machine learning techniques. The project involves multiple stages, including data cleaning, exploratory data analysis, feature engineering, application of Artificial Neural Network, Random Forest Classifier and XGBoost algorithms for the prediction of various risks involved at the time of pregnancy.

2 Project Tasks

The main objective of PINE (Pre-Natal Investigation Engine) is to develop a system involving modern machine learning techniques for maternal serum screen assessment to evaluate the risk of chromosomal disorders in fetuses. The system leverages machine learning models to dynamically calculate test result medians, derive Multiples of the Median (MoMs), and predict the probability of conditions such as:

- Down Syndrome (Trisomy 21)
- Trisomy 18
- Open Neural Tube Defects (ONTD)
- Smith-Lemli-Opitz Syndrome (SLOS)

PINE leverages maternal and pregnancy-related data to generate highly accurate, data-driven assessments of the chromosomal risks. The backend system collects and processes clinical data, while machine learning models (Random Forest, XG Boost) enhance the precision of median calculations and risk predictions, ensuring better early detection of fetal abnormalities.

3 Project Solution Steps

DATA: This data gives maternal information, pregnancy related information, test results and computed parameters.

This project mainly predicts the probability of four chromosomal risks:

- Down syndrome
- Open neural tube defects (NTDs)
- Trisomy 18
- Smith-Lemli-Opitz syndrome (SLOS)

To achieve this, we utilize a machine learning model with carefully selected input variables. These variables include:

- Maternal factors: Weight, age, diabetic status
- Gestational age: At ultrasound (USG) and at the time of the test
- Number of fetuses
- Test results: AFP, hCG, and estriol (EST)
- Calculated MoMs: AFP MoM, hCG MoM, EST MoM

The steps are as follows:

1. The provided data is adequately privatized to prevent sensitive PII/PHI information leak. These models are built on privatized data.
2. Removed columns having too many null values.
3. Checked for missing data and replaced them (if possible), else removed the whole column (in case of too many missing data).
4. Feature engineering is used to clean the data further.
5. Did an exploratory data analysis for the given set of data.
6. Removed outliers, if any.
7. Plotted a correlation matrix to check the relation among the various attributes
8. Applied neural network to identify the risks after checking the relationship of various parameters with the outputs (risks).
9. Graphs and evaluation metrics were used to visualize the patterns and calculate the accuracies.
10. The median values are calculated by dividing the result by the Multiple of Median value (MoM).
 - Unlike the original software, which uses inputted medians adjusted by maternal factors, this system dynamically calculates medians.
 - To achieve this, we first derive medians by dividing existing test results by their corresponding MoMs (multiples of the median). Then, we train a machine learning model on 80% of the data (collected up to September 2024) to predict these medians.
 - For a new patient, the system first predicts the medians using the trained ML model and the patient's specific input variables. MoMs are then calculated by dividing the patient's test results by these predicted medians. Finally, these MoMs, along with other clinically relevant variables, are used to predict risks.
11. Since the accuracies were not high enough, other machine learning models like Random Forest and XG Boost were used. These algorithms provide an almost perfect R2 Score.

4 Solution Details and Results

1. Exploratory data analysis

- Dataset structure

The given data contains patient information of the past 10 years. It has around 10,255 rows and 143 columns. It contains the following information:

Input Columns:

Maternal Information:

- Date of birth (DOB)
- Race
- Weight
- Diabetic status
- Smoking habit
- Family history of open neural tube defects (NTDs)

Pregnancy-Related Information:

- Last menstrual period (LMP) date
- Ultrasound (USG) date
- Gestational age estimated by USG
- Confirmation of LMP date based on USG estimation
- Date of sample collection

Test Results:

- Alpha-fetoprotein (AFP)
- Human chorionic gonadotropin (hCG)
- Unconjugated estriol (uE3)

- Inhibin A (INHA)

Computed Parameters:

Based on a subset of the above parameters, the following are calculated using rule-based methods:

- Maternal age at the time of the test
- Gestational age at the time of the test

Output Columns:

- Down syndrome
- Open neural tube defects (NTDs)
- Trisomy 18
- Smith-Lemli-Opitz syndrome (SLOS)

- Data cleaning

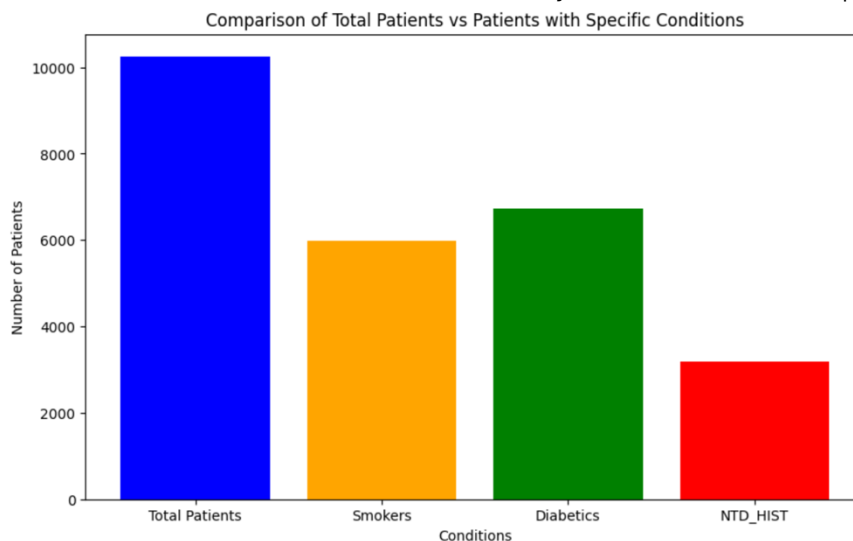
Cleaning the data to remove unnecessary features. This includes:

- Missing Data check
- Dropped null valued columns
- Reduced the final shape of the data to 10254 x 32

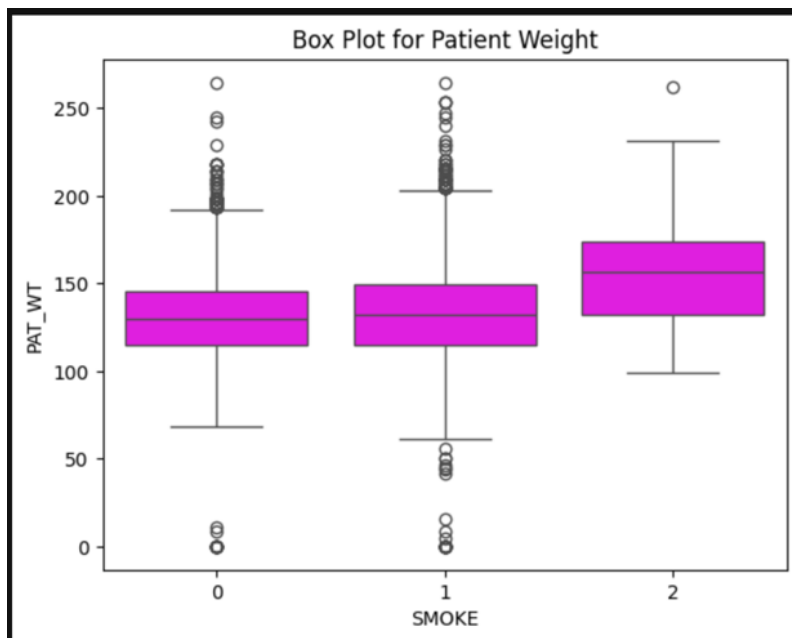
- Feature extraction

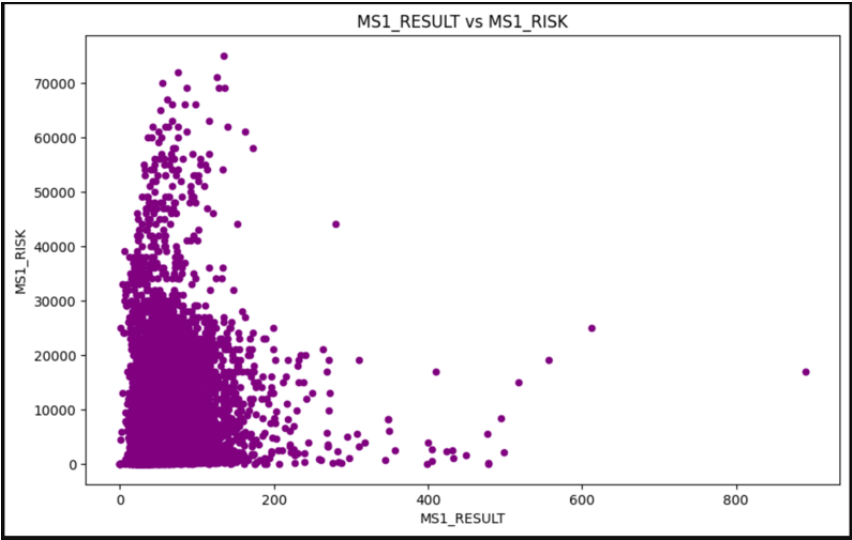
After extracting the important features, I found that Down syndrome, Open neural tube defects (NTDs), Trisomy 18 and Smith-Lemli-Opitz syndrome (SLOS)

are the most important columns which were later analyzed (from the graphs) to obtain the results.



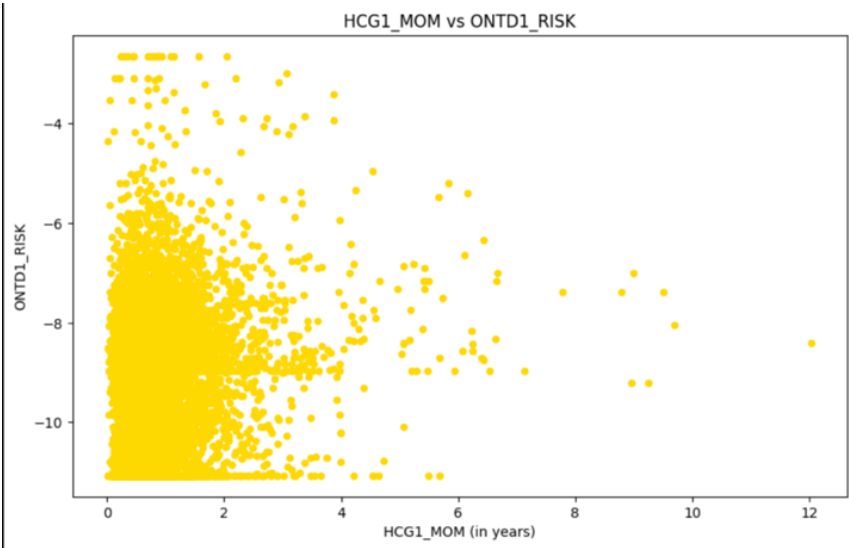
Shows the various conditions affecting a patient and compares them with the total number of patients



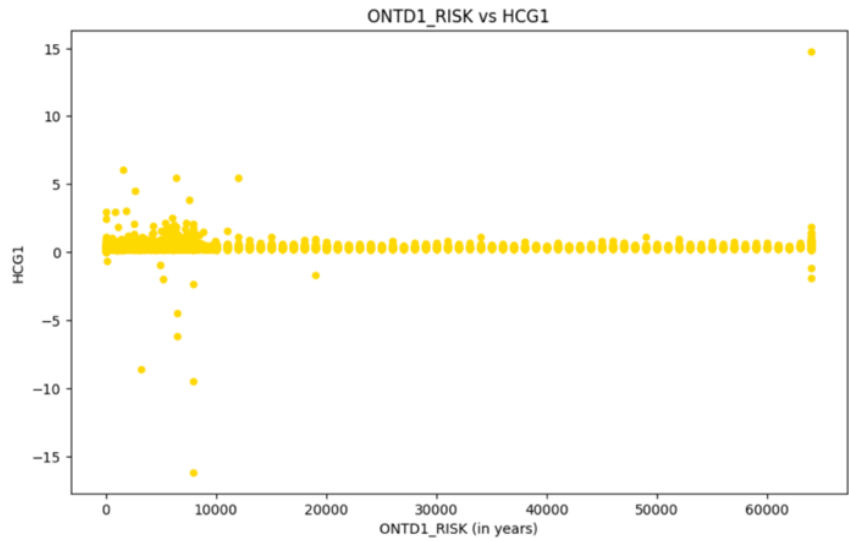


Displays the outliers and range of patients weights that might affect them

Plot of Down Syndrome



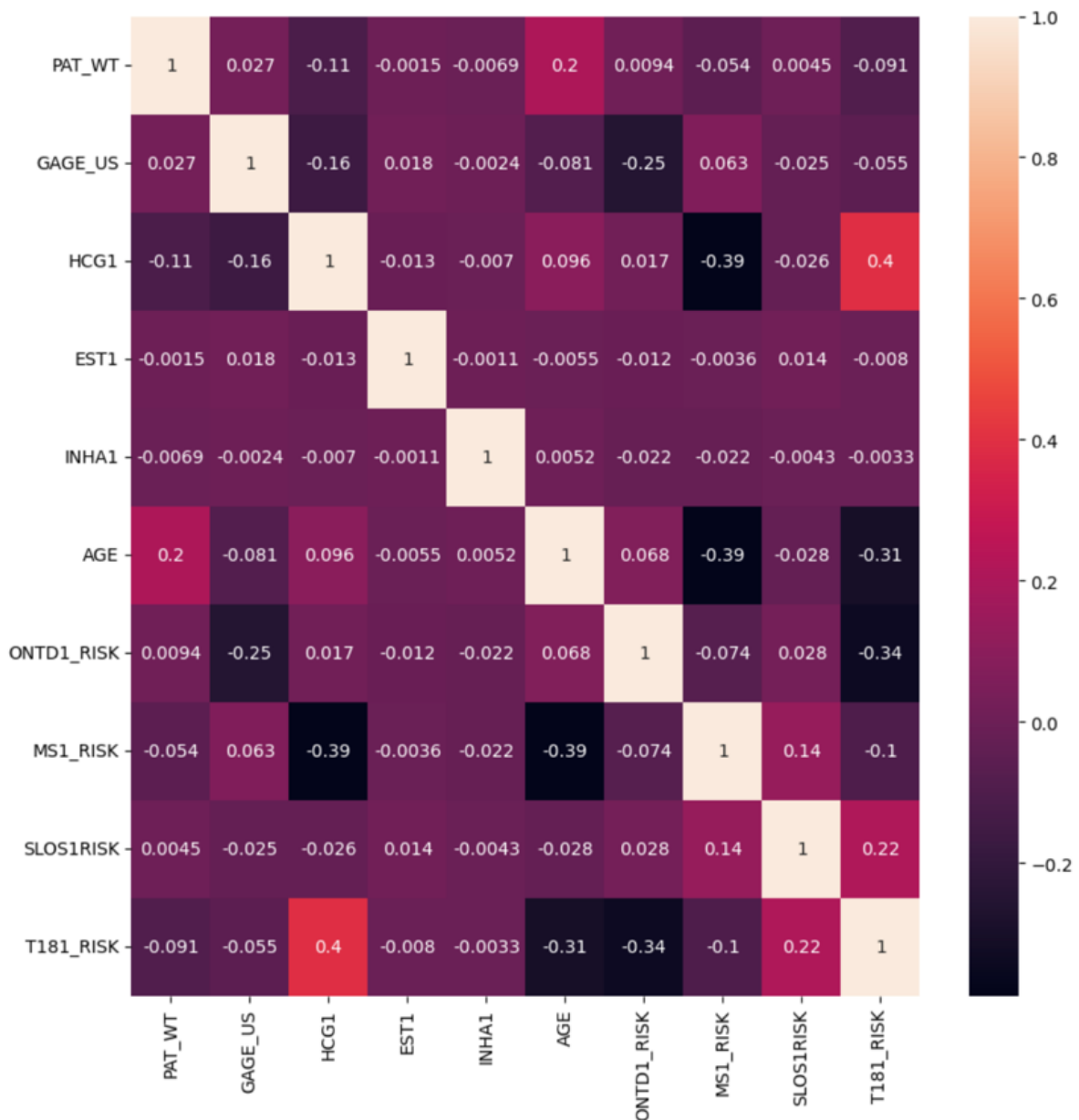
Result vs Down Syndrome Risk



Plot of Human Chorionic Gonadotropin Multiple of Median vs inverse of the logarithm of Open Neural Tube Defects Risk

Plot of Open Neural Tube Defects Risk vs inverse of the logarithm of Human Chorionic Gonadotropin

- Correlation Heatmap
Relation between various features in the dataset.



2. Artificial Neural Network

Multiple ANN models are trained taking similar inputs (X) each time, but different output (y) values.

- Training using T181_RISK as the output value
 - Hidden Layers: 64, 32, 1
 - Activation Functions: Relu, Relu, Linear
 - Loss: Mean Squared Error
 - Mean Squared Error: 615481489.7044336
 - R2 Score: 0.24281211363695843
- Hidden Layers: 32, 32, 1
 - Activation Functions: Relu, Relu, Linear

- Loss: Mean Squared Error
 - Mean Squared Error: 0.0023197734526408005
 - R2 Score: 0.009681165316187679
- Training using T181_RISK, MS1_RISK, ONTD1_RISK and SLOS1RISK as the output values
 - Hidden Layers: 64, 32, 4
 - Activation Functions: Relu, Relu, Linear
 - Loss: Mean Squared Error
 - Mean Squared Error: 1854669652.854947
 - R2 Score: -3.9931356505435236
- Training using the inverse of T181_RISK as the output value
 - Hidden Layers: 64, 32, 4
 - Activation Functions: Relu, Relu, Linear
 - Loss: Mean Squared Error
 - Mean Squared Error: 463505171.12258166
 - R2 Score: 0.4208764792631603
- Training using the inverse of MS1_RISK as the output value
 - Hidden Layers: 64, 32, 4
 - Activation Functions: Relu, Relu, Linear
 - Loss: Mean Squared Error
 - Mean Squared Error: 38872069.950368926
 - R2 Score: 0.5737472145168001
- Training using the inverse of ONTD1_RISK as the output value
 - Hidden Layers: 64, 32, 4
 - Activation Functions: Relu, Relu, Linear
 - Loss: Mean Squared Error
 - Mean Squared Error: 143425041.8956394
 - R2 Score: 0.6942693037225225
- Training using the inverse of SLOS1RISK as the output value
 - Hidden Layers: 64, 32, 4
 - Activation Functions: Relu, Relu, Linear
 - Loss: Mean Squared Error
 - Mean Squared Error: 485922556.6807032
 - R2 Score: 0.20813126312315866

3. Random Forest

- Training using MS1_GEST, MULT_FETUS, PAT_WT as X values and HCG1_median as the y value
 - Cross Validation: 5
 - Scoring: R2
 - Best Score: 0.9922884862376448
- Training using MS1_GEST, MULT_FETUS, PAT_WT as X values and EST1_median as the output value
 - Cross Validation: 5
 - Scoring: R2

- Best Score: 0.9985930264300213

4. XG Boost

- Training using T181_RISK probability logarithmic transformation with a constant shift as the output value
 - Cross Validation: 5
 - Scoring: R2
 - Best Score: 0.995658689879636
 - R2 Score: 0.9959282559693705
- Training using SLOS1RISK probability logarithmic transformation with a constant shift as the output value
 - Cross Validation: 5
 - Scoring: R2
 - Best Score: 0.9796149690383462
 - R2 Score: 0.9937048883107146
- Training using MS1_RISK probability logit as the output value
 - Cross Validation: 5
 - Scoring: R2
 - Best Score: 0.9541522789275897
 - R2 Score: -49.425412835874575
- Training using MS1_RISK probability logarithmic transformation with a constant shift as the output value
 - Cross Validation: 5
 - Scoring: R2
 - Best Score: 0.9156724154669511
 - R2 Score: 0.9470154384738998
- Training using ONTD1_RISK probability logit as the output value
 - Cross Validation: 5
 - Scoring: R2
 - Best Score: 0.9924501908947148
 - R2 Score: 0.9922373787175732

5 Conclusion

After evaluating several machine learning algorithms, Random Forest emerged as the most precise, achieving an R-squared value of at least 99% on the final 20% of the test data. For a new patient, the system initially forecasts the medians using the trained machine learning model in conjunction with the patient's specific input variables. The Multiples of the Median (MoMs) are subsequently calculated by dividing the patient's test results by these predicted medians. Ultimately, these MoMs, together with other clinically relevant variables, are utilized to predict risks.