BrainStation Capstone Project Report - Gestational Diabetes Mellitus (GDM) Prediction

Problem Statement
This project primarily arose from my personal experience with Gestational Diabetes. Gestational Diabetes Mellitus (GDM) is a type of diabetes that occurs in the second to third trimester of pregnancy when the body is unable to produce enough insulin leading to a rise in the blood sugar levels. Early diagnosis and treatment of GDM can prevent adverse pregnancy outcomes for both mother and baby and reduce health cost. The Canadian Diabetes Association and the American Diabetes Association recommend that all pregnant women be screened for gestational diabetes between 24 and 28 weeks gestation, and earlier if high risk factors are present using the oral glucose tolerance test (OGTT).

B**ackground**
In Canada, between 3 – 20% of pregnant women develop gestational diabetes, depending on their risk factors. Women with GDM are at an increased risk of later developing type 2 diabetes. This project aims to better predict GDM based on medical, physical, social and other risk factors using supervised machine learning methods in order to improve birth outcomes and prevent associated infant and child morbidity. Machine learning algorithms are increasingly being used to identify risk factors and predict GDM .

Data source, format, structure and quality
The data was freely available on Kaggle as .xlsx at https://www.kaggle.com/datasets/sumathisanthosh/gestational-diabetes-mellitus-gdm-data-set. The original source of the data is from a research paper by A. Sumathi and S. Meganathan on predicting GDM using an Ensemble Classifier. "A. Sumathi and S. Meganathan, "Ensemble classifier technique to predict gestational diabetes mellitus (gdm)," Computer Systems Science and Engineering, vol. 40, no.1, pp. 313–325, 2022. https://www.techscience.com/csse/v40n1/44217". The data set has 3525 rows and 17 columns: 1 unique identifier, 15 predictor and 1 target variable. The data has a lot of missing values for 4 columns.

Data Cleaning and preprocessing

Data cleaning
The data cleaning process mainly involved filling missing values, dropping the rows was not an option as I would be losing over 50% of the data. There were no duplicate columns, an examination of the 4 duplicate rows showed they were actually unique. The columns were renamed to show uniformity.

Exploratory Data Analysis
I checked the distribution of the target variable(GDM) and examined the relationship between variables with correlation coefficients of r > 4.0 and GDM. This was to understand the relationship between those variables and GDM in relation to the real world. I discovered that these variables were indeed predictive of GDM. (Figure 1) shows the proportion of the GDM(1) and non GDM(0) classes, there are more 0 classes. (Figure 2) shows the relationship between GDM and prediabetes, pregnant women with prediabetes are more likely to have GDM. (Figure 3) shows the relationship between GDM and Oral Glucose Tolerance Test(OGTT), pregnant women

with a high OGTT result have GDM.

Figure 1

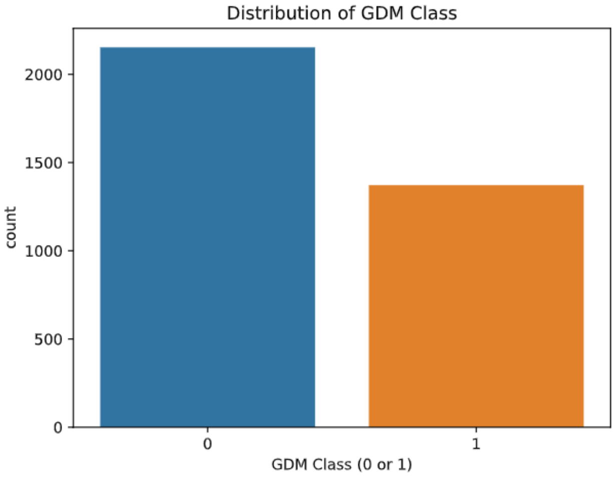


Figure 2

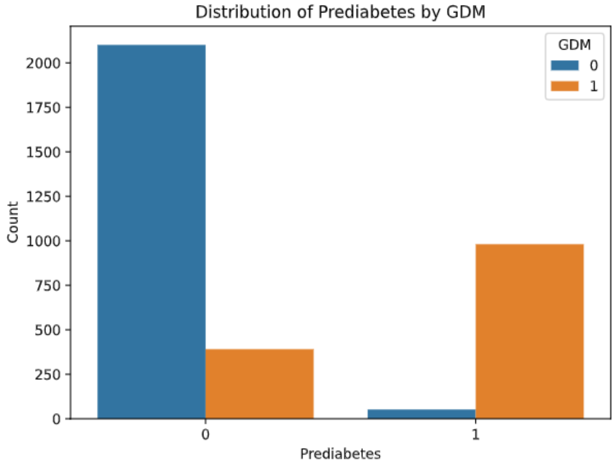
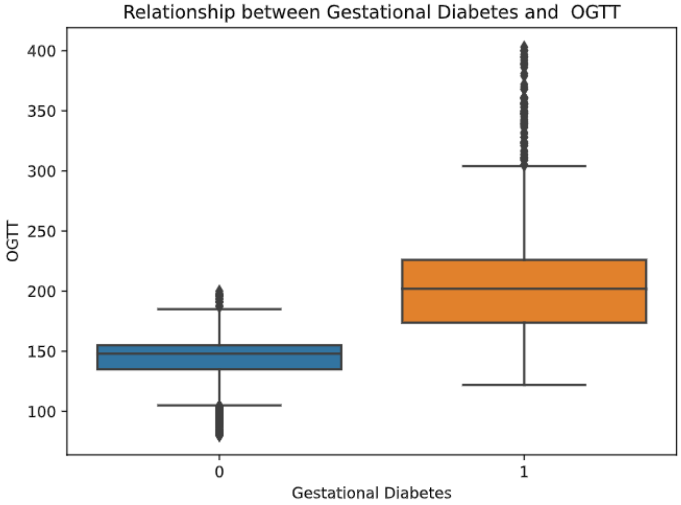


Figure 3



Feature Engineering

I created two new features, blood pressure and weight categories to understand how each category relates to GDM as high blood pressure and being obese or overweight were positively associated with GDM.

Insights, modeling, and results

Modeling and results

A baseline Logistic regression model built before feature engineering. A Logistic Regression, K Nearest Neighbors and Random Forest classifier models were built after feature engineering as they work with datasets with large features and allows for ebay optimization to find the best parameters for the best model. For further optimization, a GridSearch of Logistic Regression, K Nearest Neighbors and Random Forest, Gradient Boosting and XGBoost classifier models were built to find the best overall model. Data for all models were scaled except for Random Forest which does not require scaling. All models performed quite well with a higher recall score. Having a higher recall score compared to precision score in the best model is important to correctly identifying all GDM cases .

Insights

It was interesting to see the top 10 and bottom 10 predictive features of GDM, OGTT was the most predictive and HDL the least predictive. In examining the relationship between GDM and the weight categories, an obese pregnant woman was 3 times more likely to be diagnosed with GDM and a pregnant woman with Hypertension stage 2 was also 3 times more likely to be diagnosed with GDM.

Findings and conclusions

It was interesting to see that all models performed well with a high recall score which is good as we want to correctly predict all GDM cases as GDM. Next steps for this project would be to build a web app to apply the model and predict GDM. I would also like to explore deep learning algorithms in better predicting GDM for early detection and commencement of treatment.