Diabetes Early Prediction

Problem Statement

Diabetes is a chronic disease that results in high blood glucose. According to Statista.com, in 2021, 10.5% percent of global adult population suffered from diabetes, and the number is expected to rise over 12% by year 2045.

In this capstone project, the primary goal is to build statistic models to accurately predict diabetes based on symptoms. The second goal is to discover the leading factors to improve awareness so that the public, especially people with family history of diabetes, can be better educated to be alert of the symptoms.

Data Preprocessing

This dataset used contains the sign and symptom data of newly diabetic or would be diabetic patient. This has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor.

The dataset consist of total 15 features and one target variable named class.
1. Age: Age in years ranging from (20years to 65 years)
2. Gender: Male / Female
3. Polyuria: Yes / No
4. Polydipsia: Yes/ No
5. Sudden weight loss: Yes/ No
6. Weakness: Yes/ No
7. Polyphagia: Yes/ No
8. Genital Thrush: Yes/ No
9. Visual blurring: Yes/ No
10. Itching: Yes/ No
11. Irritability: Yes/No
12. Delayed healing: Yes/ No
13. Partial Paresis: Yes/ No
14. Muscle stiffness: yes/ No
15. Alopecia: Yes/ No
16. Obesity: Yes/ No
Class: Positive / Negative

Feature 2-16 are categorical variables and transformed to 1/0 before modelling.
Feature 1, ‘Age’ variable can be used as is for tree based machine learning models; for distance based model such as KNN, it needs to be scaled before modelling.
Class variable is transformed to 1/0 before modelling.

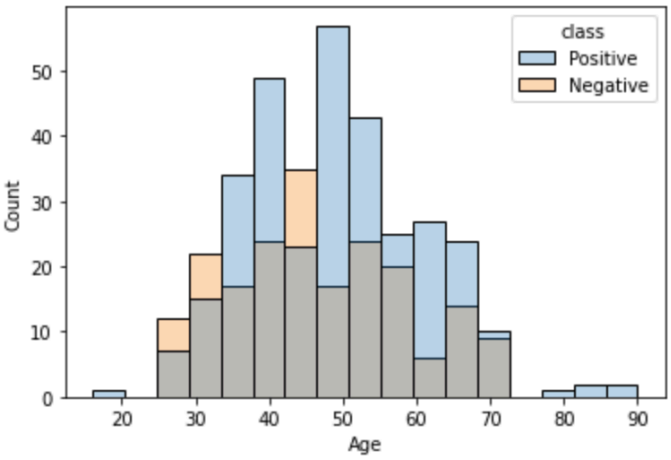
Duplicates are checked. Given the nature of the problem, unique identifier of patient is not available. There are ‘duplicate’ rows from the values perspective, but they are not removed due to: 1. they are not necessarily the same patient since lack of patient ID 2. The positive/negative ratio is at 60/40, removing ‘duplicates’ will make it 69/31; for the purpose of having relatively more balanced data, we shouldn't delete any 'duplicates'.

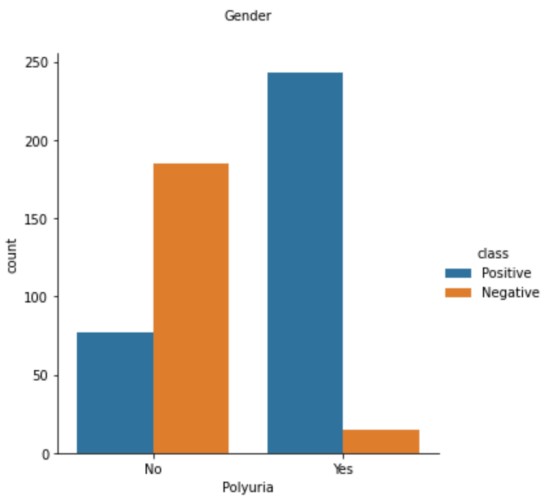
Check missing data, none exist.
Check outliers, none exist.

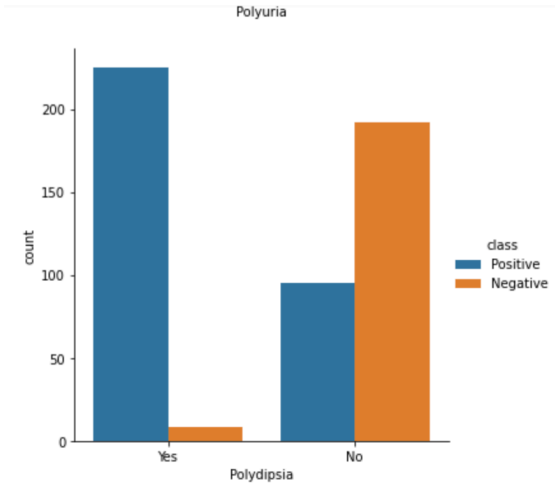
Exploratory Data Analysis

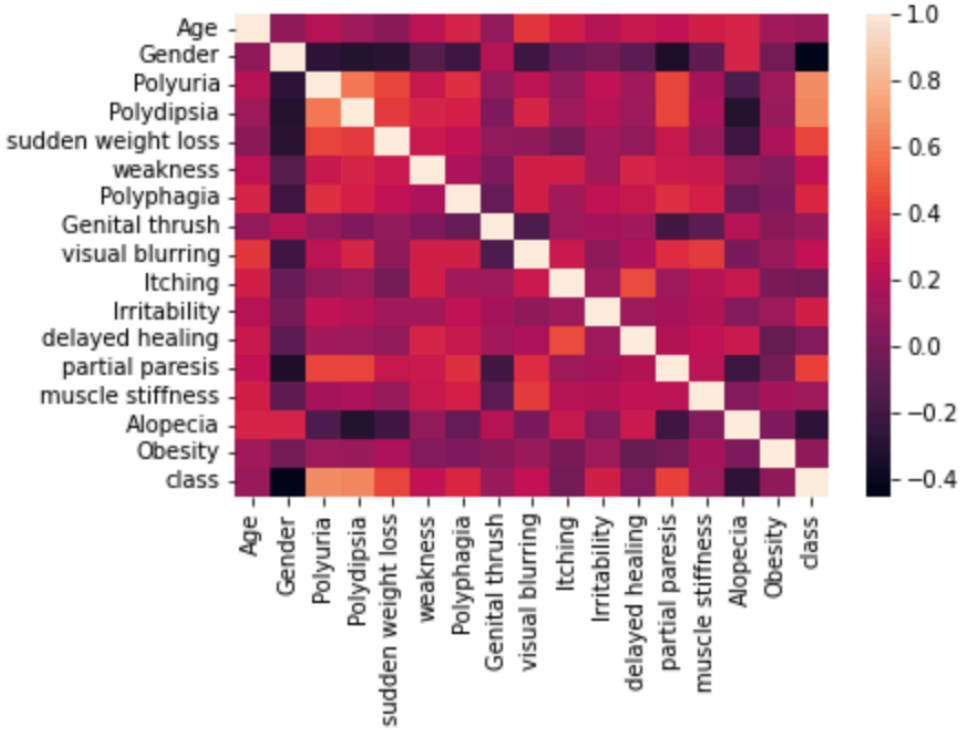
Age doesn't appear to have strong relationship with diabetes except for the few above 75 or below 20 years old showing positive.
Some categorical variables appear to be correlated with positive diabetes.

- Positive correlation includes Polyuria/ Polydipsia/Sudden weight loss/ Weakness/Polyphagia/visual blurring/ Irritability/partial paresis/muscle stiffness
- Negative correlation includes Genital thrush/Alopecia
- No correlation includes Itching/ delayed healing/obesity









Model Selection

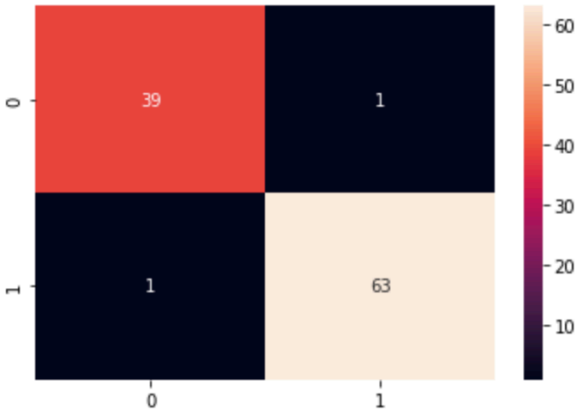
In total 4 different types of models have been trained and validated; Gradient Boosting, XGB, Random Forest and K Nearest Neighbour.

First, training dataset was used to tune hyper parameters through cross validation, using ‘recall’ score as the target metric. ‘Recall’ is used instead of ‘accuracy’ or ‘precision’ to align with the goal of minimise the type 2 error of failing to detect.
With the best parameters, all but KNN can reach a recall score of 98%, KNN is trailing with 96%.

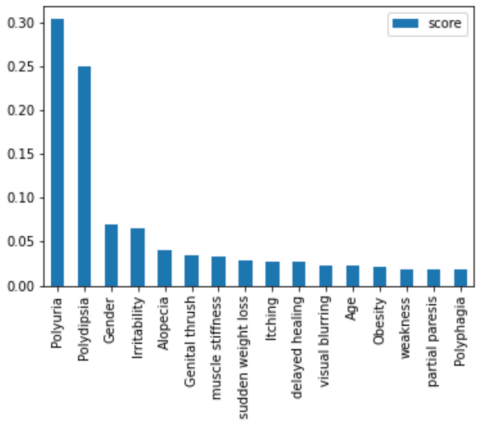
Then, full training dataset was used to retrain the GB, XGB, RF model using the best parameters. Test dataset was used to evaluate the result.

As shown on right, all 3 models result in the same confusion matrix, where 63 are correctly predicted to be positive, 39 correctly predicted to be negative, and 1 false positive and 1 false negatives.

In this case, all three models can be deployed. But there is a performance difference. XGB takes much less time to both train and predict, and thus more favoured in the situation where we need to constantly update the model.



Takeaways



We have successfully constructed a model that can predict diabetes with high accuracy, high precision, and high recall.

By plotting feature importance using ‘gain’ method, the most important 2 symptoms to look for is Polyuria and Polydipsia.