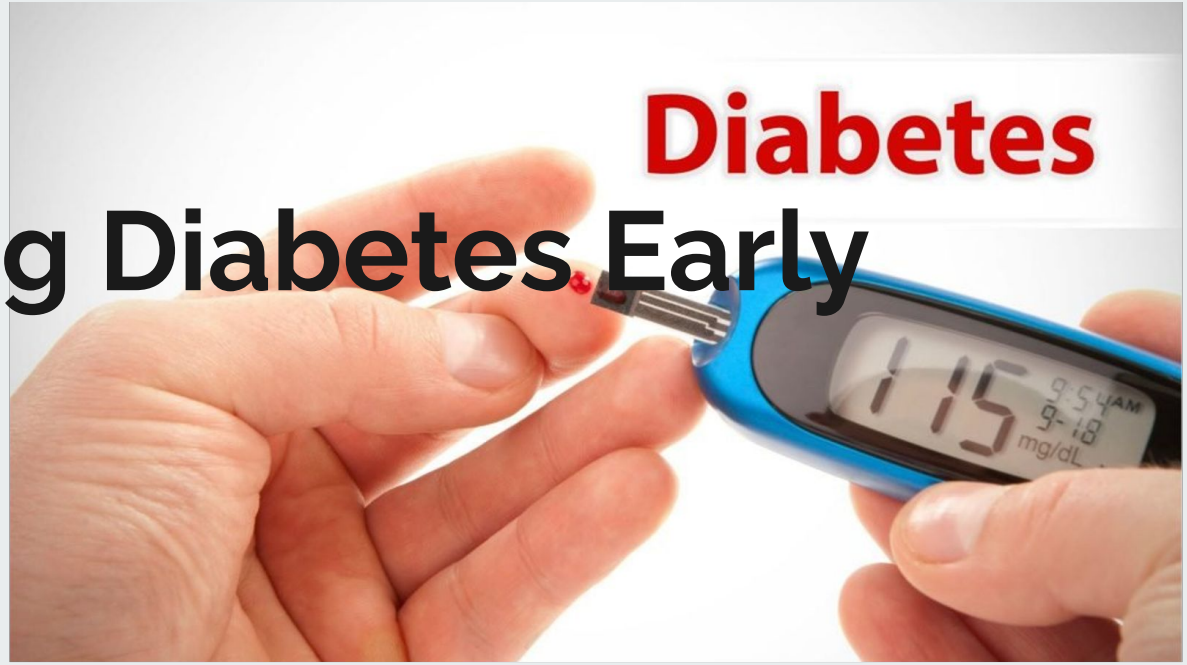
Detecting Diabetes Early

Daniel Kim



DIABETES SYMPTOMS

Problem Statement & Goal - How to Detect it?

Blurry Eyesight

Feeling Hungry

Sudden Weight Loss

Feeling Thirsty

Frequent Urination

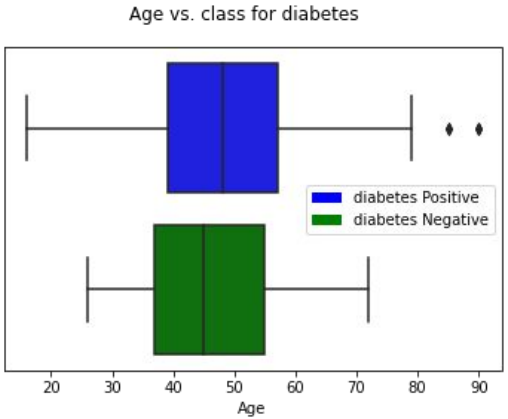
Data Wrangling

(https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.)

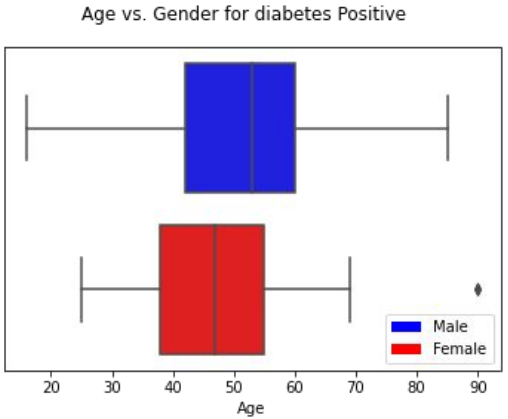
1. Detecting NaN & Redundancy
    a. There were no NaN
    b. Although there were more than 50% of duplicates, it is necessary to include the duplicates because the redundancies are not the same but different individuals having same symptoms even with the same 'Gender' and the same 'Age'..
    c. Except for Age, all other features are binary, so it was needed to do one hot encoding to make a machine learning understand the data better.
2. One Hot Encoding

    After one -hot encoding, all the other binary columns ending with such as '_Female', '_Negative', or '_No' were removed.

Exploratory Data Analysis - Box Plots



Older ‘Age’ group tends to have diabetes a little bit more.

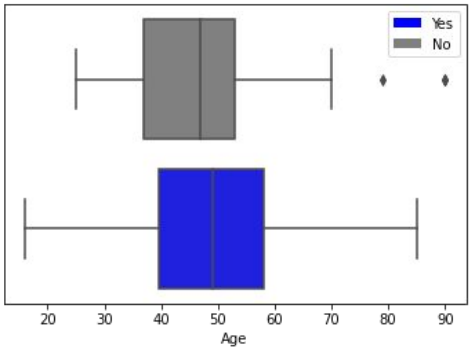


Male ‘Gender’ group with diabetes Positive tends to be older than Female one..

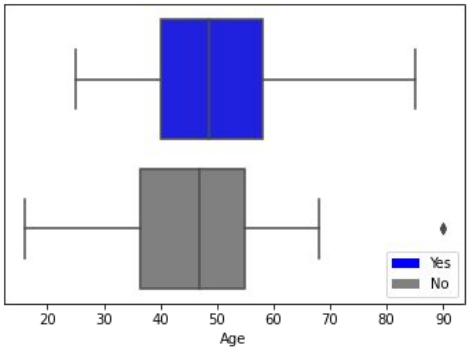
Exploratory Data Analysis - Box Plots

Different Tendencies in Age Group per Symptom in Diabetes Positive Group

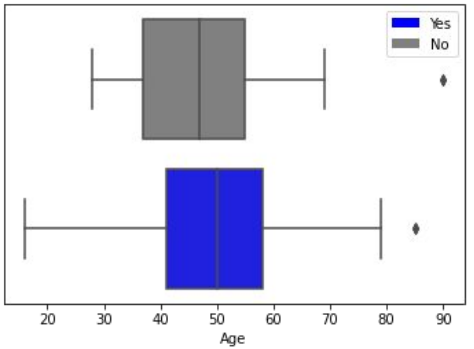
Age vs. Polyuria for diabetes Positive



Age vs. weakness for diabetes Positive



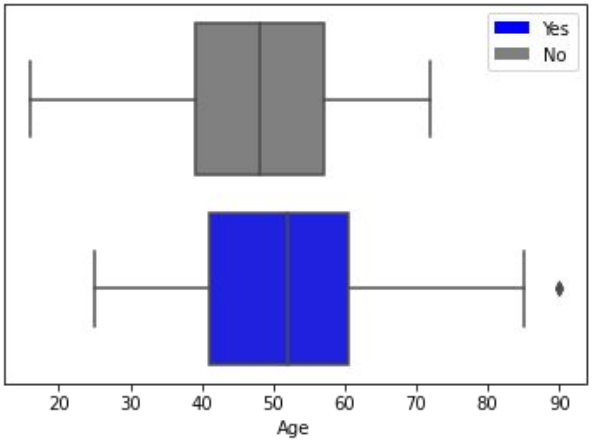
Age vs. Polyphagia for diabetes Positive



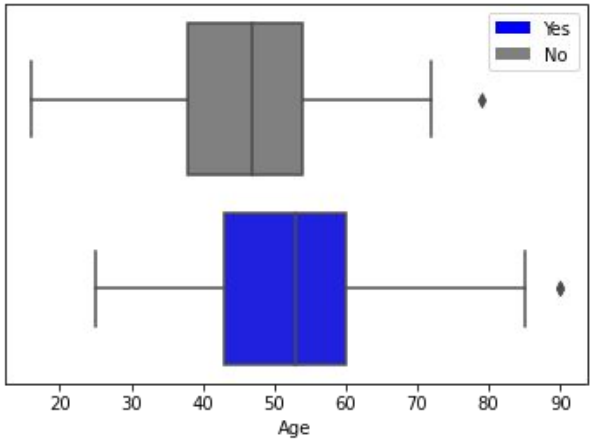
Exploratory Data Analysis - Box Plots

Different Tendencies in Age Group per Symptom in Diabetes Positive Group

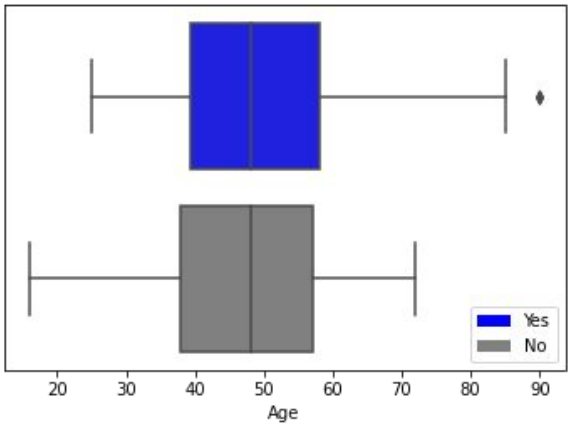
Age vs. Genital thrush for diabetes Positive



Age vs. visual blurring for diabetes Positive



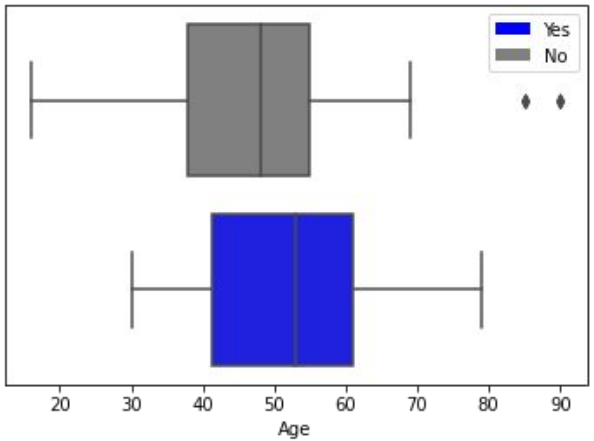
Age vs. Itching for diabetes Positive



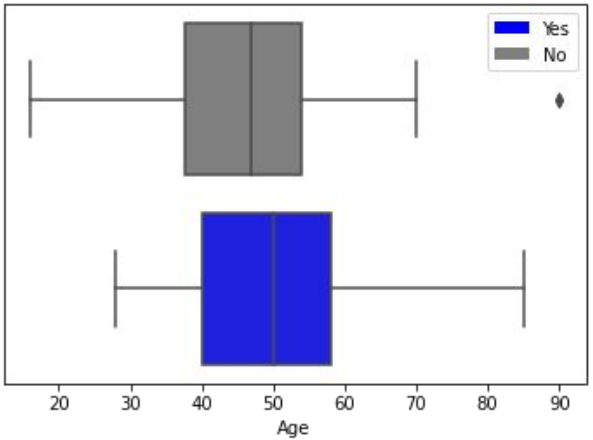
Exploratory Data Analysis - Box Plots

Different Tendencies in Age Group per Symptom in Diabetes Positive Group

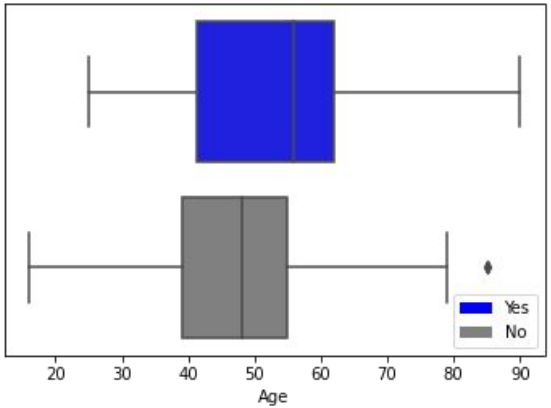
Age vs. Irritability for diabetes Positive



Age vs. partial paresis for diabetes Positive

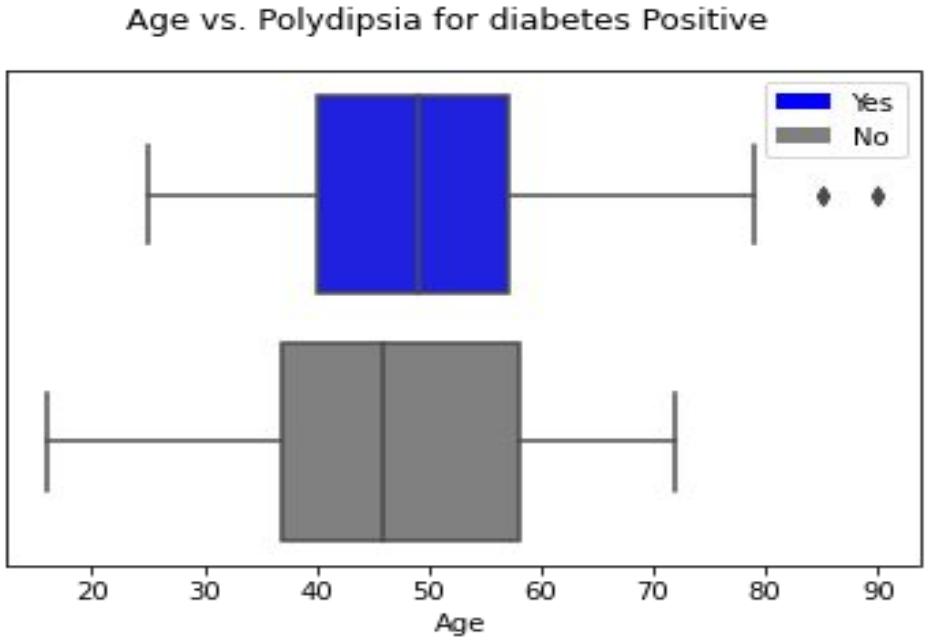


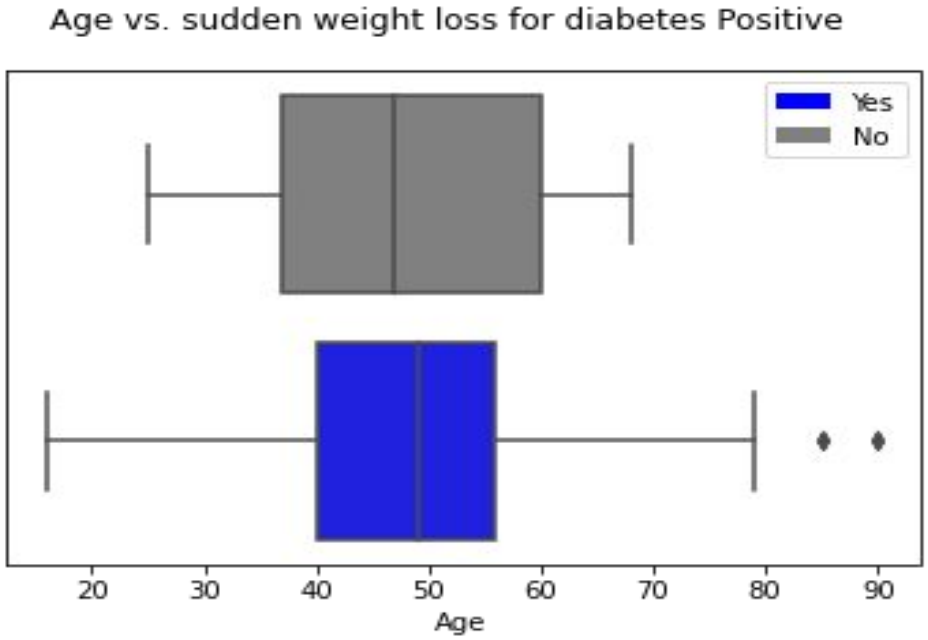
Age vs. Alopecia for diabetes Positive



Exploratory Data Analysis - Box Plots

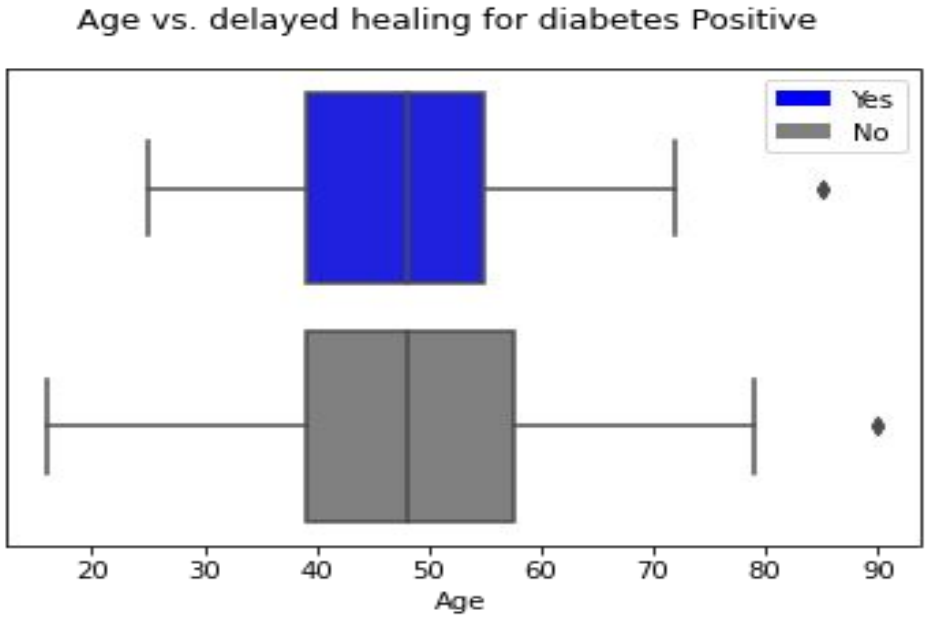
Different Variances in the Same Age Group per Symptom in Diabetes Positive Group

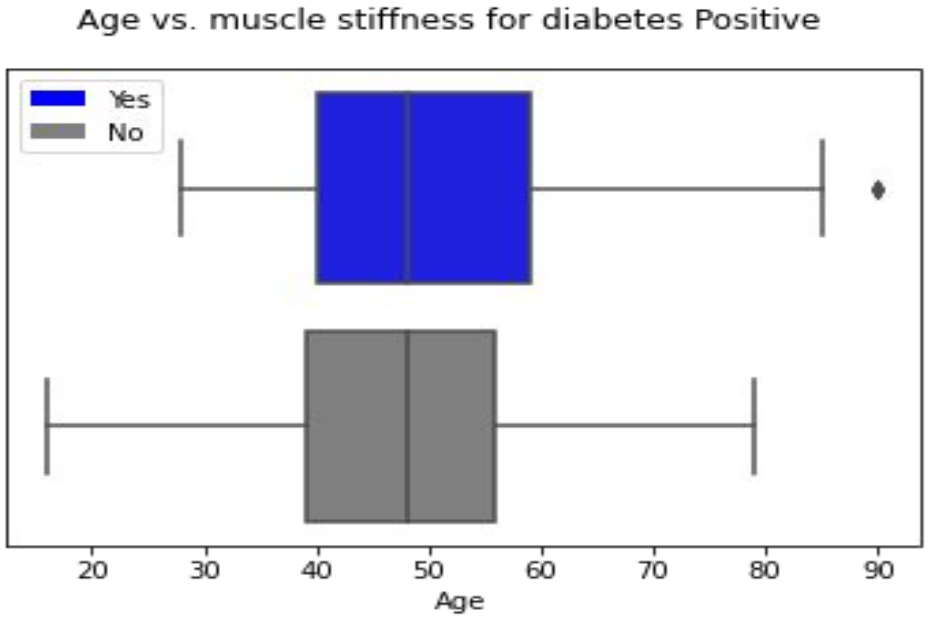




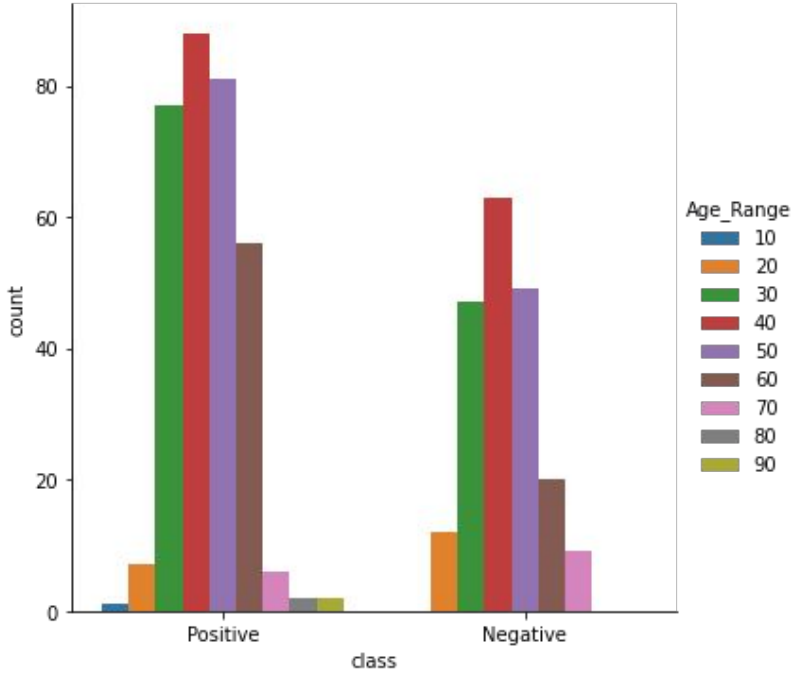
Exploratory Data Analysis - Box Plots

Different Variances in the Same Age Group per Symptom in Diabetes Positive Group





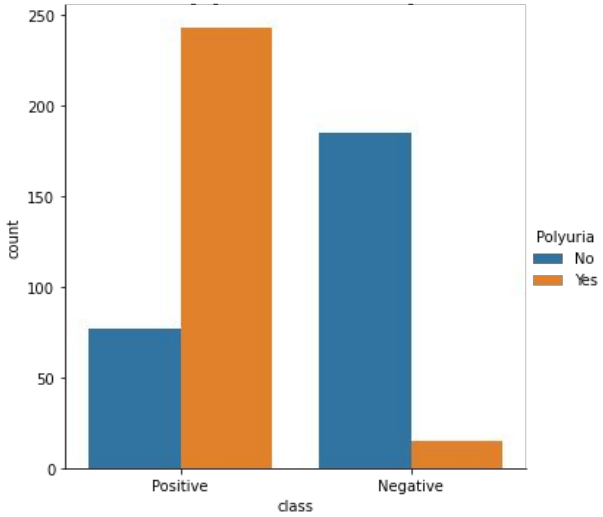
Exploratory Data Analysis - Histograms

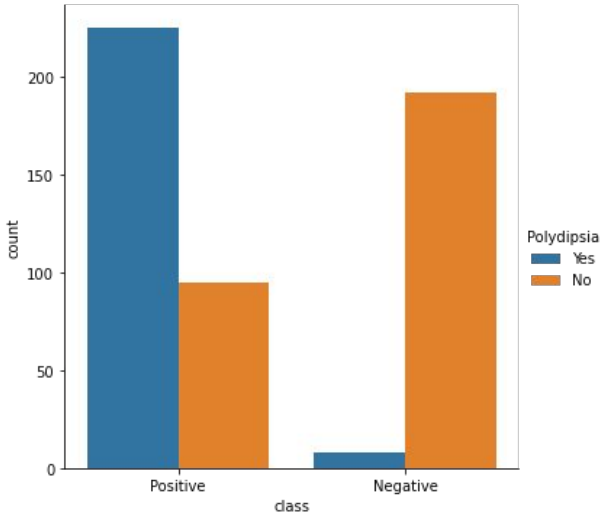


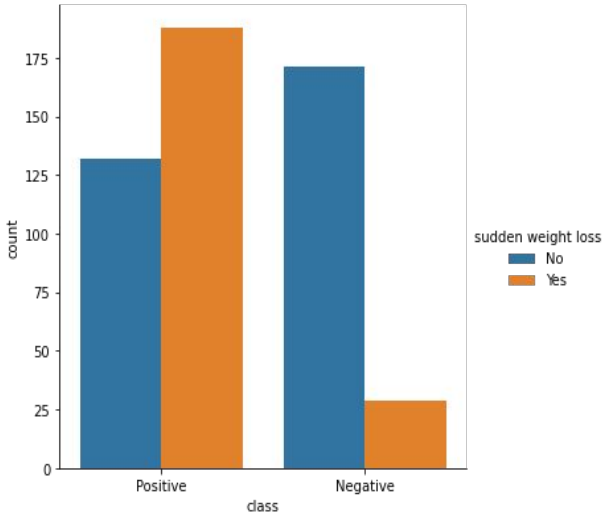
Although each distribution size is different, their age brackets distribute similarly.

Exploratory Data Analysis - Catplots

The below Catplots show distinct differences per symptoms whether diabetes is diagnosed as

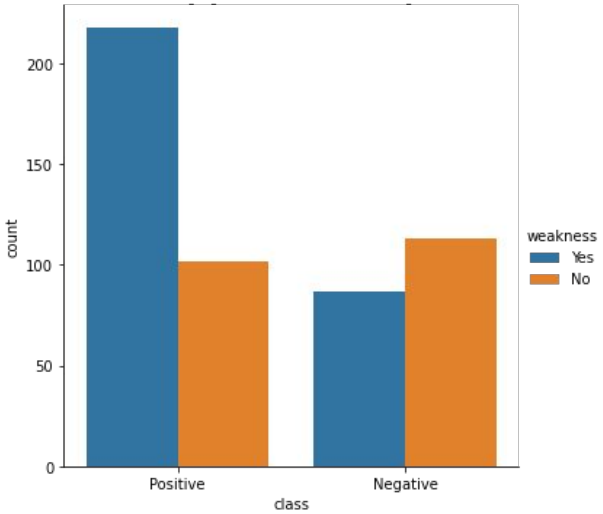


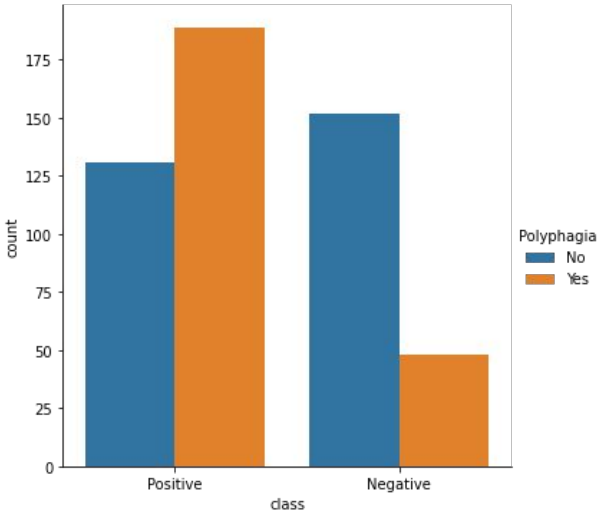


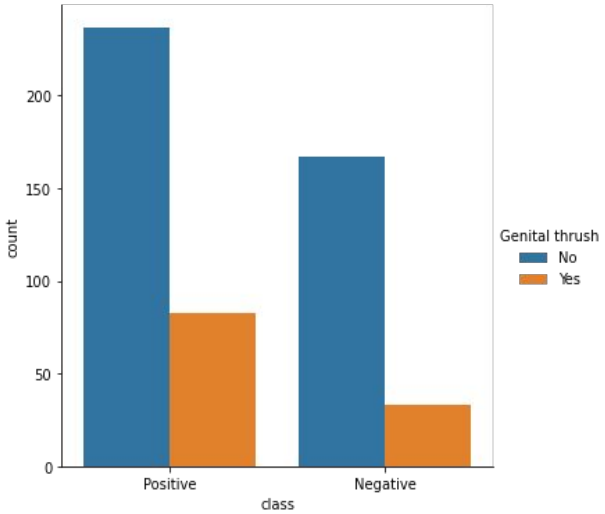


Exploratory Data Analysis - Catplots

The below Catplots show distinct differences per symptoms whether diabetes is diagnosed as

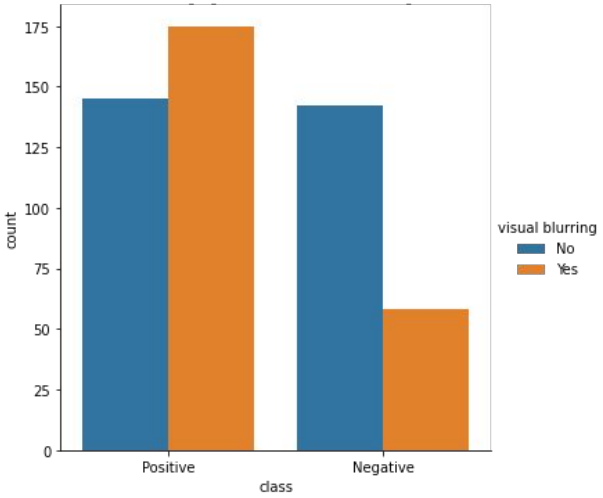


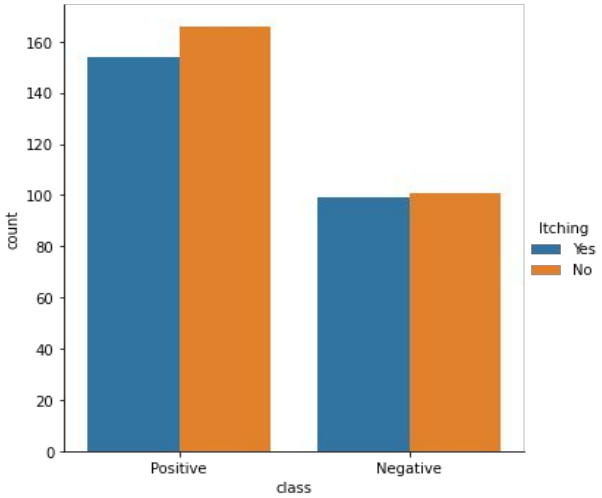


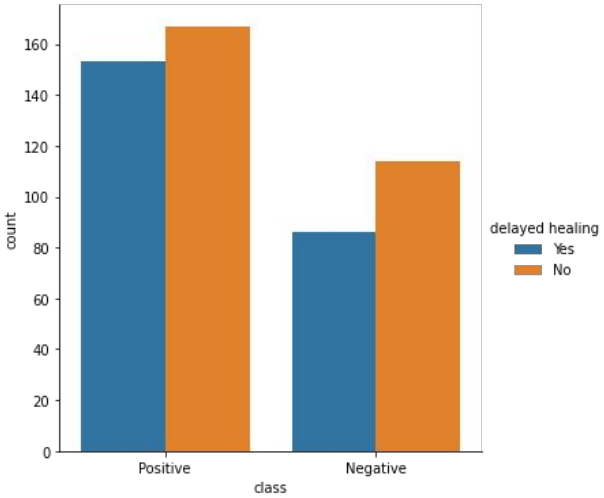


Exploratory Data Analysis - Catplots

The below Catplots show distinct differences per symptoms whether diabetes is diagnosed as

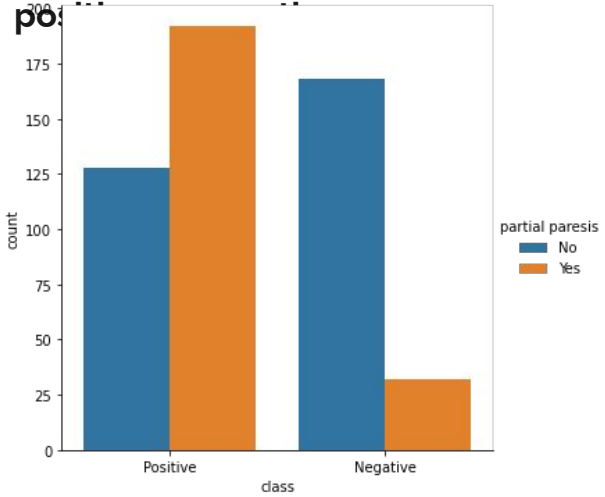


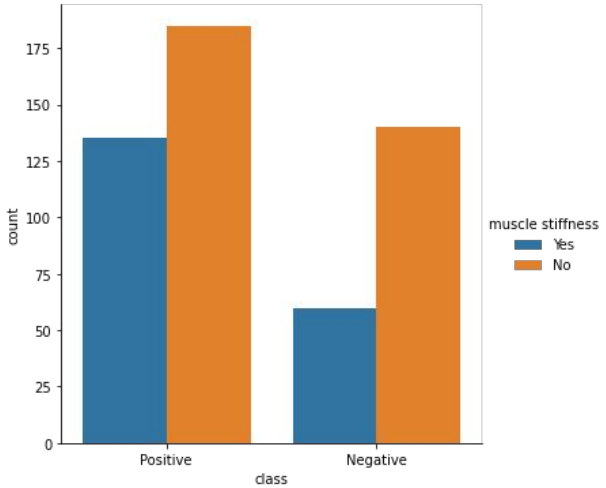




Exploratory Data Analysis - Catplots

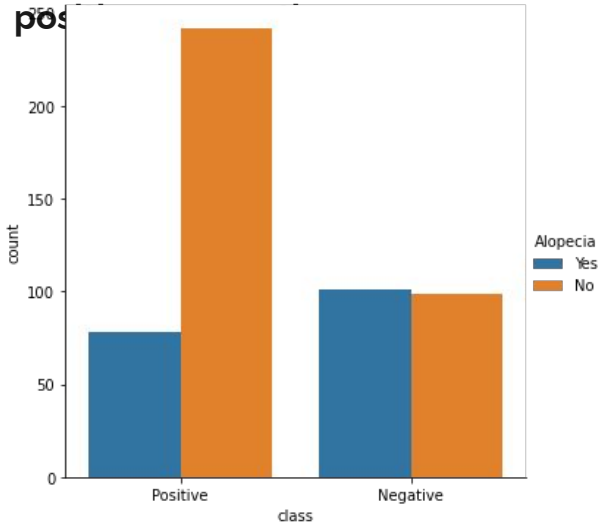
The below Catplots show distinct differences per symptoms whether diabetes is diagnosed as po

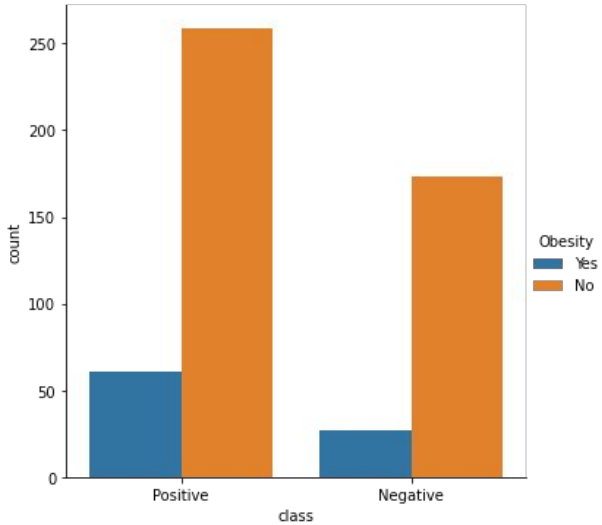




Exploratory Data Analysis - Catplots

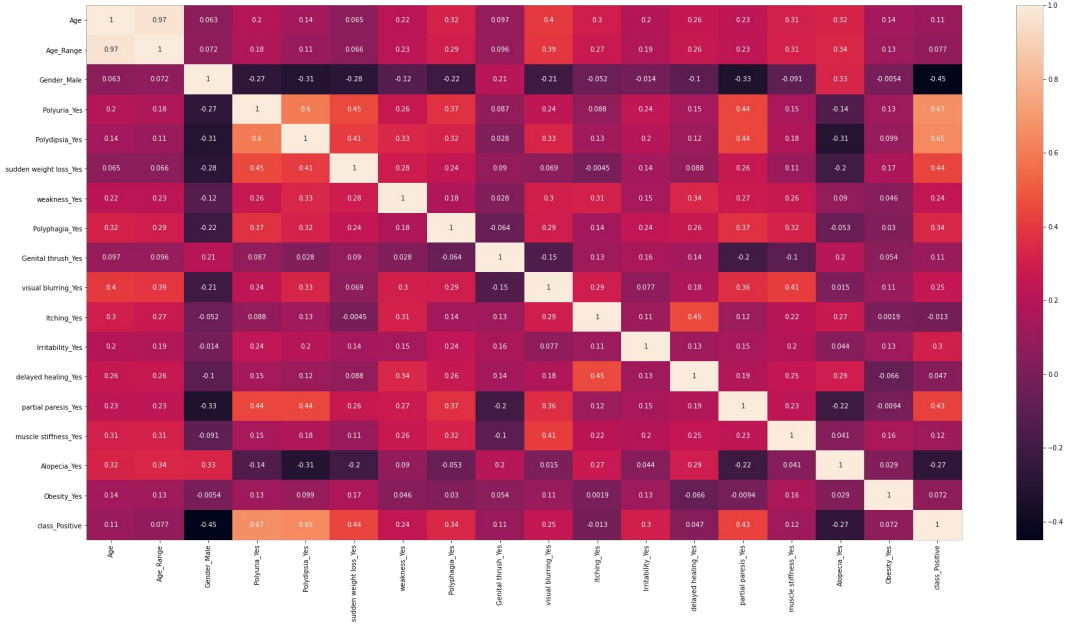
The below Catplots show distinct differences per symptoms whether diabetes is diagnosed as pos





Exploratory Data Analysis - Heat Map

Too many features seem correlated to each other, so it is needed to simplify their correlation by feature importance of several machine learning algorithms.



Machine Learning - Feature Importance

'Polyuria' showed the strongest feature when it comes to identifying 'class_positive'. So we can conclude that 'Polyuria' plays a key role in detect diabetes.

Besides 'Polyuria', there are the following common strong features to detect diabetes even if their importance rates vary: Age, Gender_Male, and Polydipsia.

Algorithms	Decision Tree Classifier	Random Forest Classifier	Gradient Classifier
Feature Importance %	Age: 6% Gender_Male: 11% **Polyuria: 45%** Polydipsia: 8% Alopecia: 7%	Age: 9% Gender_Male: 10% **Polyuria: 20%** Polydipsia: 19% sudden weight loss: 7% partial paresis: 5%	Age: 6% Gender_Male: 11% **Polyuria: 39%** Polydipsia: 22% Alopecia: 5%

Machine Learning - Detection Accuracy

The Decision Tree Classifier showed the best score (97%). Other algorithms showed quite remarkable accuracy scores, but their scores are still lower than the Decision Tree Classifier.

Algorithms	**Decision Tree Classifier**	Random Forest Classifier	Gradient Boost Classifier
Best Score	**97% by Precision**	95% by Recall	96% by Precision
Best Parameters by Randomized Search CV	**'splitter': 'random', 'random_state': 42, 'min_samples_split': 5, 'min_samples_leaf': 2, 'max_features': None, 'max_depth': 15, 'criterion': 'entropy', 'class_weight': 'balanced'**	'n_estimators': 1200, 'min_samples_split': 5, 'min_samples_leaf': 2, 'max_features': ‘sqrt’, 'max_depth': 15, 'criterion': 'entropy', 'class_weight': 'balanced'	'n_estimators': 1200, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'auto', 'max_depth': 15, 'loss': 'deviance', 'criterion': 'mse'

Recommendation - For Better Detection of Diabetes

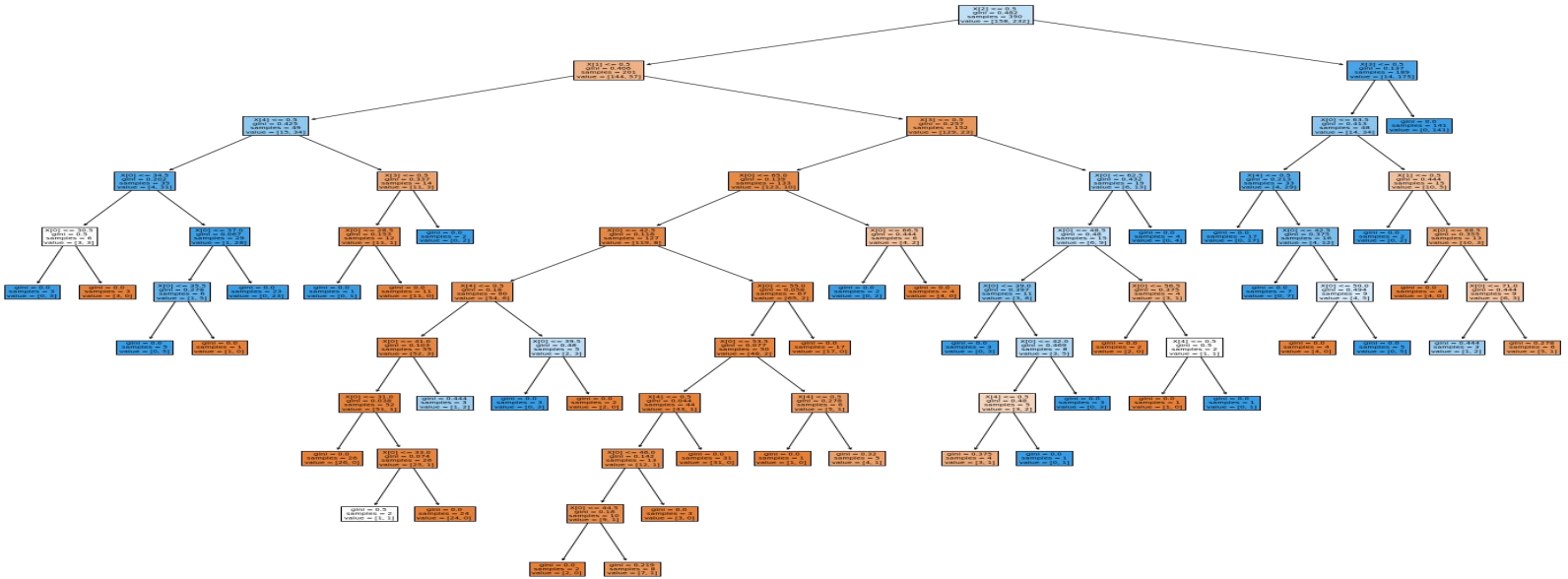
In order to identify diabetes, you may conduct either of the following testing methods on a trade-off basis for both efficiency and effectiveness.

- Combinations of the above-mentioned features can give you optimal diabetes procedures. For example, a test of ‘Polyuria’ and confirmation of ‘Gender’ should be an independent variable for detecting diabetes. ‘Polydipsia’ should be dependent variables with higher weights than others. And then, the other features should be optional test variables with lower weights.

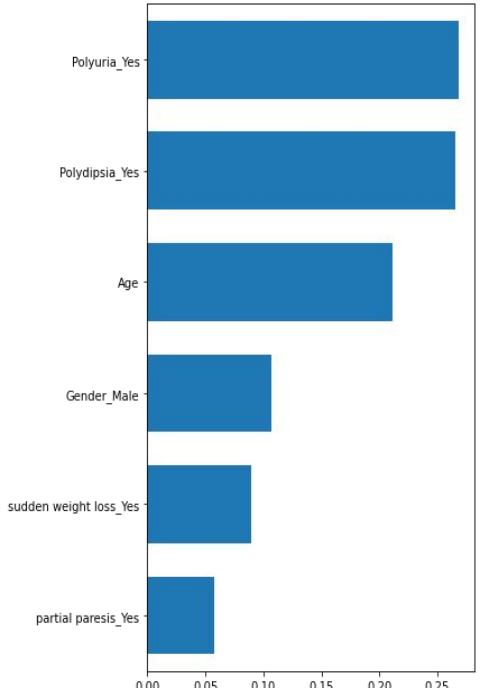
- You can just conduct a diabetes test by the Decision Tree Classifier method.

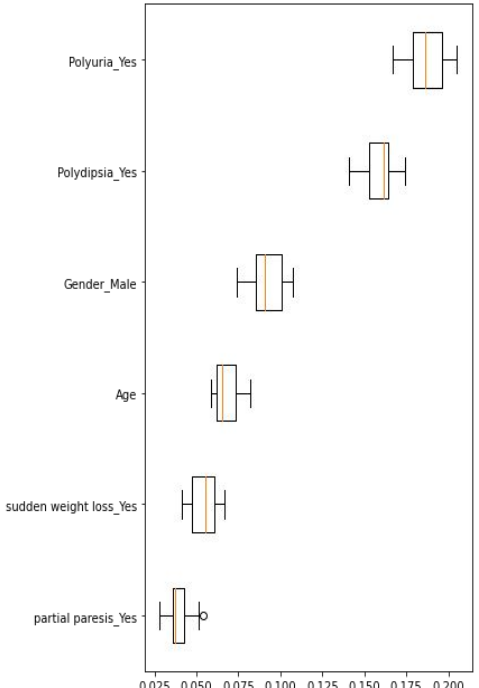
The recommended solution can be both effective and efficient for medical labs to detect diabetes by reducing not only the number of tests but also lead time to test diabetes.

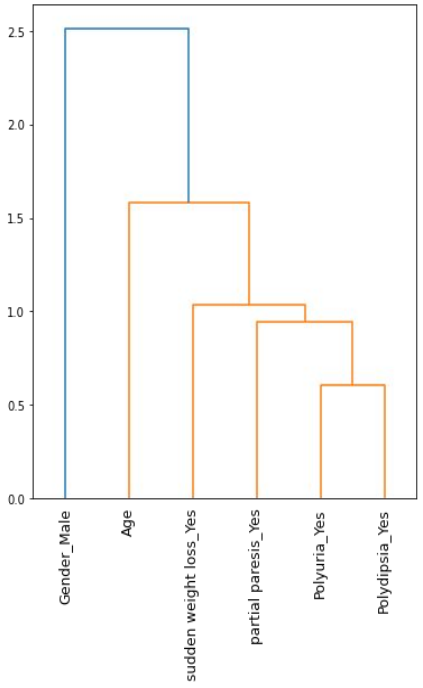
In-Depth Analysis - Decision Tree

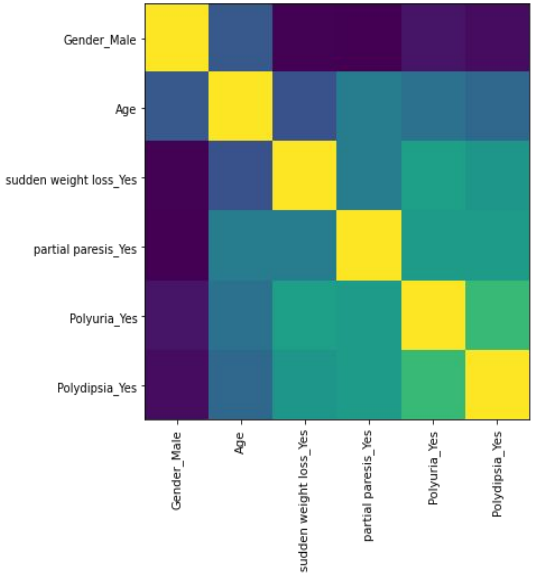


In-Depth Analysis - Random Forest Classifier









In-Depth Analysis - Gradient Boost Classifier