Diabetes Early Prediction

Lucas Peng, Jan 4th 2022

10.5%

of population worldwide have diabetes in 2021

12%

…by 2045

Goal

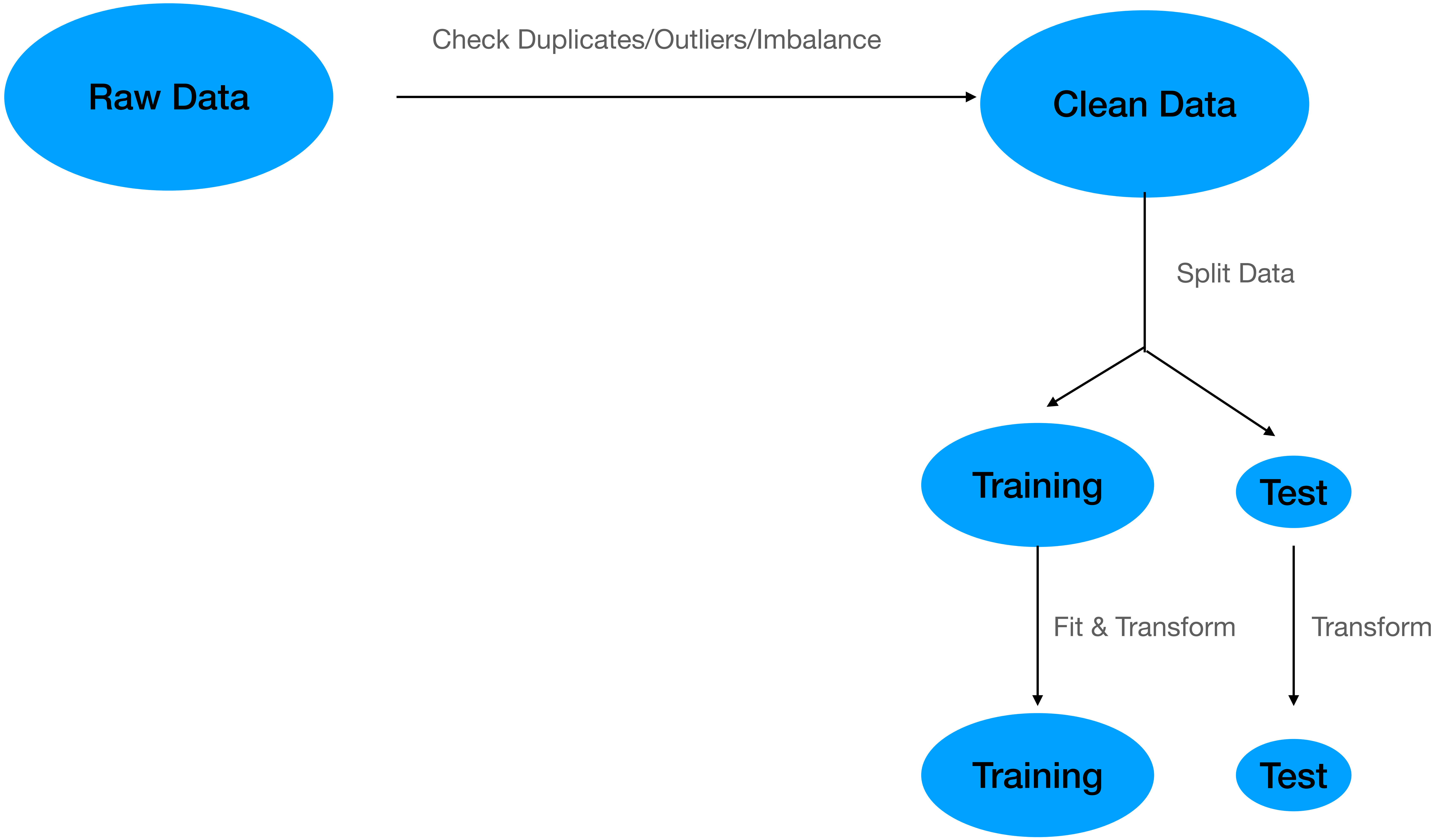
- **Predict** diabetes based on symptoms
- Raise awareness, especially with people whose family had diabetes, about the **most important** symptoms to be alerted for

Data Preprocessing

Sample Data

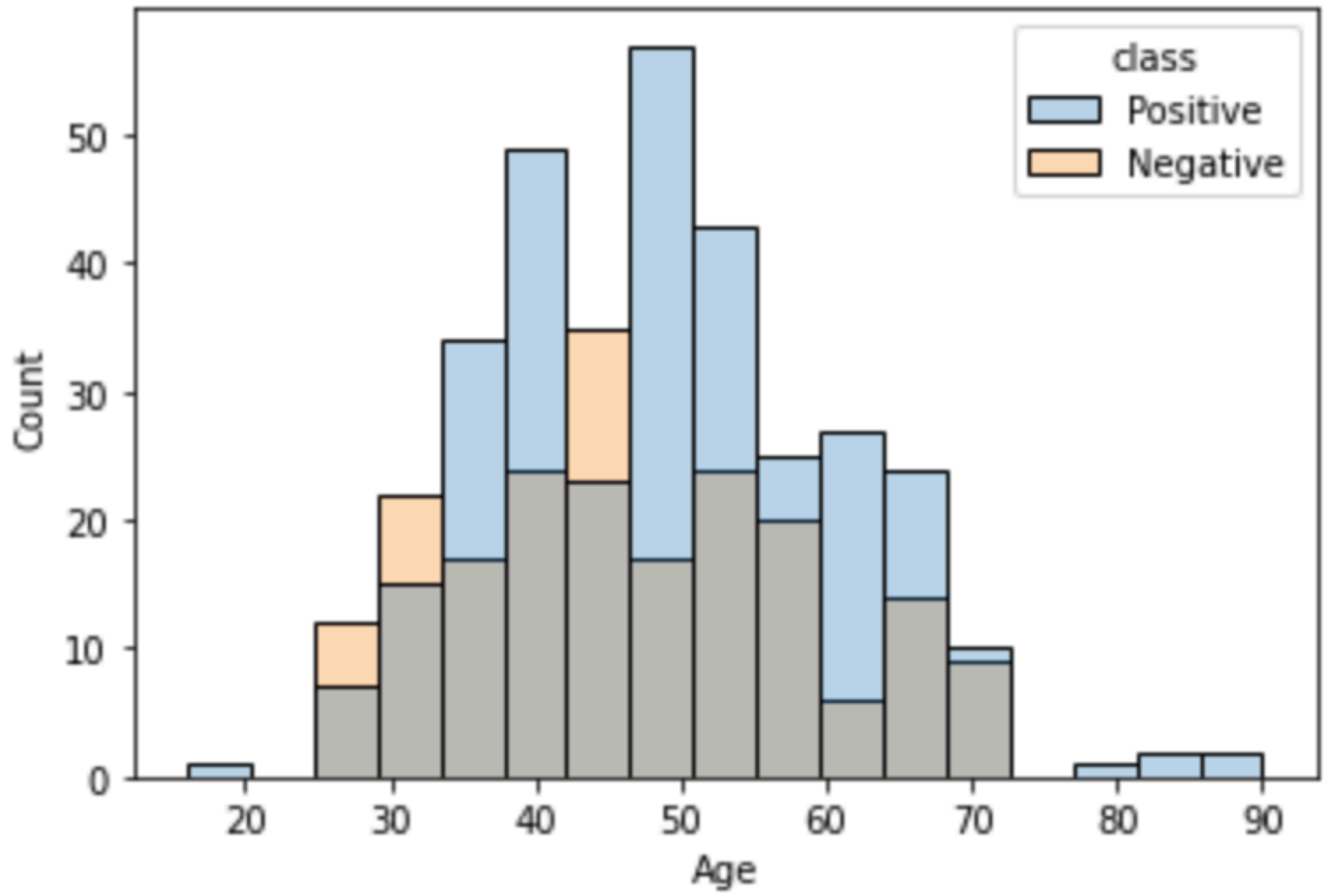
- 520 records

Age	40
Gender	Male
Polyuria	No
Polydipsia	Yes
sudden weight loss	No
weakness	Yes
Polyphagia	No
Genital thrush	No
visual blurring	No
Itching	Yes
Irritability	No
delayed healing	Yes
partial paresis	No
muscle stiffness	Yes
Alopecia	Yes
Obesity	Yes
Class	Positive


Raw Data
Check Duplicates/Outliers/Imbalance
Clean Data
Split Data
Training
Test
Fit & Transform
Transform
Training
Test

Exploratory Data Analysis

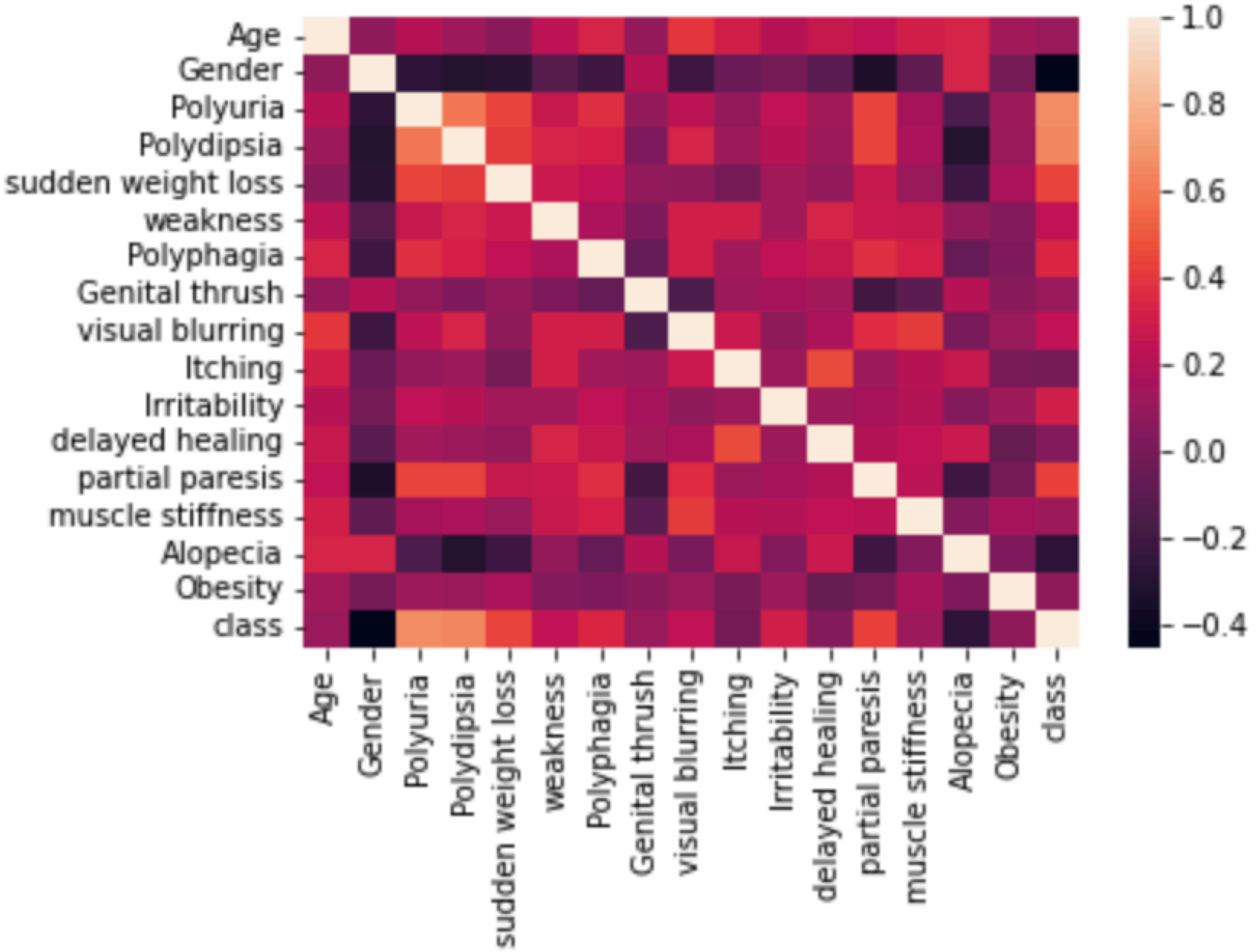
Age effect on Diabetes



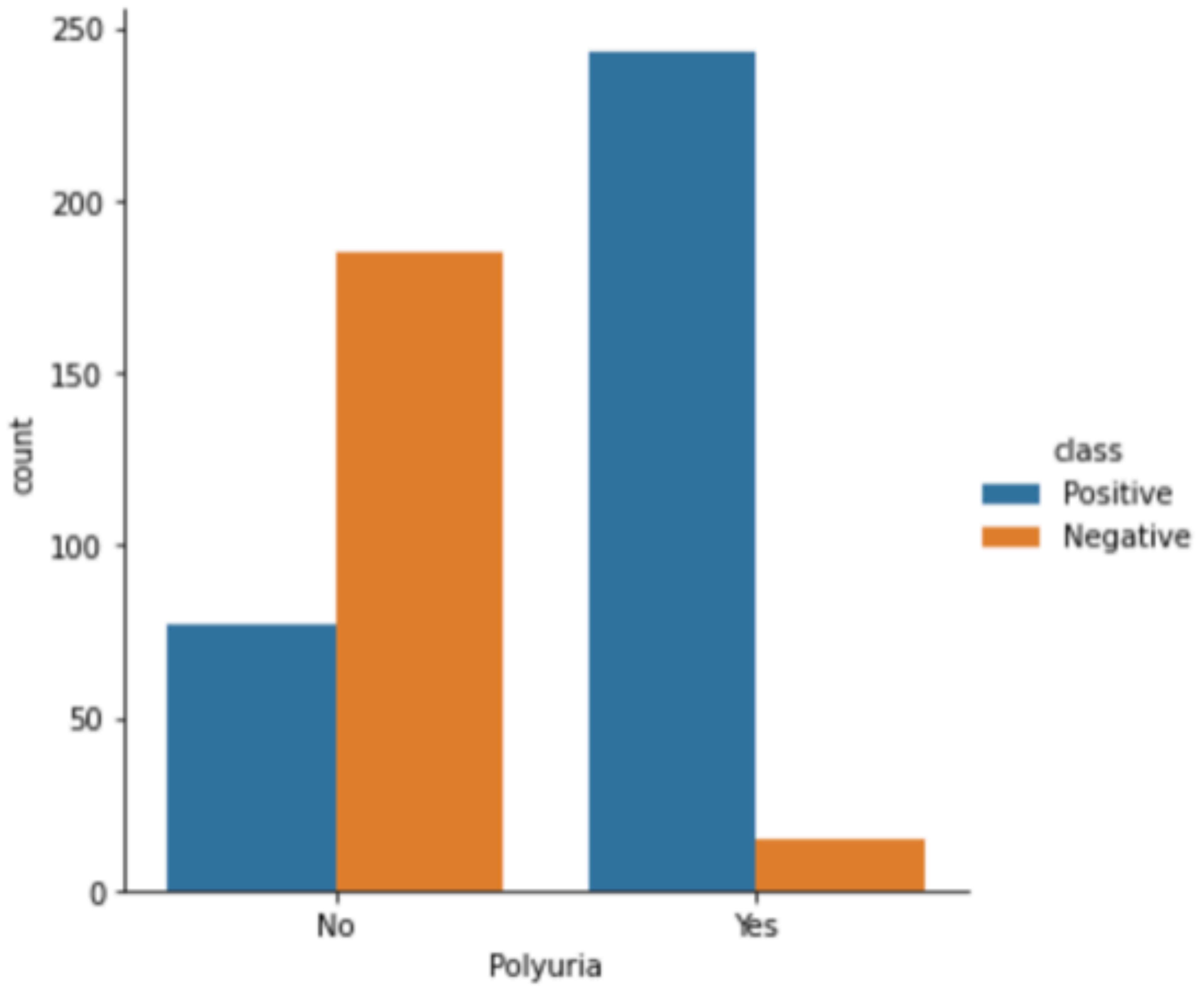
Aside from several positive cases observed for younger (<20) and older (>80) people, the positive and negative case seem to be more or less evenly distributed for people in the middle age, with positive slightly tilting towards older people.

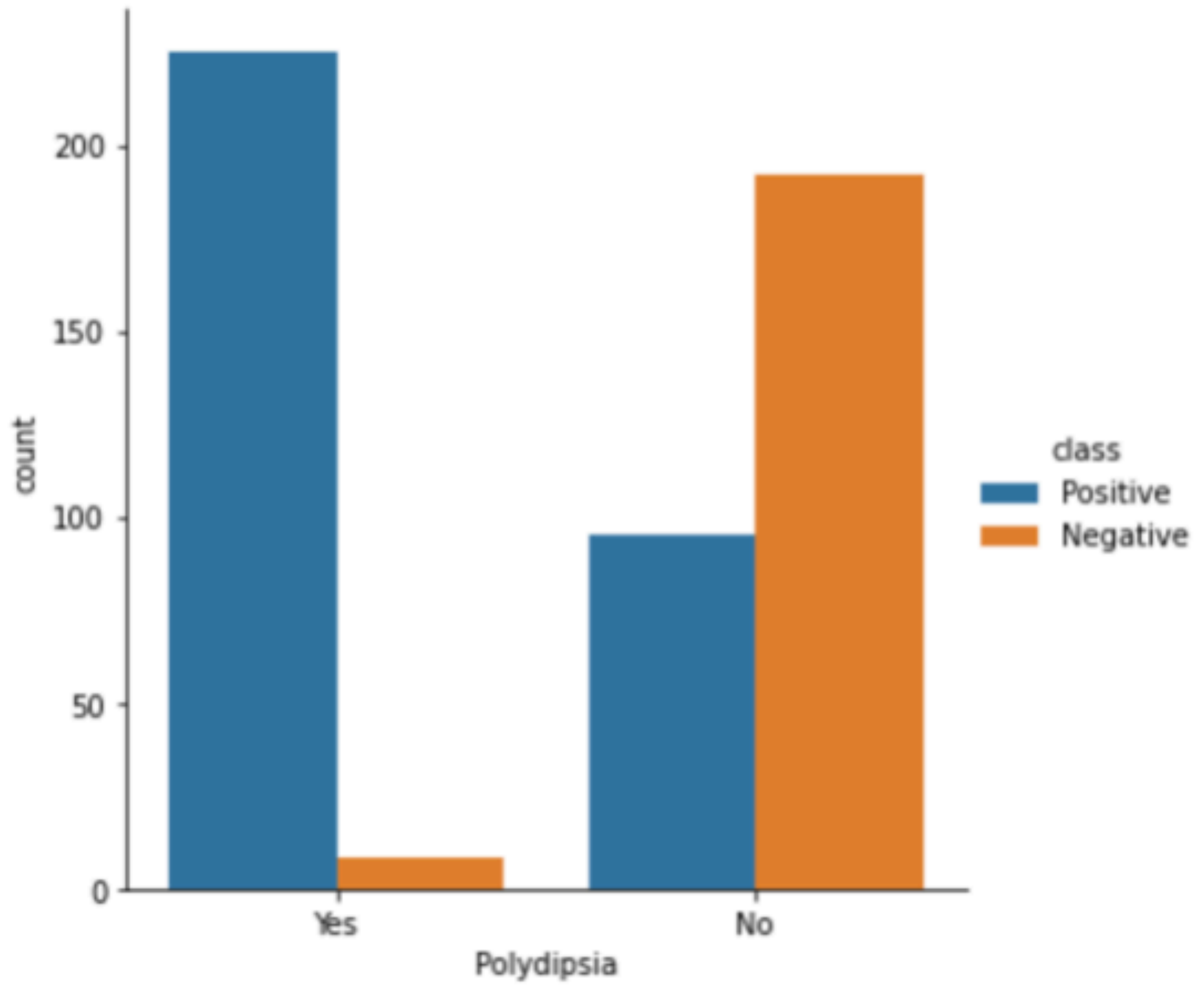
This is important to know because there was a widely held mis-perception that younger people have very low risk of diabetes.

Correlation Heatmap



Polyuria | Polydipsia



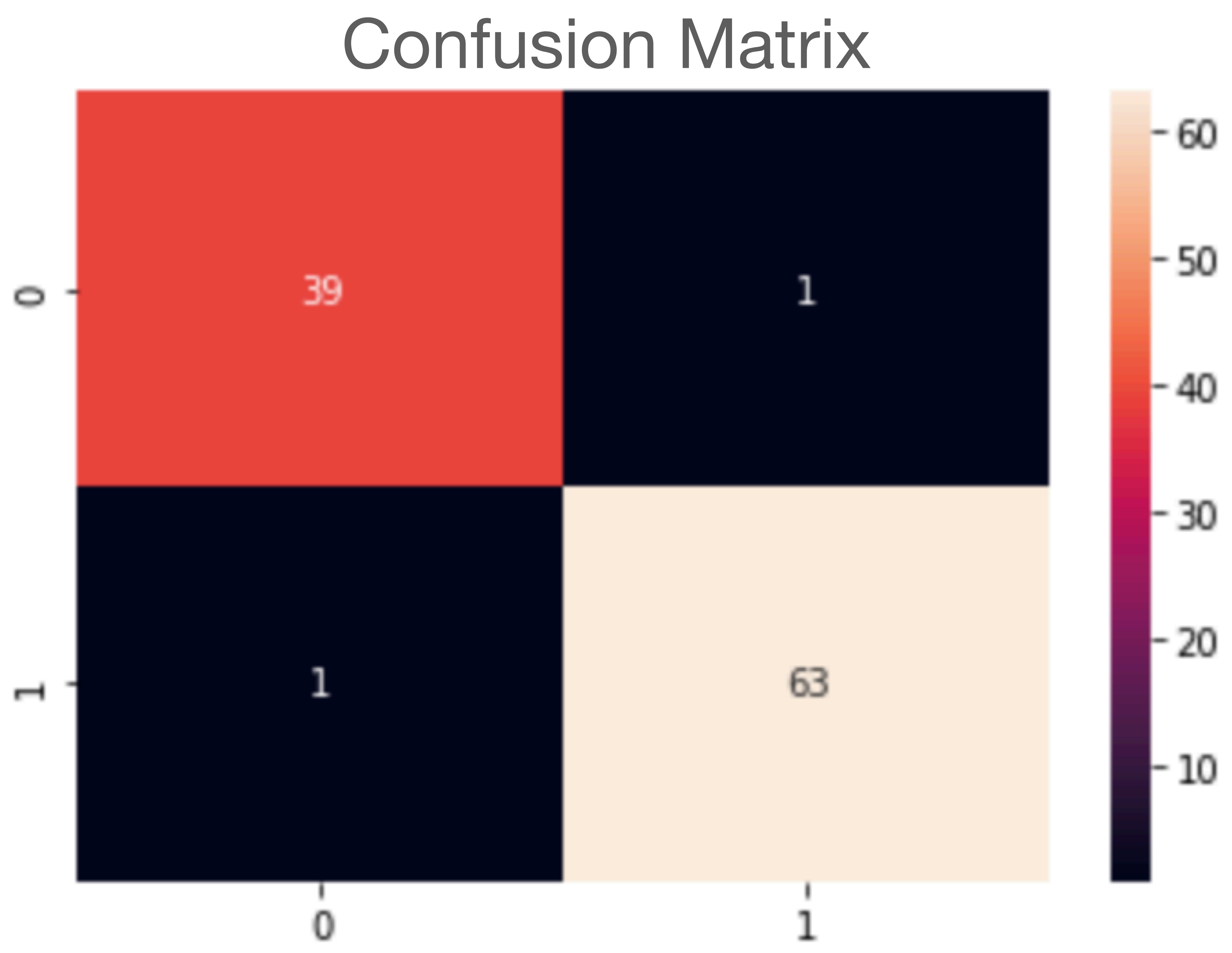


Model Selection

Model Comparison

	CV Recall	Training Time (RandomSearchCV)	Prediction Time
Gradient Boosting	98.4%	5.31 sec	2.48 ms
XGBoost	97.7%	1.74 sec	1.34 ms
Random Forest	97.7%	8.78 sec	18.7 ms
K Nearest Neighbour	~~96.1%~~	~~0.25 sec~~	~~N/A~~

XGBoost closer look

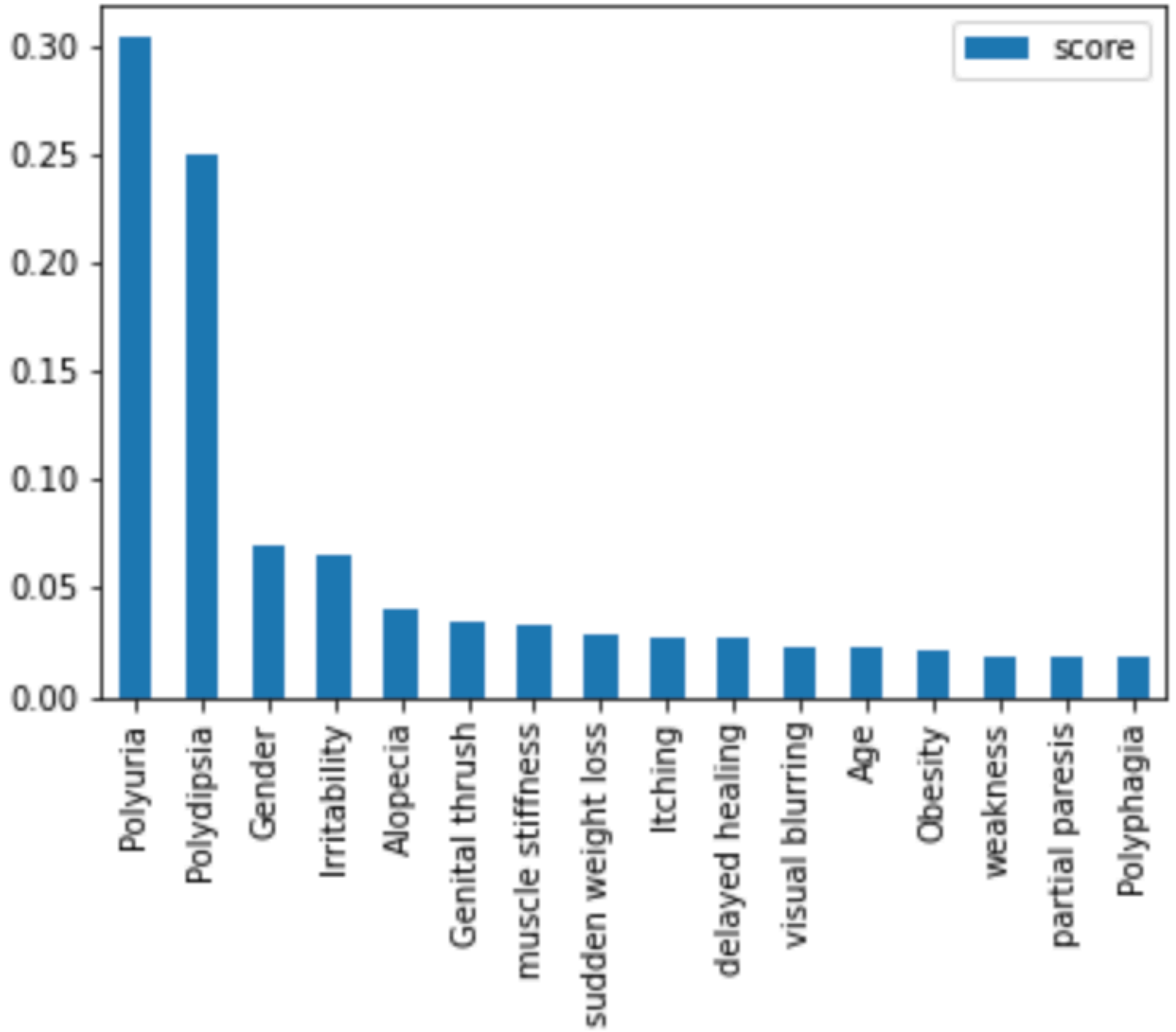


Classification Report

	precision	recall	f1-score	support
0.0	0.97	0.97	0.97	40
1.0	0.98	0.98	0.98	64
accuracy			0.98	104
macro avg	0.98	0.98	0.98	104
weighted avg	0.98	0.98	0.98	104

Takeaway

Feature importance



Polyuria & Polydipsia
are two most important
indicators,

**ranked by average gain when used
in splitting the tree.**

Thank you

Lucas Peng, Jan 4th 2022
Github: https://github.com/bbltxl/CapstoneProject_Diabetes