Diabetes Machine Learning Project

Dallas Strandell

February 1, 2022

1 Introduction

This document describes the method to model symptoms of patients to predict for diabetes. The source of the data is:

https://www.kaggle.com/andrewmvd/early-diabetes-classification/version/1

The data was modeled in the notebook modeling.ipynb

2 Initial Data Analysis

The first step was to understand the starting data. With plotting, each feature was analysed. Most of the features were binary integers with age and gender being the exceptions. Gender was stored as the strings Male and Female and therefore was changed to 1 and 0 respectively. Additionally, age was normalized to between 0 and 1. A subplot was made with all of the features except age vs class (Figure 1). Class is whether the patient tests positive for diabetes: 0 being negative and 1 being positive. The correlations of the features with class were also calculated and are shown in Table 1; polyuria had the highest at 0.667 while alopecia had the lowest at -0.268 (not including gender). For the definitions of each symptom see the Kaggle URL above.

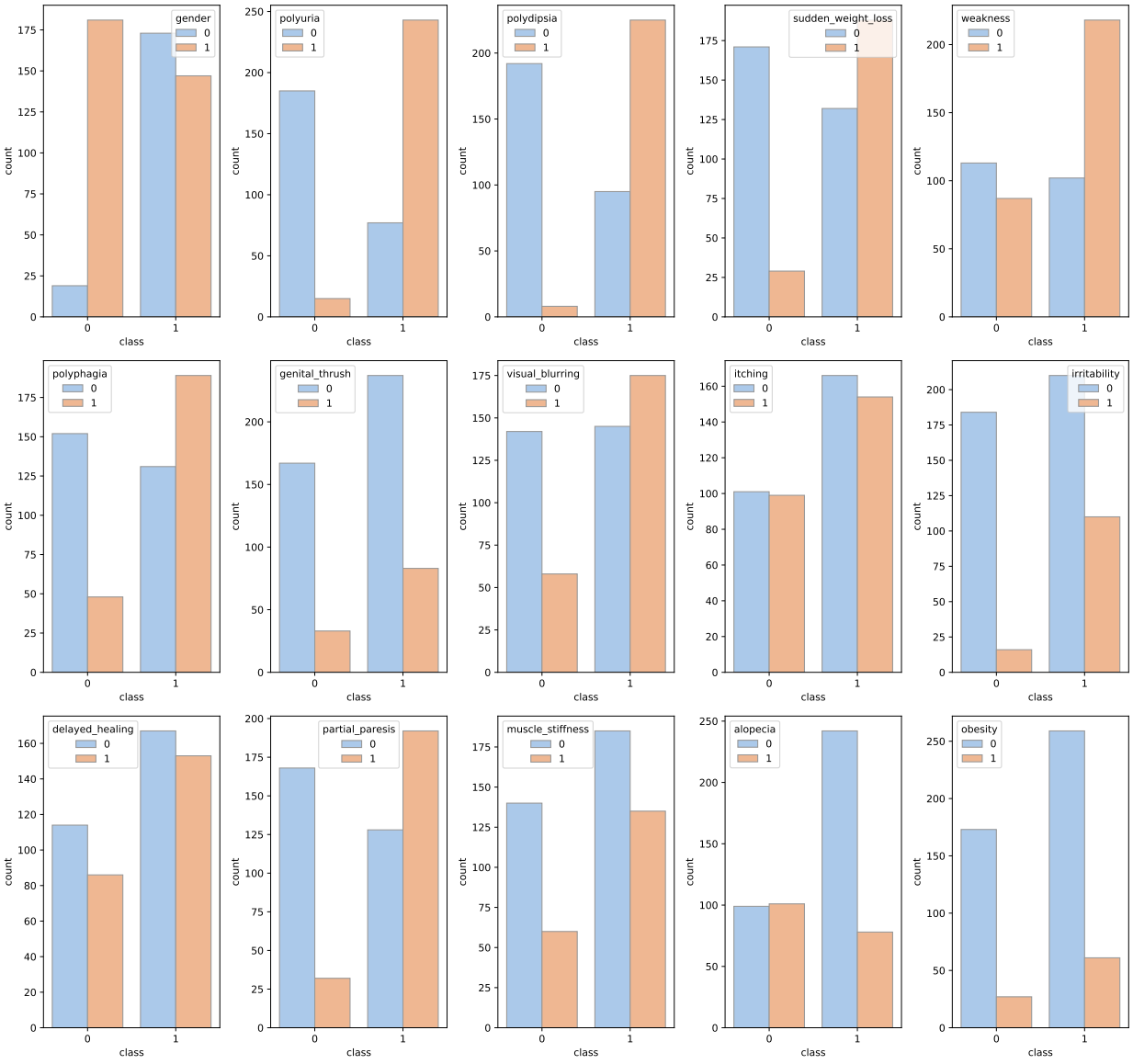

Figure 1: Initial data analysis plots. A class equal to 1 means the patient had diabetes.

Feature	Correlation
age	0.109
gender	-0.449
polyuria	0.666
polydipsia	0.649
sudden weight loss	0.437
weakness	0.243
polyphagia	0.343
genital thrush	0.110
visual blurring	0.251
itching	-0.0133
irritability	0.299
delayed healing	0.047
partial paresis	0.432
muscle stiffness	0.122
alopecia	-0.268
obesity	0.072

Table 1: Feature correlation with having diabetes.

3 Model Results

Metrics	Decision Tree	Decision Tree Optimized	Random Forest
accuracy	0.97	0.97	0.99
false positives	2	2	0
false negatives	3	2	1

Table 2: Selected modeling results

Modeling was performed mainly with Decision Tree and Random Forest classifiers. Other models were tested and the results can be viewed in the modeling.ipynb notebook. Random forest gave the best results with only one false negative with an accuracy of 0.99 as shown in Table 2. The default hyperparameters of decision tree gave 5 false results with an accuracy of 0.97. Optimizing made little difference: decreasing the false negatives by 1. Feature importance for the random forest model shows that polyuria and polydipsia (excess urination and thirst) are the two most important symptoms (Figure 2). This matches well with the correlation for the features where they have the highest values (Table 1).

4 Conclusion

The models described in this report can be used to predict if a patient has diabetes based on common symptoms. Random forest classification gave the best results with an accuracy of 0.99.

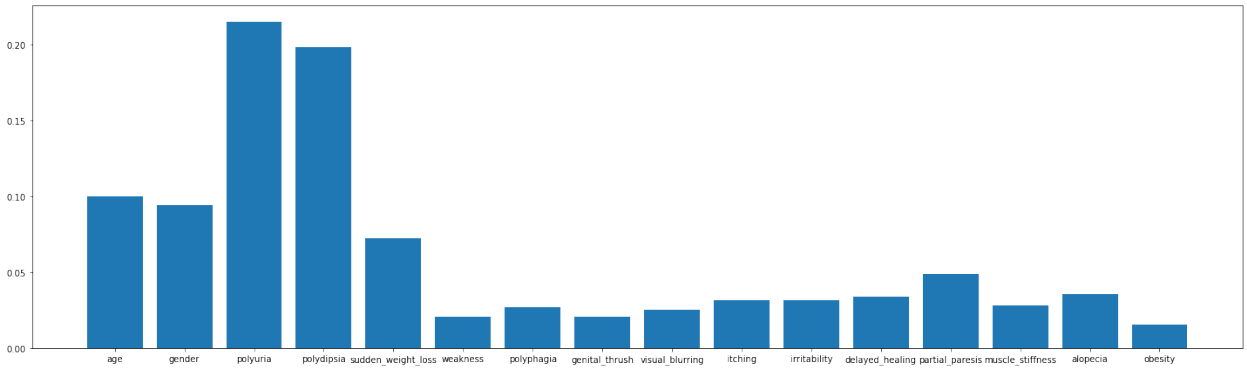

Figure 2: Feature importance of the random forest model.