Finding early Diabetes in Patients

Camilo Garcia

Abstract

Dataset for prediction early stage diabetes on patients and the second one is used to find important factors in people that already had the disease.

My main questions are: Which factors have the most influence in developing diabetes? Can you predict which patience will have diabetes from an early stage diagnosis? Can you classify patients in groups so there can be an effective treatment for each group?

The methods used for solving the questions where simple statistics and plotting analysis, a ML classification algorithm with sklearn and a clustering algorithm with sklearn KMeans.

It seems that the features with more influence to detect diabetes at an early stage are levels of Polyuria, levels of Polydipsia and plasma glucose concentration, also you can indeed classify patients at an early stage but it is not so clear to classify them into groups with clustering.

Motivation

My parents are doctor and I find a lot of pride in doctors daylife. Saving others lives and risking their own is a noble job. Diabetes is a serious disease that kills more than 1.5 million people every year and affects other 400 million more. I, as an engineer, always wanted to help improve life somehow. My motivation behind this project is trying to identify and prevent diabetes at an early age and trying to classify people in groups that helps them fight against this disease.

Dataset(s)

For this project only 2 databases were used:

- UCI Early stage diabetes risk prediction dataset
- Pima Indians Diabetes Database

Data Preparation and Cleaning

Checked if there were any missing values, fast visualization and plotting of data and deleting some columns that hat few values.

There were not many problems in this phase

Research Question(s)

In this project 3 questions were tried to be solved:

1. Which factors have the most influence in developing diabetes?
2. Can you predict which patience will have diabetes from an early stage diagnosis?
3. Can you classify patients in groups so there can be an effective treatment for each group?

Methods

For solving the first question I converted most of the categorical features to numbers with the function LabelEncoder(), then I calculated the correlation between features and finally I plotted the most correlated features that afterwards developed diabetes.

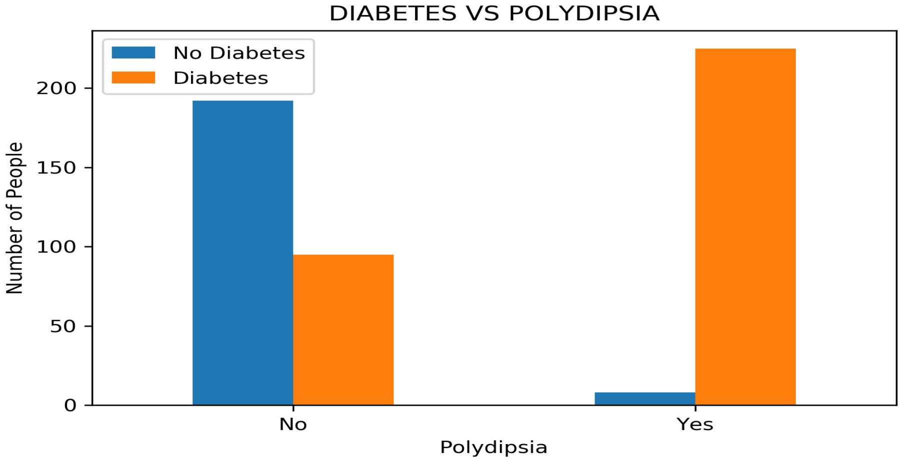
Then I did a ML classification task for predicting if patients were going to develop diabetes. For that I used a Decision Tree Classifier and compared the results with a svm.SVC classifier.

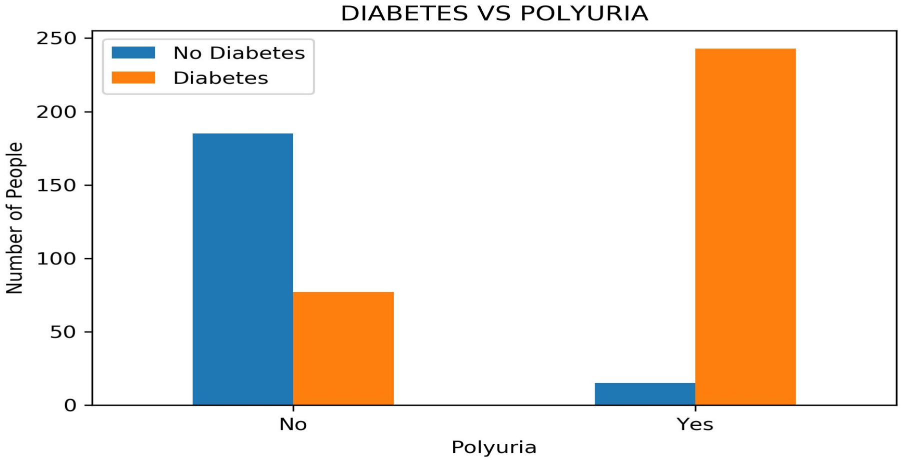
Next, I analyzed the second dataset to find out another important feature for detecting diabetes in patients. Lastly I tried to classify the second dataset into cluster trying to find groups which could have different treatments for the disease. The clustering algorithm was made with SKLearn Kmeans.

Findings

As we can see from the 2 graphics, the two most important characteristics for detecting diabetes early on are the presence of Polydipsia and Polyuria. If a patient's start symptoms is almost certain they could be developing diabetes soon.

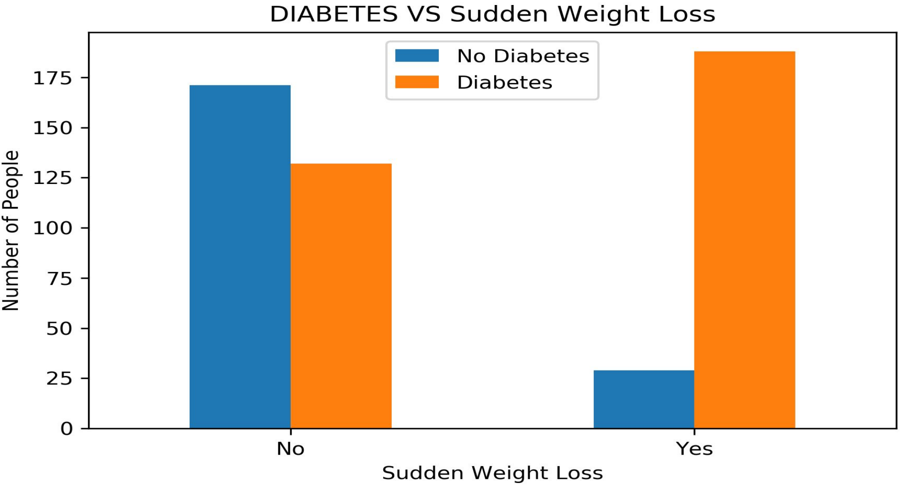
Polydipsia means excessive thirst and Polyuria means excessive production of urine.

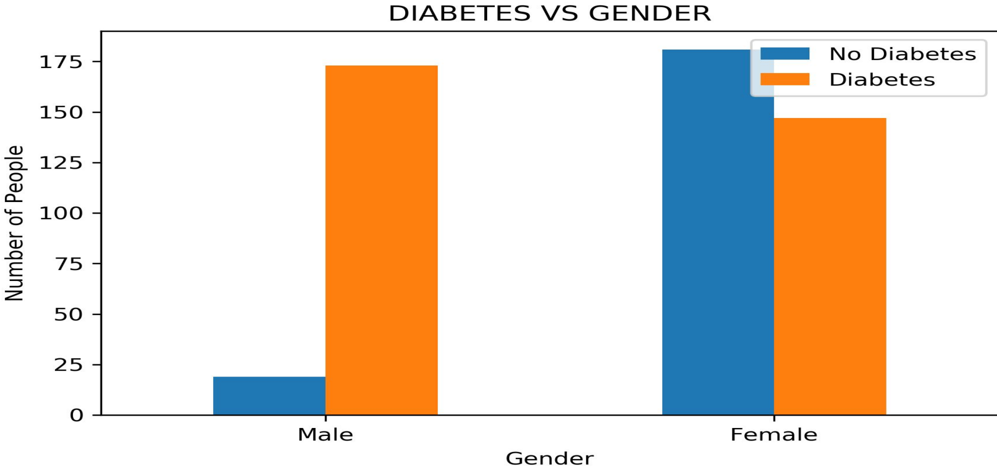




Findings

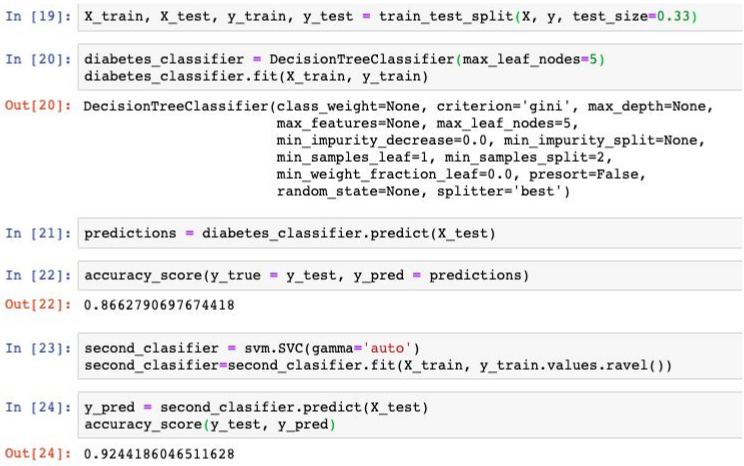
- I analyzed also the features weight loss and gender in patients. As we can see from the graphics, sudden weight loss could mean the patient is developing diabetes (overweight in other hand isn't as correlated to diabetes as weight loss).
- Finally gender has nothing to do with diabetes, the graphic just shows there were more females than males in this study.





Findings

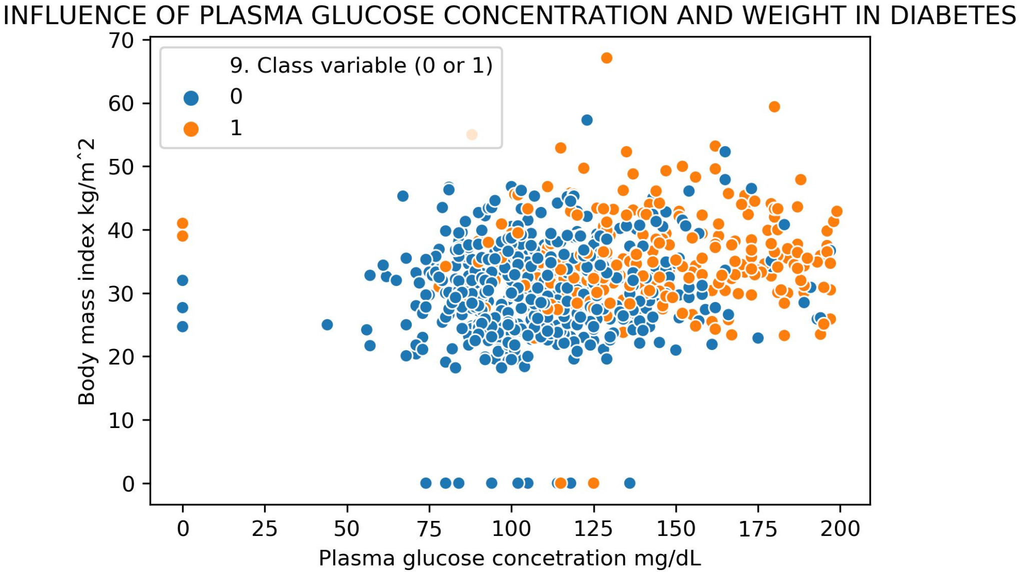
Here are the results of my classification task. As we can see, both classifiers hat a very high accuracy with 86% and 92%. Also, with many iterations I found out that the SVC Classifier had almost always higher accuracy than the Decision Tree. That could be because of the small nodes in the Decision Tree but also it could mean that SVC is overfitting.



```
In [19]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)

In [20]: diabetes_classifier = DecisionTreeClassifier(max_leaf_nodes=5)
         diabetes_classifier.fit(X_train, y_train)

Out[20]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
                     max_features=None, max_leaf_nodes=5,
                     min_impurity_decrease=0.0, min_impurity_split=None,
                     min_samples_leaf=1, min_samples_split=2,
                     min_weight_fraction_leaf=0.0, presort=False,
                     random_state=None, splitter='best')

In [21]: predictions = diabetes_classifier.predict(X_test)

In [22]: accuracy_score(y_true = y_test, y_pred = predictions)

Out[22]: 0.8662790697674418

In [23]: second_clasifier = svm.SVC(gamma='auto')
         second_clasifier=second_clasifier.fit(X_train, y_train.values.ravel())

In [24]: y_pred = second_clasifier.predict(X_test)
         accuracy_score(y_test, y_pred)

Out[24]: 0.9244186046511628
```

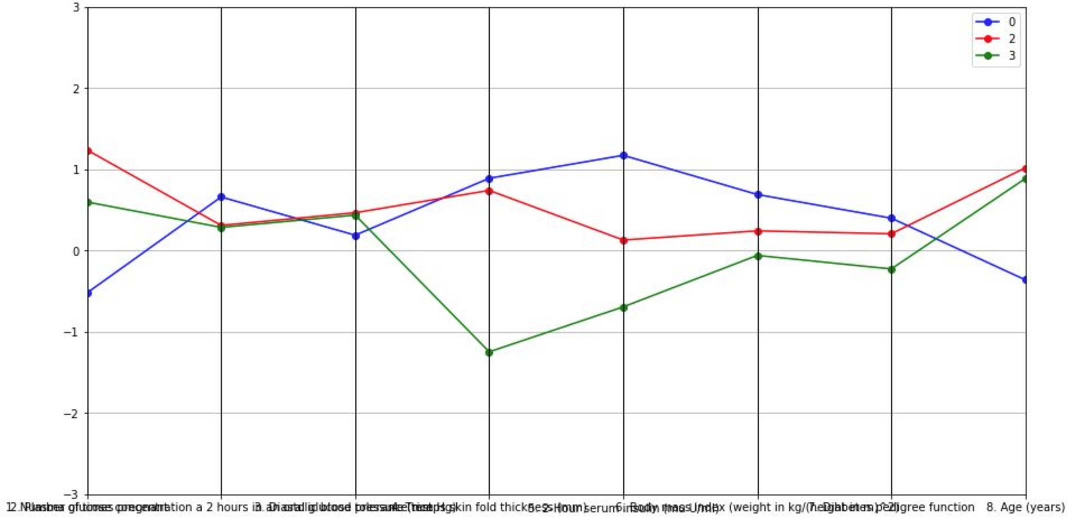
Findings

This is a graphic from the second dataset, as we can see from the graph weight or body mass isn't an indicator of diabetes (Class 1). On the other hand, levels of plasma glucose concentrations (especially over 125mg/dL is a big factor to find diabetes)



Findings

This are the results of clustering the patients into 3 groups. I did it to put patients into groups with better treatments for each group but it is difficult to define a group and their characteristics without more medical knowledge, here in the graph we can see 3 groups that differ in skin fold thickness, age, number of times pregnant, and insulin levels in blood.



Limitations

Dasates had very few entries, meaning the findings are not as precise and meaningful as if I had a bigger dataset.

Conclusions

- Polydipsia and Polyuria, and high levels of plasma glucose concentration are important in detecting diabetes early on in patients.
- Gender has nothing to with diabetes. Also sudden weight loss can be an important factor to take into account.
- With a simple dataset of patients with features like age, gender, Polyuria, Polydipsia, weight loss, and others it is possible to classify patients into having diabetes or not with a certain level of certainty.
- SMV has more accuracy than a decision tree for this task, but one must be careful with overfitting.
- Classification can be useful for special treatments but it is also difficult to find insights.

Acknowledgements

First data set: UCI Machine learning repository

Second data set: Kaggle.com

A colleague from university gave me feedback (data engineer) but I collected all the data myself.

References

I did all the work on my own.