Background

How good a model can be developed to predict diabetes from relevant symptoms?

The dataset used for this investigation was from the UCI Machine Learning Repository and was collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor. It includes the gender, age and 14 symptoms of a possibly diabetic patient and whether they tested positive or negative for diabetes. (Note - gender and age will be referred to as symptoms - making a total of 16 "symptoms" for testing. There are 520 patients/subjects/rows in the dataset.*) The aim is to build a model to predict whether a patient providing that data would be more likely to test positive or negative for diabetes.

Although the dataset is limited in number of patients and location, a model developed would certainly be relevant for that location and the process might reveal information that could be a basis for more inclusive modeling.

*Note: the data-wrangling notebook incorrectly said 550 at top, although showed 520 when working through the data.

Source and Definitions

The link to the UCI Machine Learning Repository dataset is:
https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.

Definitions of medical term symptoms:

polyuria: production of abnormally large volumes of dilute urine.

polydipsia: abnormally great thirst as a symptom of disease (such as diabetes) or psychological disturbance.

polyphagia: also known as hyperphagia, is the medical term for excessive or extreme hunger. It's different than having an increased appetite after exercise or other physical activity. While your hunger level will return to normal after eating in those cases, polyphagia won't go away if you eat more food.

alopecia: alopecia areata is a condition that causes hair to fall out in small patches, which can be unnoticeable. These patches may connect, however, and then become noticeable. The condition develops when the immune system attacks the hair follicles, resulting in hair loss.

partial paresis: paresis involves the weakening of a muscle or group of muscles. It may also be referred to as partial or mild paralysis. Unlike paralysis, people with paresis can still move their muscles. These movements are just weaker than normal. Paresis occurs when nerves are damaged.

Discoveries

Three models with excellent results were developed, indicating that, indeed the data from the given questionnaires had information that could be used to develop a model that would make good predictions of the presence of diabetes, certainly among patients of the data source hospital, but perhaps more globally.

It is also possible that the limitation of data to patients coming from a single hospital might mean the model was overfitted to match such patients. It should also be noted that the search for a best model was not

exhaustive. Different models could have been looked at, as well as more hyperparameters could have been tried for each of the models.

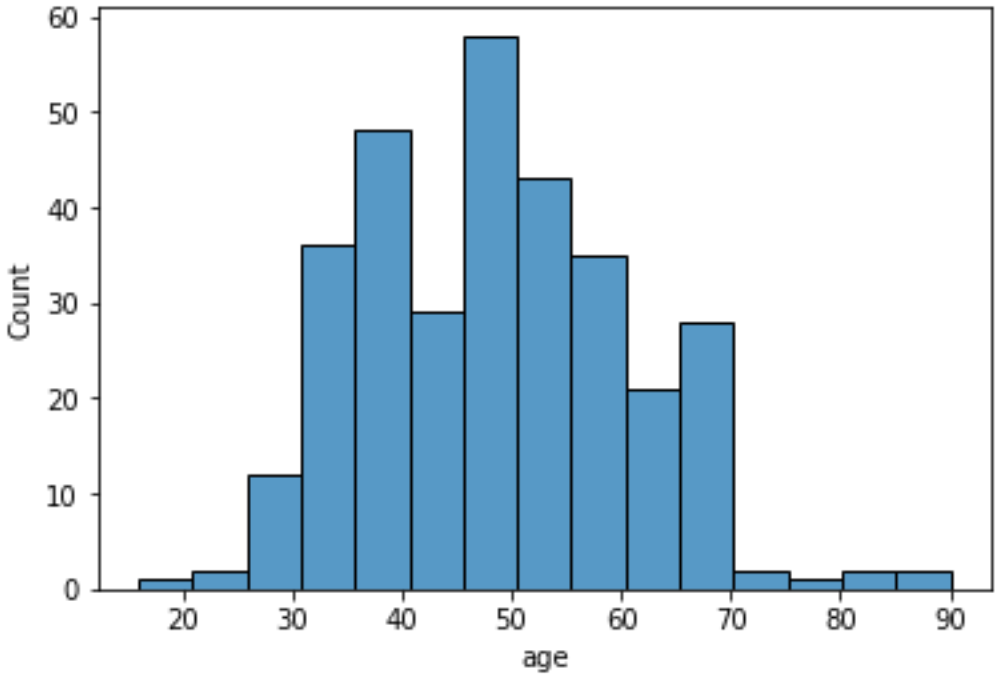
However, keeping all that in mind, the models did perform very well and thus have the possibility to be useful.

Basis

Completion of binary dataset – get the right age division:

Most of the data in the dataset was binary in its initial form. The one exception was age. Since that was the only exception, data exploration was done to try to get a good binary division of age to be able to take advantage of the use of models that work well with all binary data.

The mean age of the entire dataset was about 48 and the median was 47.5. Below is the count of how many there were of each age where diabetes_found=true.



This does not reveal a clear case for an age division that would help with prediction. A qcut division into four age groups was done to look at what percentage of each age group had higher percentages of diabetes:

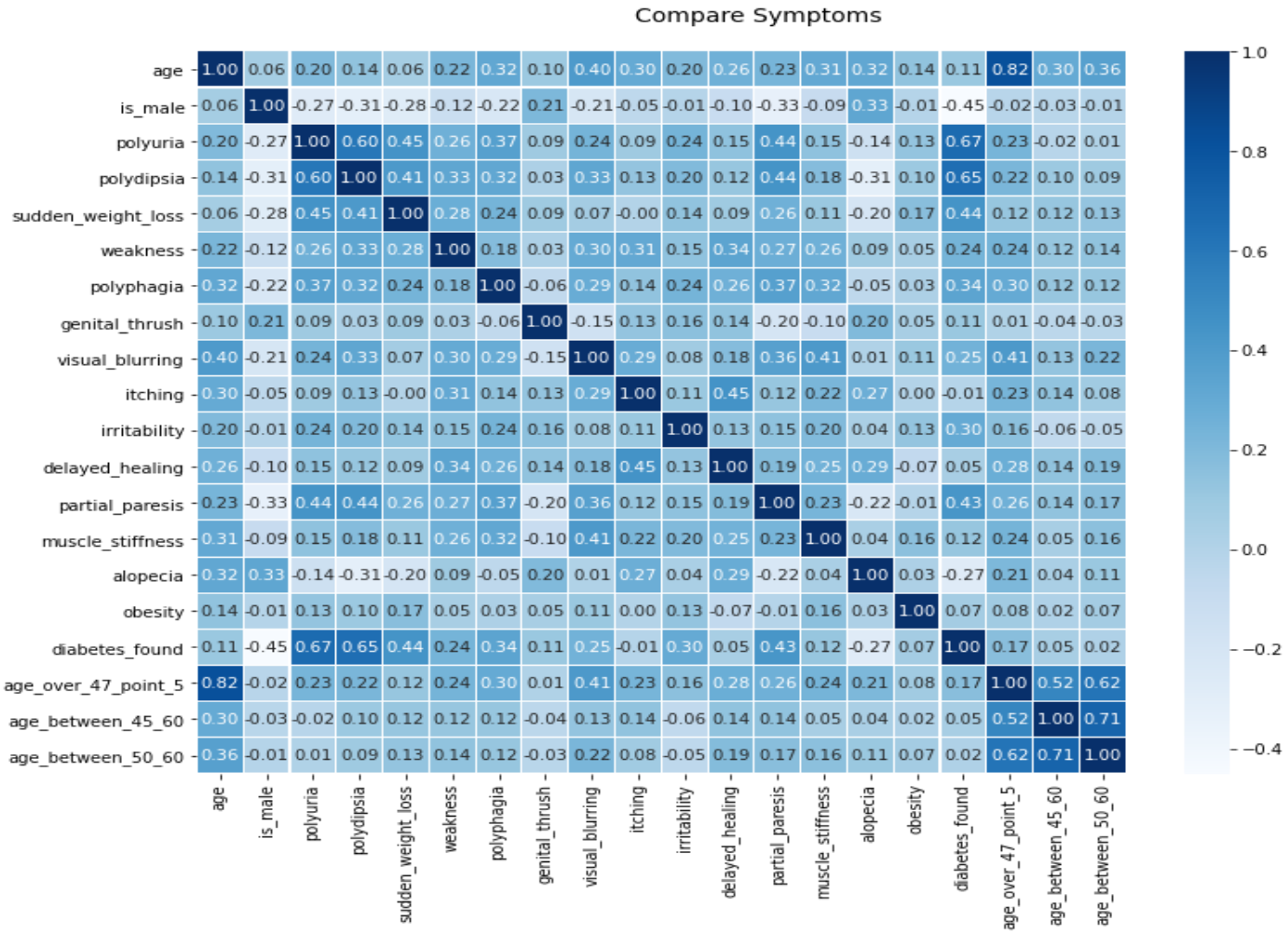
Age range	diabetes found only	all	percentage_diabetes_found
(15.999, 39.0]	85	144	59.027778
(39.0, 47.5]	54	116	46.551724
(47.5, 57.0]	103	143	72.027972
(57.0, 90.0]	78	117	66.666667

Since the last two age groups had relatively high percentages of those with diabetes compared to the other age groups, yet a full count of the two age groups together contained half of the entire population, it seemed like it might be a useful binary cutting point for the use in the prediction model. Thus, a column was created called age_over_47_point_5. (Note, if you look in the notebook, I use “>= 47.5” instead of “> 47.5”. This is technically

incorrect but, in practicality, it didn't matter because the ages in the dataset were all integers so, in the case of ages in this dataset, both ">= 47.5" and "> 47.5" are equivalent to "> 47").

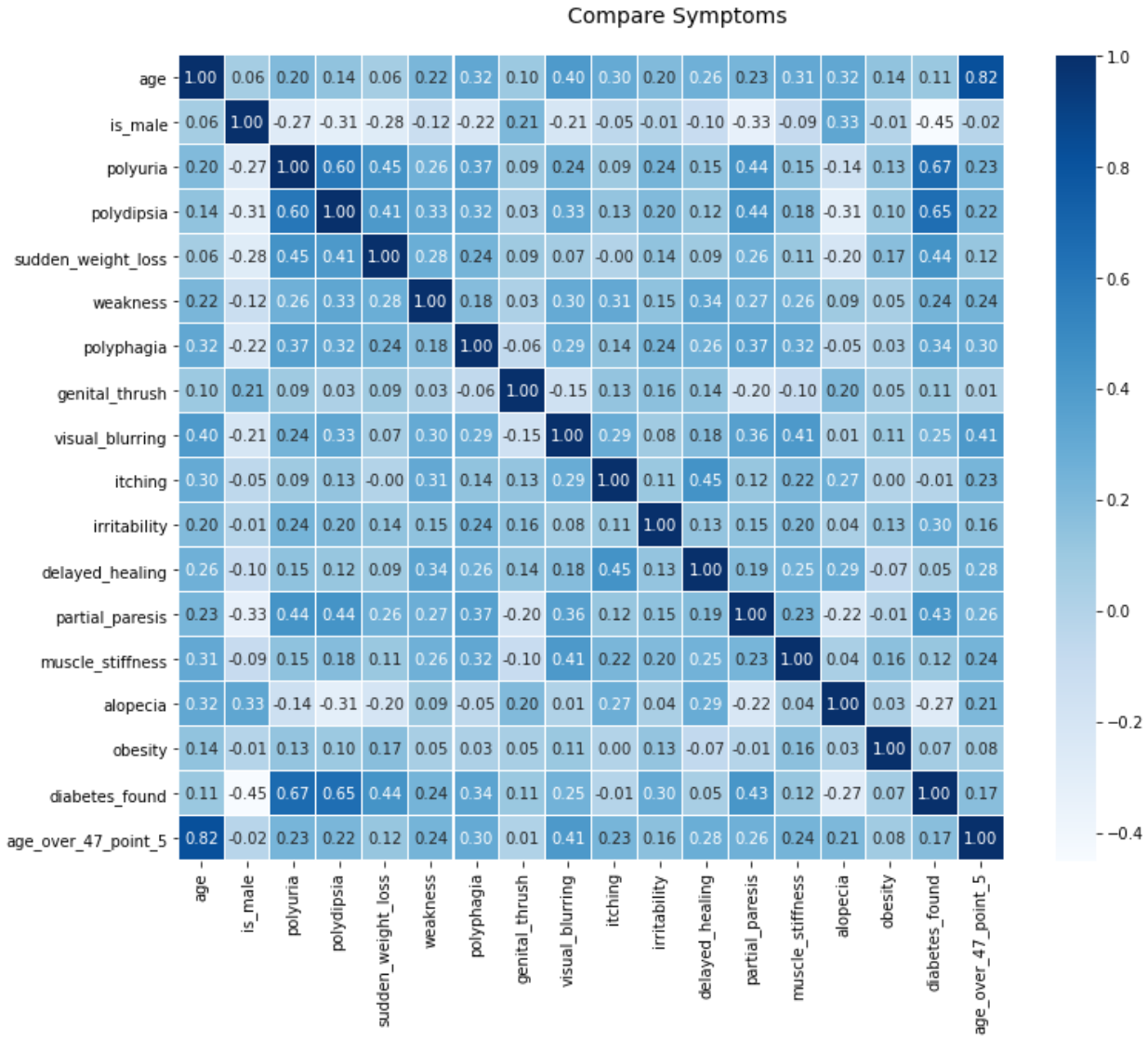
Further investigation was done to see if this division was a good choice. Looking at age count of those with diabetes bar graph above, a high concentration of the population appears to be between ages 45 and 60. Thus, for comparison's sake, columns for those were age between 45 and 60, and those between ages 50 and 60 were created for a correlation test. Only one of the three age division was going to survive in the final binary dataset.

Below is Pearson coefficient correlation heatmap showing all the columns in the dataset. More will be done with a slightly different version of the correlation heatmap later. However, for now, looking at just the last three lines, we'll see if any of the age groups have a strong correlation with the target column diabetes_found. The column age_over_47_point_5 only has a correlation with it of 0.17, which is weak. However, age_between_45_and_60 has a correlation of only 0.05, even worse, and age_between_50_and_60 is worse still at 0.02. Thus, although not looking to be as helpful in prediction as one would hope age would be, age_over_47_point_5 seemed the best bet based on the experimentation done.



Find columns to exclude

Other data discoveries that helped with the development of the model were aided by the same Pearson coefficient correlation heatmap, only done without the unnecessary age columns. One thing that was looked at was the correlation between any two "symptoms" in the dataset. One of the "symptoms" needs to be considered in its own special category because it was the target (diabetes_found – indicating whether the patient tested positive for diabetes), whereas the others were independent variables. The goals were to
1) see if there were any correlation between any two non-target "symptoms" and
2) to see the correlation between the target (diabetes_found) and each of the other "symptoms".
The reasons for each goal and the results of obtaining them are discussed below the heatmap.



1. Initial examination of the independent variable symptoms showed that there was only one somewhat significant correlation (0.60 for polyuria with polydipsia), but it was not a high enough correlation to warrant getting rid of one of them. Thus, there was no immediately obvious reason to get rid of any of the symptoms.

2. Only two of the symptoms showed a high moderate or low strong correlation* with diabetes_found. These were polyuria (0.67) and polydipsia (0.65). All the other symptoms had lower 0.45 or lower correlations with diabetes_found. Based on the heatmap, a list of the top eight correlated symptoms

was produced, including the two mentioned, as well as is_male (-0.45), sudden_weight_loss (0.44), partial_paresis (0.43), polyphagia (0.34), irritability (0.30), visual blurring (0.25). However, it certainly was not clear that only those eight or any number of "top" correlations should be used in the model or, in fact, that any of the sixteen should be excluded. Further kbest testing showed that it was indeed best to include all sixteen symptoms.

*Note: some sources label a Pearson coefficient of absolute(r) = 0.5 or more as "strong", some sources label 0.70 or more as "strong".

See if there is enough data

Although it is not conclusive, the indication is that a training set above would not provide significant improvement to the model 235 (see graph of cross-validation score testing below). Therefore, the training set size of 364 might have been large enough. However, if more data became available, I would certainly attempt to use it. However, there was not a strong case to generate extra data.

The benefits of the training set size seems to level off at about 235 or so but the leveling off is not part of a clearcut pattern that indicates that no larger training set could provide a significant gain. Thus, further exploration might be desired to see if getting more training data is desirable, if it were a situation where that was possible

Model testing

From there, several models were examined – gradient boosting classifier, linear SVC, and logistic regression. The best model (gradient boosting classifier with hyperparameters learning_rate = 0.25, max_depth = 3 and n_estimators = 88) resulted in a Mean Absolute Error (MAE) of only 0.03 when run on the test data.

The two other models that were developed did not do quite as well, but each had very reasonable MAEs of 0.06.

Conclusion

The dataset was from a very limited source (a single hospital) and had only 520 rows, which is not the ideal. However, it also did not have any columns/symptoms that needed to be eliminated to improve the results. This indicates that the data being captured from the questionnaires is useful for diabetes prediction, certainly within the context of that hospital but very possibly more universally. There might be additional symptoms that would have improved the diabetes prediction models developed here, but the symptoms provided offered a promising start. Also, for the same or similar datasets, the three model types, but particularly Gradient Boosting Classifier, would be a good place to start.