Ethan Greenwood, William Zhu MATH 4050

ML Data Project - Diabetes Risk Prediction

Introduction:

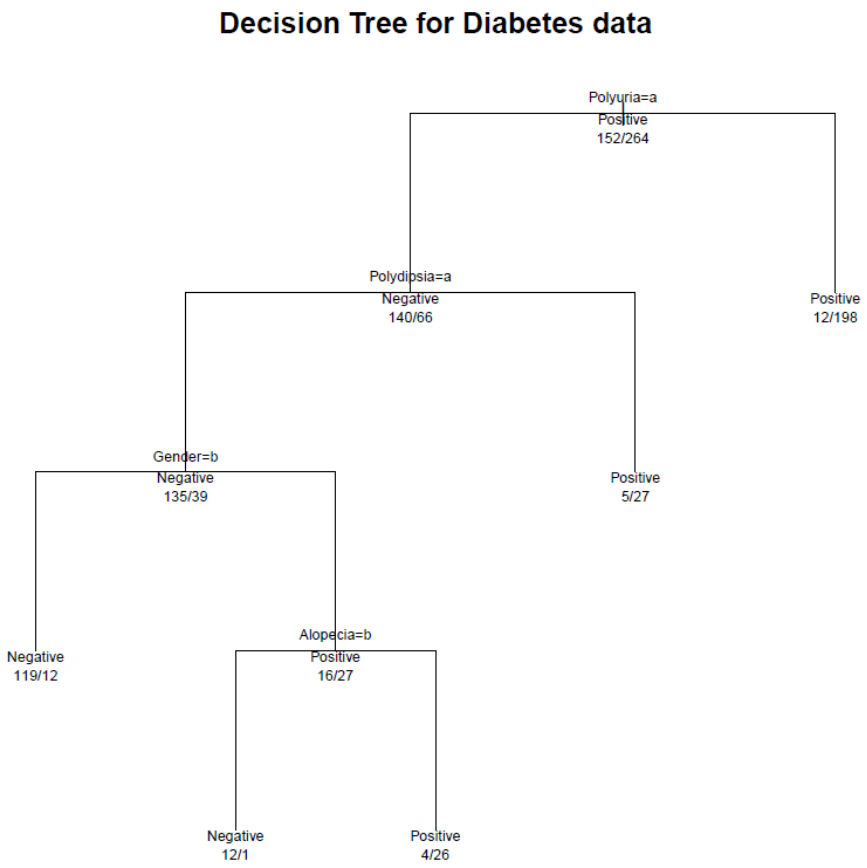
The data that we chose was an early-stage diabetes risk prediction dataset. There are 16 attributes that are all symptoms of diabetes, and the output is whether or not they are positive for diabetes. This doesn't distinguish between type 1 and type 2 diabetes. There are 520 data points in total. This data was stated to be collected by "using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor". The data is not completely reliable because it relies on patient symptom reporting.

What we are intending to learn is how useful the dataset is. In order to examine the data we will use K-Nearest Neighbors and Decision Trees. We choose these algorithms because Decision Trees will tell us the most useful attributes and K-Nearest Neighbors will tell us how useful the dataset is as a whole in predictions.

Results and Discussion:

We cleaned up the data by changing all binary data. We changed the "no" and "yes" to 0 and 1, "male" and "female" to 0 and 1 repsectivaly, and "positive" and "negative" to 1 and 0 respectively. In order to do KNN we also had to normalize the age data. We used an 80 - 20 split of training to test data using the sample function in R.

Decision Tree:



	Negative	Positive
Negative	36	3
Positive	6	59

Overall Accuracy: 0.9134

Negative: 0.9230

Positive: 0.9076

The overall accuracy of the classes were pretty high at 91.34%. Although there were 16 features only a few were used and highlights some symptoms are more of a significant sign than others.

KNN:

	Negative	Positive
Negative	38	0
Positive	3	63

Overall Accuracy: 0.9712

Negative: 1.0000

Positive: 0.9545

KNN was performed with a k of 4. The resulting predictions excel at predicting Negatives (100%), but are not as good at predicting Positives from our features (95.45%).

Conclusion:

From the decision tree we learned that being positive for polyuria is a large indicator or being diabetic. In addition we can have good prediction rates using only a few parameters. This would be useful in real life situations where it's not feasible to test for so many symptoms, and we could still get an accurate prediction. KNN performed much better than decision trees. KNN takes into account features that do not have large information gain but still can influence some cases, whereas the decision tree prunes these features and loses that information. KMeans may have been able to outperform KNN for this data, but with the high accuracy shown by KNN, such a small change does not seem worth the extra time input. For a similar reason, a neural

network also does not seem to be worth the greatly increased amount of effort despite the binary nature of our features providing a seemingly ideal input data set for one.

Citations:

slam, MM Faniqul, et al. 'Likelihood prediction of diabetes at early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125.

R scripts:

See Decision_Tree.R
See Knn.R