Rowan Lochrin
COMP3308
Project 2 Report
21/5/17

1 Introduction

This study was on diabetes in female members of the Pima Indian tribe above the age of 21. The aim of this study was to catalog various attributes of an individuals medical history (Number of times pregnant, Diastolic blood pressure, etc.) as well as weather that individual showed signs of diabetes with the hope that this information could be used to develop tools to help to predict the risk factor for certain individuals, and understand which of the attributes measured predicts an individuals risk of diabetes. This study is relevant as American Indian's suffer from a genetic predisposition to diabetes and analysis of this data may be useful in diagnosing it and making preventative treatment accessible for people with a high risk factor before they develop diabetes.

2 Data

This data set contains 8 numeric attributes as well as weather or not that individual showed signs of diabetes for 768 female members of the Pima Indian tribe over the age of 21. 268 of them showed signs of diabetes 500 of them did not. the attributes are:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)

The attributes that were selected as having the highest correlation with the probability of developing diabetes (CFS) were:

1. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
2. Body mass index (weight in kg/(height in m)^2)
3. Age (years)

3 Results and Discusion

Below are the average accuracies of the classifiers over 10 fold satisfied cross validation.
Weka Classifiers:

	ZeroR	1R	1NN	5NN	NB	DT	MLP	SVM
No-CFS	65.2%	70.8%	69.8%	75.6%	74.7%	74.6%	75.1%	76.4%
CFS	65.1%	70.3%	69.3 %	72.8%	76.8%	72.5%	76.9%	76.2%

My Classifiers:

	1NN	5NN	NB
No-CFS	68.3 %	74.1%	74.8%
CFS	69.3%	72.1%	77.4%

Overall it looks like with some of the classifiers we tested we were able to achieve about the same degree of accuracy as discussed in the study (76%). What was interesting to note is how well the Naive Bayes Classifier preformed in comparison to some of the more complex classifiers in both the CFS and Non-CFS datasets. This suggests that a individuals chance of developing diabetes may be based on a linear combination of these attributes rather than a complicated interaction between them.

Another surprising thing is how well nearest neighbor performs without CFS even though nearest neighbor is usually less reliable in higher dimension spaces. Both my own and Weka's classifier even preforms significantly better without CFS when run with $K = 5$. This leads me to believe that even the values that were not selected by CFS have some predictive power over the class.

Comparing the results from my classifiers it looks like for the most part there were no statistically significant differences in performance. However the numbers are not exactly the same this could be due to a difference in the way Weka implements these algorithms, however it also may be due to the way Weka implements the folding in CFS. I believe Weka randomizes the list

before splitting it into folds and I did not, instead I simply put every line in the fold corresponding to it's line number modulo 10. Weka's method may be better as there may be some underlying pattern in the order of the lines in the data that could produce more homogenous folds leading to artificially high or artificially low accuracies from our classifiers. However in this case it does not look like it screwed results in any particular direction.

4 Conclusion

The best performing classifiers studied in this project were able to achieve around 76% accuracy an improvement of about 10% from a ZeroR classifier (the classifier that just gives the most common result in the training data to every object in the testing data). This score was achieved by the Support Vector Machine, Multilayer Perceptron, and Naive Bayes classifiers. The fact that the Naive Bayes classifier worked as well as more complicated classifiers indicates that the attributes may be somewhat independent in nature.

I'm unsure how this compares to the current methods of predicting diabetes but this strikes me as a fairly low degree of accuracy. Given the variety of classifiers we used it may be that the data we have is insufficient to create an accurate classifier, or that the chance of developing diabetes is partially random in nature. In particular I'd be interested building a classifier that takes into account weekly exercise and how many family members have had diabetes for individuals.

As far as refining the classifiers to work on the current data set, the performance of Nearest neighbor on the No-CFS data set indicates there may be something useful in the attributes we through out in the CFS set. It may be that weighting all attributes to reflect there predictive power rather then throwing out entirely those with lower predictive power, may give us better results.