SID: 450163258
COMP3308 Assignment 2 17/5/2018

Aim

The aim of this study is to determine if the features in Pima Indians Diabetes Database can be used as an indicator for diabetes with good certainty. This information could be used to go through an existing dataset with similar features and find patients who may have or be at risk of developing diabetes.

Results and Discussion

Feature selection was tested. Weka was used to determine features that are highly correlated to the diabetes problem. These were the selected features:

Kept in CFS	**Removed**
Plasma glucose concentration	Number of times pregnant
2-Hour serum insulin	Diastolic blood pressure
Body mass index	Triceps skin fold thickness
Diabetes pedigree function	
Age	

The features generated by weka look like they have high correlation with diabetes. Glucose, insulin, BMI and age are all closely related to existing processes of detecting diabetes

Several Weka classifiers were tested against the custom Nearest neighbour and Naïve Bayes classifiers:

	ZeroR	**1R**	**1NN**	**5NN**	**NB**	**MLP**	**My5NN**	**MyNB**
No feature selection	0.651	0.698	0.679	0.741	0.743	0.755	0.671	0.710
CFS	0.651	0.698	0.690	0.741	0.758	0.757	0.647	0.736

The test data shows the probability of getting a classification correct.

All classifiers were able to work with certainties much higher than 50%. This means that the data is definitely useful for detecting diabetes.

The custom classifiers did not perform as well as the corresponding Weka classifiers. The custom Nearest Neighbour classifier produced results with 7% lower certainty. The custom Naïve Bayes classifier resulted in 3.3% lower certainty.

Naïve Bayes with CFS was most accurate however, the custom Naïve Bayes with CFS was not far behind (2.2%).
CFS significantly (2%) improved performance in Naïve Bayes classifiers. They did not affect other classifiers significantly.

Conclusion

It is possible to predict diabetes in patients with this dataset. There is significant correlation with some of the features in the dataset and the likelihood of diabetes to allow about a 75% correct prediction. More data and training in the future on a multilayer perceptron could provide much better results as the features can be dynamically weighted.

Reflection

I really liked this assignment. Developing the Naïve Bayes and Nearest Neighbour algorithms in python really helped me understand some of the maths concepts that I was unsure about in this unit. A specific thing that I learned in this assignment was the process of training and testing in classifiers. I was also able to practice some algorithms/patterns and learn how to do mathematical calculations in Python