Dataset

The dataset we are using is the Pima Indians Diabetes data (pima.csv) that provides a sample size of 768 people who shared information relevant to diabetes and whether they have diabetes or not (500 do have, 267 do not have). The source and owner of the data is the National Institute of Diabetes and Digestive and Kidney Diseases. The data has been modified by replacing missing values with average values for the purpose of the assignment.

All patient are female and at least 21 years old, and they all provided the following information:

* Number of times they were pregnant
* Plasma glucose concentration
* Blood pressure
* Skin fold thickness
* Insulin they received
* BMI
* Diabetes pedigree function
* Age
* Whether they have diabetes or not

With this dataset, the aim was to predict diabetes outcome in individuals with the above information. Apart from that, this dataset was modified with Correlation-based feature selection using the method Best-First Search in Weka software. With that, 6 variables were kept:

* Plasma glucose concentration
* Insulin they received
* BMI
* Diabetes pedigree function
* Age
* Whether they have diabetes or not

During our work, we worked with both dataset and compared results.

Result and Discussion

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ZeroR | 1R | 1NN | 5NN | NB | DT | MLP | SVM | RF |
| No Feature Selection | 65.1042 | 70.8333 | 67.8385 | 74.4792 | 75.1302 | 71.7448 | 75.3906 | 76.3021 | 74.8698 |
| CFS | 65.1042 | 70.8333 | 69.0104 | 74.4792 | 76.3021 | 73.3073 | 75.7813 | 76.6927 | 75.9115 |

Comparison between classifiers (Weka)

Overall, the SVM algorithm performed the best with 76-77% accuracy. While the ZeroR algorithm performed the worst.

ZeroR - 1R

Since there are 500 non-diabetes and 268 diabetes the ZeroR algorithm label every value as non-diabetes for majority reason. However, if we were testing on a different dataset, this accuracy could easily be worse which means that the classifier is not reliable.

Although 1R algorithm performed better, it only uses 1 feature which realistically not a good sign for predicting diabetes. There are many factors that together cause diabetes and relying one 1 factor is a bad strategy. On top of that, the selection of 1 attribute is not perfect if we work with continuous values (real numbers) instead of nominal values. That is because the feature selection is based on the best frequency table in terms of prediction that the 1R algorithm creates for every feature.

NB

While 1R classifies based on 1 feature, NB uses all features so the accuracy is significantly higher as expected. However, we do not know if the features we are using here are independent from each other and that could be an issue. On top of that we made the assumption that all variables here are equally important. For example BMI could be more important than age or blood pressure and glucose concentration could be related to one another. Hence, we may not want to use this algorithm in medicine practice.

RF-DT

When it comes to Decision Tree and Random Forest, RF is almost like an extension of DT. It is fair to expect that RF performs better as it uses multiple trees to make prediction, but it seems like the difference between DT and RF in accuracy is around 1-2%. Considering that the computation time of RF is longer than for the DT but the result is similar, we might only stick with the DT algorithm.

Effect of Feature Selection

As we can see in the table, using the CFS dataset, Weka gave us an overall better accuracy for almost every algorithm than for the original dataset where we did not select features. That could be due to the fact that there were many unnecessary variables included in the original dataset that make noises for the classification algorithms. Also, working with less variable reduces the computational time of our program. So removing them helps the performance overall.

Reflection

One of the most interesting and new experience during the assignment was definitely related to Weka. I have never used it before, but found it convenient to use. I would suggest it to people who are interested in AI but do not have experience in coding. Overall, it was an enjoyable assignment that was challenging and interesting at the same time as one of the few occasions where we really dig into some advanced algorithms.

Conclusion

Although we see that different algorithms has different result, it does not mean one is better than the other. In this work we considered a lot of classifiers but in practice, it is crucial to understand the project we are working with. We must select classifier based on the domain knowledge and information we gather so that we save time from unnecessary work.

In our case, SVM seemed to performed the best after removing some attributes, but it does not mean the classifier works in every situation. In other topics, such as image recognition or gene expression analysis, we might want to use other classifier. On top of that, even the best algorithm in this assignment has room for improvement. Though 76% accuracy is good, it is not great especially in medicine where people’s life are on the line.

Hence, for future work, it is suggested to work with different dataset – maybe include male patients or children. There could be other factors that causes diabetes that were not in our dataset, and it is important to understand the interaction between the attributes. That means using the classifiers with some context and domain knowledge could further improve the prediction of diabetes in pantients.