COMP3608: Assignment 2

Due on Friday 14th May, 2021 at 11:59pm

490379581, 490379075

Contents

1 Introduction

The aim of the study is to evaluate the effectiveness of a selection of different classifiers on the Pima Indians dataset [Smith et al., 1988]. In doing so, we will also compare the effects of pre-processing techniques by using numeric and nominal data as well as data with and without feature selection. In comparing our implementations of the Naïve Bayes (NB) and Decision Tree (DT) classification algorithms to the ones implemented in Weka [Frank et al., 2016], we assess the reproducibility of results for fundamental algorithmic concepts.

The study of classification algorithms is important as even simple classifiers can be extremely powerful and applicable to problems across many sectors, such as health diagnosis. While it is possible to implement highly accurate classifiers using complex state-of-the-art techniques such as deep learning, these techniques are often time and resource consuming, as well as difficult to justify results due to their black-box nature. Therefore, it is a valuable exercise to investigate if we can create simple yet powerful classifiers that work efficiently and logically, which is the aim of this study. Furthermore, if simple pre-processing techniques can provide better classification performance at the cost of little time or computational effort, then basic classifiers such as those used in this study can offer cheap, fast and meaningful results where complex classifiers are infeasible. If efficient yet simple classifiers offer competition to more complex ones in terms of real-world performance, the range of problems to which they might be applied grows, from which, the various insights offered by machine learning can be harnessed in more instances. For example, the prediction of diabetes in Pima Indians could lead to preventative or more informed treatment, which is of particular importance for disadvantaged communities.

2 Data

The dataset used is a modified version of the Pima Indians Diabetes Database. Any missing values in a row were replaced with the average value of that attribute in the dataset. All subjects are females at least 21 years of age and all are of Pima Indian heritage. There are a total of 768 subjects and 8 predictive features plus the class attribute, as described below:

Attribute	Description	Units
Features		
Pregnancies	Number of times pregnant	*integer*
Plasma Glucose Concentration	Level after 2 hours in an oral glucose tolerance test	*integer*
Diastolic Blood Pressure		$mmHg$
Triceps Skin Fold Thickness		mm
2-hour Serum Insulin		$\mu U/ml$
Body Mass Index		kg/m^2
Diabetes Pedigree Function		$float$
Age	Age in years	*integer*
Class		
Diabetes	Whether this person tested positive for diabetes	*yes* or *no*

Table 1: Description of dataset attributes.

To perform feature selection we used Weka's inbuilt Correlation-based Feature selection (CFS), the central idea of CFS being that good predictive features are highly correlated with the class attribute, but are uncorrelated with each other. The highlighted cells in Table 1 are the features selected using Weka's CFS with the Best First search method option.

The data was also provided in a discretised format, where all attributes had been transformed into ordered, nominalised values. Running Weka's inbuilt CFS method on this dataset gave the same subset of

features as in the numeric data. This format of the data was used for the DTs, Bagging, Boosting and Random Forest (RF) algorithms.

3 Results and Discussion

We present results from running the various classification algorithms on Weka and compare these to our own results. We used 10-fold Cross Validation (CV) to evaluate all algorithms, both in Weka and in our own implementations of the classifiers. All algorithms run in the Weka environment use standard settings except for Weka's Bagging and Boosting algorithms where we have used J48 trees as the base tree instead of the default recommended trees for more meaningful comparison of standardised results. The classification algorithm with the highest accuracy in each row is highlighted green, and that with the lowest accuracy is highlighted red:

Classification Algorithm (% Accuracy 2 d.p.)

Numeric	ZeroR	1R	1NN	5NN	NB	MLP	SVM	MyNB
No feature selection	65.10	70.83	67.84	74.48	75.13	75.39	76.30	74.61
CFS	65.10	70.83	69.01	74.48	76.30	75.78	76.69	76.31

Nominal Data	DT unpruned	DT pruned	MyDT	Bagg	Boost	RF
No feature selection	75.39	75.00	74.61	74.87	76.17	73.18
CFS	79.42	79.42	78.53	78.52	78.65	78.91

Table 2: Algorithm classification accuracies using 10-fold CV.

3.1 Effect of CFS

From these results, we see that CFS either improved or maintained the accuracy for every algorithm. By only retaining a subset of five of the original features, CFS has likely led to an increase in accuracies by filtering out irrelevant features, and thus preventing overfitting of the classifiers. We note that this makes no change to the ZeroR algorithm, since this unintelligently guesses the class attribute. The selected subset of features from CFS - Plasma Glucose Concentration, 2-hour Serum Insulin, Body Mass Index, Diabetes Pedigree Function and Age - can be validated as sound features for classification as they have been shown to have correlation with diabetes in other studies [Abdul-Ghani and DeFronzo, 2009], [Saxena et al., 2011], [Gray et al., 2015], [Geman et al., 2017]. Furthermore, using CFS simplified both the dataset and the classifiers, as the classifiers only had to operate on five attributes instead of eight. This is especially meaningful for the DTs, as the trees will have a lower depth and less leaves. Overall, using CFS to perform feature selection was beneficial to the construction of simplistic yet powerful classifiers.

3.2 Effect of Discretisation

The classifiers that used the nominal data had a higher accuracy on average, with the best performing classifier using nominal data. Discretisation may have led to an increase in accuracies for a similar reason to CFS but through different methods. By subsetting data into fewer categories within each feature, this may have increased classification accuracies through better generalisation to data outside the training set and by reducing overfitting.

3.3 Comparison of Classifiers

For the classifiers using the numeric data, the most accurate algorithm was the Support Vector Machine (SVM), and the least accurate was ZeroR, both irrespective of CFS. For the classifiers using the nominal data, the most accurate algorithm without CFS was Boosting, and the least accurate was the Random Forest (RF). With CFS, the most accurate algorithm was the pruned/unpruned DTs, and the least accurate was Bagging. Overall, the most accurate classifier was the pruned/unpruned DTs with CFS. For this study, the important difference between these algorithms is their complexity and computational effort. Here, we note that algorithms such as 1R are incredibly simplistic and easy to implement yet still provide a reasonable accuracy. We can compare this to the Multi-Layer Perceptron (MLP), which is relatively more complex and took longer to build in Weka, yet still has the advantage of providing a higher accuracy. We compare this to the optimal algorithms, the pruned/unpruned DTs, as they are both reasonably simplistic to implement, easy to interpret and understand and yet provide a substantially better accuracy. Therefore, despite being more complex than 1R, the additional complexity is worth the added performance. With regards to the tree-based classifiers, while RF added more complexity, it had a lower accuracy than both Bagging and Boosting when run on data without CFS, but when run on data with CFS, it had the highest accuracy out of the three.

3.4 Comparison of Our Implementation to Weka

Our implementation of both the NB and DT algorithms all had marginally lower accuracies than those in Weka, except for our NB with CFS, which had an accuracy 0.01% higher than the NB in Weka. The difference in accuracies being less than 1% between implementations may be due to the randomness of the 10-fold CV folds, the stratification process, or minor differences in implementation. However, it is worth noting that all except the NB with CFS having lower accuracies could indicate a systematic difference between them. In the appendix, we include a text representation of our implementation of the DT without CFS (Appendix A), our implementation of the DT with CFS (Appendix B), the unpruned DT in Weka without CFS (Appendix C), the unpruned DT in Weak with CFS (Appendix D), the pruned DT in Weka without CFS (Appendix E) and the pruned DT in Weka with CFS (Appendix F).

3.5 Effect of Pruning DTs

For algorithms run on CFS nominal data, the unpruned and pruned Decision Trees were equally the most accurate and Bagging on the J48 trees the least accurate. Pruning of the DT led to a minor reduction in accuracy by 0.39% when run on data with no feature selection, but had no significant difference when run on data with CFS. As seen in the table, the unpruned DT in Weka with CFS and the pruned DT in Weka with CFS have the same accuracy, but as seen in the appendix, their tree diagrams are different (see Appendix D and Appendix F). On further inspection in Weka, the root mean squared error and receiver operating characteristic curve are different, so there are subtle differences in the classifiers that is masked by using accuracy as the sole performance metric. Based off the results, the unpruned trees performed better than the pruned trees, which suggests that pruning is ineffective. However, the simplicity of the pruned DTs, which reduced the size of the tree without CFS by 75% from 59 nodes to 15 nodes, outweighs the slight gain in accuracy. In addition, the computational cost of pruning was minimal. Therefore, pruning had an overall beneficial effect on the construction of the DT classifiers in terms of building simple yet powerful classifiers.

3.6 Assessing the Performance of Classifiers

Accuracy of correctly classified examples is the simplest metric to evaluate the performance of the classifiers. Using accuracy as a performance metric is somewhat useful for this dataset as there are 500 subjects without diabetes and 268 with diabetes. However, this means that an unintelligent classifier that simply guesses "no" for all testing examples can perform with 65.10% accuracy and should be considered a baseline, which is how the ZeroR algorithm in Weka performed. Therefore, the accuracies of all other algorithms should be compared to this baseline, and since they are higher, we infer that the algorithms are identifying features in a meaningful way which lead to more accurate classifications. Using 10 fold CV was used in order to

provide a more robust measure for accuracy as it eliminates the randomness of the train-test split and a better indication of how it will perform on unseen data.

4 Conclusions and Future Works

The aim of the study was to evaluate the effectiveness of a selection of different classifiers on the Pima Indians dataset, as well as evaluating the difference in performance when using numeric as opposed to discretised data, and data with and without feature selection. The above results show that algorithms that used the nominal Pima Indians data with CFS had the highest accuracies, and algorithms run on the numerical data with no feature selection had the lowest accuracies. Our implementations of NB and DT performed reasonably well in comparison to the corresponding classifier in Weka, with all differences in accuracies between implementations being less than 1%. This gives two meaningful conclusions. Firstly, the results indicate that it is possible to build simple classifier algorithms that can predict classes with relatively high accuracy. Such algorithms can offer insights in most contexts, but may not be suitable in making significant decisions with substantial consequences. However, given their value in some contexts, the second meaningful conclusion drawn is the value of programs such as Weka to easily and efficiently implement such algorithms. The accessibility of programs such as Weka, that are able to build simple yet powerful classifiers with beneficial pre-processing, is increasingly facilitating the use of machine learning for widespread benefit.

Possible work for the future mainly includes running the same algorithms on different datasets in both numerical and nominally discretised formats, and both with and without CFS in order to determine if these findings can be extrapolated to other datasets and applications. In doing this, we could test our hypothesis that using discretised data and CFS in classification algorithms results in better performance. Many other feature selection methods exist, such as Pearsons Correlation, Linear Discriminant Analysis and ANOVA. It would be also be useful to compare the effect of these methods on the same data. Further investigation into whether or not there is a systematic difference between our implementations of the NB and DT algorithms and Weka is also possible for future work.

As discussed, using accuracy as the sole metric of performance is a naïve approach to estimating an algorithms real-world performance as it is not robust to datasets with highly skewed classes. Future work might also include the implementation of and comparison between different measures of performance, such as Recall, Precision and F1 Score.

5 Reflection

In this assignment, collaboration through Git and Overleaf played a crucial role in formulating the code and report, especially in order to submit by the assignment deadline. Being able to access the most up-to-date version of both the code and the report were essential in distributing the workload evenly and ensuring both members of the team were on the same page about what stage of the assignment we were at. This was definitely the biggest takeaway that we both had from the assignment.

Regarding the implementation of these algorithms ourselves, it was enjoyable and interesting to compare our results to the black-box implementations from Weka. Though NB is a very simple algorithm to implement, it was surprising how powerful it was, which really underpinned the purpose and direction of the report. We found it quite amazing that you can achieve high classification accuracy just from the mean and variance of features. On another note, implementing the DTs ourselves allowed us to fine-tune them, such as using gain ratio instead of information gain, or splitting unseen categories in testing down previously known categories and taking weighted average or simply taking the majority class of that leaf. Having this control over the logic of the algorithm, which we wouldn't have had or understood if we didn't implement this ourselves, made this assignment much more informative and fun.

References

[Abdul-Ghani and DeFronzo, 2009] Abdul-Ghani, M. A. and DeFronzo, R. A. (2009). Plasma glucose concentration and prediction of future risk of type 2 diabetes. *Diabetes Care*, 32(suppl_2):S194–S198.

[Frank et al., 2016] Frank, E., Hall, M. A., and Witten, I. H. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Fourth edition.

[Geman et al., 2017] Geman, O., Chiuchisan, I., and Toderean, R. (2017). Application of adaptive neuro-fuzzy inference system for diabetes classification and prediction. In *2017 E-Health and Bioengineering Conference (EHB)*. IEEE.

[Gray et al., 2015] Gray, N., Picone, G., Sloan, F., and Yashkin, A. (2015). Relation between BMI and diabetes mellitus and its complications among US older adults. *Southern Medical Journal*, 108(1):29–36.

[Saxena et al., 2011] Saxena, P., Prakash, A., and Nigam, A. (2011). Efficacy of 2-hour post glucose insulin levels in predicting insulin resistance in polycystic ovarian syndrome with infertility. *Journal of Human Reproductive Sciences*, 4(1):20.

[Smith et al., 1988] Smith, J., Everhart, J., Dickson, W., Knowler, W., and Johannes, R. (1988). Using the ADAP Learning Algorithm to Forcast the Onset of Diabetes Mellitus. *Proceedings - Annual Symposium on Computer Applications in Medical Care*, 10.

Appendices

A MyDT without CFS

```
Plasma Glucose Concentration = high
|   Body Mass Index = high
|   |   Age = high
|   |   |   Diabetes Pedigree Function = high
|   |   |   |   Diastolic Blood Pressure = high
|   |   |   |   |   Pregnancies = low
|   |   |   |   |   |   Triceps Skin Fold Thickness = high
|   |   |   |   |   |   |   2-hour Serum Insulin = high: yes (9/4)
|   |   |   |   |   |   |   2-hour Serum Insulin = low: yes (1/0)
|   |   |   |   |   |   Triceps Skin Fold Thickness = low: yes (1/0)
|   |   |   |   |   Pregnancies = high: yes (12/2)
|   |   |   |   Diastolic Blood Pressure = low
|   |   |   |   |   Pregnancies = low: yes (1/1)
|   |   |   |   |   Pregnancies = high: no (2/0)
|   |   |   Diabetes Pedigree Function = low
|   |   |   |   2-hour Serum Insulin = high
|   |   |   |   |   Triceps Skin Fold Thickness = high
|   |   |   |   |   |   Diastolic Blood Pressure = high
|   |   |   |   |   |   |   Pregnancies = high: yes (9/6)
|   |   |   |   |   |   |   Pregnancies = low: no (12/11)
|   |   |   |   |   |   Diastolic Blood Pressure = low
|   |   |   |   |   |   |   Pregnancies = low: yes (2/1)
|   |   |   |   |   |   |   Pregnancies = high: yes (3/0)
|   |   |   |   |   Triceps Skin Fold Thickness = low
|   |   |   |   |   |   Pregnancies = low
|   |   |   |   |   |   |   Diastolic Blood Pressure = high: yes (1/1)
|   |   |   |   |   |   |   Diastolic Blood Pressure = low: no (1/0)
|   |   |   |   |   |   Pregnancies = high: no (1/0)
|   |   |   |   2-hour Serum Insulin = low: yes (1/0)
|   |   Age = low
|   |   |   Triceps Skin Fold Thickness = high
|   |   |   |   Diabetes Pedigree Function = low
|   |   |   |   |   Diastolic Blood Pressure = high: no (12/8)
|   |   |   |   |   Diastolic Blood Pressure = low: yes (4/3)
|   |   |   |   Diabetes Pedigree Function = high
|   |   |   |   |   Diastolic Blood Pressure = high: yes (6/5)
|   |   |   |   |   Diastolic Blood Pressure = low: no (3/2)
|   |   |   Triceps Skin Fold Thickness = low
|   |   |   |   Diastolic Blood Pressure = high: no (3/0)
|   |   |   |   Diastolic Blood Pressure = low
|   |   |   |   |   2-hour Serum Insulin = high
|   |   |   |   |   |   Diabetes Pedigree Function = high: yes (1/1)
|   |   |   |   |   |   Diabetes Pedigree Function = low: no (1/0)
|   |   |   |   |   2-hour Serum Insulin = low: no (1/0)
|   Body Mass Index = low
|   |   Triceps Skin Fold Thickness = high
|   |   |   2-hour Serum Insulin = high
|   |   |   |   Diabetes Pedigree Function = high: no (5/0)
|   |   |   |   Diabetes Pedigree Function = low
```

```
|   |   |   |   |   Age = high
|   |   |   |   |   |   Diastolic Blood Pressure = high: no (5/1)
|   |   |   |   |   |   Diastolic Blood Pressure = low: no (2/0)
|   |   |   |   |   Age = low
|   |   |   |   |   |   Diastolic Blood Pressure = low: yes (2/2)
|   |   |   |   |   |   Diastolic Blood Pressure = high: no (1/0)
|   |   |   2-hour Serum Insulin = low
|   |   |   |   Diabetes Pedigree Function = high: yes (1/0)
|   |   |   |   Diabetes Pedigree Function = low: no (1/0)
|   |   Triceps Skin Fold Thickness = low: no (9/0)
Plasma Glucose Concentration = low
|   Body Mass Index = low: no (66/0)
|   Body Mass Index = high
|   |   2-hour Serum Insulin = high
|   |   |   Age = low
|   |   |   |   Diastolic Blood Pressure = low
|   |   |   |   |   Triceps Skin Fold Thickness = low: no (7/0)
|   |   |   |   |   Triceps Skin Fold Thickness = high
|   |   |   |   |   |   Diabetes Pedigree Function = low: no (9/3)
|   |   |   |   |   |   Diabetes Pedigree Function = high: no (5/1)
|   |   |   |   Diastolic Blood Pressure = high: no (18/0)
|   |   |   Age = high
|   |   |   |   Diabetes Pedigree Function = low
|   |   |   |   |   Triceps Skin Fold Thickness = high
|   |   |   |   |   |   Pregnancies = high
|   |   |   |   |   |   |   Diastolic Blood Pressure = high: no (8/1)
|   |   |   |   |   |   |   Diastolic Blood Pressure = low: no (1/0)
|   |   |   |   |   |   Pregnancies = low
|   |   |   |   |   |   |   Diastolic Blood Pressure = high: no (11/2)
|   |   |   |   |   |   |   Diastolic Blood Pressure = low: no (3/1)
|   |   |   |   |   Triceps Skin Fold Thickness = low: no (1/0)
|   |   |   |   Diabetes Pedigree Function = high
|   |   |   |   |   Diastolic Blood Pressure = high: yes (3/3)
|   |   |   |   |   Diastolic Blood Pressure = low: yes (1/0)
|   |   2-hour Serum Insulin = low
|   |   |   Diastolic Blood Pressure = high
|   |   |   |   Age = high: no (12/0)
|   |   |   |   Age = low
|   |   |   |   |   Triceps Skin Fold Thickness = high
|   |   |   |   |   |   Diabetes Pedigree Function = low: no (5/1)
|   |   |   |   |   |   Diabetes Pedigree Function = high: yes (1/0)
|   |   |   |   |   Triceps Skin Fold Thickness = low: no (6/0)
|   |   |   Diastolic Blood Pressure = low: no (23/0)
Plasma Glucose Concentration = very high
|   2-hour Serum Insulin = high
|   |   Body Mass Index = low
|   |   |   Age = high
|   |   |   |   Triceps Skin Fold Thickness = high
|   |   |   |   |   Pregnancies = high
|   |   |   |   |   |   Diabetes Pedigree Function = high
|   |   |   |   |   |   |   Diastolic Blood Pressure = low: yes (1/0)
|   |   |   |   |   |   |   Diastolic Blood Pressure = high: yes (1/1)
|   |   |   |   |   |   Diabetes Pedigree Function = low: yes (2/0)
|   |   |   |   |   Pregnancies = low: yes (2/2)
```

```
|   |   |   |   Triceps Skin Fold Thickness = low: yes (3/0)
|   |   |   Age = low
|   |   |   |   Diastolic Blood Pressure = low: no (1/0)
|   |   |   |   Diastolic Blood Pressure = high
|   |   |   |   |   Triceps Skin Fold Thickness = low: yes (1/1)
|   |   |   |   |   Triceps Skin Fold Thickness = high: no (1/0)
|   |   Body Mass Index = high
|   |   |   Pregnancies = low
|   |   |   |   Age = high
|   |   |   |   |   Diabetes Pedigree Function = low
|   |   |   |   |   |   Diastolic Blood Pressure = high
|   |   |   |   |   |   |   Triceps Skin Fold Thickness = high: yes (14/2)
|   |   |   |   |   |   |   Triceps Skin Fold Thickness = low: yes (3/0)
|   |   |   |   |   |   Diastolic Blood Pressure = low
|   |   |   |   |   |   |   Triceps Skin Fold Thickness = low: yes (1/1)
|   |   |   |   |   |   |   Triceps Skin Fold Thickness = high: yes (3/0)
|   |   |   |   |   Diabetes Pedigree Function = high
|   |   |   |   |   |   Triceps Skin Fold Thickness = high
|   |   |   |   |   |   |   Diastolic Blood Pressure = high: yes (10/5)
|   |   |   |   |   |   |   Diastolic Blood Pressure = low: yes (1/1)
|   |   |   |   |   |   Triceps Skin Fold Thickness = low: yes (1/0)
|   |   |   |   Age = low
|   |   |   |   |   Diabetes Pedigree Function = high: yes (12/0)
|   |   |   |   |   Diabetes Pedigree Function = low
|   |   |   |   |   |   Triceps Skin Fold Thickness = high
|   |   |   |   |   |   |   Diastolic Blood Pressure = high: yes (7/2)
|   |   |   |   |   |   |   Diastolic Blood Pressure = low: yes (3/0)
|   |   |   |   |   |   Triceps Skin Fold Thickness = low
|   |   |   |   |   |   |   Diastolic Blood Pressure = high: yes (1/0)
|   |   |   |   |   |   |   Diastolic Blood Pressure = low: no (1/0)
|   |   |   Pregnancies = high
|   |   |   |   Diabetes Pedigree Function = high: yes (16/0)
|   |   |   |   Diabetes Pedigree Function = low
|   |   |   |   |   Diastolic Blood Pressure = high: yes (12/3)
|   |   |   |   |   Diastolic Blood Pressure = low: yes (3/1)
|   2-hour Serum Insulin = low
|   |   Diabetes Pedigree Function = low: no (2/0)
|   |   Diabetes Pedigree Function = high: yes (1/0)
Plasma Glucose Concentration = medium
|   Age = high
|   |   Body Mass Index = low
|   |   |   Diastolic Blood Pressure = high
|   |   |   |   Pregnancies = low
|   |   |   |   |   Diabetes Pedigree Function = low
|   |   |   |   |   |   Triceps Skin Fold Thickness = high: no (2/1)
|   |   |   |   |   |   Triceps Skin Fold Thickness = low: no (2/0)
|   |   |   |   |   Diabetes Pedigree Function = high: no (3/0)
|   |   |   |   Pregnancies = high: no (13/0)
|   |   |   Diastolic Blood Pressure = low
|   |   |   |   Pregnancies = low
|   |   |   |   |   Triceps Skin Fold Thickness = high: no (5/0)
|   |   |   |   |   Triceps Skin Fold Thickness = low: no (2/1)
|   |   |   |   Pregnancies = high: yes (1/0)
|   |   Body Mass Index = high
```

```
|   |   |   Diabetes Pedigree Function = low
|   |   |   |   2-hour Serum Insulin = high
|   |   |   |   |   Diastolic Blood Pressure = high
|   |   |   |   |   |   Pregnancies = high: no (14/12)
|   |   |   |   |   |   Pregnancies = low
|   |   |   |   |   |   |   Triceps Skin Fold Thickness = high: no (18/11)
|   |   |   |   |   |   |   Triceps Skin Fold Thickness = low: yes (1/1)
|   |   |   |   |   Diastolic Blood Pressure = low
|   |   |   |   |   |   Triceps Skin Fold Thickness = high
|   |   |   |   |   |   |   Pregnancies = high: yes (3/3)
|   |   |   |   |   |   |   Pregnancies = low: yes (5/4)
|   |   |   |   |   |   Triceps Skin Fold Thickness = low: no (2/1)
|   |   |   |   2-hour Serum Insulin = low: no (5/0)
|   |   |   Diabetes Pedigree Function = high
|   |   |   |   Pregnancies = low
|   |   |   |   |   Triceps Skin Fold Thickness = high
|   |   |   |   |   |   Diastolic Blood Pressure = high: yes (9/7)
|   |   |   |   |   |   Diastolic Blood Pressure = low: yes (3/3)
|   |   |   |   |   Triceps Skin Fold Thickness = low: yes (2/0)
|   |   |   |   Pregnancies = high: yes (13/0)
|   Age = low
|   |   Body Mass Index = high
|   |   |   Triceps Skin Fold Thickness = high
|   |   |   |   Pregnancies = low
|   |   |   |   |   Diabetes Pedigree Function = high
|   |   |   |   |   |   Diastolic Blood Pressure = high
|   |   |   |   |   |   |   2-hour Serum Insulin = high: no (12/2)
|   |   |   |   |   |   |   2-hour Serum Insulin = low: no (3/0)
|   |   |   |   |   |   Diastolic Blood Pressure = low
|   |   |   |   |   |   |   2-hour Serum Insulin = high: yes (3/3)
|   |   |   |   |   |   |   2-hour Serum Insulin = low: yes (1/0)
|   |   |   |   |   Diabetes Pedigree Function = low
|   |   |   |   |   |   Diastolic Blood Pressure = low
|   |   |   |   |   |   |   2-hour Serum Insulin = high: no (18/2)
|   |   |   |   |   |   |   2-hour Serum Insulin = low: no (5/1)
|   |   |   |   |   |   Diastolic Blood Pressure = high
|   |   |   |   |   |   |   2-hour Serum Insulin = high: no (20/5)
|   |   |   |   |   |   |   2-hour Serum Insulin = low: no (3/0)
|   |   |   |   Pregnancies = high: yes (1/1)
|   |   |   Triceps Skin Fold Thickness = low
|   |   |   |   Diabetes Pedigree Function = high
|   |   |   |   |   Diastolic Blood Pressure = low
|   |   |   |   |   |   2-hour Serum Insulin = high: no (3/1)
|   |   |   |   |   |   2-hour Serum Insulin = low: no (2/0)
|   |   |   |   |   Diastolic Blood Pressure = high: no (4/0)
|   |   |   |   Diabetes Pedigree Function = low: no (14/0)
|   |   Body Mass Index = low
|   |   |   Diabetes Pedigree Function = low: no (34/0)
|   |   |   Diabetes Pedigree Function = high
|   |   |   |   2-hour Serum Insulin = high: no (5/0)
|   |   |   |   2-hour Serum Insulin = low
|   |   |   |   |   Diastolic Blood Pressure = low: yes (1/1)
|   |   |   |   |   Diastolic Blood Pressure = high: no (1/0)
```

B MyDT with CFS

```
Plasma Glucose Concentration = high
|    Body Mass Index = high
|    |    Age = high
|    |    |    Diabetes Pedigree Function = high
|    |    |    |    2-hour Serum Insulin = high: yes (23/9)
|    |    |    |    2-hour Serum Insulin = low: yes (1/0)
|    |    |    Diabetes Pedigree Function = low
|    |    |    |    2-hour Serum Insulin = high: yes (26/22)
|    |    |    |    2-hour Serum Insulin = low: yes (1/0)
|    |    Age = low
|    |    |    2-hour Serum Insulin = high
|    |    |    |    Diabetes Pedigree Function = low: no (18/12)
|    |    |    |    Diabetes Pedigree Function = high: no (10/9)
|    |    |    2-hour Serum Insulin = low: no (1/0)
|    Body Mass Index = low
|    |    2-hour Serum Insulin = high
|    |    |    Diabetes Pedigree Function = high: no (6/0)
|    |    |    Diabetes Pedigree Function = low
|    |    |    |    Age = high: no (8/1)
|    |    |    |    Age = low: no (8/2)
|    |    2-hour Serum Insulin = low
|    |    |    Diabetes Pedigree Function = high
|    |    |    |    Age = low: no (1/0)
|    |    |    |    Age = high: yes (1/0)
|    |    |    Diabetes Pedigree Function = low: no (2/0)
Plasma Glucose Concentration = low
|    Body Mass Index = low: no (66/0)
|    Body Mass Index = high
|    |    2-hour Serum Insulin = high
|    |    |    Age = low
|    |    |    |    Diabetes Pedigree Function = low: no (28/3)
|    |    |    |    Diabetes Pedigree Function = high: no (11/1)
|    |    |    Age = high
|    |    |    |    Diabetes Pedigree Function = low: no (24/4)
|    |    |    |    Diabetes Pedigree Function = high: yes (4/3)
|    |    2-hour Serum Insulin = low
|    |    |    Age = high: no (16/0)
|    |    |    Age = low
|    |    |    |    Diabetes Pedigree Function = low: no (21/1)
|    |    |    |    Diabetes Pedigree Function = high: no (9/1)
Plasma Glucose Concentration = 'very high'
|    2-hour Serum Insulin = high
|    |    Body Mass Index = low
|    |    |    Age = high: yes (9/3)
|    |    |    Age = low
|    |    |    |    Diabetes Pedigree Function = high: no (1/0)
|    |    |    |    Diabetes Pedigree Function = low: no (2/1)
|    |    Body Mass Index = high
|    |    |    Age = high
|    |    |    |    Diabetes Pedigree Function = low: yes (36/7)
|    |    |    |    Diabetes Pedigree Function = high: yes (28/6)
|    |    |    Age = low
```

```
|    |    |    |    Diabetes Pedigree Function = high: yes (12/0)
|    |    |    |    Diabetes Pedigree Function = low: yes (11/3)
|    2-hour Serum Insulin = low
|    |    Diabetes Pedigree Function = low: no (2/0)
|    |    Diabetes Pedigree Function = high: yes (1/0)
Plasma Glucose Concentration = medium
|    Age = high
|    |    Body Mass Index = low
|    |    |    2-hour Serum Insulin = high
|    |    |    |    Diabetes Pedigree Function = low: no (13/2)
|    |    |    |    Diabetes Pedigree Function = high: no (11/1)
|    |    |    2-hour Serum Insulin = low: no (3/0)
|    |    Body Mass Index = high
|    |    |    Diabetes Pedigree Function = low
|    |    |    |    2-hour Serum Insulin = high: no (42/33)
|    |    |    |    2-hour Serum Insulin = low: no (5/0)
|    |    |    Diabetes Pedigree Function = high: yes (27/10)
|    Age = low
|    |    Body Mass Index = high
|    |    |    Diabetes Pedigree Function = high
|    |    |    |    2-hour Serum Insulin = high: no (22/6)
|    |    |    |    2-hour Serum Insulin = low: no (5/1)
|    |    |    Diabetes Pedigree Function = low
|    |    |    |    2-hour Serum Insulin = high: no (51/8)
|    |    |    |    2-hour Serum Insulin = low: no (10/1)
|    |    Body Mass Index = low
|    |    |    Diabetes Pedigree Function = low: no (34/0)
|    |    |    Diabetes Pedigree Function = high
|    |    |    |    2-hour Serum Insulin = high: no (5/0)
|    |    |    |    2-hour Serum Insulin = low: no (2/1)
```

C Weka DT Unpruned without CFS

```
Plasma Glucose Concentration = high
|   Body Mass Index = high
|   |   Triceps Skin Fold Thickness = high
|   |   |   Pregnancies = low
|   |   |   |   Diabetes Pedigree Function = high
|   |   |   |   |   Age = high: yes (16.0/5.0)
|   |   |   |   |   Age = low
|   |   |   |   |   |   Diastolic Blood Pressure = high: yes (11.0/5.0)
|   |   |   |   |   |   Diastolic Blood Pressure = low: no (5.0/2.0)
|   |   |   |   Diabetes Pedigree Function = low
|   |   |   |   |   Diastolic Blood Pressure = high: no (43.0/19.0)
|   |   |   |   |   Diastolic Blood Pressure = low: yes (10.0/4.0)
|   |   |   Pregnancies = high
|   |   |   |   Diastolic Blood Pressure = high: yes (29.0/8.0)
|   |   |   |   Diastolic Blood Pressure = low
|   |   |   |   |   Diabetes Pedigree Function = high: no (2.0)
|   |   |   |   |   Diabetes Pedigree Function = low: yes (3.0)
|   |   Triceps Skin Fold Thickness = low: no (13.0/4.0)
|   Body Mass Index = low: no (29.0/4.0)
Plasma Glucose Concentration = low
```

```
|   Body Mass Index = high
|   |   2-hour Serum Insulin = high
|   |   |   Age = high
|   |   |   |   Diabetes Pedigree Function = high: yes (7.0/3.0)
|   |   |   |   Diabetes Pedigree Function = low: no (28.0/4.0)
|   |   |   Age = low: no (43.0/4.0)
|   |   2-hour Serum Insulin = low: no (48.0/2.0)
|   Body Mass Index = low: no (66.0)
Plasma Glucose Concentration = very high
|   2-hour Serum Insulin = high
|   |   Body Mass Index = high: yes (103.0/16.0)
|   |   Body Mass Index = low
|   |   |   Age = high: yes (12.0/3.0)
|   |   |   Age = low: no (4.0/1.0)
|   2-hour Serum Insulin = low: no (3.0/1.0)
Plasma Glucose Concentration = medium
|   Age = high
|   |   2-hour Serum Insulin = high
|   |   |   Body Mass Index = high
|   |   |   |   Diabetes Pedigree Function = high: yes (37.0/10.0)
|   |   |   |   Diabetes Pedigree Function = low
|   |   |   |   |   Diastolic Blood Pressure = high: no (57.0/24.0)
|   |   |   |   |   Diastolic Blood Pressure = low
|   |   |   |   |   |   Triceps Skin Fold Thickness = high: yes (15.0/7.0)
|   |   |   |   |   |   Triceps Skin Fold Thickness = low: no (3.0/1.0)
|   |   |   Body Mass Index = low: no (27.0/3.0)
|   |   2-hour Serum Insulin = low: no (8.0)
|   Age = low
|   |   Body Mass Index = high
|   |   |   Pregnancies = low
|   |   |   |   Triceps Skin Fold Thickness = high
|   |   |   |   |   Diabetes Pedigree Function = high
|   |   |   |   |   |   Diastolic Blood Pressure = high: no (17.0/2.0)
|   |   |   |   |   |   Diastolic Blood Pressure = low: yes (7.0/3.0)
|   |   |   |   |   Diabetes Pedigree Function = low: no (54.0/8.0)
|   |   |   |   Triceps Skin Fold Thickness = low: no (24.0/1.0)
|   |   |   Pregnancies = high: yes (2.0/1.0)
|   |   Body Mass Index = low: no (42.0/1.0)
```

D Weka DT Unpruned with CFS

```
Plasma Glucose Concentration = high
|   Body Mass Index = high
|   |   Age = high: yes (82.0/31.0)
|   |   Age = low: no (50.0/21.0)
|   Body Mass Index = low: no (29.0/4.0)
Plasma Glucose Concentration = low
|   Body Mass Index = high
|   |   2-hour Serum Insulin = high
|   |   |   Age = high
|   |   |   |   Diabetes Pedigree Function = high: yes (7.0/3.0)
|   |   |   |   Diabetes Pedigree Function = low: no (28.0/4.0)
|   |   |   Age = low: no (43.0/4.0)
```

```
|   |   2-hour Serum Insulin = low: no (48.0/2.0)
|   Body Mass Index = low: no (66.0)
Plasma Glucose Concentration = very high
|   2-hour Serum Insulin = high
|   |   Body Mass Index = high: yes (103.0/16.0)
|   |   Body Mass Index = low
|   |   |   Age = high: yes (12.0/3.0)
|   |   |   Age = low: no (4.0/1.0)
|   2-hour Serum Insulin = low: no (3.0/1.0)
Plasma Glucose Concentration = medium
|   Age = high
|   |   Body Mass Index = high
|   |   |   Diabetes Pedigree Function = high: yes (37.0/10.0)
|   |   |   Diabetes Pedigree Function = low: no (80.0/33.0)
|   |   Body Mass Index = low: no (30.0/3.0)
|   Age = low: no (146.0/17.0)
```

E Weka DT Pruned without CFS

```
Plasma Glucose Concentration = high
|   Body Mass Index = high
|   |   Triceps Skin Fold Thickness = high: yes (119.0/51.0)
|   |   Triceps Skin Fold Thickness = low: no (13.0/4.0)
|   Body Mass Index = low: no (29.0/4.0)
Plasma Glucose Concentration = low: no (192.0/14.0)
Plasma Glucose Concentration = very high: yes (122.0/24.0)
Plasma Glucose Concentration = medium
|   Age = high
|   |   Body Mass Index = high
|   |   |   Diabetes Pedigree Function = high: yes (37.0/10.0)
|   |   |   Diabetes Pedigree Function = low: no (80.0/33.0)
|   |   Body Mass Index = low: no (30.0/3.0)
|   Age = low: no (146.0/17.0)
```

F Weka DT Pruned with CFS

```
Plasma Glucose Concentration = high
|   Body Mass Index = high
|   |   Age = high: yes (82.0/31.0)
|   |   Age = low: no (50.0/21.0)
|   Body Mass Index = low: no (29.0/4.0)
Plasma Glucose Concentration = low: no (192.0/14.0)
Plasma Glucose Concentration = very high: yes (122.0/24.0)
Plasma Glucose Concentration = medium
|   Age = high
|   |   Body Mass Index = high
|   |   |   Diabetes Pedigree Function = high: yes (37.0/10.0)
|   |   |   Diabetes Pedigree Function = low: no (80.0/33.0)
|   |   Body Mass Index = low: no (30.0/3.0)
|   Age = low: no (146.0/17.0)
```