Predicting Diabetes in the Pima Indians: An Investigation into Classification Strategies

Group 23: 490424010, 490390494

May 14, 2021

Contents

List of Figures

List of Tables

1 Introduction

1.1 Aim

The aim of this study is to investigate methods for predicting the future onset of diabetes mellitus (or simply diabetes), with relevance to females over 21 and of Pima Indian heritage.

1.2 Relevance

This study is important because the classifiers created have the potential to forewarn individuals of their risk to diabetes, and to be used as an easy clinical tool for early prevention. This is particularly important because, if left untreated, diabetes can lead to many serious long-term health implications such as cardiovascular disease, stroke, diabetic ketoacidosis and even death [2].

2 Data

The dataset used throughout this paper originates from the National Institute of Diabetes and Digestive and Kidney Diseases and was first used in a demonstration of the ADAP Learning Algorithm in 1988 [12]. It consists of 768 non-diabetic females aged at least 21 years old and of Pima Indian heritage. There are 9 columns per row, the first 8 of which are biometric measurement attributes whilst the final one is the class consisting of whether or not the individual with be diagnosed with diabetes. A description of each column in the dataset is shown in Table 1. To maintain consistency the dataset has been cleaned to remove any missing values.

Table 1: A synopsis of the dataset's columns with those selected by CFS highlighted.

Description	**Units**
Number of times pregnant	n/a
Plasma glucose concentration at 2 hours in an oral glucose tolerance test	mg/dL
Diastolic blood pressure	mm Hg
Triceps skin fold thickness	mm
Serum insulin level	μU/mL
Body mass index (BMI)	kg/m^2
Diabetes pedigree function (likelihood of diabetes based on family history)	n/a
Age	years
Is diabetes diagnosed between 1 and 5 years after the above measurements are recorded?	n/a

2.1 Attribute Selection

The Correlation-based Feature Selection (CFS) method is a way of determining a representative set of attributes which are highly correlated with the class but uncorrelated with each other. This can improve the training of a classification model by removing features that are not predictive of the class.

Using the CFS algorithm [7] implemented in Weka 3.8.5 [5], the attributes that were selected are plasma glucose concentration, serum insulin level, BMI, diabetes pedigree function and age, and are additionally highlighted in Table 1.

3 Results and Discussion

3.1 Classifier Accuracy

The canonical Naïve Bayes (NB) and Decision Tree (DT) classification algorithms were implemented with tie decisions resulting in a 'yes' and are hereafter referred to as MyNB and MyDT respectively. 10-fold stratified cross validation was then performed on these algorithms and 12 other inbuilt Weka algorithms using the dataset described in section 2 after normalisation and discretisation for the numeric and nominal classification algorithms respectively.

Tables 2 and 3 present all the resulting accuracy figures for each tested classification algorithm, shown in percentage (%) to 4 d.p., using both the full dataset and the dataset after CFS, and coloured for ease of comparison.

Table 2: The 10-fold stratified cross validation accuracy in percentage (%) of each tested *numeric* classification algorithm using the dataset with and without CFS.

Numeric Data	ZeroR	1R	1NN	5NN	NB	MLP	SVM	MyNB
No feature selection	65.1042	70.8333	67.8385	74.4792	75.1302	75.3906	76.3021	75.2614
CFS	65.1042	70.8333	69.0104	74.4792	76.3021	75.7813	76.6927	76.0407

Table 3: The 10-fold stratified cross validation accuracy in percentage (%) of each tested *nominal* classification algorithm using the dataset with and without CFS.

Nominal Data	DT unpruned	DT pruned	MyDT	Bagg	Boost	RF
No feature selection	75.0000	75.3906	73.4484	74.8698	76.1719	73.1771
CFS	79.4271	79.4271	78.3869	78.5156	78.6458	78.9063

3.2 DT Diagrams

Decision trees were built on the full discretised dataset using three different algorithms: MyDT, and two DT classifiers from Weka (DT unpruned and DT pruned). The MyDT tree was built using the ID3 algorithm (without pruning), which recursively builds a tree based on maximum information gain. The two Weka variants were built using J48 (an implementation of the C4.5 algorithm) with default parameters, but differ in that one has been pruned in addition to the other [5]. The DT diagrams are displayed in Figures 2, 3 and 4 in section 6.

3.3 Discussion

3.3.1 Comparison of Classifiers

Overall, the accuracy of the 14 classifiers ranged roughly between 65% and 80% with a mean of ~74.5%.

The best performing numeric classifier was the SVM, both with and without feature selection, where it achieved an accuracy of ~76.7% and ~76.3% respectively. Similar in performance were MyNB, NB and MLP, with accuracies roughly within 1% of the SVM. This small difference in accuracies ranging

from 75% to 77% is not necessarily indicative of algorithmic superiority but may be the effect of random noise in the testing dataset.

On the other hand, the worst performing numeric classifiers were ZeroR, 1R and 1NN, achieving accuracies between 65% and 71%. These simple algorithms are clearly not complex enough to capture patterns in the data, but are instead good points for comparison as to what is easily achievable (for example by predicting the majority class in ZeroR).

Within the nominal classifiers, the highest accuracy was ∼79.4%, and was obtained by both the pruned and unpruned DT using feature selection. Despite this, all of the nominal classifiers performed well using feature selection, with accuracies ranging roughly between 78% and 79.5%. Without feature selection, the best performing nominal classifier was Boost with an accuracy of ∼76.2%.

The worst performing nominal classifier was MyDT, with and without feature selection, where it achieved an accuracy of ∼78.3% and ∼73.4% respectively.

The 6 nominal classifiers clearly performed much better than the 8 numeric ones with a mean accuracy of ∼76.8% compared to ∼72.8%. In addition, using CFS improved or equalled the performance of every classifier, with an average improvement in accuracy of ∼2.1%.

The implementations of MyNB and Weka's NB only differ in terms of their running time performance. In fact, the minimal differences in accuracies evident in Table 2 are most likely the result of different 10-fold data stratifications used in the cross validation accuracy calculations. On the other hand, the implementations of MyDT and Weka's two DTs differ profoundly. MyDT is built using the ID3 algorithm without pruning, whilst Weka uses J48 (an implementation of the 8th revision of the C4.5 algorithm [5]) which is very similar to ID3 but using the normalised information gain ratio as its splitting criterion. This resulted in Weka's two DTs performing better than MyDT with and without feature selection.

3.3.2 Feature Selection

The feature selection method (CFS) selected a subset of 5 features from the original 9. As is highlighted in Table 1, these were, in no particular order:

- Plasma glucose concentration
- Serum insulin level
- BMI
- Diabetes pedigree function
- Age

It makes sense that this selected subset is highly correlated with the onset of diabetes but also mutually uncorrelated.

In fact, having diabetes is defined for the dataset in question as obtaining a plasma glucose concentration of at least 200 mg/dL, 2 hours after the ingestion of 75mg of carbohydrate solution [12]. So it is no surprise that CFS claims that glucose concentration is a strong predictor of the onset of diabetes. Furthermore, the level of insulin has been found to be one of the strongest predictors of the onset of diabetes in numerous studies containing a wide range of attributes [3, 1, 9, 8]. Although glucose concentration and insulin levels often exhibit a significant inversely-correlated relationship when measured via an oral glucose tolerance test (OGTT) and insulin release test (IRT) respectively [14], the original dataset description is vague as to how the insulin level was recorded. In fact, Figure 1 shows that all of the 5 selected attributes in the CFS subset are mutually uncorrelated, as expected.

Furthermore, BMI and other weight-related features have also previously been found to be good predictors of the onset of diabetes [3, 9, 8]. Specifically within our dataset, the weight-related features

are BMI and triceps skin fold thickness, which are known to be significantly correlated [4]. However, BMI has a higher association with health-related risk factors than triceps skin fold thickness [11, 6], which explains why it was prioritised in the CFS subset.

The two other features selected by CFS are also known to have a high correlation with diabetes. In particular, age has been shown to be both a strong predictor [13] and also relatively uncorrelated with other features, as demonstrated by its inclusion in principal component analysis (PCA) in a different paper using the same dataset [9]. Family history of diabetes has also been shown to be a relatively strong and independent predictor of diabetes [10].

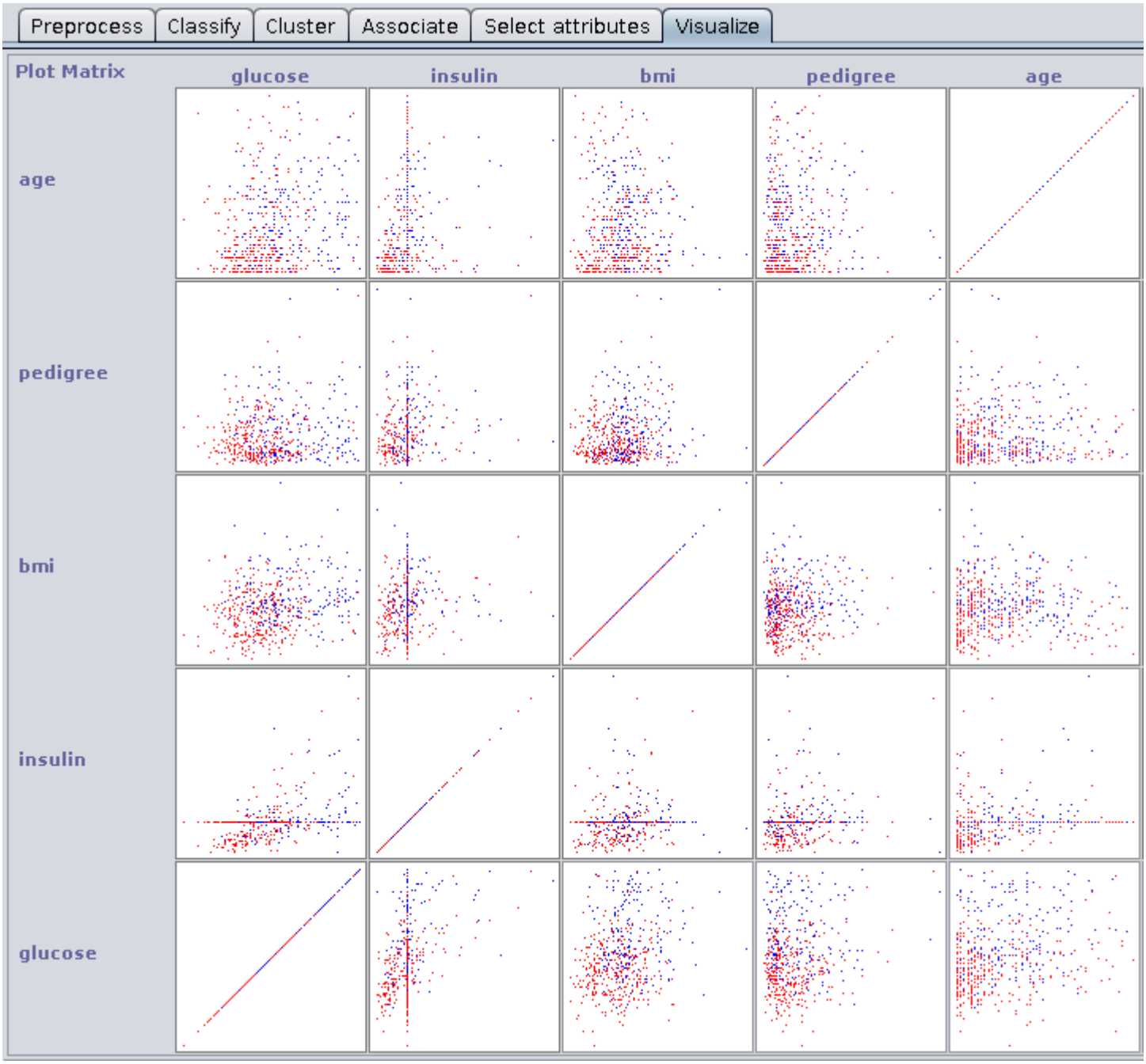

Figure 1: Scatter plot matrix using Weka with each variable from CFS.

Across all models, the accuracy either improved or stayed the same after using CFS compared to no feature selection, with an average improvement in accuracy of ∼2.1%. This effect was most apparent with tree-based models, with the accuracy improving by ∼5% for DT pruned and unpruned, MyDT and RF. This is likely due to a reduction in over fitting, stemming from the removal of features that did not add much information or were otherwise terrible predictors of the onset of diabetes.

Another advantage of CFS is the reduction in memory size and computational time. This is espe-

cially noticeable when pruning large datasets that have many features which are highly correlated with each other or poorly correlated with the target classification. In our dataset for example, the number of features was reduced by over 40%. Not only does this allow for smaller file sizes and therefore faster training and testing, but it can also increase the interpretability of models as there are less features being used when visualising decision trees.

3.3.3 Decision Trees

Although all three decision tree algorithms (MyDT, DT unpruned and DT pruned) differ in both the method used and 10-fold accuracy, each algorithm uses a splitting method that tends to reduce the information entropy of the partitioned data. As a result, we see strong similarities in the first couple splits of each tree. In particular, the diagrams in section 6 reveal that each tree found glucose to be the best feature to first split on, as its high correlation with diabetes caused it to produce in the largest reduction in information entropy compared to other features. The second splits are also quite similar, with each tree subsequently splitting on either BMI, insulin, or age, depending on the value of the parent glucose attribute.

After a few splits however, slight differences in the algorithms start to become apparent in the tree diagrams. For one, Weka's J48 DT with pruning (Figure 4) reaches leaf nodes on the third and fourth split, whereas Weka's unpruned J48 tree (Figure 3) generally continues until split six or seven. In comparison, our unpruned ID3 implementation (Figure 2) reaches 8 splits in some cases. Interestingly, the depth of the tree seems to be directly negatively correlated with the relative performance of the algorithm. This is likely due to the presence of over-fitting that occurs when you begin splitting on smaller and smaller partitions of the original data, causing the tree to start splitting on noise rather than patterns.

One method to prevent such over-fitting, as implemented by Weka's best performing J48 tree, is the powerful concept of pruning. This strategy involves iteratively removing leaf nodes from a tree by replacing the sub-tree from which a family of leaves stem with a new leaf node, that then classifies as the majority of class of that subtree. At each step, the new validation accuracy of the resulting tree is compared and the process is repeated until all subtree reductions result in a loss of accuracy. As clear with the superior performance of the pruned J48 tree, such a procedure can allow the size of a tree to be significantly reduced whilst simultaneously increasing overall accuracy.

3.3.4 Tree-based Ensemble Classifiers

Although tree-based classifiers exist for numeric features, our study focuses on the use of tree-based classifiers for exclusively nominal data. Out of the six tree-based classifiers tested, three are variations of single decision trees (discussed in section 3.3.3), and the other three use ensemble methods. Ensemble methods involve the creation of a number of single classifiers, where each classifier makes a prediction, and the results are combined similar to a voting mechanism. As we are combining multiple predictions from different classifiers, ideally, the mistakes made by one classifier will be outvoted by the predictions from other classifiers which are correct on those particular samples. In the case of our study, the ensemble methods tested were bagging, boosting, and random forest.

Bagging involves 'bootstrapping' a number of new datasets by sampling from the original data with replacement, and then distributing the bootstrapped datasets among a number of single classifiers for training. Predictions on new data are then calculated by combining the predictions from each classifier through a majority vote with equal weighting. While bagging did have some improvements over MyDT and RF without feature selection, this particular ensemble method did not prove particularly beneficial in predicting diabetes over single J48 trees.

Boosting was the second ensemble method used in our study. This involves creating a series of classifiers, where the data of the next classifier is weighted towards the misclassified examples in the previous classifier. Unlike bagging, individuals trees are then given different weighting in the final vote for new predictions, typically depending on their overall performance on the training data. In our study, boosting performed particularly well on the dataset without feature selection, as it beat the second best tree-based classifier by almost 1%. Interestingly, this advantage diminished after CFS was applied.

The last ensemble method used was Random Forest. In contrast to boosting, this method was the worst tree-based method without feature selection, but outperformed both bagging and boosting after CFS was applied. Similar to bagging, this algorithm generates multiple new datasets which are then distributed among single trees for training and majority-vote predictions subsequently. The key difference being that random forests use a random subset of the original features, rather than sampling from individual examples.

4 Conclusion

The main findings of our results include that CFS is hugely beneficial for all non-trivial classifiers but especially for tree-based classifiers. CFS also highlighted the attributes with the best predictive power of the onset of diabetes, including age which is quite interesting. In fact, younger Pima Indian females were less likely to develop diabetes given a medium glucose level according to the pruned decision tree in Figure 4 than older females.

Furthermore, nominal classifiers performed significantly better than numeric classifiers on this dataset with a $\sim$4% higher mean accuracy. This was interesting as raw numeric data should in theory contain more useful or accessible information than equivalent discrete date, yet the model performance suggests otherwise.

Future work to be done includes using more complex and powerful classification methods, such as deep learning, to try and develop a more accurate classifier on this dataset. As well as, investigating the same 14 classifiers but on a completely new and unrelated dataset in order to draw more well-rounded conclusions. Furthermore, we would like to investigate why some ensemble methods (particularly RF and Bagging) performed worse than single tree methods. This subverted our expectations that ensembles should often perform equally well or better than the single tree methods, and so we believe it deserves further exploration.

5 Reflection

One important thing that we have learned throughout this assignment is the power of feature selection, namely CFS, in increasing the performance of classifiers whether numeric or nominal. We also learned a lot about Latex formatting and working in a group using GitHub. We also discovered how interesting ensemble methods can be such as boosting and bagging.

However, overall we believe the best part about this assignment was learning how to use the powerful tool Weka. At first we were put off as the program seemed old and had a confusing user interface, but after getting the hang of it we began to realise its true potential as a tool for quickly processing data and then training powerful ML models.

References

[1] Abdul-Ghani, M. A., Williams, K., DeFronzo, R. A., and Stern, M. What is the best predictor of future type 2 diabetes? *Diabetes Care 30*, 6 (2007), 1544–1548.

[2] AE, K., GE, U., JM, M., and JN, F. Hyperglycemic crises in adult patients with diabetes. *Diabetes Care 32*, 7 (Jul 2009), 1335–1343.

[3] Balkau, B., Lange, C., Fezeu, L., Tichet, J., de Lauzon-Guillain, B., Czernichow, S., Fumeron, F., Froguel, P., Vaxillaire, M., Cauchi, S., Ducimetière, P., and Eschwège, E. Predicting diabetes: clinical, biological, and genetic approaches: data from the epidemiological study on the insulin resistance syndrome (desir). *Diabetes Care 31* (2008), 2056–61.

[4] Chavhan, S., and Chandrachood, M. Correlation of body mass index with biceps and triceps skin fold thickness. *International Journal Of Community Medicine And Public Health 7*, 4 (2020), 1475–1479.

[5] Frank, E., Hall, M. A., and Witten, I. H. *Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, 4 ed. Morgan Kaufmann, 2016.

[6] Freedman, D. S., Katzmarzyk, P. T., Dietz, W. H., Srinivasan, S. R., and Berenson, G. S. Relation of body mass index and skinfold thicknesses to cardiovascular disease risk factors in children: the bogalusa heart study. *The American journal of clinical nutrition 90* (2009), 210–6.

[7] Hall, M. A. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, The University of Waikato, Apr 1999.

[8] Lai, H., Huang, H., Keshavjee, K., Guergachi, A., and Gao, X. Predictive models for diabetes mellitus using machine learning techniques.

[9] Mahboob Alam, T., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Imtiaz Baig, T., Hussain, A., Malik, M. A., Raza, M. M., Ibrar, S., and Abbas, Z. A model for early prediction of diabetes. *Informatics in Medicine Unlocked 16* (2019), 100204.

[10] Scott, R. A., Langenberg, C., and Sharp, S. J. The link between family history and risk of type 2 diabetes is not explained by anthropometric, lifestyle or genetic risk factors: the epic-interact study. *Diabetologia 56* (2013), 60–9.

[11] Sivrikaya, K., Ziyagil, M., and Çebï, M. Relationship between body mass index and skinfold thickness in exercised and sedentary boys and girls. *Universal Journal of Educational Research 7* (01 2019), 48–54.

[12] Smith, J., Everhart, J., Dickson, W., Knowler, W., and Johannes, R. Using the adap learning algorithm to forcast the onset of diabetes mellitus. *Proceedings - Annual Symposium on Computer Applications in Medical Care 10* (Nov 1988).

[13] Vuvor, F., and Egbi, G. Correlation of diabetes mellitus and body weight of adults above the age of 30 years in a medical facility in ghana. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews 11* (2017), S407–S409. SI: Online Supplement - 1.

[14] Xu, S.-h., Jin, W.-s., and Lin, Y.-d. Relationship between plasma glucose level and insulin secretion in type 2 diabetic patients. 859–62.

Nomenclature

μU/mL Micro enzyme units per millilitre

1R One Rule

Bagg Bagging

Boost Boosting

CFS Correlation-based feature selection

d.p. decimal points

DT Decision Tree

IRT Insulin release test

kg/m^2 Weight in kilograms per height in metres squared

kNN k-Nearest Neighbours

mg/dL Milligrams per decilitre

MLP Multilayer Perceptron

mm Millimetres

mm Hg Millimetres of mercury

n/a Not applicable

NB Naïve Bayes

OGTT Oral glucose tolerance test

PCA Principal component analysis

RF Random Forest

SVM Support Vector Machines

ZeroR Classifier that always predicts the majority class

6 Appendix

```
glucose = very high
| insulin = high
| | bmi = high
| | | npreg = high
| | | | pedigree = high: yes (16.0/0.0)
| | | | pedigree = low
| | | | | blood = high: yes (12.0/3.0)
| | | | | blood = low: yes (3.0/1.0)
| | | npreg = low
| | | | age = high
| | | | | pedigree = high
| | | | | | triceps = high
| | | | | | | blood = high: yes (10.0/5.0)
| | | | | | | blood = low: yes (1.0/1.0)
| | | | | | triceps = low: yes (1.0/0.0)
| | | | | pedigree = low
| | | | | | blood = high
| | | | | | | triceps = high: yes (14.0/2.0)
| | | | | | | triceps = low: yes (3.0/0.0)
| | | | | | blood = low
| | | | | | | triceps = high: yes (3.0/0.0)
| | | | | | | triceps = low: yes (1.0/1.0)
| | | | age = low
| | | | | pedigree = high: yes (12.0/0.0)
| | | | | pedigree = low
| | | | | | triceps = high
| | | | | | | blood = high: yes (7.0/2.0)
| | | | | | | blood = low: yes (3.0/0.0)
| | | | | | triceps = low
| | | | | | | blood = high: yes (1.0/0.0)
| | | | | | | blood = low: no (1.0/0.0)
| | bmi = low
| | | age = high
| | | | triceps = high
| | | | | npreg = high
| | | | | | pedigree = high
| | | | | | | blood = high: yes (1.0/1.0)
| | | | | | | blood = low: yes (1.0/0.0)
| | | | | | pedigree = low: yes (2.0/0.0)
| | | | | npreg = low
| | | | | | blood = high: yes (1.0/1.0)
| | | | | | blood = low: yes (1.0/1.0)
| | | | triceps = low: yes (3.0/0.0)
| | | age = low
| | | | blood = high
| | | | | triceps = high: no (1.0/0.0)
| | | | | triceps = low: yes (1.0/1.0)
| | | | blood = low: no (1.0/0.0)
| insulin = low
| | pedigree = high: yes (1.0/0.0)
| | pedigree = low: no (2.0/0.0)
```

```
glucose = high
| bmi = high
| | age = high
| | | pedigree = high
| | | | blood = high
| | | | | npreg = high: yes (12.0/2.0)
| | | | | npreg = low
| | | | | | triceps = high
| | | | | | | insulin = high: yes (9.0/4.0)
| | | | | | | insulin = low: yes (1.0/0.0)
| | | | | | triceps = low: yes (1.0/0.0)
| | | | blood = low
| | | | | npreg = high: no (2.0/0.0)
| | | | | npreg = low: yes (1.0/1.0)
| | | pedigree = low
| | | | insulin = high
| | | | | triceps = high
| | | | | | blood = high
| | | | | | | npreg = high: yes (9.0/6.0)
| | | | | | | npreg = low: no (12.0/11.0)
| | | | | | blood = low
| | | | | | | npreg = high: yes (3.0/0.0)
| | | | | | | npreg = low: yes (2.0/1.0)
| | | | | triceps = low
| | | | | | npreg = high: no (1.0/0.0)
| | | | | | npreg = low
| | | | | | | blood = high: yes (1.0/1.0)
| | | | | | | blood = low: no (1.0/0.0)
| | | | insulin = low: yes (1.0/0.0)
| | age = low
| | | triceps = high
| | | | pedigree = high
| | | | | blood = high: yes (6.0/5.0)
| | | | | blood = low: no (3.0/2.0)
| | | | pedigree = low
| | | | | blood = high: no (12.0/8.0)
| | | | | blood = low: yes (4.0/3.0)
| | | triceps = low
| | | | blood = high: no (3.0/0.0)
| | | | blood = low
| | | | | insulin = high
| | | | | | pedigree = high: yes (1.0/1.0)
| | | | | | pedigree = low: no (1.0/0.0)
| | | | | insulin = low: no (1.0/0.0)
| bmi = low
| | triceps = high
| | | insulin = high
| | | | pedigree = high: no (5.0/0.0)
| | | | pedigree = low
| | | | | age = high
| | | | | | blood = high: no (5.0/1.0)
| | | | | | blood = low: no (2.0/0.0)
| | | | | age = low
```

| | | | | | blood = high: no (1.0/0.0)
| | | | | | blood = low: yes (2.0/2.0)
| | | insulin = low
| | | | pedigree = high: yes (1.0/0.0)
| | | | pedigree = low: no (1.0/0.0)
| | triceps = low: no (9.0/0.0)
glucose = medium
| age = high
| | bmi = high
| | | pedigree = high
| | | | npreg = high: yes (13.0/0.0)
| | | | npreg = low
| | | | | triceps = high
| | | | | | blood = high: yes (9.0/7.0)
| | | | | | blood = low: yes (3.0/3.0)
| | | | | triceps = low: yes (2.0/0.0)
| | | pedigree = low
| | | | insulin = high
| | | | | blood = high
| | | | | | npreg = high: no (14.0/12.0)
| | | | | | npreg = low
| | | | | | | triceps = high: no (18.0/11.0)
| | | | | | | triceps = low: yes (1.0/1.0)
| | | | | blood = low
| | | | | | triceps = high
| | | | | | | npreg = high: yes (3.0/3.0)
| | | | | | | npreg = low: yes (5.0/4.0)
| | | | | | triceps = low: no (2.0/1.0)
| | | | insulin = low: no (5.0/0.0)
| | bmi = low
| | | blood = high
| | | | npreg = high: no (13.0/0.0)
| | | | npreg = low
| | | | | pedigree = high: no (3.0/0.0)
| | | | | pedigree = low
| | | | | | triceps = high: no (2.0/1.0)
| | | | | | triceps = low: no (2.0/0.0)
| | | blood = low
| | | | npreg = high: yes (1.0/0.0)
| | | | npreg = low
| | | | | triceps = high: no (5.0/0.0)
| | | | | triceps = low: no (2.0/1.0)
| age = low
| | bmi = high
| | | triceps = high
| | | | npreg = high: yes (1.0/1.0)
| | | | npreg = low
| | | | | pedigree = high
| | | | | | blood = high
| | | | | | | insulin = high: no (12.0/2.0)
| | | | | | | insulin = low: no (3.0/0.0)
| | | | | | blood = low
| | | | | | | insulin = high: yes (3.0/3.0)

| | | | | | | insulin = low: yes (1.0/0.0)
| | | | | pedigree = low
| | | | | | blood = high
| | | | | | | insulin = high: no (20.0/5.0)
| | | | | | | insulin = low: no (3.0/0.0)
| | | | | | blood = low
| | | | | | | insulin = high: no (18.0/2.0)
| | | | | | | insulin = low: no (5.0/1.0)
| | | triceps = low
| | | | pedigree = high
| | | | | blood = high: no (4.0/0.0)
| | | | | blood = low
| | | | | | insulin = high: no (3.0/1.0)
| | | | | | insulin = low: no (2.0/0.0)
| | | | pedigree = low: no (14.0/0.0)
| | bmi = low
| | | pedigree = high
| | | | insulin = high: no (5.0/0.0)
| | | | insulin = low
| | | | | blood = high: no (1.0/0.0)
| | | | | blood = low: yes (1.0/1.0)
| | | pedigree = low: no (34.0/0.0)
glucose = low
| bmi = high
| | insulin = high
| | | age = high
| | | | pedigree = high
| | | | | blood = high
| | | | | | npreg = high
| | | | | | | triceps = high: yes (2.0/1.0)
| | | | | | | triceps = low: no (1.0/0.0)
| | | | | | npreg = low
| | | | | | | triceps = high: no (1.0/0.0)
| | | | | | | triceps = low: yes (1.0/0.0)
| | | | | blood = low: yes (1.0/0.0)
| | | | pedigree = low
| | | | | triceps = high
| | | | | | npreg = high
| | | | | | | blood = high: no (8.0/1.0)
| | | | | | | blood = low: no (1.0/0.0)
| | | | | | npreg = low
| | | | | | | blood = high: no (11.0/2.0)
| | | | | | | blood = low: no (3.0/1.0)
| | | | | triceps = low: no (1.0/0.0)
| | | age = low
| | | | blood = high: no (18.0/0.0)
| | | | blood = low
| | | | | triceps = high
| | | | | | pedigree = high: no (5.0/1.0)
| | | | | | pedigree = low: no (9.0/3.0)
| | | | | triceps = low: no (7.0/0.0)
| | insulin = low
| | | blood = high

```
| | | | age = high: no (12.0/0.0)
| | | | age = low
| | | | | triceps = high
| | | | | | pedigree = high: yes (1.0/0.0)
| | | | | | pedigree = low: no (5.0/1.0)
| | | | | triceps = low: no (6.0/0.0)
| | | blood = low: no (23.0/0.0)
| bmi = low: no (66.0/0.0)
```

Figure 2: The DT diagram of MyDT trained on the full discretised dataset.

```
glucose = high
| bmi = high
| | triceps = high
| | | npreg = low
| | | | pedigree = high
| | | | | age = high: yes (16.0/5.0)
| | | | | age = low
| | | | | | blood = high: yes (11.0/5.0)
| | | | | | blood = low: no (5.0/2.0)
| | | | pedigree = low
| | | | | blood = high: no (43.0/19.0)
| | | | | blood = low: yes (10.0/4.0)
| | | npreg = high
| | | | blood = high: yes (29.0/8.0)
| | | | blood = low
| | | | | pedigree = high: no (2.0)
| | | | | pedigree = low: yes (3.0)
| | triceps = low: no (13.0/4.0)
| bmi = low: no (29.0/4.0)
glucose = low
| bmi = high
| | insulin = high
| | | age = high
| | | | pedigree = high: yes (7.0/3.0)
| | | | pedigree = low: no (28.0/4.0)
| | | age = low: no (43.0/4.0)
| | insulin = low: no (48.0/2.0)
| bmi = low: no (66.0)
glucose = very high
| insulin = high
| | bmi = high: yes (103.0/16.0)
| | bmi = low
| | | age = high: yes (12.0/3.0)
| | | age = low: no (4.0/1.0)
| insulin = low: no (3.0/1.0)
glucose = medium
| age = high
| | insulin = high
| | | bmi = high
| | | | pedigree = high: yes (37.0/10.0)
| | | | pedigree = low
| | | | | blood = high: no (57.0/24.0)
```

```
| | | | | blood = low
| | | | | | triceps = high: yes (15.0/7.0)
| | | | | | triceps = low: no (3.0/1.0)
| | | bmi = low: no (27.0/3.0)
| | insulin = low: no (8.0)
| age = low
| | bmi = high
| | | npreg = low
| | | | triceps = high
| | | | | pedigree = high
| | | | | | blood = high: no (17.0/2.0)
| | | | | | blood = low: yes (7.0/3.0)
| | | | | pedigree = low: no (54.0/8.0)
| | | | triceps = low: no (24.0/1.0)
| | | npreg = high: yes (2.0/1.0)
| | bmi = low: no (42.0/1.0)
```

Figure 3: The DT diagram of the Weka J48 algorithm *without* pruning and trained on the full discretised dataset.

```
glucose = high
| bmi = high
| | triceps = high: yes (119.0/51.0)
| | triceps = low: no (13.0/4.0)
| bmi = low: no (29.0/4.0)
glucose = low: no (192.0/14.0)
glucose = very high: yes (122.0/24.0)
glucose = medium
| age = high
| | bmi = high
| | | pedigree = high: yes (37.0/10.0)
| | | pedigree = low: no (80.0/33.0)
| | bmi = low: no (30.0/3.0)
| age = low: no (146.0/17.0)
```

Figure 4: The DT diagram of the Weka J48 algorithm *with* pruning and trained on the full discretised dataset.