THE UNIVERSITY OF SYDNEY

COMP3608 Introduction to Artificial Intelligence (Adv)

Assignment 2: Classification

Name	**SID**
Aditya Pillai	490428362
Raymond Huang	490407806

Table of Contents

1 Introduction

1.1 Aim of Study

This study aims to assess the performance of classifiers that predict the presence of diabetes for individuals of Pima Indian heritage, with a focus on the effect of Correlation-based Feature Selection.

1.2 Significance of Study

1.2.1 Diabetes

Diabetes is a chronic disorder in which a person has high blood sugar, either because the body does not produce enough insulin, or because cells do not respond adequately to the insulin that is produced (Shaw & Tanamas, 2012). There are two types of diabetes and with no cure for the disorder, “the condition requires lifelong management” (Shaw & Tanamas, 2012). Consequently, there are a range of health complications that are associated with diabetes affecting the feet, eyes, kidneys, and cardiovascular health of an individual that compromises their quality of life.

1.2.2 Relevance to Australia

As a priority health issue, early detection and identification of diabetes is crucial. In Australia, there are approximately one million people diagnosed with diabetes, with a growing number of children and adolescents with type 2 diabetes (Shaw & Tanamas, 2012). Indigenous Australians particularly are at a significantly higher risk, and are three times more likely to have type 2 diabetes (Shaw & Tanamas, 2012). Hence, by conducting this study, we hope to assess the performance of diabetes prediction classifiers and deliver improved health outcomes for Indigenous Australians.

2 Data

2.1 Dataset

This dataset has been obtained from the National Institute of Diabetes and Digestive and Kidney Diseases (1990). The Pima Indian Diabetes dataset is composed of 768 patient records with 8 numeric attributes per record. All patients in this dataset were females of at least 21 years of age and of Pima Indian heritage. These attributes are:

- Number of times pregnant
- Plasma glucose concentration in an oral glucose tolerance test
- Diastolic blood pressure (mm Hg)
- Tricep skinfold thickness (mm)
- 2-Hour serum insulin (mu U/ml)
- Body mass index (weight in kg/height in m^2)
- Diabetes pedigree function
- Age (years)

2.2 Attribute Selection

2.2.1 Correlation-Based Feature Selection

Correlation-based feature (CFS) selection was used to select a subset of the original attributes that are highly correlated with the class, reducing the number of predictors required and dimensionality of the dataset. This technique employs a heuristic for "evaluating the merit of a subset of features" based on the hypothesis that good feature subsets contain features that are highly correlated with the class (Hall & Smith, 1997). The following equation (Ghiselli, 1964) formalises the heuristic:

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

where $Merit_s$is the heuristic "merit" of a feature subset S containing k features, $\bar{r_{cf}}$ the average feature class correlation, and $\bar{r_{ff}}$ the average feature-feature intercorrelation (Hall, n.d.). Interpreting this equation can be considered in two parts, where the numerator indicates how predictive a group of features are and the denominator denotes how redundant they are. Best first search was employed to search the feature subset space, starting with an empty set of features and generating all possible single feature expansions. The subset with the highest evaluation is chosen and expanded in the same manner adding single features, with this process iterating until the best subset is found (Hall, 1999).

2.2.2 Selected Attributes

Performing CFS on the Pima Indian Diabetes dataset selected the following attributes for both numeric and nominal data:

- Plasma glucose concentration in an oral glucose tolerance test
- 2-Hour serum insulin (mu U/ml)
- Body mass index (weight in kg/(height in m)^2)
- Diabetes pedigree function
- Age (years)

3 Results and Discussion

3.1 Classifier Accuracy Results

The following tables describe the accuracy results for Weka's classifiers where MyNB and MyDT are our implementation of the Naive Bayes and Decision Tree algorithms, respectively. All values are reported as percentages.

3.1.1 Numeric Data

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	75.2495
CFS	65.1042	70.8333	69.0104	74.4792	76.3021	75.7813	76.6927	76.5550

3.1.2 Nominal Data

Nominal Data	DT Unpruned	DT Pruned	MyDT	Bagging	Boosting	RF
No feature selection	75.0000	75.3906	73.6945	74.8698	76.1719	73.1771
CFS	79.4271	79.4271	77.7290	78.5156	78.6458	78.9063

3.2 Decision Tree Diagrams

In the decision tree diagrams the following encoding scheme was used:

No.	**Attribute**
1	Number of times pregnant
2	Plasma glucose concentration in an oral glucose tolerance test
3	Diastolic blood pressure (mm Hg)
4	Tricep skinfold 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)

Note:
In only the MyDT diagram with CFS (3.2.1), the following encoding scheme was used:

No.	Attribute
1	Plasma glucose concentration in an oral glucose tolerance test
2	2-Hour serum insulin (mu U/ml)
3	Body mass index (weight in kg/(height in m)^2)
4	Diabetes pedigree function
5	Age (years)

3.2.1 MyDT

No Feature Selection	CFS
2 = low | 6 = low: no | 6 = high | | 5 = high | | | 8 = low | | | | 3 = low | | | | | 4 = low: no | | | | | 4 = high | | | | | | 7 = low: no | | | | | | 7 = high: no | | | | 3 = high: no | | | 8 = high | | | | 7 = low | | | | | 4 = high | | | | | | 1 = high | | | | | | | 3 = high: no | | | | | | | 3 = low: no | | | | | | 1 = low | | | | | | | 3 = high: no | | | | | | | 3 = low: no | | | | | 4 = low: no | | | | 7 = high | | | | | 3 = high | | | | | | 1 = low | | | | | | | 4 = low: yes | | | | | | | 4 = high: no | | | | | | 1 = high | | | | | | | 4 = high: yes | | | | | | | 4 = low: no | | | | | 3 = low: yes | | 5 = low | | | 3 = high | | | | 8 = high: no | | | | 8 = low | | | | | 4 = high | | | | | | 7 = low: no | | | | | | 7 = high: yes | | | | | 4 = low: no | | | 3 = low: no	1 = low | 3 = low: no | 3 = high | | 2 = high | | | 5 = low | | | | 4 = low: no | | | | 4 = high: no | | | 5 = high | | | | 4 = low: no | | | | 4 = high: yes | | 2 = low | | | 5 = high: no | | | 5 = low | | | | 4 = low: no | | | | 4 = high: no

```
2 = medium
|  8 = high
|  |  6 = low
|  |  |  3 = high
|  |  |  |  1 = low
|  |  |  |  |  7 = low
|  |  |  |  |  |  4 = high: no
|  |  |  |  |  |  4 = low: no
|  |  |  |  |  7 = high: no
|  |  |  |  1 = high: no
|  |  |  3 = low
|  |  |  |  1 = low
|  |  |  |  |  4 = high: no
|  |  |  |  |  4 = low: no
|  |  |  |  1 = high: yes
|  |  6 = high
|  |  |  7 = low
|  |  |  |  5 = high
|  |  |  |  |  3 = high
|  |  |  |  |  |  1 = high: no
|  |  |  |  |  |  1 = low
|  |  |  |  |  |  |  4 = high: no
|  |  |  |  |  |  |  4 = low: yes
|  |  |  |  |  3 = low
|  |  |  |  |  |  4 = high
|  |  |  |  |  |  |  1 = high: yes
|  |  |  |  |  |  |  1 = low: yes
|  |  |  |  |  |  4 = low: no
|  |  |  |  5 = low: no
|  |  |  7 = high
|  |  |  |  1 = low
|  |  |  |  |  4 = high
|  |  |  |  |  |  3 = high: yes
|  |  |  |  |  |  3 = low: yes
|  |  |  |  |  4 = low: yes
|  |  |  |  1 = high: yes
|  8 = low
|  |  6 = high
|  |  |  4 = high
|  |  |  |  1 = low
|  |  |  |  |  7 = high
|  |  |  |  |  |  3 = high
|  |  |  |  |  |  |  5 = high: no
|  |  |  |  |  |  |  5 = low: no
|  |  |  |  |  |  3 = low
|  |  |  |  |  |  |  5 = high: yes
|  |  |  |  |  |  |  5 = low: yes
|  |  |  |  |  7 = low
|  |  |  |  |  |  3 = low
|  |  |  |  |  |  |  5 = high: no
|  |  |  |  |  |  |  5 = low: no
|  |  |  |  |  |  3 = high
|  |  |  |  |  |  |  5 = high: no
|  |  |  |  |  |  |  5 = low: no
|  |  |  |  1 = high: yes
|  |  |  4 = low
|  |  |  |  7 = high
|  |  |  |  |  3 = low
|  |  |  |  |  |  5 = high: no
|  |  |  |  |  |  5 = low: no
|  |  |  |  |  3 = high: no
|  |  |  |  7 = low: no
|  |  6 = low
|  |  |  7 = low: no
|  |  |  7 = high
|  |  |  |  5 = high: no
|  |  |  |  5 = low
|  |  |  |  |  3 = low: yes
|  |  |  |  |  3 = high: no
```

```
1 = medium
|  5 = high
|  |  3 = low
|  |  |  2 = high
|  |  |  |  4 = low: no
|  |  |  |  4 = high: no
|  |  |  2 = low: no
|  |  3 = high
|  |  |  4 = low
|  |  |  |  2 = high: no
|  |  |  |  2 = low: no
|  |  |  4 = high: yes
|  5 = low
|  |  3 = high
|  |  |  4 = high
|  |  |  |  2 = high: no
|  |  |  |  2 = low: no
|  |  |  4 = low
|  |  |  |  2 = high: no
|  |  |  |  2 = low: no
|  |  3 = low
|  |  |  4 = low: no
|  |  |  4 = high
|  |  |  |  2 = high: no
|  |  |  |  2 = low: no
```

```
2 = high
|  6 = high
|  |  8 = high
|  |  |  7 = high
|  |  |  |  3 = high
|  |  |  |  |  1 = low
|  |  |  |  |  |  4 = high
|  |  |  |  |  |  |  5 = high: yes
|  |  |  |  |  |  |  5 = low: yes
|  |  |  |  |  |  4 = low: yes
|  |  |  |  |  1 = high: yes
|  |  |  |  3 = low
|  |  |  |  |  1 = low: yes
|  |  |  |  |  1 = high: no
|  |  |  7 = low
|  |  |  |  5 = high
|  |  |  |  |  4 = high
|  |  |  |  |  |  3 = high
|  |  |  |  |  |  |  1 = high: yes
|  |  |  |  |  |  |  1 = low: no
|  |  |  |  |  |  3 = low
|  |  |  |  |  |  |  1 = low: yes
|  |  |  |  |  |  |  1 = high: yes
|  |  |  |  |  4 = low
|  |  |  |  |  |  1 = low
|  |  |  |  |  |  |  3 = high: yes
|  |  |  |  |  |  |  3 = low: no
|  |  |  |  |  |  1 = high: no
|  |  |  |  5 = low: yes
|  |  8 = low
|  |  |  4 = high
|  |  |  |  7 = low
|  |  |  |  |  3 = high: no
|  |  |  |  |  3 = low: yes
|  |  |  |  7 = high
|  |  |  |  |  3 = high: yes
|  |  |  |  |  3 = low: no
|  |  |  4 = low
|  |  |  |  3 = high: no
|  |  |  |  3 = low
|  |  |  |  |  5 = high
|  |  |  |  |  |  7 = high: yes
|  |  |  |  |  |  7 = low: no
|  |  |  |  |  5 = low: no
|  6 = low
|  |  4 = high
|  |  |  5 = high
|  |  |  |  7 = high: no
|  |  |  |  7 = low
|  |  |  |  |  8 = high
|  |  |  |  |  |  3 = high: no
|  |  |  |  |  |  3 = low: no
|  |  |  |  |  8 = low
|  |  |  |  |  |  3 = low: yes
|  |  |  |  |  |  3 = high: no
|  |  |  5 = low
|  |  |  |  7 = high: yes
|  |  |  |  7 = low: no
|  |  4 = low: no
```

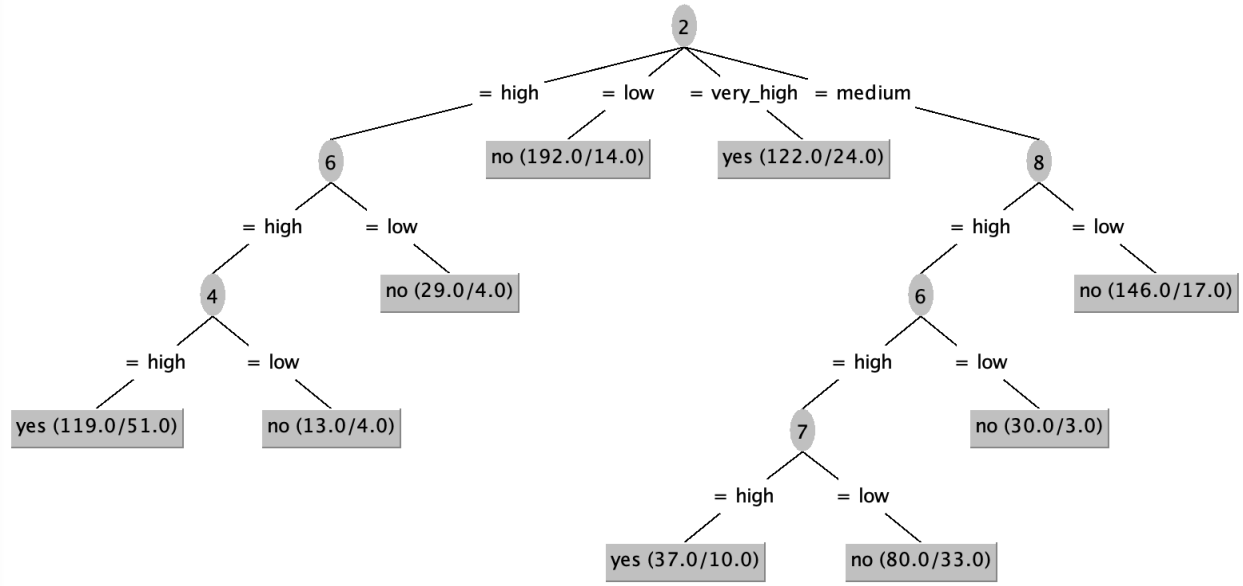
```
1 = high
|  3 = high
|  |  5 = high
|  |  |  4 = high
|  |  |  |  2 = high: yes
|  |  |  |  2 = low: yes
|  |  |  4 = low
|  |  |  |  2 = high: yes
|  |  |  |  2 = low: yes
|  |  5 = low
|  |  |  2 = high
|  |  |  |  4 = low: no
|  |  |  |  4 = high: no
|  |  |  2 = low: no
|  3 = low
|  |  2 = high
|  |  |  4 = high: no
|  |  |  4 = low
|  |  |  |  5 = high: no
|  |  |  |  5 = low: no
|  |  2 = low
|  |  |  4 = high
|  |  |  |  5 = low: no
|  |  |  |  5 = high: yes
|  |  |  4 = low: no
```

```
2 = very_high
|  5 = high
|  |  6 = low
|  |  |  8 = high
|  |  |  |  4 = high
|  |  |  |  |  1 = high
|  |  |  |  |  |  7 = high
|  |  |  |  |  |  |  3 = low: yes
|  |  |  |  |  |  |  3 = high: yes
|  |  |  |  |  |  7 = low: yes
|  |  |  |  |  1 = low
|  |  |  |  |  |  3 = low: yes
|  |  |  |  |  |  3 = high: yes
|  |  |  |  4 = low: yes
|  |  |  8 = low
|  |  |  |  3 = low: no
|  |  |  |  3 = high
|  |  |  |  |  4 = low: yes
|  |  |  |  |  4 = high: no
|  |  6 = high
|  |  |  1 = low
|  |  |  |  8 = high
|  |  |  |  |  7 = low
|  |  |  |  |  |  3 = high
|  |  |  |  |  |  |  4 = high: yes
|  |  |  |  |  |  |  4 = low: yes
|  |  |  |  |  |  3 = low
|  |  |  |  |  |  |  4 = low: yes
|  |  |  |  |  |  |  4 = high: yes
|  |  |  |  |  7 = high
|  |  |  |  |  |  4 = high
|  |  |  |  |  |  |  3 = high: yes
|  |  |  |  |  |  |  3 = low: yes
|  |  |  |  |  |  4 = low: yes
|  |  |  8 = low
|  |  |  |  7 = high: yes
|  |  |  |  7 = low
|  |  |  |  |  4 = high
|  |  |  |  |  |  3 = high: yes
|  |  |  |  |  |  3 = low: yes
|  |  |  |  |  4 = low
|  |  |  |  |  |  3 = high: yes
|  |  |  |  |  |  3 = low: no
|  |  |  1 = high
|  |  |  |  7 = high: yes
|  |  |  |  7 = low
|  |  |  |  |  3 = high: yes
|  |  |  |  |  3 = low: yes
|  5 = low
|  |  7 = low: no
|  |  7 = high: yes
```

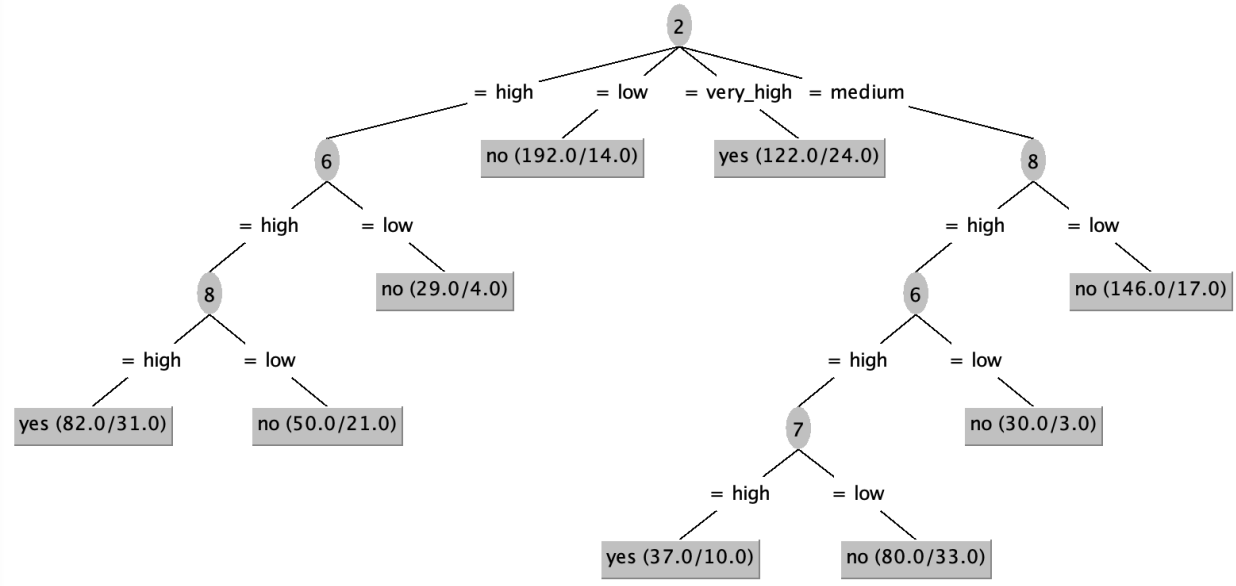
```
1 = very_high
|  2 = high
|  |  3 = low
|  |  |  5 = high
|  |  |  |  4 = high: yes
|  |  |  |  4 = low: yes
|  |  |  5 = low
|  |  |  |  4 = high: no
|  |  |  |  4 = low: no
|  |  3 = high
|  |  |  5 = high
|  |  |  |  4 = low: yes
|  |  |  |  4 = high: yes
|  |  |  5 = low
|  |  |  |  4 = high: yes
|  |  |  |  4 = low: yes
|  2 = low
|  |  4 = low: no
|  |  4 = high: yes
```

3.2.2 Pruned DT

No feature selection



Correlation-based Feature Selection



3.2.3 Unpruned DT

No Feature Selection	Correlation-based Feature Selection

```
2 = high
|   6 = high
|   |   4 = high
|   |   |   1 = low
|   |   |   |   7 = high
|   |   |   |   |   8 = high: yes (16.0/5.0)
|   |   |   |   |   8 = low
|   |   |   |   |   |   3 = high: yes (11.0/5.0)
|   |   |   |   |   |   3 = low: no (5.0/2.0)
|   |   |   |   7 = low
|   |   |   |   |   3 = high: no (43.0/19.0)
|   |   |   |   |   3 = low: yes (10.0/4.0)
|   |   |   1 = high
|   |   |   |   3 = high: yes (29.0/8.0)
|   |   |   |   3 = low
|   |   |   |   |   7 = high: no (2.0)
|   |   |   |   |   7 = low: yes (3.0)
|   |   4 = low: no (13.0/4.0)
|   6 = low: no (29.0/4.0)
2 = low
|   6 = high
|   |   5 = high
|   |   |   8 = high
|   |   |   |   7 = high: yes (7.0/3.0)
|   |   |   |   7 = low: no (28.0/4.0)
|   |   |   8 = low: no (43.0/4.0)
|   |   5 = low: no (48.0/2.0)
|   6 = low: no (66.0)
2 = very_high
|   5 = high
|   |   6 = high: yes (103.0/16.0)
|   |   6 = low
|   |   |   8 = high: yes (12.0/3.0)
|   |   |   8 = low: no (4.0/1.0)
|   5 = low: no (3.0/1.0)
2 = medium
|   8 = high
|   |   5 = high
|   |   |   6 = high
|   |   |   |   7 = high: yes (37.0/10.0)
|   |   |   |   7 = low
|   |   |   |   |   3 = high: no (57.0/24.0)
|   |   |   |   |   3 = low
|   |   |   |   |   |   4 = high: yes (15.0/7.0)
|   |   |   |   |   |   4 = low: no (3.0/1.0)
|   |   |   6 = low: no (27.0/3.0)
|   |   5 = low: no (8.0)
|   8 = low
|   |   6 = high
|   |   |   1 = low
|   |   |   |   4 = high
|   |   |   |   |   7 = high
|   |   |   |   |   |   3 = high: no (17.0/2.0)
|   |   |   |   |   |   3 = low: yes (7.0/3.0)
|   |   |   |   |   7 = low: no (54.0/8.0)
|   |   |   |   4 = low: no (24.0/1.0)
|   |   |   1 = high: yes (2.0/1.0)
|   |   6 = low: no (42.0/1.0)
```

```
2 = high
|   6 = high
|   |   8 = high: yes (82.0/31.0)
|   |   8 = low: no (50.0/21.0)
|   6 = low: no (29.0/4.0)
2 = low
|   6 = high
|   |   5 = high
|   |   |   8 = high
|   |   |   |   7 = high: yes (7.0/3.0)
|   |   |   |   7 = low: no (28.0/4.0)
|   |   |   8 = low: no (43.0/4.0)
|   |   5 = low: no (48.0/2.0)
|   6 = low: no (66.0)
2 = very_high
|   5 = high
|   |   6 = high: yes (103.0/16.0)
|   |   6 = low
|   |   |   8 = high: yes (12.0/3.0)
|   |   |   8 = low: no (4.0/1.0)
|   5 = low: no (3.0/1.0)
2 = medium
|   8 = high
|   |   6 = high
|   |   |   7 = high: yes (37.0/10.0)
|   |   |   7 = low: no (80.0/33.0)
|   |   6 = low: no (30.0/3.0)
|   8 = low: no (146.0/17.0)
```

3.3 Discussion

3.3.1 Classifier Performance Comparison

The following table compares the worst and best performing classifiers and their accuracies, with and without feature selection.

	Accuracy (%)	
	Lowest	Highest
No feature selection	ZeroR: 65.1042	Bagging: 74.8698
CFS	ZeroR: 65.1042	DT: 79.4271

It is evident that there was variability in accuracy results dependent on the classifier and whether CFS was applied. ZeroR performed the worst in terms of accuracy and even with CFS failed to achieve an improvement in accuracy. The classifier with the highest accuracy when no feature selection was applied was Bagging. Bagging reported a higher accuracy when CFS was used, with an increase of 3.6458%.

The algorithm that performed best with CFS was the Decision Tree, both pruned and unpruned with equal highest accuracies. Both classifiers achieved significantly higher accuracy with CFS compared to no feature selection with an improvement of 4.4271% and 4.0365% for unpruned and pruned decision trees, respectively.

Our implementation of Naive Bayes achieved a higher accuracy compared to Weka's NB classifier with and without feature selection. Alternatively, our implementation of Decision Tree achieved lower accuracy than Weka's pruned and unpruned DT algorithms, with and without CFS (Discussion at 3.3.3.2).

3.3.2 Feature Selection

3.3.2.1 Feature Analysis

Conducting CFS selected a subset of the original features that made intuitive sense. Examining the attributes selected by CFS reveals relations with many traditional predictors of diabetes.

<u>Plasma Glucose Concentration</u>
Plasma glucose concentration measurements are an important tool used to screen diabetes.
It is measured through an oral glucose tolerance test with a concentration of 200 mg/dL considered diagnostic (Gurung & Jialal, 2020). Plasma glucose levels can be used as a tool to monitor diabetes as "higher glucose correlates with more significant co-morbidities and risk for mortality" (Gurung & Jialal, 2020). Hence, it is evident that this feature is strongly correlated with the onset of diabetes and was included in the CFS subset of original features.

2-Hour Serum Insulin

Insulin is an anabolic hormone that promotes glucose uptake, glycogenesis, lipogenesis, and protein synthesis of skeletal muscle and fat tissue through the tyrosine kinase receptor pathway (Buppajarntham, 2019). Insulin is one of the key factors in the regulation of plasma glucose homeostasis, as it counteracts glucagon and other catabolic hormones (Buppajarntham, 2019). Increased insulin levels have been associated with the early stages of type 2 diabetes as therefore are positively correlated with the onset of diabetes and consequently selected by CFS.

Body Mass Index

Body mass index (BMI) is a value derived from the mass and height of a person. The BMI is defined as the body mass divided by the square of the body height, and is expressed in units of kg/m^2. Research indicates there is a correlation between BMI and diabetes, with "Increased BMI was associated with increased prevalence of diabetes mellitus" (Bays, Chapman & Grandy, 2007). Therefore, it is unsurprising that CFS has selected this attribute due to the correlation between BMI and diabetes.

Diabetes Pedigree Function

The diabetes pedigree function provides "a synthesis of the diabetes mellitus history in relatives and the genetic relationship of those relatives to the subject" (Shanker, Hu & Hung, 2000). This function provides a measure of the expected genetic influence of affected and unaffected relatives on the subject's eventual risk (Shanker, Hu & Hung, 2000). As a result, the CFS selection of this attribute is appropriate for this study.

Age

The increased prevalence of glucose intolerance with advancing age indicates there exists a correlation between age and diabetes. This is possibly attributed to the "decline in lean body mass and increase in body fat, particularly visceral adiposity that often accompanies aging, may contribute to the development of insulin resistance" (Suastika, Dwipayana, Siswadi & Tuty, 2012). Unsurprisingly, CFS has selected age as an important attribute to predict the onset of diabetes for an individual.

3.3.2.2 Feature Selection Effect

For both numeric and nominal data, on par or increased accuracy of the Weka classifiers suggests that feature selection was largely beneficial. In particular for the nominal data, feature selection significantly improved the accuracy of the classifiers and in some cases up to 5%. This is best highlighted in the comparison of unpruned decision trees (Section 3.2.3), where the size of the CFS unpruned decision tree is substantially smaller than its non-feature selection counterpart.

Advantages

Accuracy

Accuracy refers to the proportion of correctly classified examples and is used to determine if the classifier is performing well. CFS maintained or increased classifier accuracy for all Weka machine learning algorithms. Specifically, nominal data classifiers experienced the greater

increase in accuracy (Section 3.3.4). As CFS aims to discard redundant data, overfitting is reduced leading to more robust algorithms that generalise well. Greater classifier accuracy enables us to derive conclusions with greater certainty and communicate these conclusions to stakeholders. In the context of this study, medical professionals must have a level of trust that these machine learning algorithms are able to predict the onset of diabetes with accuracy as these predictions impact the lives of individuals. Therefore, as feature selection increases accuracy, this is a key advantage of its usage.

Dimensionality

Feature selection reduced the number of predictor attributes from eight to five, and in turn the dimensionality of the dataset. This assists in preventing overfitting when the model works very well on a set of training data but fails to predict new examples. Reducing the number of attributes eliminates irrelevant features without incurring a loss in information (Xu, 2018). For example, CFS failed to select the feature *number of pregnancies*. Research suggests that pregnancy is adversely affected by diabetes, resulting in increased risk of “preterm delivery, preeclampsia, macrosomia, shoulder dystocia, intrauterine fetal demise, fetal growth restriction, cardiac and renal malformations” (Vargas, Repke & Ural, 2010). However, there is little evidence to suggest the converse is true; that the number of pregnancies has any bearing on an individual's risk of diabetes. Therefore, feature selection reduces the dimensionality of the dataset, preventing overfitting and the inclusion of redundant attributes in predictive models.

Processing Time

As feature selection reduces dataset dimensionality, there are computational advantages that increase processing efficiency. As there is less data to process due to less features, the algorithms train and are able to learn faster. This reduces the computational cost of running these algorithms, improving the efficiency of conducting research and obtaining results. In this study, obtaining results at a faster rate leads to earlier potential diagnosis of diabetes which is an advantageous medical outcome. Thus, improved computational efficiency of running machine learning algorithms with feature selection offers considerable advantages.

3.3.3 Decision Tree Pruning

3.3.3.1 Role and Effect

Pruning aims to address issues of overfitting in decision trees. Overfitting occurs when the classifier performs very well with training examples but struggles to classify new examples. This can occur in decision trees when the tree becomes very specific where each branch of the tree is deep enough to perfectly classify training examples, likened to a look-up table (Koprinska, 2021). Therefore, pruning can be used to stop growing the decision tree earlier to the point when it perfectly classifies training data or pruning it once it is fully grown, with the latter the most common approach (Koprinska, 2021).

3.3.3.2 Classifier Comparison

Examining classifier accuracy results (Section 3.1) indicates our implementation of the Decision Tree classifier compares slightly less accurately to the pruned and unpruned Weka decision trees.

MyDT achieved an accuracy of 1.3055% less than the Weka unpruned decision tree and 1.6961% less than the pruned decision tree. When CFS was applied, MyDT achieved an increased accuracy of 4.0345% from the decision tree with no feature selection - a significant improvement. However, in comparison to Weka's pruned and unpruned decision trees, it performed 1.6981% worse in terms of accuracy.

There are a range of possible explanations for this lower accuracy level in comparison to Weka decision trees. Weka classification algorithms have been optimised at a higher level due to the technical expertise of machine learning algorithm developers. Extensive testing and optimisations are to be expected when developing software such as Weka as these algorithms are designed to cater for a range of datasets. Therefore, accuracy levels of Weka decision tree algorithms are higher in comparison to our implementation.

3.3.4 Tree-Based Classifiers Accuracy Comparison

Examining the accuracy levels of the tree-based classifiers reveals that all were benefited by CFS to various extents. Boosting, which aims to make the classifiers complement each other, performed the best when no feature selection was applied and was least impacted by CFS with an accuracy improvement after CFS of 2.4739%. Alternatively, Random Forest performed the best when CFS was applied, experiencing the largest improvement in accuracy between no feature selection and CFS, with an accuracy improvement of 5.7292%.

Weka's AdaBoost algorithm initially assigns equal weights to each training observation and uses multiple weak models, assigning higher weights to those observations that were misclassified. As it uses multiple weak models, combining the results of the decision boundaries achieved during multiple iterations, the accuracy of the misclassified observations is improved, and hence the accuracy of the overall iterations is improved (Rahman et al., 2020). As this algorithm incorporates feature selection in it's iterative process, this may explain the relative lack of impact on accuracy that CFS has on the Boosting algorithm.

The Random Forest algorithm is an extension of bagging for decision trees that can be used for classification. It improves upon bagged decision trees "that disrupts the greedy splitting algorithm during tree creation so that split points can only be selected from a random subset of the input attributes. This simple change can have a big effect decreasing the similarity between the bagged trees and in turn the resulting predictions" (Brownlee, 2016). Therefore, as bagging and random feature selection are used to generate diversity and reduce correlation, CFS has a greater effect on the accuracy by selecting the subset of the original attributes that are highly correlated with the class.

4 Conclusion

4.1 Findings

This study has assessed the performance of various classifiers in an attempt to determine the best predictor for diabetes. It is evident that Correlation-based Feature Selection is beneficial to classifiers, maintaining or improving accuracy and reducing the dimensionality of datasets leading to improved algorithm performance. Comparative analysis of classifier accuracy reveals decision trees, particularly those that used CFS, were the most accurate classifiers by a considerable margin. Based on these results, the decision trees with CFS selected *Plasma glucose concentration, 2-Hour serum insulin, Body mass index, Diabetes pedigree function* and *Age* as the best subset of features for diabetes prediction. Thus, it can be concluded that CFS can enhance the performance of machine learning algorithms to predict the onset of diabetes for an individual.

4.2 Future Work

4.2.1 Algorithm Optimisation

Future work is required to optimise our implementations of machine learning algorithms for greater robustness. Our implementation of the Naive Bayes algorithm assumed an ideal state of input data, which is not applicable to real world datasets. For example, when the conditional probability of some event being zero. In this case, a Laplace correction or as a generalisation an M-estimate could be implemented to avoid this issue (Koprinska, 2021). Additionally, considering cases of missing data can be handled by not including the attribute when the attribute value in the new example is missing and for training examples not including the value in counts (Koprinska, 2021). There is scope for improving the Decision Tree algorithm to avoid overfitting using pruning techniques such as subtree replacement. This approach involves evaluating all candidate nodes and accepting the best pruning that will improve the accuracy most, iterating until any further pruning decreases accuracy on the validation set (Koprinska, 2021). Therefore future work to achieve greater robustness in our algorithms is required to apply these classifiers on a range of datasets.

4.2.2 Correlation Measures

Exploring alternative measures of correlation used in CFS is an area for future work. In this study, it was observed that CFS has a significantly greater impact in terms of accuracy on nominal data classifiers rather than numeric data. One explanation for this difference in accuracy levels may be the unequal treatment of nominal features, that is non-uniform when compared to numeric features with a common basis for computing correlation (Hall, 1997). One area of research to consider is performing reverse discretisation as described by Ting (1995). which involves converting nominal attributes to numeric attributes, where each nominal value of an attribute is replaced with its estimated prior probability from the training data (Hall, 1999). As all attributes including the class are numeric, Pearson’s linear correlation can be used with CFS (Hall, 1999).

There is scope for future work to examine how this anti-discretisation process will affect the accuracy of machine learning algorithms. Hence, there is scope for future work to explore alternative measures of correlation to ensure consistency in accuracy results across both numeric and nominal data types.

5 Reflection

1)

In developing our own implementations of the Naive Bayes and Decision Tree classifiers we gained an appreciation for open source machine learning software such as Weka. Implementing our own algorithms was a relatively inefficient process that required abstraction of the problem, development and testing of code for correctness. The statistical output of the model processing and Weka tools such as preprocessing and visualisation significantly reduced the development time of machine learning models. Additionally, the ability to apply various models on the same dataset allowed us to compare the performance of the models with ease and select ones best suited for the purpose of our study. Thus, with the increased efficiency of model development we were able to critically analyse the results and derive meaningful conclusions.

2)

Observing the benefits of feature selection indicated it is an effective tool that can improve the accuracy of machine learning algorithms whilst reducing dimensionality. In particular, CFS led to increased classifier accuracy whilst simultaneously reducing overfitting and improving the performance of algorithms. Performance improvements are desirable as they reduce the time required to train algorithms and time taken to collect results. This allows researchers to conduct deeper analysis of results and consider a range of classifiers as opposed to only a limited selection of algorithms. Where appropriate, feature selection should be considered for the advantages discussed in this study. Consequently, feature selection was a beneficial technique applied in this study and we have gained a deeper understanding of its processes and advantages.

References

Bays, H., Chapman, R., & Grandy, S. (2021). The relationship of body mass index to diabetes mellitus, hypertension and dyslipidaemia: comparison of data from two national surveys. Retrieved 12 May 2021, from

Brownlee, J. (2021). How to Use Ensemble Machine Learning Algorithms in Weka. Retrieved 13 May 2021, from https://machinelearningmastery.com/use-ensemble-machine-learning-algorithms-weka/

Buppajarntham, S. (2019). Insulin: Reference Range, Interpretation, Collection and Panels. Retrieved 14 May 2021, from https://emedicine.medscape.com/article/2089224-overview

Gurung, P., & Jialal, I. (2020). Plasma Glucose. Retrieved 13 May 2021, from https://www.ncbi.nlm.nih.gov/books/NBK541081/#:~:text=Normal%20plasma%20glucose%20levels%20are,individuals%20can%20vary%20with%20age.

Hall, M., & Smith, L. Practical Feature Subset Selection for Machine Learning.

Hall, M. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. Retrieved 13 May 2021, from https://www.cs.odu.edu/~mukka/cs795sum10dm/Lecturenotes/Day4/10.1.1.148.6025%5B1%5D.pdf

Koprinska, I. (2021). *COMP3308/3608 - Introduction to Artificial Intelligence*. Lecture, University of Sydney.

Rahman, S., Irfan, M., Raza, M., Moyeezullah Ghori, K., Yaqoob, S., & Awais, M. (2020). Performance Analysis of Boosting Classifiers in Recognizing Activities of Daily Living. *International Journal Of Environmental Research And Public Health*, *17*(3), 1082. doi: 10.3390/ijerph17031082

Shanker, M., Hu, M., & Hung, M. (2000). Estimating Probabilities of Diabetes Mellitus Using Neural Networks. *SAR And QSAR In Environmental Research*, *11*(2), 133-147. doi: 10.1080/10629360008039119

Shaw, J., & Tanamas, S. (2012). diabetes: the silent pandemic and its impact on Australia. Retrieved 9 May 2021, from https://www.diabetesaustralia.com.au/wp-content/uploads/Diabetes-the-silent-pandemic-and-its-impact-on-Australia.pdf

Suastika, K., Dwipayana, P., Siswadi, M., & Tuty, R. (2012). Age is an Important Risk Factor for Type 2 Diabetes Mellitus and Cardiovascular Diseases. Retrieved 14 May 2021, from

Vargas, R., Repke, J., & Ural, S. (2010). Type 1 Diabetes Mellitus and Pregnancy. Retrieved 13 May 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3046748/

Xu, C. (2018). Why is Dimensionality Reduction so Important?. Retrieved 13 May 2021, from https://medium.com/@cxu24/why-dimensionality-reduction-is-important-dd60b5611543#:~:text=In%20addition%20to%20avoiding%20overfitting,feature%20extraction%20in%20next%20blogs.