Title: Accuracy of Simple Classifiers

Authors: 460251747

Date: May 28, 2018

Aim

In this study, we applied 12 different machine learning algorithms to a medical dataset using weka to predict whether a patient will be tested positive for diabetes or not. In addition, we implemented a naive bayes and decision tree algorithms and compare those against weka implementations. Finally to preprocess the data we used correlation-based feature selection (CFS) which selects the most relevant attributes of the data. Each algorithm was ran both with and without using CFS. The accuracy of each algorithm for predicting diabetes was measured and the most accurate algorithm will the suggested. This study is important because it will suggest the best algorithm to be used on more medical related data sets.

Data

The data used for this project comes from the National Institute of Diabetes and Digestive Kidney Diseases in the United States. The data contains 768 instances (each one representing a patient), each with eight different attributes. These ones are described below .

Attributes

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold 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)
9. Class variable ("yes" or "no")

The Correlation-based feature selection (CFS) selects a subset of attributes which are deemed the most relevant by a heuristic. This heuristic considers how good

the individual attributes are at predicting if a patient will be tested positive for diabetes and how much each attribute correlates to the other attributes. In other words, the heuristic selects features which are highly correlated with testing positive for diabetes or not but are uncorrelated with each other. The following were the attributes selected by the CFS method.

Selected Attributes

1. Plasma glucose concentration a 2 hours 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)

Results and discussion

Classifier Accuracy

Numeric Data	ZeroR	1R	1NN	5NN	NB	MLP	SVM	RF	MyNB
No feature selection	65.1042 %	70.8333 %	67.8385 %	74.4792 %	75.1302 %	75.3906 %	76.3021 %	74.8698 %	75.2604 %
CFS	65.1042 %	70.8333 %	69.0104 %	74.4792 %	76.3021 %	75.7813 %	76.6927 %	75.9115 %	76.8229 %

Nominal Data	DT unpruned	DT pruned	MyDT
No feature selection	75 %	75.3906 %	82.1615 %
CFS	80.0781 %	79.5573 %	80.3385 %

Decision Tree Diagrams

Weka Unpruned DT

```
 Plasma glucose concentration a 2 hours in an oral glucose tolerance test = high
|    Body mass index = high
|   |    Triceps skin fold thickness = high
|   |   |   Number of times pregnant = 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)
|   |   |   Number of times pregnant = 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 a 2 hours in an oral glucose tolerance test = 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 a 2 hours in an oral glucose tolerance test = 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 a 2 hours in an oral glucose tolerance test = 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
|   |   |   Number of times pregnant = 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)
|   |   |   Number of times pregnant = high: yes (2.0/1.0)
|   |    Body mass index = low: no (42.0/1.0)
```

Weka Pruned DT

```
 Plasma glucose concentration a 2 hours in an oral glucose tolerance test = 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 a 2 hours in an oral glucose tolerance test = low: no (192.0/1
 Plasma glucose concentration a 2 hours in an oral glucose tolerance test = very high: yes (
 Plasma glucose concentration a 2 hours in an oral glucose tolerance test = 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)
```

MyDT

```
Plasma glucose concentration a 2 hours in an oral glucose tolerance test = high
|    Body mass index = high
|    |    age = high
|    |    |    Diabetes pedigree function = high
|    |    |    |    Diastolic blood pressure = high
|    |    |    |    |    Number of times pregnant = high
|    |    |    |    |    |    riceps skin fold thickness = high
|    |    |    |    |    |    |    2-Hour serum insulin = high: yes
|    |    |    |    |    Number of times pregnant = low
|    |    |    |    |    |    riceps skin fold thickness = high
|    |    |    |    |    |    |    2-Hour serum insulin = high: yes
|    |    |    |    |    |    |    2-Hour serum insulin = low: yes
|    |    |    |    |    |    riceps skin fold thickness = low: yes
|    |    |    |    Diastolic blood pressure = low
|    |    |    |    |    Number of times pregnant = high: no
|    |    |    |    |    Number of times pregnant = low
|    |    |    |    |    |    riceps skin fold thickness = high
|    |    |    |    |    |    |    2-Hour serum insulin = high: yes
|    |    |    Diabetes pedigree function = low
|    |    |    |    2-Hour serum insulin = high
```

```
|   |   |   |   |   |   riceps skin fold thickness = high
|   |   |   |   |   |   |   Diastolic blood pressure = high
|   |   |   |   |   |   |   |   Number of times pregnant = high: yes
|   |   |   |   |   |   |   |   Number of times pregnant = low: no
|   |   |   |   |   |   |   Diastolic blood pressure = low
|   |   |   |   |   |   |   |   Number of times pregnant = high: yes
|   |   |   |   |   |   |   |   Number of times pregnant = low: yes
|   |   |   |   |   |   riceps skin fold thickness = low
|   |   |   |   |   |   |   Number of times pregnant = high: no
|   |   |   |   |   |   |   Number of times pregnant = low
|   |   |   |   |   |   |   |   Diastolic blood pressure = high: yes
|   |   |   |   |   |   |   |   Diastolic blood pressure = low: no
|   |   |   |   2-Hour serum insulin = low: yes
|   |   |   age = low
|   |   |   |   riceps skin fold thickness = high
|   |   |   |   |   Diabetes pedigree function = high
|   |   |   |   |   |   Diastolic blood pressure = high
|   |   |   |   |   |   |   Number of times pregnant = low
|   |   |   |   |   |   |   |   2-Hour serum insulin = high: yes
|   |   |   |   |   Diastolic blood pressure = low
|   |   |   |   |   |   Number of times pregnant = low
|   |   |   |   |   |   |   |   2-Hour serum insulin = high: no
|   |   |   |   Diabetes pedigree function = low
|   |   |   |   |   Diastolic blood pressure = high
|   |   |   |   |   |   Number of times pregnant = low
|   |   |   |   |   |   |   |   2-Hour serum insulin = high: no
|   |   |   |   |   Diastolic blood pressure = low
|   |   |   |   |   |   Number of times pregnant = low
|   |   |   |   |   |   |   |   2-Hour serum insulin = high: yes
|   |   |   |   riceps skin fold thickness = low
|   |   |   |   |   Diastolic blood pressure = high: no
|   |   |   |   |   Diastolic blood pressure = low
|   |   |   |   |   |   2-Hour serum insulin = high
|   |   |   |   |   |   |   Diabetes pedigree function = high
|   |   |   |   |   |   |   |   Number of times pregnant = low: yes
|   |   |   |   |   |   |   Diabetes pedigree function = low: no
|   |   |   |   |   |   2-Hour serum insulin = low: no
|   |   Body mass index = low
|   |   |   riceps skin fold thickness = high
|   |   |   |   2-Hour serum insulin = high
|   |   |   |   |   Diabetes pedigree function = high: no
|   |   |   |   |   Diabetes pedigree function = low
|   |   |   |   |   |   age = high
|   |   |   |   |   |   |   Diastolic blood pressure = high
|   |   |   |   |   |   |   |   Number of times pregnant = low: no
|   |   |   |   |   |   |   Diastolic blood pressure = low: no
```

```
|    |    |    |    |    age = low
|    |    |    |    |    |    Diastolic blood pressure = high: no
|    |    |    |    |    |    Diastolic blood pressure = low
|    |    |    |    |    |    |    Number of times pregnant = low: yes
|    |    |    2-Hour serum insulin = low
|    |    |    |    Diabetes pedigree function = high: yes
|    |    |    |    Diabetes pedigree function = low: no
|    |    riceps skin fold thickness = low: no
Plasma glucose concentration a 2 hours in an oral glucose tolerance test = low
|    Body mass index = high
|    |    2-Hour serum insulin = high
|    |    |    age = high
|    |    |    |    Diabetes pedigree function = high
|    |    |    |    |    Diastolic blood pressure = high
|    |    |    |    |    |    Number of times pregnant = high
|    |    |    |    |    |    |    riceps skin fold thickness = high: yes
|    |    |    |    |    |    |    riceps skin fold thickness = low: no
|    |    |    |    |    |    Number of times pregnant = low
|    |    |    |    |    |    |    riceps skin fold thickness = high: no
|    |    |    |    |    |    |    riceps skin fold thickness = low: yes
|    |    |    |    |    Diastolic blood pressure = low: yes
|    |    |    |    Diabetes pedigree function = low
|    |    |    |    |    riceps skin fold thickness = high
|    |    |    |    |    |    Number of times pregnant = high
|    |    |    |    |    |    |    Diastolic blood pressure = high: no
|    |    |    |    |    |    |    Diastolic blood pressure = low: no
|    |    |    |    |    |    Number of times pregnant = low
|    |    |    |    |    |    |    Diastolic blood pressure = high: no
|    |    |    |    |    |    |    Diastolic blood pressure = low: no
|    |    |    |    |    riceps skin fold thickness = low: no
|    |    |    age = low
|    |    |    |    Diastolic blood pressure = high: no
|    |    |    |    Diastolic blood pressure = low
|    |    |    |    |    riceps skin fold thickness = high
|    |    |    |    |    |    Diabetes pedigree function = high
|    |    |    |    |    |    |    Number of times pregnant = low: no
|    |    |    |    |    |    Diabetes pedigree function = low
|    |    |    |    |    |    |    Number of times pregnant = low: no
|    |    |    |    |    riceps skin fold thickness = low: no
|    |    2-Hour serum insulin = low
|    |    |    Diastolic blood pressure = high
|    |    |    |    age = high: no
|    |    |    |    age = low
|    |    |    |    |    riceps skin fold thickness = high
|    |    |    |    |    |    Diabetes pedigree function = high: yes
|    |    |    |    |    |    Diabetes pedigree function = low
```

```
|    |    |    |    |    |    |    Number of times pregnant = low: no
|    |    |    |    |    |    riceps skin fold thickness = low: no
|    |    |    Diastolic blood pressure = low: no
|    Body mass index = low: no
Plasma glucose concentration a 2 hours in an oral glucose tolerance test = medium
|    age = high
|    |    Body mass index = high
|    |    |    Diabetes pedigree function = high
|    |    |    |    Number of times pregnant = high: yes
|    |    |    |    Number of times pregnant = low
|    |    |    |    |    riceps skin fold thickness = high
|    |    |    |    |    |    Diastolic blood pressure = high
|    |    |    |    |    |    |    2-Hour serum insulin = high: yes
|    |    |    |    |    |    Diastolic blood pressure = low
|    |    |    |    |    |    |    2-Hour serum insulin = high: yes
|    |    |    |    |    riceps skin fold thickness = low: yes
|    |    |    Diabetes pedigree function = low
|    |    |    |    2-Hour serum insulin = high
|    |    |    |    |    Diastolic blood pressure = high
|    |    |    |    |    |    Number of times pregnant = high
|    |    |    |    |    |    |    riceps skin fold thickness = high: no
|    |    |    |    |    |    Number of times pregnant = low
|    |    |    |    |    |    |    riceps skin fold thickness = high: no
|    |    |    |    |    |    |    riceps skin fold thickness = low: yes
|    |    |    |    |    Diastolic blood pressure = low
|    |    |    |    |    |    riceps skin fold thickness = high
|    |    |    |    |    |    |    Number of times pregnant = high: yes
|    |    |    |    |    |    |    Number of times pregnant = low: yes
|    |    |    |    |    |    riceps skin fold thickness = low
|    |    |    |    |    |    |    Number of times pregnant = low: no
|    |    |    |    2-Hour serum insulin = low: no
|    |    Body mass index = low
|    |    |    Diastolic blood pressure = high
|    |    |    |    Number of times pregnant = high: no
|    |    |    |    Number of times pregnant = low
|    |    |    |    |    Diabetes pedigree function = high: no
|    |    |    |    |    Diabetes pedigree function = low
|    |    |    |    |    |    riceps skin fold thickness = high
|    |    |    |    |    |    |    2-Hour serum insulin = high: no
|    |    |    |    |    |    riceps skin fold thickness = low: no
|    |    |    Diastolic blood pressure = low
|    |    |    |    Number of times pregnant = high: yes
|    |    |    |    Number of times pregnant = low
|    |    |    |    |    riceps skin fold thickness = high: no
|    |    |    |    |    riceps skin fold thickness = low
|    |    |    |    |    |    2-Hour serum insulin = high
```

```
|   |   |   |   |   |   |   Diabetes pedigree function = high: no
|   age = low
|   |   Body mass index = high
|   |   |   riceps skin fold thickness = high
|   |   |   |   Number of times pregnant = high
|   |   |   |   |   Diastolic blood pressure = high
|   |   |   |   |   |   2-Hour serum insulin = high
|   |   |   |   |   |   |   Diabetes pedigree function = low: yes
|   |   |   |   Number of times pregnant = low
|   |   |   |   |   Diabetes pedigree function = high
|   |   |   |   |   |   Diastolic blood pressure = high
|   |   |   |   |   |   |   2-Hour serum insulin = high: no
|   |   |   |   |   |   |   2-Hour serum insulin = low: no
|   |   |   |   |   |   Diastolic blood pressure = low
|   |   |   |   |   |   |   2-Hour serum insulin = high: yes
|   |   |   |   |   |   |   2-Hour serum insulin = low: yes
|   |   |   |   |   Diabetes pedigree function = low
|   |   |   |   |   |   Diastolic blood pressure = high
|   |   |   |   |   |   |   2-Hour serum insulin = high: no
|   |   |   |   |   |   |   2-Hour serum insulin = low: no
|   |   |   |   |   |   Diastolic blood pressure = low
|   |   |   |   |   |   |   2-Hour serum insulin = high: no
|   |   |   |   |   |   |   2-Hour serum insulin = low: no
|   |   |   riceps skin fold thickness = low
|   |   |   |   Diabetes pedigree function = high
|   |   |   |   |   Diastolic blood pressure = high: no
|   |   |   |   |   Diastolic blood pressure = low
|   |   |   |   |   |   2-Hour serum insulin = high
|   |   |   |   |   |   |   Number of times pregnant = low: no
|   |   |   |   |   |   2-Hour serum insulin = low: no
|   |   |   |   Diabetes pedigree function = low: no
|   |   Body mass index = low
|   |   |   Diabetes pedigree function = high
|   |   |   |   2-Hour serum insulin = high: no
|   |   |   |   2-Hour serum insulin = low
|   |   |   |   |   Diastolic blood pressure = high: no
|   |   |   |   |   Diastolic blood pressure = low
|   |   |   |   |   |   Number of times pregnant = low
|   |   |   |   |   |   |   riceps skin fold thickness = low: yes
|   |   |   Diabetes pedigree function = low: no
Plasma glucose concentration a 2 hours in an oral glucose tolerance test = very high
|   2-Hour serum insulin = high
|   |   Body mass index = high
|   |   |   Number of times pregnant = high
|   |   |   |   Diabetes pedigree function = high: yes
|   |   |   |   Diabetes pedigree function = low
```

```
|   |   |   |   |   Diastolic blood pressure = high
|   |   |   |   |   |   riceps skin fold thickness = high
|   |   |   |   |   |   |   age = high: yes
|   |   |   |   |   Diastolic blood pressure = low
|   |   |   |   |   |   riceps skin fold thickness = high
|   |   |   |   |   |   |   age = high: yes
|   |   |   |   Number of times pregnant = low
|   |   |   |   |   age = high
|   |   |   |   |   Diabetes pedigree function = high
|   |   |   |   |   |   riceps skin fold thickness = high
|   |   |   |   |   |   |   Diastolic blood pressure = high: yes
|   |   |   |   |   |   |   Diastolic blood pressure = low: yes
|   |   |   |   |   |   riceps skin fold thickness = low: yes
|   |   |   |   |   Diabetes pedigree function = low
|   |   |   |   |   |   Diastolic blood pressure = high
|   |   |   |   |   |   |   riceps skin fold thickness = high: yes
|   |   |   |   |   |   |   riceps skin fold thickness = low: yes
|   |   |   |   |   |   Diastolic blood pressure = low
|   |   |   |   |   |   |   riceps skin fold thickness = high: yes
|   |   |   |   |   |   |   riceps skin fold thickness = low: yes
|   |   |   |   |   age = low
|   |   |   |   |   |   Diabetes pedigree function = high: yes
|   |   |   |   |   |   Diabetes pedigree function = low
|   |   |   |   |   |   riceps skin fold thickness = high
|   |   |   |   |   |   |   Diastolic blood pressure = high: yes
|   |   |   |   |   |   |   Diastolic blood pressure = low: yes
|   |   |   |   |   |   riceps skin fold thickness = low
|   |   |   |   |   |   |   Diastolic blood pressure = high: yes
|   |   |   |   |   |   |   Diastolic blood pressure = low: no
|   |   |   Body mass index = low
|   |   |   |   age = high
|   |   |   |   |   riceps skin fold thickness = high
|   |   |   |   |   |   Number of times pregnant = high
|   |   |   |   |   |   Diabetes pedigree function = high
|   |   |   |   |   |   |   Diastolic blood pressure = high: yes
|   |   |   |   |   |   |   Diastolic blood pressure = low: yes
|   |   |   |   |   |   Diabetes pedigree function = low: yes
|   |   |   |   |   Number of times pregnant = low
|   |   |   |   |   |   Diastolic blood pressure = high
|   |   |   |   |   |   |   Diabetes pedigree function = low: yes
|   |   |   |   |   |   Diastolic blood pressure = low
|   |   |   |   |   |   |   Diabetes pedigree function = low: yes
|   |   |   |   |   riceps skin fold thickness = low: yes
|   |   |   |   age = low
|   |   |   |   |   Diastolic blood pressure = high
|   |   |   |   |   |   riceps skin fold thickness = high: no
```

```
|    |    |    |    |    riceps skin fold thickness = low
|    |    |    |    |    |    Number of times pregnant = low
|    |    |    |    |    |    |    Diabetes pedigree function = low: yes
|    |    |    |    Diastolic blood pressure = low: no
|    2-Hour serum insulin = low
|    |    Diabetes pedigree function = high: yes
|    |    Diabetes pedigree function = low: no
```

MyDT - CFS

```
Plasma glucose concentration a 2 hours in an oral glucose tolerance test = 'very high'
|    2-Hour serum insulin = high
|    |    Body mass index = high
|    |    |    Age = high
|    |    |    |    Diabetes pedigree function = high: yes
|    |    |    |    Diabetes pedigree function = low: yes
|    |    |    Age = low
|    |    |    |    Diabetes pedigree function = high: yes
|    |    |    |    Diabetes pedigree function = low: yes
|    |    Body mass index = low
|    |    |    Age = high
|    |    |    |    Diabetes pedigree function = high: yes
|    |    |    |    Diabetes pedigree function = low: yes
|    |    |    Age = low
|    |    |    |    Diabetes pedigree function = high: no
|    |    |    |    Diabetes pedigree function = low: no
|    2-Hour serum insulin = low
|    |    Diabetes pedigree function = high: yes
|    |    Diabetes pedigree function = low: no
Plasma glucose concentration a 2 hours in an oral glucose tolerance test = high
|    Body mass index = high
|    |    Age = high
|    |    |    Diabetes pedigree function = high
|    |    |    |    2-Hour serum insulin = high: yes
|    |    |    |    2-Hour serum insulin = low: yes
|    |    |    Diabetes pedigree function = low
|    |    |    |    2-Hour serum insulin = high: yes
|    |    |    |    2-Hour serum insulin = low: yes
|    |    Age = low
|    |    |    2-Hour serum insulin = high
|    |    |    |    Diabetes pedigree function = high: no
|    |    |    |    Diabetes pedigree function = low: no
|    |    |    2-Hour serum insulin = low: no
|    Body mass index = low
|    |    2-Hour serum insulin = high
```

```
|    |    |    Diabetes pedigree function = high: no
|    |    |    Diabetes pedigree function = low
|    |    |    |    Age = high: no
|    |    |    |    Age = low: no
|    |    2-Hour serum insulin = low
|    |    |    Diabetes pedigree function = high
|    |    |    |    Age = high: yes
|    |    |    |    Age = low: no
|    |    |    Diabetes pedigree function = low: no
Plasma glucose concentration a 2 hours in an oral glucose tolerance test = low
|    Body mass index = high
|    |    2-Hour serum insulin = high
|    |    |    Age = high
|    |    |    |    Diabetes pedigree function = high: yes
|    |    |    |    Diabetes pedigree function = low: no
|    |    |    Age = low
|    |    |    |    Diabetes pedigree function = high: no
|    |    |    |    Diabetes pedigree function = low: no
|    |    2-Hour serum insulin = low
|    |    |    Age = high: no
|    |    |    Age = low
|    |    |    |    Diabetes pedigree function = high: no
|    |    |    |    Diabetes pedigree function = low: no
|    Body mass index = low: no
Plasma glucose concentration a 2 hours in an oral glucose tolerance test = medium
|    Age = high
|    |    Body mass index = high
|    |    |    Diabetes pedigree function = high
|    |    |    |    2-Hour serum insulin = high: yes
|    |    |    Diabetes pedigree function = low
|    |    |    |    2-Hour serum insulin = high: no
|    |    |    |    2-Hour serum insulin = low: no
|    |    Body mass index = low
|    |    |    2-Hour serum insulin = high
|    |    |    |    Diabetes pedigree function = high: no
|    |    |    |    Diabetes pedigree function = low: no
|    |    |    2-Hour serum insulin = low: no
|    Age = low
|    |    Body mass index = high
|    |    |    Diabetes pedigree function = high
|    |    |    |    2-Hour serum insulin = high: no
|    |    |    |    2-Hour serum insulin = low: no
|    |    |    Diabetes pedigree function = low
|    |    |    |    2-Hour serum insulin = high: no
|    |    |    |    2-Hour serum insulin = low: no
|    |    Body mass index = low
```

```
|   |   |   Diabetes pedigree function = high
|   |   |   |   2-Hour serum insulin = high: no
|   |   |   |   2-Hour serum insulin = low: no
|   |   |   Diabetes pedigree function = low: no
```

Firstly, the effect of the feature selection (CFS) for each algorithm will be discussed below:

- The first thing to note is that there was no effect on the accuracy of ZeroR and IR algorithms by firstly doing a feature selection. ZeroR wasn't affected by CFS because it ignores all features except the target (which is whether a patient has diabetes or not). 1R wasn't affected by CFS because this one selects the attribute which has the highest predictability (and creates a rule based on that attribute). This one is the same attribute for both before and after CFS.
- 1NN was affected by CFS due to the fact that the nearest neighbor of a data point changes whenever an attribute is removed. This is regardless of the distance heuristic used to classify the points. Also, 5NN wasn't affected by CFS possibly due to the fact that it uses the 5 closest points instead of the closest one which reduces the effect of the reduction of attributes.
- Naive Bayes calculates the probability of all attributes given that the data Both Weka and our implementation of Naive Bayes was affected by CFS due to the fact it calculates the probability of the all the attributes given that a person was tested positive (or negative) for Diabetes.
- Both Weka's and our implementation of decision trees were affected by CFS. This can be explained due to the fact that the tree is build by taking all attributes into consideration. Having a smallest set of attributes would produce a different decision tree.
- There was also a minimal effect of pruning on the accuracy of the decision trees. Also, the decision trees accuracy was notoriously high compared to the other algorithms, suggesting that there was overfitting. WHY WOULDN"T PRUNNING AFFECT MUCH THE ACCURACY ? DISCUSS THE REST OF THE ALGORITHMS (MLP, SVM, RF)

Conclusion

The most appropriate machine learning algorithm for this data set is naive bayes. This is due to its high accuracy compared to the other algorithms and its low likelihood of overfitting the data. Also, note that the CFS selected attributes have low correlation between each other. for example, plasma glucose concentration isn't related to the body mass index. This might explain why naive bayes is so accurate. Additionally, the difference between out implementation and Weka implementation of naive bayes could be due to the difference in breaking the ties.

Reflection