COMP3308 Assignment 1
Predicting Diabetes

Woo Hyun Jung 310250811
Khanh Cao Quoc Nguyen 311253865

1 Aim

The aim of our study is to predict whether a new patient will test positive for diabetes (class 1). The study is important because blah blah dicks

2 Data

2.1 Dataset

The data set we used was the Pima Indians Diabetes Database found at `http://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/`.
There are nine 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 (0 or 1)

and two classes:

1. class1 (testing positive for diabetes)
2. class0 (testing negative for diabetes)

2.2 Data preperation

The raw `.data` file was preprocessed in multiple ways. Firstly, we added a header row to the file and converted to `.csv`. This was necessary to be able to open it in Weka. Before we loaded it, however, we had to change some values in the data as there were missing values (denoted by 0). We wrote a script `csv_scripts\missing_values.py` which took the average of each attribute column and replaced the 0 values. This was done because ...
Once this new file was loaded into Weka, each of the attributes were normalised in the range $[0, 1]$. The final output file was saved as `pima.csv`

2.3 Attribute selection

3 Results and Discussion

4 Conclusions

5 Reflection

6 Instructions