SECOND ASSIGNMENT REPORT: CLASSIFICATION

Introduction to Artificial Intelligence COMP3308
School of Computer Science

THE UNIVERSITY OF
SYDNEY

Students:
Tatenda Chivasa
Alvaro Balvin

Professor: Dr. Irena Koprinska
May 2020

1.-Project Description

This assignment consists on the implementation of two classification algorithms, then evaluation on a real dataset by using a method of stratified cross validation. The algorithms' performance is then compared with external classifiers used on the same dataset but applied by Weka. Lastly we look into the effects of feature selection specifically the Correlation-based Feature Selection method (CFS) from Weka.

1.1.-Aim of the study

The aim of the study is to implement two classification algorithms, the K-Nearest Neighbor and Naïve Bayes algorithms. The K-Nearest Neighbor algorithm is implemented so that it is able to classify for any given value of K. The Naïve Bayes algorithm is implemented for numeric attributes with the use of a probability density function assuming the data is normally distributed. These classifiers will be trained on the given dataset, the Pima Indian Diabetes dataset and then used to predict the classes (yes or no) to which new examples belong. The algorithms are evaluated on the dataset using stratified cross validation. The performance of our classifiers will then be evaluated using 10-fold stratified cross-validation in order to obtain the average accuracy of our classifiers. The 10-fold stratified cross-validation helps us check that our data has not been over-fitted. We will also perform Correlation-based Feature Selection (CFS) on the Pima Indian Diabetes dataset in order to find the attributes that best represent the data and can be used for classification.

1.2.-Importance of the study

This study is important because it familiarizes us with classification which is an important part of supervised learning. Classification is important because it helps us learn more and understand a given dataset. As such, classification has a wide range of applications in many different domains such as spam detection, medical diagnosis and credit card approval among many others. This study in particular helps us explore two kinds of learning in classification. The K-Nearest Neighbor algorithm helps us explore the lazy learning classifiers that store training data until test data appears. The Naïve Bayes algorithm helps us explore the eager learning classifiers that make a classification model prior to receiving test data. Classification data can help us extract useful models and help us predict future trends.

2. - Data description

The dataset used in this study is the Pima Indian Diabetes dataset obtained from the National Institute of Diabetes and Digestive and Kidney Diseases.

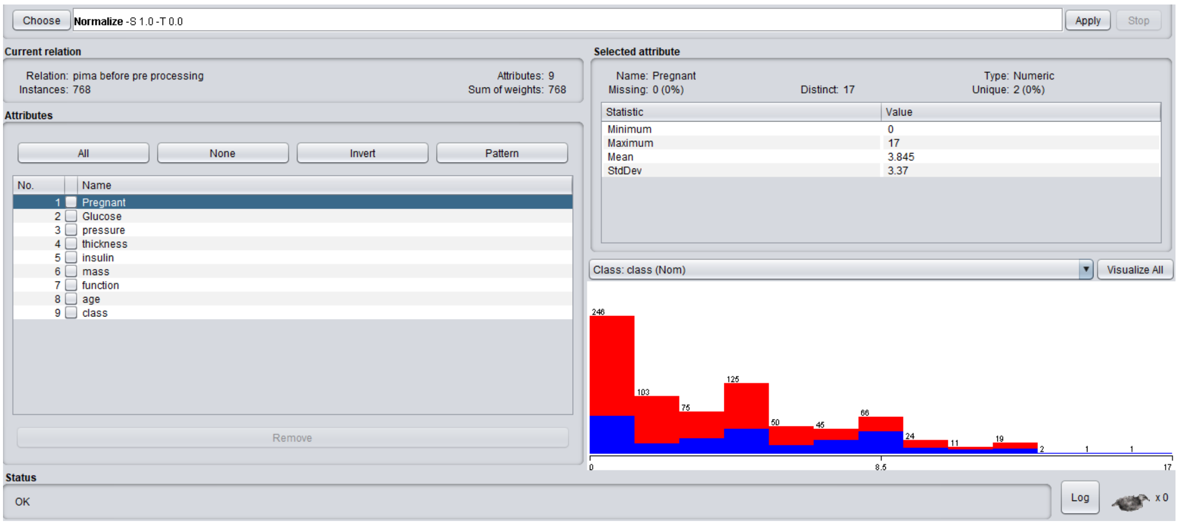
The attributes given from the data are the following:

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")

2.1. - Dataset

The data is obtained from female patients and each entry in the dataset describes patient data, the attributes are measurements taken and tests done on that patient. There are two classes, yes and no to show whether a patient has diabetes or not. The data has 768 entries each with 9 attributes including the class attribute. All the attributes except the class attribute are numeric.



Normalize -S 1.0 -T 0.0

Current relation

Relation: pima before pre processing
Instances: 768
Attributes: 9
Sum of weights: 768

Attributes

No.		Name
1	□	Pregnant
2	□	Glucose
3	□	pressure
4	□	thickness
5	□	insulin
6	□	mass
7	□	function
8	□	age
9	□	class

Selected attribute

Name: Pregnant
Missing: 0 (0%)
Distinct: 17
Type: Numeric
Unique: 2 (0%)

Statistic	Value
Minimum	0
Maximum	17
Mean	3.845
StdDev	3.37

Class: class (Nom)

Status

OK

Figure 1.-Data inserted in the statistics software to proceed with the normalization

2.2. - Dataset normalization

In order to make the coding for the classifiers easier and efficient, the data acquired was normalized. In the normalization, the values for the attributes will be changed and put into a range of [0:1]. Consequently, the methods to classify the data will also vary from the procedures that would take place if the data was, for example, qualitative.

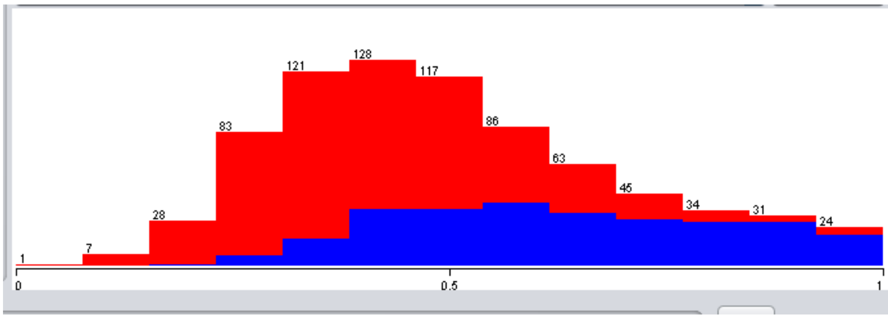


Figure 2.-Example of the processed data from the Glucose attributes

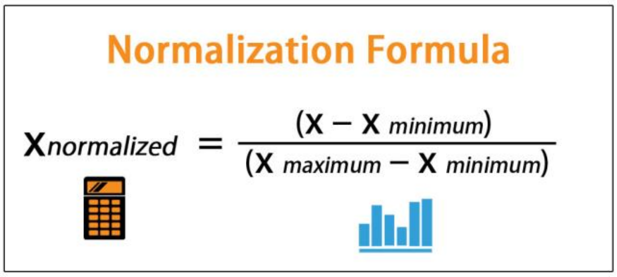


Figure 3.-Normalization Formula shown in the picture adobe

2.2. - The Correlation-based Feature Selection method (CFS) summary

This process is done in machine learning in order to simplify the operations, which also helps reducing the training times, by removing the data (in this case, attributes), that are consider irrelevant or redundant.

An attribute is considered admissible when is hardly correlated with other attributes but is strongly correlated with the class to predict. That indicates that in can help predict the class of the object selected with more accuracy without relaying in the other attributes with predominance.

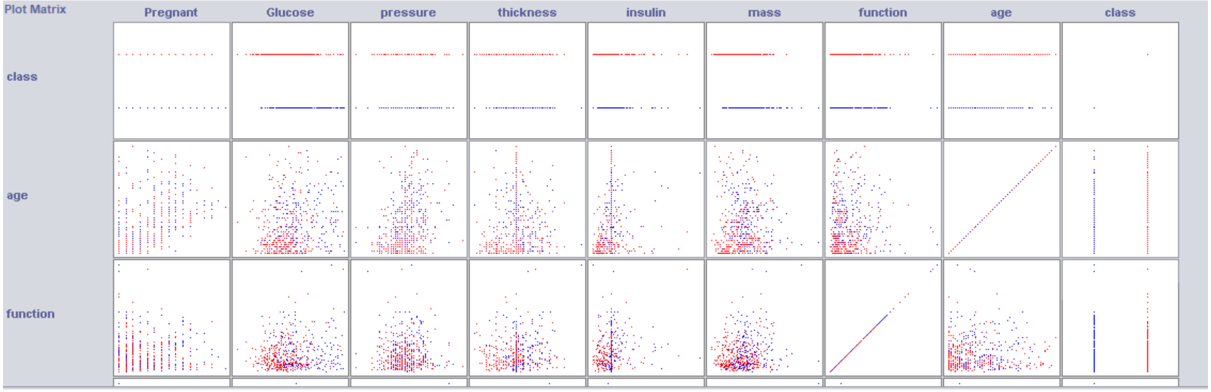


Figure 4.-Weka shows the relationship between the attributes

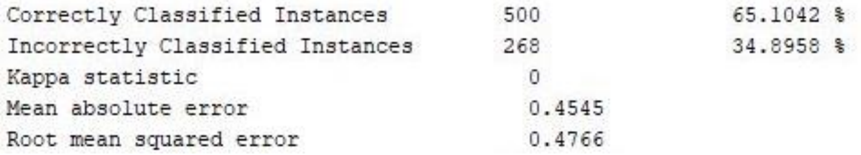
3.-Results and discussion

In Weka, with 10-fold cross validation selected as the way of testing the accuracy, we run the following algorithms: ZeroR, 1R, k-Nearest Neighbor (k-NN; IBk in Weka), Naïve Bayes (NB), Decision Tree (DT; J48 in Weka), Multi-Layer Perceptron (MLP), Support Vector Machine (SVM; SMO in Weka) and Random Forest (RF).

In the next lines, Weka results are presented in order to analyze the benefits of each classifier technique and compare the data obtained with the normal data and the selected data with de CFS.

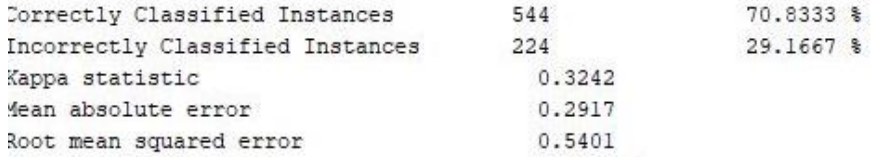
3.1.-Performance of Weka's classifiers without feature selection

3.1.1.-ZeroR Classifier without feature selection



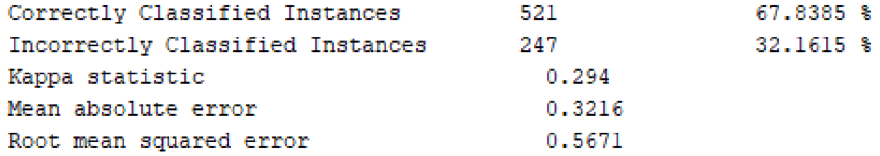
```
Correctly Classified Instances         500               65.1042 %
Incorrectly Classified Instances       268               34.8958 %
Kappa statistic                          0
Mean absolute error                      0.4545
Root mean squared error                  0.4766
```

3.1.2.-OneR Classifier without feature selection



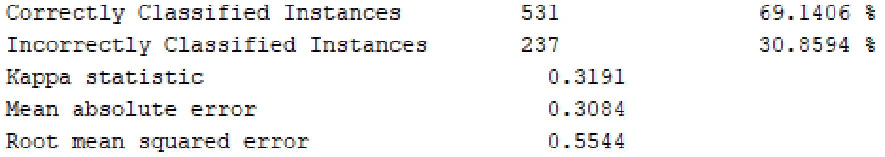
```
Correctly Classified Instances         544               70.8333 %
Incorrectly Classified Instances       224               29.1667 %
Kappa statistic                          0.3242
Mean absolute error                      0.2917
Root mean squared error                  0.5401
```

3.1.3.-k-Nearest Neighbor Classifier without feature selection (1NN)



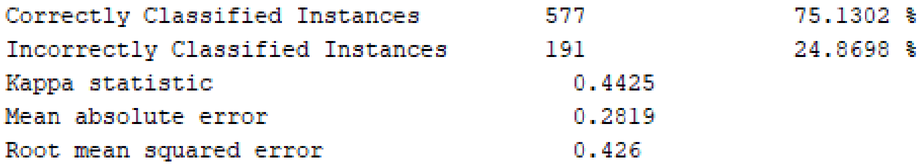
```
Correctly Classified Instances         521               67.8385 %
Incorrectly Classified Instances       247               32.1615 %
Kappa statistic                          0.294
Mean absolute error                      0.3216
Root mean squared error                  0.5671
```

3.1.4.-k-Nearest Neighbor Classifier without feature selection (5NN)



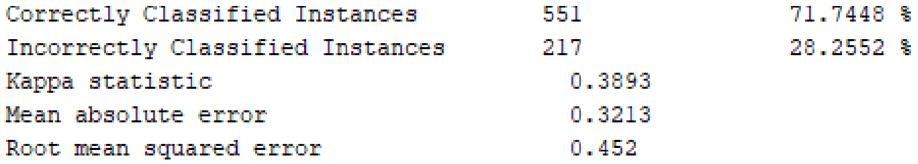
```
Correctly Classified Instances         531               69.1406 %
Incorrectly Classified Instances       237               30.8594 %
Kappa statistic                          0.3191
Mean absolute error                      0.3084
Root mean squared error                  0.5544
```

3.1.5.- Naïve Bayes Classifier without feature selection (NB)



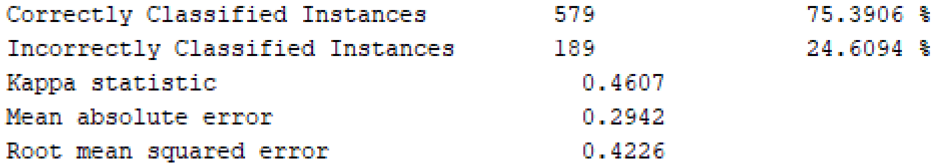
```
Correctly Classified Instances         577               75.1302 %
Incorrectly Classified Instances       191               24.8698 %
Kappa statistic                          0.4425
Mean absolute error                      0.2819
Root mean squared error                  0.426
```

3.1.6.- Decision Tree Classifier without feature selection (DT)



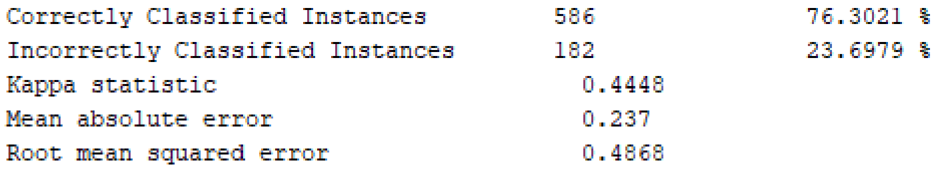
```
Correctly Classified Instances         551               71.7448 %
Incorrectly Classified Instances       217               28.2552 %
Kappa statistic                          0.3893
Mean absolute error                      0.3213
Root mean squared error                  0.452
```

3.1.7.-Multi-Layer Perceptron Classifier without feature selection (MLP)



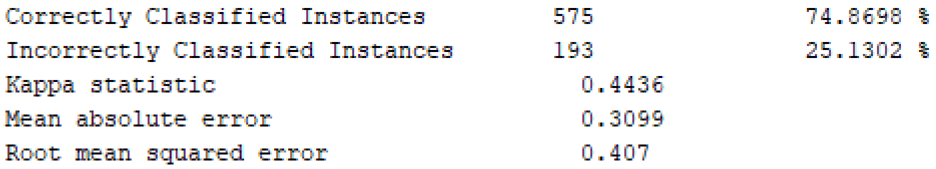
```
Correctly Classified Instances         579               75.3906 %
Incorrectly Classified Instances       189               24.6094 %
Kappa statistic                          0.4607
Mean absolute error                      0.2942
Root mean squared error                  0.4226
```

3.1.8.-Support Vector Machine Classifier without feature selection (SVM)



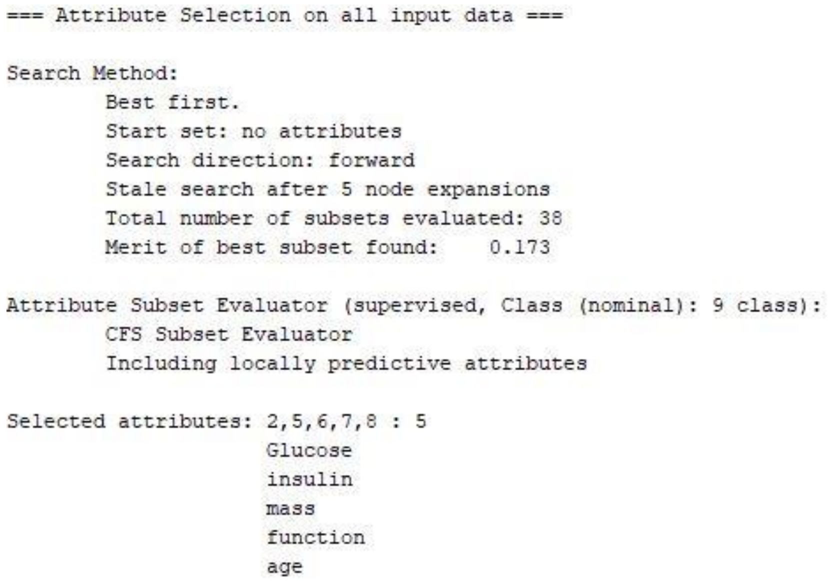
```
Correctly Classified Instances         586               76.3021 %
Incorrectly Classified Instances       182               23.6979 %
Kappa statistic                          0.4448
Mean absolute error                      0.237
Root mean squared error                  0.4868
```

3.1.9.-Random Forest Classifier without feature selection (RF)



```
Correctly Classified Instances         575               74.8698 %
Incorrectly Classified Instances       193               25.1302 %
Kappa statistic                          0.4436
Mean absolute error                      0.3099
Root mean squared error                  0.407
```

3.2.- Performance of Weka's classifiers with feature selection



```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 38
        Merit of best subset found:    0.173

Attribute Subset Evaluator (supervised, Class (nominal): 9 class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,5,6,7,8 : 5
                     Glucose
                     insulin
                     mass
                     function
                     age
```

Figure 5.-Weka shows the attributes selected after de CFS

The figure above shows the attributes that were selected with the feature selection algorithm as the recommended subset to predict the class of the dataset. With these 5 designated attributes, the classifiers are run again to compare the accuracy results of these algorithms with and without the feature selection.

	ZeroR	1R	1NN	5NN	NB	DT	MPL	SVM	RF
No feature selection	65.1%	70.8%	67.8%	69.1%	75.1%	71.7%	75.4%	76.3%	74.8%
CFS	65.1%	70.8%	69.0%	74.5%	76.3%	73.3%	75.8%	76.7%	75.9%

	My1NN	My5NN	MyNB
No feature selection	92%	91%	93%
CFS	93%	92%	93%

By comparing the accuracy done by Weka in their classifiers, we observe that the accuracy is even or greater with the data processed by the CFS algorithm for both cases.

4.-Conclusion

4.1.-Conclusions based on the results

We can conclude that the accuracy didn't have any important changes while applying the CFS, so its better to processed the data this way.

4.2.-Future Work

In the future we suggest that classification methods that use different kinds of learning can be used on the data set such as decision trees or neural networks or more classification methods can be used. This would allow exploration of unsupervised learning algorithms on the dataset for which we would have to remove the class labels on the given dataset. Furthermore it would allow a comparison of performance between two algorithms that use different kinds of learning, supervised and unsupervised.

5.-Reflection

5.1. - Lessons

Throughout this assignment we learnt to apply two classification algorithms to a given dataset using the stratified cross validation method. We learnt how to apply the probability density function to the numeric attributes for Naïve Bayes classification. In addition to this we learnt to apply feature selection to the given dataset in order to investigate which attributes contributed the most to our prediction of the class in which new data belonged to.