Weka Experiment
Prepared by: **Arvy Olarcos Llave**

Dataset Reference:

https://github.com/renatopp/arff-datasets/blob/master/classification/diabetes.arff

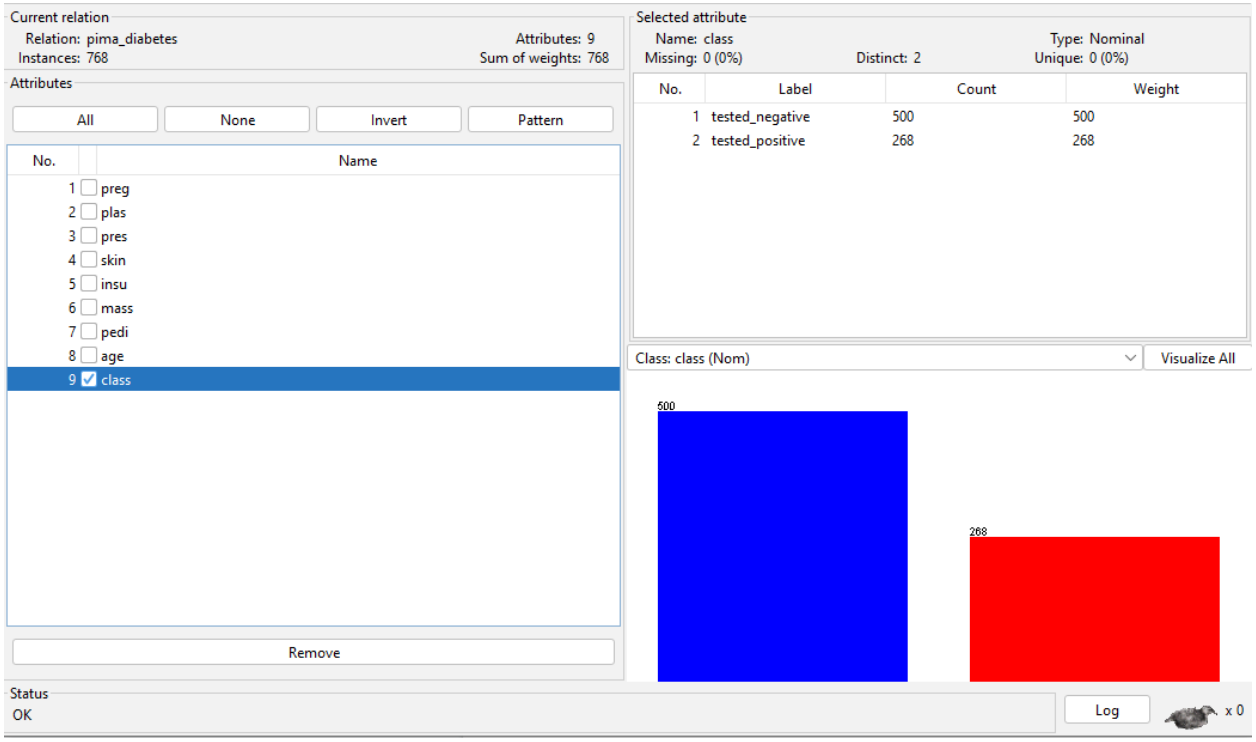
The dataset used in the experiment is the Pima Indians Diabetes dataset. The dataset will predict whether the patient is prone to be diabetic. The patients in this dataset are all females of at least 21 years of age from Pima Indian Heritage. The Dataset has 768 instances and 8 numerical attributes plus a class. For each attribute all are numeric-valued and the output variable predicted is nominal, comprising two classes. The dataset contains no missing values.

Attributes:

1. **preg -** number of times pregnant
2. **plas -** Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. **pres -** Diastolic blood pressure (mm Hg)
4. **skin -** Triceps skin fold thickness (mm)
5. **snsu -** 2 Hour serum insulin (mu U/ml)
6. **mass -** Body mass index (weight in kg/(height in m)^2)
7. **pedi -** Diabetes pedigree function
8. **age -** Age (years)
9. **class -** Class variable (0 or 1)

Class Distribution: (class value 1 is interpreted as "tested positive for diabetes")

Class Value	Number of Instances
0	500
1	268



Brief Statistical Analysis:

Attribute Number	Mean	Standard Deviation
1	3.8	3.4
2	120.9	32.0
3	69.1	19.4
4	20.5	16.0
5	79.8	115.2
6	32.0	7.9
7	0.5	0.3
8	33.2	11.8

Results: Using the Different Classification Techniques

The experiment uses four Classification Algorithms which are Random Forest, J48, Naive Bayes and Support Vector Machine. Under the training of the four algorithms, the 10 fold cross-validation is selected as an evaluation approach.

Random Forest

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         582               75.7813 %
Incorrectly Classified Instances       186               24.2188 %
Kappa statistic                          0.4566
Mean absolute error                      0.3106
Root mean squared error                  0.4031
Relative absolute error                 68.3405 %
Root relative squared error             84.5604 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.836    0.388    0.801      0.836    0.818      0.458    0.820     0.886     tested_negative
                 0.612    0.164    0.667      0.612    0.638      0.458    0.820     0.679     tested_positive
Weighted Avg.    0.758    0.310    0.754      0.758    0.755      0.458    0.820     0.814

=== Confusion Matrix ===

   a   b   <-- classified as
 418  82 |   a = tested_negative
 104 164 |   b = tested_positive
```

Figure 1: Summary of the Results of Random Forest Classifier

J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         567               73.8281 %
Incorrectly Classified Instances       201               26.1719 %
Kappa statistic                          0.4164
Mean absolute error                      0.3158
Root mean squared error                  0.4463
Relative absolute error                 69.4841 %
Root relative squared error             93.6293 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.814    0.403    0.790      0.814    0.802      0.417    0.751     0.811     tested_negative
                 0.597    0.186    0.632      0.597    0.614      0.417    0.751     0.572     tested_positive
Weighted Avg.    0.738    0.327    0.735      0.738    0.736      0.417    0.751     0.727

=== Confusion Matrix ===

   a   b   <-- classified as
 407  93 |   a = tested_negative
 108 160 |   b = tested_positive
```

Figure 2: Summary of the Results of J48 Classifier

Naive Bayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         586               76.3021 %
Incorrectly Classified Instances       182               23.6979 %
Kappa statistic                          0.4664
Mean absolute error                      0.2841
Root mean squared error                  0.4168
Relative absolute error                 62.5028 %
Root relative squared error             87.4349 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.844    0.388    0.802      0.844    0.823      0.468    0.819     0.892     tested_negative
                 0.612    0.156    0.678      0.612    0.643      0.468    0.819     0.671     tested_positive
Weighted Avg.    0.763    0.307    0.759      0.763    0.760      0.468    0.819     0.815

=== Confusion Matrix ===

   a   b   <-- classified as
 422  78 |   a = tested_negative
 104 164 |   b = tested_positive
```

Figure 3: Summary of the Results of Naive Bayes Classifier

Support Vector Machine

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         594               77.3438 %
Incorrectly Classified Instances       174               22.6563 %
Kappa statistic                          0.4682
Mean absolute error                      0.2266
Root mean squared error                  0.476
Relative absolute error                 49.848  %
Root relative squared error             99.862  %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.898    0.459    0.785      0.898    0.838      0.480    0.720     0.771     tested_negative
                 0.541    0.102    0.740      0.541    0.625      0.480    0.720     0.560     tested_positive
Weighted Avg.    0.773    0.334    0.769      0.773    0.763      0.480    0.720     0.698

=== Confusion Matrix ===

   a   b   <-- classified as
 449  51 |   a = tested_negative
 123 145 |   b = tested_positive
```

Figure 4: Summary of the Results of SVM Classifier

Analysis of the Results

Summing all values of the class distribution with 768 instances to compare all the Classification Algorithms the Random Forest correctly classified 582 instances with 186 incorrectly classified instances. The J48 correctly classified 567 instances with 201 incorrectly classified instances. The Naive Bayes correctly classified 586 instances with 182 incorrectly classified instances. Lastly, the SVM correctly classified 594 instances with 174 incorrectly classified instances.

Detailed Accuracy of the Class: **Tested Positive**

	Random Forest	**J48**	**Naive Bayes**	**SVM**
Accuracy*(Correctly Classified Instances)*	75.78%	73.83%	76.30%	77.34%
Recall	61.2%	59.7%	61.2%	54.1%
Precision	66.7%	63.2%	67.8%	74%
F-Measure	63.8%	73.6%	64.3%	62.5%

Detailed Accuracy of the Class: **Tested Negative**

	Random Forest	**J48**	**Naive Bayes**	**SVM**
Accuracy*(Incorrectly Classified Instances)*	24.22%	26.17%	23.70%	22.66%
Recall	83.6%	81.4%	84.4%	90%
Precision	80.1%	79%	80.2%	79%
F-Measure	81.8%	80.2%	82.3%	84%

Conclusions

In the experiment, the classification of diabetes dataset applied four classification algorithms which are Random Forest, J48, Naive Bayes and SVM. The four algorithms are used based on the performance factors classification accuracy to know the most effective classification technique in predicting if a patient is prone to diabetes. The result of the experiment concluded that the best Classification algorithm is Support Vector Machine algorithm with 77.34% accuracy rate.