**Neural and Evolutionary Computation (NEC)**

**A2: Classification with SVM, BP and MLR**

Professors: SERGIO GÓMEZ JIMÉNEZ, JORDI DUCH GAVALDÀ

Students: SEBASTIAN BUZDUGAN, ERIKA LOZADA MARTINEZ

**Main Concepts:**

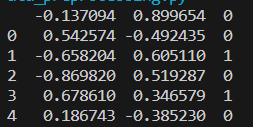
The task involves data classification using Support Vector Machines (SVM), Back-Propagation (BP), and Multiple Linear Regression (MLR) on different datasets. Key aspects include parameter selection via cross-validation, data preprocessing, normalization, and evaluation through error rates, confusion matrices, ROC curves, and AUC. The assignment emphasizes finding optimal parameters for each algorithm, handling both binary and multifeatured datasets, and includes comprehensive documentation of methods and analyses.

**Part 1: Selecting and analyzing the datasets**

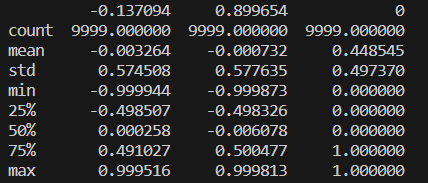
**Ring dataset**

In this dataset the two first columns are the input and the last one is the classifier:

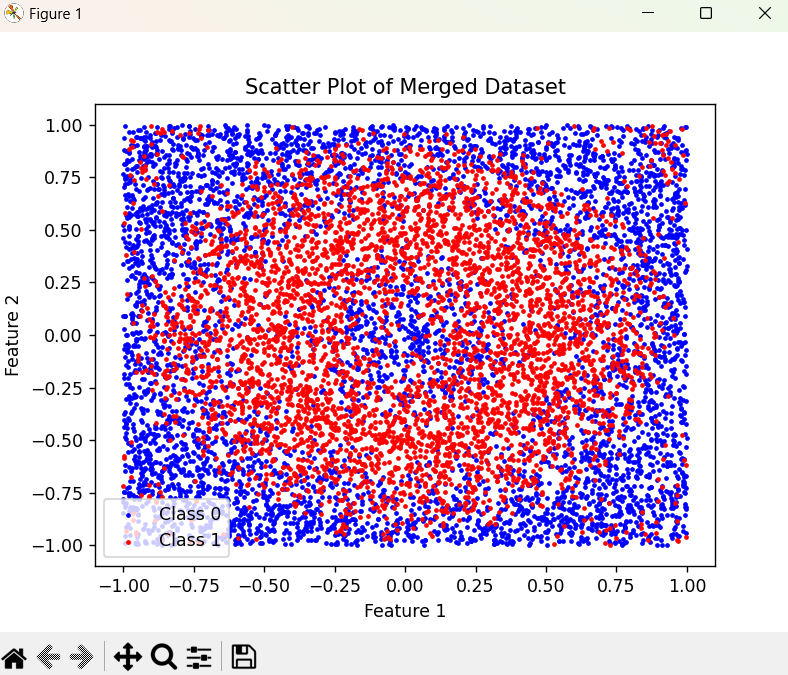
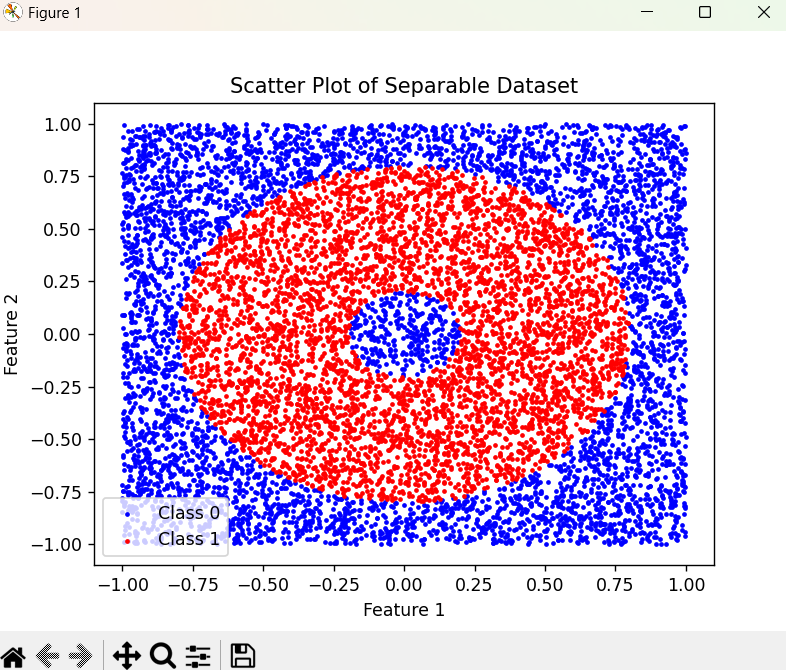
The head of the dataset is the following:



The descriptive statistics of the dataset are the following:



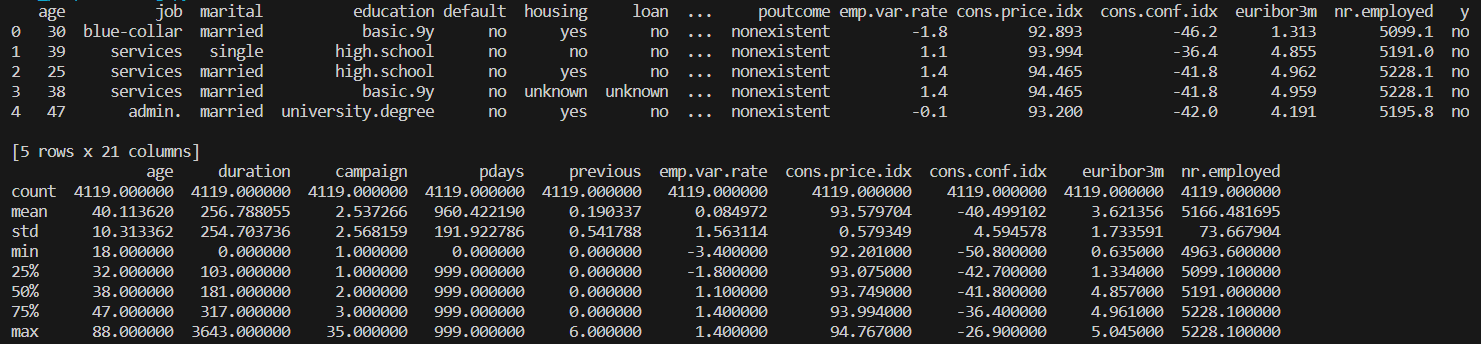
Data seems to be normalized. Here below there is the plot of the ring dataset:



**Bank dataset**

This dataset has 20 features, the first 19 columns are input and the last one is classifier for prediction. From this dataset the chosen file was the bank-additional.csv.

The head of the dataset and the descriptive statistics are the following:

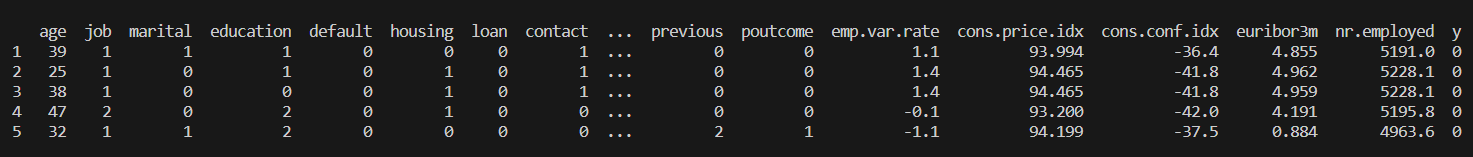
****

The variables of the dataset are:

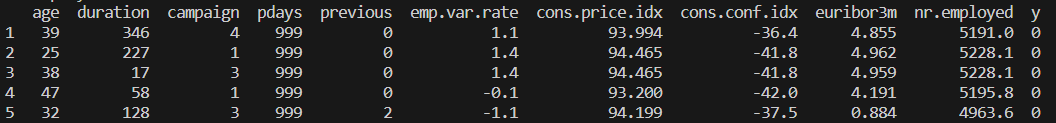
* "age": numerical
* "job":categorical
* "marital":categorical
* "education":categorical
* "default":categorical
* "housing"categorical
* "loan":categorical
* "contact":categorical
* "month":categorical
* "day\_of\_week":categorical
* "duration":numerical
* "campaign":numerical
* "pdays":numerical
* "previous":numerical
* "poutcome":categorical
* "emp.var.rate":numerical
* "cons.price.idx":numerical
* "cons.conf.idx":numerical
* "euribor3m":numerical
* "nr.employed":numerical
* "y":categorical

In this dataset there were unknown values that were filled with the mode of the column. The yes/no column was converted to numerical values that represent the clases, yes values were converted to 1 and no to 0. The rest of categorical columns were converted to numbers.

Then, the processed dataset is the following:



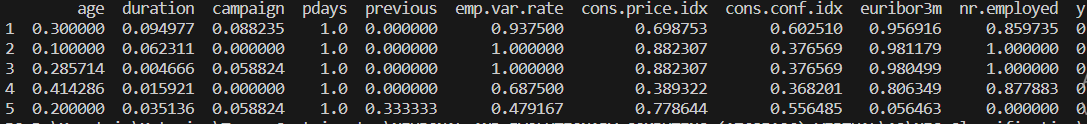
Since, the last column is using as classifier, the rest of categorical variables of the dataset are not considered.

****

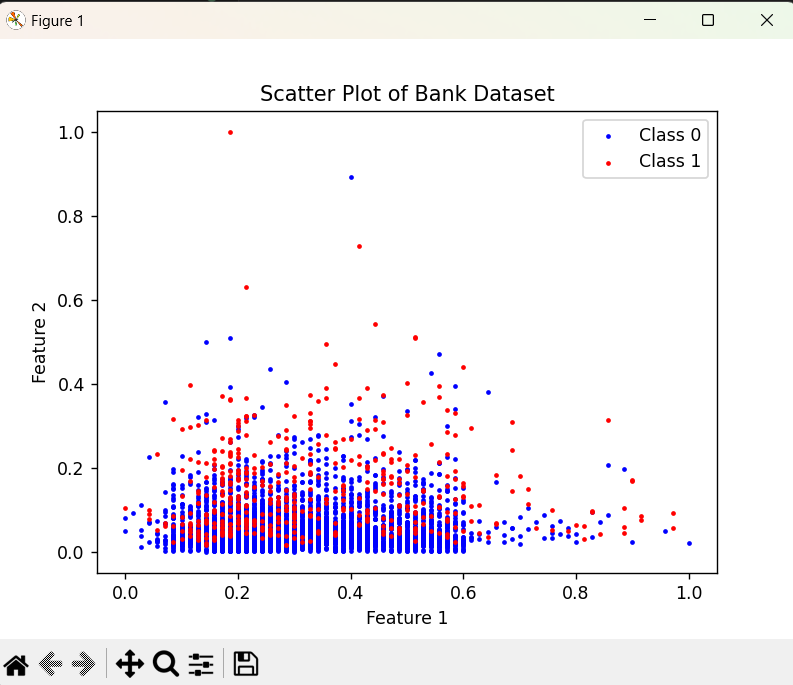
Finally data was normalized as follows:

* age: There are positive values that vary widely, so it will be normalized with Min-Max scaling between 0 and 1.
* duration: There are positive values that vary widely, so it will be normalized with Min-Max scaling between 0 and 1.
* campaign: There are positive values that vary widely, so it will be normalized with Min-Max scaling between 0 and 1.
* pdays: There are positive values that vary widely, so it will be normalized with Min-Max scaling between 0 and 1.
* previous: There are positive values that vary widely, so it will be normalized with Min-Max scaling between 0 and 1.
* emp.var.rate: There are positive values that vary widely, so it will be normalized with Min-Max scaling between 0 and 1.
* cons.price.idx: There are positive values that vary widely, so it will be normalized with Min-Max scaling between 0 and 1.
* cons.conf.idx: There are positive values that vary widely, so it will be normalized with Min-Max scaling between 0 and 1.
* euribor3m: There are positive values that vary widely, so it will be normalized with Min-Max scaling between 0 and 1.
* nr.employed: There are positive values that vary widely, so it will be normalized with Min-Max scaling between 0 and 1.

The head of the resulting dataset is:



Here below there is the plot of the Banknote dataset:

****

**Personalized dataset: Banknote authentication**

**Dataset variables**

The personalized dataset was obtained from the Univesity of California, Irvine Machine Learning Repository. This dataset contains data extracted from images that were taken for the evaluation of an authentication procedure for banknotes. The dataset was downloaded from <https://archive.ics.uci.edu/dataset/267/banknote+authentication>.

1. variance of Wavelet Transformed image (continuous)

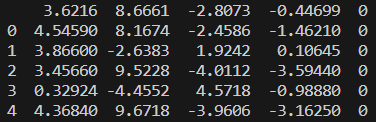
2. skewness of Wavelet Transformed image (continuous)

3. curtosis of Wavelet Transformed image (continuous)

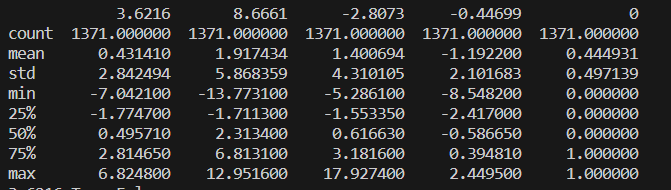
4. entropy of image (continuous)

5. class (integer)

Here below there is the head of the dataset:



In this dataset the first four columns are the input and the last one is the classification one. The following image contains the descriptive statistics of the dataset:

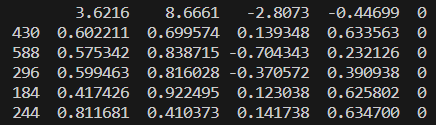


Then data normalization was applied in numerical variables as follows:

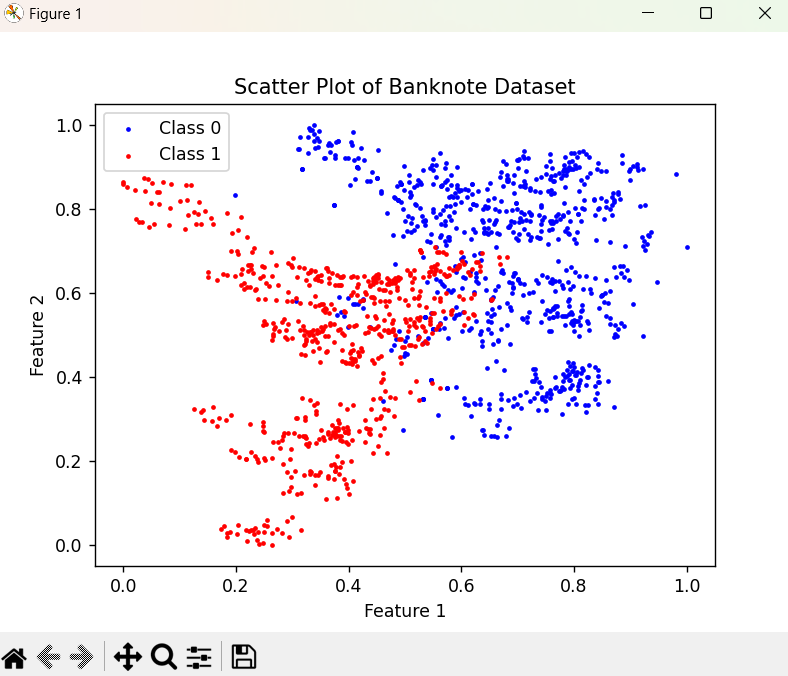
* Variance: There are positive values that vary widely, so it will be normalized with Min-Max scaling between 0 and 1.
* Skewness: There are positive values that vary widely, so it will be normalized with Min-Max scaling between 0 and 1.
* Curtosis: It follows normal distribution so Z-score normalization will be applied.
* Entropy: It follows normal distribution so Z-score normalization will be applied.
* Class: No normalization is needed for this variable since it is categorical.

Data was shuffled to avoid sorting,

The head of the dataset after normalization is the following:



Here below there is the plot of the Banknote dataset:



**Part 2: Classification problem**

1. **Support Vector Machines (SVM)**

Support Vector Machines (SVM) are a supervised learning algorithm primarily used for classification tasks. They work by finding a hyperplane in a multi-dimensional space to distinctly classify data points. SVM is particularly effective due to its ability to use kernel functions, making it capable of handling both linear and non-linear data.

The SVM model has the following parameters:

* C: Is the regularization.
* kernel: Specifies the kernel type that can be 'linear', 'poly', 'rbf', 'sigmoid' and 'precomputed'.
* gamma: Kernel coefficient influences the shape of the decision boundary.
* degree: Degree of the polynomial kernel function.
* coef0: Independent term in the kernel function used in 'poly' and 'sigmoid'.
* shrinking: Whether to use the shrinking heuristic to speed up the training process.
* probability: Whether to enable probability estimates.
* tol: Tolerance for stopping criterion.
* class\_weight: Weights associated with classes.
* max\_iter: Hard limit on iterations within solver.

**Ring dataset**

The supervised training of the dataset was performed for the following parameters:

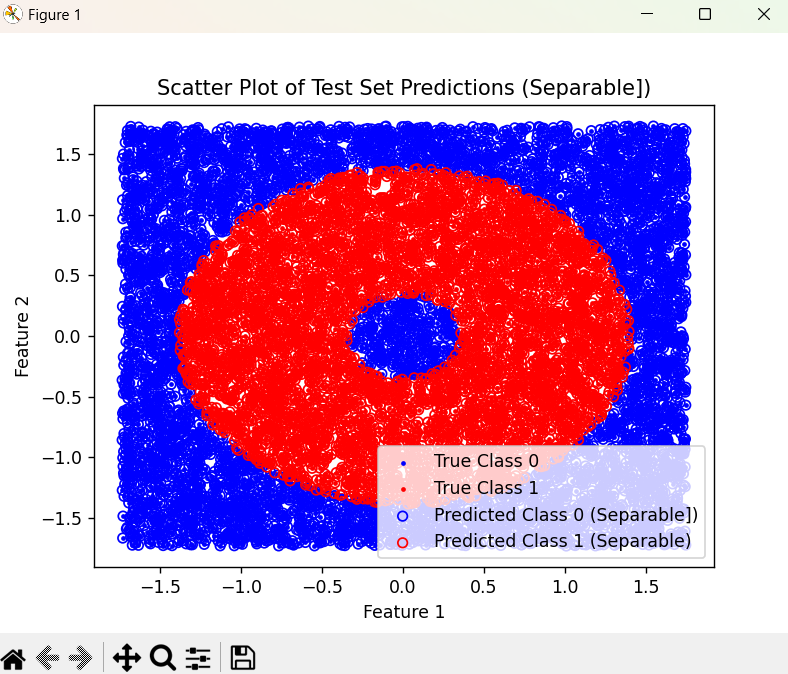
* Kernel:'linear', 'poly', 'rbf', 'sigmoid', 'precomputed'
* Constant:100,200,300,400,500,600,700,800,900

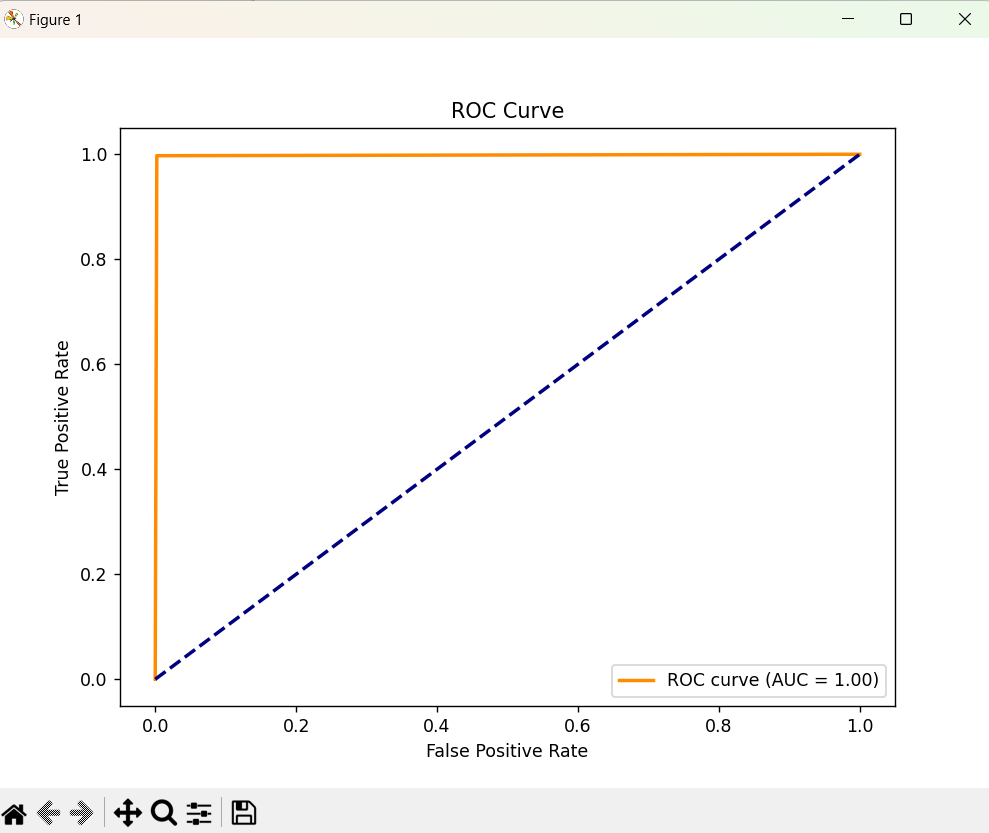
From the execution it was possible to identify that the parameters in which the accuracy is the maximun and the classification error is the minimum is:

* Kernel: rbf
* Constant:701

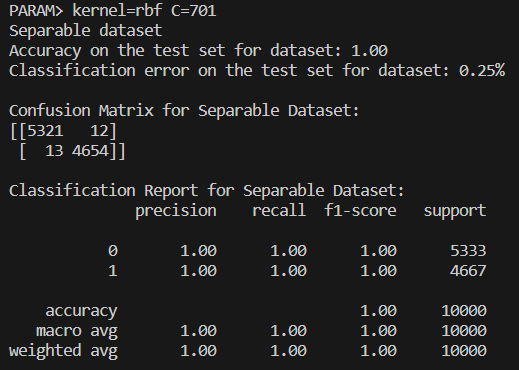
With the best parameters the prediction was performed with the 80% of the dataset for training and the other 20% for test:

* Separable dataset

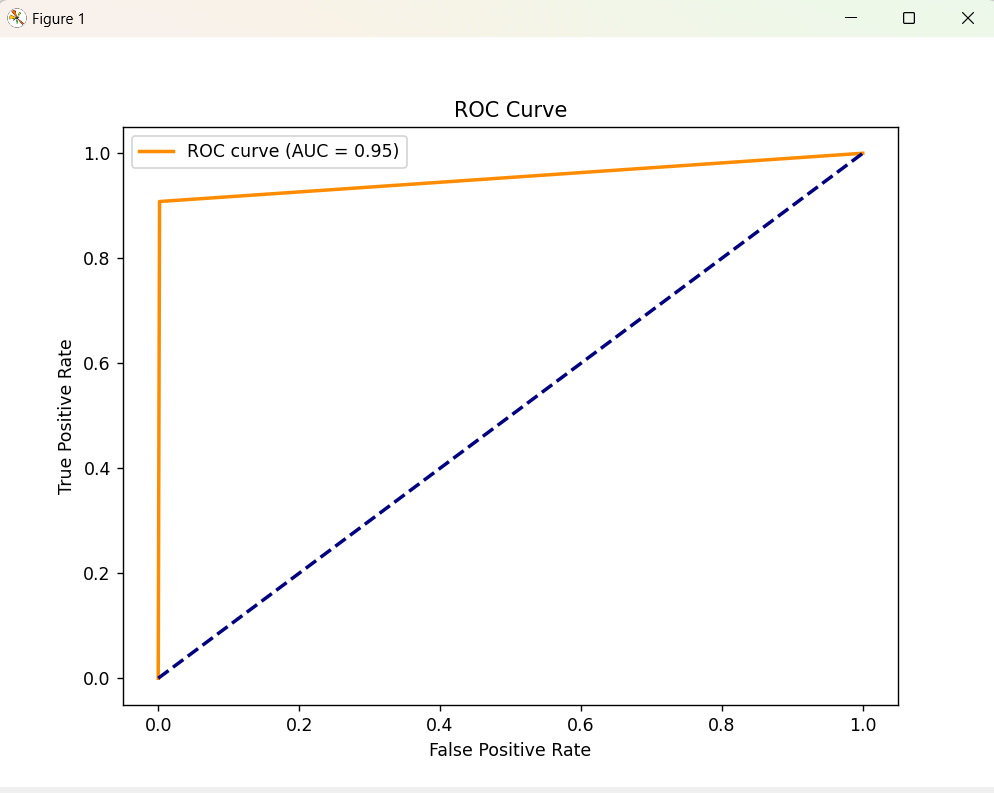
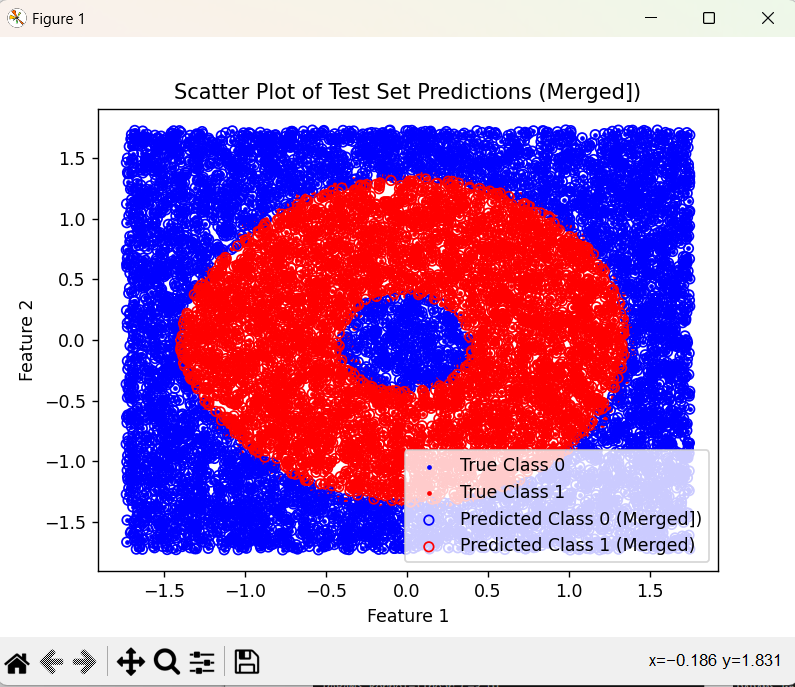




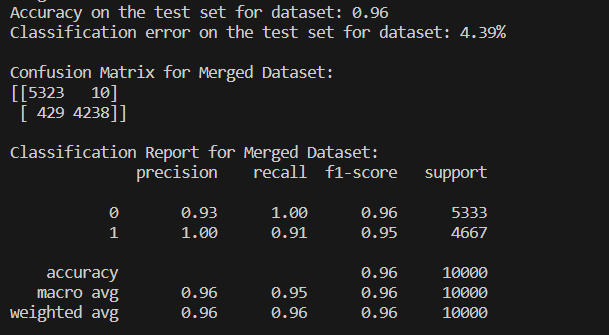




* Merged dataset:



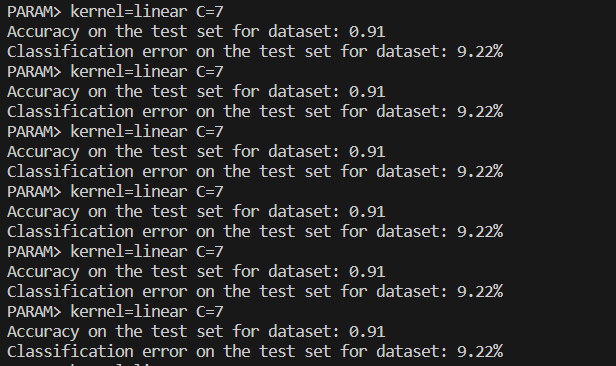


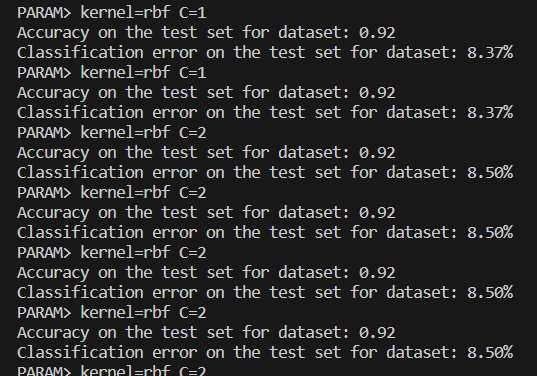
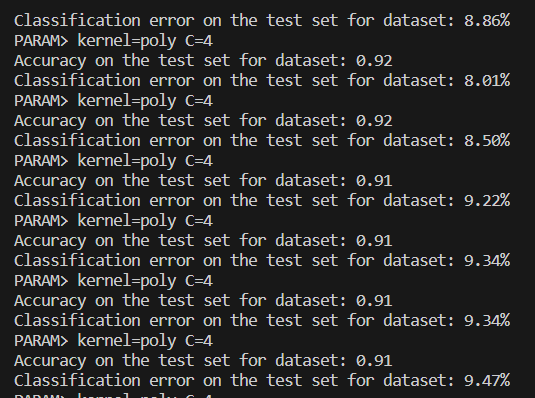


**Bank dataset**

The supervised training of the dataset was performed for the following parameters:

* Kernel:'linear', 'poly', 'rbf', 'sigmoid', 'precomputed'
* Constant:1,2,3,4,5,6,7,8,9,10
* Degree:1,2,3,45

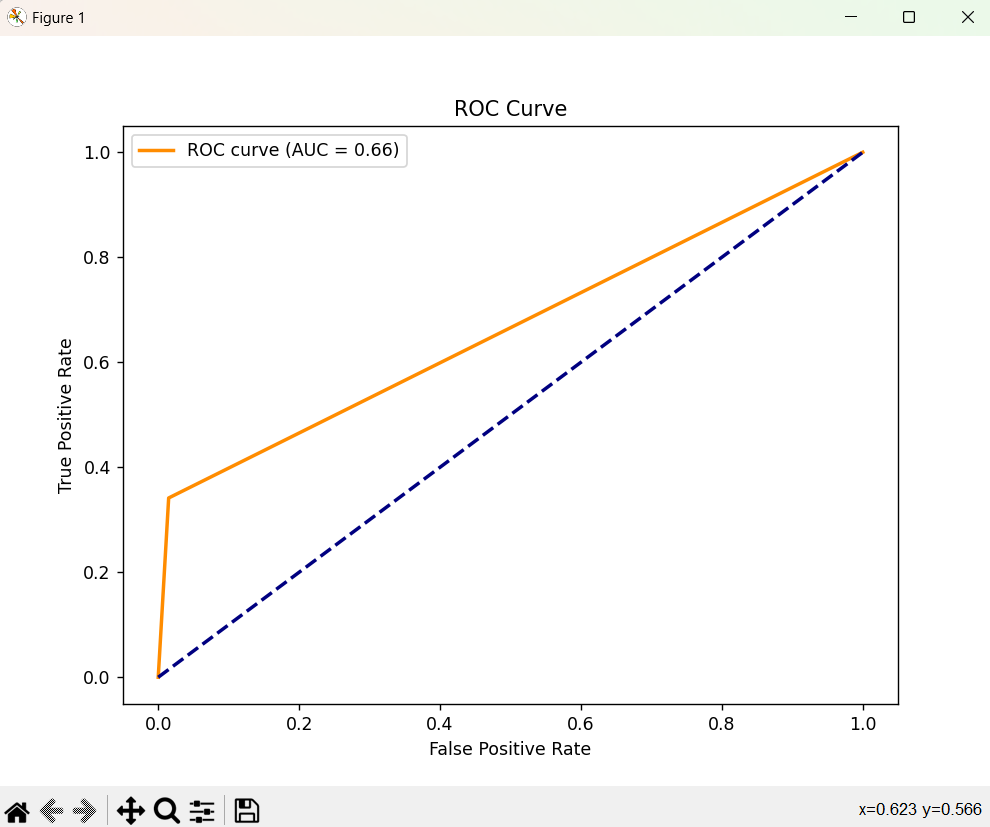


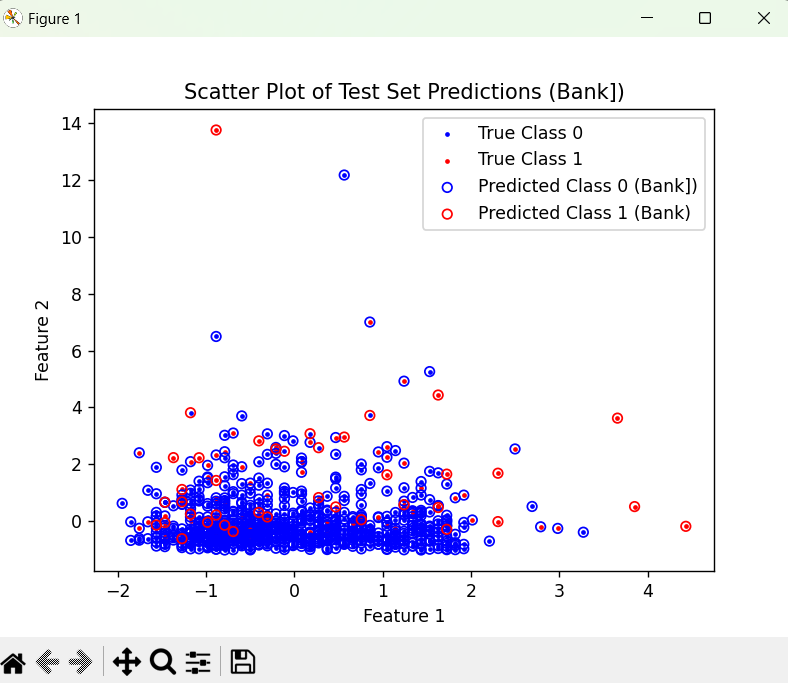


From the execution it was possible to identify that the parameters in which the accuracy is the maximun and the classification error is the minimum is:

* Kernel: poly
* Constant:1
* Degree:4

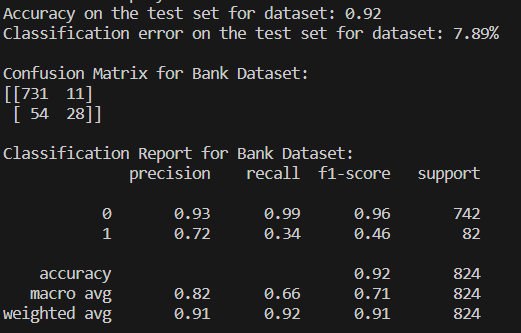
With the best parameters the prediction was performed with the 80% of the dataset for training and the other 20% for test:





The accuracy, the classification error, the confusion matrix and the classification report from the execution of the supervised training with the best parameters is the following:

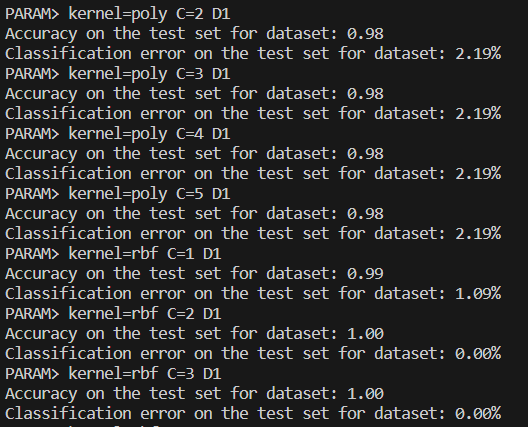
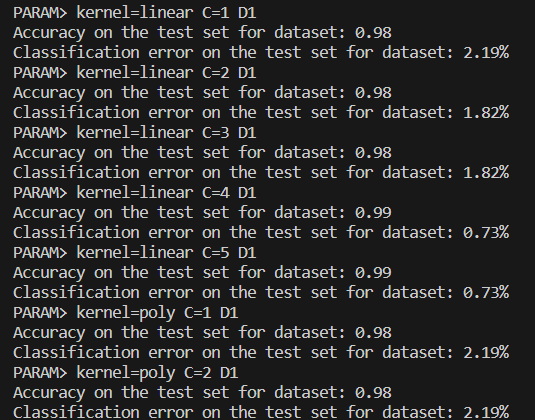


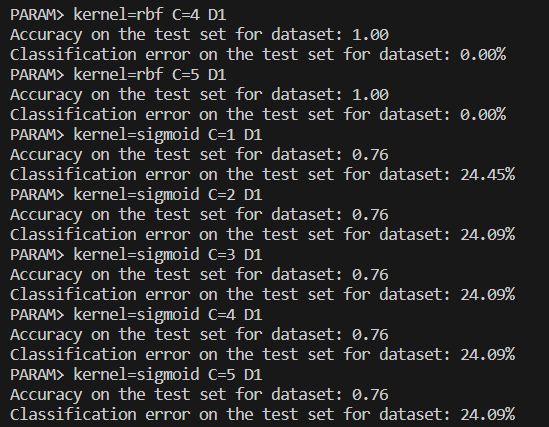


**Banknote dataset**

The supervised training of the dataset was performed for the following parameters:

* Kernel:'linear', 'poly', 'rbf', 'sigmoid', 'precomputed'
* Constant:1,2,3,4,5,6,7,8,9,10
* Degree:1,2,3

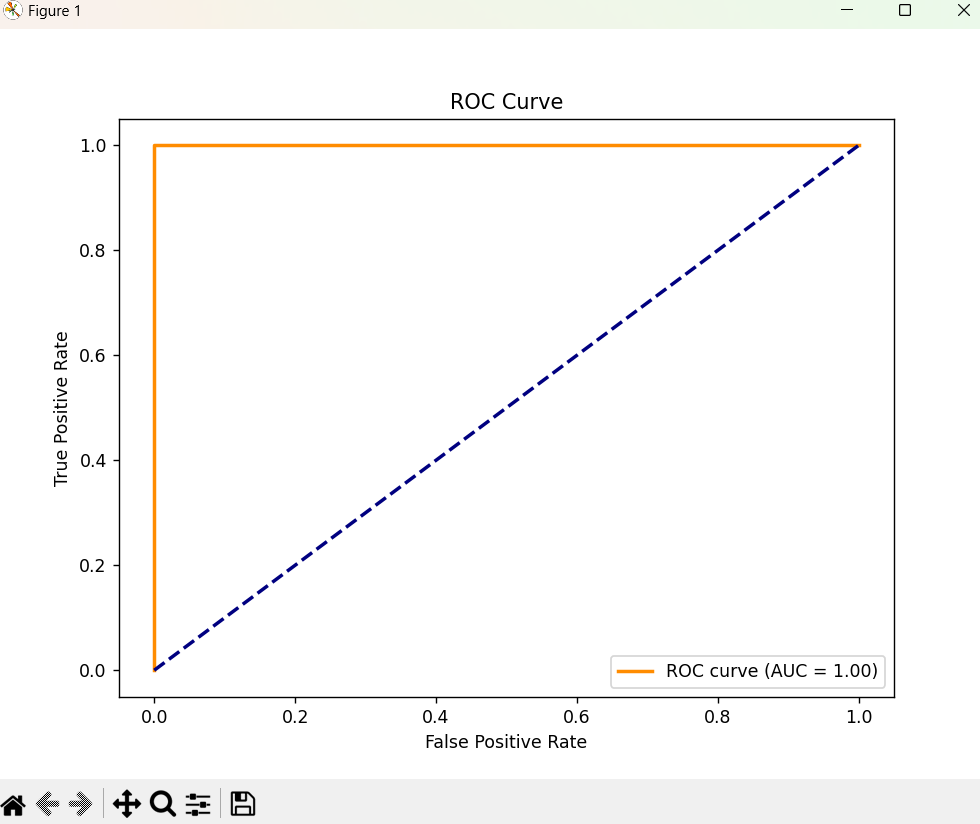
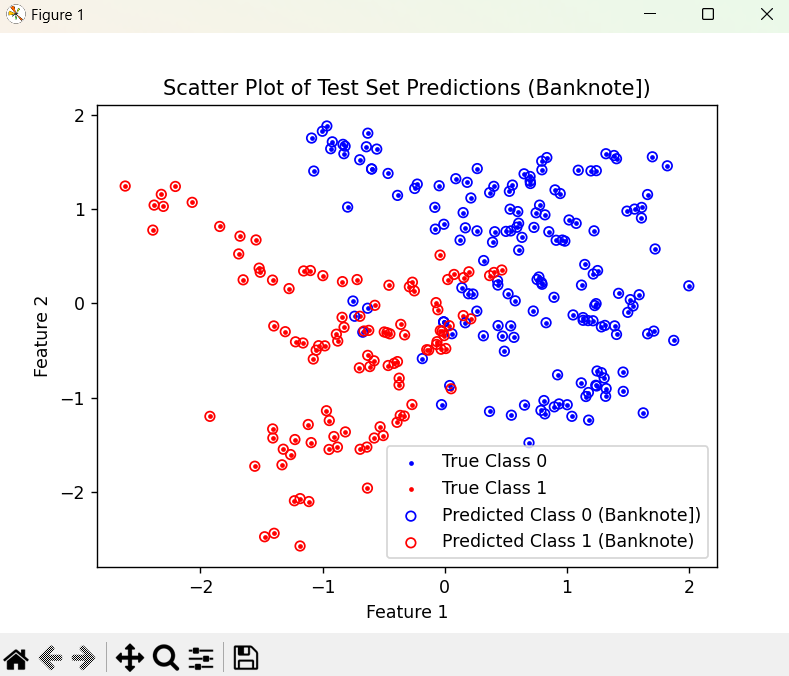




From the execution it was possible to identify that the parameters in which the accuracy is the maximun and the classification error is the minimum is:

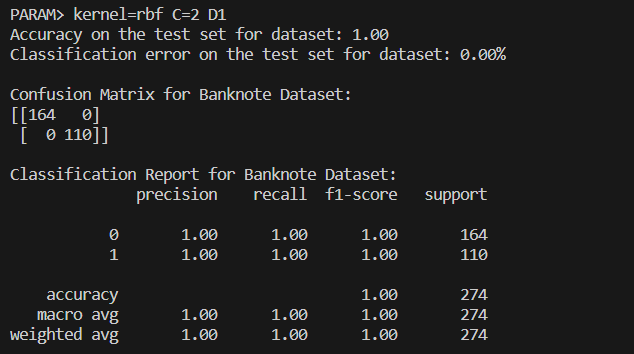
* Kernel: rbf
* Constant:2

With the best parameters the prediction was performed with the 80% of the dataset for training and the other 20% for test:



The accuracy, the classification error, the confusion matrix and the classification report from the execution of the supervised training with the best parameters is the following:





1. **Multiple Linear Regression (MLR)**

For the MLR case there was used **Weka**, a popular data mining software with libraries fror LinearRegresion. Dataset were loaded directly from files.

Then we train the data using the default method:

A screenshot of a computer

Description automatically generated

Then we compare the results with the one for the cross-validation with 10 folds:

A screenshot of a computer

Description automatically generated

By creating a table to compare the values if they are bigger or smaller then 0.5 (threshold value) we can detect the approximation of the function. For 0.5 => 4667 (sum of errors) – which is very fast.

A screenshot of a computer

Description automatically generated

Results from each dataset are in the github repository.

1. **Back-Propagation (BP)**

We have used TensorFlow, one of the most popular when it comes to BP algorithm and Keras which is a library that provides the interface for artificial neural networks.

The classification error is a critical metric in evaluating the performance of our model. It measures the proportion of incorrect predictions out of the total number of instances. In a binary classification task, an error occurs when the predicted class does not match the actual class. For example, predicting that a client will subscribe ('yes') when they actually will not ('no'), or vice versa.

A screenshot of a computer

Description automatically generated

Figure 1 - Bank dataset

This matrix reveals that:

* 698 instances were correctly predicted as 'no' (True Negatives).
* 43 instances were incorrectly predicted as 'yes' (False Positives).
* 43 instances were incorrectly predicted as 'no' (False Negatives).
* 40 instances were correctly predicted as 'yes' (True Positives).

The classification error from the same fold is calculated to be approximately 10.44%. This error rate is a significant metric, considering the imbalance in the dataset; it reflects the proportion of misclassifications.

For the **banknote\_authentication** dataset, we've designed a neural network with the following specifications:

* **Learning Rate:** 0.05, which governs the speed at which the model learns. A moderate rate helps in converging to a global minimum without overshooting.
* **Momentum:** 0.1, to help accelerate the optimizer in the correct direction and prevent oscillations.
* **Neural Network Architecture:** The network consists of an input layer with 4 neurons (one for each feature), two hidden layers with 20 and 5 neurons respectively, and an output layer with a single neuron.

A screenshot of a computer

Description automatically generated

Figure 2 - Banknote dataset

This confusion matrix tells us that:

* 156 instances of genuine banknotes were correctly identified (True Negatives).
* 5 instances of genuine banknotes were incorrectly classified as forgeries (False Positives).
* 0 instances of forged banknotes were incorrectly classified as genuine (False Negatives).
* 114 instances of forged banknotes were correctly identified (True Positives).

From the given fold, the classification error is remarkably low, at approximately 1.82%. This low error rate indicates a high level of accuracy in the model's ability to classify banknotes as genuine or counterfeit.

A 4-fold cross-validation technique is also used for the ring dataset to validate the model's performance and ensure its ability to generalize to new data. This technique involves partitioning the dataset into four subsets, training on three subsets, and validating on the fourth. This process is repeated four times, with each subset used once as the validation set.

A screenshot of a computer

Description automatically generated

Figure 3 - Ring dataset

Interpreting the confusion matrix:

* 1233 True Negatives (TN): Correctly predicted negative class instances.
* 82 False Positives (FP): Negative class instances incorrectly predicted as positive.
* 7 False Negatives (FN): Positive class instances incorrectly predicted as negative.
* 1178 True Positives (TP): Correctly predicted positive class instances.

The classification error, derived from the confusion matrix, is approximately 3.56%, indicating the model's high accuracy in classifying the instances. This low error rate is particularly noteworthy in a dataset with a complex, non-linear decision boundary like that of the ring dataset.

The custom BP algorithm has been updated from its original form in the first assignment (A1), leading to enhanced functionality and now works. Here are the key improvements:

* **Enhanced Cross-Validation Methodology:** The addition of the 'trainingFolds' parameter to the BP algorithm is a major change. The 'folds' refer to the different segments that this new feature creates from the data and epochs. Training and validation sets are rotated using each fold in a methodical manner, which results in a more complete and reliable training procedure.
* **Modification to the "CostFunction" Method:** A new parameter has been added to the algorithm's "costFunction" method. This modification is essential for precisely computing the classification error, particularly in light of the 'forward' method's adoption of a threshold. The formula..... is used to calculate the error, where..... stands for various classification counts.
* **Normalization Procedures Are Removed:** The normalization and denormalization stages of data processing are eliminated in the current version of the BP algorithm. The'readDataAndNormalize' method has been modified due to the dataset being utilized already having a normalized format.
* **Threshold Application in the 'Forward' Method:** A notable change is the 'forward' method's addition of a new parameter. This update is necessary for testing since it applies a threshold to the final layer's output, turning it into binary form (0 or 1).

**Compute and compare**

Analyzing the data in the table below, it's evident that Multiple Linear Regression isn't the best fit for this scenario. Its predictions seem quite random, with errors nearing 50% in both cases. This is likely because linear regressions are designed to model linear relationships, and they falter when the relationships are nonlinear, leading to poor model performance.

|  |  |  |  |
| --- | --- | --- | --- |
|  | SVM | MLR | BP |
| ring-separable | 4.18% | 46.67% | 1.85% |
| ring-merged | 6.91% | 46.67% | 1.85% |

Neural networks typically excel when dealing with well-segregated and easily segmented data, which holds true in this case. Despite needing to modify our code and adjust the architecture and parameters, neural networks yielded quite impressive results. However, a notable downside is the longer execution time, probably due to the vast amount of data processed. This factor should be considered when opting for this method.

The most effective technique in this instance appears to be the Support Vector Machines (SVM). SVMs demonstrated lower classification error rates and overall better performance. This success can be attributed to the relative simplicity of creating hyperplanes in SVMs compared to the complexities involved in training a neural network. The nature of the data, which could be readily separated using lines in a plane, likely played a significant role in the effectiveness of SVMs in this context.

The BP algorithm demonstrates exceptional performance. This indicates the neural network's capability to capture complex, non-linear patterns in data. However, its performance on the bank dataset shows room for improvement.

|  |  |  |  |
| --- | --- | --- | --- |
| X | ring | bank | banknote |
| SVM | 4.18% | 9.22% | ? |
| MLR | 46.67% | 10.68% | 1.09% |
| BP | 1.85% | 11.04% | 0.36% |

**Performance Ratings**

* Green (Good): Error rates below 5% are considered excellent. BP's performance on the ring and banknote datasets falls into this category, showcasing high accuracy.
* Orange (Average): Error rates between 5% to 10% represent average performance. SVM's performance on the bank dataset and MLR on the banknote dataset are in this range, indicating a need for potential improvements or adjustments in model parameters.
* Red (Poor): Error rates above 10% are deemed below par. MLR's performance on the ring dataset is particularly concerning, indicating a poor fit of the model to the dataset's structure.

Based on these results, it seems that the BP algorithm performs better overall, especially for datasets with intricate boundaries or patterns. SVM remains a dependable technique across various datasets, whereas MLR's efficacy appears to be highly contingent on the dataset. The comparison will be completed and a complete picture of the capabilities of each method will be provided by the final error rate for SVM on the banknote authentication dataset, which is still to be determined (TBD).

**Part 2.1: Parameter selection**

**SVM:**

* Ring dataset

|  |  |  |
| --- | --- | --- |
|  | **Separable dataset** | **Merged dataset** |
| **Parameters** | * Kernel: rbf * Constant:701 | * Kernel: rbf * Constant:701 |
| **Cross-validation score** | 0.997 | 0.7789 |
| **Classification error on test set** | 0.25% | 4.39% |
| **Accuracy** | 1.0 | 0.96 |

* Bank dataset

|  |  |
| --- | --- |
| **Parameters** | • Kernel: poly  • Constant:1  • Degree:4 |
| **Cross-validation score** | 0.9077 |
| **Classification error on test set** | 7.89% |
| **Accuracy** | 0.92 |

* Banknote dataset

|  |  |
| --- | --- |
| **Parameters** | • Kernel: rbf  • Constant:2 |
| **Cross-validation score** | 0.9973 |
| **Classification error on test set** | 0.00% |
| **Accuracy** | 1.00 |

**BP:**

* Ring dataset

|  |  |  |
| --- | --- | --- |
|  | **Separable dataset** | **Merged dataset** |
| **Parameters** |  |  |
| **Cross-validation score** |  |  |
| **Classification error on test set** |  |  |

* Bank dataset

|  |  |
| --- | --- |
| **Parameters** |  |
| **Cross-validation score** |  |
| **Classification error on test set** |  |

* Banknote dataset

|  |  |
| --- | --- |
| **Parameters** |  |
| **Cross-validation score** |  |
| **Classification error on test set** |  |

**Part 2.2: Evaluation of the results**

Confusion matrix comparation

|  |  |  |  |
| --- | --- | --- | --- |
|  | SVM | MLP | BP |
| Ring | [[5321 12]  [ 13 4654]]  [[5323 10]  [ 429 4238]] |  | [[5188 145]  [ 40 4627]]  [[983 402]  [192 923]] |
| Bank | [[731 11]  [ 54 28]] |  | [[698 43]  [43 40]] |
| Banknote | [[164 0]  [ 0 110]] |  | [[156 5]  [0 114]] |

GITHUB – REPOSITORY:

<https://github.com/sebastianbuzdugan/A2-NeuralNetworks>