**Question 1. Decision Tree Classifier [10 Points]**

Data: The zip file “*hw2.q1.data.zip*” contains 3 CSV files:

* “*hw2.q1.train.csv*” contains 10,000 rows and 26 columns. The first column ‘*y*’ is the output variable with 2 classes: 0, 1. The remaining 25 columns contain input features: *x\_1*, …, x*\_25*.
* “*hw2.q1.test.csv*” contains 5,000 rows and 41 columns. The first column ‘*y*’ is the output variable with 2 classes: 0, 1. The remaining 25 columns contain input features: *x\_1*, …, x*\_25*.
* “*hw2.q1.new.csv*” contains 30 rows and 26 columns. The first column ‘*ID*’ is an identifier for 30 unlabeled samples. The remaining 25 columns contain input features: *x\_1*, …, x*\_25*.

**Task 1**. [4 points]

Use 5-fold cross-validation with the 10,000 labeled exampled from “*hw2.q1.train.csv*” to determine the fewest number of rules using which a decision tree classifier can achieve mean cross-validation accuracy of at least 0.96. Report the number of rules needed, the cross-validation accuracy obtained, and all the hyper-parameter values for the *DecisionTreeClassifier*.

*Number of rules needed*: **4**

*Mean cross-validation accuracy:* **0.9695** (*rounded to 4 decimal places*)

*Hyper-parameter values for selected DecisionTreeClassifier model:*

**Text

Description automatically generated**

**Task 2**. [2 Points]

Train a *DecisionTreeClassifier* with the hyper-parameter values determined in Task 1 on all 10,000 training samples and use it to predict the output class ‘*y*’ for the 2,000 examples in “*hw2.q1.test.csv*”*.* Report the following:

* *Accuracy on 2,000 test examples*: **0.9710** (rounded to 4 decimal places)
* *Classification report* for the 2,000 test examples:

Classification report on test samples:

precision recall f1-score support

0.0 0.9679 0.9771 0.9725 1048 1.0 0.9745 0.9643 0.9694 952

accuracy 0.9710 2000 macro avg 0.9712 0.9707 0.9709 2000 weighted avg 0.9710 0.9710 0.9710 2000

* Of the 952 test samples that belong to class *y*=1, how many are correctly predicted (according to your classification report)?

TOTAL = 0.9643\*952 = 918 are correctly predicted.

**Predicted class = 0**

**Predicted class = 1**

**Class label = 0**

1024

24

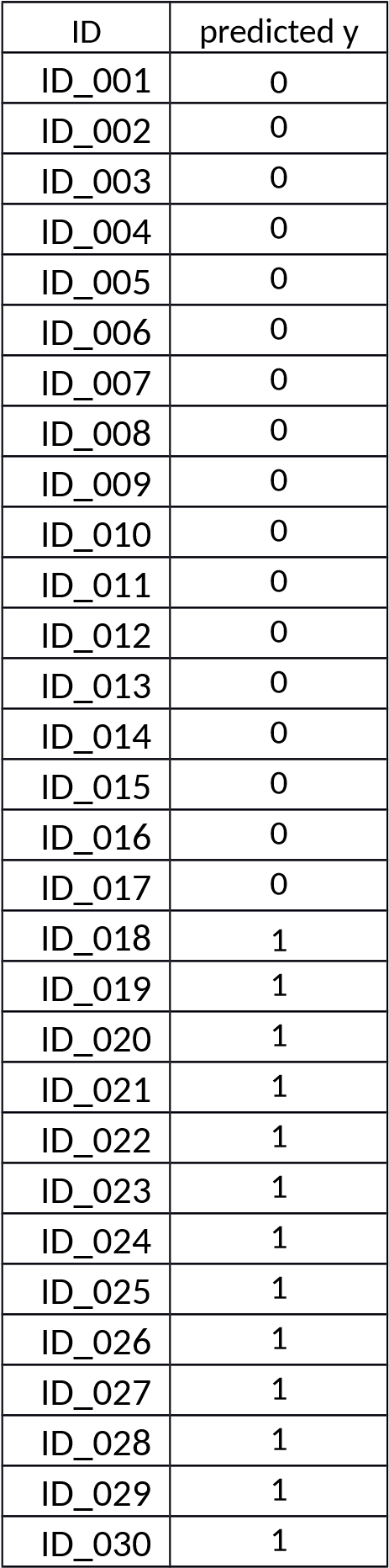
**Class label = 1**

34

918

**Task 3.** [2 Points]

Use the model trained in Task 2 to predict the output class ‘*y’* for the 30 examples in “*hw2.q1.new.csv*”. Specify the predicted classes in the table below:

**Task 4.** [2 Points]

Of the 25 input variables which ones are relevant for this classification task?

The following input variables are relevant for this classification task: **x\_4, x12, x\_22**

Interpret your trained model and specify the rules that can be used to classify the output based on the inputs.

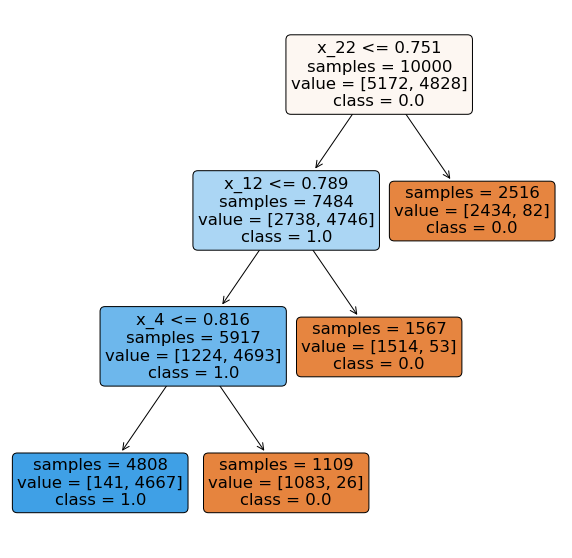
Rules:

*Rule 1. IF x\_22 > 0.751 THEN y = 0.*

*Rule 2. IF x\_22 ≤ 0.751 AND x\_12 > 0.789 THEN y = 0*

*Rule 3. IF x\_22 ≤ 0.751 AND x\_12 ≤ 0.789 AND x\_4 > 0.816 THEN y = 0.*

*Rule 4. IF x\_22 ≤ 0.751 AND x\_12 ≤ 0.789 AND x\_4 ≤ 0.816 THEN y = 1.*



**Question 2. Supervised machine learning classifiers [10 Points]**

Data: The zip file “*hw2.q2.data.zip*” contains 3 CSV files:

* “*hw2.q2.train.csv*” contains 8,000 rows and 11 columns. The first column ‘*y*’ is the output variable with 4 classes: 0, 1, 2, 3. The remaining 10 columns contain input features: *x1*, …, x*10*.
* “*hw2.q2.test.csv*” contains 2,000 rows and 11 columns. The first column ‘*y*’ is the output variable with 4 classes: 0, 1, 2, 3. The remaining 10 columns contain input features: *x1*, …, x*10*.
* “*hw2.q1.new.csv*” contains 30 rows and 10 columns. The first column ‘*ID*’ is an identifier for 30 unlabeled samples. The remaining 10 columns contain input features: *x1*, …, x*10*.

**Task 1**. [6 points]

Use 4-fold cross-validation with the 8,000 labeled exampled from “*hw2.q2.train.csv*” to identify a classifier that achieves mean cross-validation accuracy of at least 0.96. You should try several *Scikit-Learn* classifiers, including: *GaussianNB, DecisionTreeClassifier, RandomForestClassifier, ExtraTreesClassifier, KNeighborsClassifier, LogisticRegression, SVC, and MLPClassifier*. Try different hyper-parameter values for the better performing classifiers to obtain a good set of hyper-parameter values. Then select the best performing model. Report the following:

***Selected model with hyper-parameter values****:*

*Text

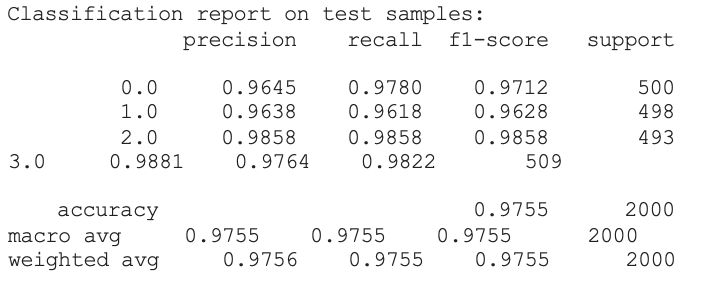
Description automatically generated*

*Mean cross-validation accuracy:* ***0.9698***(*rounded to 4 decimal places*)

**Task 2**. [2 Points]

Train the classifier with the hyper-parameter values determined in Task 1 on all 8,000 training samples and use it to predict the output class ‘*y*’ for the 2,000 examples in “*hw2.q2.test.csv*”*.* Report the following:

* *Accuracy on 2,000 test examples*: **0.9755** (rounded to 4 decimal places)
* *Classification report* for the 2,000 test examples:

****

* Of the 500 test samples that belong to class *y*=0, how many are correctly predicted (according to your classification report)?

Total = 0.9780\*500 = 489 are correctly predicted

**Predicted class =**

**0**

**Predicted class =**

**1**

**Predicted class =**

**2**

**Predicted class =**

**3**

**Class label = 0**

489

7

1

3

**Class label = 1**

11

479

6

2

**Class label = 2**

3

3

486

1

**Class label = 3**

4

8

0

497

**Task 3.** [2 Points]

Use the model trained in Task 2 to predict the output class ‘*y’* for the 30 examples in “*hw2.q2.new.csv*”. Specify the predicted classes in the table below:

