

PW1: CN2 algorithm

Santiago del Rey Juárez

April 2022

Contents

1	CN2 pseudo-code	2
2	Results	4
2.1	Iris dataset	4
2.1.1	Rules	4
2.2	Heart dataset	5
2.2.1	Rules	5
2.3	Rice dataset	9
2.3.1	Rules	10
2.4	Performance	13
3	How to execute the code	14
3.1	Using the CN2 class	14
3.2	Using the runner script	14

Chapter 1

CN2 pseudo-code

In this chapter short, we see the pseudo-code of the CN2 algorithm. It is broken down into two procedures. The first one (see Algorithm 1) corresponds to the main loop of the algorithm where we obtain and save the rules. The second one (see Algorithm 2) corresponds to the procedure that generates the rules and finds the best current rule.

Algorithm 1 CN2 Algorithm - Part 1

```
1: procedure CN2( $E$ )
2:   Discretize  $E$  with user-defined  $\#bins$ 
3:   Replace missing values in  $E$  with the mode of the feature

4:    $rules \leftarrow$  Empty list of rules
5:    $selectors \leftarrow$  Set of all possible selectors
6:    $default\_rule\_class \leftarrow$  Most common class in  $E$ 
7:    $N \leftarrow |E|$ 
8:   ( $best\_cpx$ ,  $covered\_examples$ ,
     $most\_common\_class$ ,  $precision$ )  $\leftarrow find\_best\_complex(E)$ 
9:   while  $best\_cpx$  not nil and  $E$  not empty do
10:     $coverage \leftarrow |covered\_examples|/N$ 
11:    Remove  $covered\_examples$  from  $E$ 
12:     $rules.append(best\_cpx, most\_common\_class, precision, coverage)$ 
13:    ( $best\_cpx$ ,  $covered\_examples$ ,
     $most\_common\_class$ ,  $precision$ )  $\leftarrow find\_best\_complex(E)$ 
14:  end while
15:   $precision \leftarrow \#instances\ with\ default\_rule\_class/|E|$ 
16:   $coverage \leftarrow |E|/N$ 
17:   $rules.append(*, default\_rule\_class, precision, coverage)$ 
18: end procedure
```

Algorithm 2 CN2 Algorithm - Part 2

```
1: procedure FIND_BEST_COMPLEX( $E$ )
2:    $best\_cpx \leftarrow nil$ 
3:    $best\_entropy \leftarrow \infty$ 
4:    $best\_significance \leftarrow -\infty$ 
5:    $best\_cpx\_covered\_examples \leftarrow nil$ 
6:    $best\_cpx\_class \leftarrow nil$ 
7:    $best\_cpx\_precision \leftarrow 0$ 
8:    $star \leftarrow$  The set containing the empty complex
9:   while  $star$  is not empty do
10:     $entropies \leftarrow []$ 
11:     $significances \leftarrow []$ 
12:     $new\_star \leftarrow \{x \wedge y | x \in star, y \in selectors\}$ 
13:    Remove all complexes from  $new\_star$  that are either in  $star$ 
      or null
14:    for  $cpx$  in  $new\_star$  do
15:       $E' \leftarrow$  Set of covered examples by  $cpx$ 
16:       $cpx\_entropy \leftarrow entropy(E')$ 
17:       $cpx\_significance \leftarrow significance(E')$ 
18:       $entropies.append(cpx\_entropy)$ 
19:       $significances.append(cpx\_significance)$ 
20:      if  $cpx\_significance \geq$  user defined significance then
21:        if  $cpx\_entropy < best\_entropy$  and
           $cpx\_significance \geq best\_significance$  then
22:           $best\_cpx \leftarrow cpx$ 
23:           $best\_entropy \leftarrow cpx\_entropy$ 
24:           $best\_significance \leftarrow cpx\_significance$ 
25:           $best\_cpx\_covered\_examples \leftarrow E'$ 
26:           $best\_cpx\_class \leftarrow$  Most common class in  $E'$ 
27:           $best\_cpx\_precision \leftarrow \#instances\ with\ most\ common\ class / |E'|$ 
28:        end if
29:      end if
30:    end for
31:    repeat
32:      Remove from  $star$  the worst complex based on the  $entropies$ 
        and  $significances$  lists
33:    until  $size(new\_star) \leq$  user defined maximum
34:     $star \leftarrow new\_star$ 
35:  end while
36: end procedure
```

Chapter 2

Results

In this chapter, we will see the results of applying our implementation of the CN2 algorithm in three datasets of small, medium and large sizes.

Since the algorithm has a random component when computing the rule significance, the results have been obtained after running the algorithm 5 times for each dataset.

2.1 Iris dataset

The following section presents the results obtained from applying the algorithm to the Iris dataset¹. This is a small-sized dataset with 150 instances and 4 real attributes plus the class attribute.

2.1.1 Rules

In this ruleset, we can see how, in most cases, the algorithm can classify the instances only requiring one of the attributes. Another thing to note is how specific the rules are. This can be observed by looking at the precision, which usually is 1, and coverage, which is no greater than 0.125, of each rule. One last comment about this ruleset is how the algorithm can cover almost all the training data without needing the default rule, with only 3.3% of the training examples not being classified by any rule but the default.

Below we can see the ruleset (in brackets we can see each rule's coverage and precision respectively).

```
IF petal_length = (5.61, 6.9]  $\implies$  virginica [0.125, 1.000]  
IF petal_length = (0.99, 1.4]  $\implies$  setosa [0.125, 1.000]  
IF petal_length = (3.76, 4.35]  $\implies$  versicolor [0.125, 1.000]  
IF petal_length = (5.12, 5.61]  $\implies$  virginica [0.125, 1.000]  
IF sepal_width = (3.5, 4.4]  $\implies$  setosa [0.075, 1.000]  
IF petal_length = (4.8, 5.12]  $\wedge$  sepal_width = (2.5, 2.8]  $\implies$  virginica [0.058, 0.857]
```

¹<https://archive.ics.uci.edu/ml/datasets/iris>

IF sepal_width = (3.3, 3.5] \wedge petal_length = (1.4, 1.6] \implies setosa [0.050, 1.000]
 IF petal_length = (1.4, 1.6] \implies setosa [0.050, 1.000]
 IF petal_length = (1.6, 3.76] \wedge petal_width = (0.3, 1.0] \implies versicolor [0.050, 0.833]
 IF petal_width = (1.3, 1.5] \wedge petal_length = (4.35, 4.8] \implies versicolor [0.042, 1.000]
 IF petal_length = (4.8, 5.12] \implies virginica [0.033, 0.750]
 IF sepal_length = (6.71, 7.7] \implies versicolor [0.025, 1.000]
 IF sepal_width = (2.8, 2.9] \implies versicolor [0.025, 1.000]
 IF sepal_width = (2.9, 3.0] \implies virginica [0.025, 1.000]
 IF petal_width = (1.0, 1.3] \implies versicolor [0.017, 1.000]
 IF sepal_length = (5.8, 6.14] \implies versicolor [0.017, 1.000]
 IF * \implies virginica [0.033, 0.500]

2.2 Heart dataset

The following section presents the results obtained from applying the algorithm to the Heart dataset². This is a medium-sized dataset with 918 instances and 12 attributes both numerical and categorical.

2.2.1 Rules

In this ruleset, we can see how the algorithm needs at least two attributes to classify the instances in contrast with the one seen before. Probably because this dataset contains more data both in terms of instances and attributes. Also, as in the previous ruleset, the rules are very specific which is demonstrated by their high precision and low coverage. The last point to note is that the algorithm was not able to produce any rule only for 0.1% of the instances.

Below we can see the ruleset (in brackets we can see each rule's coverage and precision respectively).

IF ChestPainType = ATA \wedge Age = (27.0, 42.0] \implies No [0.045, 1.000]
 IF ST.Slope = Flat \wedge Oldpeak = (-2.61, 0.0] \wedge Age = (48.0, 53.0] \implies Yes [0.025, 1.000]
 IF Oldpeak = (1.4, 2.0] \wedge MaxHR = (59.0, 110.0] \implies Yes [0.023, 1.000]
 IF ChestPainType = ATA \wedge Cholesterol = (182.0, 212.0] \implies No [0.018, 1.000]
 IF Age = (59.0, 63.0] \wedge RestingBP = (79.0, 115.0] \implies Yes [0.018, 1.000]
 IF MaxHR = (59.0, 110.0] \wedge RestingBP = (140.0, 150.0] \implies Yes [0.016, 1.000]
 IF Age = (53.0, 56.0] \implies No [0.016, 0.583]
 IF Oldpeak = (2.0, 6.2] \wedge RestingBP = (135.0, 140.0] \implies Yes [0.014, 1.000]
 IF FastingBS = 1 \wedge Age = (27.0, 42.0] \wedge ChestPainType = ASY \implies Yes [0.014, 1.000]
 IF RestingBP = (115.0, 120.0] \implies No [0.014, 0.700]
 IF Oldpeak = (2.0, 6.2] \wedge Cholesterol = (289.0, 603.0] \implies Yes [0.012, 1.000]
 IF Oldpeak = (2.0, 6.2] \wedge Age = (53.0, 56.0] \implies Yes [0.012, 1.000]
 IF ST.Slope = Up \wedge Cholesterol = (212.0, 233.0] \wedge Age = (42.0, 48.0] \implies No [0.012, 1.000]
 IF ExerciseAngina = Y \wedge FastingBS = 1 \implies Yes [0.012, 1.000]
 IF ChestPainType = ATA \wedge RestingBP = (130.0, 135.0] \implies No [0.011, 1.000]
 IF ChestPainType = ATA \wedge RestingBP = (79.0, 115.0] \implies No [0.011, 1.000]

²<https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>

IF ChestPainType = ATA \wedge Age = (48.0, 53.0] \implies No [0.011, 1.000]
 IF ExerciseAngina = Y \wedge RestingBP = (140.0, 150.0] \wedge Age = (63.0, 77.0] \implies Yes [0.011, 1.000]
 IF ExerciseAngina = Y \wedge FastingBS = 1 \wedge Age = (48.0, 53.0] \implies Yes [0.011, 1.000]
 IF ChestPainType = ATA \wedge Cholesterol = (289.0, 603.0] \wedge Age = (53.0, 56.0] \implies No [0.010, 1.000]
 IF Oldpeak = (2.0, 6.2] \wedge Age = (56.0, 59.0] \implies Yes [0.010, 1.000]
 IF ST.Slope = Up \wedge Sex = F \wedge Age = (53.0, 56.0] \implies No [0.010, 1.000]
 IF ChestPainType = NAP \wedge Age = (27.0, 42.0] \wedge RestingBP = (120.0, 130.0] \implies No [0.010, 1.000]
 IF ChestPainType = ATA \wedge Age = (53.0, 56.0] \implies No [0.010, 0.857]
 IF ST.Slope = Up \wedge Age = (59.0, 63.0] \implies No [0.010, 0.857]
 IF ChestPainType = NAP \wedge MaxHR = (165.0, 202.0] \implies No [0.010, 1.000]
 IF ExerciseAngina = Y \wedge Age = (59.0, 63.0] \wedge RestingBP = (115.0, 120.0] \implies Yes [0.008, 1.000]
 IF ChestPainType = ATA \wedge RestingBP = (120.0, 130.0] \wedge Age = (42.0, 48.0] \implies No [0.008, 1.000]
 IF RestingBP = (150.0, 200.0] \wedge Age = (53.0, 56.0] \implies Yes [0.008, 1.000]
 IF ST.Slope = Up \wedge Cholesterol = (289.0, 603.0] \wedge Age = (27.0, 42.0] \implies No [0.008, 1.000]
 IF Sex = F \wedge Cholesterol = (260.0, 289.0] \wedge ChestPainType = NAP \implies No [0.008, 1.000]
 IF RestingECG = ST \wedge Age = (53.0, 56.0] \implies Yes [0.008, 1.000]
 IF Oldpeak = (1.4, 2.0] \wedge RestingBP = (130.0, 135.0] \implies Yes [0.007, 1.000]
 IF MaxHR = (165.0, 202.0] \wedge RestingECG = ST \wedge Oldpeak = (-2.61, 0.0] \implies No [0.007, 1.000]
 IF ST.Slope = Up \wedge ChestPainType = ATA \wedge RestingBP = (135.0, 140.0] \implies No [0.007, 1.000]
 IF MaxHR = (165.0, 202.0] \wedge RestingBP = (140.0, 150.0] \implies No [0.007, 1.000]
 IF Oldpeak = (2.0, 6.2] \wedge RestingECG = Normal \implies Yes [0.007, 1.000]
 IF Sex = F \wedge Age = (48.0, 53.0] \wedge ExerciseAngina = N \wedge Oldpeak = (-2.61, 0.0] \implies No [0.007, 1.000]
 IF RestingBP = (140.0, 150.0] \wedge Age = (63.0, 77.0] \implies Yes [0.007, 1.000]
 IF RestingBP = (130.0, 135.0] \implies No [0.007, 0.800]
 IF ST.Slope = Up \wedge Cholesterol = (212.0, 233.0] \wedge ChestPainType = NAP \implies No [0.007, 1.000]
 IF ExerciseAngina = Y \wedge Oldpeak = (1.0, 1.4] \implies Yes [0.007, 1.000]
 IF Sex = F \wedge Age = (42.0, 48.0] \implies No [0.007, 1.000]
 IF Oldpeak = (1.4, 2.0] \wedge Age = (42.0, 48.0] \implies Yes [0.007, 1.000]
 IF MaxHR = (110.0, 121.0] \wedge Age = (42.0, 48.0] \implies Yes [0.007, 1.000]
 IF MaxHR = (110.0, 121.0] \wedge RestingBP = (120.0, 130.0] \wedge ExerciseAngina = Y \implies Yes [0.007, 0.800]
 IF Cholesterol = (182.0, 212.0] \wedge ChestPainType = ASY \implies No [0.007, 0.800]
 IF MaxHR = (59.0, 110.0] \wedge Sex = M \implies Yes [0.007, 1.000]
 IF RestingBP = (120.0, 130.0] \implies Yes [0.007, 1.000]
 IF Age = (48.0, 53.0] \implies Yes [0.007, 1.000]
 IF ExerciseAngina = Y \wedge RestingBP = (130.0, 135.0] \wedge Age = (59.0, 63.0] \implies Yes [0.005, 1.000]
 IF Sex = F \wedge ChestPainType = NAP \wedge RestingBP = (115.0, 120.0] \implies No [0.005, 1.000]
 IF ChestPainType = ATA \wedge Age = (42.0, 48.0] \implies No [0.005, 1.000]
 IF ExerciseAngina = Y \wedge Oldpeak = (1.0, 1.4] \wedge Age = (53.0, 56.0] \implies Yes [0.005, 1.000]

IF ST_Slope = Flat \wedge Cholesterol = (260.0, 289.0] \wedge Age = (42.0, 48.0] \implies Yes [0.005, 1.000]
 IF ChestPainType = NAP \wedge Cholesterol = (260.0, 289.0] \implies No [0.005, 1.000]
 IF ExerciseAngina = Y \wedge Cholesterol = (-1.0, 182.0] \wedge Age = (53.0, 56.0] \implies Yes [0.005, 1.000]
 IF MaxHR = (121.0, 132.0] \wedge Oldpeak = (1.0, 1.4] \implies Yes [0.005, 1.000]
 IF ChestPainType = TA \wedge RestingBP = (135.0, 140.0] \implies Yes [0.005, 1.000]
 IF RestingBP = (140.0, 150.0] \wedge Age = (53.0, 56.0] \wedge ExerciseAngina = N \implies No [0.005, 1.000]
 IF FastingBS = 1 \wedge Age = (53.0, 56.0] \implies Yes [0.005, 1.000]
 IF ExerciseAngina = Y \wedge MaxHR = (143.0, 153.0] \wedge Age = (42.0, 48.0] \implies Yes [0.005, 1.000]
 IF RestingBP = (140.0, 150.0] \wedge ChestPainType = NAP \implies No [0.005, 1.000]
 IF FastingBS = 1 \wedge Age = (63.0, 77.0] \implies Yes [0.005, 1.000]
 IF MaxHR = (165.0, 202.0] \implies Yes [0.005, 0.750]
 IF Age = (59.0, 63.0] \wedge ChestPainType = ASY \implies Yes [0.005, 1.000]
 IF FastingBS = 1 \wedge Cholesterol = (-1.0, 182.0] \implies Yes [0.005, 1.000]
 IF RestingBP = (120.0, 130.0] \wedge Sex = F \implies No [0.005, 1.000]
 IF Oldpeak = (0.0, 0.1] \wedge Age = (48.0, 53.0] \implies No [0.004, 1.000]
 IF MaxHR = (165.0, 202.0] \wedge ST_Slope = Down \implies No [0.004, 1.000]
 IF Sex = F \wedge Age = (42.0, 48.0] \wedge ChestPainType = ATA \implies No [0.004, 1.000]
 IF ExerciseAngina = Y \wedge Age = (53.0, 56.0] \wedge Sex = F \implies Yes [0.004, 1.000]
 IF Sex = F \wedge Age = (42.0, 48.0] \wedge RestingECG = LVH \implies No [0.004, 1.000]
 IF ExerciseAngina = Y \wedge Oldpeak = (2.0, 6.2] \wedge RestingBP = (140.0, 150.0] \implies Yes [0.004, 1.000]
 IF MaxHR = (110.0, 121.0] \wedge RestingBP = (130.0, 135.0] \implies Yes [0.004, 1.000]
 IF ST_Slope = Down \wedge RestingBP = (115.0, 120.0] \implies Yes [0.004, 1.000]
 IF ST_Slope = Down \wedge ExerciseAngina = N \implies No [0.004, 1.000]
 IF Age = (59.0, 63.0] \wedge Cholesterol = (260.0, 289.0] \implies Yes [0.004, 1.000]
 IF Sex = F \wedge Age = (63.0, 77.0] \wedge RestingBP = (79.0, 115.0] \implies No [0.004, 1.000]
 IF RestingBP = (130.0, 135.0] \wedge ChestPainType = ASY \wedge ST_Slope = Up \implies No [0.004, 1.000]
 IF RestingBP = (130.0, 135.0] \wedge Age = (42.0, 48.0] \implies Yes [0.004, 1.000]
 IF ChestPainType = NAP \wedge Oldpeak = (1.0, 1.4] \implies No [0.004, 1.000]
 IF ChestPainType = ATA \wedge Age = (63.0, 77.0] \implies No [0.004, 1.000]
 IF MaxHR = (110.0, 121.0] \wedge RestingBP = (135.0, 140.0] \implies Yes [0.004, 1.000]
 IF RestingBP = (150.0, 200.0] \wedge ST_Slope = Up \wedge FastingBS = 1 \implies No [0.004, 1.000]
 IF Age = (59.0, 63.0] \wedge RestingECG = LVH \wedge Sex = M \implies Yes [0.004, 1.000]
 IF RestingBP = (140.0, 150.0] \wedge RestingECG = ST \implies Yes [0.004, 1.000]
 IF MaxHR = (165.0, 202.0] \wedge Cholesterol = (182.0, 212.0] \implies Yes [0.004, 1.000]
 IF RestingBP = (140.0, 150.0] \wedge Age = (59.0, 63.0] \implies Yes [0.004, 1.000]
 IF FastingBS = 1 \wedge Age = (59.0, 63.0] \implies Yes [0.004, 1.000]
 IF Cholesterol = (182.0, 212.0] \wedge Age = (42.0, 48.0] \implies No [0.004, 1.000]
 IF Cholesterol = (182.0, 212.0] \wedge Age = (48.0, 53.0] \implies No [0.004, 1.000]
 IF ST_Slope = Up \wedge RestingBP = (135.0, 140.0] \wedge Age = (48.0, 53.0] \implies No [0.004, 1.000]
 IF Cholesterol = (289.0, 603.0] \wedge Oldpeak = (1.4, 2.0] \wedge Sex = M \implies Yes [0.004, 1.000]
 IF MaxHR = (59.0, 110.0] \wedge Age = (63.0, 77.0] \implies Yes [0.004, 1.000]
 IF ST_Slope = Up \wedge MaxHR = (59.0, 110.0] \implies No [0.004, 1.000]
 IF Sex = F \wedge Cholesterol = (-1.0, 182.0] \implies Yes [0.004, 1.000]
 IF Age = (63.0, 77.0] \wedge Sex = F \implies No [0.004, 1.000]
 IF Oldpeak = (0.0, 0.1] \wedge MaxHR = (153.0, 165.0] \implies Yes [0.003, 1.000]
 IF Sex = F \wedge ChestPainType = NAP \wedge Age = (59.0, 63.0] \implies No [0.003, 1.000]
 IF ExerciseAngina = Y \wedge Oldpeak = (0.0, 0.1] \implies Yes [0.003, 1.000]

IF Oldpeak = (0.0, 0.1] \implies No [0.003, 1.000]
 IF ST.Slope = Up \wedge Oldpeak = (2.0, 6.2] \wedge RestingBP = (140.0, 150.0] \implies Yes [0.003, 1.000]
 IF Sex = F \wedge Cholesterol = (212.0, 233.0] \wedge Age = (42.0, 48.0] \implies No [0.003, 1.000]
 IF ExerciseAngina = Y \wedge Oldpeak = (2.0, 6.2] \wedge Age = (27.0, 42.0] \implies Yes [0.003, 1.000]
 IF Oldpeak = (2.0, 6.2] \wedge RestingBP = (130.0, 135.0] \implies Yes [0.003, 1.000]
 IF ST.Slope = Up \wedge MaxHR = (143.0, 153.0] \wedge Age = (56.0, 59.0] \implies No [0.003, 1.000]
 IF ExerciseAngina = Y \wedge ChestPainType = TA \implies No [0.003, 1.000]
 IF MaxHR = (165.0, 202.0] \wedge Oldpeak = (0.1, 1.0] \wedge Age = (27.0, 42.0] \implies No [0.003, 1.000]
 IF ST.Slope = Flat \wedge ChestPainType = ATA \wedge Age = (59.0, 63.0] \implies Yes [0.003, 1.000]
 IF ST.Slope = Down \wedge ChestPainType = TA \implies No [0.003, 1.000]
 IF ST.Slope = Down \wedge Age = (42.0, 48.0] \implies Yes [0.003, 1.000]
 IF Oldpeak = (2.0, 6.2] \wedge Sex = F \implies Yes [0.003, 1.000]
 IF Oldpeak = (2.0, 6.2] \wedge RestingBP = (115.0, 120.0] \implies Yes [0.003, 1.000]
 IF ChestPainType = NAP \wedge Oldpeak = (2.0, 6.2] \wedge Age = (63.0, 77.0] \implies Yes [0.003, 1.000]
 IF MaxHR = (59.0, 110.0] \wedge RestingBP = (130.0, 135.0] \implies No [0.003, 1.000]
 IF ST.Slope = Up \wedge ChestPainType = TA \wedge Age = (63.0, 77.0] \implies No [0.003, 1.000]
 IF ST.Slope = Flat \wedge ChestPainType = TA \wedge Age = (42.0, 48.0] \implies Yes [0.003, 1.000]
 IF ExerciseAngina = Y \wedge Cholesterol = (182.0, 212.0] \wedge Age = (27.0, 42.0] \implies Yes [0.003, 1.000]
 IF ST.Slope = Down \wedge Cholesterol = (260.0, 289.0] \implies Yes [0.003, 1.000]
 IF RestingBP = (130.0, 135.0] \wedge Age = (53.0, 56.0] \implies No [0.003, 1.000]
 IF Oldpeak = (1.4, 2.0] \wedge ChestPainType = TA \implies No [0.003, 1.000]
 IF FastingBS = 1 \wedge Cholesterol = (260.0, 289.0] \wedge RestingBP = (150.0, 200.0] \implies Yes [0.003, 1.000]
 IF RestingBP = (130.0, 135.0] \wedge MaxHR = (121.0, 132.0] \implies Yes [0.003, 1.000]
 IF MaxHR = (165.0, 202.0] \wedge Age = (59.0, 63.0] \implies Yes [0.003, 1.000]
 IF ST.Slope = Down \wedge RestingBP = (120.0, 130.0] \implies Yes [0.003, 1.000]
 IF ST.Slope = Down \implies No [0.003, 1.000]
 IF ChestPainType = TA \wedge RestingECG = LVH \wedge RestingBP = (150.0, 200.0] \implies Yes [0.003, 1.000]
 IF MaxHR = (165.0, 202.0] \wedge Cholesterol = (212.0, 233.0] \wedge RestingECG = Normal \wedge RestingBP = (135.0, 140.0] \implies No [0.003, 1.000]
 IF ChestPainType = ATA \wedge ExerciseAngina = Y \wedge Sex = M \implies Yes [0.003, 1.000]
 IF Oldpeak = (2.0, 6.2] \wedge MaxHR = (59.0, 110.0] \implies Yes [0.003, 1.000]
 IF ChestPainType = TA \wedge Sex = F \implies No [0.003, 1.000]
 IF Oldpeak = (2.0, 6.2] \wedge ExerciseAngina = Y \implies No [0.003, 1.000]
 IF Age = (27.0, 42.0] \wedge RestingBP = (120.0, 130.0] \implies Yes [0.003, 1.000]
 IF Oldpeak = (1.4, 2.0] \wedge FastingBS = 1 \wedge Age = (59.0, 63.0] \implies Yes [0.003, 1.000]
 IF MaxHR = (165.0, 202.0] \wedge Age = (56.0, 59.0] \implies Yes [0.003, 1.000]
 IF Oldpeak = (2.0, 6.2] \wedge MaxHR = (143.0, 153.0] \implies Yes [0.003, 1.000]
 IF ChestPainType = TA \wedge FastingBS = 0 \implies No [0.003, 1.000]
 IF ChestPainType = TA \wedge Cholesterol = (-1.0, 182.0] \implies Yes [0.003, 1.000]
 IF ST.Slope = Up \wedge Oldpeak = (1.0, 1.4] \implies No [0.003, 1.000]
 IF Oldpeak = (1.0, 1.4] \wedge MaxHR = (165.0, 202.0] \implies No [0.003, 1.000]
 IF Cholesterol = (260.0, 289.0] \wedge Age = (42.0, 48.0] \implies No [0.003, 1.000]
 IF ChestPainType = ATA \wedge Cholesterol = (260.0, 289.0] \implies Yes [0.003, 1.000]
 IF Oldpeak = (2.0, 6.2] \wedge FastingBS = 1 \implies No [0.003, 1.000]
 IF ChestPainType = ATA \wedge RestingBP = (120.0, 130.0] \implies No [0.003, 1.000]
 IF MaxHR = (121.0, 132.0] \wedge Age = (53.0, 56.0] \implies Yes [0.003, 1.000]
 IF RestingECG = ST \wedge Age = (48.0, 53.0] \implies Yes [0.003, 1.000]

IF Cholesterol = (260.0, 289.0] \wedge Age = (63.0, 77.0] \implies Yes [0.003, 1.000]
 IF MaxHR = (110.0, 121.0] \wedge Cholesterol = (260.0, 289.0] \implies Yes [0.003, 1.000]
 IF MaxHR = (110.0, 121.0] \wedge RestingBP = (79.0, 115.0] \wedge RestingECG = Normal \implies Yes [0.003, 1.000]
 IF Sex = F \wedge FastingBS = 1 \implies No [0.003, 1.000]
 IF Cholesterol = (260.0, 289.0] \wedge RestingBP = (79.0, 115.0] \implies No [0.003, 1.000]
 IF FastingBS = 1 \wedge Age = (42.0, 48.0] \implies Yes [0.003, 1.000]
 IF FastingBS = 1 \wedge Age = (48.0, 53.0] \implies Yes [0.003, 1.000]
 IF Age = (59.0, 63.0] \wedge Cholesterol = (289.0, 603.0] \implies No [0.003, 1.000]
 IF RestingBP = (140.0, 150.0] \wedge Cholesterol = (289.0, 603.0] \implies Yes [0.003, 1.000]
 IF MaxHR = (165.0, 202.0] \wedge Sex = M \wedge ST.Slope = Up \wedge Oldpeak = (-2.61, 0.0] \wedge RestingECG = LVH \implies No [0.003, 1.000]
 IF Age = (59.0, 63.0] \wedge Sex = M \wedge ST.Slope = Flat \wedge FastingBS = 0 \wedge Oldpeak = (1.4, 2.0] \implies No [0.003, 1.000]
 IF MaxHR = (121.0, 132.0] \wedge Age = (42.0, 48.0] \implies Yes [0.003, 1.000]
 IF MaxHR = (110.0, 121.0] \wedge Age = (63.0, 77.0] \wedge Sex = M \implies No [0.003, 1.000]
 IF Cholesterol = (182.0, 212.0] \wedge Age = (63.0, 77.0] \implies Yes [0.003, 1.000]
 IF Cholesterol = (182.0, 212.0] \wedge Sex = F \implies No [0.003, 1.000]
 IF FastingBS = 1 \wedge RestingBP = (135.0, 140.0] \implies No [0.003, 1.000]
 IF Age = (42.0, 48.0] \wedge MaxHR = (143.0, 153.0] \implies Yes [0.003, 1.000]
 IF Age = (42.0, 48.0] \wedge Sex = M \wedge ChestPainType = NAP \implies No [0.003, 1.000]
 IF Cholesterol = (-1.0, 182.0] \wedge MaxHR = (143.0, 153.0] \implies No [0.003, 1.000]
 IF MaxHR = (121.0, 132.0] \wedge ChestPainType = NAP \implies Yes [0.003, 1.000]
 IF MaxHR = (110.0, 121.0] \wedge Cholesterol = (-1.0, 182.0] \wedge FastingBS = 0 \wedge Sex = M \wedge ST.Slope = Flat \implies Yes [0.003, 1.000]
 IF RestingBP = (79.0, 115.0] \wedge Age = (56.0, 59.0] \implies No [0.003, 1.000]
 IF RestingBP = (79.0, 115.0] \wedge Age = (63.0, 77.0] \implies Yes [0.003, 1.000]
 IF Oldpeak = (1.4, 2.0] \wedge Age = (63.0, 77.0] \implies Yes [0.003, 1.000]
 IF RestingECG = ST \wedge Cholesterol = (233.0, 260.0] \implies No [0.003, 1.000]
 IF Cholesterol = (233.0, 260.0] \wedge Age = (56.0, 59.0] \implies Yes [0.003, 1.000]
 IF Cholesterol = (233.0, 260.0] \wedge MaxHR = (121.0, 132.0] \implies Yes [0.003, 1.000]
 IF Cholesterol = (233.0, 260.0] \implies Yes [0.003, 1.000]
 IF Age = (42.0, 48.0] \implies Yes [0.003, 1.000]
 IF Age = (48.0, 53.0] \wedge ChestPainType = NAP \implies Yes [0.003, 1.000]
 IF Age = (63.0, 77.0] \wedge Cholesterol = (212.0, 233.0] \implies No [0.003, 1.000]
 IF MaxHR = (110.0, 121.0] \implies No [0.003, 1.000]
 IF MaxHR = (121.0, 132.0] \implies No [0.003, 1.000]
 IF Cholesterol = (289.0, 603.0] \implies Yes [0.003, 1.000]
 IF ChestPainType = NAP \implies No [0.003, 1.000]
 IF Oldpeak = (-2.61, 0.0] \implies Yes [0.003, 1.000]
 IF MaxHR = (153.0, 165.0] \implies No [0.003, 1.000]
 IF * \implies Yes [0.001, 1.000]

2.3 Rice dataset

The following section presents the results obtained from applying the algorithm to the Rice dataset³. This is a large-sized dataset with 3810 instances and 7 numerical attributes and the class.

³<https://www.kaggle.com/datasets/muratkokludataset/rice-dataset-commeo-and-osmancik>

2.3.1 Rules

In this last ruleset, we see how happens the same as in the previous two. We have very specific rules again, which are characteristic of this algorithm. Also, we see how most of the rules are composed by the conjunction of multiple selectors. The most noticeable thing in this ruleset is the fact that the algorithm has not been able to classify 20.4% of the training instances. This is a considerable amount of instances, even more compared to the two previous cases. Moreover, by looking at the accuracy of the default rule we can see that most of the remaining instances are misclassified by this rule.

Below we can see the ruleset (in brackets we can see each rule's coverage and precision respectively).

```
IF Major_Axis_Length = (147.79000000000002, 174.46]  $\implies$  Osmancik [0.250, 0.999]
IF Major_Axis_Length = (203.33, 239.01]  $\wedge$  Minor_Axis_Length = (90.01, 107.54]  $\implies$ 
Cammeo [0.051, 0.981]
IF Major_Axis_Length = (203.33, 239.01]  $\wedge$  Minor_Axis_Length = (86.34, 90.01]  $\wedge$  Area
= (13932.0, 18913.0]  $\implies$  Cammeo [0.050, 1.000]
IF Eccentricity = (0.77, 0.87]  $\implies$  Osmancik [0.043, 0.931]
IF Major_Axis_Length = (203.33, 239.01]  $\wedge$  Extent = (0.64, 0.73]  $\wedge$  Area = (13932.0,
18913.0]  $\implies$  Cammeo [0.042, 1.000]
IF Major_Axis_Length = (174.46, 185.63]  $\wedge$  Minor_Axis_Length = (82.65, 86.34]  $\implies$ 
Osmancik [0.026, 0.886]
IF Major_Axis_Length = (203.33, 239.01]  $\wedge$  Extent = (0.49, 0.6]  $\wedge$  Eccentricity = (0.89,
0.9]  $\implies$  Cammeo [0.025, 1.000]
IF Major_Axis_Length = (174.46, 185.63]  $\wedge$  Extent = (0.6, 0.64]  $\wedge$  Convex_Area = (11627.75,
12676.5]  $\implies$  Osmancik [0.022, 0.941]
IF Major_Axis_Length = (203.33, 239.01]  $\wedge$  Extent = (0.73, 0.86]  $\wedge$  Eccentricity = (0.89,
0.9]  $\implies$  Cammeo [0.018, 1.000]
IF Major_Axis_Length = (174.46, 185.63]  $\wedge$  Eccentricity = (0.77, 0.87]  $\wedge$  Area = (11375.25,
12405.5]  $\implies$  Osmancik [0.015, 0.979]
IF Perimeter = (359.09000000000003, 426.42]  $\wedge$  Eccentricity = (0.89, 0.9]  $\implies$  Osmancik
[0.011, 1.000]
IF Eccentricity = (0.77, 0.87]  $\wedge$  Area = (11375.25, 12405.5]  $\wedge$  Extent = (0.6, 0.64]  $\implies$ 
Osmancik [0.011, 1.000]
IF Major_Axis_Length = (174.46, 185.63]  $\wedge$  Minor_Axis_Length = (82.65, 86.34]  $\wedge$  Extent
= (0.64, 0.73]  $\implies$  Osmancik [0.011, 0.971]
IF Perimeter = (483.02, 548.45]  $\wedge$  Area = (12405.5, 13932.0]  $\wedge$  Minor_Axis_Length =
(82.65, 86.34]  $\implies$  Cammeo [0.010, 1.000]
IF Convex_Area = (14274.25, 19099.0]  $\wedge$  Extent = (0.73, 0.86]  $\implies$  Cammeo [0.008,
0.880]
IF Major_Axis_Length = (203.33, 239.01]  $\wedge$  Minor_Axis_Length = (82.65, 86.34]  $\wedge$  Con-
vex_Area = (14274.25, 19099.0]  $\implies$  Cammeo [0.008, 1.000]
IF Major_Axis_Length = (203.33, 239.01]  $\wedge$  Minor_Axis_Length = (82.65, 86.34]  $\wedge$  Area
= (12405.5, 13932.0]  $\implies$  Cammeo [0.008, 1.000]
IF Perimeter = (359.09000000000003, 426.42]  $\wedge$  Minor_Axis_Length = (82.65, 86.34]  $\implies$ 
Osmancik [0.008, 1.000]
IF Perimeter = (359.09000000000003, 426.42]  $\implies$  Osmancik [0.008, 0.957]
IF Convex_Area = (14274.25, 19099.0]  $\wedge$  Extent = (0.49, 0.6]  $\implies$  Cammeo [0.007, 1.000]
IF Perimeter = (359.09000000000003, 426.42]  $\wedge$  Eccentricity = (0.87, 0.89]  $\implies$  Osmancik
[0.007, 1.000]
```

IF Perimeter = (483.02, 548.45] \wedge Extent = (0.6, 0.64] \implies Cammeo [0.007, 1.000]
 IF Eccentricity = (0.9, 0.95] \wedge Convex_Area = (12676.5, 14274.25] \wedge Minor_Axis.Length = (82.65, 86.34] \wedge Extent = (0.64, 0.73] \implies Cammeo [0.006, 0.895]
 IF Perimeter = (483.02, 548.45] \wedge Minor_Axis.Length = (59.52, 82.65] \implies Cammeo [0.006, 1.000]
 IF Major_Axis.Length = (203.33, 239.01] \wedge Extent = (0.6, 0.64] \wedge Eccentricity = (0.9, 0.95] \implies Cammeo [0.006, 1.000]
 IF Major_Axis.Length = (203.33, 239.01] \wedge Extent = (0.73, 0.86] \wedge Perimeter = (448.66, 483.02] \implies Cammeo [0.005, 1.000]
 IF Eccentricity = (0.77, 0.87] \wedge Major_Axis.Length = (174.46, 185.63] \wedge Minor_Axis.Length = (90.01, 107.54] \wedge Perimeter = (426.42, 448.66] \wedge Extent = (0.64, 0.73] \implies Osmancik [0.005, 1.000]
 IF Convex_Area = (14274.25, 19099.0] \wedge Area = (12405.5, 13932.0] \implies Cammeo [0.005, 1.000]
 IF Perimeter = (426.42, 448.66] \wedge Minor_Axis.Length = (82.65, 86.34] \wedge Major_Axis.Length = (185.63, 203.33] \implies Osmancik [0.005, 1.000]
 IF Area = (7550.99, 11375.25] \wedge Extent = (0.49, 0.6] \wedge Eccentricity = (0.89, 0.9] \implies Osmancik [0.005, 1.000]
 IF Perimeter = (359.09000000000003, 426.42] \wedge Extent = (0.64, 0.73] \implies Osmancik [0.004, 1.000]
 IF Eccentricity = (0.77, 0.87] \wedge Convex_Area = (11627.75, 12676.5] \wedge Minor_Axis.Length = (90.01, 107.54] \implies Osmancik [0.004, 1.000]
 IF Area = (13932.0, 18913.0] \wedge Minor_Axis.Length = (86.34, 90.01] \implies Cammeo [0.004, 1.000]
 IF Major_Axis.Length = (174.46, 185.63] \wedge Minor_Axis.Length = (59.52, 82.65] \wedge Eccentricity = (0.87, 0.89] \wedge Area = (7550.99, 11375.25] \implies Osmancik [0.004, 1.000]
 IF Perimeter = (483.02, 548.45] \wedge Major_Axis.Length = (185.63, 203.33] \wedge Minor_Axis.Length = (90.01, 107.54] \wedge Convex_Area = (14274.25, 19099.0] \implies Cammeo [0.004, 0.667]
 IF Major_Axis.Length = (174.46, 185.63] \wedge Perimeter = (448.66, 483.02] \implies Osmancik [0.004, 0.917]
 IF Major_Axis.Length = (203.33, 239.01] \wedge Minor_Axis.Length = (59.52, 82.65] \wedge Extent = (0.49, 0.6] \implies Cammeo [0.004, 1.000]
 IF Major_Axis.Length = (174.46, 185.63] \wedge Eccentricity = (0.77, 0.87] \wedge Extent = (0.49, 0.6] \implies Osmancik [0.004, 1.000]
 IF Convex_Area = (14274.25, 19099.0] \wedge Eccentricity = (0.87, 0.89] \wedge Area = (13932.0, 18913.0] \wedge Extent = (0.6, 0.64] \wedge Minor_Axis.Length = (90.01, 107.54] \implies Cammeo [0.004, 0.818]
 IF Major_Axis.Length = (174.46, 185.63] \wedge Extent = (0.6, 0.64] \implies Osmancik [0.004, 0.818]
 IF Major_Axis.Length = (203.33, 239.01] \wedge Extent = (0.49, 0.6] \wedge Area = (12405.5, 13932.0] \implies Cammeo [0.003, 1.000]
 IF Convex_Area = (14274.25, 19099.0] \wedge Extent = (0.6, 0.64] \implies Cammeo [0.003, 0.900]
 IF Convex_Area = (14274.25, 19099.0] \implies Cammeo [0.003, 0.900]
 IF Minor_Axis.Length = (82.65, 86.34] \wedge Eccentricity = (0.9, 0.95] \wedge Extent = (0.6, 0.64] \implies Cammeo [0.003, 0.900]
 IF Major_Axis.Length = (174.46, 185.63] \wedge Minor_Axis.Length = (86.34, 90.01] \wedge Area = (11375.25, 12405.5] \wedge Extent = (0.73, 0.86] \implies Osmancik [0.003, 1.000]
 IF Major_Axis.Length = (174.46, 185.63] \wedge Eccentricity = (0.89, 0.9] \wedge Extent = (0.49, 0.6] \implies Osmancik [0.003, 1.000]
 IF Major_Axis.Length = (174.46, 185.63] \wedge Convex_Area = (7722.99, 11627.75] \wedge Area = (11375.25, 12405.5] \implies Osmancik [0.003, 1.000]
 IF Major_Axis.Length = (174.46, 185.63] \wedge Extent = (0.49, 0.6] \wedge Minor_Axis.Length = (86.34, 90.01] \wedge Area = (11375.25, 12405.5] \implies Osmancik [0.003, 1.000]

IF Major_Axis_Length = (203.33, 239.01] \wedge Eccentricity = (0.77, 0.87] \implies Cammeo [0.002, 1.000]
 IF Perimeter = (359.09000000000003, 426.42] \wedge Major_Axis_Length = (174.46, 185.63] \wedge Extent = (0.73, 0.86] \wedge Area = (7550.99, 11375.25] \implies Osmancik [0.002, 1.000]
 IF Major_Axis_Length = (174.46, 185.63] \wedge Extent = (0.6, 0.64] \wedge Area = (7550.99, 11375.25] \implies Osmancik [0.002, 1.000]
 IF Major_Axis_Length = (203.33, 239.01] \wedge Area = (12405.5, 13932.0] \wedge Minor_Axis_Length = (59.52, 82.65] \implies Cammeo [0.002, 0.857]
 IF Major_Axis_Length = (174.46, 185.63] \wedge Extent = (0.6, 0.64] \wedge Minor_Axis_Length = (59.52, 82.65] \implies Osmancik [0.002, 1.000]
 IF Perimeter = (483.02, 548.45] \wedge Extent = (0.6, 0.64] \wedge Area = (12405.5, 13932.0] \implies Cammeo [0.002, 1.000]
 IF Major_Axis_Length = (174.46, 185.63] \wedge Extent = (0.6, 0.64] \wedge Eccentricity = (0.89, 0.9] \implies Osmancik [0.002, 1.000]
 IF Major_Axis_Length = (174.46, 185.63] \wedge Minor_Axis_Length = (59.52, 82.65] \wedge Area = (7550.99, 11375.25] \wedge Extent = (0.64, 0.73] \wedge Eccentricity = (0.9, 0.95] \wedge Perimeter = (426.42, 448.66] \implies Osmancik [0.002, 1.000]
 IF Major_Axis_Length = (203.33, 239.01] \wedge Area = (11375.25, 12405.5] \implies Cammeo [0.002, 1.000]
 IF Eccentricity = (0.77, 0.87] \wedge Area = (11375.25, 12405.5] \wedge Minor_Axis_Length = (82.65, 86.34] \implies Osmancik [0.002, 1.000]
 IF Major_Axis_Length = (174.46, 185.63] \wedge Minor_Axis_Length = (90.01, 107.54] \wedge Area = (11375.25, 12405.5] \implies Osmancik [0.002, 1.000]
 IF Perimeter = (426.42, 448.66] \wedge Area = (7550.99, 11375.25] \wedge Minor_Axis_Length = (82.65, 86.34] \implies Osmancik [0.002, 1.000]
 IF Convex_Area = (7722.99, 11627.75] \wedge Extent = (0.73, 0.86] \wedge Major_Axis_Length = (174.46, 185.63] \wedge Eccentricity = (0.9, 0.95] \implies Osmancik [0.002, 1.000]
 IF Perimeter = (483.02, 548.45] \wedge Eccentricity = (0.77, 0.87] \wedge Extent = (0.64, 0.73] \implies Cammeo [0.001, 1.000]
 IF Convex_Area = (14274.25, 19099.0] \wedge Eccentricity = (0.89, 0.9] \implies Cammeo [0.001, 1.000]
 IF Area = (7550.99, 11375.25] \wedge Eccentricity = (0.89, 0.9] \wedge Convex_Area = (11627.75, 12676.5] \implies Osmancik [0.001, 1.000]
 IF Major_Axis_Length = (174.46, 185.63] \wedge Perimeter = (448.66, 483.02] \wedge Area = (11375.25, 12405.5] \implies Osmancik [0.001, 1.000]
 IF Area = (13932.0, 18913.0] \implies Cammeo [0.001, 0.750]
 IF Major_Axis_Length = (174.46, 185.63] \wedge Perimeter = (448.66, 483.02] \wedge Extent = (0.49, 0.6] \implies Osmancik [0.001, 1.000]
 IF Major_Axis_Length = (203.33, 239.01] \wedge Minor_Axis_Length = (59.52, 82.65] \wedge Extent = (0.6, 0.64] \implies Cammeo [0.001, 1.000]
 IF Major_Axis_Length = (203.33, 239.01] \wedge Eccentricity = (0.89, 0.9] \wedge Area = (12405.5, 13932.0] \implies Cammeo [0.001, 1.000]
 IF Perimeter = (483.02, 548.45] \wedge Minor_Axis_Length = (86.34, 90.01] \wedge Convex_Area = (12676.5, 14274.25] \wedge Major_Axis_Length = (185.63, 203.33] \implies Cammeo [0.001, 1.000]
 IF Eccentricity = (0.77, 0.87] \wedge Convex_Area = (11627.75, 12676.5] \implies Osmancik [0.001, 1.000]
 IF Perimeter = (426.42, 448.66] \wedge Minor_Axis_Length = (86.34, 90.01] \wedge Major_Axis_Length = (185.63, 203.33] \implies Osmancik [0.001, 1.000]
 IF Convex_Area = (7722.99, 11627.75] \wedge Area = (11375.25, 12405.5] \implies Cammeo [0.001, 1.000]
 IF Major_Axis_Length = (174.46, 185.63] \wedge Minor_Axis_Length = (82.65, 86.34] \wedge Eccentricity = (0.89, 0.9] \wedge Extent = (0.64, 0.73] \wedge Perimeter = (426.42, 448.66] \implies Osmancik [0.001, 1.000]

IF Perimeter = (483.02, 548.45] \wedge Eccentricity = (0.9, 0.95] \wedge Major_Axis_Length = (185.63, 203.33] \implies Cammeo [0.001, 1.000]
 IF Major_Axis_Length = (203.33, 239.01] \wedge Area = (12405.5, 13932.0] \wedge Extent = (0.64, 0.73] \implies Cammeo [0.001, 1.000]
 IF Major_Axis_Length = (174.46, 185.63] \wedge Minor_Axis_Length = (82.65, 86.34] \wedge Eccentricity = (0.89, 0.9] \wedge Extent = (0.73, 0.86] \implies Osmancik [0.001, 1.000]
 IF Eccentricity = (0.9, 0.95] \wedge Minor_Axis_Length = (86.34, 90.01] \wedge Extent = (0.6, 0.64] \implies Cammeo [0.001, 1.000]
 IF Eccentricity = (0.9, 0.95] \wedge Convex_Area = (12676.5, 14274.25] \wedge Area = (11375.25, 12405.5] \implies Cammeo [0.001, 1.000]
 IF * \implies Osmancik [0.204, 0.476]

2.4 Performance

In this section, we will briefly discuss the performance of the algorithm in the three datasets.

In Table 2.1 we can see how well the algorithm performs in the different datasets in terms of accuracy and training time. We can see that in all the datasets the best accuracy is between 70-80%, which is a fairly good result. However, the average accuracy is not so promising for the Rice dataset, which obtains a 53.5%. This could be due to the discretization step since all its features are numeric and we are grouping ranges of values in one category. Thus, it might be the case that two instances of different classes end up with the same feature values, leading to misclassifications.

In terms of time performance, we see how the greater the dataset the greater the time needed to train. One interesting thing is that the largest dataset needs less time to train than the medium one. This is probably due to two factors.

First, the Heart dataset has more features and different values per feature than the Rice dataset. This means that the number of selectors will be bigger in the former and that it will take more time to explore all the possible rules.

Second, it is possible that in the Rice dataset the algorithm can find rules surpassing the significance threshold more easily than in the Heart dataset. Thus, the algorithm will cover the whole set of examples faster than in the Heart dataset.

Table 2.1: Accuracy and training time for each dataset.

	Iris	Heart	Rice
Best accuracy	0.8	0.739	0.787
Avg. accuracy	0.767	0.701	0.535
Avg. training time (s)	3.152	72.048	28.972

Chapter 3

How to execute the code

3.1 Using the CN2 class

To create an instance of the CN2 class we can use the default parameters or we can specify the maximum star size, the significance threshold and seed used in the significance computation to make the results reproducible.

Once we have an instance of the algorithm we must call the `fit` method, which receives the training data without the labels, the labels, and the number of bins used in the discretization of numerical variables.

After training the algorithm we can use the `predict` method, which receives a set of examples without the labels and returns its predicted labels.

To obtain the ruleset we can access the `rule_list` class attribute. Additionally, the CN2 class implements two methods to obtain the formatted rules. The `print_rules` method displays the rules in the standard output and the `save_rules` method, which receives a file name without the extension and a format (i.e. text or latex) and saves the rules in a folder named “results”.

3.2 Using the runner script

To facilitate the testing of the algorithm the `runner.py` scrip is provided.

To run the test you need to create a python virtual environment and install in it the following dependencies:

```
pandas = "^1.4.1"
numpy = "^1.22.3"
scipy = "^1.8.0"
sklearn = "^0.0"
```

After the installation, you can run the test with the `runner.py` script.

WARNING: The `runner.py` must be in the same location that the source and data folders.

```
usage: runner.py [-h] [--short] [--medium] [--long] [--iterations ITERATIONS]
                [--seed SEED]
```

optional arguments:

```
-h, --help            show this help message and exit
--iterations ITERATIONS, -i ITERATIONS
                        number of times the algorithm is executed (default 5)
--seed SEED           seed for reproducible results
```

Datasets:

By default all datasets are used

```
--short, -s          run CN2 with short dataset
--medium, -m         run CN2 with medium dataset
--long, -l           run CN2 with long dataset
```