Prof. Heitor Silvério Lopes
Prof. Thiago H. Silva

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

CPGEI

Data Mining & Knowledge Discovery

Class 2b – Data Classification (Rules)

2025

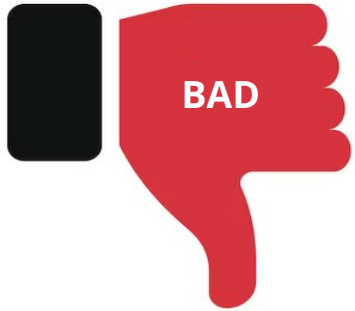
Tasks x Methods in Data Mining

Tasks	Methods
Classification	Decision trees (C4.5), Classification rules, k-nearest-neighboors, Random forest, Support vector machine, Bayesian classifier, Neural network, Adaboost
Association Rules	Apriori, FP-growth, Eclat, Zigzag
Regression	Linear Regression, Polynomial regression, Logistic regression
Feature Selection & Dimensionality Reduction	Principal component analysis (PCA), Chi-square, Entropy, Information gain
Clustering	K-means, Kohonen's self-organized map, Density-based scan, Hierarchical grouping, t-SNE
Data visualization *	Silhouette plot, scatter plot, heatmap, box plot, clusters, t-SNE

Advantages and drawback sof decision trees

- Advantages:
  - Visual representation
  - Compact representation of a rule set
  - Fast classification of new instances
  - It can deal with continuous of discrete atributes

- Drawbacks:
  - Irrelevant atributes can negativelly affect the construction of the tree and its understanding
  - Small variations in the data can result in significantly different trees
  - A subtree can be replicated several times
  - Decision trees are not adequate when having many classes



Rule-based classifiers

- Classify instances of a dataset using a set of *n* rules (*n*≥ 1) such as:
  IF **antecedent** THEN **consequent**
- A set of "*R*" rules is a disjunction* of *1..n* rules

- The **antecedent** is a conjunction* of *1..k* triplets of the type: {Attribute, Operator, Value}: *{Aj,Oj, Vj}, j=1..k*
  - The **Operator** depends on the type of atribute, it can be: {=, ≠, >, <, ≤, ≥, etc}
  - Each **Value** is defined within the limits {*max*,*min*} of each attribute

- The **consequent** represents the Class to which the instance belongs

* conjunction="AND" (^), disjunction="OR" (v)

Advantages of a rule-based classifier

- They are as comprehensive as Decision Trees
- They are intuitively easy to interpret
- They are easy to generate from a dataset
- They can classify new instances quickly
- Their classification performance is comparable to Decision Trees

Decision Trees X Decision Rules

- Decision trees are criated top-down (“divide-and-conquer”), while Decision Rules are created botton-up (by “coverage”)
- A decision tree has a set of equivalent decision rules and vice-versa
- BUT:
  - The set of rules generated by traversing a tree can be very large !!!
  - Small sets of rules with many attributes can generate very complex trees

Important properties for a rule induction algorithm

- Quality:
  - The set of induced rules should have a good accuracy (or other quality metric), even in the presence of noise in the data
- Simplicity:
  - In order to be humanly compreensible, the set of induced rules must be as simple as possible
- Escalability:
  - In real-world applications, datasets may be large and high-dimensional.
  - The rule induction algorithm **must be** computational efficient

Coverage X Accuracy

- The **Coverage** of a rule is the fraction of instances that satisfy the antecedent of the rule:

$$\text{Coverage}(R) = |A| \ / \ |T|$$

  - Where: |A| is the number of instances that satisfy rule R, and |T| is the total number of instances

- The **Accuracy** of a rule is the fraction of instances that satisfy both, the antecedent and the consequent of the rule:

$$\text{Precision}(R) = |A \cap c| \ / \ |T|,$$

  - Where $|A \cap c|$ is the number of instances that satisfy the rule

Example of a rule-based classifier

- Rule **R1** covers 1/16 of instances and hits 1/1

	Nome	Temperatura corporal	Cobertura de pele	Dá cria	Ser Aquático	Ser Aéreo	Possui Pernas	Hiberna	Rótulo da Classe
1	Humano	Sangue quente	Cabelo	Sim	Não	Não	Sim	Não	Mamífero
2	Píton	Sangue frio	Escamas	Não	Não	Não	Não	Sim	Réptil
3	Salmão	Sangue frio	Escamas	Não	Sim	Não	Não	Não	Peixe
4	Baleia	Sangue quente	Cabelo	Sim	Sim	Não	Não	Não	Mamífero
5	Sapo	Sangue frio	Nenhuma	Não	Sim	Não	Sim	Sim	Anfíbio
6	Dragão de Komodo	Sangue frio	Escamas	Não	Não	Não	Sim	Não	Réptil
7	Morcego	Sangue quente	Cabelo	Sim	Não	Sim	Sim	Sim	Mamífero
8	Pomba	Sangue quente	Penas	Não	Não	Sim	Sim	Não	Ave
9	Gato	Sangue quente	Pêlo	Sim	Não	Não	Sim	Não	Mamífero
10	Leopardo	Sangue frio	Pêlo	Sim	Sim	Não	Sim	Não	Mamífero
11	Tubarão	Sangue frio	Escamas	Não	Sim	Não	Não	Não	Peixe
12	Tartaruga	Sangue frio	Escamas	Não	Semi	Não	Sim	Não	Réptil
13	Pingüim	Sangue quente	Penas	Não	Semi	Não	Sim	Não	Ave
14	Porco-espinho	Sangue quente	Espinhos	Sim	Não	Não	Sim	Sim	Mamífero
15	Enguia	Sangue frio	Escamas	Não	Sim	Não	Não	Não	Peixe
16	Salamandra	Sangue frio	Nenhuma	Não	Semi	Não	Sim	Sim	Anfíbio

R1: (Dá cria= não) AND (Ser aéreo = sim) → Ave
R2: (Dá cria= não) AND (Ser aquático= sim) → Peixe
R3: (Dá cria = sim) AND (Temp.corporal = quente) → Mamífero
R4: (Dá cria = não) AND (Ser aéreo = não) → Réptil
R5: (Ser aquático = semi) → Anfíbio

Example of a rule-based classifier

- Rule **R2** covers 4/16 of instances and hits 3/4

	Nome	Temperatura corporal	Cobertura de pele	Dá cria	Ser Aquático	Ser Aéreo	Possui Pernas	Hiberna	Rótulo da Classe
1	Humano	Sangue quente	Cabelo	Sim	Não	Não	Sim	Não	Mamífero
2	Píton	Sangue frio	Escamas	Não	Não	Não	Não	Sim	Réptil
3	Salmão	Sangue frio	Escamas	Não	Sim	Não	Não	Não	Peixe
4	Baleia	Sangue quente	Cabelo	Sim	Sim	Não	Não	Não	Mamífero
5	Sapo	Sangue frio	Nenhuma	Não	Sim	Não	Sim	Sim	Anfíbio
6	Dragão de Komodo	Sangue frio	Escamas	Não	Não	Não	Sim	Não	Réptil
7	Morcego	Sangue quente	Cabelo	Sim	Não	Sim	Sim	Sim	Mamífero
8	Pomba	Sangue quente	Penas	Não	Não	Sim	Sim	Não	Ave
9	Gato	Sangue quente	Pêlo	Sim	Não	Não	Sim	Não	Mamífero
10	Leopardo	Sangue frio	Pêlo	Sim	Sim	Não	Sim	Não	Mamífero
11	Tubarão	Sangue frio	Escamas	Não	Sim	Não	Não	Não	Peixe
12	Tartaruga	Sangue frio	Escamas	Não	Semi	Não	Sim	Não	Réptil
13	Pingüim	Sangue quente	Penas	Não	Semi	Não	Sim	Não	Ave
14	Porco-espinho	Sangue quente	Espinhos	Sim	Não	Não	Sim	Sim	Mamífero
15	Enguia	Sangue frio	Escamas	Não	Sim	Não	Não	Não	Peixe
16	Salamandra	Sangue frio	Nenhuma	Não	Semi	Não	Sim	Sim	Anfíbio

R1: (Dá cria= não) AND (Ser aéreo = sim) → Ave
R2: (Dá cria= não) AND (Ser aquático= sim) → Peixe
R3: (Dá cria = sim) AND (Temp.corporal = quente) → Mamífero
R4: (Dá cria = não) AND (Ser aéreo = não) → Réptil
R5: (Ser aquático = semi) → Anfíbio

Example of a rule-based classifier

- Rule **R3** covers 6/16 of instances and hits 5/6
- Rule **R4** covers 9/16 of instances and hits 3/9
- Rule **R5** covers 3/16 of instances and hits 1/3

R1: (Dá cria= não) AND (Ser aéreo = sim) → Ave
R2: (Dá cria= não) AND (Ser aquático= sim) → Peixe
R3: (Dá cria = sim) AND (Temp.corporal = quente) → Mamífero
R4: (Dá cria = não) AND (Ser aéreo = não) → Réptil
R5: (Ser aquático = semi) → Anfíbio

Example of a rule-based classifier

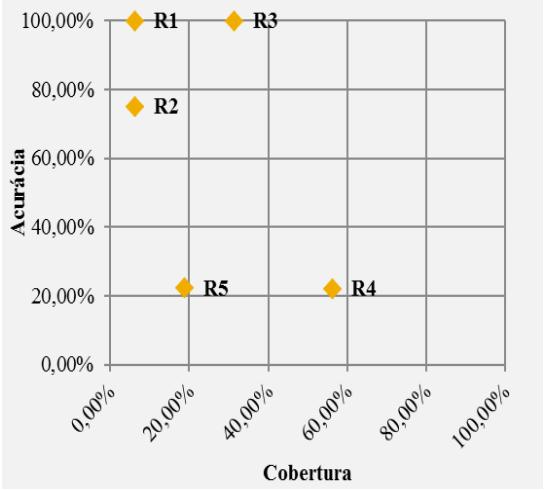
- Analysing the whole dataset with all 5 rules:

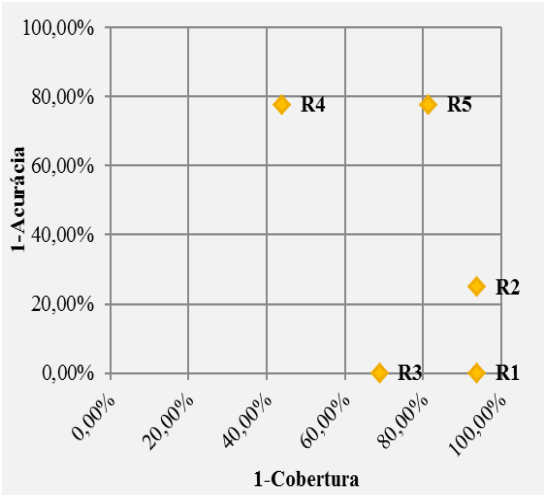
Rule	Coverage	Accuracy
R1	1/16 = 6,25%	**1/1 = 100,0%**
R2	4/16 = 25,00%	3/4 = 75,0%
R3	5/16 = 31,25%	**5/5 = 100,0%**
R4	**9/16 = 56,25%**	2/9 = 22,2%
R5	3/16 = 18,75%	1/3 = 33,3%

Example of a rule-based classifier

- Which is the **best** rule (R1..R5) ??
- The best means the rule that has, at the **same time**, the maximal coverage and maximal accuracy:

Regra	Cobertura	Acurácia
R1	1/16 = 6,25%	**1/1 = 100,0%**
R2	4/16 = 25,00%	3/4 = 75,0%
R3	5/16 = 31,25%	**5/5 = 100,0%**
R4	**9/16 = 56,25%**	2/9 = 22,2%
R5	3/16 = 18,75%	1/3 = 33,3%





Problems with rule-based classifiers

- Although classification rules are interesting for Descriptive Analysis, it does not always work correctly:
- Consider the following set of rules and the **new** instances to be classified:

R1: (Dá cria= não) AND (Ser aéreo = sim) → Ave
R2: (Dá cria= não) AND (Ser aquático= sim) → Peixe
R3: (Dá cria = sim) AND (Temp.corporal = quente) → Mamífero
R4: (Dá cria = não) AND (Ser aéreo = não) → Réptil
R5: (Ser aquático = semi) → Anfíbio

Animal	Temp.	Cobertura	Dá cria	Aquático	Aéreo	Pernas	Hiberna
lêmure	quente	pêlos	sim	não	não	sim	não
tartaruga	frio	escamas	não	semi	não	sim	não
tubarão	frio	escamas	sim	sim	não	não	não

- The “lêmure” instance triggers rule R3 (→ “mamífero”): **OK !**
- The “tartaruga” instance triggers rule R4 (→”réptil”) and R5 (→”anfíbio”): **CONFLICT !**
- The “tubarão” instance do not trigger any rule, so there is an **INDETERMINATION !**

Rule ordering

- If a set of rules is NOT mutually exclusive, then an instance could be covered by multiple rules

- There are two approaches for solving a possible rule conflict problem:
  - Ordered Rules: rules are ordered in descending order of priorit (based on a quality metric such as accuracy or coverage)

  - Unordered Rules: an instance can trigger multiple rules and each consequent has a vote. The highest number of votes for the class labels determines the instance's class

Rule ordering

- Individual rules are ranked according to their quality, considering a specific metric, but this may be difficult to interpret in the real-world
- Example: consider the original rules ordered by Coverage or Accuracy

R7: (Dá cria=sim) ^ (Temp.corporal=quente) → Mamífero
R11: (Cobertura=escamas) ^ (Aquático=sim) → Peixe
R8: (Dá cria=não) ^ (Temp.corporal=quente) → Ave
R12: (Cobertura=nenhuma) → Anfíbio
R10: (Cobertura=escamas) ^ (Aquático=não) → Réptil
R6: (Cobertura=penas) ^ (Ser aéreo=sim) → Ave
R9: (Ser aquático=semi) → Anfíbio

Accuracy

Regra	C	A	ordem C	ordem A
R7	31,25%	100%	1	1
R11	18,75%	100%	3	2
R8	12,50%	100%	5	3
R12	12,50%	100%	6	4
R10	25%	50%	2	5
R6	12,50%	50%	7	6
R9	18,75%	33,30%	4	7

Different ordering !!!

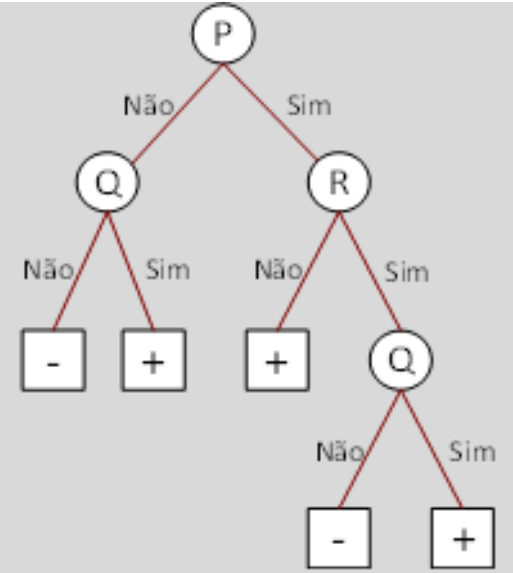
Rule ordering

- In general, ordering by classes is the most usual method for classification rules
- Rules that belong to the same class are grouped together
- Relative ordering **inside the same class** is not importante, as long as one of the rules is triggered
- Sorting from the original rules:

R6: (Cobertura=penas) ^ (Ser aéreo=sim) → Ave
R8: (Dá cria=não) ^ (Temp.corporal=quente) → Ave
R7: (Dá cria=sim) ^ (Temp.corporal=quente) → Mamífero
R12: (Cobertura=nenhuma) → Anfíbio
R9: (Ser aquático=semi) → Anfíbio
R10: (Cobertura=escamas) ^ (Aquático=não) → Réptil
R11: (Cobertura=escamas) ^ (Aquático=sim) → Peixe

Rule extraction from decision trees

- Each rule is obtained b starting from the root node and ending with a leaf node in the tree. This procedure is repeated for all leaf nodes
- Using first-order logic it is possible to simplify the set of rules



R1: (P=Não) ^(Q=Não) → classe -
R2: (P=Não) ^ (Q=Sim) → classe +
R3: (P=Sim) ^ (R=Não) → classe +
R4: (P=Sim) ^ (R=Sim) ^ (Q=Não) → classe -
R5: (P=Sim) ^ (R=Sim) ^ (Q=Sim) → classe +

Baseline Algorithms

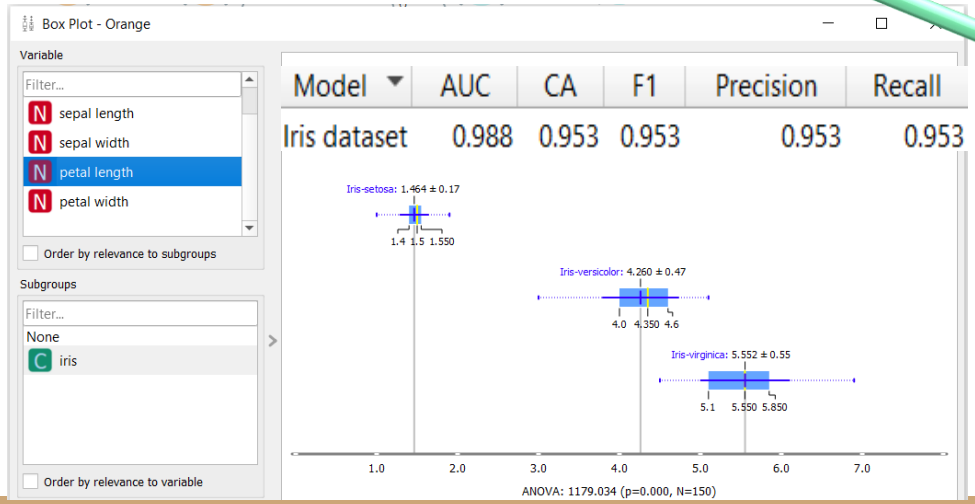
IMPORTANT

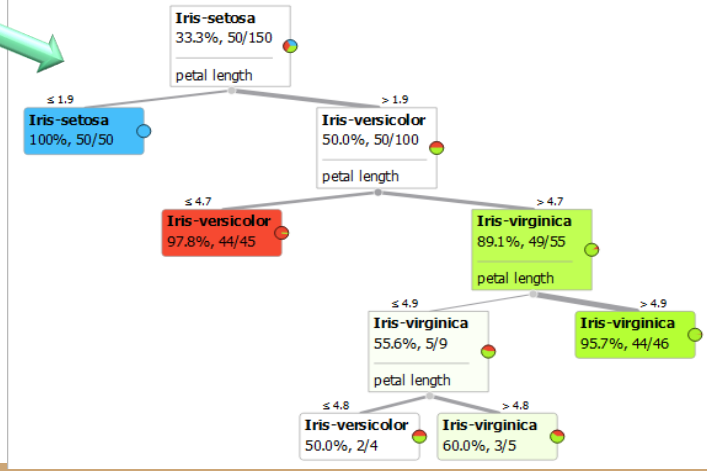
- Objective: to provide a benchmark against which to compare other classification algorithms
  - **ZeroRule** algorithm (or “no rule”)
    - This algorith Simply predicts the class of the majority of instances
    - If the number of instances per class is balanced, any class can be used
  - **OneRule** algorithm (or “single rule”)
    - Applies the classifier using only the atribute of greatest importance (which minimizes the entropy or othe measure)

Case study #3: Iris dataset with baselines

- ZeroRule algorithm (“no rule”)
  - Since classes are balanced, it is enough to choose any class
  - Accuracy= 50/150 = 30%
- OneRule algorithm (“single rule”)
  - Use Only the most importante atribute
  - The one at the top of the decision-tree

		#	In...in	Gini	ANOVA	χ^2	ReliefF	FCBF
1	petal length		1.086	0.423	1179.034	98.946	0.369	1.542
2	petal width		1.059	0.407	959.324	94.162	0.389	1.451
3	sepal length		0.624	0.247	119.265	79.243	0.115	0.000
4	sepal width		0.361	0.154	47.364	50.082	0.146	0.255



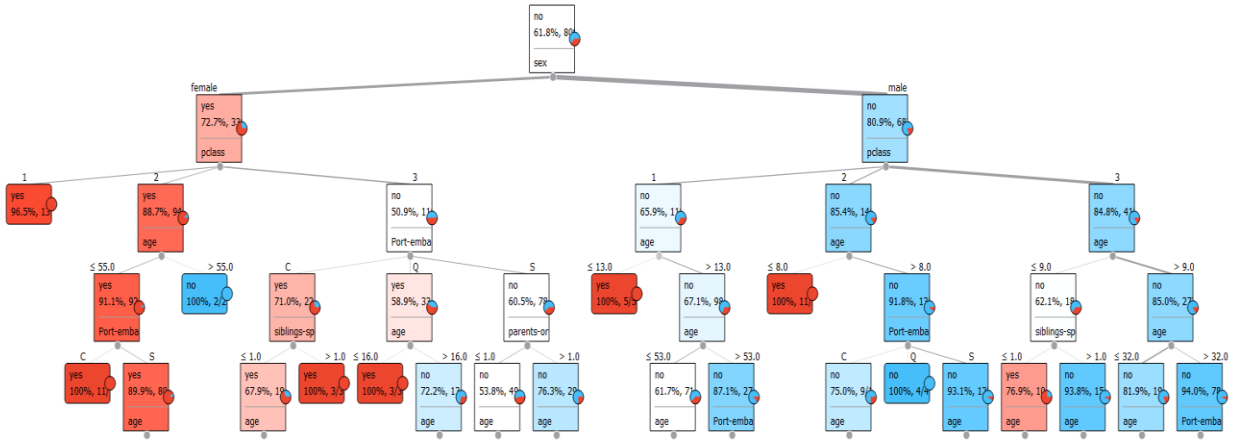


Case study #4: Titanic dataset with baselines

- ZeroRule: *survived=no* 809/1309=**61,8**%
- OneRule: *If sex=female then survived=yes* 1021/1309=**78**%
- Decision-tree

Model	AUC	CA	F1	Precision	Recall
Tree	0.930	0.868	0.865	0.869	0.868
CN2 rule inducer	0.883	0.845	0.841	0.847	0.845

Tree size: 275 nodes, 141 leaves
Edge widths: Relative to root
Target class: None



Case study #4: Titanic dataset with baselines

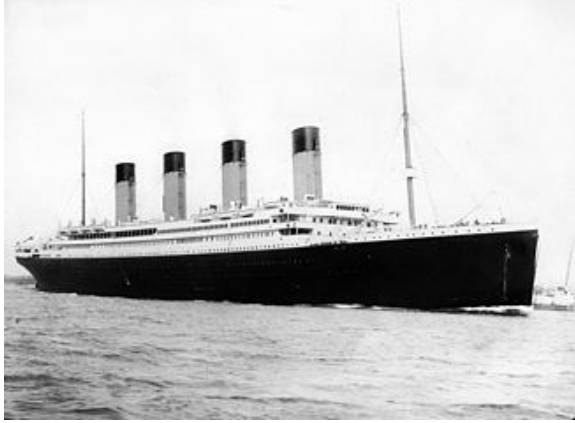
- CN2 algorithm

Model	AUC	CA	F1	Precision	Recall
Tree	0.930	0.868	0.865	0.869	0.868
CN2 rule inducer	0.883	0.845	0.841	0.847	0.845

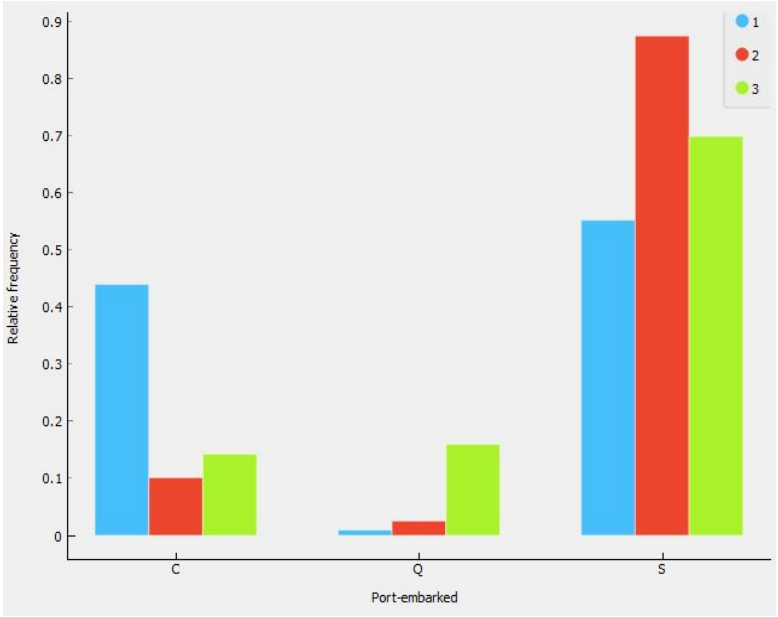
Rule ordering: ordered
Covering algorithm: exclusive
Gamma: 0.7
Evaluation measure: entropy
Beam width: 5
Minimum rule coverage: 5
Maximum rule length: 3

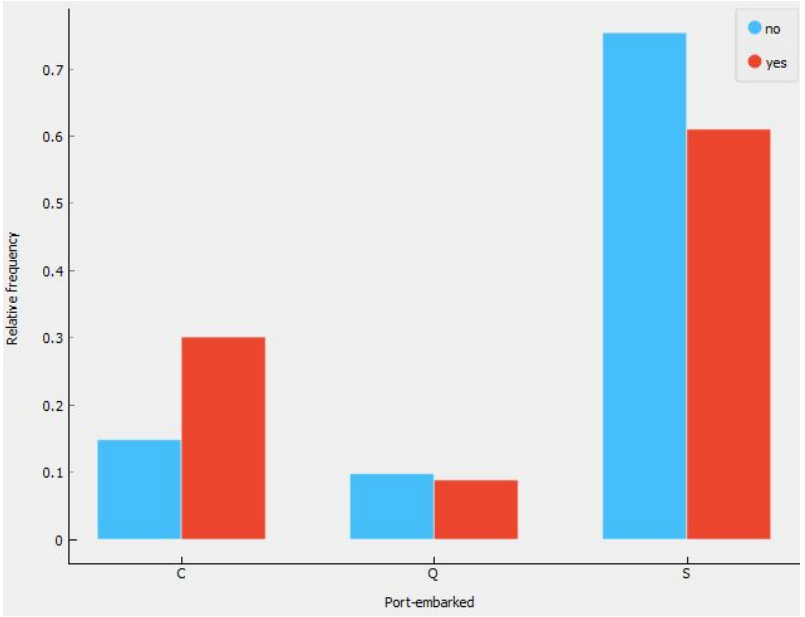
	IF conditions		THEN class	Distribution	Probabilities [%]	Quality	Length
6	sex=female AND Port-embarked=C AND age≥52.0	→	survived=1.0	[0, 12]	7 : 93	-0.00	3
8	sex=female AND pclass=1 AND age≥14.0	→	survived=1.0	[4, 119]	4 : 96	-0.207	3
11	sex≠female AND pclass=2 AND age≤14.0	→	survived=1.0	[1, 11]	14 : 86	-0.414	3
17	pclass=2 AND parents-or-children-aboard≥2.0	→	survived=1.0	[0, 17]	5 : 95	-0.00	2
18	pclass=2 AND parents-or-children-aboard≥1.0 AND age≥31.0	→	survived=1.0	[0, 7]	11 : 89	-0.00	3
24	pclass=2 AND parents-or-children-aboard≥1.0	→	survived=1.0	[2, 10]	21 : 79	-0.650	2
26	pclass=2 AND age≤57.0 AND age≤22.0	→	survived=1.0	[1, 10]	15 : 85	-0.439	3
27	pclass=2 AND Port-embarked≠C AND age≤57.0	→	survived=1.0	[8, 34]	20 : 80	-0.702	3
28	pclass≠3 AND age≤17.0 AND sex≠female	→	survived=1.0	[1, 6]	22 : 78	-0.592	3
31	age≤17.0 AND Port-embarked=C AND age≤6.0	→	survived=1.0	[1, 6]	22 : 78	-0.592	3
32	sex≠female AND pclass=3 AND age≥3.0	→	survived=1.0	[0, 8]	10 : 90	-0.00	3
35	Port-embarked=Q AND age≤31.0 AND siblings-spouses-aboard≥1.0	→	survived=1.0	[1, 6]	22 : 78	-0.592	3
39	Port-embarked≠S AND age≤17.0	→	survived=1.0	[4, 9]	33 : 67	-0.890	2
40	Port-embarked=C AND age≤29.89770554493308 AND parents-or-children-aboard≥1.0	→	survived=1.0	[2, 6]	30 : 70	-0.811	3
42	pclass=1	→	survived=1.0	[1, 4]	29 : 71	-0.722	1
43	age≥33.0 AND Port-embarked=S	→	survived=1.0	[1, 5]	25 : 75	-0.650	2
46	Port-embarked=C	→	survived=1.0	[1, 5]	25 : 75	-0.650	1
47	age≥22.0 AND age≤23.0	→	survived=1.0	[5, 11]	33 : 67	-0.896	2
48	age≥29.89770554493308 AND Port-embarked≠S	→	survived=1.0	[9, 19]	33 : 67	-0.906	2
51	parents-or-children-aboard≥1.0 AND sex=female AND parents-or-children-aboard≥2.0	→	survived=1.0	[2, 4]	38 : 62	-0.918	3
52	parents-or-children-aboard≥1.0 AND sex≠male	→	survived=1.0	[4, 7]	38 : 62	-0.946	2
54	Port-embarked=S	→	survived=1.0	[8, 12]	41 : 59	-0.971	1
0	siblings-spouses-aboard≥4.0 AND Port-embarked≠S	→	survived=0.0	[5, 0]	86 : 14	-0.00	2
1	sex≠female AND siblings-spouses-aboard≥5.0	→	survived=0.0	[9, 0]	91 : 9	-0.00	2
5	siblings-spouses-aboard≥3.0 AND siblings-spouses-aboard≥5.0	→	survived=0.0	[6, 0]	88 : 12	-0.00	2
7	sex≠female AND Port-embarked=Q AND pclass=2	→	survived=0.0	[5, 0]	86 : 14	-0.00	3

Case study #2: Titanic dataset



- Case distribution by port of embarkation and by survival

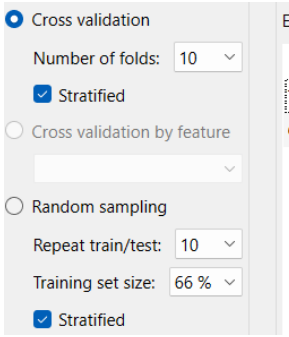




Southampton
Cheerbourg
Queenstown

Case study #2: Titanic dataset

- Pruned decision tree obtained with the full dataset



Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Tree (1) (1)	0.775	0.778	0.775	0.775	0.778	0.521
CN2 Rule Induction	0.815	0.765	0.764	0.763	0.765	0.498