

Université de Sherbrooke
Computer Science Department

IFT870/BIN710
Forage de données / Forage de données pour la bio-informatique
Data mining / Data mining for computational biology

Winter 2025

Mid-term Exam

Lecturer:
Davy Ouedraogo

Saturday, March 1st, 2025
9 am to 5 pm

Submit your work through turnin.dinf.usherbrooke.ca

This exam is individual. During grading, a score of zero (0) will be assigned to any work found to contain plagiarism. The use of Generative AI is prohibited. Any resources used in completing the exam must be cited. To submit your work, open a web browser and navigate to <http://turnin.dinf.usherbrooke.ca>. Select the course IFT870 (BIN710) and the project titled "ExamenIntra." Then, upload the file `examenintra.ipynb` and submit it. The filename must be exactly `examenintra.ipynb`. Your file must include a section with your name, surname, and your matriculation number or CIP.

This exam includes 4 questions and 4 pages.

Question 1: /20 points

Question 2: /25 points

Question 3: /45 points

Question 4: /10 points

Total: /100 points

NAME : _____.

SURNAME : _____.

MATRICULATION NUMBER : _____.

SIGNATURE : _____.

Data Collection:

A dataset containing information about paper publications in Open Access journals was provided. This data has been used to create the resource <http://flourishoa.org/>.

Retrieve the three (3) tables from the dataset on GitHub: <https://github.com/FlourishOA/Data>

Question 1: Exploration-Description (20 pts)

- a) Present a summary of any attribute in the three tables after conducting a quantitative analysis and visualizing the statistical descriptions, if necessary **(20 pts)**

Question 2: Exploration-Preprocessing (25 pts)

- a) Perform data preprocessing by removing duplicates and correcting any inconsistencies, if present. **(15 points)**
- b) Is there a correlation between the journal categories (feature 'category') and the publication cost (feature 'price')? Justify your answer. **(10 pts)**

Question 3: Representation-Classification-Regression (45 pts)

- a) Build an accurate and robust model to predict the values of journal categories. This includes :
- i.* The adapted feature selection;
 - ii.* The choice of a classification model and its parameters.
 - The goal is to determine the best parameter value of « n_neighbors » among the integers 1, 3 and 5 for the « KNeighborsClassifier » model implemented by the scikit-learn library. The classification based-distance model must use the distance *dString* as defined in the Annex section [**Annex**] on page 4.
 - iii.* The computation of the model's score (the one retained in the previous step with the best parameters)
 - iv.* The application of the model to predict the missing category values.

In your opinion, why should the best parameter selected be appropriate for this dataset? Justify all the choices made. **(20 pts)**

- b) Remove the features with more than 50% of missing values. **(5pts)**
- c) Build an accurate and robust model to predict the current publication cost (feature « price ») considering others features. This includes:
- i.* The adapted feature selection.
 - ii.* The choice of a regression model and its parameters
 - The goal is to determine the best parameter values for both of the following models:

- The parameter « n_neighbors » in the range [1, 10] and the parameter « metric » for the « KNeighborsRegressor » model as implemented in the scikit-learn library. The distances to consider for inferring the best metric are Minkowski, Manhattan and Euclidean.
- The parameters « max_depth » in the range [5, 20] and « max_leaf_nodes » in the range [10, 30] for the « DecisionTreeRegressor » model as implemented in the scikit-learn library.

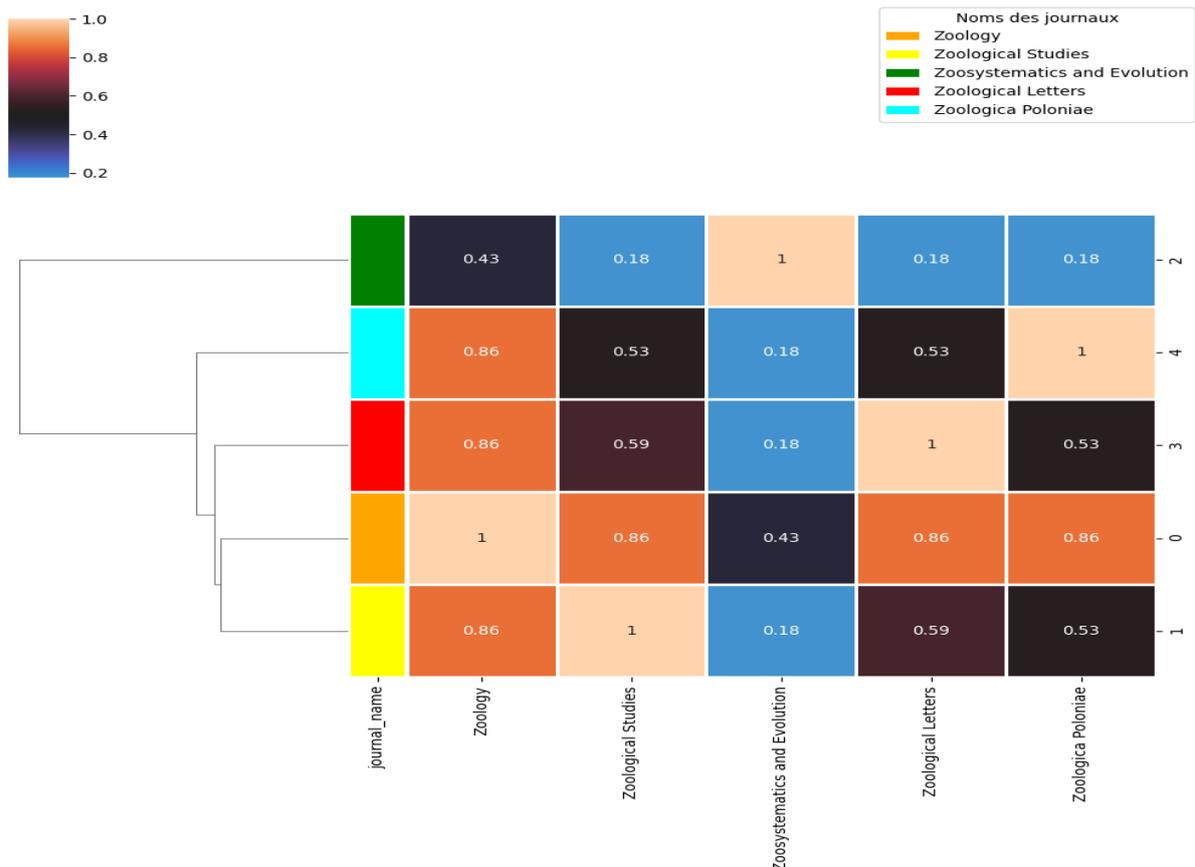
iii. The computation of the model's score (the one retained in the previous step with the best parameters). List the ten (10) journals with the highest deviation (+ or -) values from the predicted ones. Which conclusions can you draw?

iv. The application of the model to predict the missing publication costs.

Justify all your choices. **(20 pts)**

Question 4: Representation-Visualization (10 pts)

a) A figure generated from the data preprocessing of the data is provided. What does figure represent? What observations or conclusions can you draw? **(5pts)**



b) Propose a methodological pipeline and a Python script to generate the figure above **(5pts)**

Annex (*dString*)

Let a et b , two chains of words. $dString(a, b)$ represent the distance between a and b , defined as following:

$$dString(a, b) = 1 - \frac{common_words(a,b)}{\min(|words(a)|, |words(b)|)} \text{ with}$$

$$common_words(a, b) = |words(a) \cap words(b)|$$

and

$\forall x \in \{a, b\}$, $words(x)$ represents the set of words in x .

Example

Note the exclusion of the words: and, &... which are used as delimiters and should not be consider in your distance computation.

$a_1 = Physics \text{ and } Chemistry$

$b_1 = Physics$

$c_1 = Languages \text{ and } Linguistics$

$$dString(a_1, b_1) = 1 - \frac{|{\{Physics\}}|}{\min(|{\{Physics, Chemistry\}}|, |{\{Physics\}}|)} = 1 - \frac{1}{1} = 0$$

$$dString(a_1, c_1) = 1 - \frac{|\emptyset|}{\min(|{\{Physics, Chemistry\}}|, |{\{Languages, Linguistics\}}|)} = 1 - \frac{0}{2} = 1$$

-End of the exam-