CMSC 435 Assignment 3

Fall 2023
(individual work; 8 pts total)

This assignment asks you to design, evaluate and compare models for the prediction of proteins that interact with nucleic acids using a provided dataset.

Dataset
The dataset (*dataset_a3.csv* file) is in the text-based, comma-separated format where each protein is represented by 10 numeric features and 1 symbolic outcome. The outcome feature, which is called "Class", annotates each protein as *Yes* (interacting with nucleic acids) vs. *No* (non-interacting). The dataset includes 8795 proteins, with 936 labeled *Yes* and 7859 labeled *No*.

Development of predictive models
You are required to develop models with the RapidMiner Studio version 10.2 using **four different algorithms**. Two of these four algorithms must be the **Logistic Regression and k-Nearest Neighbor** (these operators are called "Logistic Regression" and "k-NN" in RapidMiner). You can choose any of the other predictive algorithms for the other two selections, however, at **least one of these algorithms must have parameters that can be adjusted** (details below). You should parametrize each of these four algorithms (i.e., select the best possible combination of values of their parameters), to the best of your ability, in order to **maximize** predictive performance that you will quantify with accuracy ("% of correctly classified instances"). You will need to make an educated guess and/or use trial-and-error approach to figure out which **key parameters** make a difference and how to use them, i.e., you will use the key parameters to increase accuracy when compared with the accuracy generated using default parameter values. While you should consider all parameters, select **no more than 3 key parameters** from among all available parameters for a given algorithm. **Do not use the "advanced parameters"**, which means that you do not need to parametrize methods that do not have non-advanced parameters. Do not attempt to sample or modify the dataset, i.e., do not perform feature or sample/object selection; there will be opportunities to do that in the class project.

Testing of predictive models
For each algorithm you must perform three types of tests:
- on the entire dataset ("use training dataset")
- on 50% of the dataset; you will use the other 50% to compute the model ("50% split")
- using the three-fold cross-validation

The 50% split can be implemented in RapidMiner with the "Split Data" operator. The three-fold cross-validation divides the dataset at random into three equal-size subsets, where one subset is used to test the model and the remaining two to compute the prediction model. This is repeated three times, each time using a different subset as the test set. Consequently, this results in predicting every protein in the dataset. This test type is implemented in the RapidMiner Studio with the "Cross Validation" operator where the number of folds is **set to three**.

Deliverables

1. The .rmp file that you created in RapidMiner to solve this assignment. You can create this file by selecting the "Export Process" option under the "File" in the top menu. The file **must be named** a3.rmp.
2. Answers to the following five questions which you will submit in a pdf file:
   2.1. **List and briefly describe** the two algorithms that you selected. You should **name** the algorithms and briefly explain **why** you selected them and what **type of models** they produce.

2.2. Using the table shown below, **report the accuracies** for the four algorithms and the three test types. The accuracy values must be reported with two digits after the decimal point, e.g., 91.05. You must include the accuracies of the models that use the default parameters and the best values of the key parameters. In total, you have 4*3*2 = 24 results to report. **Name the key parameters and list their best selected values** for each model and each test type; leave this part of the table empty if there are no parameters. For your convenience, we provide a template of the table in Canvas.

2.3. You should obtain 100% accuracy for at least one method and one type of test. Which type of test produced this accuracy value? Do you think 100% accuracy is a good result if we assume that data in this dataset, including the yes/no Class feature, is **noisy**?

2.4. **Provide** "confusion matrix" for the **k-Nearest Neighbor model with the default parameters** based on the **three-fold cross-validation experiment**. This is the matrix in the PerformanceVector view. Use this matrix to **explain** whether this predictor would be better suited to identify proteins that interact with nucleic acids (Class = *Yes*), proteins that do not interact with nucleic acids (Class = *No*), or both types of proteins.

2.5. Using the **most accurate result from the three-fold cross-validation tests, briefly discuss** whether trying multiple algorithms and adjusting their parameters helped you in developing a more accurate predictive model compared to the baseline of 89.36%. This baseline accuracy can be achieved if we simply predict all proteins with label "No". Argue **whether or not** this amount of improvement over the baseline is large – try your best to **justify your argument**.

Notes

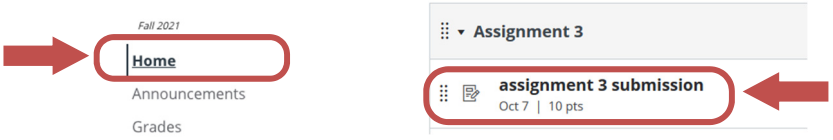
- Late submissions will be subject to deductions: 15% in first 12 hours and 30% for between 12 and 48 hours. We will not accept submissions that are over 48 hours late.
- We will check for **plagiarism**. Develop your own rmp file and provide your own answers.
- The table for question 2.2 must be in the following format; for your convenience this table is provided in the word docx format in Canvas. Example **fake** values are in green font.

Reported information	Test type	k-NN	Logistic regression		
Accuracy with default parameters	Entire dataset	12.34%			
	50%	23.45%			
	Cross-validation	34.56%			
Accuracy with best parameters	Entire dataset	45.67%			
	50%	56.78%			
	Cross-validation	67.89%			
List names of parameters		k measure types			
List selected best values of parameters (in the same order as in the list of names)	Entire dataset	12 MixedMeasures			
	50%	34 NominalMeasures			
	Cross-validation	56 NumericalMeasures			

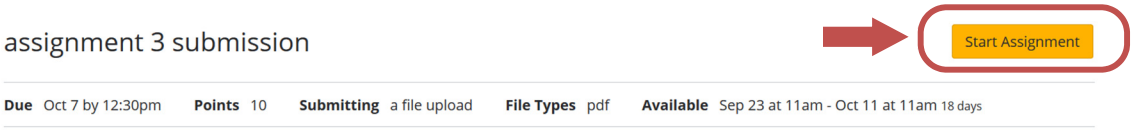
Due Date

Your assignment must be received before 12:30 pm on October 5 (Thursday), 2023. Submissions, which include the two files (a3.rmp and the pdf with the answers), must be done via the class web page in Canvas. 5-step instructions are included below.

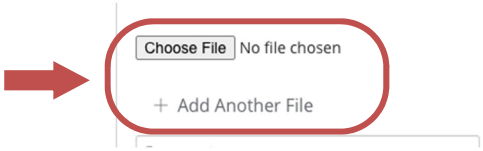
1. Go to the "Home" section, locate the "Assignment 3 submission" field and select it by clicking on the assignment title.



2. Select “Start Assignment”



3. You can select your submission files by clicking on “Choose file”.



4. [Important!] Your file(s) will be submitted only after you click the “Submit Assignment” button.

5. [Important!] Make sure your submission was completed and the correct file was sent.