CMSC 435 Project

Fall 2023
(Group work; 25 pts total)

The project asks your project group to develop, evaluate and compare models for the prediction of proteins that interact with DNA and RNA using a provided dataset. Your model must classify a given protein sequence into one of four outcomes, i.e., interacts with DNA (DNA), interacts with RNA (RNA), interacts with both DNA and RNA (DRNA), and does not interact with DNA or RNA (nonDRNA). Although each group will solve the same task, the corresponding designs must be unique, i.e., collaboration between groups is not allowed. There are 10 projects groups and they were formed at random. You can find your group assignment in Canvas. The instructor will send an email to each group to initiate the communication.

Datasets

Two datasets are provided:

- *sequences_training.txt* (*training dataset*) that includes 391 DNA proteins, 523 RNA proteins, 22 DRNA proteins, and 7859 nonDRNA proteins, for the total of 8795 proteins. This dataset is already available in Canvas.
- *sequences_test.txt* (*blind test dataset*) that includes 8794 proteins, with similar proportions between the four classes of proteins. This is an independent test set, which means that entire design procedure (including feature generation, feature selection, parameterization and selection of classifiers, etc.) should be completed using only the training dataset. The test dataset should be used to evaluate your system only once. This dataset will be posted on the class web site 2 days before the project submission deadline and it will **not** include the annotation of the outcomes. You will have to predict the outcomes and the instructor will process and assess these predictions.

The training dataset is provided in the comma-separated format where each protein is represented by:

- the amino acid sequence
- the class encoded as DNA, RNA, DRNA, and nonDRNA

Test dataset will be the same format as the training dataset, except that the outcomes will not be provided.

Evaluation of Predictions

Your group is required to perform the **5-fold cross validation** when using the *training dataset*. This cross validation divides the training dataset into 5 random, equal-size subsets, where one subset is used to test the prediction model and the remaining four to train/develop the prediction model; this is repeated 5 times, each time using a different subset as the test set. Consequently, this test results in predicting every sequence in the training dataset. This test procedure is supported by RapidMiner and other popular data science software/libraries.

For each of the four outcomes your group will convert the dataset into a binary problem, i.e., a given outcome (positive outcome) vs. all other outcomes (negative outcomes). For example, all proteins that are labeled as DNA will be considered as positive, and the remaining proteins (RNA, DRNA and nonDRNA) as negative. Next, for each of the four outcomes you will compute the following measures:

Sensitivity = SENS = 100*TP / (TP + FN)
Specificity = SPEC = 100*TN / (TN + FP)
Accuracy = 100* (TP+TN) / (TP+FP+TN+FN)
MCC = (TP*TN – FP*FN) / sqrt[(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)]

where TP is the number of true positives (correctly predicted positive outcomes), FP denotes false positives (negative outcomes that were predicted as positives), TN denotes true negatives (correctly predicted negative outcomes), FN stands for false negatives (positive outcomes that were predicted as negatives). You will also compute:

averageMCC = $(MCC_{DNA} + MCC_{RNA} + MCC_{DRNA} + MCC_{nonDRNA})/4$
accuracy4labels = $100*TP_{all}$ / (number of all protein in the dataset)

where MCC_{DNA}, MCC_{RNA}, MCC_{DRNA}, and $MCC_{nonDRNA}$ denote the MCC values when using the DNA, RNA, DRNA, and nonDRNA outcomes as the positives, and TP_{all} is the number of correctly predicted outcomes (DNA proteins predicted as DNA proteins, RNA proteins predicted as RNA proteins, etc.). These measures can be computed based on the confusion matrix. You should **round the values** to one digit after the decimal point when reporting the accuracy, sensitivities, and specificity and to three digits after the decimal point when reporting MCC. **You report must include the confusion matrix for your final/best solution**.

Your group must also provide and **summarize predictions on the *blind test dataset***. To do that you will compute your model using the entire training dataset (using the same design, i.e., features, values of parameters, etc., as in your best 5 fold cross validation result) and you will use this model to predict sequences from the blind test dataset. **In your report, you must discuss the corresponding results on both the training and the blind test dataset**; on the blind test dataset you can summarize your results by explaining and comparing how many proteins were predicted with a given outcome.

Design

Your group should **design** the model to maximize its predictive performance **evaluated based on averageMCC using the 5-fold cross validation on the training dataset**. The design may consider:

- Use of different features to encode the input protein sequence. The data mining algorithms require a rectangular dataset with a fixed size and structure of the feature vector for each object (protein). Thus, you will need to convert the input protein sequences (that have variable length) into a fixed set of (numerical) features. Lecture set 7 includes a few suggestions.
- Selection of a subset of the input features. This could potentially speed up computation of the model, remove weak/noisy features, and reduce overfitting. Feel free to combine results of multiple feature selection methods.
- Selection of a classification algorithm that you will use to compute your model from among many algorithms that are available in RapidMiner.
- Parametrization of the selected classification algorithm(s). This involves identifying and setting values of their key parameters.
- Building a system with multiple models that are used together. For instance, you could use multiple models that predict all 4 classes and combine their results together to generate one prediction. Check the methods in RapidMiner at Operators → Modeling → Predictive → Ensembles.

IMPORTANT NOTE 1: Ensure that your team perform all design activities (e.g., feature selection, selection and parametrization of the classification algorithms, etc.) using the **5-fold cross validation** on the training dataset. Otherwise you could overfit this dataset and your results on the blind test dataset will suffer. Also, make sure that you do not sample the test folds in the cross validation, i.e., the cross validation results must be done on the five test folds that collectively include the 8795 proteins.

IMPORTANT NOTE 2: Your team's design should be done **incrementally**. Start with a simple initial solution (complete the entire design, prediction, and prediction assessment process) and gradually make your design more sophisticated with the goal to improve the predictive performance. The progress report checkpoint should be based on a simple initial solution. In your final report, you should clearly indicate **one** best set of results, which must be selected based on the cross-validation results on the training dataset. Moreover, these results should be compared with your intermediate results (earlier/simpler designs, other alternatives, etc.) and with baseline results shown in Table 1, in order to justify your design choices. **In your final report, provide your results by adding them into Table 1.** This will make it easy to compare the different alternatives. **Clearly indicate which result is the best/final**. You should explain how you made decisions that led you a certain direction of redesigning/improving your model. You also should provide a convincing argument why and how your method is good/competitive in comparison to the **real baseline result** listed in Table 1.

Table 1. Predictive results based on the 5-fold cross validation on the training dataset (this table is available in Canvas).

Outcome	Quality measure	Baseline result	Design 1	Design 2	Design 3	Best Design
DNA	*Sensitivity*	6.9				
	Specificity	99.3				
	Accuracy	95.2				
	MCC	**0.132**				
RNA	*Sensitivity*	39.6				
	Specificity	98.9				
	Accuracy	95.3				
	MCC	**0.501**				
DRNA	*Sensitivity*	4.5				
	Specificity	100.0				
	Accuracy	99.7				
	MCC	**0.122**				
nonDRNA	*Sensitivity*	98.6				
	Specificity	29.8				
	Accuracy	91.3				
	MCC	**0.428**				
averageMCC		**0.296**				
accuracy4labels		90.8				

Deliverables

Each group shall provide the following five deliverables:

1. **Team project contracts** (one per team), filled in and signed.
2. **Progress report** (pdf file for parts a and b; dataset file for part c) that consists of:
   a) **Description of the first attempt to make predictions**. You should use short bullet points to list the features that you generated from the input sequences; list major data processing steps that you used to prepare the data; and name the classification algorithm that you used. This is supposed to be an initial attempt so we expect to see simple models and just a few bullet points.
   b) **4x4 confusion matrix and the four MCC values** (MCC_{DNA} + MCC_{RNA} + MCC_{DRNA} + $MCC_{nonDRNA}$) that the above model has produced.
   c) **Training dataset file** in txt/csv format. This file is the rectangular dataset with a fixed set of features for each object (protein) where the last (right-most) feature is the class. This is supposed to be an initial version of the dataset and so we expect to see a small and simple feature set.
3. **Final Report** (pdf file) that consists of:
   – **Cover page** that gives the class number and title, date of your submission, name of your group and names of all team members.
   – **Description of the design of the prediction system**. You should briefly **explain** the features that you generated from the input sequences; **how** and which features were selected; **which** classification algorithms and their parameters you tried and **why** and **which** you have chosen; and **which** other design options you considered and applied.
   – **Results** (see *Evaluation of Predictions* section). You must **organize the results in a table** using the format of Table 1. Using this format, compare your best cross validation results with the results from earlier/alternative designs and with the results shown in Table 1. Include confusion matrix for your best solution. Summarize predictions for the *blind test dataset*.
   – **Conclusions**. This is a **very important part** of your report. You should **comment on the quality of your results** and **compare** them against the baseline results from Table 1. Also, describe **your experience** in this project, and explain **advantages** and **disadvantages** of your method and why you think your results are good or bad, in comparison with the other results from Table 1.
4. **Predictions on the *blind test* dataset**. These predictions should be submitted via email (see *Deadline and Delivery* below) as a text file named with the name of your group, where each row provides prediction for a given "blind" protein. The format should be as follows:
   DNA
   DNA

RNA
nonDRNA
…

where DNA, RNA, DRNA and nonDRNA are the predicted outcomes for the protein from the same row in the *sequences_test.txt* file. The instructor will use these results to evaluate your method on the blind test dataset against the true classes, and these results will be forwarded to you as part of the evaluation of your project.

5. **In-class presentation**
   - 10 minutes long plus 3 minutes for questions&answer session
   - must describe the design, results and conclusions
   - must include the following parts:
     - Motivation for your design. Briefly explain how you arrived at your final design.
     - Description of your design. Explain (preferably with a diagram) how your method makes the predictions.
     - Discussion and comparison of the quality of the achieved best results using the results on the training dataset and Table 1.
     - Conclusions. This part is essential; see the conclusions part of your report.
6. **Statement of contributions**
   - A short document with bullet-point style list of detailed contributions to the project for each team member. The contributions cover all aspects of the project including conceptualization and design of the methodology, implementation, testing, writing the report, preparing the presentation, making the presentation, coordination of the work, notes taking, organizing group meetings, submission of deliverables, etc.
   - The contribution list for each team member should be accompanied with an estimated fraction of the total project effort, quantified in %. The effort estimates across the team members must sum up to 100%. Each team should strive to balance the effort to be equal across team member.
   - Submission of the bullet-point detailed contributions is **optional** in case when the contributions of all team members are equal.
   - Our objective for this statement is that the workload is divided evenly across team members. You should not be competing who will do more but encourage everyone to contribute equally. This statement will be used to distribute the project grade among the team members.

Marking

The evaluations of the progress report, project report and predictions constitute **15% of the final mark from the course** and it will consist of the following four parts:

1. 5% for the quality of the progress report
2. 3% for the quality of the final report
3. 3% for the quality of the design of the prediction method from the final report
4. 4% for the quality of the predictions measured using the 5-fold cross validation on the training dataset from the final report and on the blind test dataset

IMPORTANT NOTE 3: For item 4, the *averageMCC* is the main predictive quality measure that will be used to evaluate submitted solutions but the conclusions **must discuss** the other quality indices as well. **Bonuses of 3%, 2%, and 1% will be given to the project submissions that secure the highest, the second highest and the third highest value of *averageMCC* on the blind test dataset.** In case of a tie the winner will be decided based on the higher value of the *accuracy* on the blind test dataset.

IMPORTANT NOTE 4: For item 4, MCCs that are high(er) relative to other submissions or to the baseline in Table 1 are not necessary to receive a full mark. The key is to show substantial progress from the initial solution – you should show and discuss how your best design is better when compared to your own alternative solutions and explain advantages compared to the baseline results in Table 1.

The in-class presentation constitutes **10% of the final mark from the course** and will be evaluated by the instructor and your peers. The grade will consist of three parts:

1. Grade assigned by the fellow students (peer evaluation) (**4%**). Each project group will complete a short evaluation form online, see appendix A, to assess presentations of other groups. Instructor will gather and process these grades; they will be kept confidential.
2. Instructor's grade (**6%**). Instructor will grade the quality of presentations using Appendix B.

Deadlines and Delivery

- Filled in and signed **team project contracts** (one per team) must be returned to the instructor **by email** (to lkurgan@vcu.edu) by **October 26 (Thursday), 2023 before 12:30pm**. The project contract form is available in Canvas and should be editable using pdf reader applications.
- Submission of the progress report deliverables (pdf file accompanied by the file with the training dataset; two files in total; one submission per group) to the instructor **by email** (lkurgan@vcu.edu) as is due on **November 13 (Monday), 2023 before 12:30pm**. This pdf file should include the bullet-point description, confusion matrix and MCC values. Make sure to clearly identify the name and members of the submitting group. Late submission penalties apply.
- Submission of the project deliverables (final report pdf file and the test predictions; two files in total; one submission per team) to the instructor **by email** (lkurgan@vcu.edu) as is due on **November 30 (Thursday), 2023, before 12:30pm**. Late submission penalties apply.
- The presentations will be delivered on **December 5 and 7, 2023, at 12:30pm** (during the last two lectures); each date will feature five project groups. The schedule will be posted on Canvas. Presentation slides (one per team) in pdf format must be sent to the instructor **by email** (lkurgan@vcu.edu) by **December 4 (Monday), 2022, before 12:30pm**. **This is a sharp deadline – late submissions are not allowed and will receive 0.** We will share the slides via Canvas.
- The contribution statements (one per team) must be submitted to the instructor **by email** (lkurgan@vcu.edu) by **December 11 (Monday), 2023, before 6:00pm**. Submission of the bullet-point detailed contributions is **optional** in case when the contributions of all team members are equal, but the entire team still must send an email that confirms equal contributions by the deadline.

Final Notes

- While we recommend the use of RapidMiner, you can complete this project using any other data science software or language. Just make sure that you will be ready to make predictions on the blind test dataset using your selected software.
- Do not cheat (e.g., do not inflate or "tweak" the results). It is better to report honest results than to get caught cheating. In the latter case you are risking receiving 0 marks for the project.
- Your team may be asked to demonstrate how the prediction works, in case of the reported results are irregular. Thus, make sure to retain your software at least until the time of the final exam.
- Always copy the email communications to yourself so you can prove that it was sent.
- Contact the instructor immediately if problems occur.

Appendix A

CMSC 435 Intro to Data Science
Fall 2023
Peer Evaluation Form for the In-class Project Presentations

Name of the presenting group ……..……

Remarks:

- For each question enter grade between 0 and **20** or between 0 and **10** (0 being the worst, 30 or 10 being the best)
- Optionally please add comments (both positive and negative); they will be passed along to the presenting group.
- Average of these grades submitted for a given group will constitute the peer evaluation component for the project presentation.

remarks	*grade*
Quality of Presentation Did you find the presentation interesting? Were the presenters prepared? Did you understand the topics covered in the presentation? How much did you learn? Was there anything significant missing? Were the conclusions and discussion of results covered sufficiently? How would you rate handling the discussion/questions?	min 0, max **20** ……
Presentation Style Quality of presentation style – Was it finished on time? Too fast/slow? Well presented? Was the presenter just reading the slides or was (s)he presenting the material beyond the content of the slides? Was there an eye contact?	min 0, max **10** ……
Quality of Slides Quality of slides – Did you find the slides too crowded? Too brief? Too many? Easy to read? Was the layout of individual slides appropriate and consistent? How was the overall quality of the organization, in terms of the order and flow of the slides?	min 0, max **10** ……
Additional Comments	

Appendix B

CMSC 435 Intro to Data Science
Fall 2023
Instructor's Evaluation Form for the In-class Project Presentations

Name of the presenting group ……..……

TASK	***comments***	***grade***	*max grade*
Presentation finished within 8 minutes limit	(up to -10 points penalty)		Y / 0
Quality of *Motivation for the proposed design*			10
Quality of the *Description of the proposed design*			10
Quality of the *Discussion and comparison of the quality*			10
Quality of *Conclusions*			20
Quality of the Presentation and Presentation Style			10
Instructor's total mark			**60**