

Course: Introduction to Data Science (DS2006) - Class Project (5 Credits)

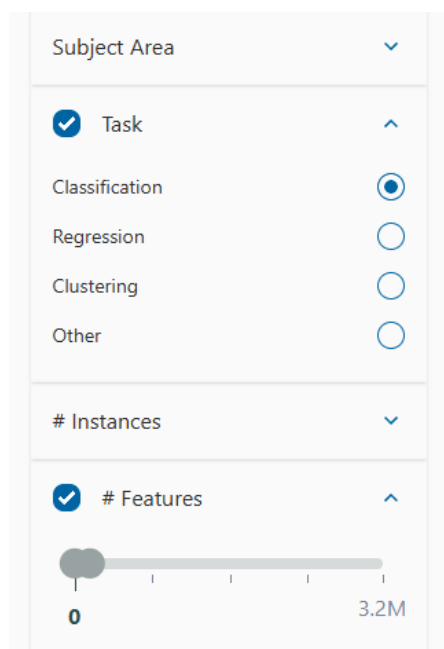
Dear Students, over this learning period we have learned many cool things about programming and data science. Now is the time to combine all that knowledge in a project that will further enhance your learning!

The idea of this year's project is that we develop a software that will perform operations on a specific dataset and perform the classification of new unseen examples during the training and validation of the classifiers.

The project has a few requisites:

1. You must choose a dataset that allows you to perform **classification**. My suggestion is to look at this link which has several UCI datasets: <https://archive.ics.uci.edu/datasets?Task=Classification&skip=0&take=10&sort=desc&orderBy=NumHits&search=&NumFeatures=0&NumFeatures=5>

Note that on that website you have useful filters (Figure 1), such as filtering by task (ensure classification is on) and also number of features (We **highly** recommend you work with a dataset with 4 to 10 features for the class project).



The image shows a vertical sidebar of filters for searching datasets. At the top is a 'Subject Area' dropdown menu. Below it is a 'Task' section with a checked checkbox and a list of radio buttons for 'Classification' (selected), 'Regression', 'Clustering', and 'Other'. Further down is a '# Instances' dropdown menu. At the bottom is a '# Features' section with a checked checkbox and a horizontal slider ranging from 0 to 3.2M, with a dark grey knob positioned near the 0 mark.

Figure 1. Example of dataset search filters from <https://archive.ics.uci.edu/datasets?>

2. It is important that one dataset (be it from UCI or somewhere else) is used only by one team (project pairs). In order to make sure no one is already using your dataset, we have a forum in Blackboard where you can mention which dataset you will be using. Datasets will be claimed on a first post, first get basis.
3. You will need to develop an application around the classification task. Ideally you want to create a menu, which will allow the user to:
 - a. Load the dataset.
 - b. Train a classification model with the current version of the dataset.
 - c. Evaluate and save the performance of the classification model.
 - d. Simulate a real environment.

Some additional observations:

Regarding 3.a: Load the Dataset: The software must allow the user to type the dataset name, tell the user it was loaded correctly and then print:

- The first top 10 rows.
- Some basic statistics.

Regarding 3.b: Train a classification model with the current version of the dataset: The software must allow the user to select between at least 2 types of supervised machine learning algorithms to train the models.

If no dataset was loaded, the software must ask the user to load a dataset first.

Regarding 3.c: Evaluate and save the performance of the classification model:

The software must enable the user to choose whether or not they want to load a specific file for evaluation.

In case the user does not provide the additional file, it performs the experiments using a suitable data partitioning strategy.

After performing the experiments it will report suitable metrics for the user which will allow them to evaluate the performance of the classification model.

After showing the metrics it will ask the user whether or not they want to save the results to a file.

If the user says yes, the software asks for the file name and saves the results information there.

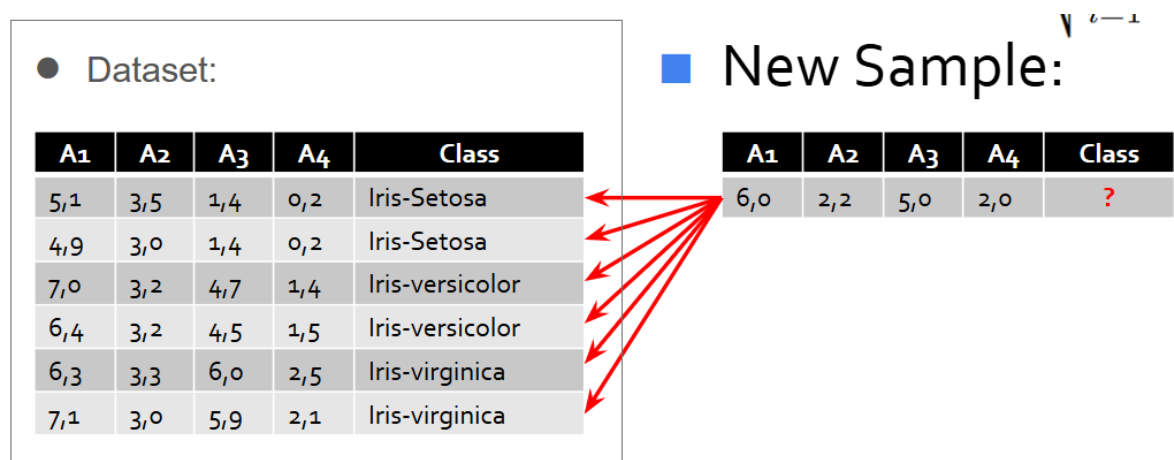
Regarding 3.d: Simulate a real environment where the user can input a new unseen example without the class information and the system will return the result from using the classification model.

If the user did not train any models yet, the software must tell the user that first they need to train a classification model.

If the user trained methods, then it asks for the user to input information related to the attributes used for the problem and uses that information to ask the classification model for the correct response.

The output of this method is the predicted class.

If for example, we were working with the Iris dataset, this method would allow us to type in the values for the attributes A1, A2, A3 and A4 as shown in the new sample in Figure 2.



● Dataset:				
A1	A2	A3	A4	Class
5,1	3,5	1,4	0,2	Iris-Setosa
4,9	3,0	1,4	0,2	Iris-Setosa
7,0	3,2	4,7	1,4	Iris-versicolor
6,4	3,2	4,5	1,5	Iris-versicolor
6,3	3,3	6,0	2,5	Iris-virginica
7,1	3,0	5,9	2,1	Iris-virginica

■ New Sample:				
A1	A2	A3	A4	Class
6,0	2,2	5,0	2,0	?

Figure 2 - Illustration of the implementation of requirement 3.d.

Project Evaluation

For evaluating the class project, we are going to have **in-person interviews** with the teams, where all members of the teams must show that they are able

to understand and modify everything about the code. During these interviews, we will ask you **individually** to do at least one of these things:

- Explain how something specific works,
- Change things in the code,
- Add things to the code,
- Remove things (with/without breaking) the code
- Fix small errors we might have introduced in your code.

During the interview you will not be allowed to consult other sources of information or your colleagues. Therefore it is really important that you own the whole code, answers such as “**I did not do that part**” will result in a **failing grade** regarding the project evaluation.

Also it is important to highlight that the code must be working properly for the students to obtain a passing grade.

Project Evaluation Logistics

For doing the interviews we are providing dates you will use to book a project interview. Note that they are all related to the **first evaluation** opportunity, and **each team gets only one time-slot** regardless of which grade they obtain in the code interviews. Therefore, even though projects are done in pairs, the evaluation is individual, and it is possible to have a situation where one student from the team obtains a **passing grade** and the other obtains a **failing grade**.

Project Evaluation Re-Examinations

In case a student obtains a **failing grade**, we will have re-examination opportunities. We will do our best to have them around the same time as the 2nd and 3rd written exams re-examinations opportunities.