**Goal:** Explore the relation among the notions of data, information and knowledge. Apply those concepts in specific scenarios and experiment (and develop) the techniques that allow to "move" from data into knowledge. Explore the classification and clustering that are nuclear to data-mining, knowledge discovery, machine learning and information retrieval. The techniques resort to statistics (e.g., 1R and Bayes rule), induction of decision trees (e.g., J4.8/C4.5, ID3) instance based (e.g., KNN with KDTree support). Develop the competence of analyzing, modeling and validating a data-mining project. Use tools to manage data (e.g., PostgreSQL), to discover knowledge (e.g., Orange data mining) and to implement specific algorithms (e.g., via Python).

**Scenario A:** The medical center "MedKnow" uses a database management system (DBMS) that contains all the data gathered, throughout time, about each patient's visit to a doctor (that works at "MedKnow"). The current goal of the ophthalmology team is to analyze all the information accumulated (throughout time) in order to extract the patterns that provide useful indicators to support the prescription (and diagnosis) activity. To achieve that goal they decided to contact the "SoftKnow" company and to send them the file "`d01_lenses.xls`" with a data snippet (related with the lenses prescription activity) and, in a line, they wrote the following challenge: "*send us a prototype of a system that would provide "MedKnow" not only the operational (daily-work) support but also the strategic perspective (useful patterns) that can be extracted from that daily-work data*".

**Project Items:**

Assume that your working group represents the "SoftKnow" company and that you teacher represents "MedKnow". In order to provide a reply to the "MedKnow" challenge your company ("SoftKnow") must develop the following items:

1. Analyze the data snippet (sent from "MedKnow") and write down (in English or Portuguese) your own assumptions about the daily-work of the "MedKnow" activity. Do not exceed 1/2 A4 page.

2. Assume that the daily-work of "MedKnow" is supported by a relational database that includes, at least, the entities (tables) `PATIENT`, `DOCTOR` and `DISEASE`. Also assume that, at least, the attributes `birthDate` (of type `date`) and `diseaseName` (of type `varchar`) are used to describe some of those entities. Notice that "myope", "hypermetrope" and "astigmatic" are the `diseaseName` represented in the received data snippet.

3. Build the conceptual data model (use Entity-Relationship notation) for the operational (daily-work) support. Be aware that this conceptual model must include the entities and attributes described in the previous item and support your own assumptions (about the daily-work of "MedKnow" activity) as you described in the first item.

4. Write down the logical model that derives from the previous item and implement using scripts to automatically create the database, the constraints and to automatically populate the database (with some predefined data).

5. Implement a script to automatically export a dataset with the proper format for Orange data-mining framework.

6. Implement the 1R classification method (use Python). Optionally integrate it in the Orange framework.

7. Provide a deployable model (solution) to "MedKnow" strategic needs using 1R method (your implementation).

8. Provide a deployable model (solution) to "MedKnow" strategic needs using ID3 and Naïve Bayes methods.

9. Provide "MedKnow" an evaluation of all methods and your own conclusions on the most reliable approach.

*Remarks*: 1) you may extend the provided data snippet "`d01_lenses.xls`" with additional examples in order to better explore and evaluate the mining methods; 2) you may explore additional methods (apart from 1R and ID3).

**>> How can your team (company) "take additional profit" from the added-value of your project?**

Recall that the "operational" daily-work of "MedKnow" can be enhanced with the classification model that your team have developed because now you can "*provide the user with a hint about the type of lenses to prescribe given the information on the remaining features*". Your classification system may communicate via shared data (e.g., your system may read the features from a table and write the expected class-value also into a table) or your system may also be made available via a service (e.g., a REST Web service). With this goal "in-mind" the operational perspective can be enhanced with "strategic-hints"; also, the feedback from those hints can be used for the system to evolve and improve the precision of such "strategic-hints" (e.g., by rebuilding the classification model from new datasets).

**Important dates (deadlines) and deliverables:**

- week 02.OCT-06.OCT (during the 3h class) – the 1 to 4 (above) items must be sent and, if possible, presented (to teacher) during the classroom. As an additional challenge present also item 5. Organize a presentation using slides and illustrate your approach by executing the implemented scripts.

- 31.OCT.2023 – deliver all the project elements described below (in the next "Rules" paragraph).

**Rules**:

- Deliver a project report, in .pdf format, with at most 10 pages; the first page must include the discipline and course names, a title, the working group number and each student's name and number.
- Deliver, in electronic format (file named AMD_XX.zip, where XX is the working group number), the project report and all the information regarding the implemented system; the ".ppt" (powerpoint), ".bat", ".exe", ".py", ".tab" (data) and any other file needed to properly execute your solution. It is very important to include a `README.txt` file with step-by-step instructions on how to properly execute your solution.