Lab IA & CD

Project 1 - Lung Cancer Classification using Computerized Tomography (CT) Data

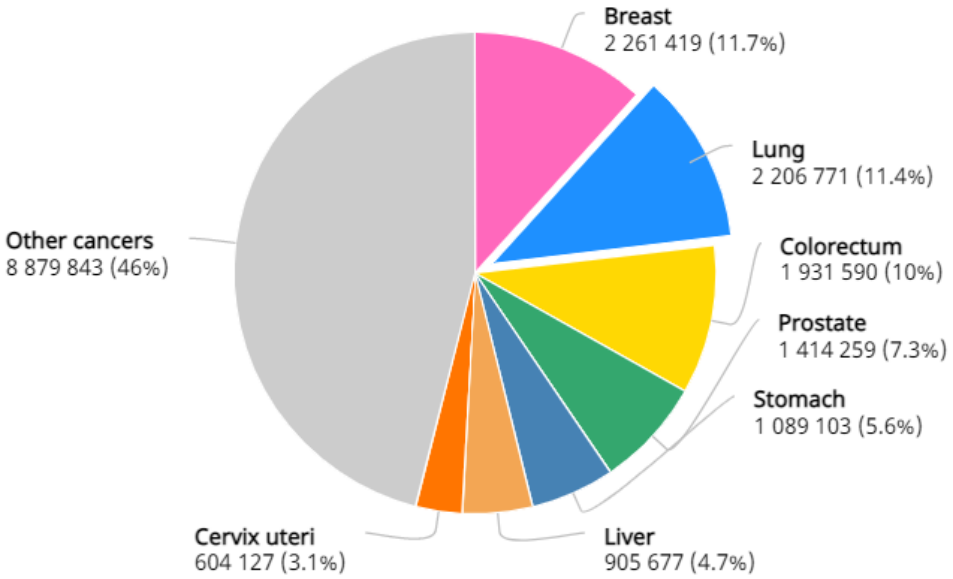
Introduction

- Provide students with:
  - software development methodologies, AI and DC projects, teamwork and communication through the implementation of projects designed for this purpose.

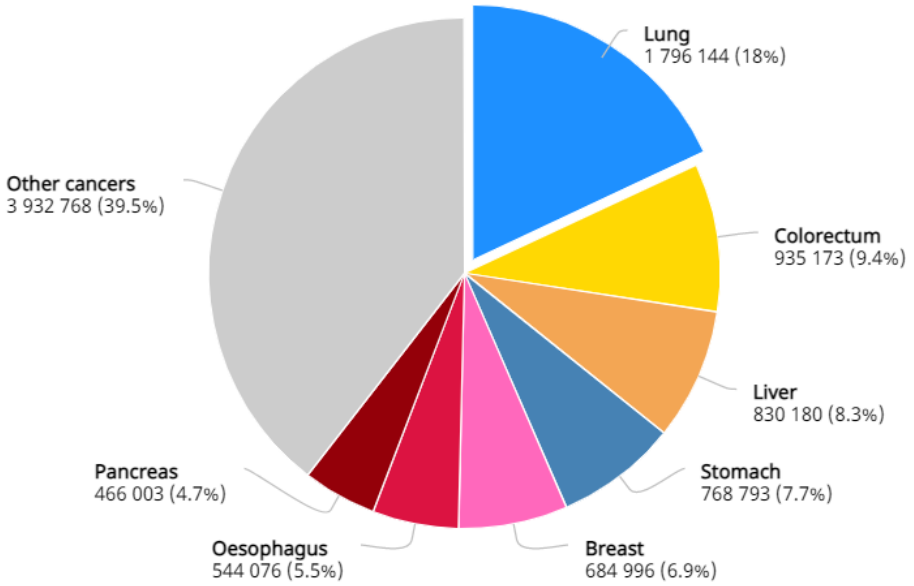
- Students should apply the knowledge obtained from the courses from previous years and research methodologies to solve the problem.

The Problem

- Lung Cancer Classification using Computerized Tomography (CT) Data



Total: 19,292,789
Cancer Incidence by Type in 2020

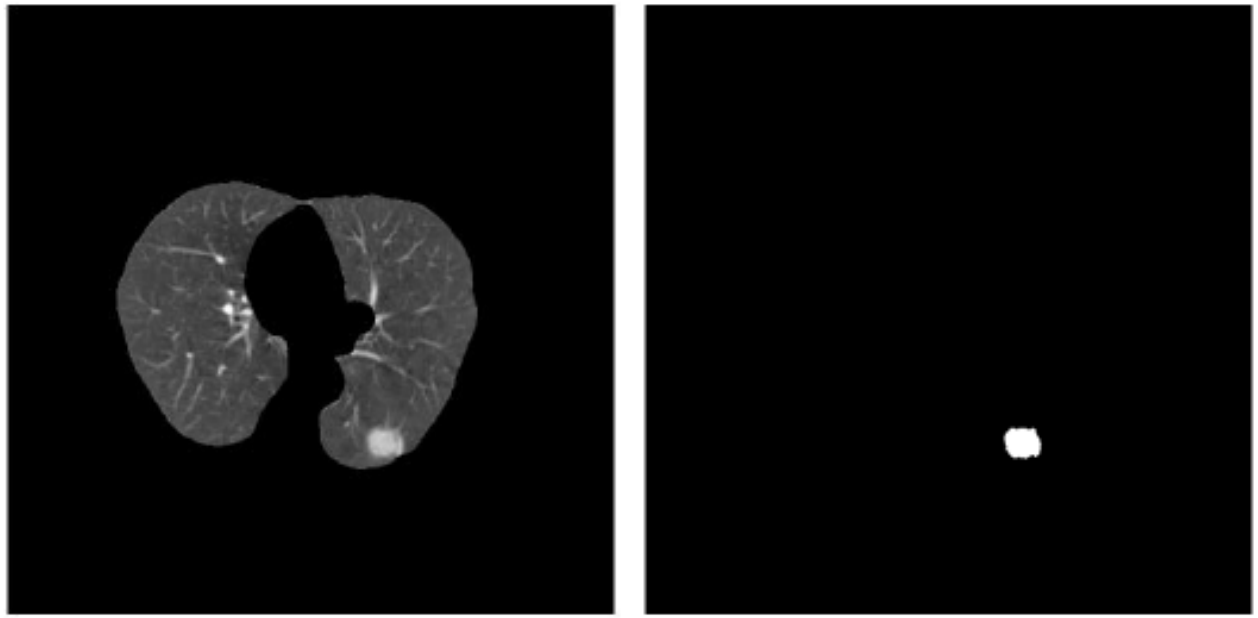


Total: 9,958,133
Cancer Mortality by Type in 2020

Data source: GLOBOCAN 2020
® International Agency for Research on Cancer 2021

Data Set

- LIDC – IDRI



CT slice

Fine Segmentation mask

Data Set

- LIDC – IDRI
  - CT scans (stack of images) of 1010 patients as a DICOM file;
  - Annotation (XML)
    - 3 categories
      - Nodule >= 3 mm
      - Nodule < 3 mm
      - Non-nodule >= 3 mm
    - Position of the nodule/non-nodule
    - Patient clinical information

Work to develop

- You should prepare a Data Science-based solution to solve the problem proposed: Lung Cancer Classification using Computerized Tomography (CT) Data.
- You should share your solution in a gitlab/github (share the link in moodle by the second practical class: 1st of October);
- The solution should be delivered in moodle by November 3, 2024, at 23:59:59.

Submision of the solution

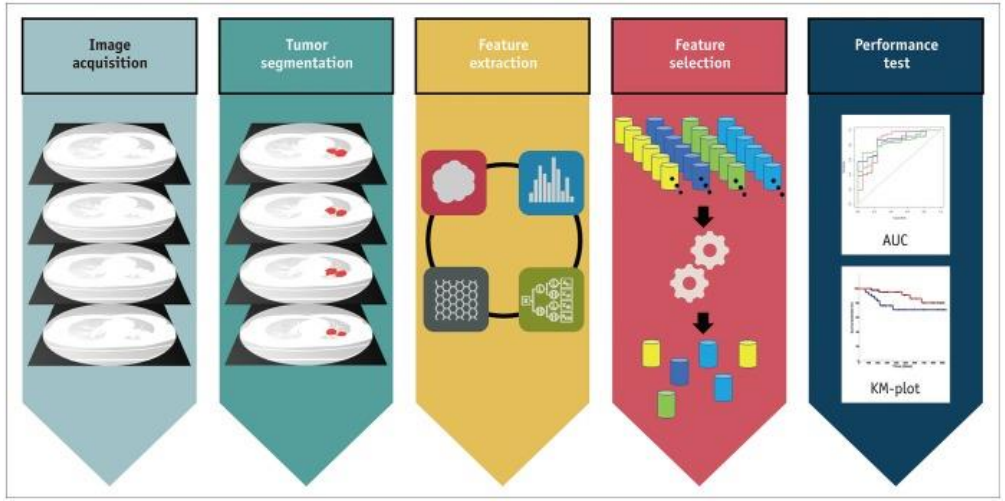
- Final code solution, as a notebook;
  - you should document your notebook, explaining your decisions and discussion about the results obtained;
- Link for a video summary. This is a team video, but each member should participate in it. This is a very short and to-the-point video (maximum of 5 minutes), summarizing the following:
  - the problem;
  - your solution;
  - the results and the impact you think this has.
- One-page document, including possible ethical and legal implications and the framework for current and future regulation issues.
- Auto-evaluation file provided by Professors

Guidelines for the solution

- Assessed data quality and the need for data cleaning. If necessary perform cleaning of the data relevant to the model;
- Perform data pre-processing steps (e.g. range of values (Hounsfield unit 5, 2D vs 3D solution);
- Performed EDA (Exploratory Data Analysis);
- Performs Feature Engineering (e.g. Radiomics, Deep Features) and Selection;
- Discusses model/algorithm and technique selection, as well as model/algorithm optimization;
- Chooses performance metrics and performs validation;
- Explores model interpretability and fairness;
- Performs visualization of results;
- Why not consider other datasets to improve the generalization of the model?;
- Shows good programming skills (best practices, code commenting, performance, speed).

Guidelines for the solution

- Some inspiration can be found in the work of Lee et. al [1].
- You can use a CRISP-DM-based methodology or other to develop your solution



[1] Bak SH Lee HY Lee G, Park H. Radiomics in Lung Cancer from Basic to Advanced: Current Status and Future Directions. Korean J Radiol, 2018

https://www.datascience-pm.com/crisp-dm-2/

Evaluation Criteria

- 15% Product: understanding the needs of the end-user and if your proposal solves that problem;
- 20% Business: understanding if the solution serves the business purpose, its applicability and impact;
- 40% Technical Skills: overall technical evaluation of the solution from a data science point-of-view;
- 15% Soft-Skills: essentially - your communication skills;
- 10% Ethical and Legal Considerations: understand if you understand it for this specific area of application.

Some Tips

Be creative in your solution! Think of how you can use certain approaches in an unusual way for example.

- Consider business constraints: understand the challenge well and identify any business constraints regarding this challenge;
- Mention the constraints you are considering for the solution in the notebook;
- Work as a team: The time is very short, our suggestion is that you distribute tasks well amongst the team;