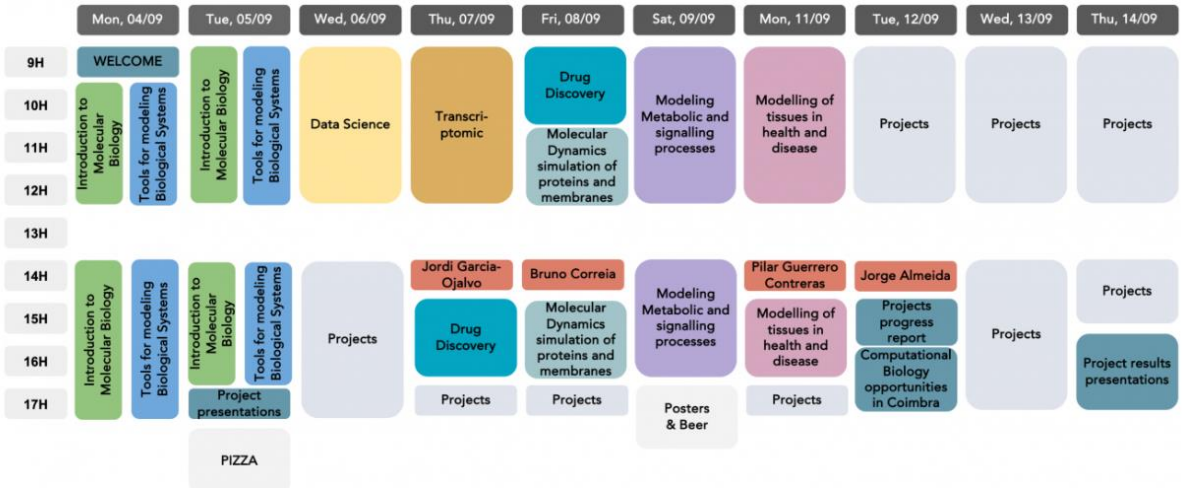
2023 SUMMER SCHOOL: COMPUTATIONAL BIOLOGY

Project: Data-driven Prediction of Alzheimer's disease.
Supervised by: Anuschka Silva-Spínola → LARN, CISUC – Prof. Joel P. Arrais

To answer two questions → 1) Who will progress to Alzheimer's disease (AD)?
2) When will these mild cognitive impairment (MCI) patients progress?

OBJECTIVES = To implement **machine learning algorithms** with the intent of characterizing MCI patients and to develop **time-to-event models** of progression to AD.

Course timeline →



WORK PLAN leading up to the project results presentation →

Wednesday 06

14h – 15h: Presentation of the Project Guidelines and Q&A + Clarifications.
15h – 16h: Discussion on the methodology and Initial planning + Brainstorming.
16h – 18h: Literature revision.

Thursday 07

17h – 18h: Data creation and Data preprocessing.

Friday 08

17h – 18h: Start building the Classification models.

Monday 11

17h – 18h: Start building the Longitudinal models.

Tuesday 12

09h – 11h: Model Refinement.
11h – 13h: Results analysis and Interpretation.
FINALIZE PROJECT REPORT

Wednesday 13

09h – 13h: Q&A + Clarifications and Create presentation slides.
14h – 18h: Finalize presentation slides.
Thursday 14

09h – 13h: Rehearse and prepare for the final presentation.
14h – 15h: Final Q&A + Clarifications.

METHODOLOGY →

1) **Literature revision – References:**

- Diagnostic criteria:

Jack CR Jr et al. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. Alzheimers Dement. 2018 Apr;14(4):535-562. doi: 10.1016/j.jalz.2018.02.018.

Albert MS et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 2011 May;7(3):270-9. doi: 10.1016/j.jalz.2011.03.008.

- Neurophysiological changes:

Jack CR Jr et al. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurol. 2013 Feb;12(2):207-16. doi: 10.1016/S1474-4422(12)70291-0.

Mattsson-Carlgren N et al The implications of different approaches to define AT(N) in Alzheimer disease. Neurology. 2020 May 26;94(21):e2233-e2244. doi: 10.1212/WNL.0000000000009485.

- Computational models:

Ezzati A et al. Optimizing Machine Learning Methods to Improve Predictive Models of Alzheimer's Disease. J Alzheimers Dis. 2019;71(3):1027-1036. doi: 10.3233/JAD-190262.

Sanz-Blasco R et al. Transition from mild cognitive impairment to normal cognition: Determining the predictors of reversion with multi-state Markov models. Alzheimers Dement. 2022 Jun;18(6):1177-1185. doi: 10.1002/alz.12448.

Tahami Monfared AA et al. Estimating Transition Probabilities Across the Alzheimer's Disease Continuum Using a Nationally Representative Real-World Database in the United States. Neurol Ther. 2023 May 31. doi: 10.1007/s40120-023-00498-1.

2) **Dataset creation:**

a) **R programming**. Based on the repository: https://github.com/aridhia/alzheimers-synthetic-data

Instructions for download:

From the repository download the zip file from the “code” button.

Save and extract the folder’s files.

In R Studio open project in new session and upload the alzheimers-synthetic-data-master.Rproj

Navigate the **Build** tab and click **Install and Restart**.

Execute the create_data.R file from the **Files**.

The create_data.R file is going to create and store the dataframes in a folder called mockep_data at the **Files**. Those .csv files need to be imported as dataframe/tab, therefore left click on them and **Import Dataset...**

If any other function needs to be called, always attach the package prior to the insertion of the name.

Datasets created:

vital_signs
socio_demographics
cdr
apoe
csf
gds
mmse
family_history
volumetric
life_style

The conformation and labels of each dataset are based on the documentation from the European Prevention of Alzheimer's Dementia Consortium.

b) Use the provided file. → Prior to it, answer the following questions: 1) What would be the appropriate sample size? 2) What variables should be transformed and why? 3) What variables would provide the most valuable information?

3) Dataset transformation:

From all those datasets available we chose the following variables/features:

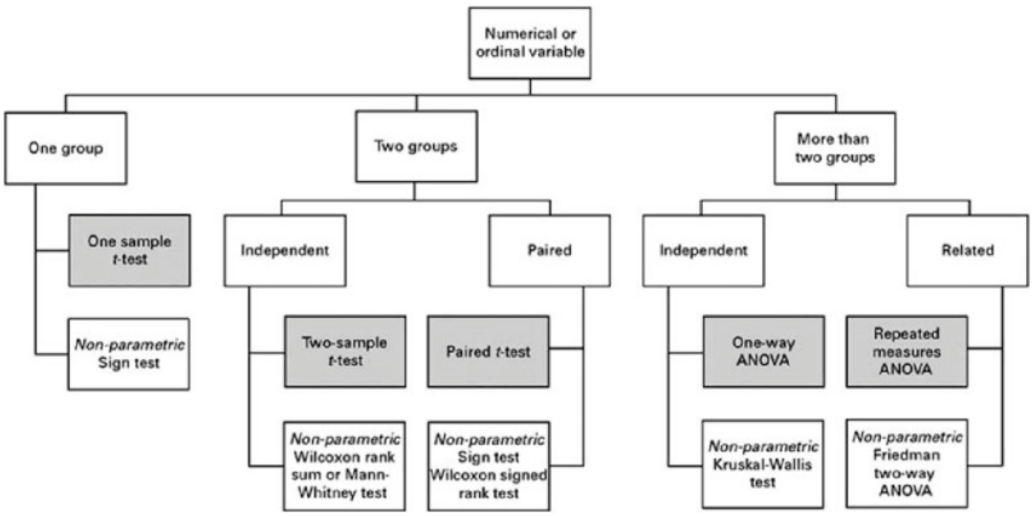
Demographic information = ID, sex, age, years of education.
Clinical information = Family history of dementia.
Genetic information = ApoE genotyping.
Fluid biomarkers = Cerebrospinal fluid (CSF) concentration of Aβ42, p-Tau181, and t-Tau.
Neuropsychological tests = Mini-Mental State Examination (MMSE), Geriatric depression scale (GDS), and Clinical dementia score (CDR) [total and sum-of-boxes score].
Neuroimaging = Hippocampal and white matter lesions volume.

- ➔ Transformation: Add the columns for the ATN scheme, presence of ApoE ε4, and the modifiable lifestyle factors: cerebrovascular disease based on white matter lesions & depression from GDS total score > 11 points.
- ➔ Diagnosis should be built upon CDR total score: controls = 0; MCI = 0.5; dementia > 0.5.

4) Descriptive statistics:

- Full analysis of the groups according to their diagnosis: summarize the central tendency, dispersion, and shape of the dataset.
- Test of normality: Shapiro-Wilk.

- Group comparison by table or figure (could be boxplot with group comparison test – p value).



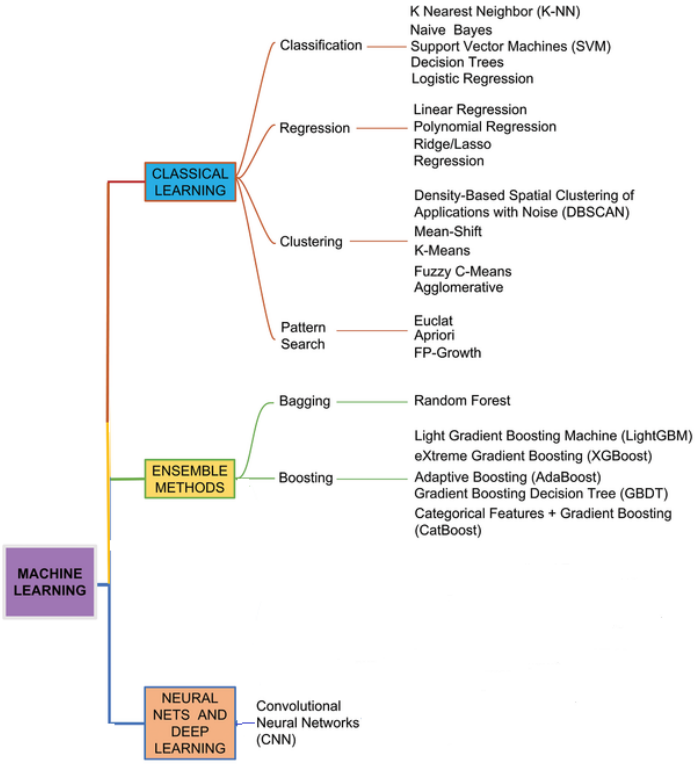
Normal distribution/non-normal distribution.

Code example:

https://www.geeksforgeeks.org/data-analysis-with-python/

5) Classification modeling:

- Choose 3 machine learning algorithms that would be appropriate for the dataset.
- Split data into training and testing sets.
- Present the results as performance metrics (table) and confusion matrix.



Code example:

https://www.geeksforgeeks.org/learning-model-building-scikit-learn-python-machine-learning-library/

https://scikit-learn.org/stable/modules/model_evaluation.html

6) **Longitudinal modeling:**

- Add “Time” column to the dataset (use years or months).
- Survival analysis with likelihood estimates.

Code example:

https://scikit-survival.readthedocs.io/en/stable/user_guide/00-introduction.html