# MATH 11205: Machine Learning in Python 2024-2025
# Project 1 Description

We will be using data collected by the Alzheimer's Disease Neuroimaging Initiative, a longitudinal, multi-center, observational study. The ADNI study aims to support the investigation and development of treatments that slow or stop the progression of Alzheimer's disease (AD). Researchers at over 60 clinical sites in the USA and Canada collect data to study the progression of AD in the human brain across normal aging, mild cognitive impairment (MCI), and Alzheimer's disease and dementia. This large-scale database combines clinical, demographic, imaging, biological, and omics data for thousands of patients. A much reduced, simplified version of the data will be used for this project, and the data have been provided in the file `adnidata.csv`.

## Assignment Goal

For the purpose of the project, consider yourself a **Data Scientist Consultant** who has been hired by the Alzheimer's Association to build a predictive model of cognitive decline. Dementia is a major international public health concern, and the cost of dementia to governments, social services and individuals has reached staggering figures. Identifying individuals at higher risk of cognitive decline is critical to potentially provide new approaches for early diagnosis of AD. In addition, this would enable the identification of subpopulations that may benefit more from therapies or drugs, or be better suited to clinical trials.

Towards this aim, you have been asked to use the provided data to build a model that makes use of information collected at a baseline visit to predict the cognitive score of individuals in a 24 month follow-up visit. In this setting, the cognitive score used is the Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) 13. The information collected at baseline includes the baseline cognitive score, diagnosis, demographic information, APOE genotype, and volumetric measurements of brain regions extracted from magnetic resonance images. Your model should be used to not only predictive future cognitive scores but also to gain insight into factors associated with higher risk of cognitive decline. In particular, individuals that are at higher risk of cognitive decline may be better suited to clinical trials for any proposed drugs or therapies aiming to slow the progression of the disease. In summary, you need to develop an **explainable, validated** model for the follow-up cognitive score as the outcome of interest using baseline features derived from the data provided.

You should start from a baseline linear regression model, including as few or as many of the provided variables. You may then explore a variety of models considering different transformations or feature engineering techniques to capture more complex relationships as well as regularization. However, your ultimate goal is to deliver a **single** model. Your final model choice should be justified and you should compare the performance of your model against the baseline model; although the main focus should be the description of your model (not the baseline or any other techniques tried).

It is important that any conclusions you draw from your model are well supported and sound and that you understand limitations of the model and the data. We explicitly **do not want a blackbox model** - you should be able to explain and justify your modeling choices and your model's predictions.

## Working as a team

This project may be completed by a team of up to 4 students (minimum of 1 student). Feel free to create your own team during workshop hours, building on the pairs for the workshops. Since we are not assigning teams, if you are a team that is looking for more members or someone looking for a team please use the pinned post on Piazza to find each other.

After the assignment is completed, we will distribute a brief peer evaluation survey. Completion of the survey is not required, but if you feel that some members contributed significantly less, this provides an opportunity for feedback and for such members to potentially have their overall mark penalized.

## Dataset Details

These are the available variables in the dataset:

- `RID` - person identifier.
- `ADAS13.bl` - ADAS-Cog 13 score at baseline. The Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) 13 includes 13 tasks that assess memory, language, and praxis. Specific tasks include word recall, naming objects and fingers, commands, constructional praxis, ideational praxis, orientation, word recognition, language, delayed word recall and a number cancellation or maze task. ADAS-Cog-13 scores range from 0 to 85, with higher scores indicates worse cognitive impairment.
- `ADAS13.m24` - ADAS-Cog 13 score at 24-month follow-up visit.
- `AGE` - age of the individual at the baseline visit.
- `DX.bl` - diagnosis of the individual at the baseline visit. The four categories (in increasing severeness) are cognitively normal (CN), early mild cognitive impairment (ECMI), late cognitive impairment (LCMI), and Alzheimer's disease (AD).
- `PTGENDER` - gender of the individual (male or female).
- `PTEDUCAT` - number of years of education.
- `PTETHCAT` - ethnicity of the individual (three levels: not Hispanic/Latino, Hispanic/Latino, or unknown).
- `PTRACCAT` - race of the individual (7 levels: white, black, asian, american indian/alaskan, hawaiian/other pacific islands, more than one, unknown).
- `PTMARRY` - martial status at baseline.
- `APOE4` - APOE genotype: indicates if the individual carries 0, 1, or 2 copies of the e4 allele.
- `Ventricles` - volume of the ventricles in cubic millimeters extracted from a magnetic resonance image at baseline.
- `Hippocampus` - volume of the hippocampus in cubic millimeters extracted from a magnetic resonance image at baseline.
- `WholeBrain` - volume of the whole brain in cubic millimeters extracted from a magnetic resonance image at baseline.

- `Entorhinal` - volume of the entorhinal cortex in cubic millimeters extracted from a magnetic resonance image at baseline.
- `Fusiform` - volume of the fusiform gyrus in cubic millimeters extracted from a magnetic resonance image at baseline.
- `MidTemp` - volume of the middle temporal lobe in cubic millimeters extracted from a magnetic resonance image at baseline.
- `ICV` - intracranial volume in cubic millimeters extracted from a magnetic resonance image at baseline.

The data provided for this project is subject to data agreements and waivers that I have signed on your behalf. Thus, **the data must NOT be shared publicly** (e.g. if your team is using GitHub, keep the repo private and do NOT share the data with any GenAI tools).

## Required Structure

A Jupyter notebook template called 'project1.ipynb' has been provided. It includes the required sections along with brief instructions on what should be included in each section. Your completed assignment must follow this structure - **you should not add or remove any of these sections, if you feel it is necessary you may add extra subsections within each**. Please remove the instructions for each section in the final document.

All of your work must be contained in the 'project1.ipynb' notebook, we will only mark what is included in this file (both the write-up and relevant coding). You may work on the notebook in whichever environment you prefer (noteable, locally, colab, codespaces,...).

There is an **upper limit of 20 pages** including code. Your notebook must include all of your work, but make sure that you are only retaining required components, e.g. remove unused code and figures (if a figure is not explicitly discussed in the text it should not be in the final document). Overall, your project will be partially assessed on your organization / presentation of the document - it should be as polished and streamlined as possible. **Try to be as concise as possible while creating your write-up. We highly recommend that you check the appearance of your rendered PDF before submitting, as its appearance can differ significantly from the notebook**.

Please submit your final PDF of project report (generated from a Jupyter notebook) to the Project assignment on Gradescope. Please ensure that you **tag all groups members** on Gradescope, and also add all group member names either in the notebook metadata or in additional markdown cell block at the beginning of the file.

## Getting Help

- **Piazza:** This forum will be used as the central location for all course related discussions and questions, and should be used over emailing course staff directly. The course lecturers will monitor and respond to questions, but feel free to provide some constructive responses to peer's questions.
- You can also ask questions at the end of lectures during any Q&A time or during workshops or office hours.

## Generative AI Policy

Please refer to the School of Mathematics Generative AI Policy. Generative AI tools may be used for project, but always as your helper or co-pilot and used responsibly, and NOT your driver! For example, GenAI may used for generating explanations for error messages and debugging, providing hints or suggestions to improve code, enhancing visualizations and the quality of the report. The data should absolutely NOT be copied and shared directly with GenAI tools, as this in breach of legal consent for this data. You must include a statement on how GenAI was used. **If the project is suspected to be heavily written by GenAI, the students may be subject to an oral presentation.**