

# DATA 505 – Computational data science and analytics

## Project 3 – Final Project – Human connectome and predicting intelligence quotient (IQ)

**Collaboration Policy:** Individual Assignment

**Total Points:** 100 (See rubric below)


**Project Topic:** In this project you'll construct machine learning pipelines that are capable of predicting the IQ of healthy adults using structural connectomes (or connectivity matrices) derived from diffusion tensor imaging (DTI) white matter (WM) tractography data.

**Project Dataset:** The dataset includes 58 healthy (i.e. no history of a neurological condition) adults and is provided in a zip file (attached to the Dropbox) named **dataset.zip**. Furthermore, for privacy purposes all identifying information about each participant has been removed (name, age, gender, etc.). The only way to identify a participant is by anonymous id NSXXX where XXX is a three digit number, e.g. NS003. For each participant, the zip archive includes three files:

1. NSXXX.csv: CSV file that defines an  $83 \times 83$  connectivity matrix for participant NSXXX. The connectivity matrix is a symmetric matrix (diagonal is zero), where the matrix values represent WM fiber density between two different gray matter regions.
2. NSXXX\_vec.csv: CSV file that defines a 3403 dimension connectivity vector for participant NSXXX. This vector is simply the upper diagonal of the connectivity matrix see example in **Fig. 1**.
3. NSXXX.txt: Text file that defines the IQ for participant NSXXX.

Example 5x5 dimension  
symmetric  
connectivity matrix

0	1	2	3	4
1	0	5	6	7
2	5	0	8	9
3	6	8	0	10
4	7	9	10	0

 = upper diagonal elements

Corresponding 10 dimension  
connectivity vector

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

**Fig. 1:** Example conversion from a 5x5 dimension symmetric connectivity matrix to a 10 dimension connectivity vector.

**Project Deliverables:** Using the project dataset described above, you'll construct the machine learning and analysis pipelines below. It is important to note, in this project you'll be making many of pipeline and analysis decisions, and each one must be described in your notebook. For instance, how many stages does

your pipeline include, and what is the purpose of each software model in each stage? Which cross-validation technique did you choose and what were the parameters (e.g. stratified, number of folds, random shuffle)? These are just a few examples. In general, assume you'll give your work to a colleague, and you're providing them descriptive information about your thought process. These "thought processes" will be outlined in your markdown section or using code comments.

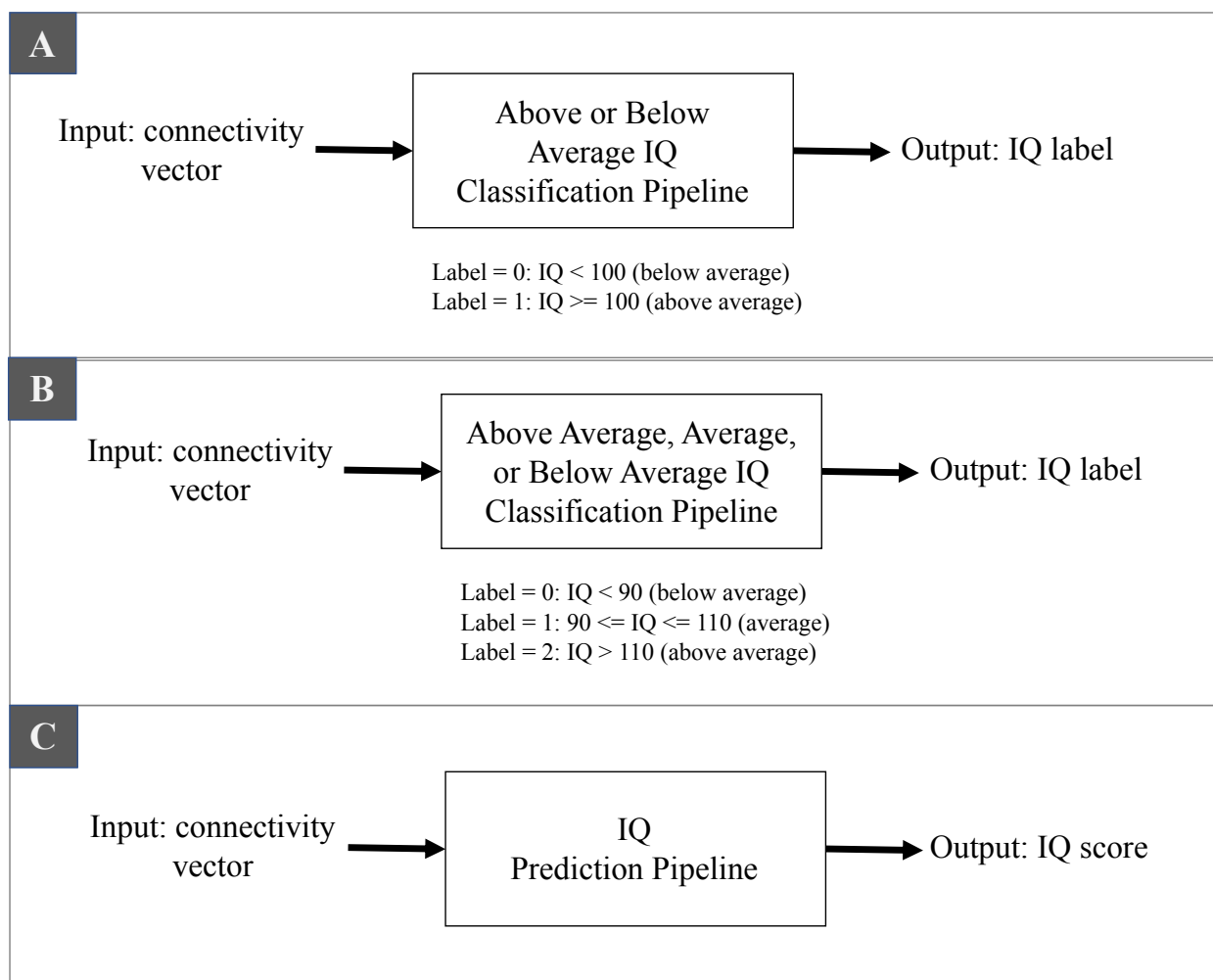
- I. Design, develop, and evaluate the performance of an above or below average IQ classification pipeline. In particular, the input to the pipeline is a connectivity vector, and the output of the pipeline is either an above or below average IQ classification (see **Fig. 2A**). Furthermore, the pipeline must include two or more stages. In this project, you may assume an average IQ is 100, therefore an **above average**  $\text{IQ} \geq 100$  and a **below average**  $\text{IQ} < 100$ . The performance evaluation must use a k-fold cross-validation technique, and for each test-fold you must compute the following metrics: accuracy, precision, recall, and F1 score.
- II. Design, develop, and evaluate the performance of a multiple-IQ classification pipeline. In particular, the input to the pipeline is a connectivity vector, and the output of the pipeline is either an above average, average, or below average IQ classification (see **Fig. 2B**). Furthermore, the pipeline must include two or more stages. In this project, you may assume an **above average**  $\text{IQ} > 110$ , **average**  $90 \leq \text{IQ} \leq 110$ , and **below average**  $\text{IQ} < 90$ . The performance evaluation must use a k-fold cross-validation technique, and for each test-fold you must compute the following metrics: accuracy, precision, recall, and F1 score.
- III. Visualize the classification results in I and II.
- IV. Design, develop, and evaluate the performance of an IQ prediction pipeline. In particular, the input to the pipeline is a connectivity vector, and the output of the pipeline is the predicted IQ value (see **Fig. 2C**). Furthermore, the pipeline must include two or more stages. The performance evaluation must use a k-fold cross-validation technique, and for each test-fold you must compute the following metrics: mean absolute error, correlation coefficient ( $r$ ), explained variance, and mean squared error.

**Approved software tools:** You may only use the Python programming language and the following technologies/software libraries: 1. Python notebook, 2. NumPy, 3. Matplotlib, 4. SciPy, 5. Scikit Learn, or 6. Pandas. Any technology that is not listed above, please ask me if it can be used in your project.

**Project Submission:** Submit your notebook to the Dropbox setup on OAKs before the due date. The name of your notebook must be, P3\_<lastname>.ipynb, where <lastname> is your last name 😊. For example, *P3\_Munsell.ipynb* would be correct if I were to submit the assignment. Please, only use one notebook to complete this assignment, i.e. no other Python files (.py files) are allowed.

## Project Grading Rubric:

Python notebook compiles and runs with no syntax errors	5 points
Your notebook must include a markdown section that describes what each cell is accomplishing. For instance, what is the cell computing or visualizing, and what mathematical, statistical, or visualization techniques is being applied.	10 points
In each notebook cell, include comments that provide more in-depth information about how the computation is performed (i.e. step-by-step approach) or how the data is stored in a particular array or matrix.	10 points
Items I, II, IV (20 points each)	60 points
Item III	15 points
	100 points



**Fig. 2:** Machine learning pipeline diagrams