Brain activity pattern differences in ADHD

January 9, 2025

1 Project Description

Neuropsychiatric disorders that occur in development, like anxiety, depression, autism, and attention deficit hyperactivity disorder (ADHD) often differ in how and to what extent they affect males and females. ADHD occurs in about 11% of adolescents, with around 14% of boys and 8% of girls having a diagnosis. There is some evidence that girls with ADHD can often go undiagnosed. Girls with ADHD who are undiagnosed will continue suffering with symptoms that burden their mental health and capacity to function.

You're a data scientist for the NHS. The data science department is trying to free doctor's time by creating a system that can diagnose ADHD in children and adolescents as a support tool for GPs.

In order to develop such a system, the fMRI department has collected functional brain imaging data of children and adolescents which are available to you together with their socio-demographic, emotions, and parenting information. As there might be gender-based differences, the task is to build a model to predict both an individual's sex and their ADHD diagnosis using this information.

The data are available here.

The company is particularly worried about misdiagnoses in female patients, as they are harder to diagnose with traditional methods. You must ensure that your model is not biased. Furthermore, due to GDPR requirements, your model should be explainable (e.g., you can use SHAP or LIME for this).

Read the following tasks in detail and make sure you understand the project.

2 Tasks

2.1 Stage 1: Data exploration

Stage 1 is about making sure you have loaded and explored your dataset, and that you understand the data and are ready (or nearly ready) to move on to modelling using appropriate methods. Make sure you spend enough time on this assignment: cleaning and understanding the data is a big part of Data Science. Your models won't work if they are not appropriate for the problem at hand.

Check the marking scheme for the 'Data Exploration' assessment. It's available on Moodle in the Assessment Information page – more details are in Section 3 of this document. Based on this marking scheme, the main tasks you should complete are:

1. Load and explore the data set, leaving a subset of data separate from the exploration to avoid overfitting.
2. Clean and preprocess the data set. By the end of Stage 1, you should have a clean and cohesive data set.
3. Think about the types of models that you can use for this project given the requirements above and the data exploration you've done. This means that I am not interested in seeing lots of plots: I want each of them to be there for a reason, and to help you make a decision about the types of models you can use for this problem. Pay attention as well to how you will evaluate the system to ensure it is not biased.

Note that you are NOT asked to perform any modelling at this stage. During your lab demo, we will check that you understand the project and have a plan for how to analyse and model the data for the second stage that is reasonable and feasible.

Remember that if you do not demonstrate your code during your assigned lab session, you will receive a mark of 0 for Stage 1.

2.2 Stage 2: Modelling and testing

In Stage 2 your task is to complete the project described in Section 1 of this document. In Stage 1 you understood and prepared your data for modelling. In Stage 2 you need to answer the question posed to you by the NHS and present your results to your manager and a board of doctors.

Check the marking scheme for the Stage 2 assessments ('Final project Code' and 'Final project Demo'). They can be found in the Assessment Information tab on Moodle and in Section 3 of this document. Based on this marking scheme, the main tasks you should complete are:

1. Some form of modelling to answer the project that your manager assigned you. The modelling is up to you, as long as it follows good practice in data science (i.e., using cross-validation and train/test splits properly and as appropriate for this data set, problem, and the chosen modelling approach). As this is a research-led module, you are encouraged to check the available literature and find out what has worked in the past, but you need to present something new/different as part of your project.

2. While you're modelling, reflect on the project description given above (Section 1 of this document) and make sure you pay special attention to the requirements set out by your manager.

3. For your presentation, make sure that all your decisions are justified and that you present your findings clearly and concisely, within the time limit. Any assumptions made and references used must be stated. As a data scientist, you must reflect on your results. Talk about your findings and what they indicate. Finally, you **must** answer the questions: Are you confident that your model fulfils the requirements for the company? Is it ready to be deployed? Are there any limitations they should consider? Is the model showing bias? Do you have any other insights from the data that can help the NHS in the future (e.g., other projects that could be tackled with these data)?

3 Deliverables and marking criteria

3.1 Stage 1: Data exploration

Your FASER submission must include:

- **A README file** with a description of the project and instructions on how to use/run the code. Any assumptions made must also be stated in the README file.
- The **code** that you used to carry out the exploration and preprocessing. One Jupyter notebook is enough, but it should not be just a stream of figures/plots with no justification of why they're informative (this is "plot dump"): make sure your notebook is properly documented and there are useful comments in the code. Explain the insights you get from each figure. There should be no errors/warnings in the notebook/s, and try to avoid repeating the same code multiple times (this is what functions are for). Headings for different subsections are a bonus.
- A **pdf file (one page maximum)** with your plans for modelling in Stage 2. This should include the types of models you are going to use (and why), and the evaluation metrics you'll use.

Do NOT upload the dataset. Your code should run with the data provided by me for this assignment.

3.1.1 Marking criteria for Data exploration [20% of final mark]

The following aspects will be assessed about the code and in the labs:

- Is there evidence of data loading? [5%]
  - Has the data been loaded?
- Is there sufficient appropriate exploration? [25%]
  - Have the inputs/features been properly explored according to the type of data?
  - Do plots have labels and legends?
  - Are the insights from the plots explained in the notebook?
- Data preprocessing: Has the data been cleaned properly? [25%]
  - Has the data been preprocessed properly, using adequate methods for the type of data that is being considered (including data splits)?
  - Is there a coherent final dataset?
- Plan for Stage 2: Is there a coherent plan based on the findings from Stage 1? [25%]
  - Does the student have a clear plan for modelling?

  - How will the model/s be evaluated?
  - Does the plan adhere to the project's requirements?

- Notebook [20%]
  - Is the notebook well presented?
  - Are there meaningful comments?
  - Is the code free of data dump and errors?
  - Does the notebook have structure?
  - Are plots explained?

Note that you will only get your mark for the Data Exploration assignment if you present your work during the lab session. Failure to attend will result in an automatic mark of 0. Failure to explain the code and discuss your results will also result in an automatic mark of 0.

3.2 Stage 2: Modelling and testing

You will receive feedback after Stage 1. **You are expected to fix any errors highlighted to you as part of Stage 2.**

There are two deliverables for Stage 2, all of which must be submitted to FASER:

3.2.1 Final project Code

Your FASER submission must include:

- A **README file** with a description of the project and instructions on how to use/run the code. Any assumptions made must also be stated in the README file.
- All the **code** that is necessary to go from the original dataset that was given to you until your final results.
  - The notebook with exploration and preprocessing.
  - The notebook used for modelling and to obtain the final results.

  Exploration/preprocessing and modelling MUST be in different jupyter notebooks. Make sure your notebooks are properly documented and there are useful comments in the code. Explain the insights you get from each figure, and justify your methods through markdown cells and comments. There should be no errors/warnings in the notebooks, and try to avoid repeating the same code multiple times (this is what functions are for). Headings for different subsections are a bonus.

Do NOT upload the dataset. Your code MUST run with the data provided by me for this assignment.

3.2.2 Final project Demo

The final project demonstration deliverable is submitted through FASER in the form of a presentation. This presentation will be seen by NHS doctors, so you should focus on a high level description of what you did and mostly on your insights, results, and what it means for them. If you think your model is good enough, your presentation should try to convince the doctors to test it. If you think there are drawbacks, you should also explain these. You will be personally liable if the model doesn't perform as you claim. Make sure you answer the questions in point 3 of Section 2.2 — what's the bottom line for the GPs?

Your FASER Submission must include:

- The document used to present your results.
- A 5-minute video presentation using the slides submitted. If the video is longer than 5 minutes, only the first 5 minutes will be watched and evaluated.

You must submit both the video and the slides or you will receive a mark of 0 for this part of the assignment. If I can't identify you from the presentation you will still get a mark of 0 - make sure your face is clearly visible through the presentation.

3.2.3 Marking criteria for Final project Code [31% of final mark]

The following aspects will be assessed about the code:

- Is the preprocessing/cleaning correct? [20%]
  - This is a chance for you to fix any errors that were identified in the Data Exploration assignment
  - Has the data been preprocessed properly, using adequate methods for the type of data that is being considered (including splitting the dataset)?
  - Have errors from Stage 1 been fixed (if any)?
  - Is there a coherent final dataset appropriate for the Final project assignment?
- Are Machine learning and Data Science conventions followed in the pipeline? [30%]
  - Are the methods appropriate for the project? — including data splits, type of modelling, metrics used, etc.
  - Are there errors in the pipeline/s?
- How well organised is the code? [10%]
  - Are the notebooks well presented?
  - Are there meaningful comments?
  - Is the code free of data dump and errors?
  - Does the notebook have structure?
  - Are plots explained?
- Is there a README file describing the project and organisation of the code? [5%]
- How much depth was achieved in the project? [20%]
  - Does the product submitted adhere to the project requirements?
  - Are the model outputs explainable?
  - How detailed is the evaluation of the model?
  - Is there a reflection/analysis of bias?
  - Do the final conclusions reflect the results?
- Has the student discussed the project with the Module Supervisor? [15%]
  - This discussion can take place over office hours or in labs, but must be face-to-face and before the deadline for this assignment.
  - Coming to ask questions is not a discussion: a discussion is a 2-way conversation — I need to see that you're working on the project.
  - It's your job to ensure I have noted down the discussion.
  - Do not leave it until the last minute!

3.2.4 Marking criteria for Final project Presentation [31% of final mark]

You must submit both the video and the slides or you will receive a mark of 0 for this part of the assignment. If I can't identify you from the presentation you will still get a mark of 0 - make sure your face is clearly visible through the presentation.

The following aspects will be assessed about the presentation video and slides:

- Introduction: Is the project introduced properly? [10%]
  - Please include 1–2 slides introducing the project and main objectives.
- Methods: Are they summarised? [15%]
  - The methods must be stated at a high level — remember that you're not presenting to data scientists, but to the board of doctors of the NHS. Just a short summary of methods is enough.
- Results: Are they presented coherently? [22.5%]

  - Has the student made appropriate use of graphs and plots to summarise the results?
  - Are there enough results for the Board to make an informed decision?

- Conclusions and recommendations [22.5%]
  - Are there conclusions for the Board?
  - Do the conclusions match the results presented?
  - Is there a final recommendation?
  - Does the recommendation match the results presented?

- Presentation style [10%]
  - Is there a title page with a catchy title and the registration number of the student?
  - Is the presentation coherent?
  - Are the slides pleasing to the eye?
  - Is there a good balance of text and figures/plots? Are plots of good quality?
  - How are the presentation skills of the student?

- Are the project questions sufficiently answered? [20%]
  - Does the model adhere to requirements?
  - Is there a reflection/analysis of the model performance?
  - Are there suggestions for future projects? — Suggesting “collecting more data” will not give you additional points; that’s not a project! Similarly, suggesting different classifiers is not a future project.

- Coherence between Demo and Code [up to -25%]
  - This section checks for consistency between the results presented in the presentation and those submitted in the code separately. If results don’t match, you might lose up to 25% of the marks for this assignment.

    The methods, figures, and results from your presentation must match those from the code.

    But we will also be looking at other general inconsistencies.