**Bioinformatics Project: Alzheimer’s Disease Risk Prediction**

**Abstract:**

Identify the possibility of developing Alzheimer’s disease based on genetic, clinical and lifestyle data.

**Data Sources:**

Some of the data sources that I used were ADNI (Alzheimer’s Disease Neuroimaging Initiative), UK Biobank and GWAS Catalog.

Signed up for biobank but need to wait for confirmation.

Key features of the Data sets used:

**Genetic:**

* APOE Genotype – APOE ε4 allele
* SNPs identified from the GWAS studies link to Alzheimer’s

**Clinical:**

* Age
* Cognitive Test Scores (MMSE – Mini-Mental State Examination)
* Family History of Alzheimer’s
* Other health conditions (diabetes and hypertension)

**Lifestyle:**

* Diet (for example, adherence to a Mediterranean diet)
* Physical Activity
* Education Level
* Smoking Status
* Alcohol Consumption

**Model Selection:**

* Logistic Regression
* Random Forrest & Gradient Boosting Machines – these are part of FEATURE IMPORTANT RANKING – machine learning models.
* Neural Networks

**My step by step:**

1. I am gathering data right now, from the GWAS catalog I went to ensemble to download SNPs. – didn’t ned up using ensemble, I used the websites and figured out to download – preferred format CSV.
2. OK, so I got genetic data from ADNI and GWAS, with additional imaging data from ADNI.
3. Now, - clean and normalise SNPs.
4. I coded in python the high missingness, the cleaning and standardising of the SNPs.
5. Now moving with feature selection.
   1. Feature selections I will go forth with, I think:
      1. Feature importance ranking – Random Forest and Gradient Boosting.
      2. Recursive Feature Elimination (RFE)
      3. L1 Regularisation (LASSO)
6. Ok, worked on the clean/ standardising of the SNPs, added the addition of saving the new data in a new file, all in one file, next is using this in a machine learning model, just as mentioned before.