Alzheimer's Detectives - Project Milestone 1

CMPT 310 - D200: Introduction to Artificial Intelligence and Machine Learning

Group Information

Manjari Prasad - mps12@sfu.ca

Wan Yu Wendy Wong - wyw9@sfu.ca

Beyzanur Kuyuk - bka81@sfu.ca

Project Summary

Our goal is to implement a machine learning classification system for the early detection of Alzheimer’s disease using patient health records. The first objective is to classify individuals as either at-risk or not at-risk for Alzheimer's using clinical and cognitive benchmarks like their age, memory test scores, familial history pertaining to the disease, and MRI indicators through logistic regression for binary classification.

What Has Been Accomplished So Far:

Data gathered or generated?

We gathered the OASIS Cross-Sectional dataset from Kaggle since the dataset contains the demographic and cognitive features essential to prediction.

Preprocessing completed?

We dropped the unnecessary columns for prediction, scaled the features, and numerically encoded gender. We use median imputation to replace missing values when MMSE, SES, and Educ values are absent and provide a binary dementia diagnosis based on CDR > 0.

Baseline model trained?

For our baseline model, we chose logistic regression because it is a good fit for binary classification, which is the focus of our project. By providing probabilities rather than just binary numbers, the model helps us understand how confident it is in each prediction. In addition, the model’s coefficients let us see how each input affects the likelihood of dementia. Being able to interpret the influence of each feature is useful in medical applications.

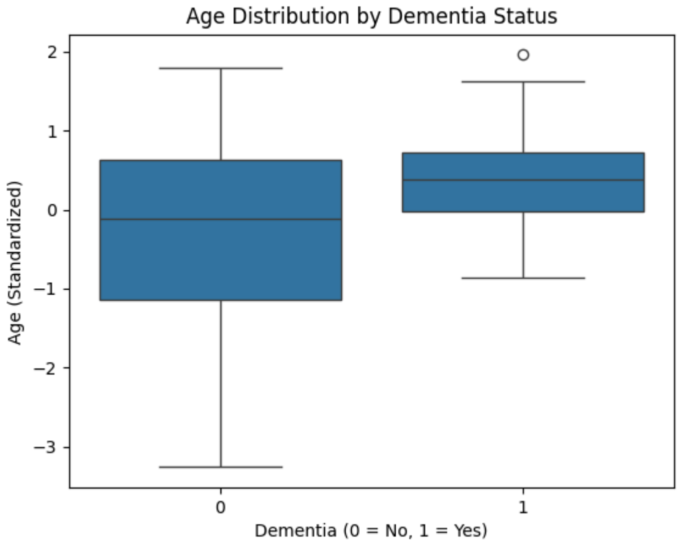
An 80-20 train-test split was used to train the logistic regression model, and an accuracy of approximately 81% was achieved. The model has been evaluated using classification report metrics, accuracy, and a confusion matrix, along with model evaluation metrics.

Visualization, pipeline, or planning work?

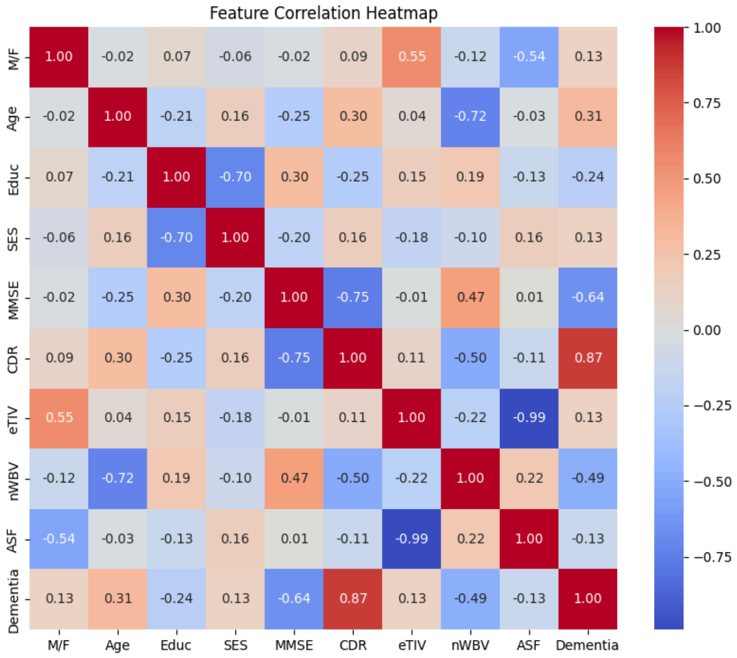
We organized our work into separate files for coding visualization, modelling, and data preprocessing to improve our workflow. We generated some critical plots for exploratory data analysis to improve our interpretability with research questions such as age distribution of dementia classes via boxplot, correlation heatmap of all features, histogram of MMSE scores by dementia and their count, and pairplot comparing important features. These allow analysis beyond description, including predictive patterns, feature correlation, class balance, etc.

Plots:

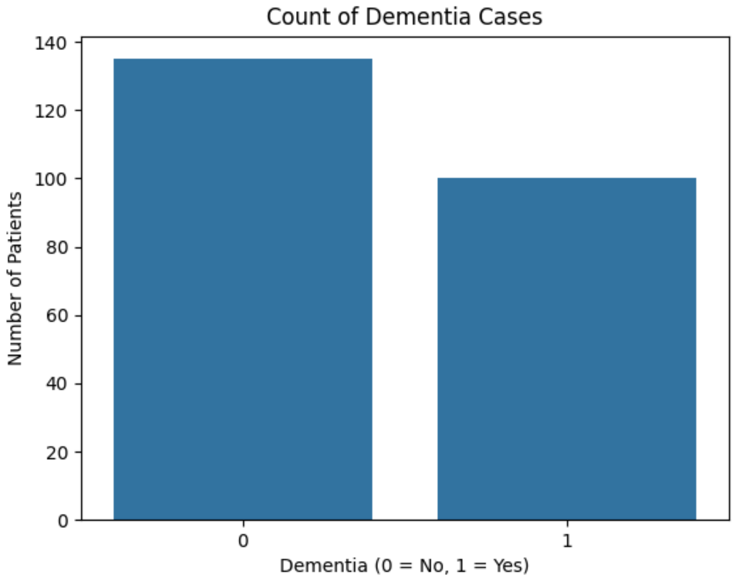
To better understand the structure of our dataset and the relationships between features, we created several exploratory plots using Seaborn and Matplotlib. These visualizations help reveal patterns relevant to dementia prediction, highlight feature correlations, and provide insight into class distribution and feature separability. Below are five key figures that support our preprocessing decisions and model interpretation.



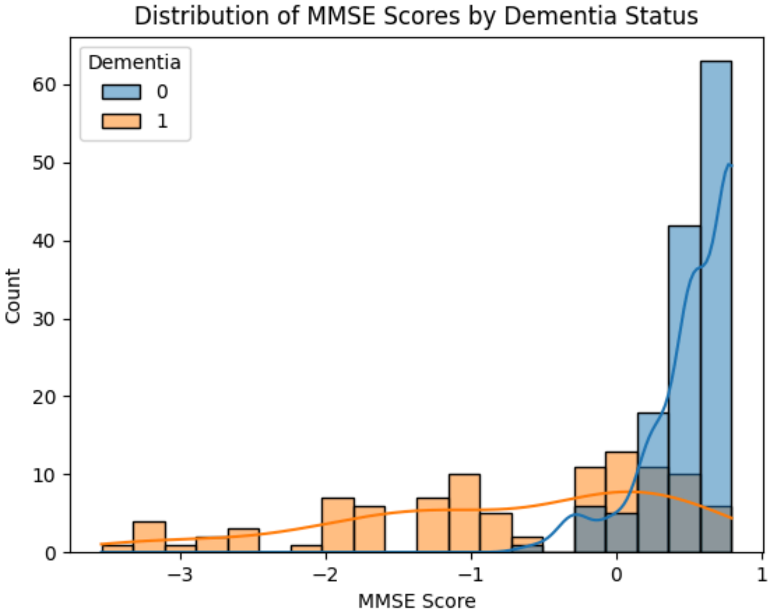
This boxplot displays the distribution of age for dementia and non-dementia groups, suggesting a slightly higher age range among dementia cases.



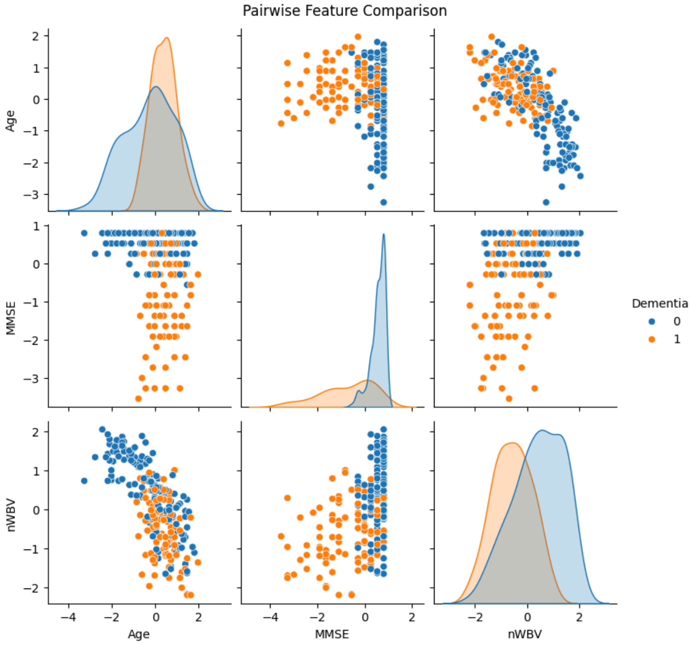
This heatmap visualizes the pairwise correlations between numeric features, helping identify which variables are strongly related.



This plot shows the number of patients with and without dementia, highlighting a class imbalance in the dataset.



This histogram compares the distribution of MMSE scores between dementia and non-dementia patients, revealing lower MMSE scores among those with dementia.



This pairplot shows scatter plots and distributions of key features, illustrating how combinations of features like MMSE and nWBV relate to dementia classification.

What Has Fallen Behind

Expected task:

The current workflow still relies on train/test split methods, and Cross-validation has not been done yet.

Challenges:

Recall of dementia detection is somewhat lower due to dataset class imbalance.
This will be implemented in future revisions.

Any Project Changes (brief):

No changes were made to the dataset or primary technique used.