Group 25: Angie Ohaeri, Melody Pak, Noah Kornfeld, Olivia Pelzek

Dataset Link:
https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset?resource=download

This dataset contains extensive health data of 2,149 unique patients and their diagnosis (or lack thereof) of Alzheimer's disease. This data was uploaded to Kaggle by Rabie El Kharoua in 2024 with the intended use of developing predictive models concerning the factors contributing to Alzheimer's disease.

There are 35 variables in this dataset, grouped into five categories: Patient Information (such as demographics and lifestyle factors), Medical History (such as history of head injuries or depression), Clinical Measurements (such as blood pressure or cholesterol levels), Cognitive and Functional Assessments, Symptoms (such as confusion or personality changes), and Diagnosis Information (which indicates if the patient has Alzheimer's disease). All of these features are numerical, either categorized by an indicator, a doctor-given score, or a numerical measurement.

We will use these variables to answer our two main questions. First, what clinical and demographic factors correlate with the likelihood of Alzheimer's disease in a patient, and what supervised learning algorithms predict the results most accurately on unseen data? Second, which category of factors (demographic, medical history, etc) is most accurate when predicting Alzheimer's disease in a patient?

For our methods, we will need to utilize feature engineering, exploratory analysis, and feature selection before our modeling. While there are no missing values, some categories must be altered using one-hot encoding (ethnicity, for example, is coded as a 0, 1, 2, or 3). We also may insert some calculated features to summarize each of the categories. We will also create correlation tables and plots to explore which variables are likely important when predicting Alzheimer's disease. We will then use a variance threshold to select the most indicative features.

For modeling, to answer our first question, we will compare a Support Vector Machine and a Logistic Regression algorithm on the same selection of features. For our second question, we will use multiple k-Nearest Neighbors algorithms, each using only the features in one category.

Code to read in data:

```
import pandas as pd
alzheimers = pd.read_csv('alzheimers_disease_data.csv')
print(alzheimers.head(5))
```

```
   PatientID  Age  Gender  Ethnicity  EducationLevel        BMI  Smoking  \
0       4751   73       0          0               2  22.927749        0
1       4752   89       0          0               0  26.827681        0
2       4753   73       0          3               1  17.795882        0
3       4754   74       1          0               1  33.800817        1
4       4755   89       0          0               0  20.716974        0
```