**Monica Cortez**

**Project: Detecting Alzheimer’s using Magnetic Resonance Imaging**

**1. Introduction:**

One of the most popular datasets to explore the brain and the effects of Alzheimer’s to it is the MRI and Alzheimer’s by the Open Access Series of Imaging Studies otherwise known as OASIS. It is a project aimed at making MRI data sets of the brain freely available to the scientific community. This dataset is MRI of demented and nondemented adults. Using this dataset, models will be created to identify with patients with mild dementia. This white paper is intended to provide an overview of the MRI and Alzheimer’s dataset including, data processing, exploratory data analysis, modeling, and evaluation.

**2. Dataset:**

The dataset is available on Kaggle. It was created by [Open](https://www.promptcloud.com/) Acess Series of Imaging Studies (OASIS) made available by Washington University Alzheimer’s Disease Research Center, Dr. Randy Buckner at the [Howard Hughes Medical Institute (HHMI)](http://www.hhmi.org/), the [Neuroinformatics Research Group (NRG)](http://nrg.wustl.edu/" \t "_blank) at Washington University School of Medicine, and the [Biomedical Informatics Research Network (BIRN)](http://www.nbirn.net/). It can be downloaded as a CSV file. The dataset is separated into two different datasets, Cross-sectional MRI Data in Young, Middle Aged, Nondemented and Demented Older Adults which contains 416 subjects aged 18 to 96, where each subject has 3 or 4 individual T1-weighted MRI scans that are obtained in a single session. The other dataset within it being Longitudinal MRI Data in Nondemented and Demented Older Adult with 150 subjects aged 60 to 96, each subject had 3 or 4 individual T1-weighed MRI scans obtained in single scan session.

**3. Data Preparation:**

The first process of data mining on the MRI and Alzheimer’s dataset is preparing the data. This includes cleaning the data for variables that are not needed, checking for missing values, inconsistencies and determining outliers. Then transforming and formatting the data so its easily analyzed. Given the nature of how the dataset was put together, the data preparation process is simple. Only one category had missing values, SES (social economic status), which only had 8 missing values. To resolve this issue, the median value of the category was taken and imputed into the missing values.

**4. Exploratory Data Analysis**

After the data has been prepared, the next step is to perform exploratory data analysis (EDA). EDA is important because it helps visualize and summarize data by showing the relationships between variables. In the case for the MRI and Alzheimer’s dataset it will focus on exploring the relationship between the features of the MRI and the dementia of the patient. This process will explicitly show through the use of graphs and a correlation matrix the correlation of the variables thus giving insight to understanding data. We can see correlations between brain density measures and scaling factors and dementia ratings.

A picture containing text, screenshot, colorfulness, pattern

Description automatically generated

**5. Modeling and Training**

Modeling techniques used will be Logistic Regression, SVM and Decision Tree. Detection of early onset Alzheimer’s being the outcome of this project as well as analyzation and discussion that could be potentially important to the disease. The modeling process involves splitting the data into train and test sets and evaluating the test set. Fivefold cross validation is used, as five is usually the standard, to determine the best parameter for the models. A logistic regression model was run, as well as an SVM model and a Decision Tree model.

**6. Evaluation**

Model evaluation is a crucial part for the MRI and Alzheimer’s dataset. The performance metric that will be used will be accuracy, we find the best parameter for accuracy, but also test for recall and AUC. This metric takes into consideration the importance of a true positive result for possible early diagnosis.