Alzheimer’s Disease Prediction

Problem Statement

- Alzheimer’s disease (AD) is a neurodegenerative disorder of uncertain cause and pathogenesis that primarily affects older adults and is the most common cause of dementia.

- The earliest clinical manifestation of AD is selective memory impairment and while treatments are available to ameliorate some symptoms, there is no cure currently available.

- Brain Imaging via magnetic resonance imaging (MRI), is used for evaluation of patients with suspected AD.

- Some studies have suggested that MRI features may predict rate of decline of AD and may guide therapy in the future.

Objective

- Using Machine Learning techniques.
- to help clinicians to accurately predict the earlier Alzheimer’s.
- Motivation is to slow down the progress of a patient from mild cognitive impairment to dementia.

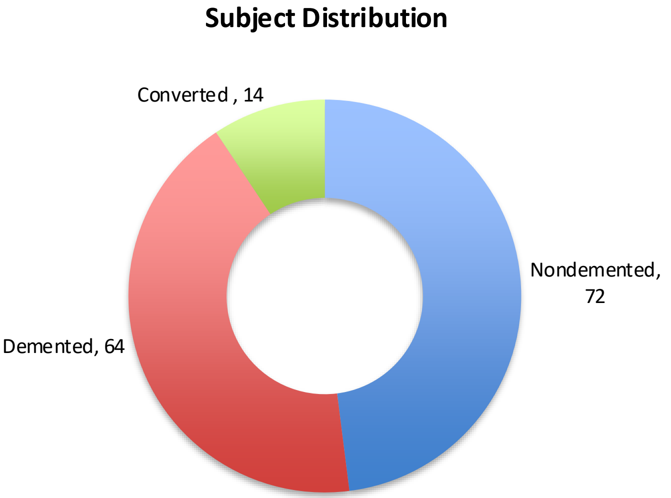
Agenda

- Methodology
  - Datasets, Models, Metrics, Tools
- Process Workflow
  - EDA
  - Data preparation
  - Data analysis
  - ML model training/evaluation
- Results
  - Accuracy F1-score, ROC Curve
- Conclusion
  - How it helps with business case
  - Recommendations
  - Interesting insight
- Future Opportunities
- Appendix

Methodology

- Source of dataset
  - From Kaggle
  - This dataset is MRI related data that was generated by the Open Access Series of Imaging Studies (OASIS)

- Data description
  - The dataset consists of a longitudinal MRI data of 150 subjects aged 60 to 96.
  - Each subject was scanned at least once.
  - Everyone is right-handed.
  - 72 of the subjects were grouped as 'Nondemented' throughout the study.
  - 64 of the subjects were grouped as 'Demented' at the time of their initial visits and remained so throughout the study.
  - 14 subjects were grouped as 'Nondemented' at the time of their initial visit and were subsequently characterized as 'Demented' at a later visit. These fall under the 'Converted' category.



Methodology

- Column descriptions

COL	FULL-FORMS
EDUC	Years of education
SES	Socioeconomic Status
MMSE	Mini Mental State Examination
CDR	Clinical Dementia Rating
eTIV	Estimated Total Intracranial Volume
nWBV	Normalize Whole Brain Volume
ASF	Atlas Scaling Factor

- Model, Metricx and Tools
    - Supervised Machine Learning Classification Problem
    - Model: Logistic Regression, Linear SVC, Random Forest Classifier and MLP Nueral Networks.
    - Metricx: accuracy, precision, recall, F1-score, ROC, AUC
    - Tools: jupyter notebook, python, pandas, numpy, matplotlib, seaborn, scikit learn, etc

Process Workflow

Exploratory Data Analysis (EDA)

```
df = pd.read_csv('oasis_longitudinal.csv')
df.head()
```

	Subject ID	MRI ID	Group	Visit	MR Delay	M/F	Hand	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
0	OAS2_0001	OAS2_0001_MR1	Nondemented	1	0	M	R	87	14	2.0	27.0	0.0	1987	0.696	0.883
1	OAS2_0001	OAS2_0001_MR2	Nondemented	2	457	M	R	88	14	2.0	30.0	0.0	2004	0.681	0.876
2	OAS2_0002	OAS2_0002_MR1	Demented	1	0	M	R	75	12	NaN	23.0	0.5	1678	0.736	1.046
3	OAS2_0002	OAS2_0002_MR2	Demented	2	560	M	R	76	12	NaN	28.0	0.5	1738	0.713	1.010
4	OAS2_0002	OAS2_0002_MR3	Demented	3	1895	M	R	80	12	NaN	22.0	0.5	1698	0.701	1.034

- 15 columns 373 rows.
- Null values.

Process Workflow

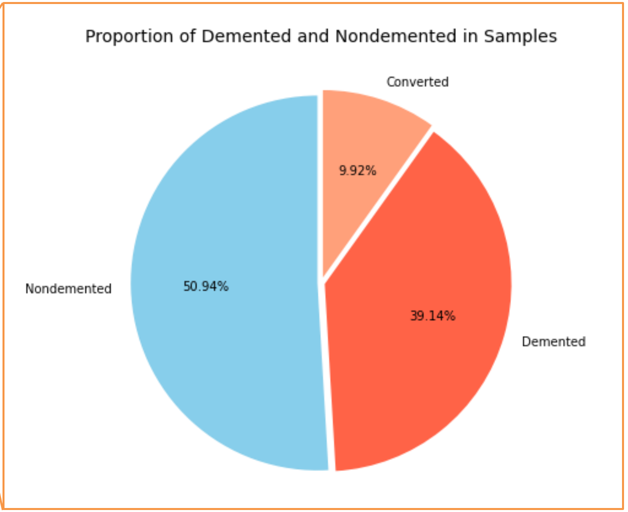
Exploratory Data Analysis (EDA)

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 373 entries, 0 to 372
Data columns (total 15 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Subject ID   373 non-null    object
 1   MRI ID       373 non-null    object
 2   Group        373 non-null    object
 3   Visit        373 non-null    int64
 4   MR Delay     373 non-null    int64
 5   M/F          373 non-null    object
 6   Hand         373 non-null    object
 7   Age          373 non-null    int64
 8   EDUC         373 non-null    int64
 9   SES          354 non-null    float64
 10  MMSE         371 non-null    float64
 11  CDR          373 non-null    float64
 12  eTIV         373 non-null    int64
 13  nWBV         373 non-null    float64
 14  ASF          373 non-null    float64
dtypes: float64(5), int64(5), object(5)
memory usage: 43.8+ KB
```

```
df.nunique()
```

```
Subject ID    150
MRI ID        373
Group           3
Visit           5
MR Delay      201
M/F             2
Hand            1
Age            39
EDUC           12
SES             5
MMSE           18
CDR             4
eTIV          286
nWBV          136
ASF           265
dtype: int64
```

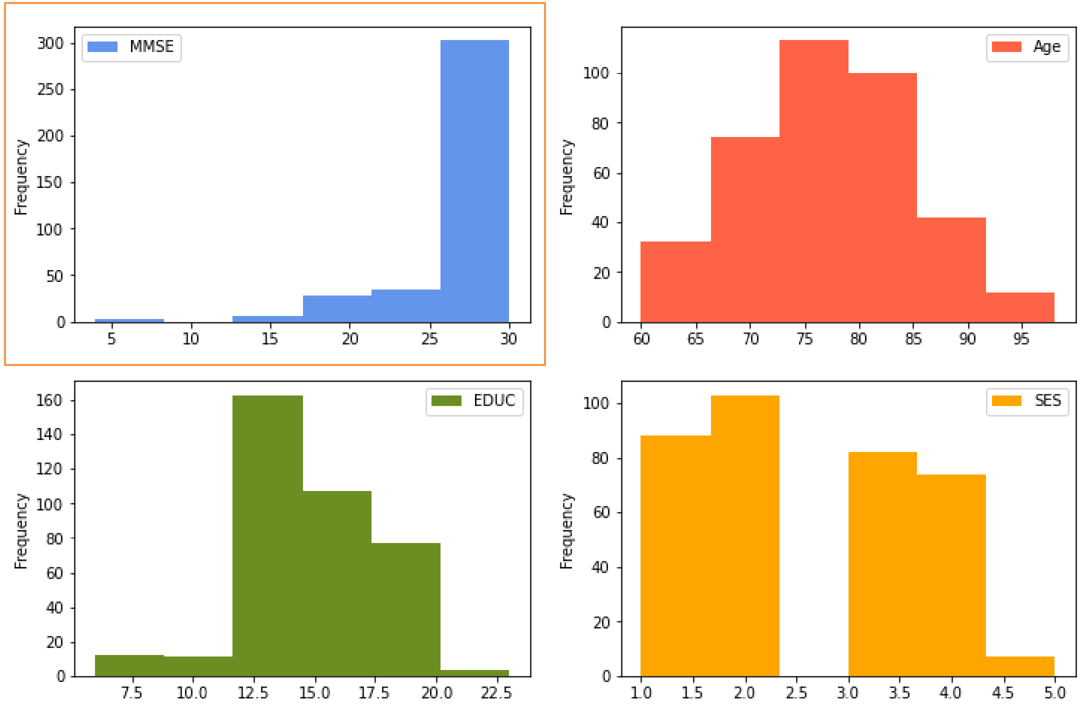


- 14 subjects were grouped as 'Nondemented' at the time of their initial visit and were subsequently characterized as 'Demented' at a later visit. These fall under the 'Converted' category.
- combine it with the Demented values.

Process Workflow

Exploratory Data Analysis (EDA)

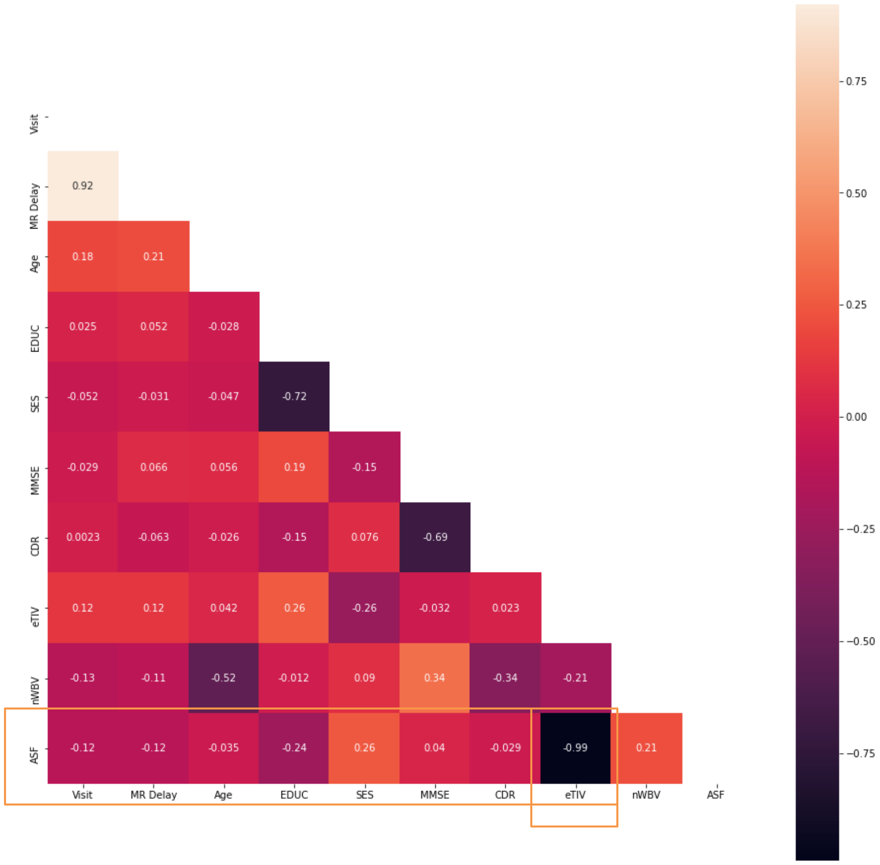
- The distribution on the integer/float type.
- Everything looks quite normalized except Mini-Mental State Examination score(MMSE) seems has some outlier.



Process Workflow

Exploratory Data Analysis (EDA)

- ASF and eTIV seem like multicollinearity so I will choose either one for training.



Process Workflow

Exploratory Data Analysis (EDA)

```
df.describe()
```

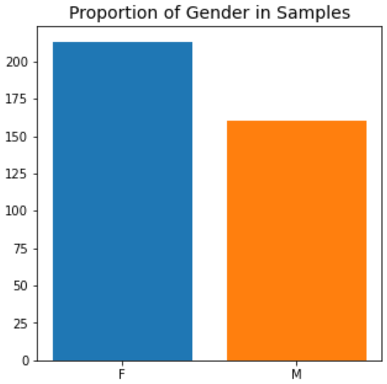
	Visit	MR Delay	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
count	373.000000	373.000000	373.000000	373.000000	354.000000	371.000000	373.000000	373.000000	373.000000	373.000000
mean	1.882038	595.104558	77.013405	14.597855	2.460452	27.342318	0.290885	1488.128686	0.729568	1.195461
std	0.922843	635.485118	7.640957	2.876339	1.134005	3.683244	0.374557	176.139286	0.037135	0.138092
min	1.000000	0.000000	60.000000	6.000000	1.000000	4.000000	0.000000	1106.000000	0.644000	0.876000
25%	1.000000	0.000000	71.000000	12.000000	2.000000	27.000000	0.000000	1357.000000	0.700000	1.099000
50%	2.000000	552.000000	77.000000	15.000000	2.000000	29.000000	0.000000	1470.000000	0.729000	1.194000
75%	2.000000	873.000000	82.000000	16.000000	3.000000	30.000000	0.500000	1597.000000	0.756000	1.293000
max	5.000000	2639.000000	98.000000	23.000000	5.000000	30.000000	2.000000	2004.000000	0.837000	1.587000

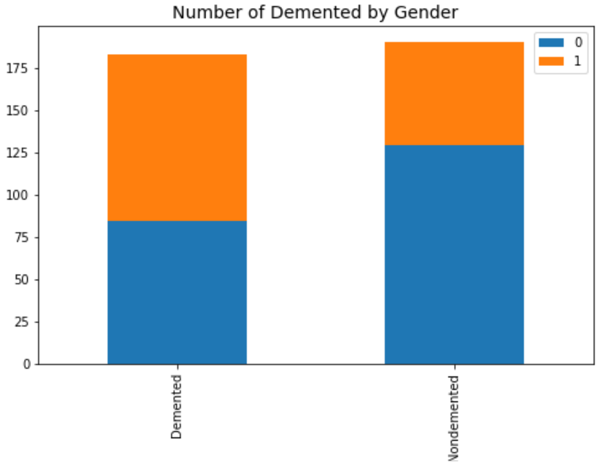
Process Workflow

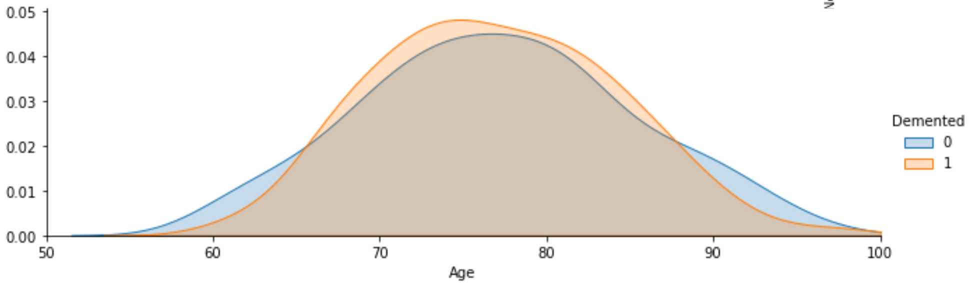
Data Analysis

- Men are slightly more likely with demented, an Alzheimer's Disease, than Women.

- There is a higher concentration of 70-80 years old in the Demented patient group than those in the nondemented patients. The patients who suffered from the disease has lower survival rate so that there are a few of 90 years old.







Process Workflow

Data preparation

- Replace null in the SES column by mean value and MESS column by median value (due to the outlier).

```
median_imputer = SimpleImputer(missing_values=np.nan, strategy='median')
mean_imputer = SimpleImputer(missing_values=np.nan, strategy='mean')
```

```
df["MMSE"] = median_imputer.fit_transform(df[["MMSE"]]).ravel()
```

```
df["SES"] = mean_imputer.fit_transform(df[["SES"]]).ravel()
```

- Encoding data (ex: combine demented and converted, rename column)
- Put the target-Demented (previously named "Group") at the last column
- Select features for training and testing
- Use stratify to preserve the proportion of target as in original dataset, in the train and test datasets as well.
- Then allocated 80% for training and 20% for testing.

```
# drop un-use column
X = df[['Gender',
        'Age',
        'EDUC',
        'SES',
        'MMSE',
        'eTIV',
        'nWBV',
       ]] # input
y = df['Demented'].values # output (dependent variable)
```

```
# split data
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    stratify=y,
                                                    test_size=0.2)
```

Process Workflow

Machine Learning model training/evaluation

- Logistic Regression
- Linear Svc
- Random Forest Classifier
- MLP Nueral Networks
- Use SelectKBest to select 5 best features.

Before Select Kbest:

	Gender	Age	EDUC	SES	MMSE	eTIV	nWBV
0	1	87	14	2.000000	27.0	1987	0.696
1	1	88	14	2.000000	30.0	2004	0.681
2	1	75	12	2.460452	23.0	1678	0.736
3	1	76	12	2.460452	28.0	1738	0.713
4	1	80	12	2.460452	22.0	1698	0.701
...	...	...	...	...	...	...	...
368	1	82	16	1.000000	28.0	1693	0.694
369	1	86	16	1.000000	26.0	1688	0.675
370	0	61	13	2.000000	30.0	1319	0.801
371	0	63	13	2.000000	30.0	1327	0.796
372	0	65	13	2.000000	30.0	1333	0.801

373 rows × 7 columns

After Select Kbest:

	Gender	EDUC	SES	MMSE	eTIV
0	1	14	2.000000	27.0	1987
1	1	14	2.000000	30.0	2004
2	1	12	2.460452	23.0	1678
3	1	12	2.460452	28.0	1738
4	1	12	2.460452	22.0	1698
...	...	...	...	...	...
368	1	16	1.000000	28.0	1693
369	1	16	1.000000	26.0	1688
370	0	13	2.000000	30.0	1319
371	0	13	2.000000	30.0	1327
372	0	13	2.000000	30.0	1333

373 rows × 5 columns

Results

After select the best feature the overall accuracy is improved.

Before Select Kbest:

	Model	Precision	Recall	f1 score	AUC
0	Logistic Regression	0.864865	0.864865	0.864865	0.866643
1	Linear SVC	0.493333	1.000000	0.660714	0.500000
2	Random Forest	0.850000	0.918919	0.883117	0.880512
3	MLP Neural Networks	0.312500	0.270270	0.289855	0.345661

After Select Kbest:

	Model	Precision	Recall	f1 score	AUC
0	Logistic Regression	0.888889	0.648649	0.750000	0.784851
1	Linear SVC	1.000000	0.081081	0.150000	0.540541
2	Random Forest	0.914286	0.864865	0.888889	0.892959
3	MLP Neural Networks	0.689655	0.540541	0.606061	0.651849

Conclusion

- Random forest classifier is the best performing model so far.
- MMSE is one of the gold standards for determining dementia and so we think it is an important feature to include.
- The estimated total intracranial volume (eTIV) is also another key feature to included.
- Need more data for more precise analysis and accuracy.

Future Opportunities

- To improve the understanding through more sophisticated EDA process with a larger sample size.
- Generation group, grade volume of brain tissue or exam score.

Appendix

- https://www.kaggle.com/jboysen/mri-and-alzheimers
- https://www.kaggle.com/ruslankl/dementia-prediction-w-tree-based-models
- https://en.wikipedia.org/wiki/Alzheimer%27s_disease
- https://www.alz.org/alzheimers-dementia/10_signs
- https://alz.org.sg/dementia/singapore/
- https://www.alzint.org/about/dementia-facts-figures/types-of-dementia/alzheimers-disease/