Alzheimer’s Disease Prediction

Tool: python, jupyter notebook, machine learning model
Source: Kaggle
Duration: 1 week

Problem Statement

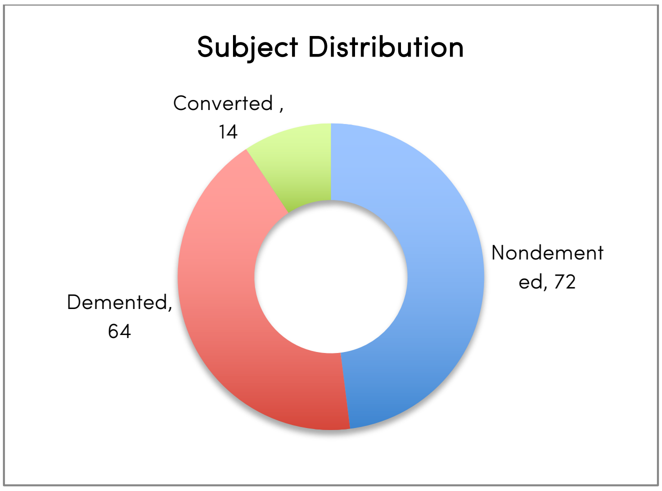
Alzheimer’s disease (AD) is a neurodegenerative disorder of uncertain cause and pathogenesis that primarily affects older adults and is the most common cause of dementia. So far there is no cure currently available but treatments to ameliorate some symptoms. Brain Imaging and MRI, are used for evaluation 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. The motivation is to slow down the progress of a patient from mild cognitive impairment to dementia.

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.



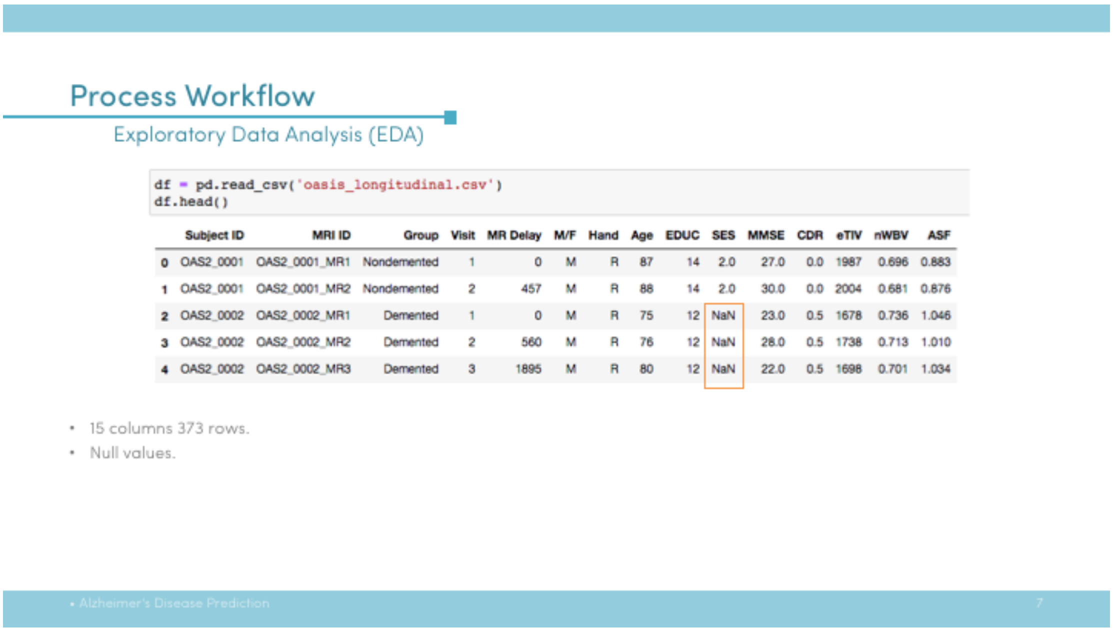
- Column descriptions

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

Exploratory Data Analysis (EDA)

This is the initial data looks like. There are 15 columns and 373 rows. From here we can see there is null value in the SES and MMSE column.



Process Workflow

Exploratory Data Analysis (EDA)

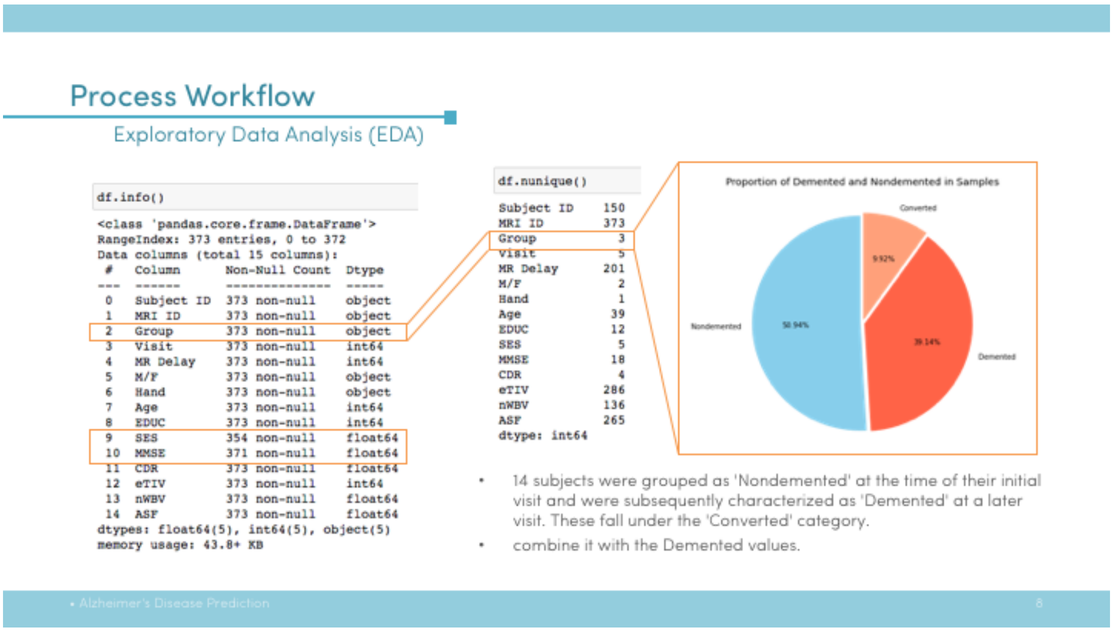
```
df = pd.read_csv('oasis_longitudinal.csv')
df.head()
```

	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.

• Alzheimer's Disease Prediction 7

There is a column named 'Group'. By looking into it, this column is consist the demented, nondemented and converted, which are actually the target for our machine learning later. Since the converted were subsequently characterized as ‘demented’, I combined it with the demented values, so this will become binary classification problem. From the pie chart here. We know that our target data is quite a balance dataset.



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
```

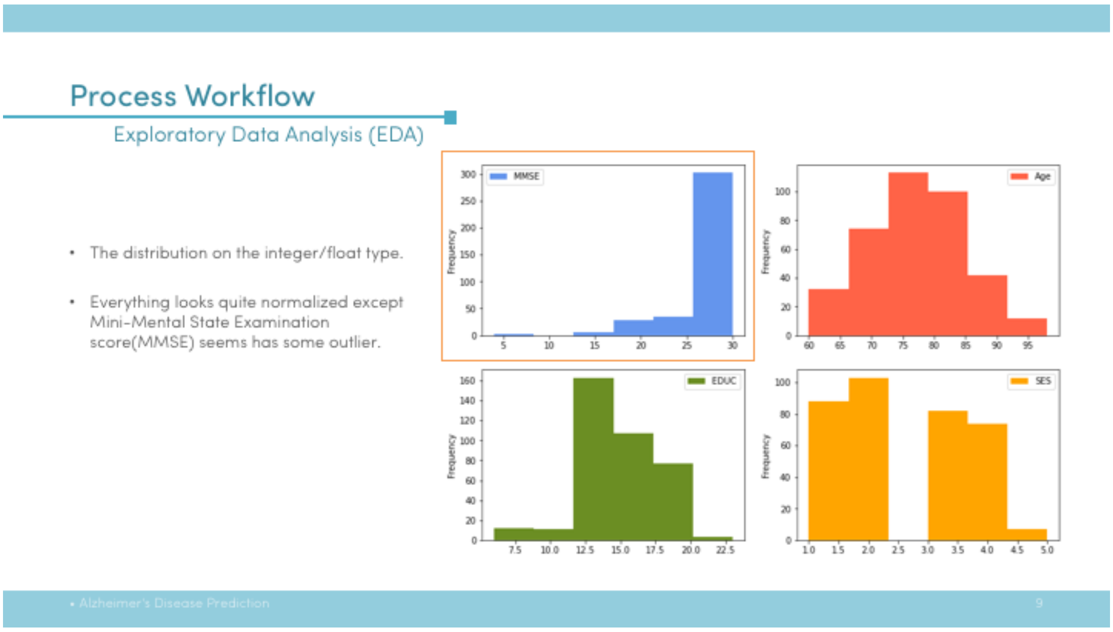
```
df.nunique()
```

```
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.

• Alzheimer's Disease Prediction 8

Next, we need to check about the distribution. Here is the distribution on the integer and float type data. As we know earlier, there are some null value in the MMSE and SES column. Everything looks quite normalized except the MMSE seems has some outlier. However, MMSE is the exam score between 0 to 30, so instead of removing the outlier I replaced the null to median.



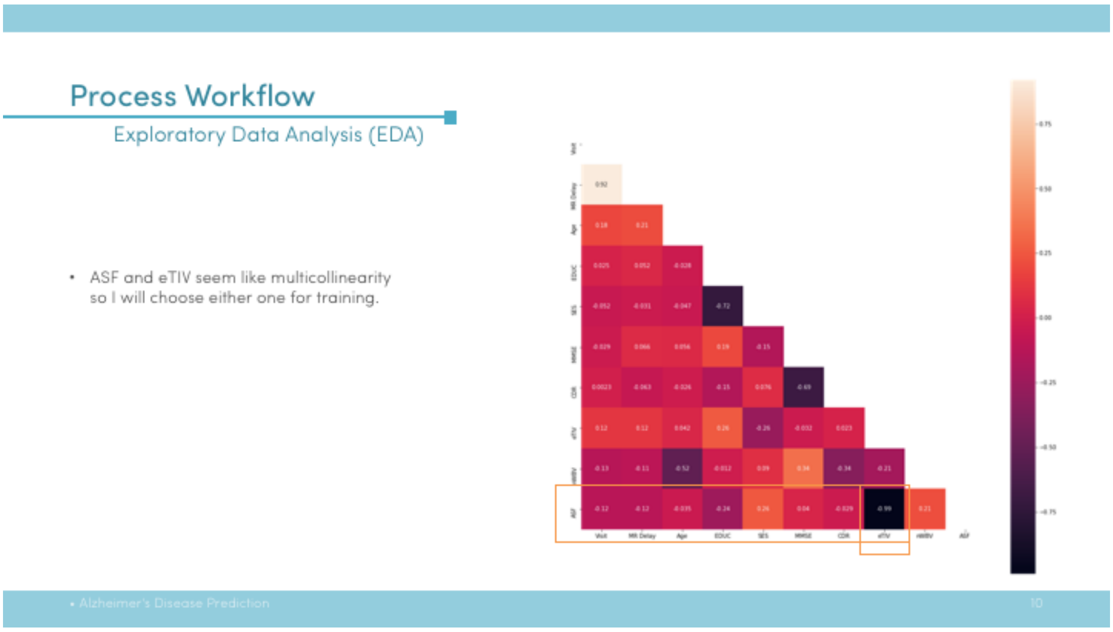
Process Workflow

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.

• Alzheimer's Disease Prediction 9

From this correlation heatmap, we found that the ASF and eTIV seem has a multicollinearity. So later I will just use either one.



Process Workflow

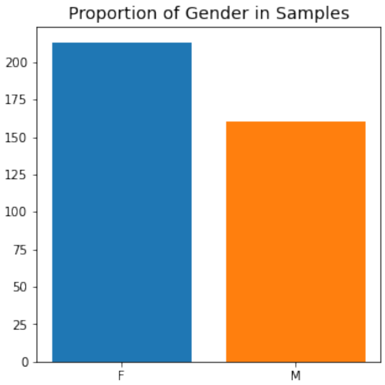
Exploratory Data Analysis (EDA)

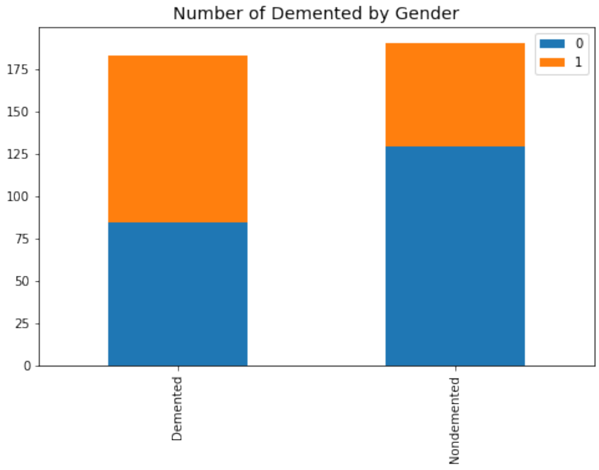
- ASF and eTIV seem like multicollinearity so I will choose either one for training.

• Alzheimer's Disease Prediction

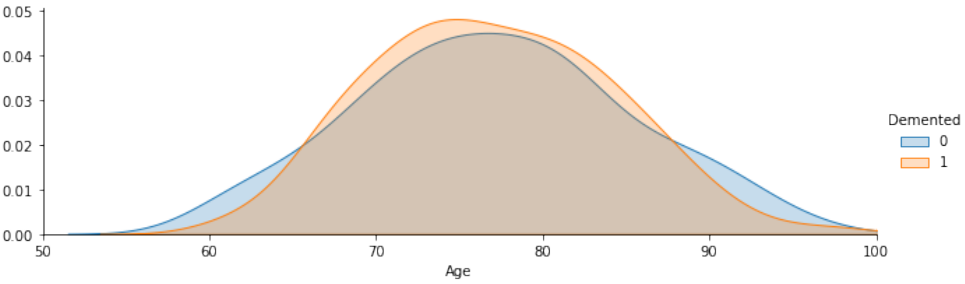
10

Before jump into data preparation, here is some data analysis on the dataset: There are more female in the dataset. Seem like Men are more likely with demented than women.





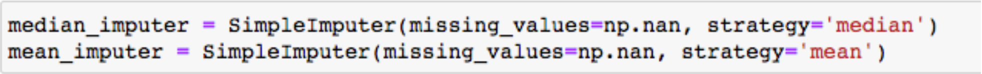
There is a higher concentration of 70-80 years old in the demented patient group than those in the nondemented patients. The patient who suffered from the disease has lower survival rate so there are a few of 90 years old.



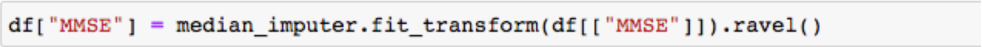
Data Preparation

It is time to prepare our data for training. Firstly is to:

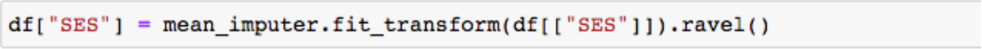
- Replace null in the SES column by mean value and MESS column by median value.



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

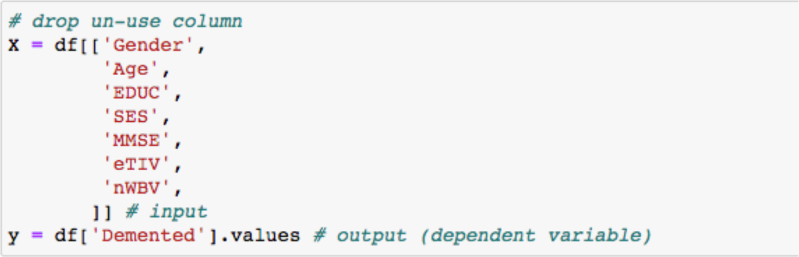


```
df["MMSE"] = median_imputer.fit_transform(df[["MMSE"]]).ravel()
```

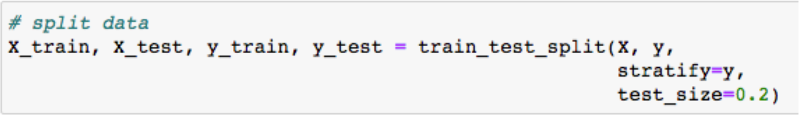


```
df["SES"] = mean_imputer.fit_transform(df[["SES"]]).ravel()
```

- Combine demented and converted and rename column.
- Encoding the object type of data, for example, the target, because our target is demented and nondemented, so need to transform to numerical data. After that will rename the column to demented for easy to understand. And then I will put this column at the last column.
- Select the initial features for training and testing. They are , Gender, Age, EDUC, SES, MMSE, eTIV, nWBV, total 7 features.
- Use stratify to preserve the proportion of target as in original dataset, in the train and test datasets as well.
- Lastly, 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)
```

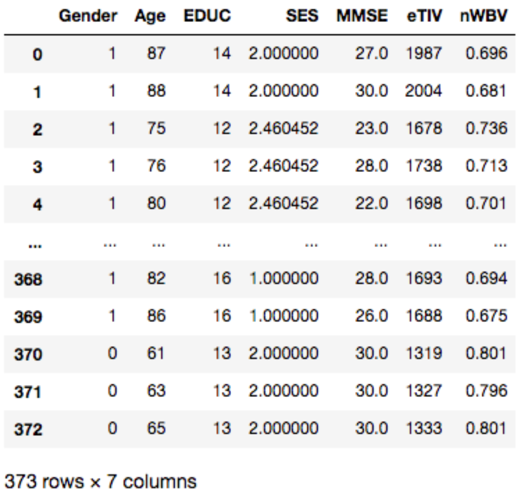


```
# split data
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    stratify=y,
                                                    test_size=0.2)
```

Machine learning model training/evaluation

1. Logistic regression
2. Linear SVC
3. Random Forest Classifier
4. MLP Nueral Networks
5. Using SelectBest to select 5 best features.

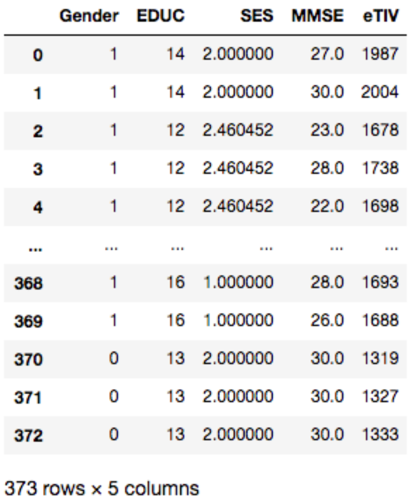
Before Select Kbest:



	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:

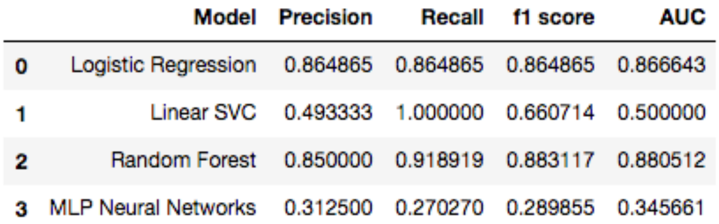


	Gender	EDUC	SES	MMSE	eTIV
0	1	14	2.000000	27.0	1987
1	1	14	2.000000	30.0	2004
2	1	12	2.460452	23.0	1678
3	1	12	2.460452	28.0	1738
4	1	12	2.460452	22.0	1698
...	...	...	...	...	...
368	1	16	1.000000	28.0	1693
369	1	16	1.000000	26.0	1688
370	0	13	2.000000	30.0	1319
371	0	13	2.000000	30.0	1327
372	0	13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

373 rows × 5 columns

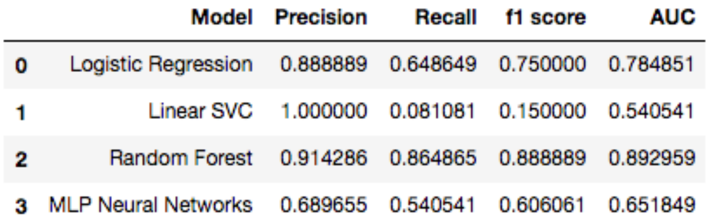
Below is the results of different model. After select the best features, the overall accuracy is improved.

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. It is an important feature to include. The estimated total intracranial volume (eTIV) is also another key feature to included. However, we need more data for more precise analysis and accuracy.

Future Opportunities

We may improve our understanding through more sophisticated EDA process with a larger sample size. Like for example, instead of age, we may group it into different generation, grade volume of brain tissue or exam scores. Then the accuracy of the prediction model can be further improved.