Machine Learning Capstone PROJECT PROPOSAL

Domain Background

[Student briefly details background information of the domain from which the project is proposed. Historical information relevant to the project should be included. It should be clear how or why a problem in the domain can or should be solved. Related academic research should be appropriately cited. A discussion of the student's personal motivation for investigating a particular problem in the domain is encouraged but not required.]

This capstone project involves machine learning modeling and analysis of clinical, demographic, and brain related derived anatomic measures from human MRI (magnetic resonance imaging) tests (http://www.oasis-brains.org/ (http://www.oasis-brains.org/)). The objectives of these measurements are to diagnose the level of Dementia in the individuals and the probability that these individuals may have Alzheimer's Disease (AD).

In published studies, Machine Learning has been applied to Alzheimers/Dementia identification from MRI scans and related data in the following academic papers:

https://www.hindawi.com/journals/jhe/2017/5485080/ (https://www.hindawi.com/journals/jhe/2017/5485080/)

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.278.3712&rep=rep1&type=pdf (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.278.3712&rep=rep1&type=pdf)

Recently, a close relative of mine had to undergo a sequence of MRI tests for cognition difficulties.The motivation for choosing this topic for the Capstone project arose from the desire to understand and analyze potential for Dementia and AD from MRI related data. This Capstone project does not use the MRI "imaging" data and does not focus on AD, focusses only on Dementia.

Problem Statement

[Student clearly describes the problem that is to be solved. The problem is well defined and has at least one relevant solution]

- Cross-Sectional and longitudinal OASIS MRI structural and demographic data (clinical, demographic, and brain related derived anatomic measures) from human MRI (magnetic resonance imaging) tests (http://www.oasis-brains.org/ (http://www.oasis-brains.org/)) will be used to train a set of linear and non-linear machine learning classification models.

- Clinical Dementia Rating (CDR) values provided in the data set will be used as "labels" for training the classification models. [Clinical Dementia Ratings (CDR values: 0=nondemented; 0.5 = very mild dementia; 1 = mild dementia; 2 = moderate dementia)]

- Pandas will be used for data loading and Python scikit-learn library for modeling.

- The goal is to train machine learning models to predict whether the individuals in the cross-validation set (test set) have dementia (CDR>0), and if they do, the severity level of dementia (CDR values of 0.5, 1, and 2). The problem will be formulated both as a binary classification problem (CDR=0, and CDR>0), and multiclass classification problem (CDR values in the dataset: 0, 0.5, 1, and 2). In the binary classification formulation, the CDR>0 the values in the sliced dataset will be relabeled as CDR=11.

- Classification Accuracy will be used as the primary metric. The results from the best model (one that provides the highest accuracy) will be reported along with those from the other models.

- About 80% of the data in the dataset will be used for training the models. About 20% of data will used prediction of the CDR label for the k-fold cross-validation with k=10. Sensitivity studies with proportion other than 80:20, e.g. 70:30, will be used to test sensitivity of this split on the accuracy.

- The base case will uses a dataset that combines the cross-sectional and the longitudinal MRI datasets.This has the benefit of having a larger dataset. The cross-sectional and the longitudinal datasets will also be trained/cross-validated separately, and classification accuracy will be reported.

- Data cleaning (e.g. removal of NaN values), dat exploration, data preparation, data visualization, and data preprocessing will be described, as needed, and the impact of the latter on prediction accuracy will be discussed.

Datasets and Inputs

[The dataset(s) and/or input(s) to be used in the project are thoroughly described. Information such as how the dataset or input is (was) obtained, and the characteristics of the dataset or input, should be included. **It should be clear how the dataset(s) or input(s) will be used in the project and whether their use is appropriate given the context of the problem.**]

Data source:

Reference 1 below provide the downloadable MRI related data in csv format. Reference 2 provides meta data and additional facts about the cross-sectional MRI. Reference 3 provides additional follow up reliability data where the test candidates continued to have no dementia. Reference 3 data may be optionally used to predict additional CDR=0 (no Dementia) cases.

Reference 1: http://www.oasis-brains.org/app/template/Index.vm;jsessionid=6926BBF18A3D5CD974E750FAC8ED01CE (http://www.oasis-brains.org/app/template/Index.vm;jsessionid=6926BBF18A3D5CD974E750FAC8ED01CE)

Reference 2: http://www.oasis-brains.org/pdf/oasis_cross-sectional_facts.pdf (http://www.oasis-brains.org/pdf/oasis_cross-sectional_facts.pdf)

Reference 3: http://www.oasis-brains.org/app/action/BundleAction/bundle/OAS1_RELIABILITY (http://www.oasis-brains.org/app/action/BundleAction/bundle/OAS1_RELIABILITY)

OASIS: Cross-sectional MRI Data in Young, Middle Aged, Nondemented and Demented Older Adults

- This dataset consists of a cross-sectional collection for 416 persons aged 18 to 96.
- For each person, 3 to 4 T1-weighted MRI scans that were obtained in single scan sessions are included.
- The persons include both men and women, and are all right-handed.
- In this dataset, one hundred persons over the age of 60 have been clinically diagnosed with very mild to moderate Alzheimer’s disease (AD).
- Also, a reliability data set (Reference 3 below) is included which contains 20 nondemented subjects imaged on a subsequent visit within 90 days of their initial session.
- See Dementia related **Additional Data** below for the cross-sectional MRI cases used this project. Features based on these **Additional Data** will be used to train classification models to predict the labels for the outcome (CDR).

OASIS: Longitudinal MRI Data in Nondemented and Demented Older Adults

- This set consists of a longitudinal collection of 150 subjects aged 60 to 96.
- Each subject was scanned on two or more visits, separated by at least one year for a total of 373 imaging sessions.
- For each subject, 3 or 4 individual T1-weighted MRI scans obtained in single scan sessions are included. **MRI image pixel data are NOT used in this problem, only related features prefixed with @ sign (below) will be used.**
- The subjects are all right-handed and include both men and women.
- 72 of the subjects were characterized as nondemented throughout the study.

- 64 of the included subjects were characterized as demented at the time of their initial visits and remained so for subsequent scans, including 51 individuals with mild to moderate Alzheimer's disease. Another 14 subjects were characterized as nondemented at the time of their initial visit and were subsequently characterized as demented at a later visit.
- See Dementia related **Additional Data** below for the longitudinal MRI cases used this project.
- Features based on the **Additional Data** are relevant to finding machine learning solutions to the problem defined above, and will be used to train classification models to predict the labels for the outcome (Critical Dementia Rating, CDR).

- **Additional data:** Specific References in parentheses below covering features are from Reference 2: http://www.oasis-brains.org/pdf/oasis_cross-sectional_facts.pdf (http://www.oasis-brains.org/pdf/oasis_cross-sectional_facts.pdf) . These features include Demographic, clinical, and derived anatomic measures related to brain that are located in the file oasis_crosssectional.csv. Features prefixed with @ will be used for the problem.
- Demographic data include:
    - @Gender (M/F), categorical data
    - Handedness (Right or eft Handed), categorical data, all of which are right hand data
    - @Age (numeric),
    - @Education (Educ). Education codes correspond to the following levels of education: 1: less than high school grad. 2: high school grad. 3: some college 4: college grad. 5: beyond college.
- Clinical data include:
    - @Mini-Mental State Examination (MMSE, Rubin et al., 1998),
    - @Clinical Dementia Rating (CDR)

      -- 0=nondemented (341 data points)

      -- 0.5 = very mild dementia (193 data points)

      -- 1 = mild dementia (69 data points);

      -- 2 = moderate dementia (5 data points), from Morris, 1993.

```
- All participants with dementia (CDR >0) were diagnosed with probable AD.
```

- Derived anatomic volumes data include:
    - @Estimated total intracranial volume (eTIV, mm3), Buckner et al., 2004
    - @Atlas scaling factor (ASF), Buckner et al., 2004
    - @Normalized whole brain volume (nWBV, mm3), Fotenos et al., 2004

Solution Statement

*Student clearly describes a solution to the problem. The solution is applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, the solution is quantifiable, measurable, and replicable.

Solution:

- Train a supervised machine learning classification model to properly classify the OASIS data according to clinical dementia ratings(CDR values).
- Train a number of candidate models from the scikit-learn library such as Logistic Regression, Linear Discriminant Analysis, KNN, Naive Bayes, CART, and SVM.
- Select the best model based on the "Accuracy" Metric.
- Combine the cross-sectional and longitudinal MRI related demographic and clinical data into a single dataset.
- Split the dataset into training dataset (80%) and the remaining data(20%) for typical ten-fold cross validation.
- Report the prediction accuracy of the models and identify the model that yields the higest classification accuracy. Report accuracy results in sklearn Confusion Matrix format(to evaluate classifier output quality) and, Classification Report format (provides, precision, recall, f1-score). See details here: http://scikit-learn.org/stable/modules/model_evaluation.html (http://scikit-learn.org/stable/modules/model_evaluation.html)).

Benchmark Model

A benchmark model is provided that relates to the domain, problem statement, and intended solution. Ideally, the student's benchmark model provides context for existing methods or known information in the domain and problem given, which can then be objectively compared to the student's solution. The benchmark model is clearly defined and measurable.

Benchmark

Student clearly defines a benchmark result or threshold for comparing performances of solutions obtained.

My primary benchmark will be a neural network model based on the same data discussed above and as used in this problem: Will use Keras frontend and tensorflow backend. My secondary benchmark will be results of the study in the two papers below:

1.Paper title: Usefulness of data from magnetic resonance imaging to improve prediction of dementia: population based cohort study

http://www.bmj.com/content/350/bmj.h2863 (http://www.bmj.com/content/350/bmj.h2863)

"Results During 10 years of follow-up, there were 119 confirmed cases of dementia, 84 of which were Alzheimer’s disease. The conventional risk model incorporated age, sex, education, cognition, physical function, lifestyle (smoking, alcohol use), health (cardiovascular disease, diabetes, systolic blood pressure), and the apolipoprotein genotype (C statistic for discrimination performance was 0.77, 95% confidence interval 0.71 to 0.82). No significant differences were observed in the discrimination performance of the conventional risk model compared with models incorporating data from MRI including white matter lesion volume (C statistic 0.77, 95% confidence interval 0.72 to 0.82; P=0.48 for difference of C statistics), brain volume (0.77, 0.72 to 0.82; P=0.60), hippocampal volume (0.79, 0.74 to 0.84; P=0.07), or all three variables combined (0.79, 0.75 to 0.84; P=0.05). Inclusion of hippocampal volume or all three MRI variables combined in the conventional model did, however, lead to significant improvement in reclassification measured by using the integrated discrimination improvement index (P=0.03 and P=0.04) and showed increased net benefit in decision curve analysis. Similar results were observed when the outcome was restricted to Alzheimer’s disease."

```
    1a. C - Statistics: http://www.statisticshowto.com/c-statistic/
```

1. Paper Title: The Use of MRI and PET for Clinical Diagnosis of Dementia and Investigation of Cognitive Impairment: A Consensus Report

https://www.alz.org/national/documents/imaging_consensus_report.pdf (https://www.alz.org/national/documents/imaging_consensus_report.pdf)

"Once the presence of dementia has been established, the role of imaging in the diagnosis of dementia subtypes is very much a function of the clinical diagnosis. The accuracy of the clinical diagnosis of Alzheimer’s disease (AD) is quite good. Pathological AD has a prevalence of about 70% (range 50% to above 80% depending upon whether the AD occurs in isolation or with other entities) among all dementias (see evidence Table 1 in reference 1); thus, even clinicians with limited neurological expertise should have a diagnostic accuracy, for AD at least, at about that level. A review of 13 published studies gave average values for sensitivity and specificity of the clinical diagnosis of AD of 81% and 70%, respectively(1). The overall accuracy of the clinical diagnosis of AD versus not-AD compared with the neuropathological standard based on those values for prevalence, sensitivity, and specificity, is 78%. "

Evaluation Metrics

- Student proposes at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model presented. The evaluation metric(s) proposed are appropriate given the context of the data, the problem statement, and the intended solution.
- The number of records for classes for CDR=0, and CDR>0 are 341 and 267 respectively, and may be considered balanced. So, accuracy may be used as a metric for this binary classification problem. As an option will consider reporting the AUC metric
- Will Report classification accuracy (number of records correctly classified divided total number of records classified). Also report accuracy results in sklearn Confusion Matrix format(to evaluate classifier output quality) and Classification Report format (provides, precision, recall, and f1-score) which are quite appropriate for the dataset used to train models for CDR classification. See details and discussion of these sklearn metrics here: http://scikit-learn.org/stable/modules/model_evaluation.html (http://scikit-learn.org/stable/modules/model_evaluation.html)).

Project Design

- Student summarizes a theoretical workflow for approaching a solution given the problem. Discussion is made as to what strategies may be employed, what analysis of the data might be required, or which algorithms will be considered. The workflow and discussion provided align with the qualities of the project. Small visualizations, pseudocode, or diagrams are encouraged but not required.

The following steps will be used. Some of the step details are patterned after recommendations from: https://machinelearningmastery.com/ (https://machinelearningmastery.com/)

1. Download the CSV data from the OASIS web site (Reference 1).
2. Load the data into a PANDAS dataframe.
3. Load libraries (from Pandas, sklearn, and matplotlib)
4. Summarize Data

```
4a) Descriptive statistics: use the Python Pandas Describe() method.
4b) Explore and visualize the dataset. Use histograms, scatter plots. Use Ma
tplotlib for visualization.
```

5. Clean the dataset

```
5a) remove NaN or replace NaN values with mean from the dataset;
5b) remove/replace missing data, if any.
```

6. Prepare Data (rename columns appropriately to be able to combine the cross-sectional and the longitudinal mri data.)
7. Preprocess and transform data, if appropriate, using normalization and/or rescaling the data. Check whether normalization and scaling will improve classification accuracy. As sensitivity study, use Data Transforms, where attributes are scaled or redistributed, in order to best expose the structure of the problem later to learning algorithms.
8. Feature selection 8a) Check correlation among features by plotting correlation matrix using the corr() method to help in selecting features that are not strongly correlated with each other. This will yield optimized model input for higher accuracy. Feature selection methods are useful where redundant features may be removed and new features developed.

   8b) As a sensitivity study check whether PCA (principle component analysis) is helpful in reducing features.
9. Evaluate Algorithms: Evaluate the classification algorithms identified earlier in this section.

   9a) Split the dataset into training dataset and validation dataset. Use seed to have reproducible results for random states.

   9b) Define test options using scikit-learn such as cross-validation and choose the evaluation metrics (see Evaluation Metrics section below)

   9c) Spot Check and Compare Algorithms: Using sklearn methods, will spot-check the suite of linear and nonlinear machine learning algorithms mention earlier in this section and compare the estimated accuracy of these algorithms. Will pick the algorithm that provides highest cross-validation accuracy.

10.Improve Accuracy: Will use the two prevalent and different ways to improve the accuracy of the models:

```
10a. Algorithm Tuning - Search for a combination of parameters for each algorithm using scikit-learn that yields
the best results.

10b. Ensembles - Combine the prediction of multiple models into an ensemble prediction using ensemble
techniques in scikit-learn.
```

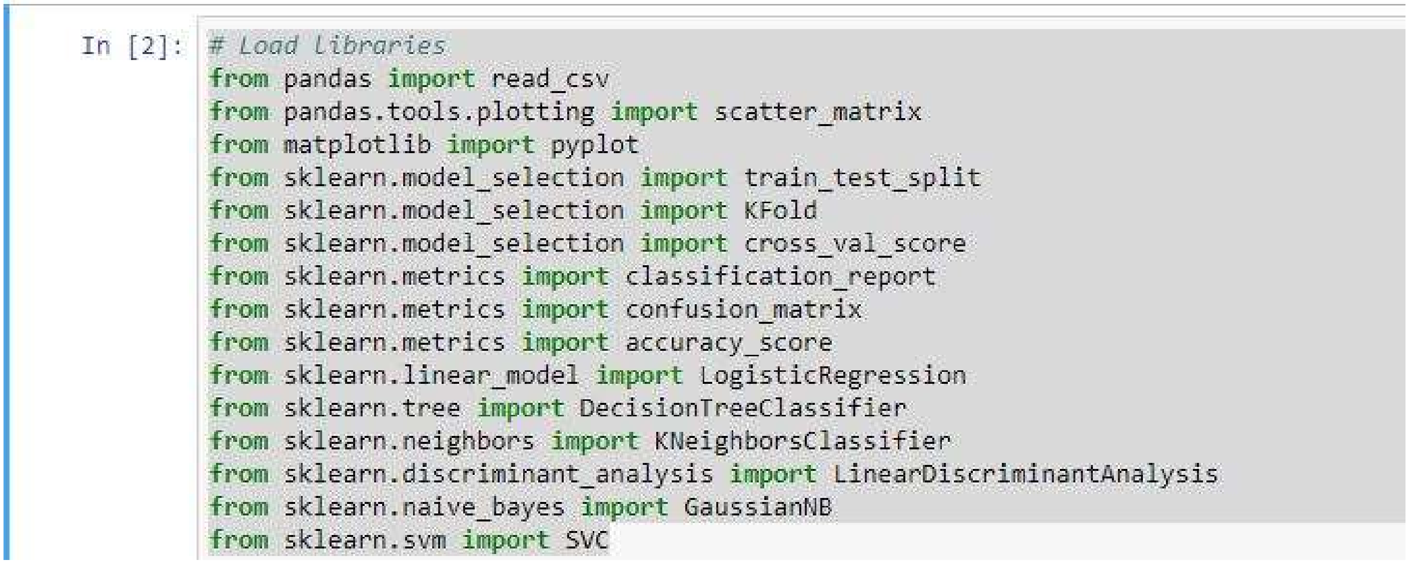
11.Finalize Model: Will use an optimal model tuned by scikit-learn to make predictions on unseen data.

```
11a. Will create a standalone model on entire training dataset using the parameters tuned by scikit-learn.

11b) Make Predictions on the validation dataset

11c) Create standalone model

11d) Save model for later use
```

Note that image (pixel or voxel) data are not used in this problem. The image data alone is not sufficient for non-radiologists like me as it requires radiological interpretation for additional feature, something beyond my expertise. As such complex CNN model will not be attempted here. This student has peviously graduated from the Udacity Deep Learning Nanodegree program where he submitted image recognition projects that used complex convolutional neural network (CNN) using tensorflow.

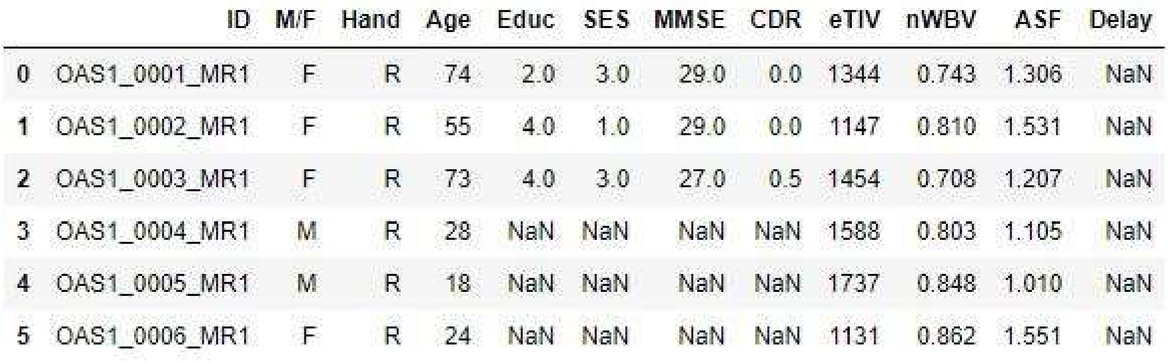
Representative screenshots and extracts of Python code and plots that will be part of the Analysis Report

- Load Libraries



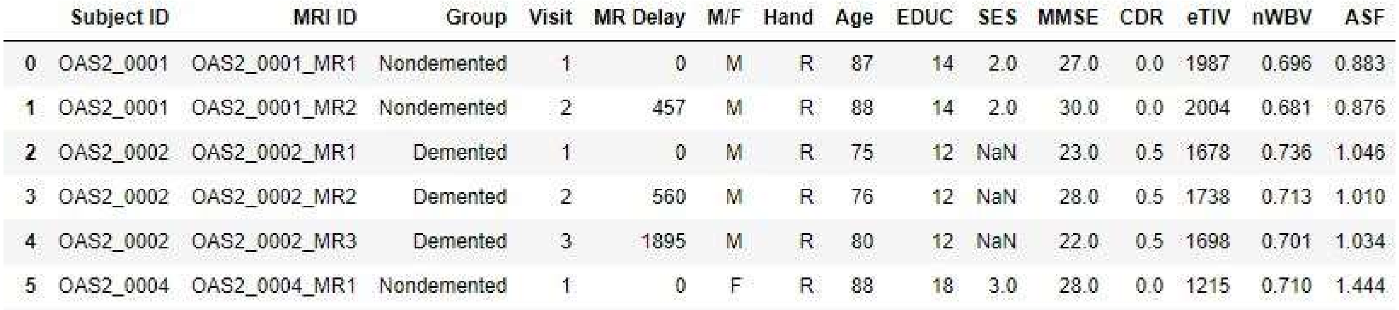
```
In [2]: # Load Libraries
        from pandas import read_csv
        from pandas.tools.plotting import scatter_matrix
        from matplotlib import pyplot
        from sklearn.model_selection import train_test_split
        from sklearn.model_selection import KFold
        from sklearn.model_selection import cross_val_score
        from sklearn.metrics import classification_report
        from sklearn.metrics import confusion_matrix
        from sklearn.metrics import accuracy_score
        from sklearn.linear_model import LogisticRegression
        from sklearn.tree import DecisionTreeClassifier
        from sklearn.neighbors import KNeighborsClassifier
        from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
        from sklearn.naive_bayes import GaussianNB
        from sklearn.svm import SVC
```

- Cross sectional data sample (Pandas data frame)



	ID	M/F	Hand	Age	Educ	SES	MMSE	CDR	eTIV	nWBV	ASF	Delay
0	OAS1_0001_MR1	F	R	74	2.0	3.0	29.0	0.0	1344	0.743	1.306	NaN
1	OAS1_0002_MR1	F	R	55	4.0	1.0	29.0	0.0	1147	0.810	1.531	NaN
2	OAS1_0003_MR1	F	R	73	4.0	3.0	27.0	0.5	1454	0.708	1.207	NaN
3	OAS1_0004_MR1	M	R	28	NaN	NaN	NaN	NaN	1588	0.803	1.105	NaN
4	OAS1_0005_MR1	M	R	18	NaN	NaN	NaN	NaN	1737	0.848	1.010	NaN
5	OAS1_0006_MR1	F	R	24	NaN	NaN	NaN	NaN	1131	0.862	1.551	NaN

- Longitudinal data sample (Pandas data frame)



	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
5	OAS2_0004	OAS2_0004_MR1	Nondemented	1	0	F	R	88	18	3.0	28.0	0.0	1215	0.710	1.444

- Merge the Cross Sectional and the Longitudinal data sets (Pandas data frames) and get descriptive statistics

```
# Merge the cross-sectional and the longitudinal MRI datasets
dfoas_merge = dfoasx.append(dfoasl2, ignore_index=True)
dfoas_merge.shape
```

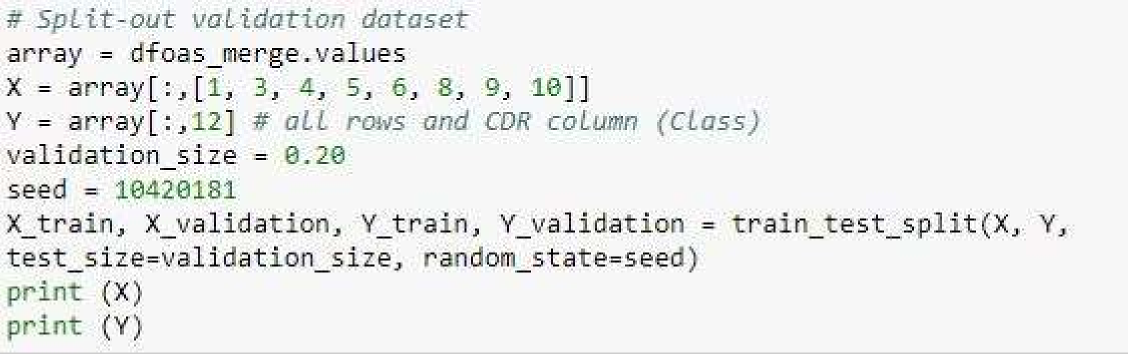
```
(809, 12)
```

```
dfoas_merge.describe()
```

	M/F	Age	Educ	SES	MMSE	eTIV	nWBV	ASF	Delay
count	809.000000	809.000000	608.000000	570.00000	606.000000	809.000000	809.000000	809.000000	393.000000
mean	0.594561	63.186650	10.184211	2.47193	27.234323	1484.782447	0.763037	1.197311	565.865140
std	0.491280	23.117511	6.058388	1.12805	3.687980	166.911689	0.059401	0.133031	631.862452
min	0.000000	18.000000	1.000000	1.00000	4.000000	1106.000000	0.644000	0.876000	0.000000
25%	0.000000	49.000000	4.000000	2.00000	26.000000	1361.000000	0.715000	1.108000	0.000000
50%	1.000000	72.000000	12.000000	2.00000	29.000000	1475.000000	0.754000	1.190000	497.000000
75%	1.000000	80.000000	16.000000	3.00000	30.000000	1583.000000	0.817000	1.290000	846.000000
max	1.000000	98.000000	23.000000	5.00000	30.000000	2004.000000	0.893000	1.587000	2639.000000

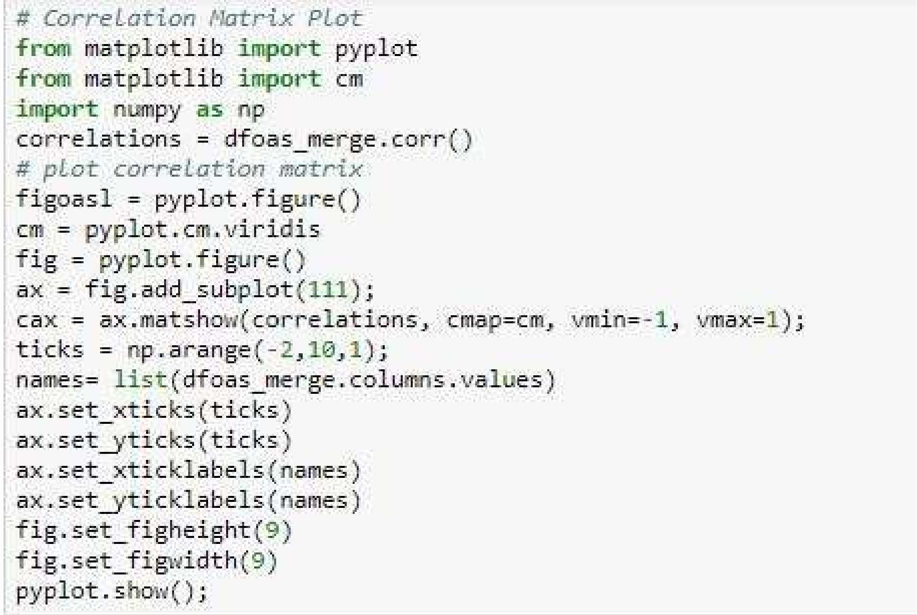
- Drop rows with NaN

  dfoas_merge=dfoas_merge.dropna(how='any')
- Code: split merged dataset into a Training set and cross-validation set

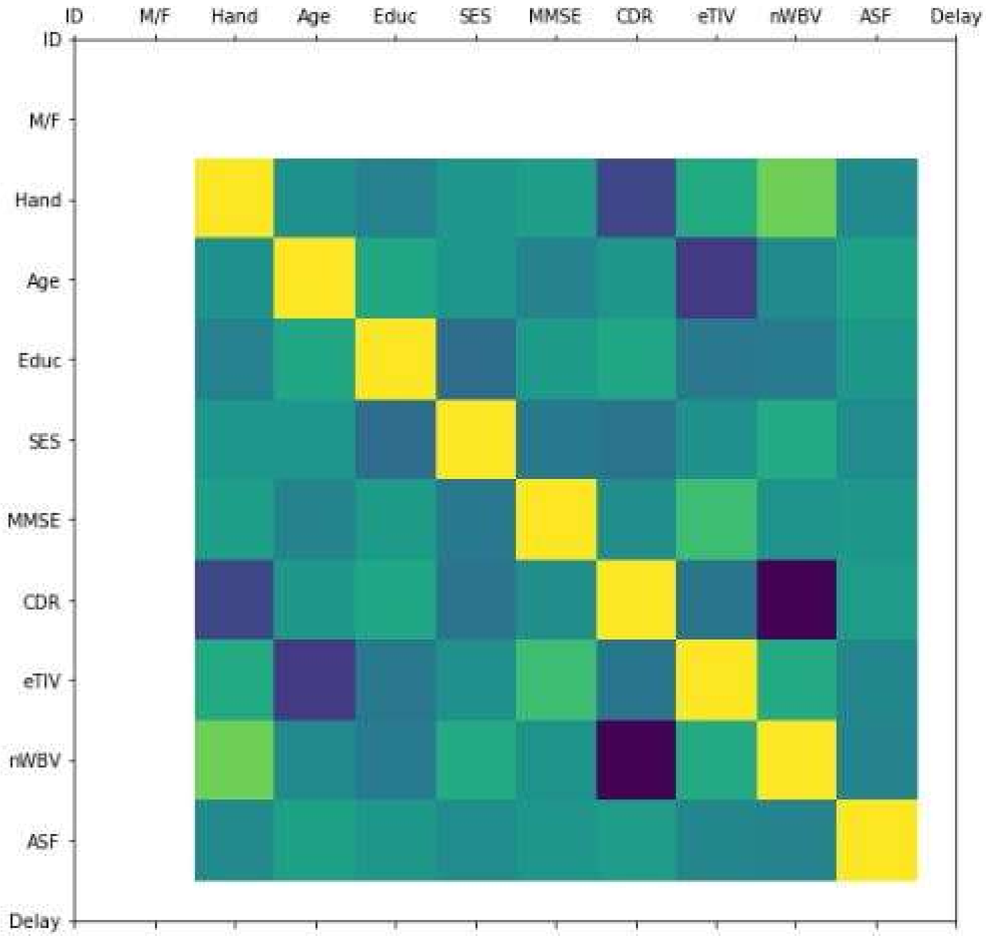


```
# Split-out validation dataset
array = dfoas_merge.values
X = array[:,[1, 3, 4, 5, 6, 8, 9, 10]]
Y = array[:,12] # all rows and CDR column (Class)
validation_size = 0.20
seed = 10420181
X_train, X_validation, Y_train, Y_validation = train_test_split(X, Y,
test_size=validation_size, random_state=seed)
print (X)
print (Y)
```

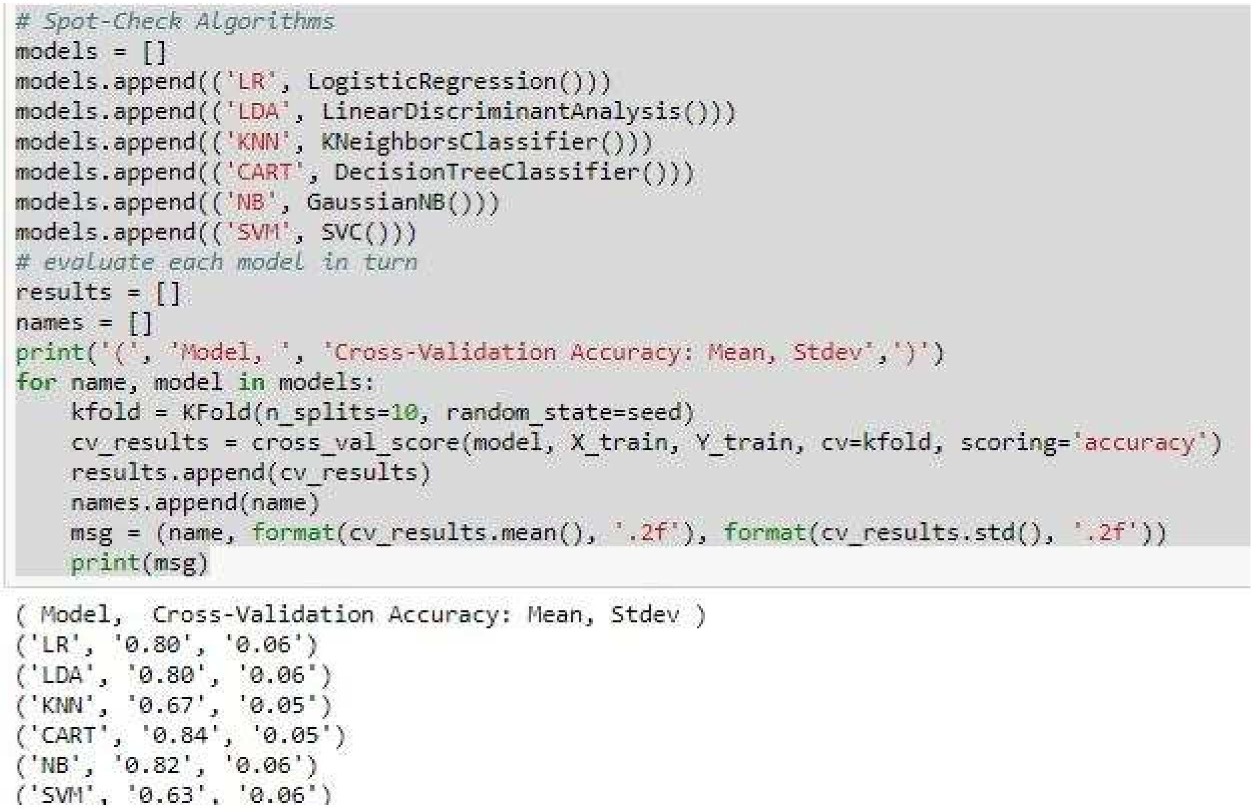
- Python Code and Plot showing correlation among features in the merged dataset



```
# Correlation Matrix Plot
from matplotlib import pyplot
from matplotlib import cm
import numpy as np
correlations = dfoas_merge.corr()
# plot correlation matrix
figoasl = pyplot.figure()
cm = pyplot.cm.viridis
fig = pyplot.figure()
ax = fig.add_subplot(111);
cax = ax.matshow(correlations, cmap=cm, vmin=-1, vmax=1);
ticks = np.arange(-2,10,1);
names= list(dfoas_merge.columns.values)
ax.set_xticks(ticks)
ax.set_yticks(ticks)
ax.set_xticklabels(names)
ax.set_yticklabels(names)
fig.set_figheight(9)
fig.set_figwidth(9)
pyplot.show();
```



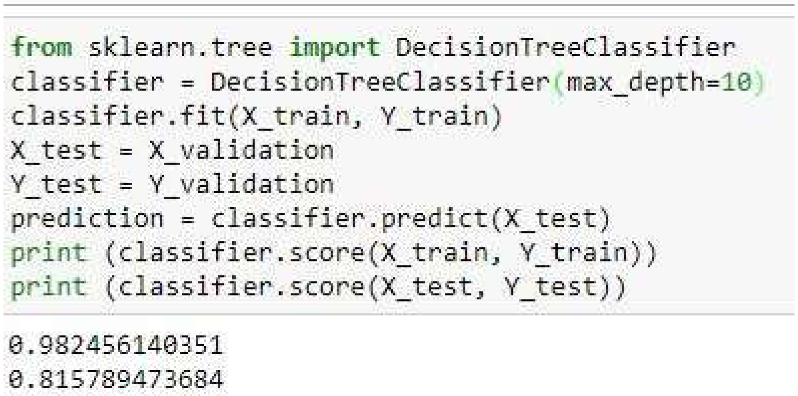
- Quick check of cross-validation accuracy of the chosen set of sklearn algorithms when applied to the merged MRI dataset



```
# Spot-Check Algorithms
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC()))
# evaluate each model in turn
results = []
names = []
print('(', 'Model, ', 'Cross-Validation Accuracy: Mean, Stdev',')')
for name, model in models:
    kfold = KFold(n_splits=10, random_state=seed)
    cv_results = cross_val_score(model, X_train, Y_train, cv=kfold, scoring='accuracy')
    results.append(cv_results)
    names.append(name)
    msg = (name, format(cv_results.mean(), '.2f'), format(cv_results.std(), '.2f'))
    print(msg)
```

```
( Model,  Cross-Validation Accuracy: Mean, Stdev )
('LR', '0.80', '0.06')
('LDA', '0.80', '0.06')
('KNN', '0.67', '0.05')
('CART', '0.84', '0.05')
('NB', '0.82', '0.06')
('SVM', '0.63', '0.06')
```

- Training and Cross-Validation Accuracy for the Decision Tree Classifier



```
from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier(max_depth=10)
classifier.fit(X_train, Y_train)
X_test = X_validation
Y_test = Y_validation
prediction = classifier.predict(X_test)
print (classifier.score(X_train, Y_train))
print (classifier.score(X_test, Y_test))
```

```
0.982456140351
0.815789473684
```

In [ ]:

Presentation

Proposal follows a well-organized structure and would be readily understood by its intended audience. Each section is written in a clear, concise and specific manner. Few grammatical and spelling mistakes are present. All resources used and referenced are properly cited.

In [ ]:

In [ ]: