Using Machine Learning to Predict Early Stage Dementia

Dementia

- Dementia is defined as “a decline in memory with impairment of any one of the cognitive functions such as skilled movements, memory loss, and a failure of motor cognition."
- It is an age related disorder with a 20% of estimated occurrence older people aged 85 and above.

Symptoms:

- Cognitive changes such as memory loss, difficulty in communicating or finding words, and difficulty with visual and spatial abilities.
- Psychological changes such as personality changes, behavioral disorders, Paranoia, Agitation, Hallucination.

Statistics:

- According to the Center for Disease Control and Prevention(CDC), within the United States, approximately 5.7 million people suffer from dementia and world-wide approximately 50 million people.
- By 2050 an estimated 135 million individuals worldwide will have dementia and by 2030 the cost of dementia care will increase by $1tr.

Research Problem

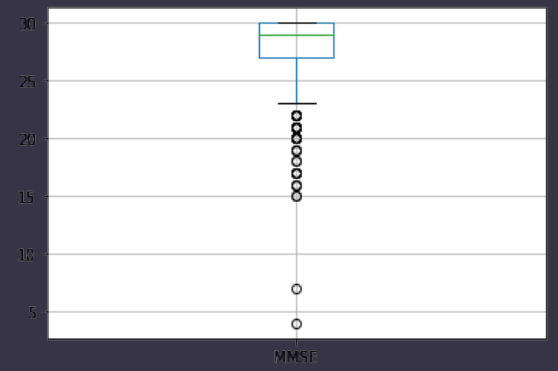
- Predicting whether a patient has dementia based on demographic, clinical and MRI data variables.
- The representative variable for predicting dementia was Clinical Dementia Rating (CDR), which is a clinically determined value.
- Python Machine learning classification tools such as Logistic regression, Decision Tree, Support Vector Machines, and Random Forest classifier were used for prediction.
- Models were assessed on their prediction score, their ability to predict CDR accurately.

Dataset

- The dataset used was MRI quantitative data for Alzheimer's disease uploaded to Kaggle from Open Access Series of Imaging Studies (OASIS).
- The dataset is a longitudinal dataset of 150 patient records ranging from 60-96 years in age.
- The dataset contains demographic information including sex, age, education level, and socioeconomic status along with clinical and MRI data including Mini-Mental State Examination (MMSE) score, Clinical Dementia Rating (CDR), estimated Total Intracranial Volume (eTIV), normalized Whole-Brain Volume (nWBV), and Atlas Scaling Factor (ASF).

Imputing missing values

- Subjects with missing values in Socioeconomic status had their missing values imputed with mean.
- Missing values in MMSE were imputed with the median values due to the data distribution so that outliers wouldn't have as much sway on the distribution(MMSE values before imputation shown below).



MMSE

- A column named *'Hand'* which refers to handedness of the patient was removed since all patients were right-handed.

Methods for Classification

Four python machine learning classification models :

1) Logistic Regression - used LogisticRegressor from the sklearn.linear_model package

2) Support Vector Machines - used support vector classifier, SVC from sklearn.svm

3) Random Forest - used RandomForestClassifier from sklearn.ensemble

4) Decision Tree - used DecisionTreeClassifier from the sklearn.tree

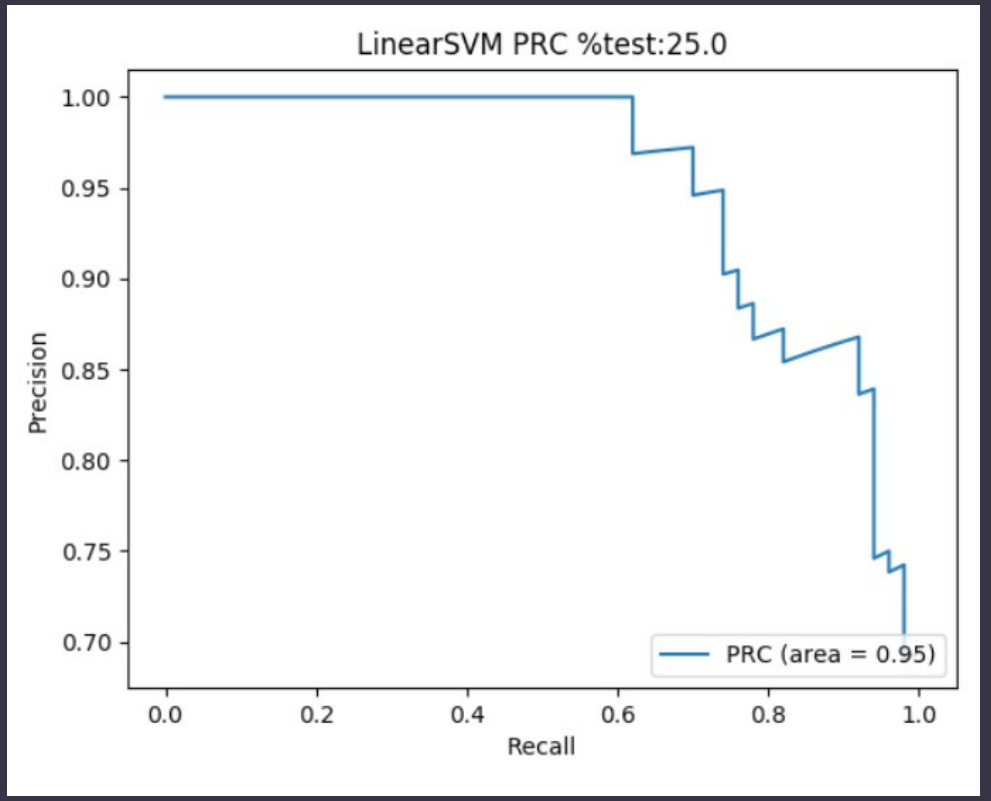
Interpretation of Results

- In order to estimate the accuracy of prediction of different machine learning models, two parameters were used.
- ACC is the percentage of correct predictions
- AUC, area under curve, of PRC, precision recall curve, which is plot between recall and precision

Example PRC Information

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$



Example of a Precision Recall Curve Plotted using Matplotlib

All Variables

Best for specific training/test set ratio
Best overall
Worst for specific training/test set ratio
Worst overall

- Accuracy of prediction (ACC) and Area Under the Precision-Recall Curve (PRC AUC) values were found for Logistic Regression, Decision Tree, SVM, and Random Forest for 50:50, 75:25, and 85:15 training/test split ratios.

For All Variables			
Machine Learning Algorithm	Training/Test Set Ratio	ACC	PRC AUC
Logistic Regression	50:50	0.802	0.9
Decision Tree	50:50	0.738	0.73
SVM	50:50	0.807	0.91
Random Forest	50:50	0.775	0.88
Logistic Regression	75:25	0.851	0.94
Decision Tree	75:25	0.777	0.75
SVM	75:25	0.851	0.95
Random Forest	75:25	0.840	0.91
Logistic Regression	85:15	0.875	0.97
Decision Tree	85:15	0.750	0.68
SVM	85:15	0.911	0.97
Random Forest	85:15	0.804	0.94

Grouped Variables

- The variables were also grouped into two separate feature lists, demographic and clinical.
- Demographic and Clinical features were used to predict CDR score to understand which features contribute more to prediction.
- The overall accuracy of the models was also assessed to determine how strong the demographic and clinical variables were as predictors for early stage dementia.

Best for specific training/test set ratio
Best overall
Worst for specific training/test set ratio
Worst overall

		For Clinical Variables		For Demographic Variables	
Machine Learning Algorithm	Training/Test Set Ratio	ACC	PRC AUC	ACC	PRC AUC
Logistic Regression	50:50	0.802	0.83	0.594	0.76
Decision Tree	50:50	0.733	0.73	0.642	0.69
SVM	50:50	0.807	0.84	0.626	0.75
Random Forest	50:50	0.765	0.85	0.658	0.76
Logistic Regression	75:25	0.83	0.89	0.638	0.76
Decision Tree	75:25	0.702	0.68	0.649	0.66
SVM	75:25	0.84	0.89	0.638	0.73
Random Forest	75:25	0.872	0.84	0.681	0.72
Logistic Regression	85:15	0.857	0.92	0.661	0.79
Decision Tree	85:15	0.821	0.77	0.643	0.66
SVM	85:15	0.839	0.94	0.696	0.76
Random Forest	85:15	0.839	0.84	0.768	0.82

Conclusion

- SVM method performed the best in all three test-train ratio's whereas Decision tree was constantly the worst performer across both ACC and AUC.
- Logistic regression also performed well and had close values in both ACC and AUC.
- In terms of training/test set ratio, Decision Tree and Random Forest performed best at a 75:25 training/test split which may indicate that increasing training set caused overtraining of the two models.
- The clinical variables were clearly better predictors for predicting CDR with ACC and AUC values relatively on target those obtained with the full set of features.