Comparison of Classifiers for Dementia Classification for an Enhanced Alzheimer's Disease Diagnosis

Ria Srivastava

1. Abstract

Recently, the healthcare industry has made significant advancements in utilizing technology/data science to enhance diagnosing diseases. With the vast amounts of data within electronic health records, genetic information, and lifestyle data available, individuals in healthcare have the opportunity to improve the accuracy and timeliness of diagnoses, influencing treatment outcomes/costs. Predictive modeling can be used as an approach to address this challenge. By analyzing electronic health records, genetic profiles, and lifestyle data, this study aims to compare different classifiers for dementia classification to further better the diagnosis of Alzheimer's disease in the future. Utilizing logistic regression, random forests, SVC. and gradient boosting algorithms on large healthcare datasets, I will compare the performance of these classifiers in classifying individuals into different dementia categories. The study will focus on developing a machine learning model that will predict an individual's likelihood of getting Alzheimer's based on their medical history, genetic profile, demographics, and lifestyle factors. Validation of the models will be conducted through cross-validation techniques and testing on independent datasets. My goal of this project is to provide an analysis comparing different classifiers for dementia classification that will offer insights into the most effective approaches to diagnose Alzheimer's. Identifying the most accurate and reliable classifiers will help healthcare professionals make better decisions regarding diagnosis and treatment, leading to earlier identification of individuals at risk of developing Alzheimer's disease and facilitating timely interventions and personalized treatment strategies.

2. Introduction

2.1 Overview of Dementia and Alzheimer's

Dementia and Alzheimers are neurodegenerative conditions that affect cognition and memory. Alzheimers is the most common cause of dementia which is a progressive brain disorder where there is a loss of brain cells and protein begins to deposit in the brain. Some symptoms of Alzheimers are difficulty remembering names, conversations, key events which slowly progresses into complete memory loss, confusion, and changes in behavior. The risk of Alzheimer's greatly increases after the age of 65, which will be explored further in this study. It is assumed that genetics, lifestyle, and the environment can greatly impact this as well, which will also be further studied in this paper. Currently, there is no cure for Alzheimer's as available treatments only slow the progression of this disease with medicines to improve cognitive function and help behavior problems. Research is still ongoing surrounding this disease, leaving scientists and researchers with many further questions.

2.2 Goals/Objectives of the Study

The overall goal of this study aims to use predictive modeling techniques to enhance the diagnosis of Alzheimers. I aim to compare different machine learning classifiers for dementia classification, specifically the performance of logistic regression, random forests, support vector classifiers, and gradient boosting algorithms. This study uses a previous study's healthcare data containing health records, genetic profiles, and lifestyle information to develop machine learning models that predicts the likelihood of an individual developing Alzheimer's disease based on various factors. I hope to find the most accurate and reliable classifiers to provide insight that can help healthcare professionals diagnose and treat individuals early. I use a longitudinal dataset

from OASIS that contains many attributes such as cognitive assessment scores and brain imaging data.

3. Methodology

3.1 Dataset and Collection

The attributes in the dataset I am using are:

1. Subject ID: Unique identifier for each subject/participant.
2. MRI ID: Unique identifier for each MRI scan.
3. Group: Classification of the subject/group, in this case, "Nondemented" indicating the absence of dementia.
4. Visit: Visit number or time point of the MRI scan.
5. MR Delay: Delay in the MRI scan (time between the previous scan and the current one).
6. M/F: Gender of the participant (M for male, F for female).
7. Hand: Handedness of the participant (R for right-handed).
8. Age: Age of the participant at the time of the MRI scan.
9. EDUC: Years of education completed by the participant.
10. SES: Socioeconomic status of the participant.
11. MMSE: Mini-Mental State Examination score, a cognitive impairment assessment.
12. CDR: Clinical Dementia Rating, a scale for characterizing cognitive and functional performance.
13. eTIV: Estimated Total Intracranial Volume, a measure of brain size.
14. nWBV: Normalized Whole Brain Volume, a normalized measure of whole brain volume.
15. ASF: Atlas Scaling Factor, used for head size normalization in morphometric analysis.

The dataset will be taken from the site OASIS. The longitudinal dataset contains 150 subjects. Some advantages of this is it follows subjects over multiple visits allowing for the analysis of disease progression. It also includes a mix of nondemented and demented subjects. Some limitations of this dataset is that it has a smaller sample size. It also has a limited number of subjects with Alzheimer's disease. This is the summary given on the OASIS site for the longitudinal dataset:

> **Summary:** 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. 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.

Description of the longitudinal dataset:

- 150 subjects were used between the ages of 60-96
- All 150 of the subjects were scanned 1-2 times
- 72 were 'nondemented' while 64 were 'demented.' 14 were initially 'demented' but then were termed 'nondemented' at a later visit. They were termed 'converted.'
- All of the 150 subjects were right handed

3.3 Data Preprocessing

Within my dataset, there were null values:

```
[2]: df.isna().sum()
```

```
[2]: Subject ID     0
     MRI ID         0
     Group          0
     Visit          0
     MR Delay       0
     M/F            0
     Hand           0
     Age            0
     EDUC           0
     SES           19
     MMSE           2
     CDR            0
     eTIV           0
     nWBV           0
     ASF            0
     dtype: int64
```

I will be replacing these null values (SES and MMSE) with the mean imputation.

```
df.isna().sum()
df['SES'].fillna(df['SES'].mean(), inplace=True)
df['MMSE'].fillna(df['MMSE'].mean(), inplace=True)
```

By performing mean imputation, I am replacing missing values with the central tendency of the respective column to help retain structure and reduce the impact of the missing values on the analysis of the data. However, this may cause bias in estimates

leading the imputed values to not accurately represent the true values of this data. It can also lead to underestimation of uncertainty or variability in the results. It also has the potential of strengthening or weakening correlations leading to incorrect interpretations.
In addition, in order to complete some of the classifiers, I had to get rid of any categorical variables and data. Since Subject ID, MRI ID, and Group were all categorical variables, they could not be used directly as features in the machine learning model. Therefore, I was able to drop subject ID and MRI IDas they don't provide any predictive value. For Group and M/F, I turned them into binary values. Either 0 or 1.

These are the outliers within the dataset:

```
Outliers:

MR Delay :  (array([ 26,  65,  69, 145, 151, 152, 253, 350], dtype=in
EDUC :  (array([], dtype=int64),)
SES :  (array([128, 129, 130, 153, 154, 171, 172], dtype=int64),)
MMSE :  (array([ 19,  20,  37,  38,  45,  46,  54,  82,  83,  84,  87
        92,  93,  94,  95,  99, 100, 130, 154, 164, 165, 176, 177, 17
       212, 215, 216, 221, 222, 224, 239, 287, 288, 304, 305, 314, 31
       347], dtype=int64),)
eTIV :  (array([  0,   1, 131], dtype=int64),)
nWBV :  (array([], dtype=int64),)
ASF :  (array([270], dtype=int64),)
```

To handle outliers, I removed them from the dataset. By removing these outliers, the model can better capture underlying patterns and relationships in the data, giving improved predictive accuracy. Outliers can also introduce noise and variability in the dataset, reducing the model's generalization ability. Removing these outliers can help create a cleaner and more representative dataset, allowing the model to focus on patterns and features. Outliers can also distort visualizations and stat summaries leading to better, interpretable results.

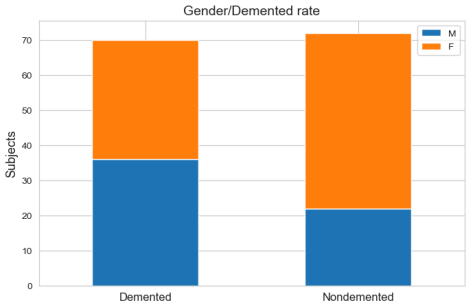
3.3 Classifier Selection

I have decided to use Random Forest Classifier, Gradient Boosting Classifier, Support Vector Machine Classifier, and Logistic Regression. Later on in the paper, I will get into the advantages and disadvantages of each.

4. Results

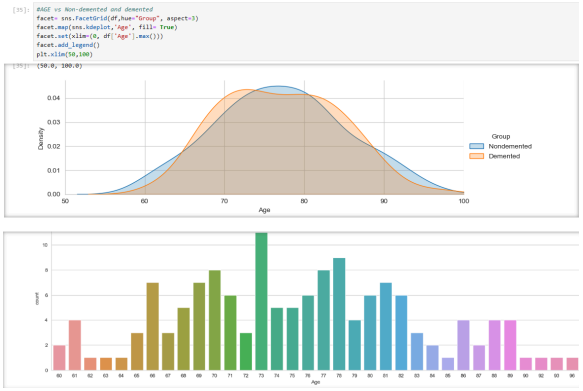
Comparing gender:

```
             M   F
Demented    36  34
Nondemented 22  50
Text(0.5, 1.0, 'Gender/Demented rate')
```



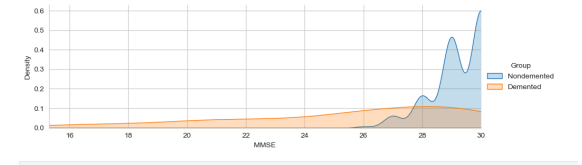
Based on the graph, the amount of Males and Females in the study who were 'demented' were around the same (36 males to 34 females). There were more females in the study who were 'nondemented' compared to 'nondemented' males (50 females to 22 males).

Comparing age:

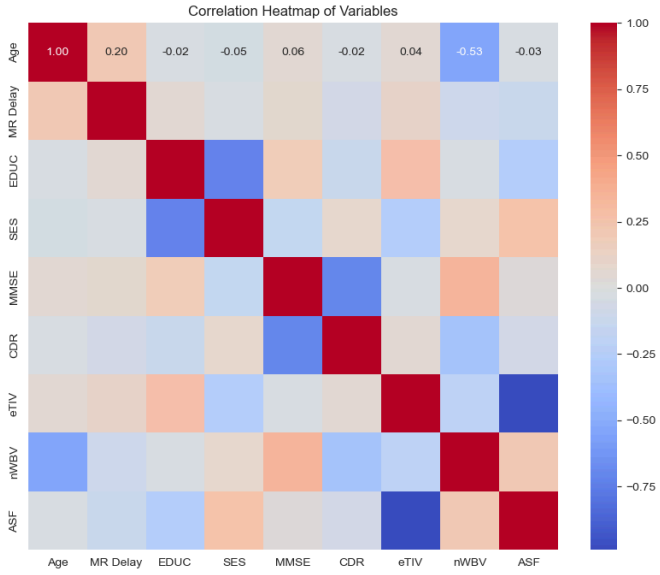


Based on these two graphs, the majority of those who are demented are 70-80 years old compared to the nondemented group. There are not many subjects < 60 or > 90.

Comparing MMSE scores:

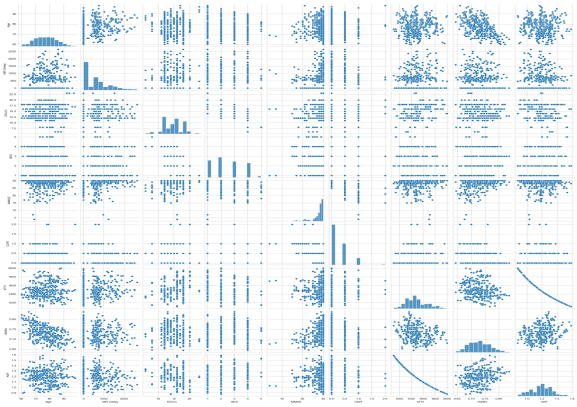


Based on the MMSE scores, the Nondemented group tended to have higher MMSE scores compared to the demented group.
Correlation Map through a heatmap:



This correlation heat map goes from -1 to 1. A correlation of 1 indicated a perfect positive correlation meaning that as one variable increases the other variable also increases (shown as red). A correlation of -1 indicated a perfect negative correlation meaning as one variable increases the other variable decreases (shown as dark blue). A correlation of 0 shows no correlation, meaning there is no linear relationship.

Pairplot to show pairwise relationship:



Summary of Longitudinal DataSet:

Summary Statistics:

	Visit	MR Delay	Age	EDUC	SES
count	373.000000	373.000000	373.000000	373.000000	354.000000
mean	1.882038	595.104558	77.013405	14.597855	2.460452
std	0.922843	635.485118	7.640957	2.876339	1.134005
min	1.000000	0.000000	60.000000	6.000000	1.000000
25%	1.000000	0.000000	71.000000	12.000000	2.000000
50%	2.000000	552.000000	77.000000	15.000000	2.000000
75%	2.000000	873.000000	82.000000	16.000000	3.000000
max	5.000000	2639.000000	98.000000	23.000000	5.000000

	MMSE	CDR	eTIV	nWBV	ASF
count	371.000000	373.000000	373.000000	373.000000	373.000000
mean	27.342318	0.290885	1488.128686	0.729568	1.195461
std	3.683244	0.374557	176.139286	0.037135	0.138092
min	4.000000	0.000000	1106.000000	0.644000	0.876000
25%	27.000000	0.000000	1357.000000	0.700000	1.099000
50%	29.000000	0.000000	1470.000000	0.729000	1.194000
75%	30.000000	0.500000	1597.000000	0.756000	1.293000
max	30.000000	2.000000	2004.000000	0.837000	1.587000

4.1 Random Forest Classifier for Longitudinal Data

A random forest classifier is based on decision trees. It combined decision trees, each on a random subset of the data and using a random subset of features. Each decision tree learns to make predictions based on the values of the input features. It can be used to predict the class labels of new data points by aggregating predictions of the individual decision trees and outputting the most commonly predicted label. Feature importance refers to the measure of the relative contribution of each feature to the prediction accuracy. It is important for understanding what features are most influential in making predictions. It calculates feature importance based on how much each feature decreases impurity across all decision trees. The high feature importance of CDR shows it is strongly correlated with the presence or absence of dementia. Changes or scores in the CDR rating are highly predictive of whether a subject is classified as having dementia or not. The severity of cognitive impairment and functional decline is a crucial factor in distinguishing between individuals or do or do not have dementia.

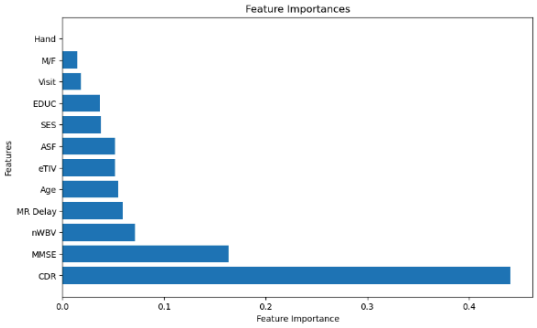
Some advantages include:

- Good performance on large datasets with high dimensionality
- Resistant to outliers and noisy data
- Automatically handles feature selection

Disadvantages:

- Computationally expensive with a large number of trees and features
- May not perform well with highly correlated features

```
Accuracy: 0.8873239436619719
```



4.2 Gradient Boosting Classifier for Longitudinal Data

The gradient boosting classifier builds a series of decision trees where each tree corrects the eros of the previous one. In classification, it combines multiple weak decision trees to create a strong classifier. It first initializes an initial estimate. Then, it iteratively fits new models to provide a more accurate estimate of the target variable. The final prediction is made by aggregating the predictions of all weak decision trees.

Advantages:

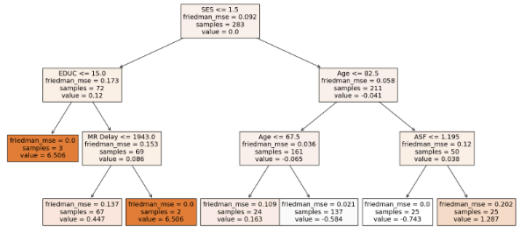
- Captures complex interactions between features
- Handles numerical and categorical data well
- Less prone to overfitting compared to Random Forest if properly tuned

Disadvantages:

- Sensitive to overfitting if number of trees is too large
- Requires careful tuning of hyperparameters to achieve optimal performance

This is the first decision tree:

```
Accuracy: 0.8732394366197183
```



4.3 Support Vector Machine Classifier for Longitudinal Data

SVM is a supervised learning algorithm that finds the hyperplane that best separates the classes in a high dimensional feature space, aiming to maximize margins between classes and minimize classification errors. It is chosen for its ability to handle high-dimensional data and nonlinear relationships between features through kernel functions. It is good for small to medium sized datasets.

Advantages:

- Good in high-dimensional space
- Memory efficient

Disadvantages:

- Not good for large datasets
- Require careful preprocessing of data, like feature scaling.

```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_squared_error
df = pd.read_csv("oasis_longitudinal.csv")
non_numeric_cols = ['Subject ID', 'MRI ID', 'Group', 'Visit', 'M/F', 'Hand']
df_numeric = df.drop(columns=non_numeric_cols)
df_numeric.dropna(inplace=True)
X = df_numeric.drop(['CDR'], axis=1)
y = df_numeric['CDR']
y_binary = y.apply(lambda x: 1 if x > 0 else 0)
X_train, X_test, y_train, y_test = train_test_split(X, y_binary, test_size=0.2, random_state=42)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
svm = SVC(kernel='linear', random_state=42)
svm.fit(X_train_scaled, y_train)
y_pred = svm.predict(X_test_scaled)
mse = mean_squared_error(y_test, y_pred)
print("Mean Squared Error:", mse)
```

```
Mean Squared Error: 0.15492957746478872
```

4.4 Logistic Regression for Longitudinal Data

Logistic Regression is a statistical method used for binary classification. It models the probability of the occurrence of a binary outcome (like the presence or absence of dementia) based on one or more predictor variables. It is chosen as a classifier because it is simple and efficient. However, it is limited in its ability to capture complex feature relationships and assume linearity between the independent variables.

Advantages:

- Simple and Interpretable making it more easy to understand the impact of each predictor variable in the outcome
- Performs well when the relationship between features and outcome is linear

Disadvantages:

- Limited in capturing complex relationships
- Assumes linearity
- Not good for high dimensional data or nonlinear relationships

```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
df = pd.read_csv("oasis_longitudinal.csv")
non_numeric_cols = ['Subject ID', 'MRI ID', 'Group', 'Visit', 'M/F', 'Hand']
df_numeric = df.drop(columns=non_numeric_cols)
df_numeric.dropna(inplace=True)
X = df_numeric.drop(['CDR'], axis=1)
y = df_numeric['CDR']
y_binary = y.apply(lambda x: 1 if x > 0 else 0)
X_train, X_test, y_train, y_test = train_test_split(X, y_binary, test_size=0.2, random_state=42)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
logreg = LogisticRegression(random_state=42)
logreg.fit(X_train_scaled, y_train)
y_pred = logreg.predict(X_test_scaled)
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
print("\nClassification Report:")
print(classification_report(y_test, y_pred))
print("\nConfusion Matrix:")
print(confusion_matrix(y_test, y_pred))
```

5. Discussion

5.1 Accuracy of each model (ROC, Accuracy, and/or Confusion Matrix)

Random Forest:

For the random forest classifier, as shown in the graph, the accuracy was 0.8873239. This is the corresponding confusion matrix:

```
Confusion Matrix:
[[ 2  2  4]
 [ 1 25  0]
 [ 0  1 36]]

Classification Report:
              precision    recall  f1-score   support

           0       0.67      0.25      0.36         8
           1       0.89      0.96      0.93        26
           2       0.90      0.97      0.94        37

    accuracy                           0.89        71
   macro avg       0.82      0.73      0.74        71
weighted avg       0.87      0.89      0.87        71
```

ROC curve cannot be applied here as it is typically used for binary classification where there are only two classes.

Gradient Boosting Classifier:

As shown in the graph, the accuracy is 0.873239436. This is the corresponding confusion matrix:

```
Confusion Matrix:
[[ 3  2  3]
 [ 2 24  0]
 [ 1  1 35]]

Classification Report:
              precision    recall  f1-score   support

   Converted       0.50      0.38      0.43         8
    Demented       0.89      0.92      0.91        26
 Nondemented       0.92      0.95      0.93        37

    accuracy                           0.87        71
   macro avg       0.77      0.75      0.76        71
weighted avg       0.86      0.87      0.87        71
```

SVR (Support Vector Regression):

In order to generate this classification report, I had to convert the continuous target variable (CDR) into a categorical one and treat it as a classification problem rather than a regression problem. The accuracy is 0.8450704225352113

```
Support Vector Machine Accuracy: 0.8450704225352113

Support Vector Machine Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.93      0.87        40
           1       0.88      0.74      0.81        31

    accuracy                           0.85        71
   macro avg       0.85      0.83      0.84        71
weighted avg       0.85      0.85      0.84        71


Support Vector Machine Confusion Matrix:
[[37  3]
 [ 8 23]]
```

Logistic Regression:

The accuracy is 0.802816901

```
Accuracy: 0.8028169014084507

Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.88      0.83        40
           1       0.81      0.71      0.76        31

    accuracy                           0.80        71
   macro avg       0.81      0.79      0.80        71
weighted avg       0.80      0.80      0.80        71


Confusion Matrix:
[[35  5]
 [ 9 22]]
```

Comparing Accuracies:

Training Random Forest...
Random Forest Accuracy: 0.8169014084507042
Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.80	0.90	0.85	40
1	0.85	0.71	0.77	31
accuracy			0.82	71
macro avg	0.82	0.80	0.81	71
weighted avg	0.82	0.82	0.81	71

Random Forest Confusion Matrix:
[[36 4]
[9 22]]

Training Gradient Boosting...
Gradient Boosting Accuracy: 0.8028169014084507
Gradient Boosting Classification Report:

	precision	recall	f1-score	support
0	0.81	0.85	0.83	40
1	0.79	0.74	0.77	31
accuracy			0.80	71
macro avg	0.80	0.80	0.80	71
weighted avg	0.80	0.80	0.80	71

Gradient Boosting Confusion Matrix:
[[34 6]
[8 23]]

Training Support Vector Machine...
Support Vector Machine Accuracy: 0.8450704225352113
Support Vector Machine Classification Report:

	precision	recall	f1-score	support
0	0.82	0.93	0.87	40
1	0.88	0.74	0.81	31
accuracy			0.85	71
macro avg	0.85	0.83	0.84	71
weighted avg	0.85	0.85	0.84	71

Support Vector Machine Confusion Matrix:
[[37 3]
[8 23]]

Training Logistic Regression...
Logistic Regression Accuracy: 0.8028169014084507
Logistic Regression Classification Report:

	precision	recall	f1-score	support
0	0.80	0.88	0.83	40
1	0.81	0.71	0.76	31
accuracy			0.80	71
macro avg	0.81	0.79	0.80	71

weighted avg 0.80 0.80 0.80 71

Logistic Regression Confusion Matrix:
[[35 5]
[9 22]]

5.2 Discussion of Results

Random Forest:

Accuracy: The random forest classifier achieved an accuracy of approximately 0.887 meaning it correctly classified individuals into dementia categories 88.7% of the time.

This high accuracy suggests that this classifier effectively captured the patterns in the dataset regarding dementia classification.

Gradient Boosting:

Accuracy: The gradient boosting classifier achieved an accuracy of about 0.873, meaning it was able to classify individuals with high accuracy.

Gradient boosting classifiers sequentially improve predictive performance by focusing on misclassified instances which leads to improved accuracy to individual decision trees.

Support Vector Machine:

Accuracy: The support vector achieved an accuracy of about 0.845. SVM classifiers are good at handling high dimensional data and are good when data is not linearly separable. This had good accuracy, though not as well as Random forest or Gradient boosting. The mean squared error was 0.1549 meaning that while misclassifications could occur, the model's predictions were generally close to the true values.

Logistic Regression:

Accuracy: The accuracy the logistic regression model achieved was about 0.803. It is often used as a baseline model for binary classification tasks. It is comparatively low to the other classifier, though it is still good for dementia classification

The accuracies of the classifiers were somewhat close with Random Forest performing better than the others. However, all of the accuracies were above 0.80 meaning they all did very well in classifying and predicting dementia in subjects.

In addition, Random Forest and Gradient Boosting show similar precision, recall, and F1 scores for both dementia and non dementia patients. SVM shows higher precision but lower recall compared to the other classifiers. Logistic Regression performs well but has lower precision and recall compared to random forest and gradient boosting classifiers. Overall, all classifiers show reasonable performance in dementia classifications with slight variations in precision, recall, and F1 scores.

Based on this information, Random Forest Classifier appears to be the best choice for this project. It achieved the highest accuracy at 0.887 with balanced precision, recall and F1-scores for dementia and non dementia patients. It effectively classified both classes without favoring one. Random forest, in general, is able to handle large datasets with high dimensional accuracy, and is very resistant to outliers.

5.3 Limitations and Assumptions

One crucial aspect of data science is finding places that I could have gone wrong in the study. Specifically, in my preprocessing step, where there could still be underlying data quality issues that were not well addressed. Inaccuracies in recorded data, missing variables, and biases could have impacted the performance of classifiers.

In addition, some limitations were the small sample size. The dataset only contained 150 subjects which limits the generalizability of my findings. A larger sample size is more representative of the population.

There were also a limited number of subjects with Alzheimer's disease. This had the potential of affecting the performance of classifiers, especially in accurately predicting this specific condition. This can

lead to an imbalance in class distribution leading to biased results.
The study also assumed the dataset is homogenous and all subjects follow a similar pattern of disease progression and risk factors. It did not take into account lifestyle factors or genetic predispositions. In addition, while mean imputation is common, it may introduce bias and affect data distribution. Imputed values do not accurately represent the true values, which potentially impact the performance of classifiers. Dropping categorical variables and converting categorical variables to binary values may oversimplify the data and discard potentially valuable information, which can lead to a loss of accurate predictive powers by the classifiers. Removing outliers to reduce noise could have possibly removed important information in the dataset. Outliers represent real data points with extreme values or anomalies that are clinically significant.

6. References

Open Access Series of Imaging Studies (OASIS), sites.wustl.edu/oasisbrains/#about. Accessed 24 Apr. 2024.

What Is Dementia? Symptoms, Types, and Diagnosis | National Institute on Aging, www.nia.nih.gov/health/alzheimers-and-dementia/what-dementia-symptoms-types-and-diagnosis. Accessed 24 Apr. 2024.

"Alzheimer's Disease and Healthy Aging Data Portal." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 9 Sept. 2019, www.cdc.gov/aging/agingdata/index.html.

"What Is Alzheimer's Disease?" *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 26 Oct. 2020, www.cdc.gov/aging/aginginfo/alzheimers.htm.

"Machine Learning Classifiers - the Algorithms & How They Work." *MonkeyLearn Blog*, 14 Dec. 2020, monkeylearn.com/blog/what-is-a-classifier/.