**Predicting the Risk of Dementia**

Garth Scheck

Bellevue University

DSC 630: Predictive Analytics

Andrew Hua

May 27, 2024

**Introduction**

Dementia, specifically Alzheimer's, cases are on the rise and expected to triple in the next twenty-five years. It is also now the seventh leading cause of death around the world.

Finding an answer to the cause and prevention of dementia is critical as the population grows because the cost of care for dementia patients is about one and a half times that of a normally aging patient which is a burden to both the families of the patients and insurance companies. This will also lead to higher costs for health care and insurance. Because of this, long-term healthcare facilities and emergency rooms would be interested in finding solutions to dementia.

This project explores machine learning as a viable approach to determining what attributes lead to dementia, and if it can be slowed down or prevented.

**Data**

The data used in this analysis is from the National Alzheimer’s Coordinating Center (NACC). NACC hosts the centralized data repository for thirty-three research centers across the United States. The data repository contains 1024 attributes and 185,831 observations. The unit of analysis is a patient, and each row represents a visit to a research facility.

The dictionary containing descriptions of the attributes can be found in the link below:

[Researchers Data Dictionary - Uniform Data Set (UDS)](https://files.alz.washington.edu/documentation/uds3-rdd.pdf)

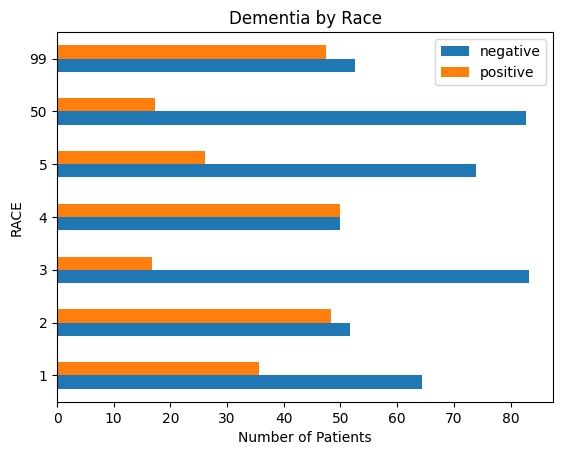
Access to the data requires approval from NACC and must be downloaded within two weeks of the request. Access can be requested through the link below:

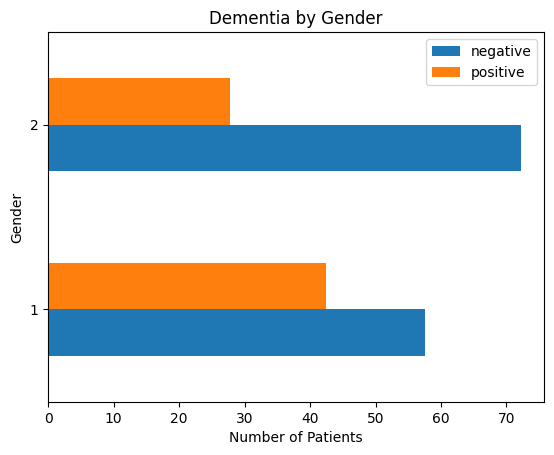
<https://naccdata.org>

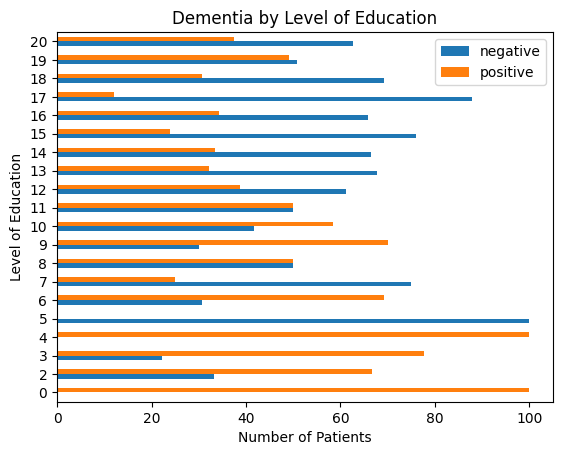
**Methods and Models**

Analyzing the data through visualizations provided a clearer understanding of the data, such as which demographics are most impacted by dementia. These can be seen in the charts below.

**A white background with black text

Description automatically generated** 

**** 



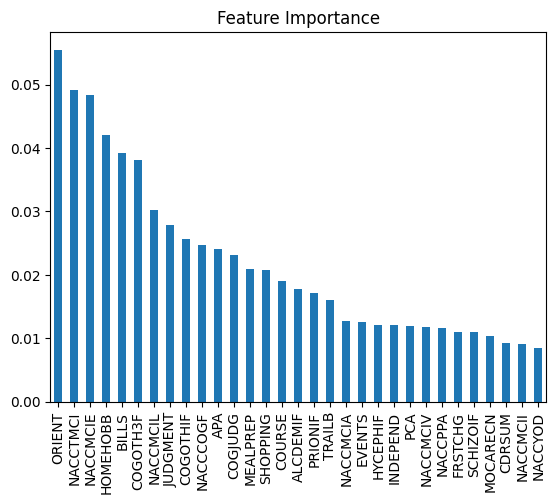
As can be seen in the charts, dementia affects Pacific Islanders, African Americans, men, and individuals with less education the most. The line graph below shows the number of dementia cases over the years compared to the total number of patients. The number of cases is pretty consistent and actually decreases some years as in 2007 and 2022. From this, it would appear there should be no concern, but an article from the Population Reference Bureau (Fact Sheet: U.S. Dementia Trends, 2021), and others, have predicted that cases will triple in 25 years. This is due to the large baby boomer population aging.

A graph showing the number of patients with dementia

Description automatically generated

Upon completing the initial analysis, a more in-depth investigation was conducted to ascertain whether machine learning could effectively identify the key attributes contributing to dementia, as mentioned in the introduction. To achieve that goal, classification algorithms were selected as the analytical approach. Data scrubbing was essential to prepare the data for the machine learning algorithms since it had not been previously cleaned. Data cleansing entailed the removal of null values through different means and removing several columns that contained dementia diagnoses; these columns would have led to overfitting. Great care had to be taken when deciding how to handle missing values. Some were replaced with hard-coded values representing values in an enumeration while others were replaced with an average, such as weight and height. Data cleansing took a majority of the time on the project.

Grid search was also utilized to tune the Random Forest Classifier model parameters and decrease the number of attributes further, simplifying the machine learning model. The grid search reduced the remaining number of attributes by more than half. The chart below displays the features resulting from the grid search contributing most to model prediction as indicated by the "Feature Importance" chart. Many of the top features had to do with measuring cognitive abilities.

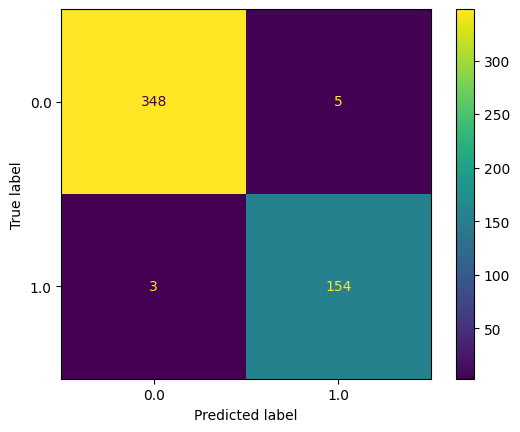


Four classification algorithms were selected to evaluate the performance and determine the most suitable model shown in the results.

**Results**

Accuracy was used to measure the performance of the models since they were all classification algorithms with the target variable being “demented”. All the models selected performed well in predicting dementia, as seen in the table below. The Random Forest Classifier achieved the highest performance with an accuracy of 98.43%, which I attribute to the fine-tuning of the model.

|  |  |  |
| --- | --- | --- |
| **Model** | **Training**  **Time** | **Accuracy %** |
| Random Forest Classifier | 6 minutes | 98.43 |
| Linear Discriminant Analysis | 44 seconds | 88.24 |
| Logistic Regression | 1 seconds | 85.70 |
| K Neighbors Classifier | 9 second | 78.43 |

The confusion matrix from the Random Forest Classifier below shows that the model produced 5 false positives and 3 false negatives out of 510 predictions. Taking the number of correct predictions divided by the total number of predictions gives us an accuracy of 98.43%. The implications of this will be discussed in the conclusion.

Confusion Matrix

**Conclusion and Recommendations**

Machine learning proved to be a viable tool for aiding in the prediction of dementia given a set of attributes. Caution should be taken when using the predictions and additional tests should be conducted to verify their accuracy, because false results may occur, as evidenced by the confusion matrix.

The analysis also showed that certain demographics are more prone to developing dementia, such as Pacific Islanders, African Americans, males, and individuals with less education. The measurements used for the attributes however seem to be more directed at detecting dementia, rather than predicting risk therefore my recommendation is that there should be more data collection regarding lifestyle such as diet, stress, smoking, etc. The NACC data is also limited to patients in America which may not represent the whole picture, therefore it would be beneficial to collect data from all around the world.

**References**

News, B. (2022, January 7). Dementia cases expected to almost triple across the world by 2050. BBC News; BBC News. <https://www.bbc.com/news/health-59896819>

How Much Does Memory Care Cost? (n.d.). National Council on Aging. <https://www.ncoa.org/adviser/local-care/memory-care-costs/>

‌Dementia becomes an emergency 1.4 million times a year. (n.d.). Ihpi.umich.edu. Retrieved March 22, 2024, from <https://ihpi.umich.edu/news/dementia-becomes-emergency-14-million-times-year#:~:text=A%20busy%2C%20crowded%2C%20confusing%20emergency>

Dementia and hospital overcrowding. (n.d.). The Irish Times. Retrieved March 22, 2024, from <https://www.irishtimes.com/opinion/letters/dementia-and-hospital-overcrowding-1.2068746>

National Alzheimer’s Coordinating Center | National Alzheimer’s Coordinating Center. (n.d.). Naccdata.org. <https://naccdata.org/>

Tembo, I. (2023, February 10). Cross-Validation using K-Fold with Scikit-Learn - Isheunesu Tembo - Medium. Medium. <https://isheunesu48.medium.com/cross-validation-using-k-fold-with-scikit-learn-cfc44bf1ce6>

Population Reference Bureau. (2021, October 21). Fact Sheet: U.S. Dementia Trends. PRB. <https://www.prb.org/resources/fact-sheet-u-s-dementia-trends/>