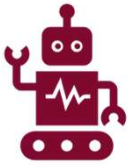
Alzheimer Classification Prediction

Laura Won
CS 677 Data Science with Python | Spring 2024

Predictive Modeling for Alzheimer's Classification:

Comparative Analysis of Machine Learning Algorithms and Ensemble Techniques.

Alzheimer’s disease is neurodegenerative disorder that affects memory and behavior.



I will use **heat-maps, logistic regression, decision trees, k-NN, and random forests** to predict Alzheimer’s disease and identify the features that most significantly affect its development.

I will **compare the accuracy and confusion matrices** of these models to determine the best classification method for this task.

Local Setup
Installation

- To follow this project, please install the following locally:
  - Pandas:
    - pip install pandas
  - Scikit- learn:
    - pip install scikit-learn
  - Matplotlib:
    - pip install matplotlib
  - Seaborn:
    - pip install seaborn
  - NumPy:
    - pip install numpy

Dataset

- Features:
  - Age
  - Years of education (EDUC)
  - Socioeconomic status (SES)
  - Clinical Dementia Rating (CDR)
    - 0-none
    - 0.5- possible
    - 1- positive
  - Mini-Mental State Examination (MMSE)

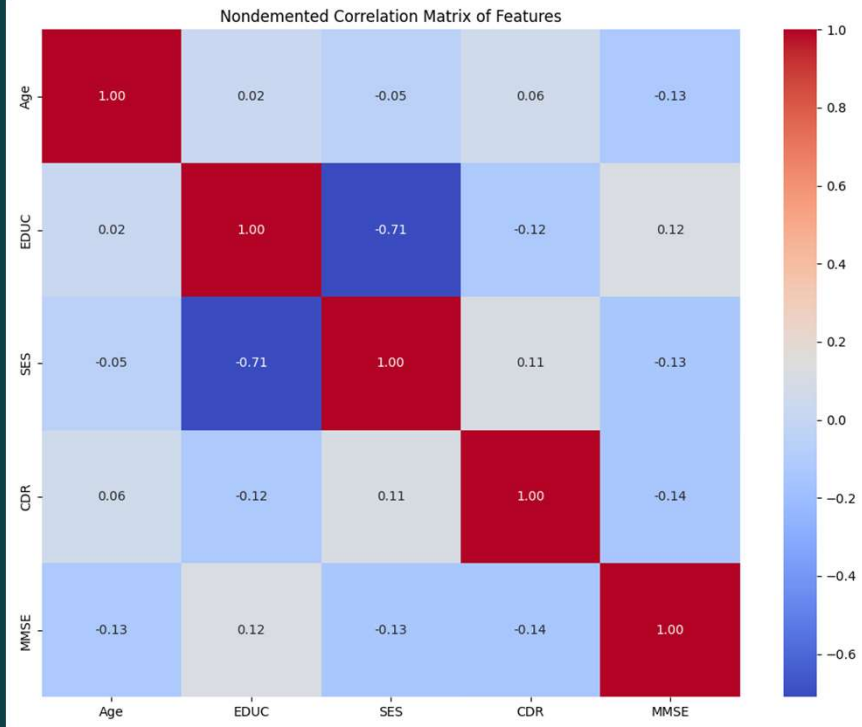
- Labels:
  - Group:
    - Nondemented (0: negative)
    - Converted (1: positive)
    - Demented (1: positive)

Methodology

1. Load data (Alzheimer.csv)
2. Data Preprocessing:
    - Combine group labels
        - Nondemented-0, Converted -0.5, Demented 1
        - Nondemented-0, Converted & Demented – 1
    - Random 50/50 splits for training and testing data
3. Visualization
    - Heatmap, Logistic Regression, Decision Trees, k-NN , Random Forest
4. Error rates in Random Forest using two hyperparameter
    - n_estimators(n) and max_depth(d)
5. Confusion matrix and Accuracy

Results

- Heatmap



Correlation matrix of features

0- Nondemented : 0.02 Age & EDUC

0.5- Converted : 0.00 MMSE & EDUC

1- Demented: 0.03 Age & CDR, SES & MMSE

Age, education, and mental state are the main factors

Results

Logistic Regression

Accuracy:88.70%

TP:152

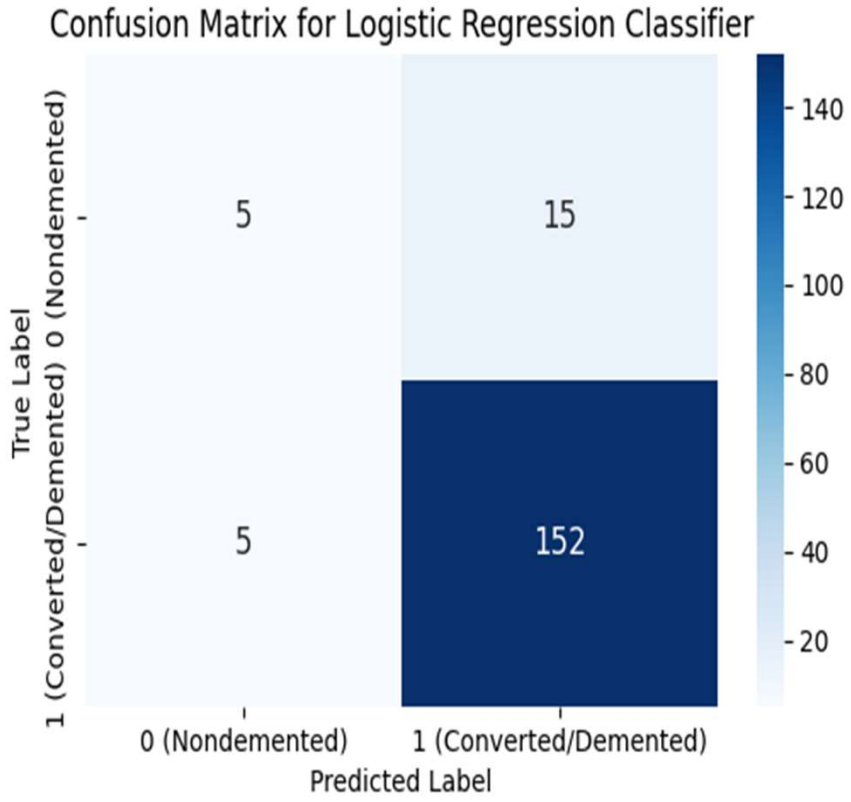
FP:15

TN:5

FN:5

TPR:96.82%

TNR:25.00%



Results

Decision Tree

Accuracy:88.14%

TP:71

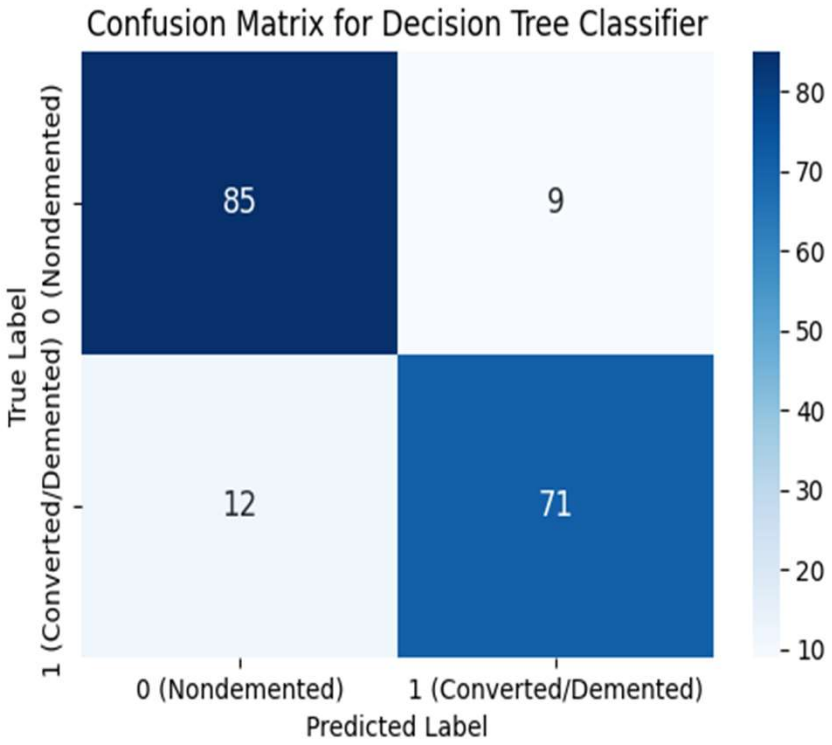
FP:9

TN:85

FN:12

TPR:85.54%

TNR:90.43%



Results

Random Forest

Accuracy:93.22%

TP:74

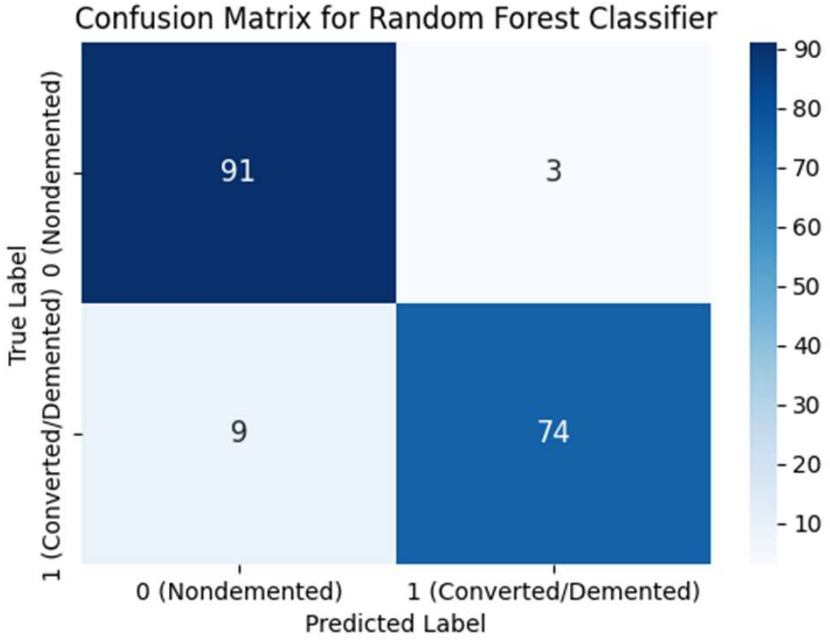
FP:3

TN:91

FN:9

TPR:89.16%

TNR:96.81%

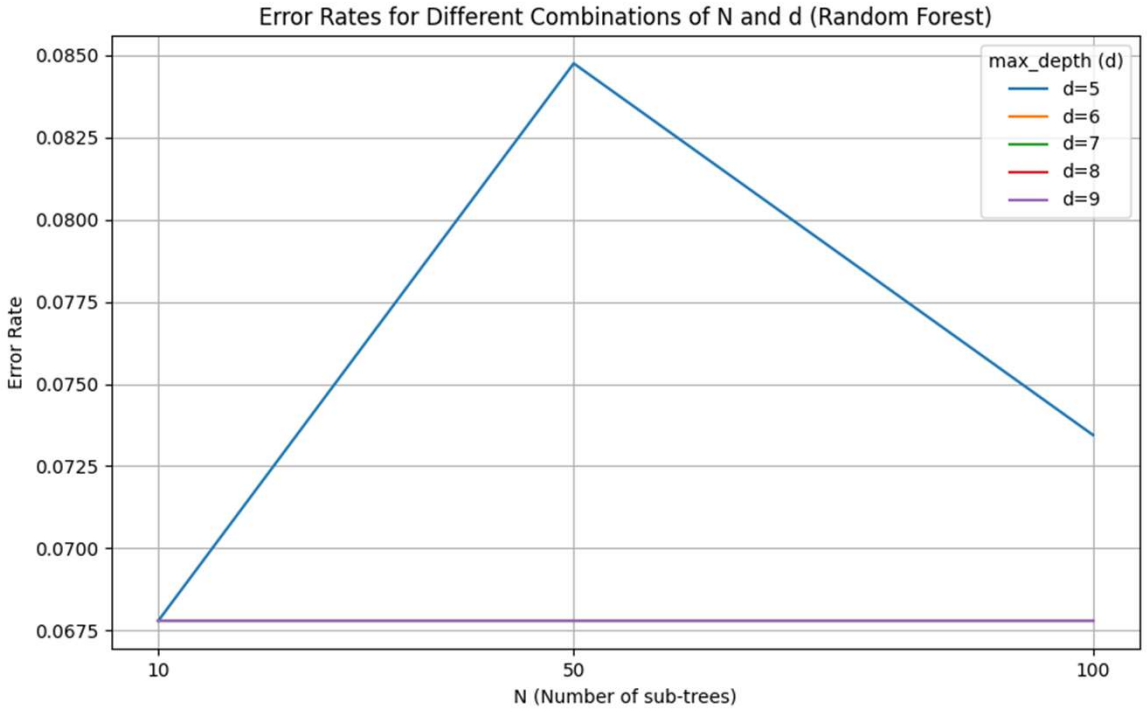


Results

Error Rate (Random Forest)

The best n is 10 and d is 5

With an error rate of 6.78%



Results

k-NN

Accuracy:76.84%

TP:47

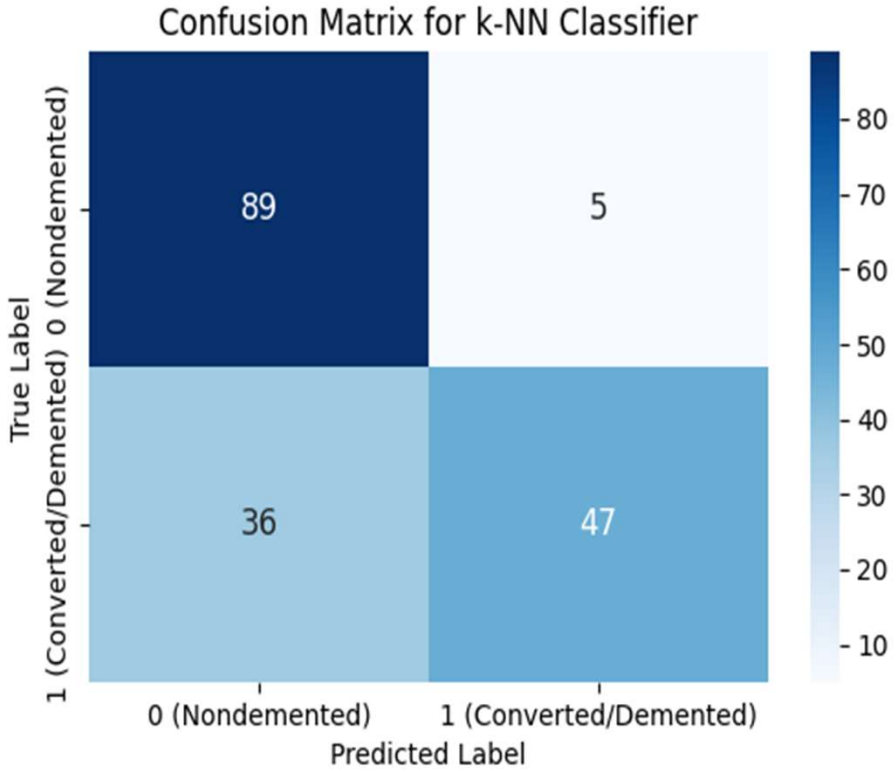
FP:5

TN:89

FN:36

TPR:56.63%

TNR:94.68%



Summary Confusion Matrix

Model	TP	FP	TN	FN	Accuracy	TRP	TNR
Logistic Regression	152	15	5	5	88.70%	96.82%	25.00%
Decision Tree	71	9	85	12	88.14%	85.54%	90.43%
Random Forest	74	3	91	9	93.22%	89.16%	96.81%
k-NN	47	5	89	36	76.84%	56.63%	94.68%

Conclusion

Early-stage detection is critical in Alzheimer's disease, and based on this project, I discovered that the converted stage correlates with mental status and years of education. I believe that prolonged exposure to intense educational environments may lead to higher levels of mental stress, potentially contributing to the disease. Demented status, age, clinical dementia rating, and socioeconomic status were also found to be significant factors. It is noteworthy that Alzheimer's patients are mostly found in older age groups. Moreover, the combination of socioeconomic status and mental state is intriguing, suggesting that lower socioeconomic status may exacerbate mental stress levels, potentially contributing to dementia.

Citation

Dataset: https://www.kaggle.com/datasets/brsdincer/alzheimer-features

Algorithms: Refer to course assignments weeks 4, 5, and 6