Alzheimer's Predictors

Stat 451

Noah Kornfeld, Angie Ohaeri, Melody Pak, Olivia Pelzek

Our Questions:

How can we best predict the presence of Alzheimer’s in a patient?

Which category of factors is most accurate when predicting Alzheimer's disease in a patient?

Our Dataset

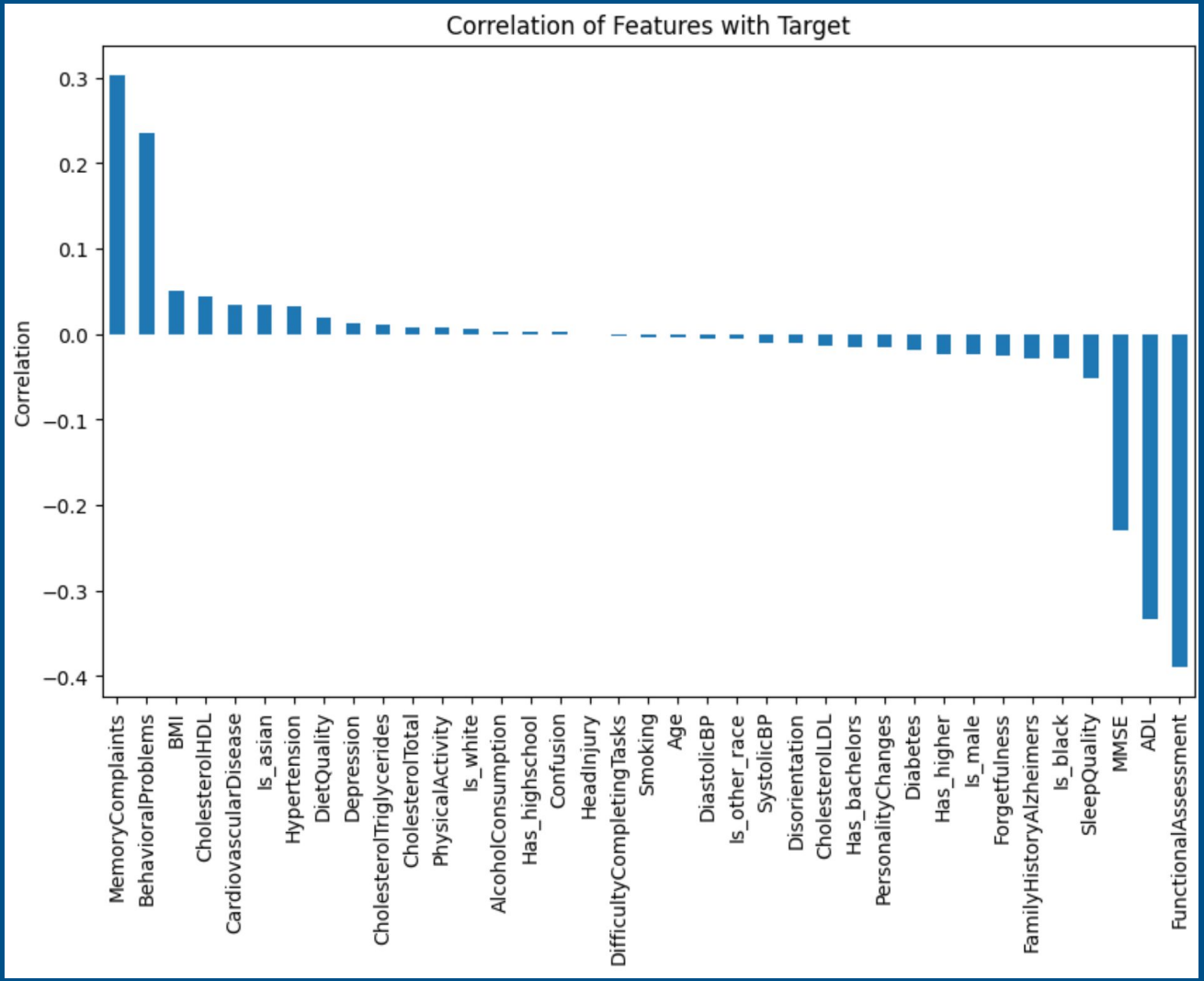
- Uploaded to Kaggle by Rabie El Kharoua
- Contains extensive health data of 2,149 unique patients and their diagnosis
- 35 variables in this dataset

- Six categories:
  - Demographic Details
  - Lifestyle Factors
  - Medical History
  - Clinical Measurements
  - Cognitive and Functional Assessments
  - Symptoms

Feature Engineering

1. No Missing Features
2. Gender, Ethnicity, Education Level need to be manipulated
   a. **Gender** ⇒ *Is_male* indicator feature
   b. **Ethnicity** ⇒ *Is_white*, *Is_black*, *Is_asian*, *Is_other_race* indicator features
   c. **Education Level** ⇒ *Has_highschool*, *Has_bachelors*, *Has_higher* indicator features
3. Dropped PatientID and DoctorInCharger Columns

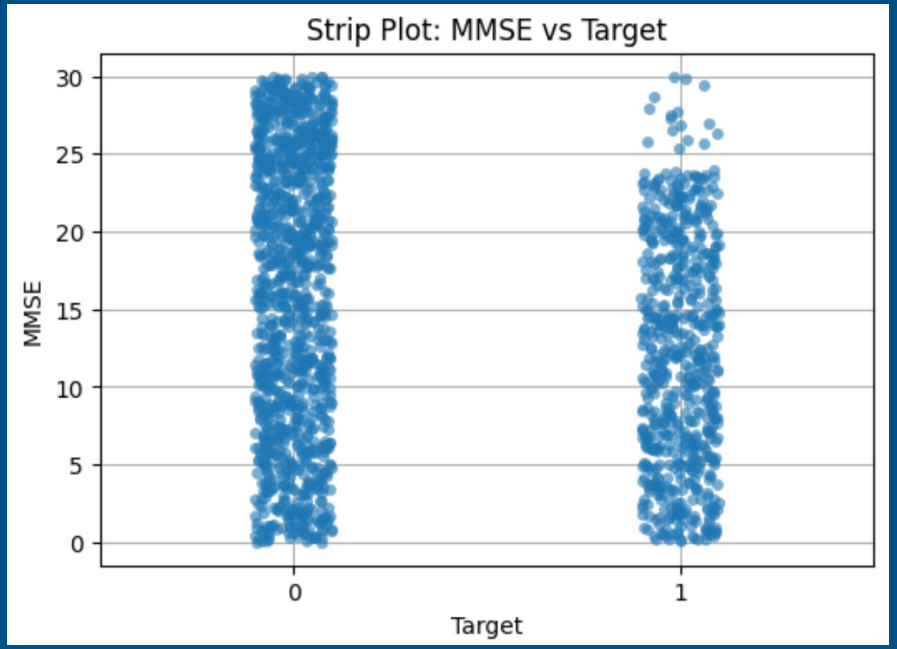
Exploring the Data

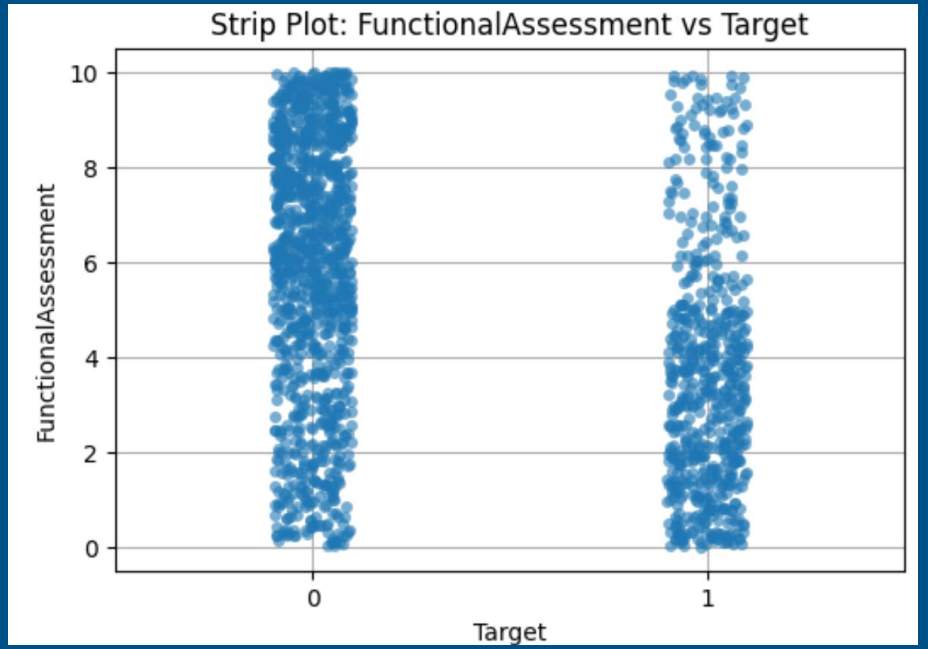


Exploring the Data

	Age	Is_male	BMI	Smoking	AlcoholConsumption	PhysicalActivity	DietQuality	SleepQuality	FamilyHistoryAlzheimers
count	2149.000000	2149.000000	2149.000000	2149.000000	2149.000000	2149.000000	2149.000000	2149.000000	2149.000000
mean	74.908795	0.506282	27.655697	0.288506	10.039442	4.920202	4.993138	7.051081	0.252210
std	8.990221	0.500077	7.217438	0.453173	5.757910	2.857191	2.909055	1.763573	0.434382
min	60.000000	0.000000	15.008851	0.000000	0.002003	0.003616	0.009385	4.002629	0.000000
25%	67.000000	0.000000	21.611408	0.000000	5.139810	2.570626	2.458455	5.482997	0.000000
50%	75.000000	1.000000	27.823924	0.000000	9.934412	4.766424	5.076087	7.115646	0.000000
75%	83.000000	1.000000	33.869778	1.000000	15.157931	7.427899	7.558625	8.562521	1.000000
max	90.000000	1.000000	39.992767	1.000000	19.989293	9.987429	9.998346	9.999840	1.000000

Exploring the Data





MMSE: Mini-Mental State Examination

Functional Assessment: “Functional assessment score. Lower scores indicate greater impairment”

Both included in “assessment factors” category

Split data into 80% training, 10% validation, and 10% testing

Feature Selection

1. Select K Best:
    a. k = 10
    b. Scoring: f regression
2. Variance Threshold:
    a. Threshold = 0.1

Features Included in Both

1. Memory Complaints
2. Behavioral Problems
3. MMSE
4. Functional Assessment
5. Sleep Quality
6. Cholesterol HDL
7. ADL
8. Cardiovascular Disease
9. BMI

Category Specific Splits

1. Select Specific Columns based on Category
2. Find best classification model for each category
3. Compare scores to find best category-model pair

```
demographic_cols = ['Age', 'Is_male','Is_white', 'Is_black', 'Is_asian',
                    'Is_other_race', 'Has_highschool', 'Has_bachelors',
                    'Has_higher' ]

lifestyle_cols = ['BMI', 'Smoking', 'AlcoholConsumption',
                  'PhysicalActivity', 'DietQuality', 'SleepQuality']

medical_cols = ['FamilyHistoryAlzheimers', 'CardiovascularDisease',
                'Diabetes','Depression', 'HeadInjury', 'Hypertension']

clinical_cols = ['SystolicBP', 'DiastolicBP','CholesterolTotal',
                 'CholesterolLDL', 'CholesterolHDL','CholesterolTriglycerides']

assessment_cols = ['MMSE', 'FunctionalAssessment','MemoryComplaints',
                   'BehavioralProblems', 'ADL']

symptoms_cols = ['Confusion','Disorientation', 'PersonalityChanges',
                 'DifficultyCompletingTasks','Forgetfulness']
```

Model Selection: Grid Search

The Model Choices

1. **Random Forest:**
   a. n_estimators: 50, 100, 200
   b. max_depth: 10, 20, None
2. **Support Vector Machine:**
   a. C: 0.1, 1, 10
   b. Kernel: linear, rbf
3. **k-Nearest Neighbors**
   a. n_neighbors: 5, 7, 9
   b. weights: uniform, distance
4. **Logistic Regression**
   a. C: 0.001, 0.01, 0.1, 1
5. **Decision Tree**
   a. max_depth: 3, 5, 10, None
   b. min_samples_split: 2, 5, 10

Model Selection and Evaluation

1. **Run Grid Search Method on following data-subsets:**
   a. All features, Variance Threshold, KBest, Feature Selection Combined, each category
2. **Method returns data frame with best model (for each model choice)**
   a. “Best” determined by accuracy and recall
   b. Maximize correctly classifying presence of Alzheimer’s (True Positive)
   c. Minimize incorrectly classifying lack of Alzheimer’s (False Negative)
3. **Compare best overall method for each data-subset**
   a. Determine best way to predict the presence of Alzheimer’s
4. **Compare category models**
   a. Determine most “predictive” patient category

Results

Features	Best Model	Validation Performance	Testing Performance
All Features	**Decision Tree Classifier** *max_depth = 20* *min_sample_split = 2*	Accuracy: 0.944 Recall: 0.895 Precision: 0.944	Accuracy: 0.944 Recall: 0.921 Precision: 0.921
Variance Threshold	**Decision Tree Classifier** *max_depth = 20* *min_sample_split = 5*	Accuracy: 0.953 Recall: 0.895 Precision: 0.971	Accuracy: 0.944 Recall: 0.921 Precision: 0.921
K Best	**Decision Tree Classifier** *max_depth = 20* *min_sample_split = 5*	Accuracy: 0.958 Recall: 0.908 Precision: 0.971	Accuracy: 0.944 Recall: 0.921 Precision: 0.921
Combined Feature Selection	**Decision Tree Classifier** *max_depth = 20* *min_sample_split = 5*	Accuracy: 0.958 Recall: 0.908 Precision: 0.971	Accuracy: 0.944 Recall: 0.921 Precision: 0.921
Demographic	**Support Vector Machine** *C = 0.1* *Kernel: Linear*	Accuracy: 0.567 Recall: 0.539 Precision: 0.414	Accuracy: 0.647 Recall: 0.013 Precision: 0.500
Lifestyle	**Support Vector Machine** *C = 0.1* *Kernel: Linear*	Accuracy: 0.507 Recall: 0.605 Precision: 0.377	Accuracy: 0.647 Recall: 0.000 Precision: 0.000
Medical	**Support Vector Machine** *C = 0.1* *Kernel: Linear*	Accuracy: 0.442 Recall: 0.895 Precision: 0.378	Accuracy: 0.353 Recall: 1.000 Precision: 0.250
Clinical	**Logistic Regression** *C = 0.01*	Accuracy: 0.526 Recall: 0.592 Precision: 0.388	Accuracy: 0.521 Recall: 0.474 Precision: 0.364
Assessment	**Decision Tree Classifier** *max_depth = 20* *min_sample_split = 2 or 5*	Accuracy: 0.958 Recall: 0.907 Precision: 0.971	Accuracy: 0.944 Recall: 0.921 Precision: 0.921
Symptoms	**Support Vector Machine** *C = 0.1* *Kernel: Linear*	Accuracy: 0.428 Recall: 0.684 Precision: 0.344	Accuracy: 0.353 Recall: 0.645 Precision: 0.204

Best Models:

1. Decision tree with max depth = 20 and min sample split = 2 or 5 produced highest recall and accuracy for validation and testing datasets for:
   - All Features
   - Variance Threshold Features
   - K (10) Best F Regression Features
   - Combined Feature Selection Features
   - Assessment Features

**(combined features: features chosen by select K best and variance threshold)

Performance on Testing Data – Accuracy, recall, precision

	Training Scenario	Classifier	Test Accuracy	Test Recall	Test Precision
1	All Features	Decision Tree	0.944186	0.921053	0.921053
2	All Features	Random Forest	0.934884	0.881579	0.930556
6	Variance Threshold	Decision Tree	0.944186	0.921053	0.921053
7	Variance Threshold	Random Forest	0.925581	0.855263	0.928571
11	SelectKBest	Decision Tree	0.944186	0.921053	0.921053
12	SelectKBest	Random Forest	0.944186	0.907895	0.932432
16	Combined Features	Decision Tree	0.944186	0.921053	0.921053
17	Combined Features	Random Forest	0.939535	0.894737	0.931507
41	Assessment Features	Decision Tree	0.944186	0.921053	0.921053
42	Assessment Features	Random Forest	0.944186	0.907895	0.932432

Result (how to best predict Alzheimer's)

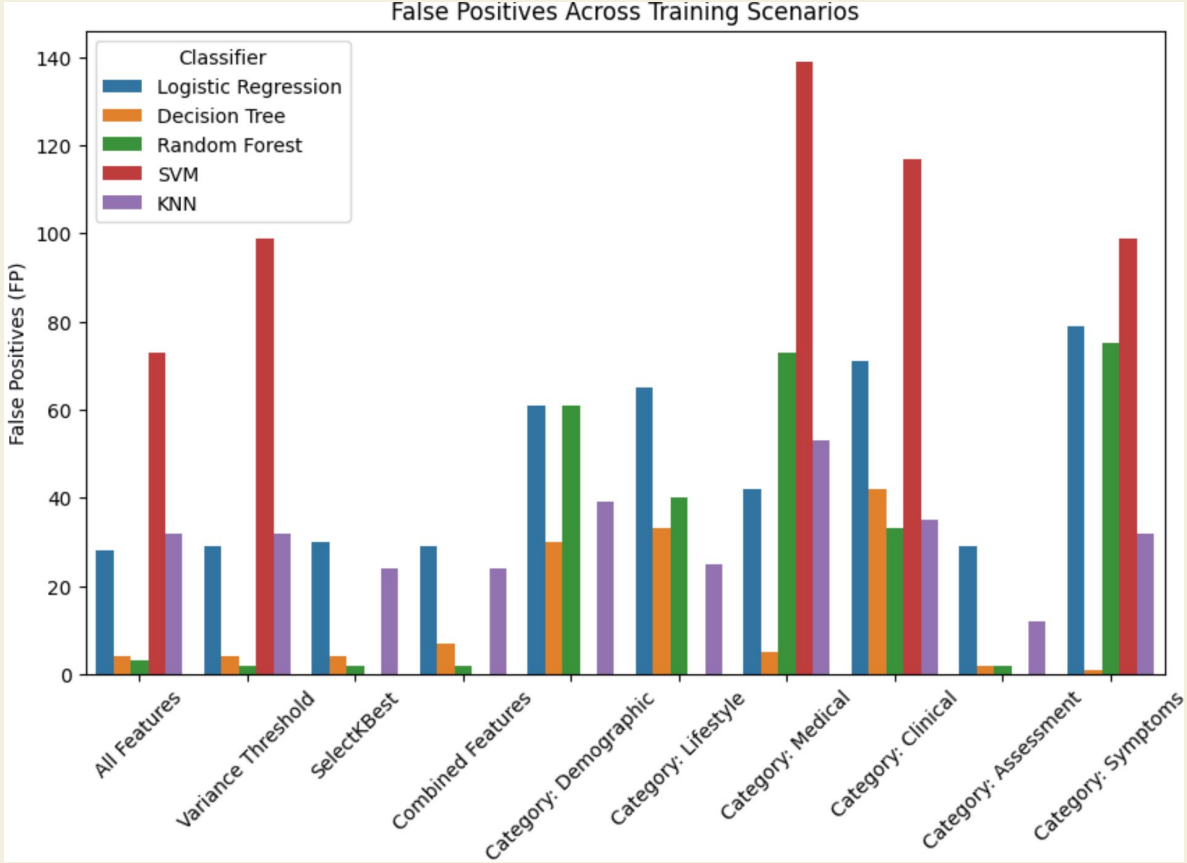
- Training on all features had great accuracy, precision, and recall but the model could be overfit
- Decision tree/random forest were best performing models based on accuracy, precision and recall
- Most realistic and well performing models based on overall accuracy:
  - Decision tree using variance threshold feature selection
  - Decision tree or random forest using Select K best
  - Decision tree using combined features
  - Decision tree or random forest using assessment features

Our Chosen Model

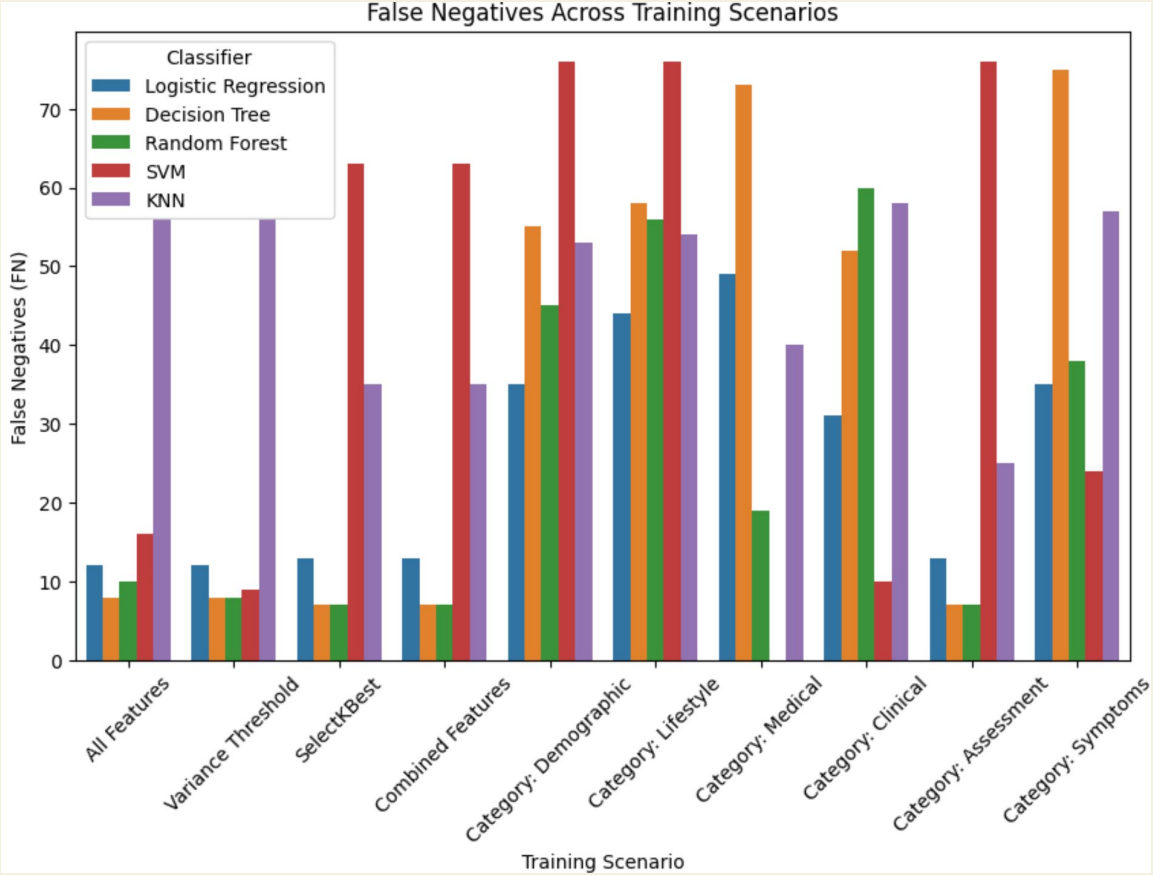
The single best model is the decision tree model using the five assessment factors. This model has a recall of .92 and an accuracy of .94 while only using 5 features.

Error Analysis

Performance on Testing Data



Performance on Testing Data



Conclusions

- The single best model is the decision tree model using the five assessment factors. This model has a recall of .92 and an accuracy of .94 while only using 5 features.
- Unsurprising that models using assessment factors performed the best given the correlation between diagnosis and memory problems, behavioral problems, MMSE, and assessment (all included in “assessment” category)

Future Implications

- Likely predict presence of Alzheimer's accurately by just using the features in both the Variance Threshold and K Best feature selection methods
- Doctors should focus on gaining accurate assessment measurements as these are the most predictive features

Q&A

Sources

- https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset?resource=download