**Introduction**

Alzheimer’s disease (AD) is a progressive neurological disorder that affects memory, thinking, and behavior. Early detection is crucial for improving patient care and management, but it remains a significant challenge due to the complexity of contributing factors and the long progression period of the disease. This project aims to develop a predictive model to assess the likelihood of an individual having Alzheimer's disease based on a range of features, including demographic details, medical history, lifestyle factors, clinical measurements, and cognitive assessments.

Below is a summary of the modeling approach for this project.

**Goals**

* To obtain a good baseline model
* Select models that produce high accuracy rates using all the features
* Adapt an approach which improves the accuracy for early detection
  + By excluding the highly correlated features in modeling
  + Restricting to the dataset with no memory complaint

**Models Explored**

We will be exploring the following models in this project.

1. Logistic Regression
2. Random forest classifier
3. XGBoost classifier
4. Gradient boosting classifier
5. Adaboost classifier
6. KNN classifier
7. Support Vector machine
8. Decision Tree
9. LDA
10. QDA
11. Naǐve Bayes

**Metrics Checked**

To evaluate the performance of the models, we will be considering the following metrics.

* Accuracy Score
* F1 Score
* Precision Score
* Recall Score
* Confusion Matrices

**Methodology**

For each method explored, we considered analysis on two broad scenarios:

1. The entire dataset
2. A restriction on the dataset based on only patients who had no memory complaints.

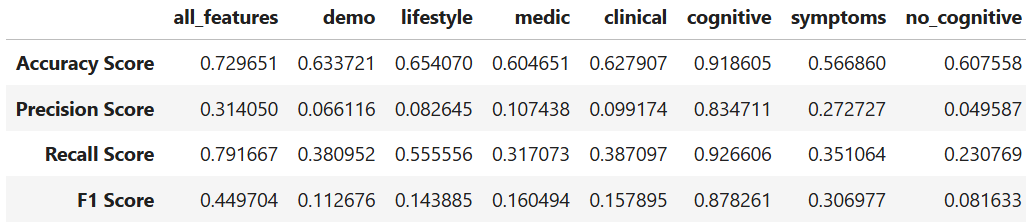
Additionally, we grouped explored the models further under these two broad scenarios by grouping the features based on similarities. The groupings are:

* All features (all\_features)
* Demographic features (demo)
* Lifestyle features (lifestyle)
* Medical history features (medic)
* Clinical features (clinical)
* Cognitive features (cognitive)
* Features related to symptoms (symptoms)
* Non-cognitive features (no\_cognitive)

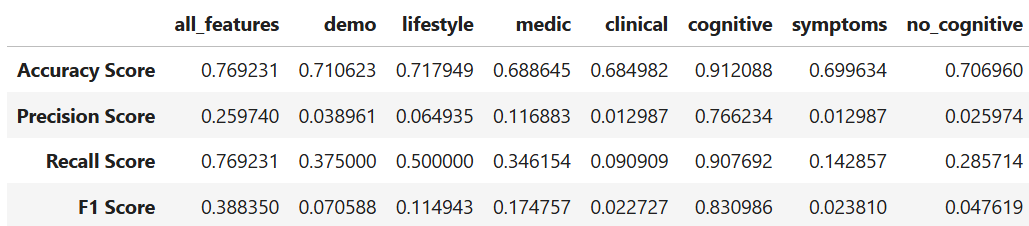
**Results**

In this section, we show the results for model exploration. The scores for the various error metrics are tabulated, while we also include confusion matrices for all features, cognitive features and non-cognitive features.

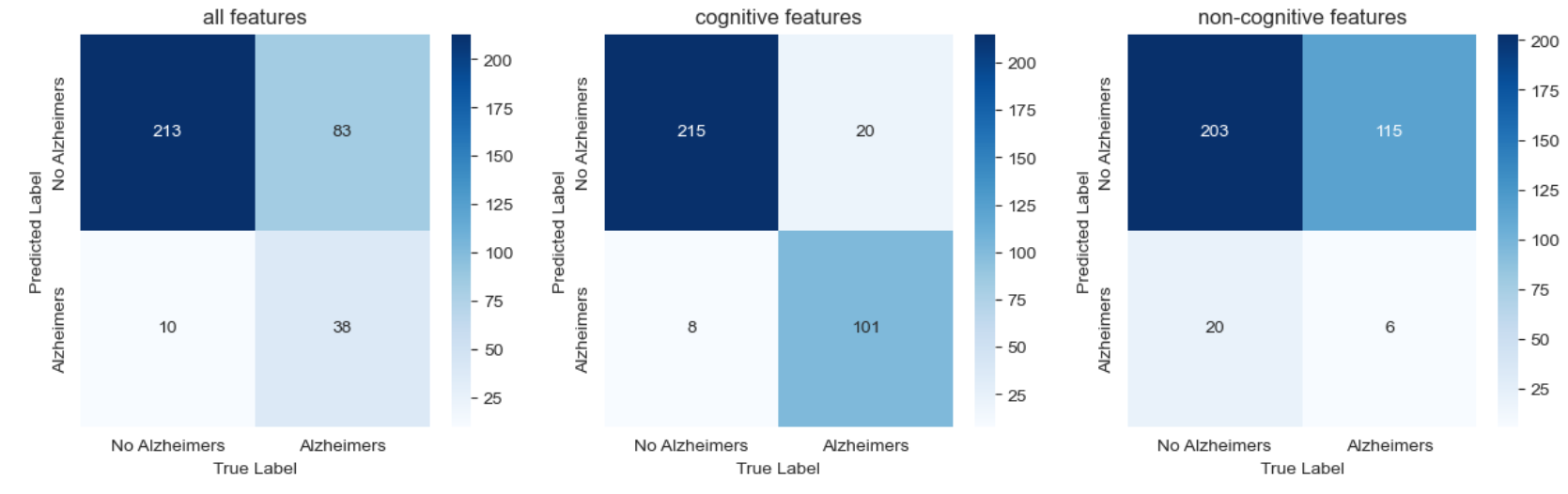
**KNN Classifier**



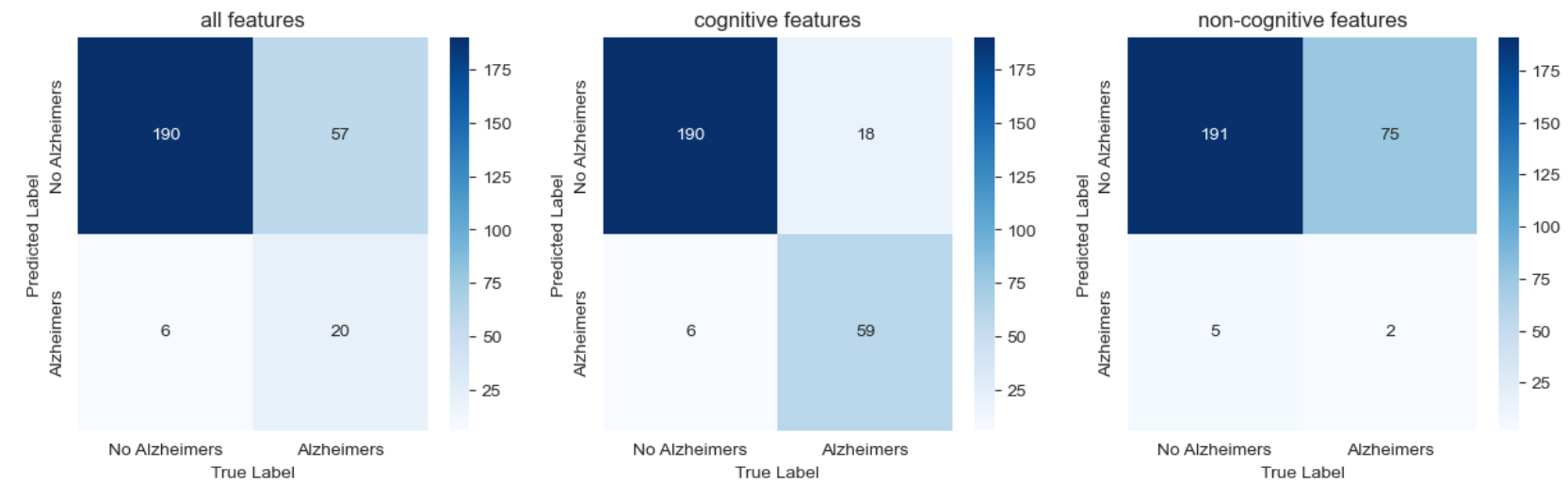
*Metrics: Entire dataset*



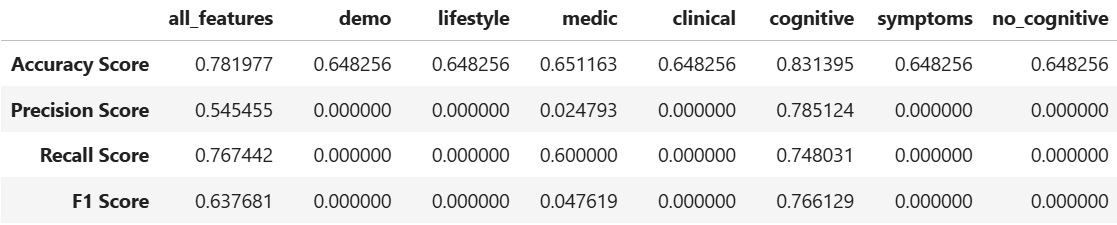
*Metrics: Dataset restricted to patients with no memory complaints*

**

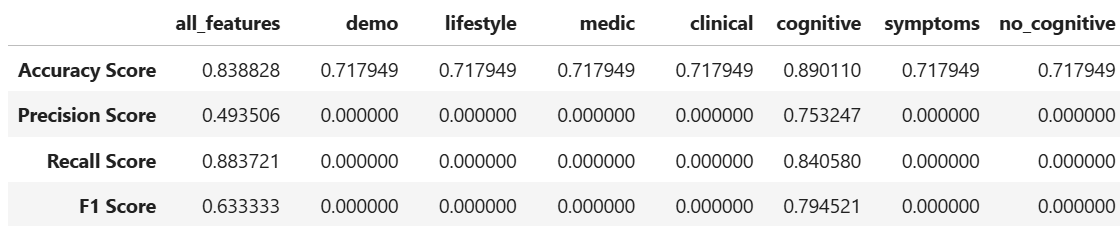
*Confusion matrices: Entire dataset*

*Confusion matrices: Dataset restricted to patients with no memory complaints*

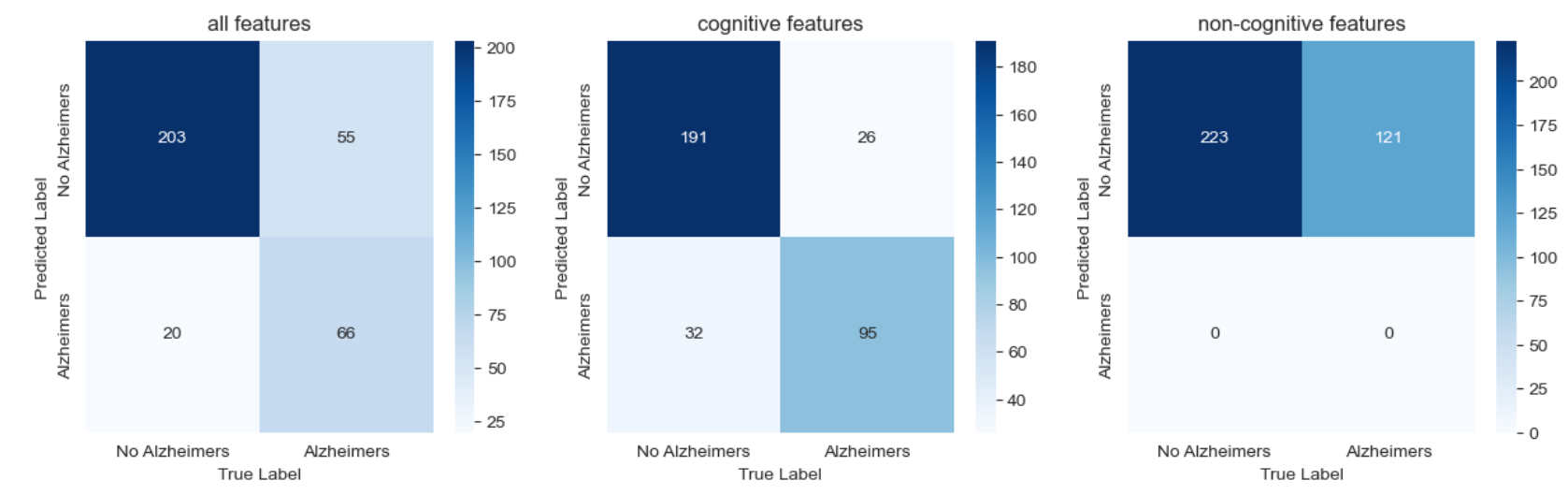
**SVM (Polynomial Kernel)**



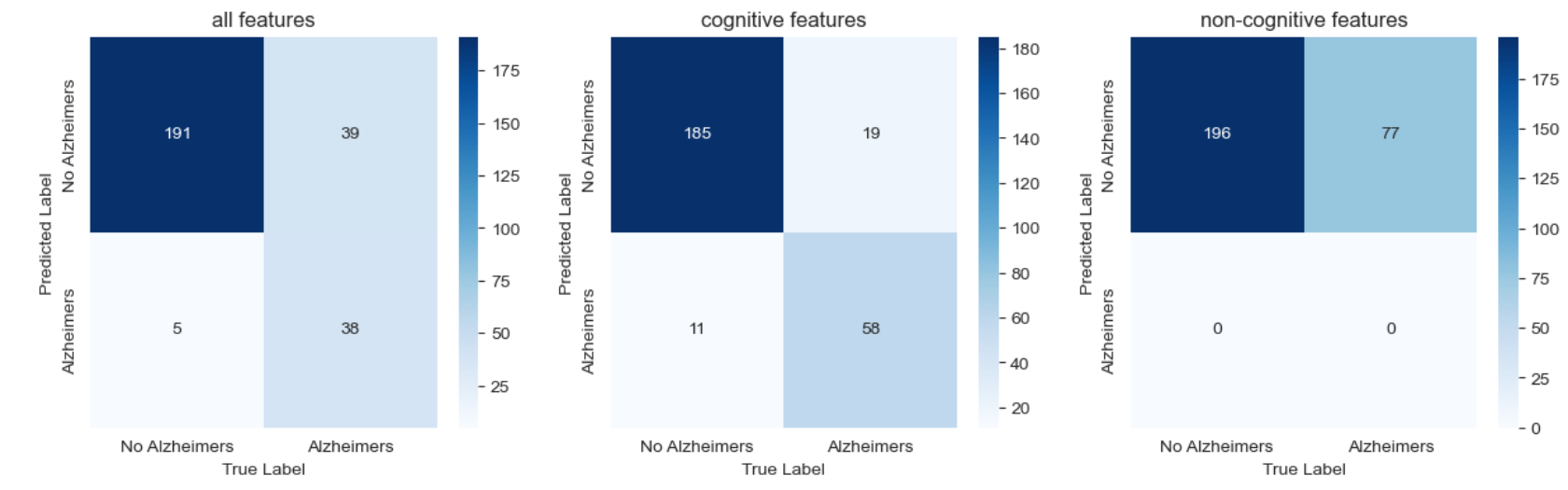
*Metrics: Entire dataset*



*Metrics: Dataset restricted to patients with no memory complaints*

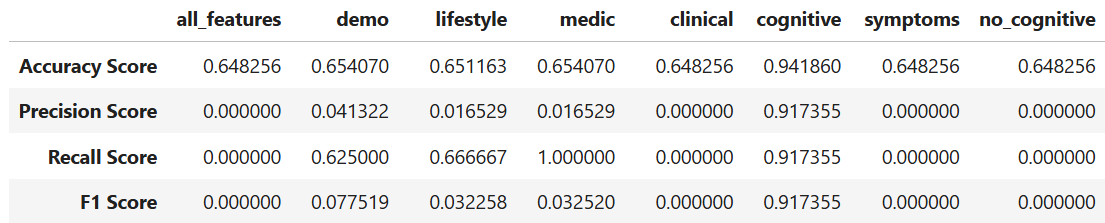
**

*Confusion matrices: Entire dataset*

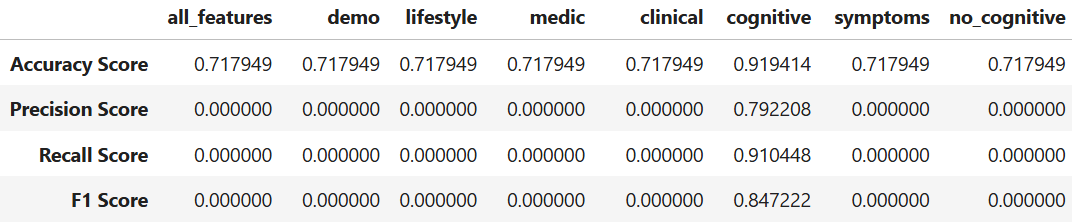
**

*Confusion matrices: Dataset restricted to patients with no memory complaints*

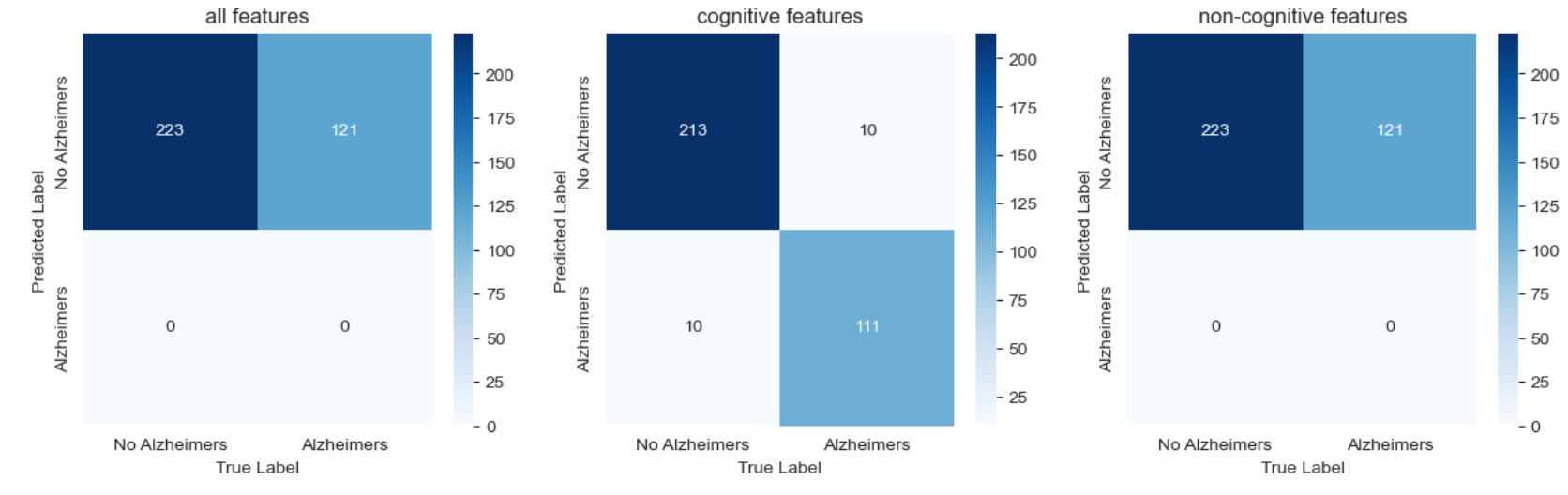
**SVM (RBF Kernel)**



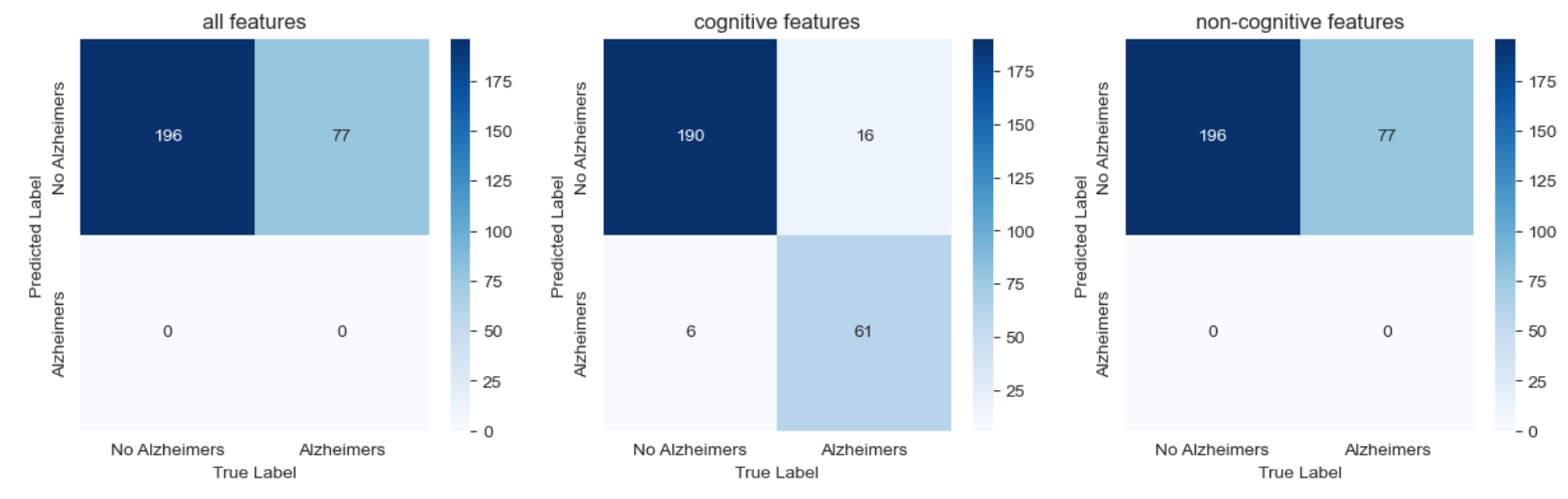
*Metrics: Entire dataset*



*Metrics: Dataset restricted to patients with no memory complaints*

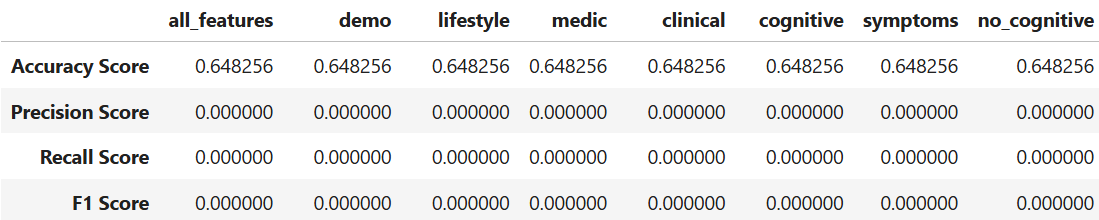
**

*Confusion matrices: Entire dataset*

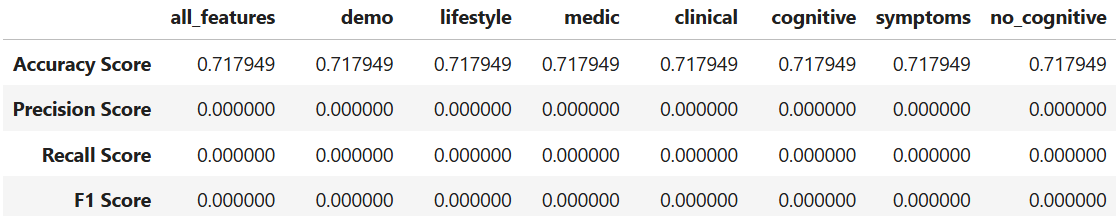
**

*Confusion matrices: Dataset restricted to patients with no memory complaints*

**SVM (Sigmoid Kernel)**



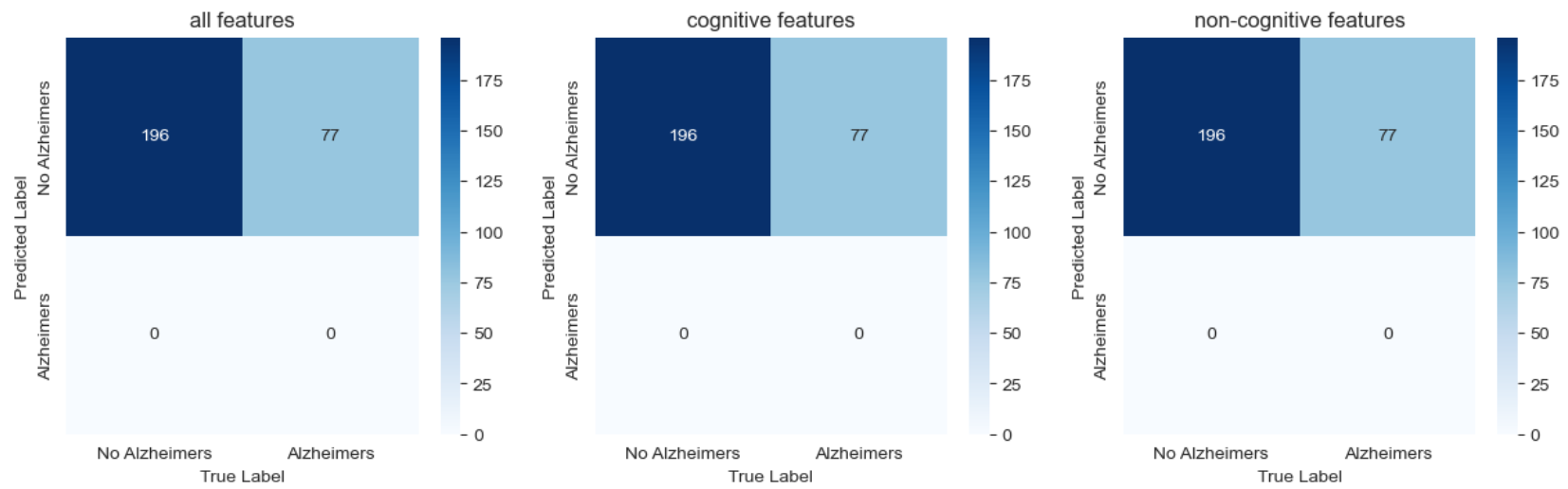
*Metrics: Entire dataset*



*Metrics: Dataset restricted to patients with no memory complaints*

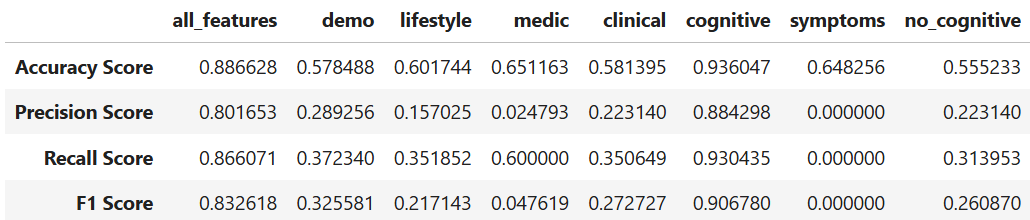
**

*Confusion matrices: Entire dataset*

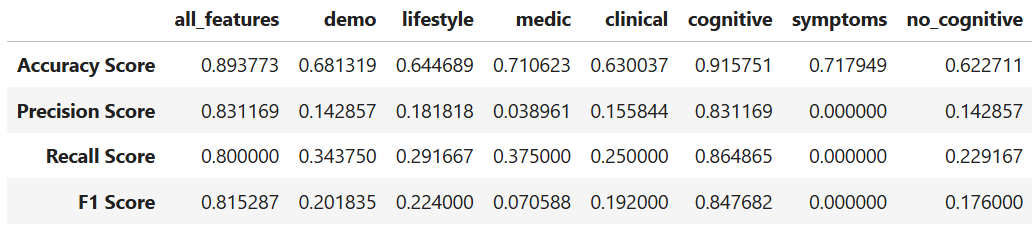
**

*Confusion matrices: Dataset restricted to patients with no memory complaints*

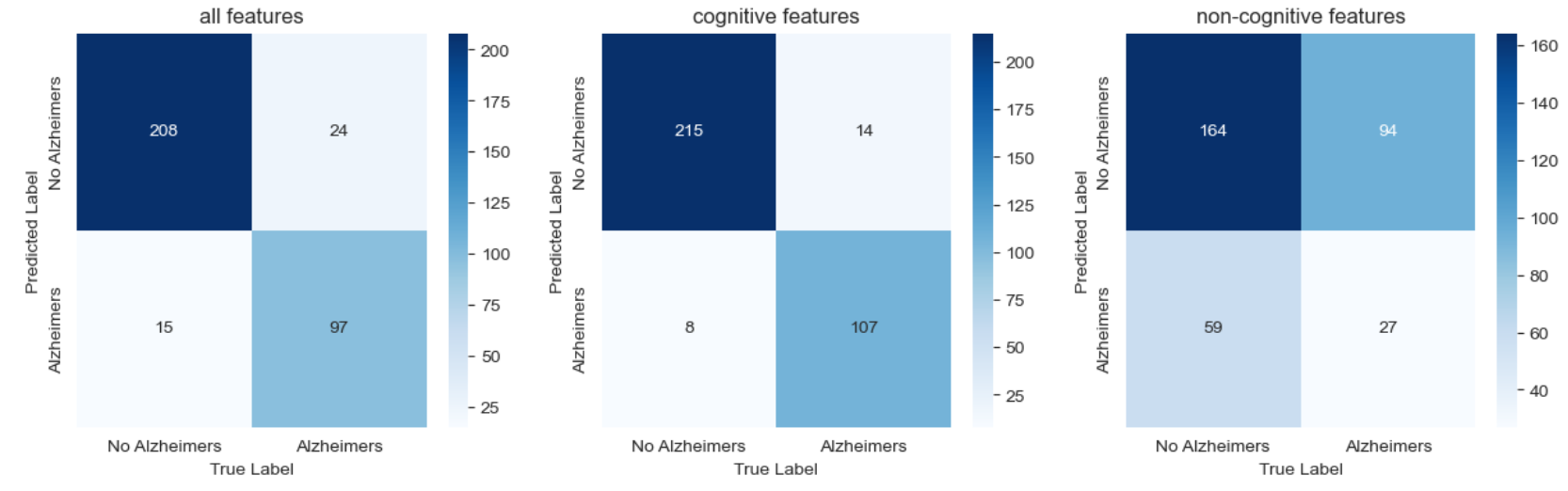
**Decision Tree Classifier**



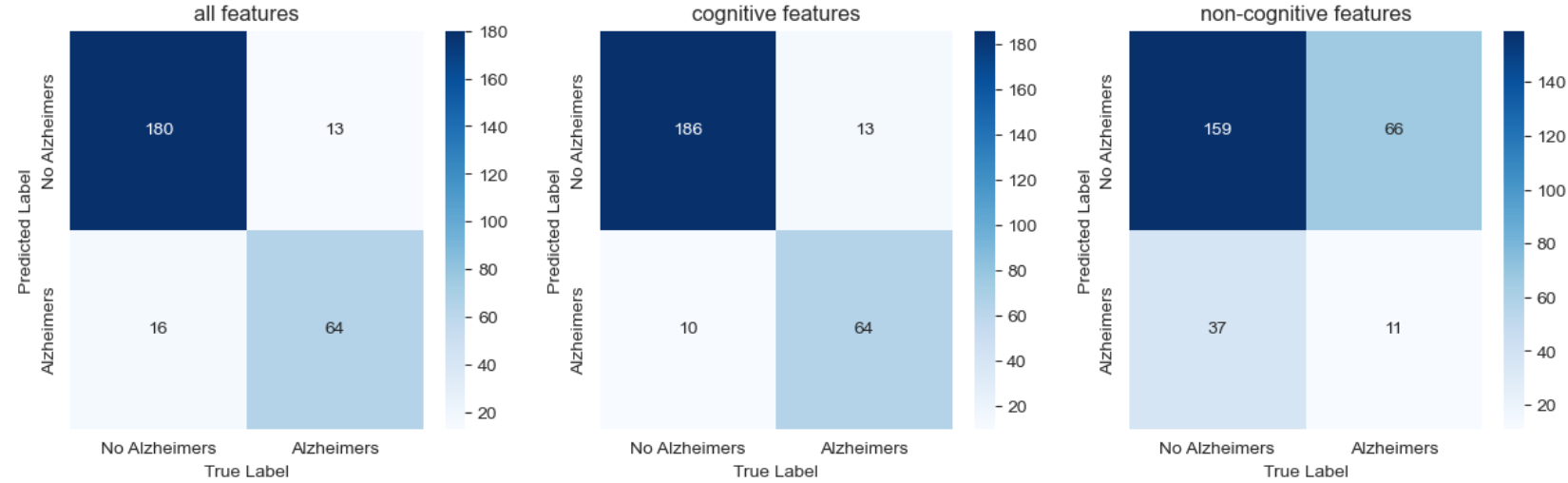
*Metrics: Entire dataset*



*Metrics: Dataset restricted to patients with no memory complaints*

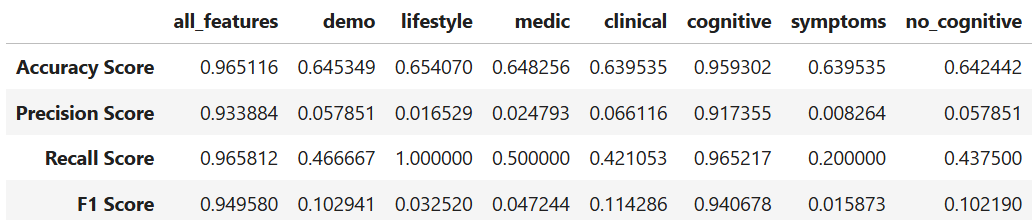
**

*Confusion matrices: Entire dataset*

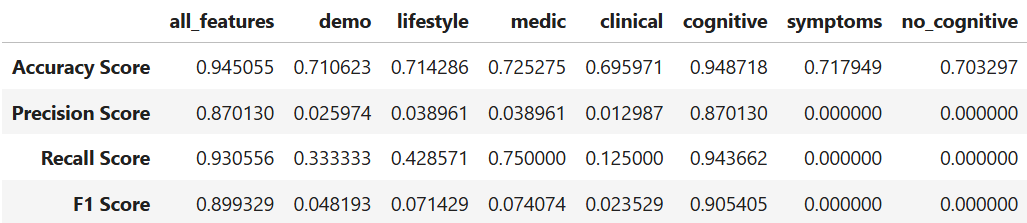
**

*Confusion matrices: Dataset restricted to patients with no memory complaints*

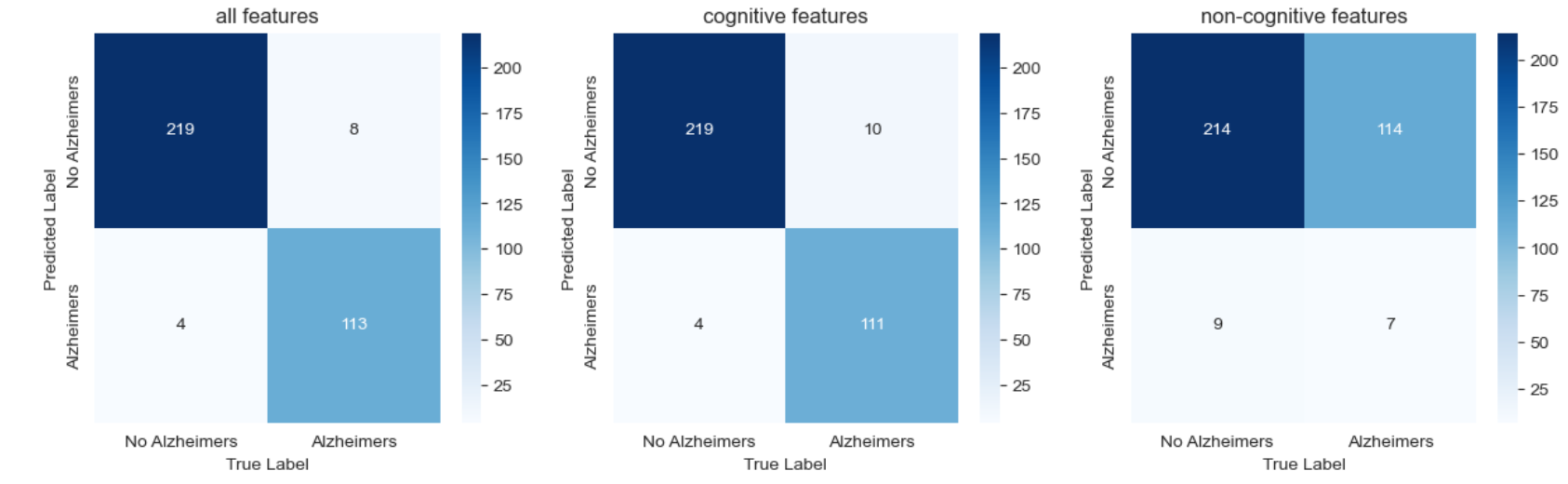
**AdaBoost Classifier**



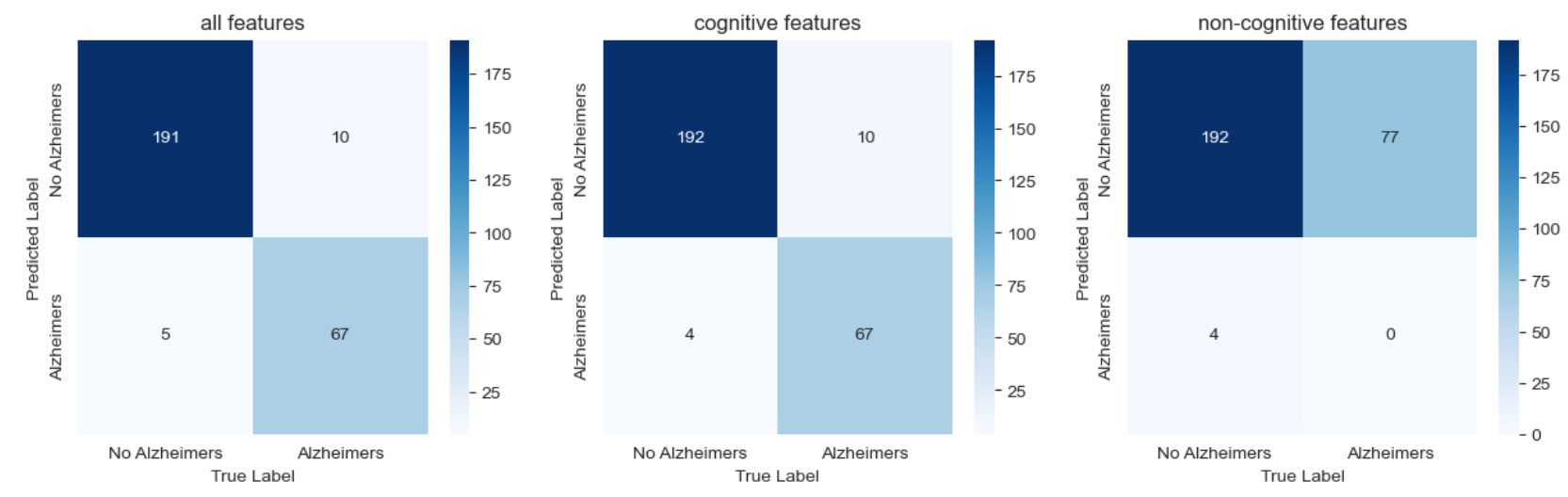
*Metrics: Entire dataset*



*Metrics: Dataset restricted to patients with no memory complaints*

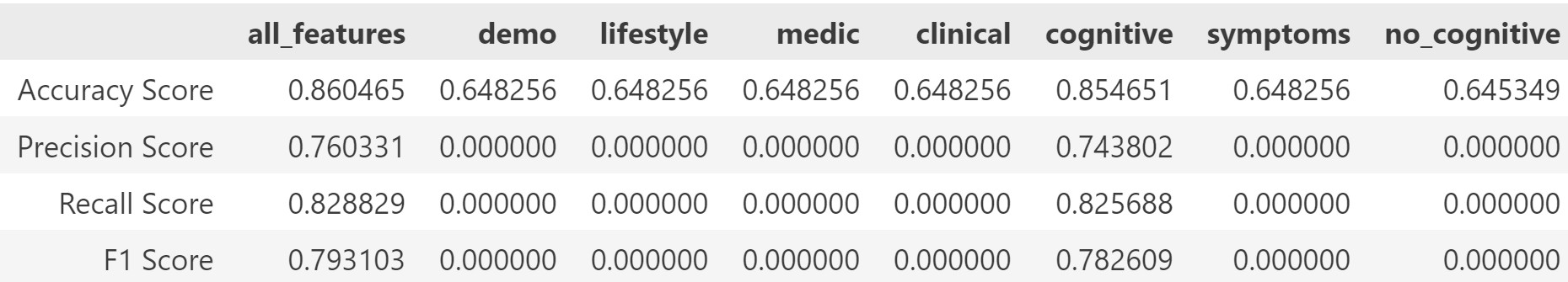
**

*Confusion matrices: Entire dataset*

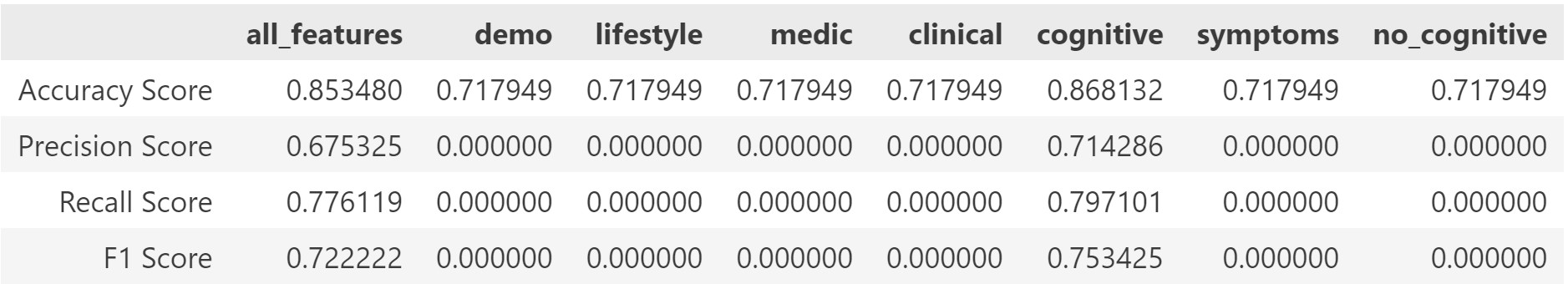
**

*Confusion matrices: Dataset restricted to patients with no memory complaints*

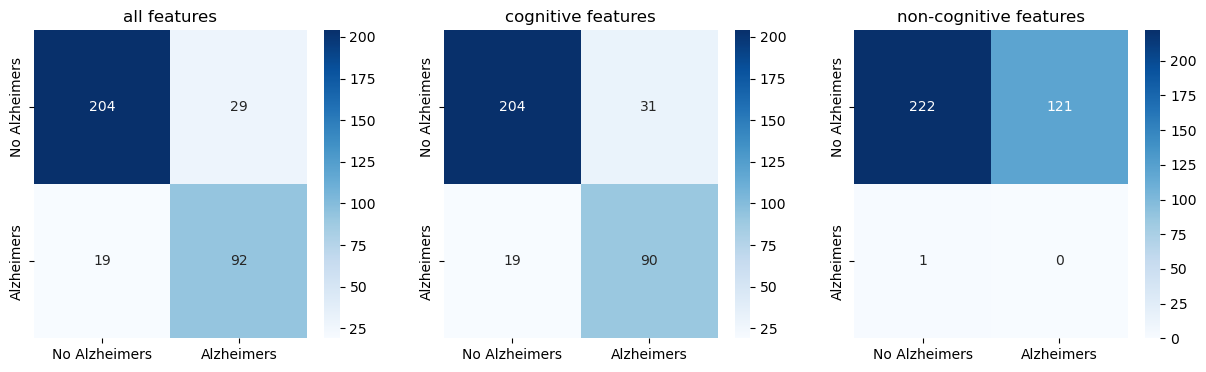
**Logistic Regression**



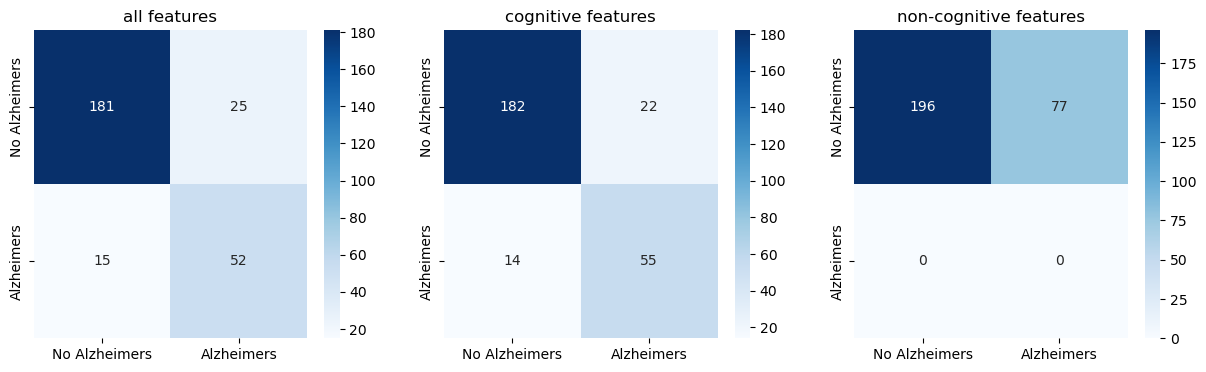
*Metrics: Entire dataset*



*Metrics: Dataset restricted to patients with no memory complaints*

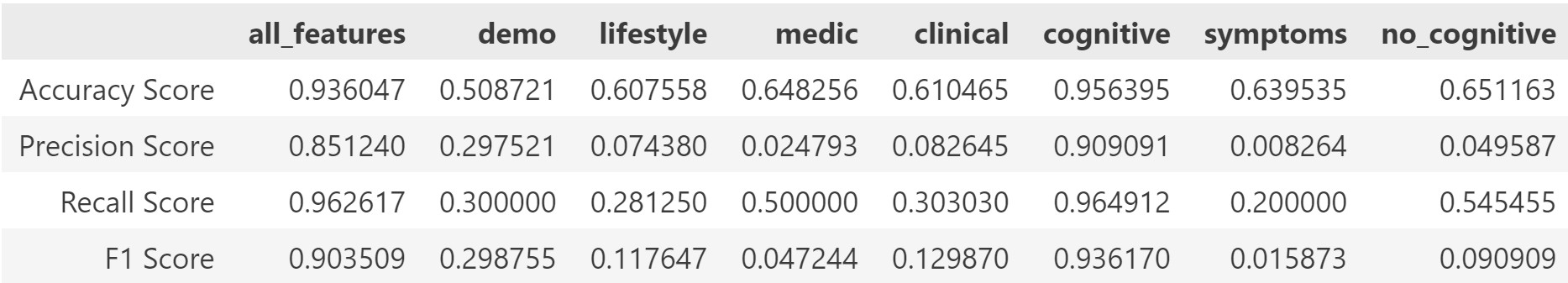


*Confusion matrices: Entire dataset*

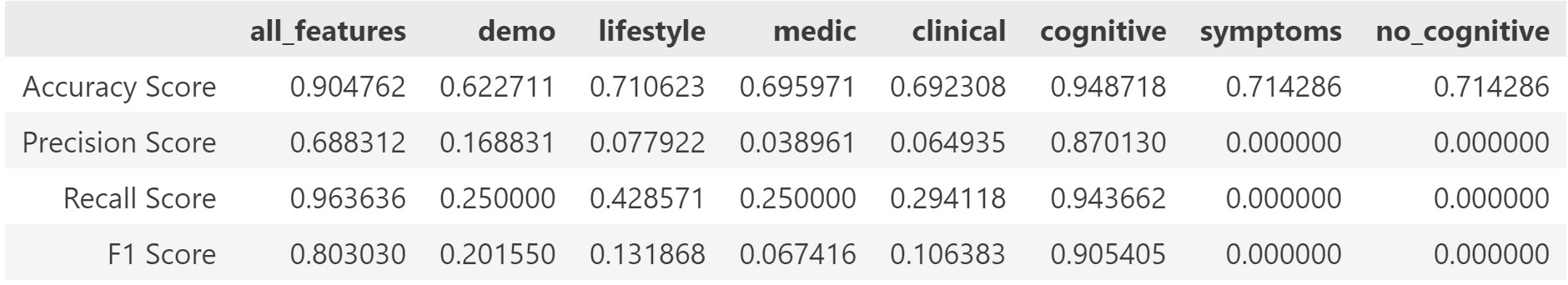


*Confusion matrices: Dataset restricted to patients with no memory complaints*

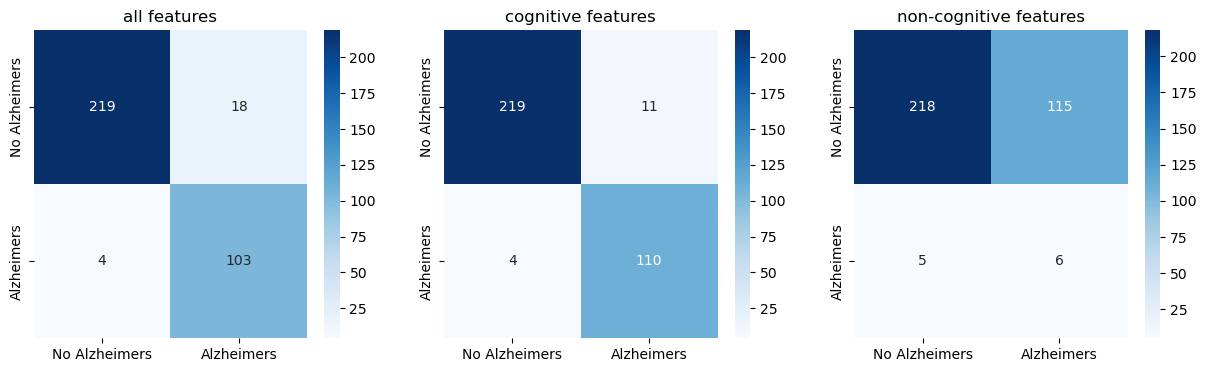
**Random Forest Classifier**



*Metrics: Entire dataset*



*Metrics: Dataset restricted to patients with no memory complaints*

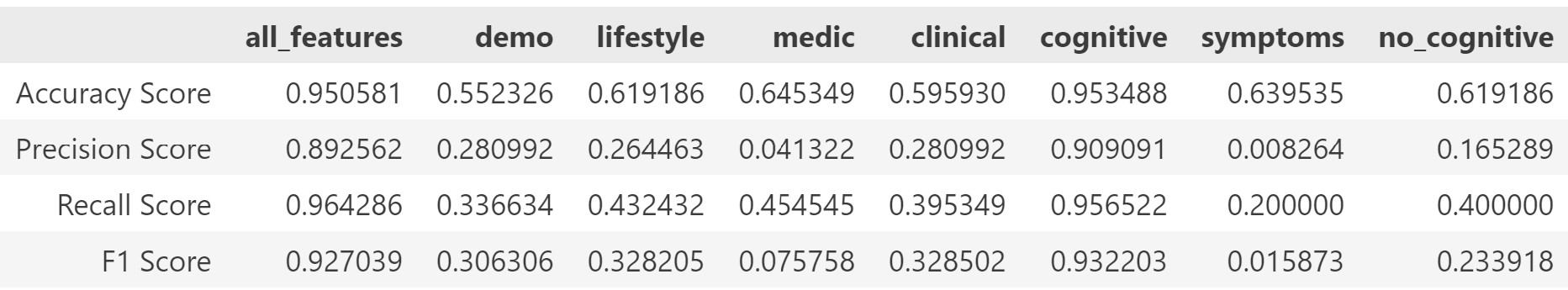


*Confusion matrices: Entire dataset*

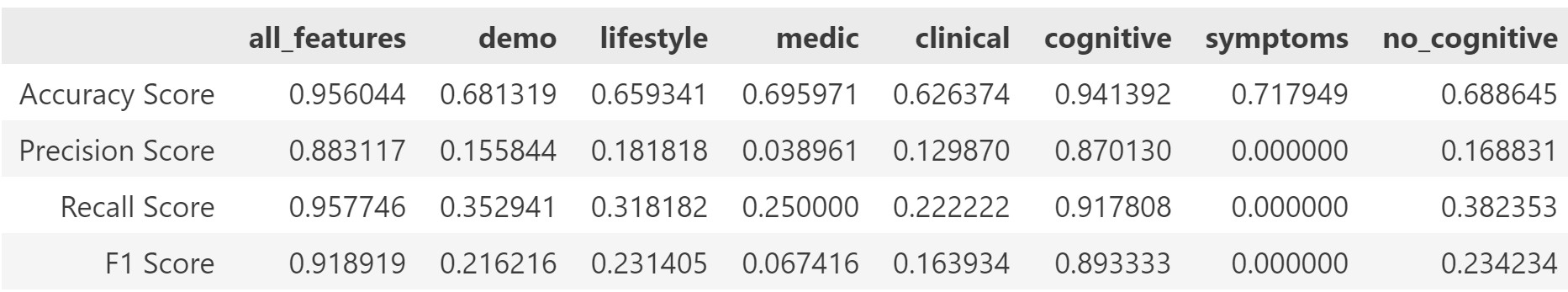


*Confusion matrices: Dataset restricted to patients with no memory complaints*

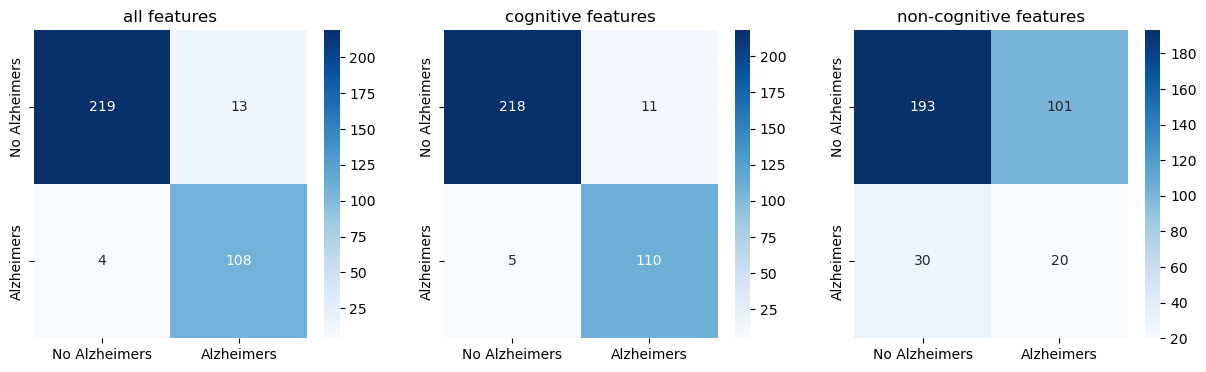
**XGBoost Classifier**



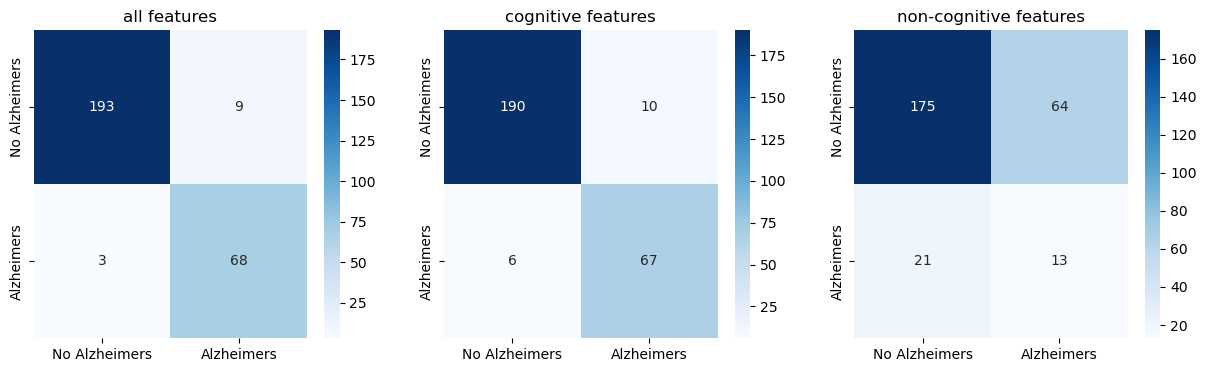
*Metrics: Entire dataset*



*Metrics: Dataset restricted to patients with no memory complaints*

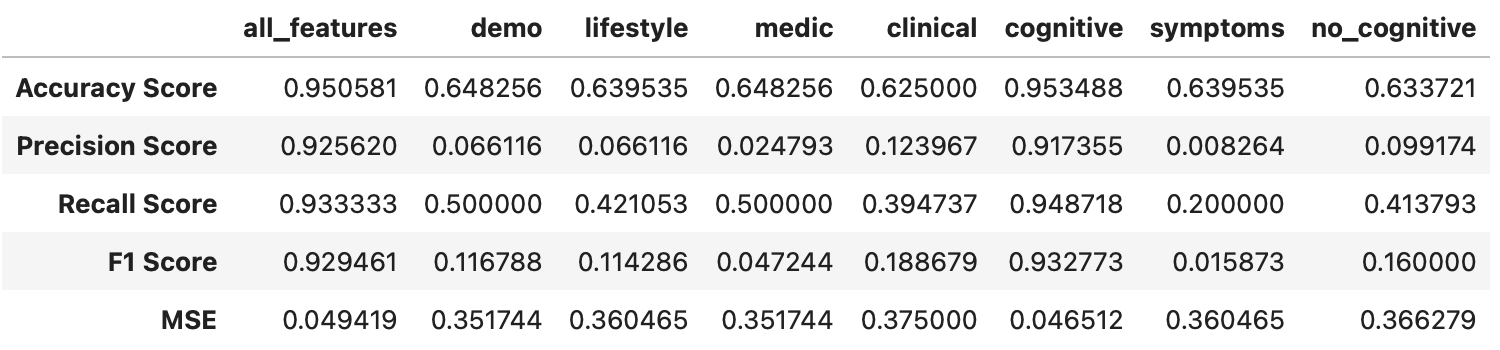


*Confusion matrices: Entire dataset*

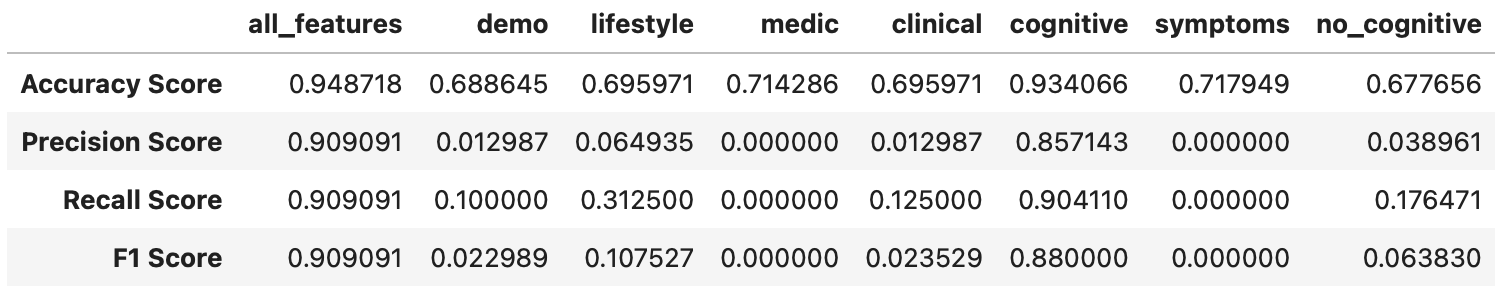


*Confusion matrices: Dataset restricted to patients with no memory complaints*

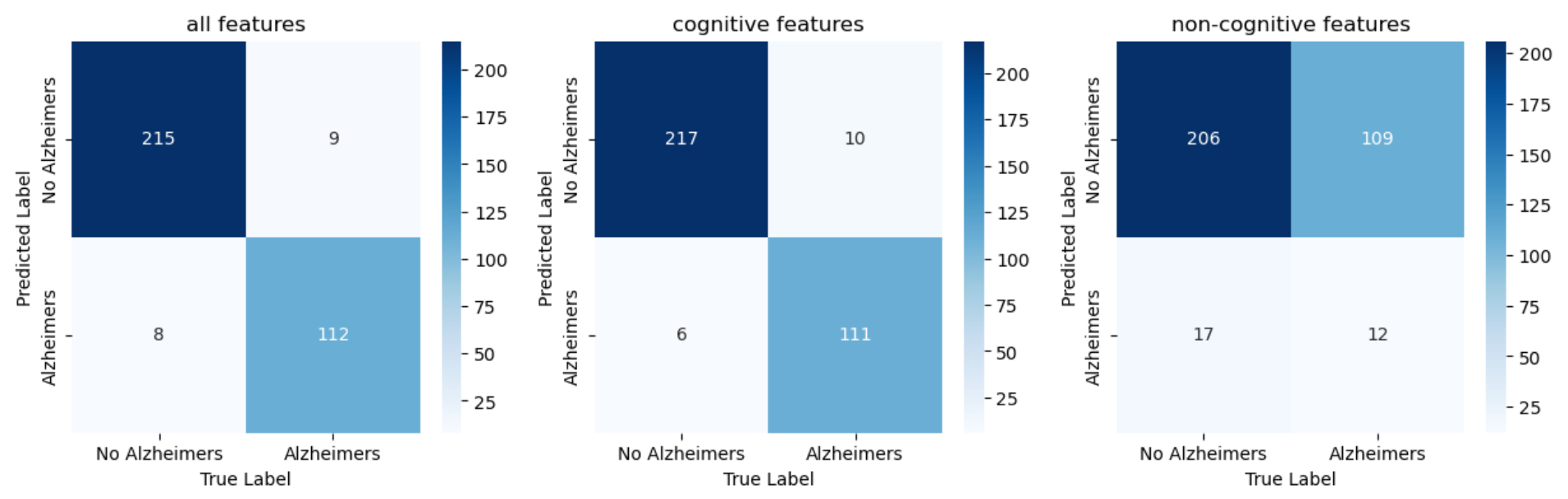
**Gradient Boosting Classifier**



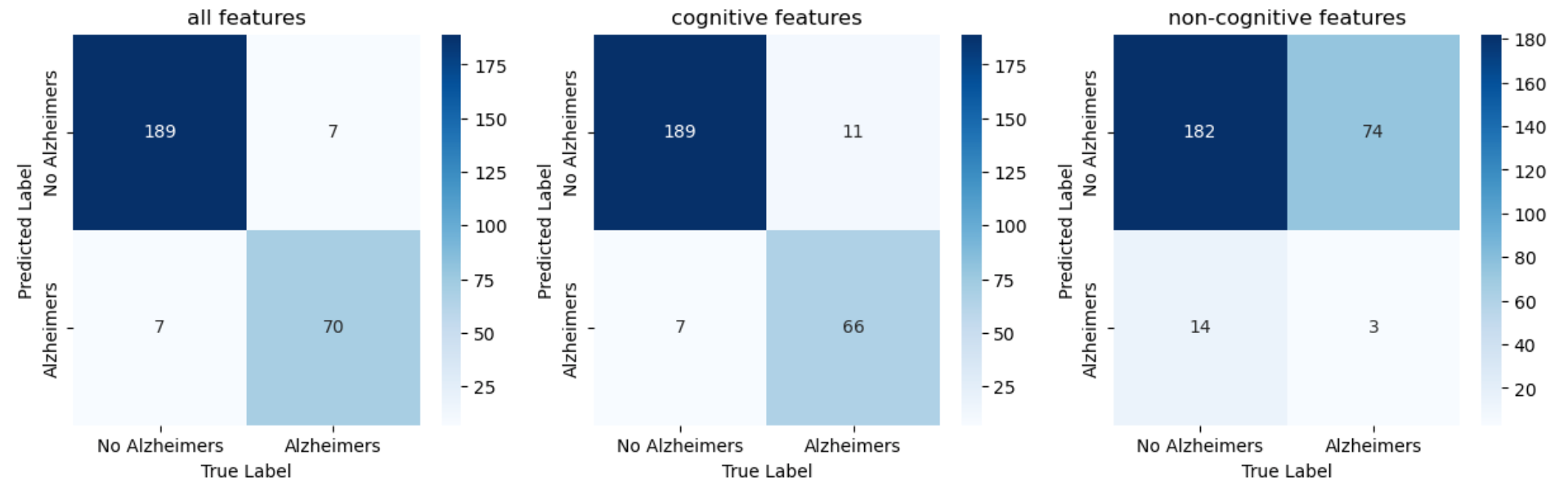
*Metrics: Entire dataset*



*Metrics: Dataset restricted to patients with no memory complaints*

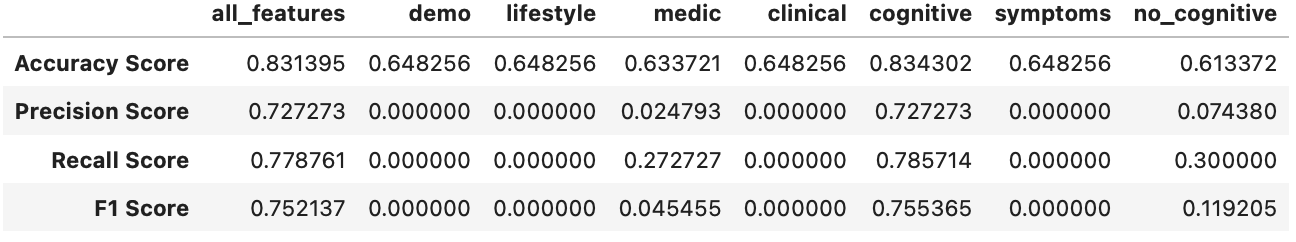


*Confusion matrices: Entire dataset*

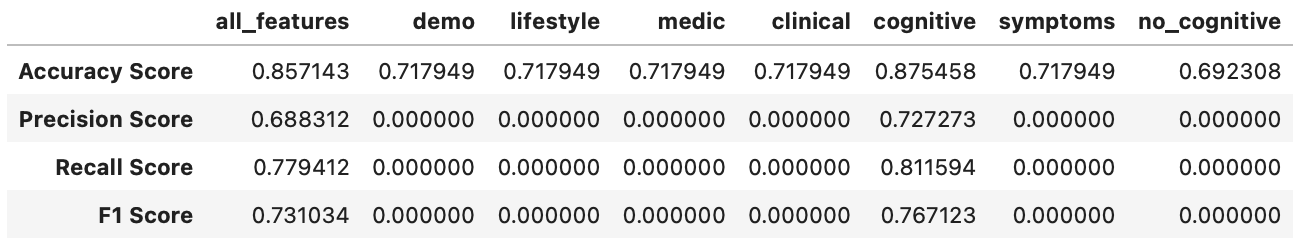


*Confusion matrices: Dataset restricted to patients with no memory complaints*

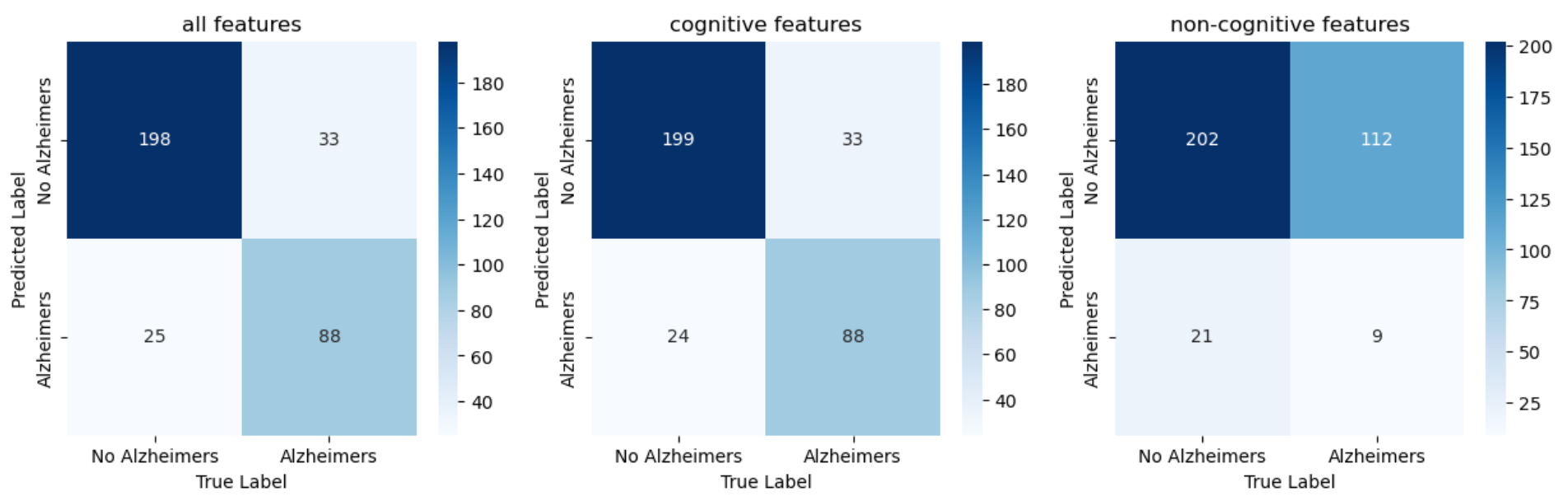
**Gaussian Naǐve Bayes**



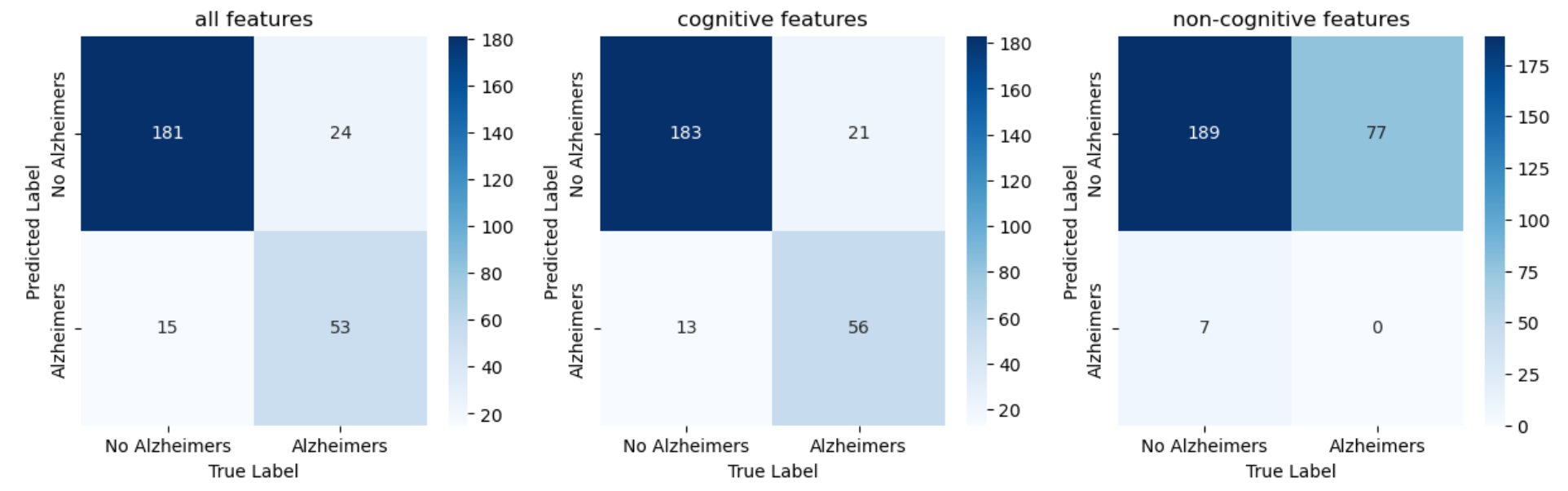
*Metrics: Entire dataset*



*Metrics: Dataset restricted to patients with no memory complaints*

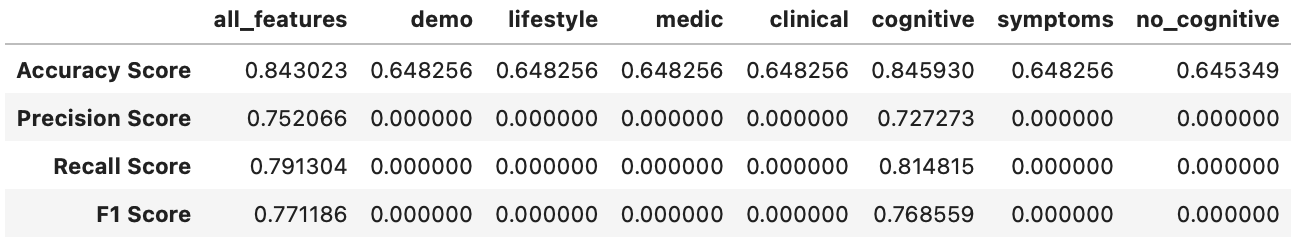


*Confusion matrices: Entire dataset*

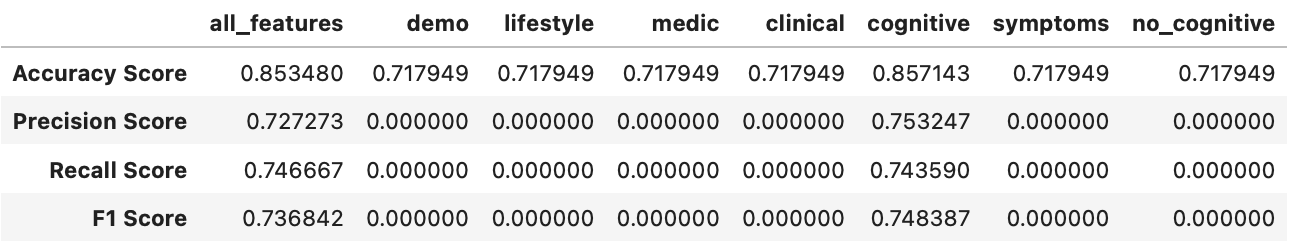


*Confusion matrices: Dataset restricted to patients with no memory complaints*

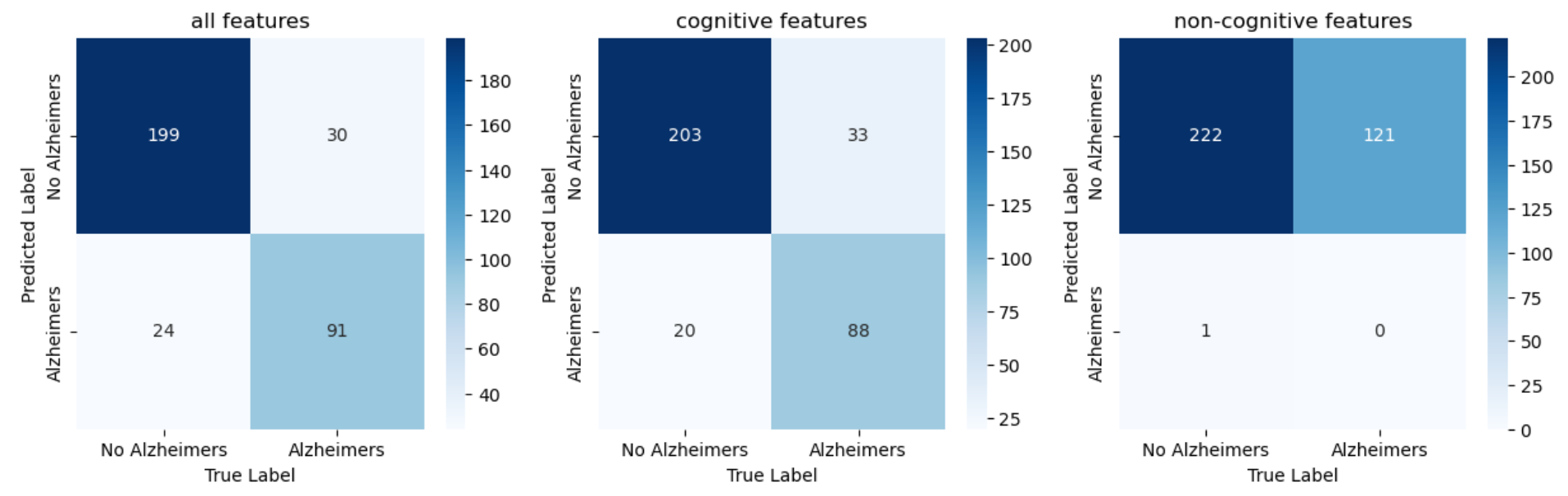
**Linear Discriminant Analysis**



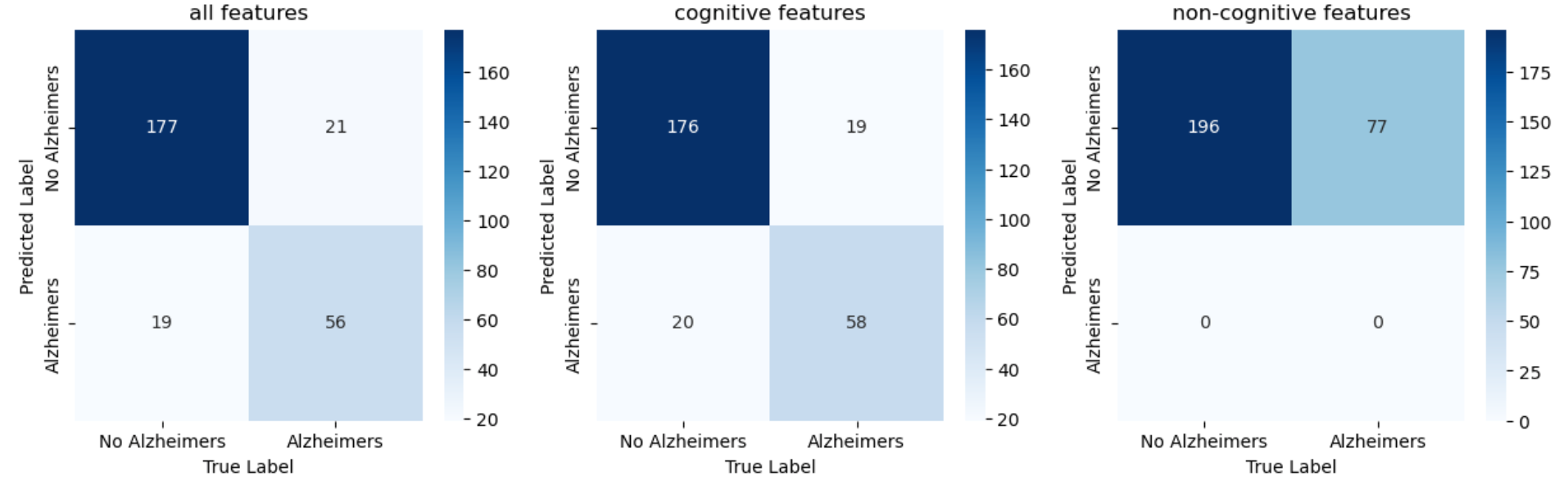
*Metrics: Entire dataset*



*Metrics: Dataset restricted to patients with no memory complaints*

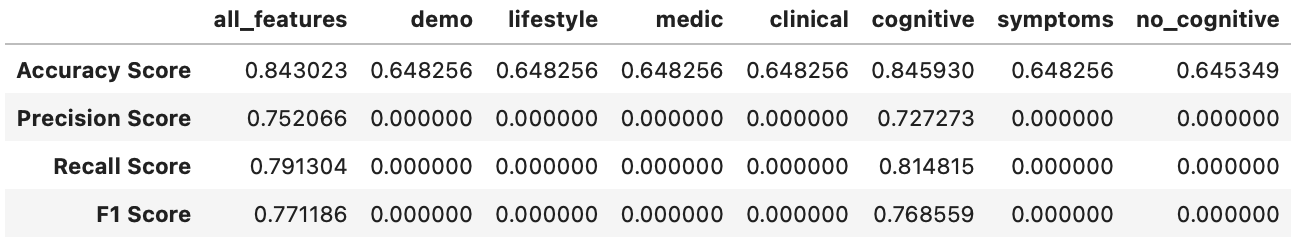


*Confusion matrices: Entire dataset*

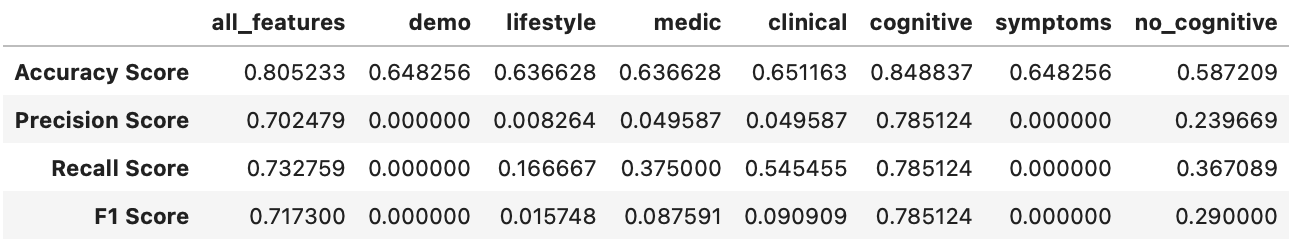


*Confusion matrices: Dataset restricted to patients with no memory complaints*

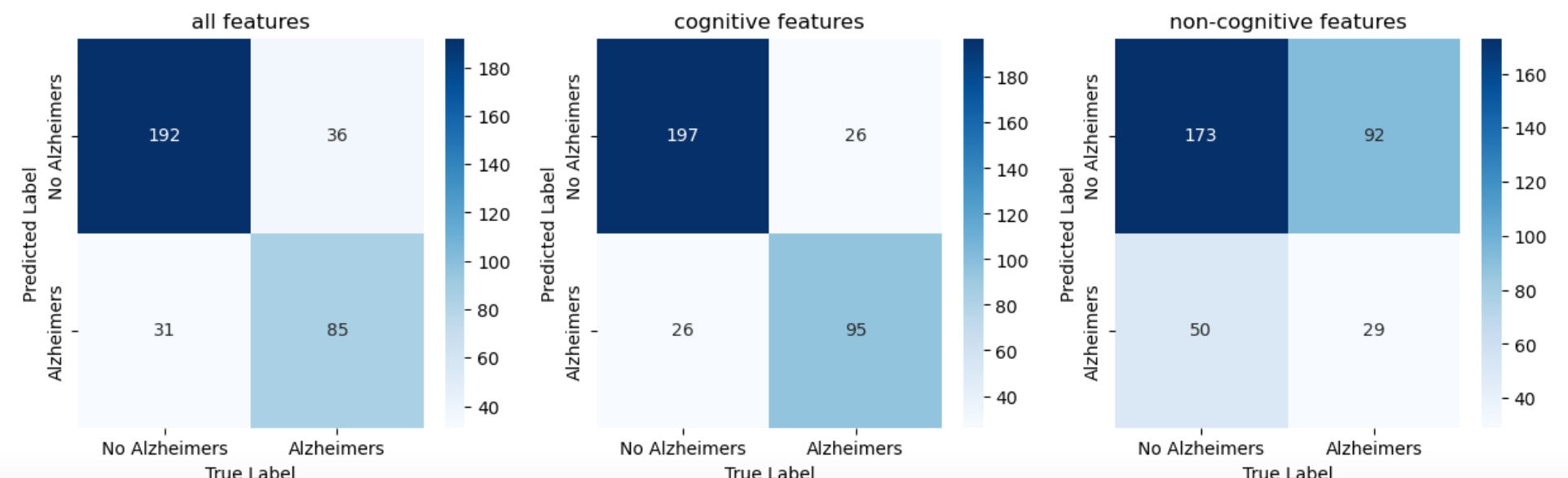
**Quadratic Discriminant Analysis**



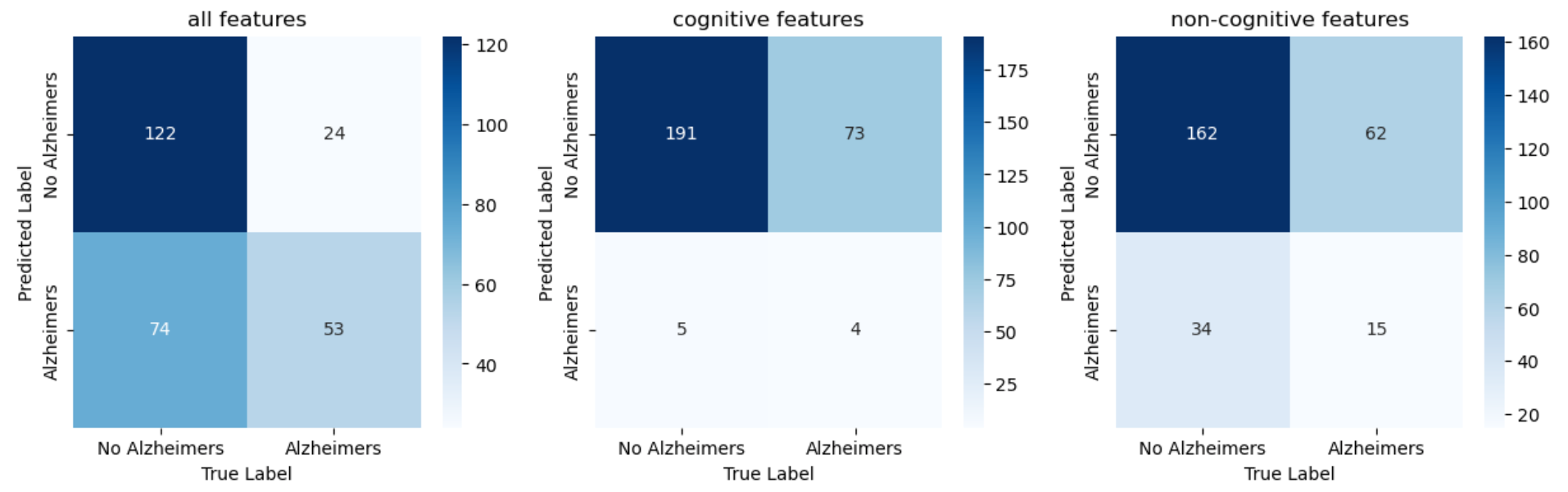
*Metrics: Entire dataset*



*Metrics: Dataset restricted to patients with no memory complaints*



*Confusion matrices: Entire dataset*



*Confusion matrices: Dataset restricted to patients with no memory complaints*

**DISCUSSION OF RESULTS**

The metrics for when we use the entire dataset seem to be slightly better than when the dataset is restricted to only records of patients with no memory complaints. There was only one case (the recall for QDA) where the difference was significant; except for that, the difference between the metrics for the two datasets were somewhat minute. This was interesting to look at as intuitively memory complaints is one main characteristic of Alzheimer's disease, so it is interesting to see that a restrictive dataset in that regard did not outperform the entire dataset. The table below gives the comparison between using all the dataset and the dataset for only patients with no memory complaints. This comparison was made for when we use all the features.

| **MODEL** | **A1** | **A2** | **P1** | **P2** | **R1** | **R2** | **F1** | **F2** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Logistic regression | 0.8605 | 0.85348 | 0.7603 | 0.67533 | 0.829 | 0.776119 | 0.7931 | 0.722222 |
| Random Forest | 0.936 | 0.90476 | 0.8512 | 0.68831 | 0.9623 | 0.963636 | 0.9304 | 0.803030 |
| Gradient boosting | 0.9506 | 0.94872 | 0.9256 | 0.90909 | 0.9333 | 0.909091 | 0.9295 | 0.909091 |
| XGBoost | 0.9506 | 0.95604 | 0.8926 | 0.88312 | 0.9643 | 0.957746 | 0.927 | 0.918919 |
| Adaboost | 0.9651 | 0.94506 | 0.8678 | 0.87013 | 0.9130 | 0.930556 | 0.8898 | 0.899329 |
| KNN | 0.7297 | 0.76923 | 0.3141 | 0.25974 | 0.7917 | 0.769231 | 0.4498 | 0.388350 |
| SVM (Poly Kernel) | 0.7820 | 0.83883 | 0.5455 | 0.49351 | 0.7674 | 0.883721 | 0.6377 | 0.633333 |
| Naǐve Bayes | 0.8314 | 0.85714 | 0.7273 | 0.68831 | 0.7788 | 0.779412 | 0.7521 | 0.731034 |
| LDA | 0.843 | 0.85348 | 0.752 | 0.72727 | 0.7913 | 0.746667 | 0.7712 | 0.736842 |
| QDA | 0.8052 | 0.64103 | 0.7025 | 0.68831 | 0.7328 | 0.417323 | 0.7173 | 0.519608 |

\*Dataset for patients with no memory complaints have better score

\*Significant difference in metric scores for when using entire dataset vs dataset with only patients with no memory complaints

**1=entire data set; 2= data set with only patients with no memory complaints; A = accuracy; P=precision R = recall F= f1 score**

Another interesting observation was that the sub categories that gave high accuracy were when we used all features and when we used only cognitive features. This cut across all models discussed and for both scenarios: using the entire dataset or the data set restricted on patients with no memory complaints.