Classifying Thyroid Cancer Tumors

CAIS++
CENTER FOR AI IN SOCIETY'S STUDENT BRANCH

Motivation

- By predicting tumor classification of Thyroid disease based off of clinical and health factors, we can predict risk and target groups at risk for earlier detection and treatment.

Data

- Our target variable, T, represents the classification of the tumor
  - 6 categories: T1a, T1b, T2, T3a, T3b, T4a, T4b
- All input variables were categorical, so we did one-hot encoding (got dummies)
- Variables N, M, Stage, Response, and Recurred were based off of T, the target variable, so we dropped them to focus on prediction of tumor classification without knowledge of the treatment afterwards

Models used

- Methods that we chose:
  - Ruwan: K-nn, Voting
  - Taran: Logistic regression, random forest
  - Alice: SVM, Gradient Boosting
- Hyperparameter tuning:
  - We performed GridSearch to get optimal hyperparameters

Performance Metrics Specific to Problem

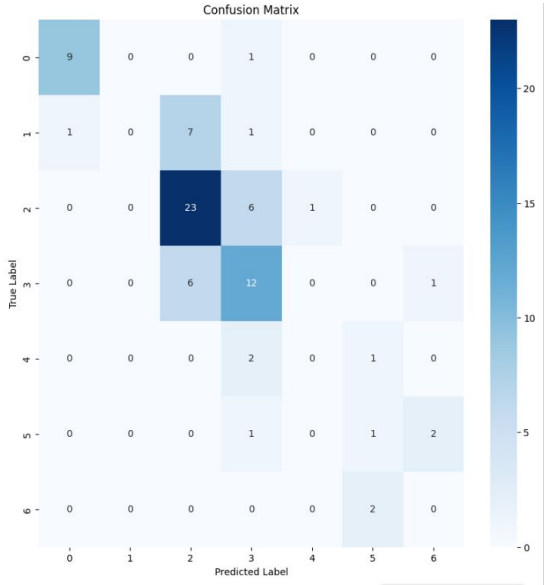
- Precision: Of the tumors classified in this category, how many were classified correctly
- Recall: How many of the tumors that were actually this classification were classified in this category
- Accuracy: Out of all tumors, how many were classified correctly

```
Mapping of numerical labels to original 'T' labels:
0 is for T1a
1 is for T1b
2 is for T2
3 is for T3a
4 is for T3b
5 is for T4a
6 is for T4b
```

Results

Support Vector Machines (SVM)

```
Best Hyperparameters:  {'C': 1, 'gamma': 1, 'kernel': 'linear'}

Accuracy:  0.5844155844155844
AUROC:  0.8244606088642603
              precision    recall  f1-score   support

           0       0.90      0.90      0.90        10
           1       0.00      0.00      0.00         9
           2       0.64      0.77      0.70        30
           3       0.52      0.63      0.57        19
           4       0.00      0.00      0.00         3
           5       0.25      0.25      0.25         4
           6       0.00      0.00      0.00         2

    accuracy                           0.58        77
   macro avg       0.33      0.36      0.35        77
weighted avg       0.51      0.58      0.54        77
```



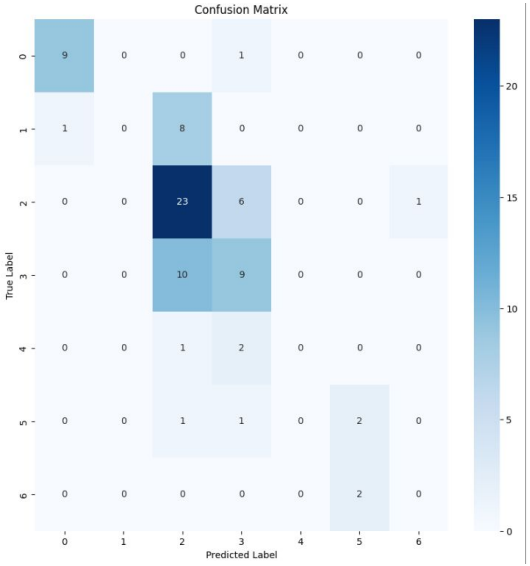
- Linear kernel: so classes are roughly linearly separable in feature space

CAIS++
CENTER FOR AI IN SOCIETY'S STUDENT BRANCH

Results

Gradient Boosting

```
Best Hyperparameters:  {'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 100}

Accuracy:  0.5584415584415584
AUROC:  0.8219494133966787
              precision    recall  f1-score   support

           0       0.90      0.90      0.90        10
           1       0.00      0.00      0.00         9
           2       0.53      0.77      0.63        30
           3       0.47      0.47      0.47        19
           4       0.00      0.00      0.00         3
           5       0.50      0.50      0.50         4
           6       0.00      0.00      0.00         2

    accuracy                           0.56        77
   macro avg       0.34      0.38      0.36        77
weighted avg       0.47      0.56      0.51        77
```

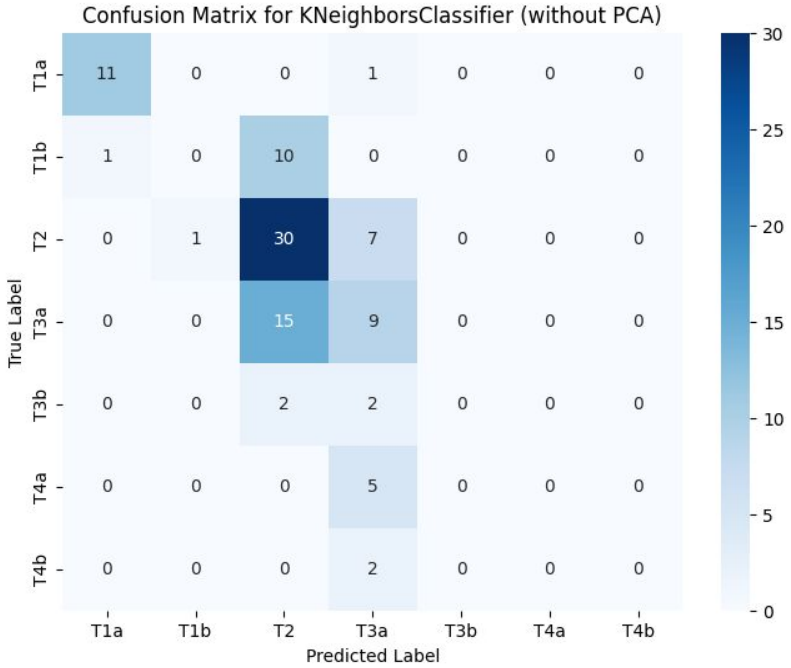


CAIS++
CENTER FOR AI IN SOCIETY'S STUDENT BRANCH

Results

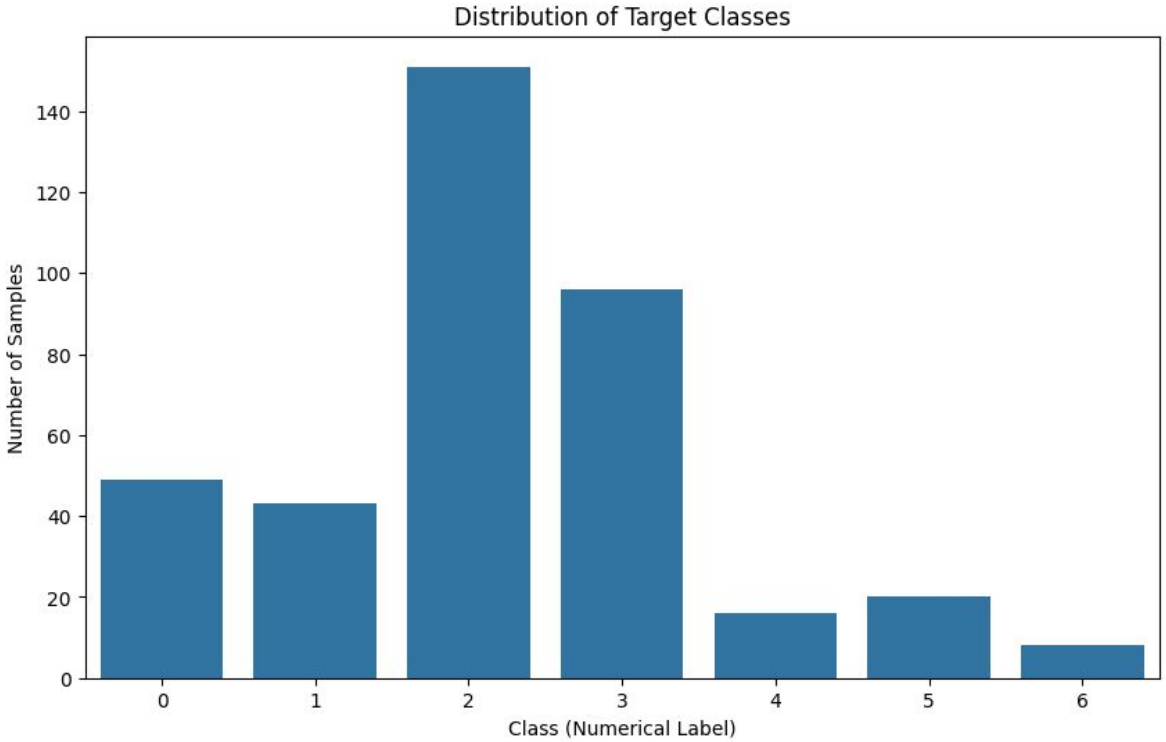
K-Nearest Neighbors

```
Best Hyperparameters:  {'n_neighbors': 19, 'p': 1, 'weights': 'uniform'}

Accuracy:  0.5208333333333334
AUROC (OvR):  0.8374711773097282
              precision    recall  f1-score   support

           0       0.92      0.92      0.92        12
           1       0.00      0.00      0.00        11
           2       0.53      0.79      0.63        38
           3       0.35      0.38      0.36        24
           4       0.00      0.00      0.00         4
           5       0.00      0.00      0.00         5
           6       0.00      0.00      0.00         2

    accuracy                           0.52        96
   macro avg       0.26      0.30      0.27        96
weighted avg       0.41      0.52      0.45        96
```



- Note: Poor precision and recall for labels 1, 4, 5, and 6

CAIS++
CENTER FOR AI IN SOCIETY'S STUDENT BRANCH



- Class imbalance
- Possible explanation for poor precision and recall for labels 4, 5, and 6
- Fewer instances of 4, 5, 6 because later stages less common and maybe more complex (?)

CAIS++
CENTER FOR AI IN SOCIETY'S STUDENT BRANCH

Results

Voting

```
Best Hyperparameters for Voting Classifier:  {'dt__criterion':
Accuracy:  0.6354166666666666
Voting Classifier Report:
              precision    recall  f1-score   support

           0       0.92      0.92      0.92        12
           1       0.00      0.00      0.00        11
           2       0.67      0.76      0.72        38
           3       0.50      0.71      0.59        24
           4       0.00      0.00      0.00         4
           5       0.50      0.60      0.55         5
           6       1.00      0.50      0.67         2

    accuracy                           0.64        96
   macro avg       0.51      0.50      0.49        96
weighted avg       0.55      0.64      0.59        96
```

```
Best Hyperparameters:
'dt  criterion': 'entropy',
'dt__max_depth': 3,
'dt  min samples split': 2,
'knn__n_neighbors': 3,
'knn  p': 1,
'knn  weights': 'uniform',
'voting': 'soft',
'weights': [1, 2]
```

- Used Knn and decision trees as base models
- greater weight given to decision trees -> confident that decision tree reason for improvement
- Poor precision and recall still; Class imbalance problem persists

CAIS++
CENTER FOR AI IN SOCIETY'S STUDENT BRANCH

Results

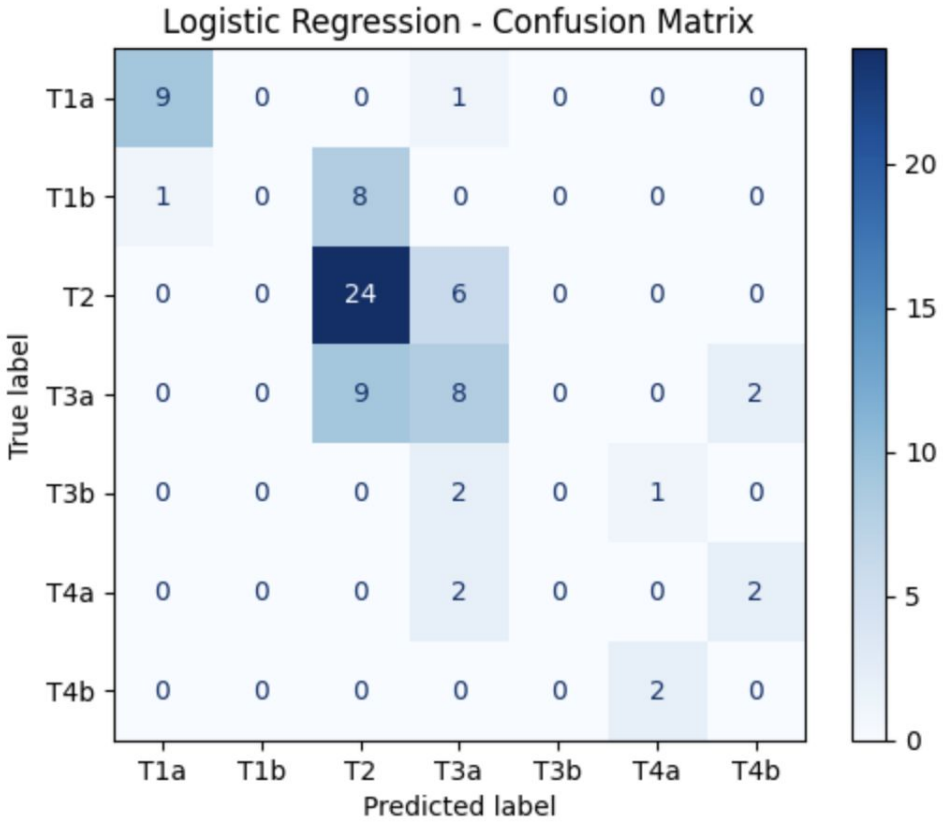
Logistic Regression

```
Accuracy 0.5324675324675324
Classification Report
              precision    recall  f1-score   support

           0       0.90      0.90      0.90        10
           1       0.00      0.00      0.00         9
           2       0.59      0.80      0.68        30
           3       0.42      0.42      0.42        19
           4       0.00      0.00      0.00         3
           5       0.00      0.00      0.00         4
           6       0.00      0.00      0.00         2

    accuracy                           0.53        77
   macro avg       0.27      0.30      0.29        77
weighted avg       0.45      0.53      0.48        77
```

CAIS++
CENTER FOR AI IN SOCIETY'S STUDENT BRANCH

Results

Logistic Regression



CAIS++
CENTER FOR AI IN SOCIETY'S STUDENT BRANCH

Results

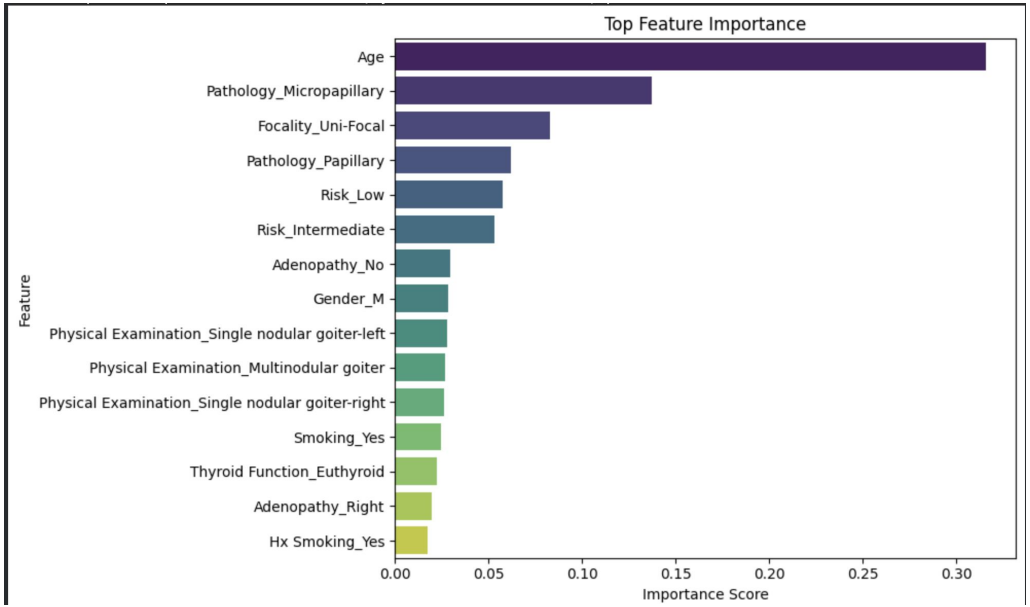
Random Forest

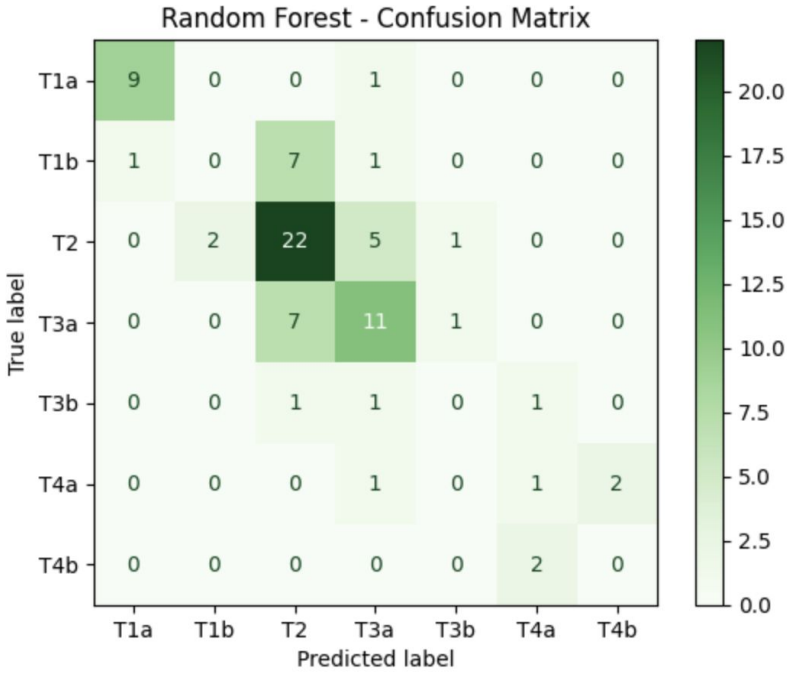
```
Grid Search Best Params {'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 100}
Grid Search Best Score 0.6565309360126917
Accuracy: 0.6233766233766234
Classification Report
              precision    recall  f1-score   support

           0       0.90      0.90      0.90        10
           1       0.00      0.00      0.00         9
           2       0.62      0.83      0.71        30
           3       0.55      0.63      0.59        19
           4       0.00      0.00      0.00         3
           5       0.50      0.50      0.50         4
           6       0.00      0.00      0.00         2

    accuracy                           0.62        77
   macro avg       0.37      0.41      0.39        77
weighted avg       0.52      0.62      0.57        77
```

CAIS++
CENTER FOR AI IN SOCIETY'S STUDENT BRANCH

Results

Random Forest





CAIS++
CENTER FOR AI IN SOCIETY'S STUDENT BRANCH

Next Steps / Limitations

- Address class imbalance
  - Would try to get more data samples, e.g., using the Synthetic Minority Over-Sampling Technique (SMOTE)
- Experiment with other base models for ensemble models