Classifying Leukemia Types Using Machine Learning

Group 17: Amanda Bell, Tim Wang, Jasmine Wong

The Problem: Classifying Leukemia Types

- American Cancer Society estimates that in 2023, almost 60,000 new cases of Leukemia and 23,710 deaths from Leukemia will occur
- 2 types of Leukemia we're looking at:
  - **Acute Myeloid Leukemia (AML)**
    - **Rapid growth of abnormal cells crowd out normal cells in bone marrow and bloodstream**
    - **More common in older adults**
  - **Acute Lymphocytic Leukemia (ALL)**
    - **Rapid growth of abnormal lymphoblasts creates immature white blood cells that are unable to fight infection**
- Chronic Myeloid Leukemia (CML) and Chronic Lymphocytic Leukemia (CLL) are two other types
- Difficult to classify types by symptoms alone
- Can we classify certain types of Leukemia using genetics?
  - With machine learning models?

About Our Features

Features

- Gene Description + Accession Number
- Numbers for each patient - values for gene expression
- Call for each gene for a patient
  - Absent
  - Present
  - Marginal

[Kaggle Gene Expression Dataset (Golub et al).]

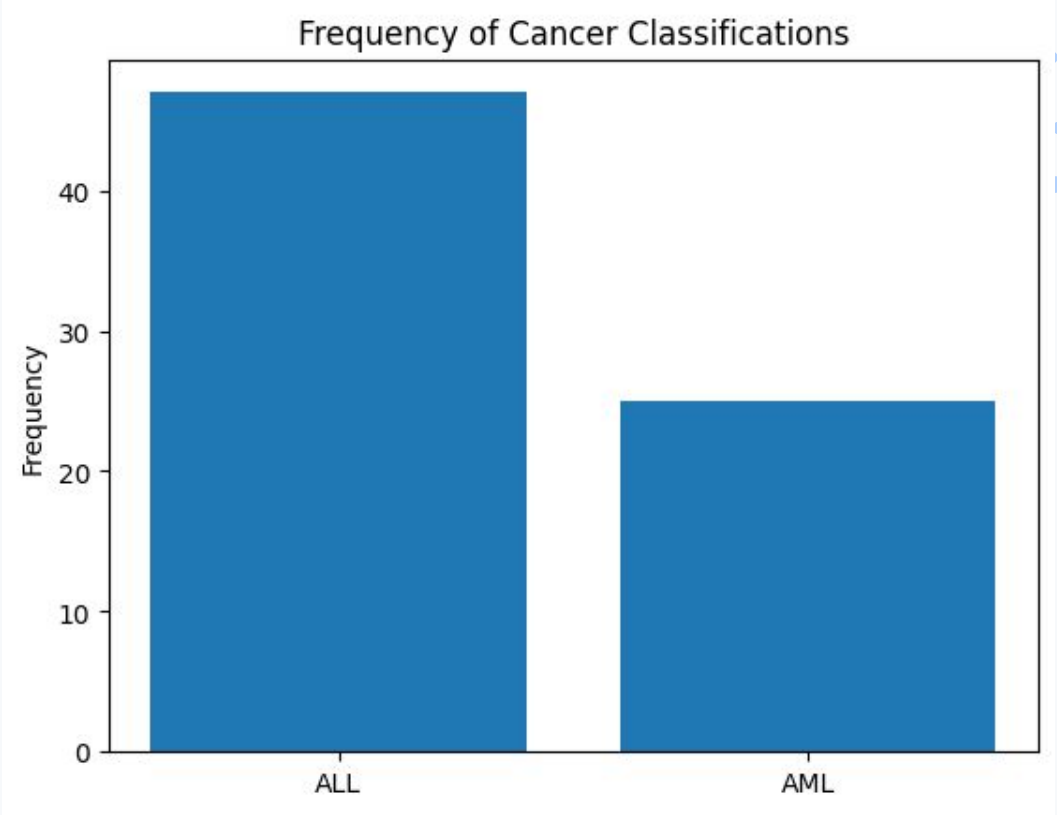
DNA Microarray Data

	Gene Description	Gene Accession Number	1	call	2	call.1	3	call.2	4	call.3
0	AFFX-BioB-5_at (endogenous control)	AFFX-BioB-5_at	-214	A	-139	A	-76	A	-135	A
1	AFFX-BioB-M_at (endogenous control)	AFFX-BioB-M_at	-153	A	-73	A	-49	A	-114	A
2	AFFX-BioB-3_at (endogenous control)	AFFX-BioB-3_at	-58	A	-1	A	-307	A	265	A

About Our Targets

Output

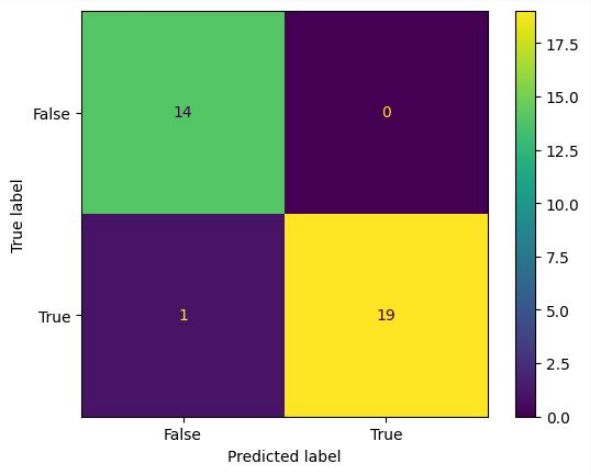
- Predict ALL or AML
- 47 ALL vs 25 AML
- More ALL Data
  - ALL is more common in older adults while AML more commonly affects children

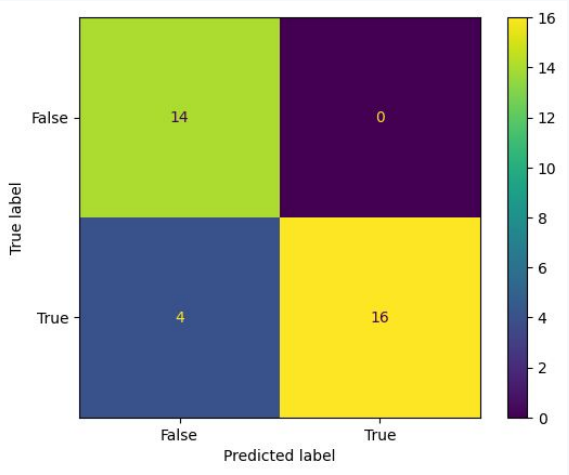


Results

Logistic Regression

- With full feature list, high accuracy (97%)
  - Only 1 misclassification
- With reduced set of features, accuracy drops (88%)
  - 4 false negatives, predicted AML but should be ALL
- Model already handled 7000+ features well, so perhaps using the reduced feature dataset resulted in a loss of some important features

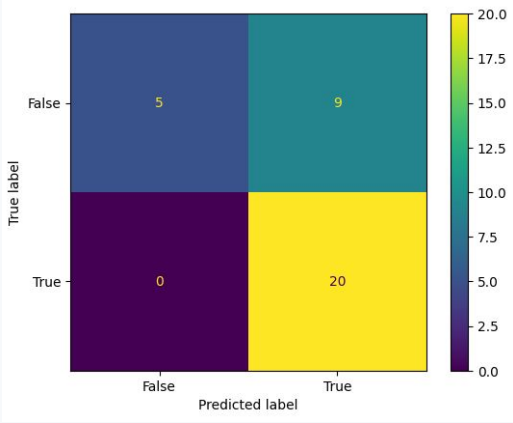




Results

Random Forest Classification

- With full feature list
  - Ok accuracy on train data but not great (76.5%)
  - Same with F1: 0.833
    - Issue with false positives, predicted ALL but should be AML
- After hyperparameter tuning
  - min sample leaf: 2 | min samples split: 9 | n_estimators: 100
  - Accuracy and F1 dropped a bit on test data?
- Before and after hyperparameter tuning, accuracy and F1 were very high on train data -> overfitting -> too many features or not enough of the other label (AML)
- Feature importance comparison to real life
  - A couple are somewhat linked to AML
  - Others seem less specific to leukemia but definitely can cause dysregulated cell processes with mutation
  - Not the genes you would see if you looked up specific ones for AML and ALL

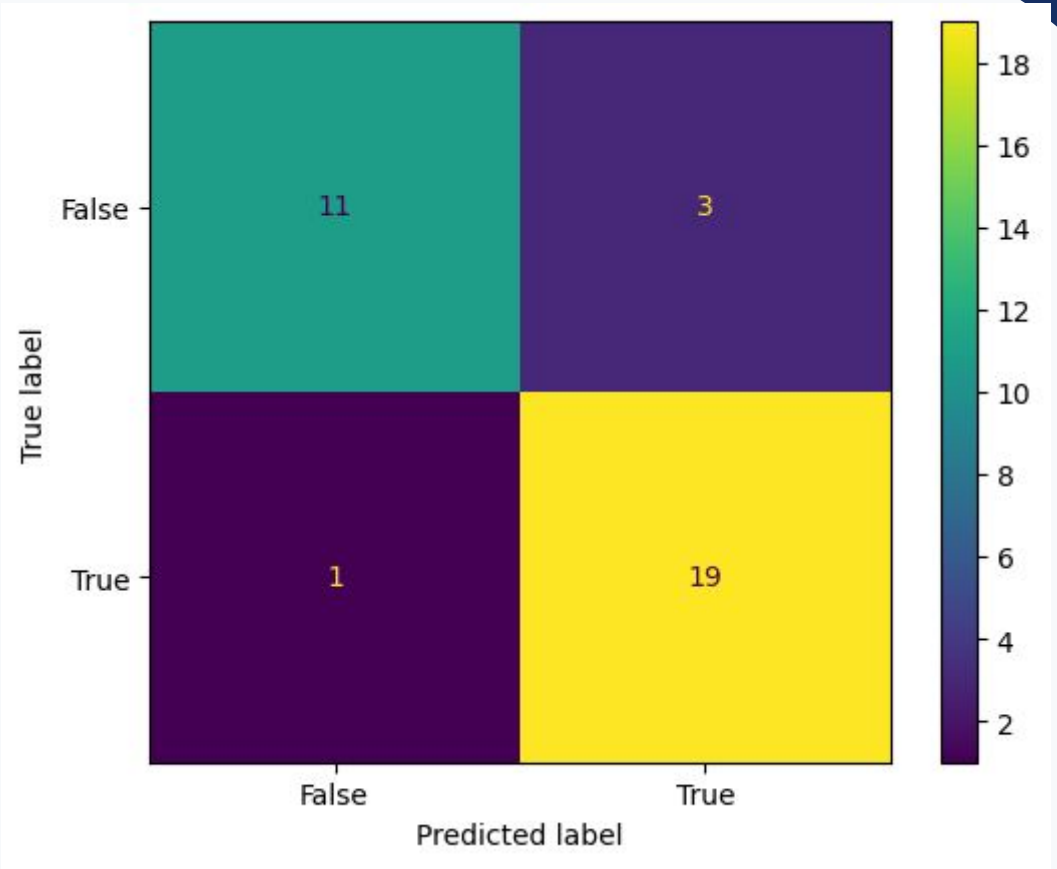


	feature	importance
2287	DF D component of complement (adipsin)	0.027093
1119	SNRPN Small nuclear ribonucleoprotein polypept...	0.024451
1806	Neuromedin B mRNA	0.020000
1143	SPTAN1 Spectrin; alpha; non-erythrocytic 1 (al...	0.020000
5898	Rhesus (Rh) Blood Group System Ce-Antigen; Alt...	0.018851
234	KIAA0022 gene	0.017618
5038	LEPR Leptin receptor	0.017276
2885	Transmembrane protein mRNA	0.017112
3257	Phosphotyrosine independent ligand p62 for the...	0.013271
1845	CTRB1 Chymotrypsinogen B1	0.011900

Results

KNN Classification

- Used GridSearch to hyperparameter tune
- Found that best neighbors = 4
- Testing Accuracy: 88.24%
- 3 False Positives, meaning predicted ALL for AML



Results

PCA and PCR

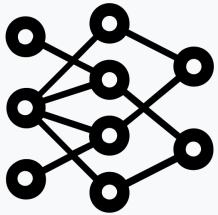
- Conducted cross-validated PCA on training set
  - Best regression model: 35 principal components
  - Mean squared error: 0.04148
- Tested PCR model using test set, with the following results
  - Mean squared error: 0.06576

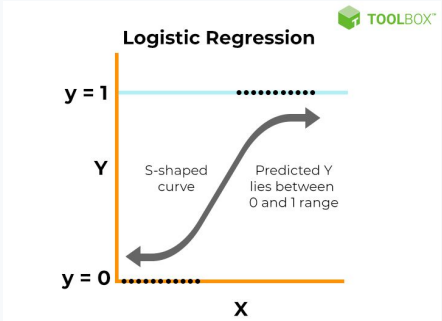
Next Steps

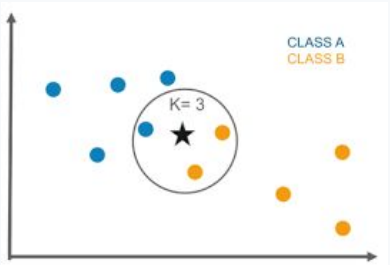
- Try a feature selection method for Random Forest Classifier
- Test algorithms on a subset of genes commonly associated with ALL and AML
    - Are they useful in differentiating between the two when not considering all other features?
- Finish coding neural network
- Finish writing documentation for GitHub repo

Our Approaches & Why We Chose Them

- Neural Network
- Logistic regression
- PCA w/ GridsearchCV
- KNN







Timeline

Date	Task Completed
3/25	Clean Data
4/1	Execute PCA, Logistic, KNN
4/7	Execute Neural Network
4/9	Finish Code Cleaning & Documentation
4/10	Start Report
4/12	Write Report
4/13	Presentation!