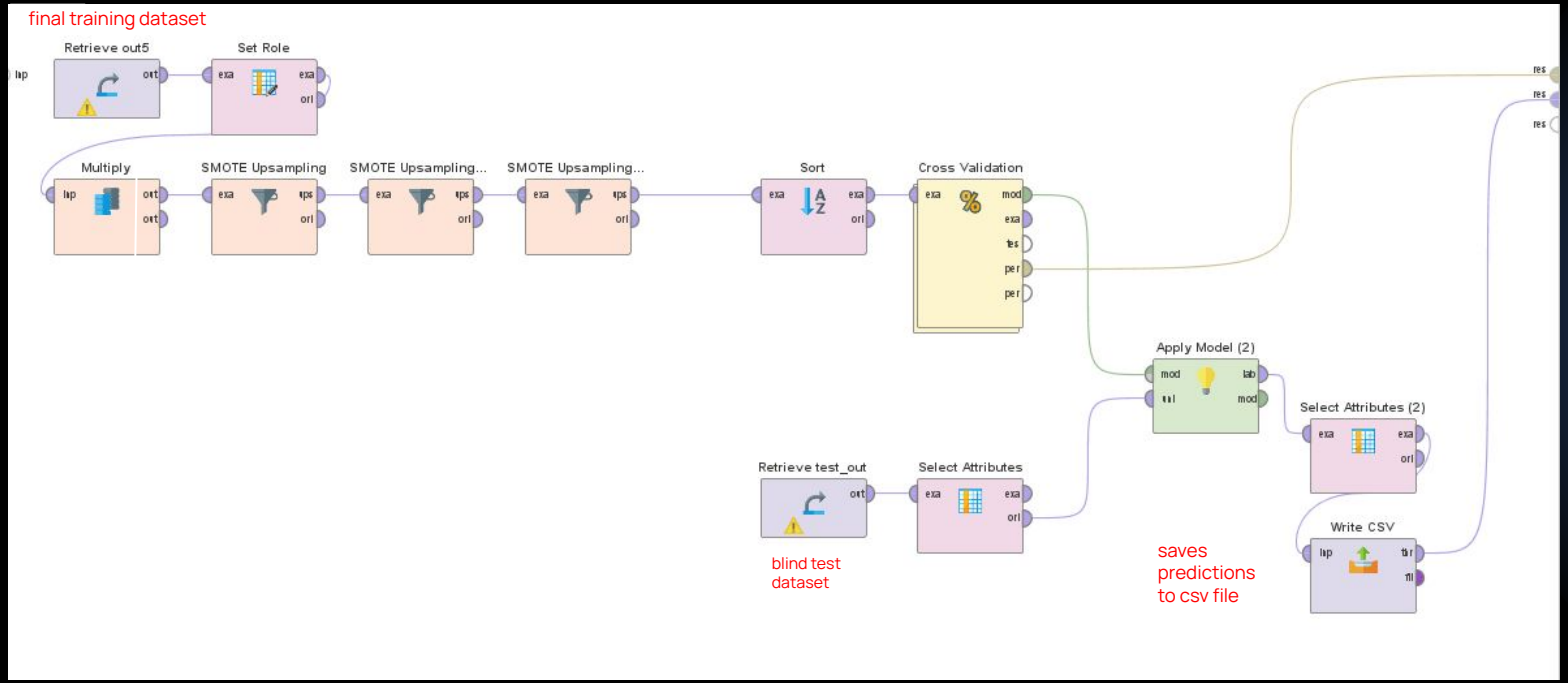
DNA Sequence Analysis

Jordan Dube, Luke Unterman, Blue Arevalo, Kiersten Kofi Adomfrimpong

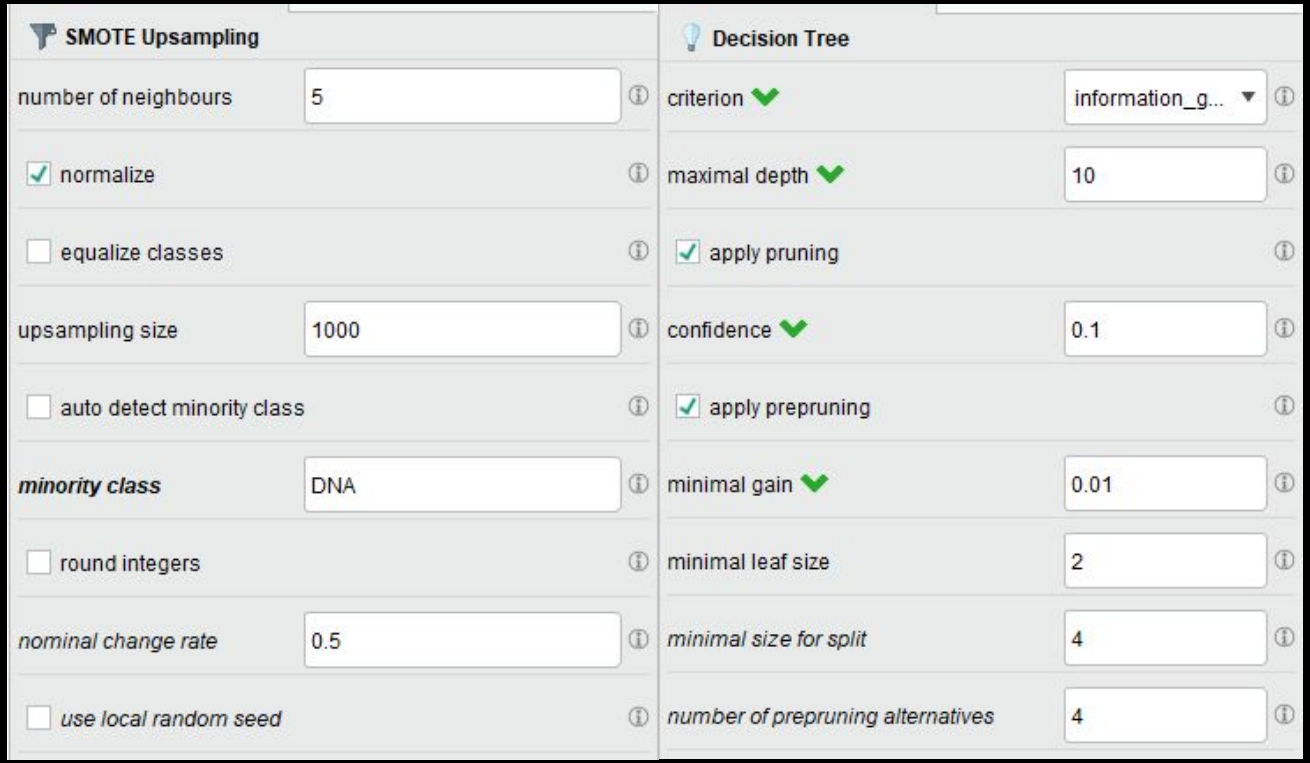
Motivation

- Generated multiple datasets from Pfeature for our model
- Noticed that average MCC values were still low as a result of an overrepresentation of nonDRNA data
- Used SMOTE oversampling operator as an attempt to correct underrepresentation of minority classes
- Refined model to reduce overfitting

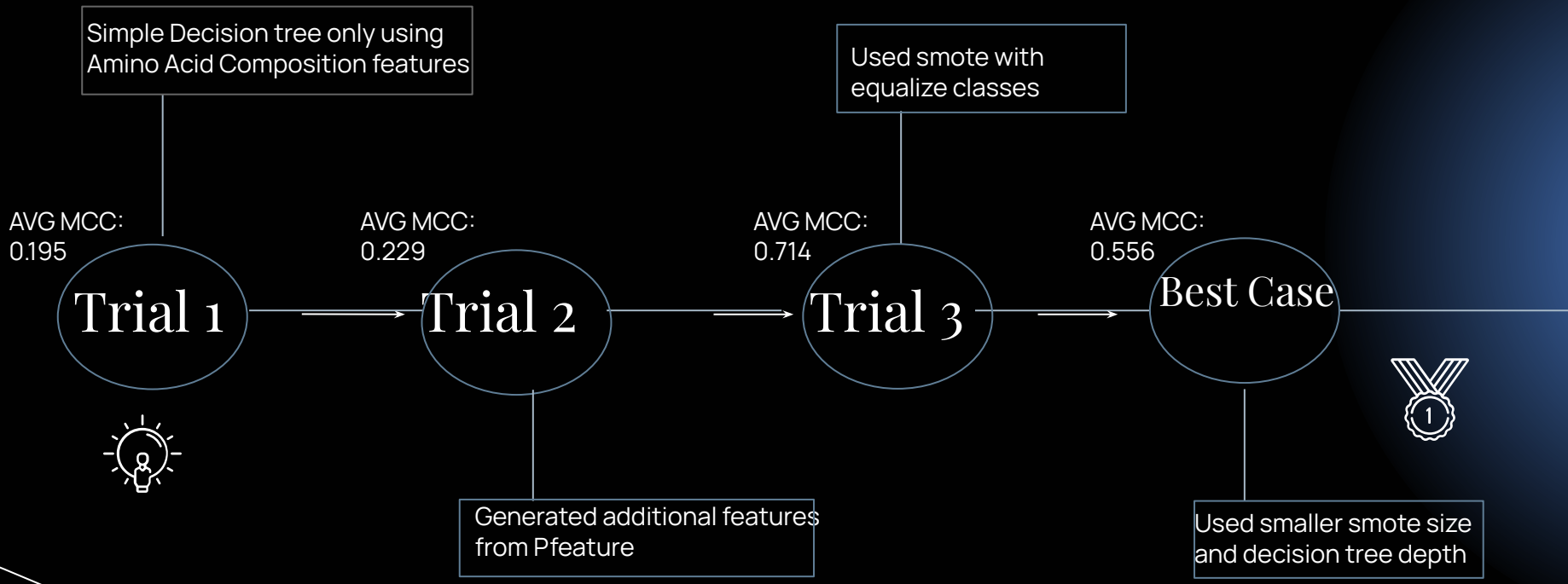
Diagram of Best Model



Final Parameters



Changes in iterations



Results

Class	Absolute Count (test set, training set)	Representation in Dataset (test set, training set)
nonDRNA	7558, 7859	85.9%, 89.4%
RNA	594, 523	6.8%, 5.9%
DNA	542, 391	6.2%, 4.4%
DRNA	100, 22	1.1%, 0.3%

Outcome	Quality measure	Baseline result	Design 1	Design 2	Design 3	Best Design
	Sensitivity	6.9%	16.4%	17.6%	93.9%	57.3%
	Specificity	99.3%	96.4%	96.4%	96.0%	97.0%
	Accuracy	95.2%	92.8%	92.9%	95.9%	95.3%
DNA	***MCC***	**0.132**	**0.131**	**0.143**	**0.681**	**0.496**
	Sensitivity	39.6%	34.6%	42.6%	94.5%	71.1%
	Specificity	98.9%	96.9%	97.1%	97.7%	98.4%
	Accuracy	95.3%	93.2%	93.9%	97.5%	96.8%
RNA	***MCC***	**0.501**	**0.343**	**0.421**	**0.814**	**0.706**
	Sensitivity	4.5%	0.0%	0.0%	81.8%	59.1%
	Specificity	100%	99.8%	99.9%	99.8%	99.6%
	Accuracy	99.7%	99.6%	99.6%	99.7%	99.5%
DRNA	***MCC***	**0.122**	**-0.002**	**-0.002**	**0.613**	**0.398**
	Sensitivity	98.6%	93.8%	94.0%	93.3%	95.2%
	Specificity	29.8%	35.4%	40.0%	96.2%	69.9%
	Accuracy	91.3%	87.6%	88.2%	93.6%	92.5%
nonDRNA	***MCC***	**0.428**	**0.309**	**0.354**	**0.748**	**0.625**
averageMCC		**0.296**	**0.195**	**0.229**	**0.714**	**0.556**
accuracy4labels		90.8%	86.6%	87.3%	93.3%	92.0%

Conclusion - Part 1

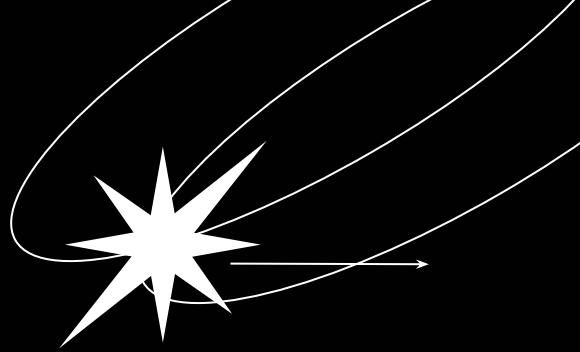
Advantages

- Uses SMOTE to oversample minority classes
- Uses a multitude of features generated from Pfeature
- Decision Tree parameters have been fine tuned

Disadvantages

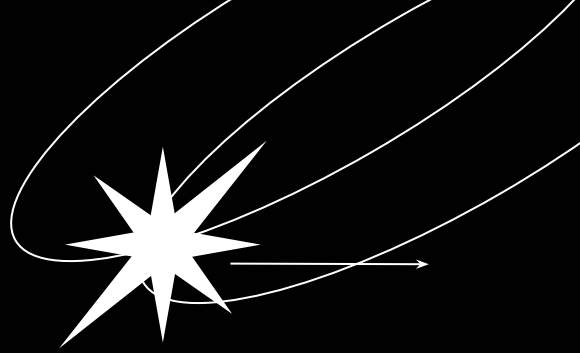
- model uses the somewhat arbitrary feature selection from Pfeature
- a small number of classification models have been explored
- Possible overfitting

Experience - Concl. Pt 2



- Ultimately pleased with results from project
- Adjusted for overfitting, underrepresentation, and noise in our model
- Results steadily improved as a result of generating more features and using SMOTE
- Confident that we came close to desired results without facing oversampling
- Project was positive experience and taught us to process, generate, and analyze real-world data
- Helped prepare us for jobs in Data Science

Q/A



Any questions?