Final Project Presentation

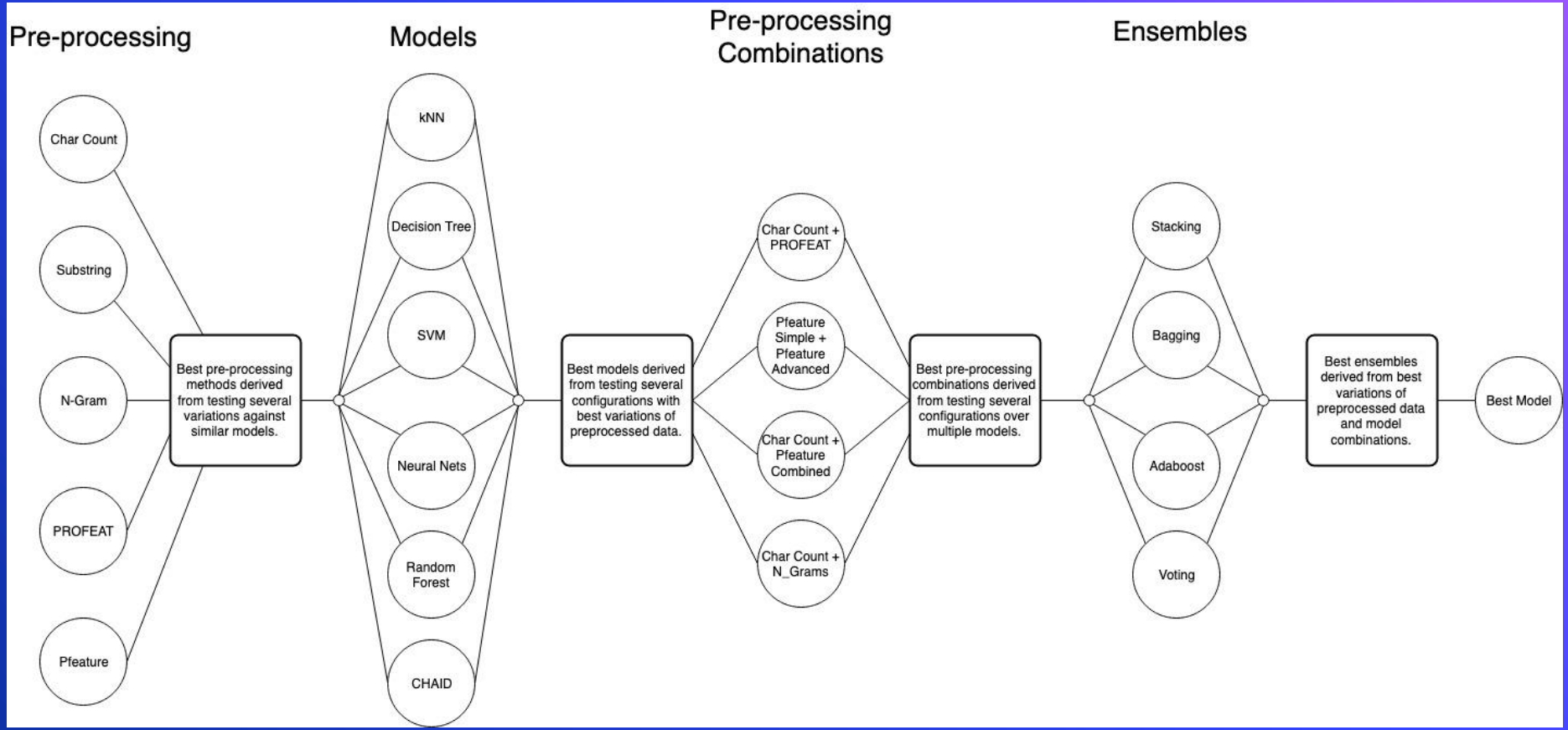
Group: *Chain Gang*
Members: Gavin, Mike P, Mike F, Jacob, Angel

1. Design Motivation

What did our design process look like?

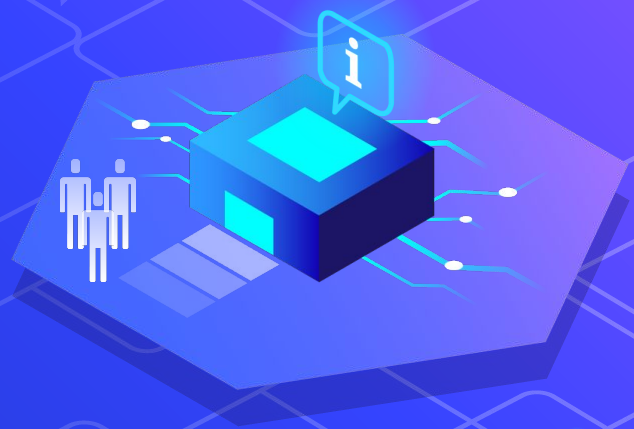
Motivation

As a team we decided to embrace the iterative approach asked of us in this project. We started by building simplistic preprocessing methods and comparing them against basic models in order to build up a solid preprocessing foundation. Then we began to swap out model types to see where we could get big performance increases. Once our team decided on a single best preprocessing method and model, we began to hybridize our preprocessing methods in order to boost performance. After this step we switched to ensemble learners which allowed us to achieve our best results. At each of these checkpoints a leading design was selected and defined. This process is depicted in the next slide.


Pre-processing
Models
Pre-processing Combinations
Ensembles
Char Count
Substring
N-Gram
PROFEAT
Pfeature
Best pre-processing methods derived from testing several variations against similar models.
kNN
Decision Tree
SVM
Neural Nets
Random Forest
CHAID
Best models derived from testing several configurations with best variations of preprocessed data.
Char Count + PROFEAT
Pfeature Simple + Pfeature Advanced
Char Count + Pfeature Combined
Char Count + N_Grams
Best pre-processing combinations derived from testing several configurations over multiple models.
Stacking
Bagging
Adaboost
Voting
Best ensembles derived from best variations of preprocessed data and model combinations.
Best Model

2. Design Description

What is the final design?

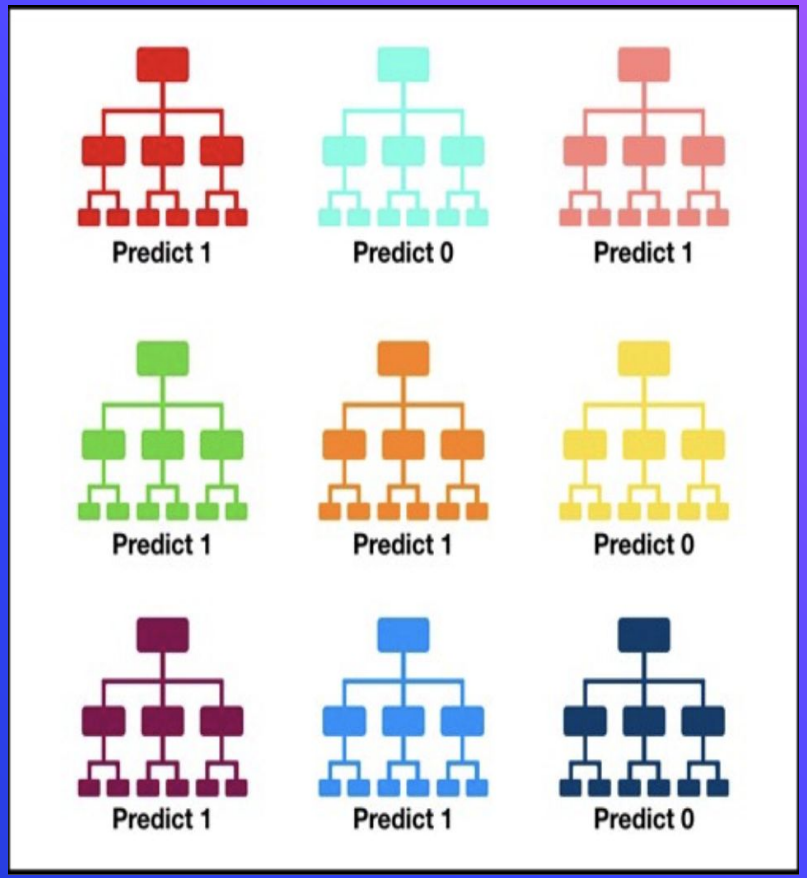


Design Description

As a team, we chose the ensemble technique of the random forest classifier.

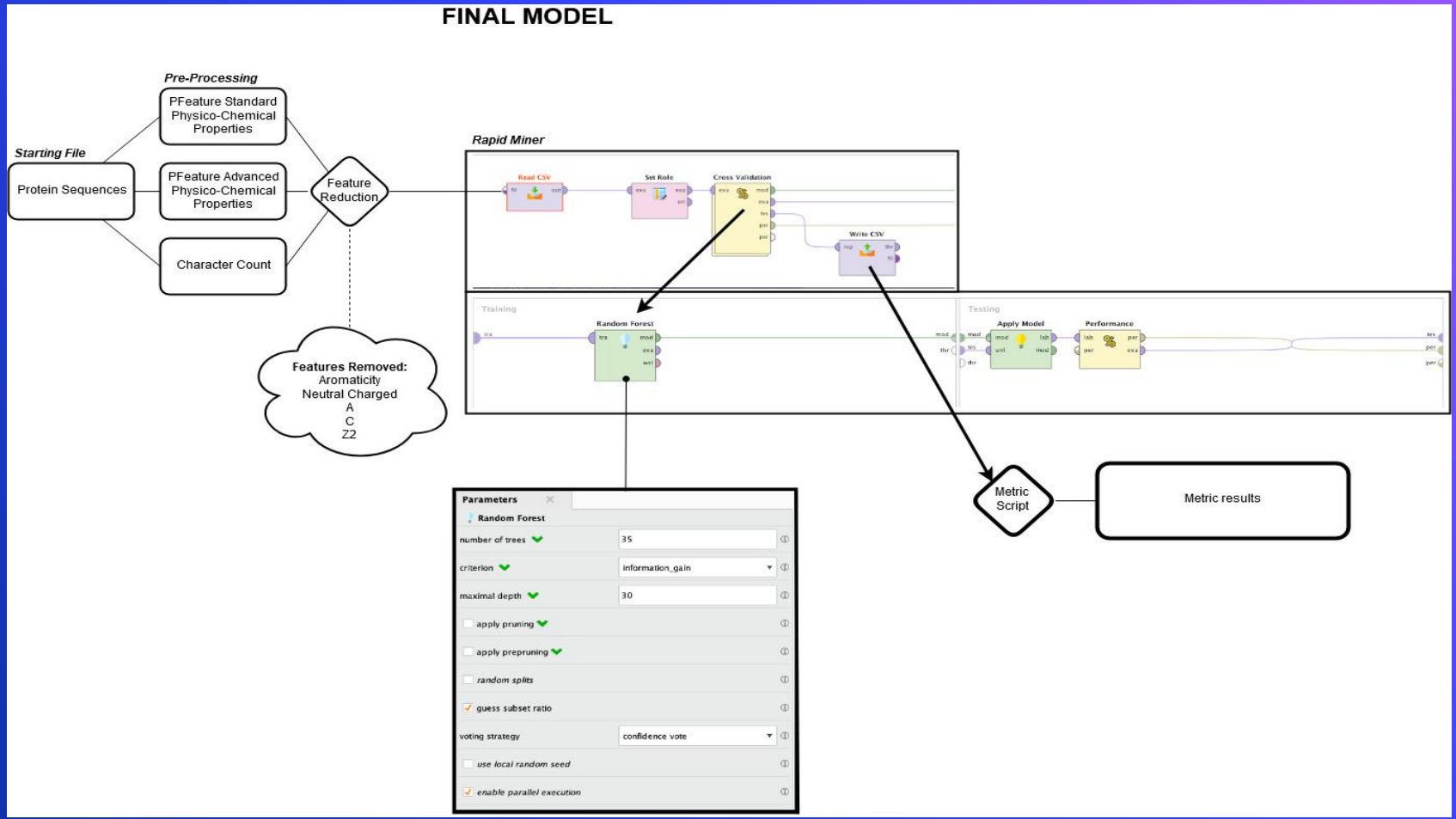
What it is:
The random forest classifier takes a set of individual decision trees that produce a class prediction and classifies according to a majority vote.

Why we chose it:
- Easily understood.
- Fast run time.
- Consistent results.



Source: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

FINAL MODEL


Pre-Processing
PFeature Standard Physico-Chemical Properties
Starting File
Protein Sequences
PFeature Advanced Physico-Chemical Properties
Character Count
Feature Reduction
Features Removed:
Aromaticity
Neutral Charged
A
C
Z2
Rapid Miner
Read CSV
Set Role
Cross Validation
Write CSV
Training
Random Forest
Testing
Apply Model
Performance
Parameters
Random Forest
number of trees
35
criterion
information_gain
maximal depth
30
apply pruning
apply prepruning
random splits
guess subset ratio
voting strategy
confidence vote
use local random seed
enable parallel execution
Metric Script
Metric results

3. Discussion and Comparison

How did we do?

Outcome	Quality Measure	Baseline Result	Design 1	Design 2	Design 3	Design 4	Best Design
DNA	Sensitivity	6.9	23.08	7.87	55.0	80.0	78.0
	Specificity	99.3	95.29	96.29	95.79	95.94	95.97
	Predictive ACC	95.2	95.17	84.84	95.60	95.86	95.87
	MCC	0.132	0.07	0.06	0.17	.26	.27
RNA	Sensitivity	39.6	45.49	11.69	77.58	80.89	78.83
	Specificity	98.9	95.29	96.29	96.49	96.21	96.40
	Predictive ACC	95.3	93.78	72.60	95.87	95.78	95.85
	MCC	0.501	0.30	0.15	0.56	.54	.55
DRNA	Sensitivity	4.5	14.29	50.00	100.0	66.67	100.0
	Specificity	100	99.76	99.76	99.76	99.77	99.77
	Predictive ACC	99.7	99.69	99.75	99.76	99.76	99.77
	MCC	0.122	0.08	0.15	0.21	.25	.30
nonDRNA	Sensitivity	98.6	91.01	92.13	92.13	91.98	92.22
	Specificity	29.8	51.78	14.68	81.57	86.39	84.97
	Predictive ACC	91.3	89.50	60.58	91.73	91.79	91.95
	MCC	0.428	0.27	0.11	0.45	.46	.47
Average MCC		0.265	0.18	0.12	0.35	.38	.40
Accuracy		90.8	89.07	58.89	91.48	91.60	91.72

4. Conclusions

What does it mean?

- **Advantages**
  - Easily reproducible.
  - Easily explainable processes.
  - Quality performance.

- **Disadvantages**
  - The key disadvantage of our model is the use of web servers to generate feature sets.
    - We have no control over when these servers go offline.
    - We don’t control the traffic of the webserver and can get slow response times when processing data.
    - Some of the features provided by these web servers have no description nor an explanation on how they are computed.

- **Summary**

  Overall, with both pros and cons considered, our team still feels that we have created a competent system. Our results point to a better predictive model not only overall but in each individual class as well. This gives us a strong indication that we are performing well.

Questions?