Using Feature Selection to Identify Gene Significance in Drug-Resistant Tuberculosis

Alan Qi and Powell Zhang

Overview

1

Overview

- Multidrug resistant tuberculosis is a form of tuberculosis that has developed resistance to *isoniazid* and *rifampin.*
- We worked with a binary dataset of resistant and susceptible tuberculosis strains.
- We utilized feature selection to determine significant genes with four models: CART, random forest, naive bayes, and genetic learning.
- We found pncA gene was most significant in determining drug resistance.
- These methods can be applied to other similar datasets in the future

Introduction

2

Tuberculosis (TB)

- Caused by *Mycobacterium tuberculosis*
- Can be cured by administering antibiotics such as *Isoniazid*, *Rifampin*, *Pyrazinamide*, and others
- Can develop resistance to certain antibiotics, making it very difficult to cure
- Approximately 50% of patients with drug resistant tuberculosis eventually die due to the disease
- 240,000 global deaths from drug resistant tuberculosis in 2017
- 4.1% of new cases of TB and 19% of previously treated cases are drug resistant

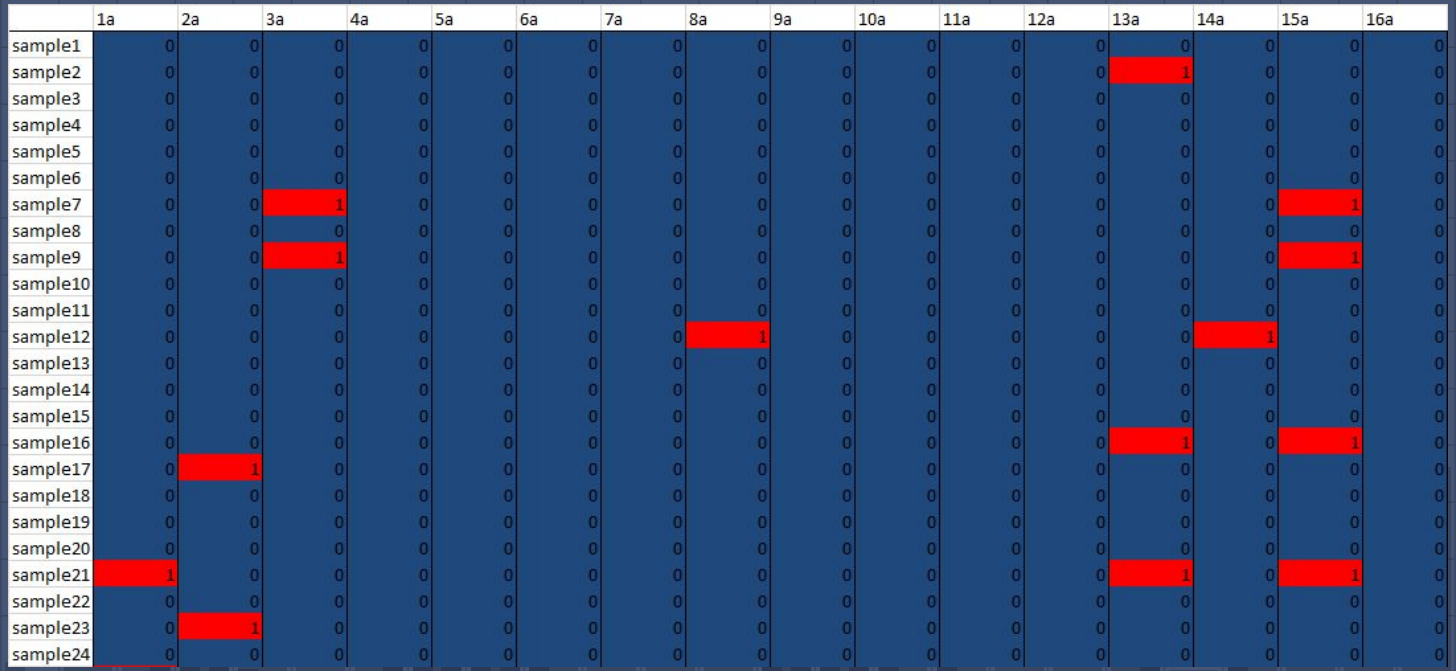
Goals of project

- Use feature analysis to determine genes significant in determining drug resistance
- Use results to screen for drug resistance and prescribe effective antibiotics
- Expand to other diseases

Data

Biomarkers

Strains



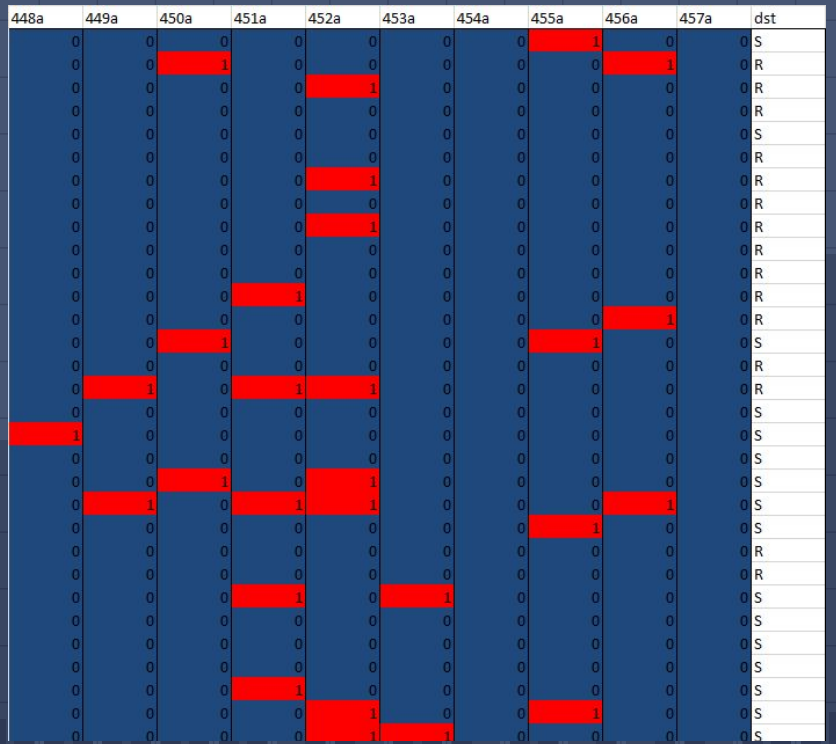
	1a	2a	3a	4a	5a	6a	7a	8a	9a	10a	11a	12a	13a	14a	15a	16a
sample1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
sample3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample7	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
sample8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample9	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
sample10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample12	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
sample13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample16	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
sample17	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample21	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
sample22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample23	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

480 strains of tuberculosis including drug resistant and susceptible strains

457 biomarkers showing mutated and wild type genes

Data

Each sample was labeled with an “R” or and “S” to show if it was resistant to drugs or susceptible



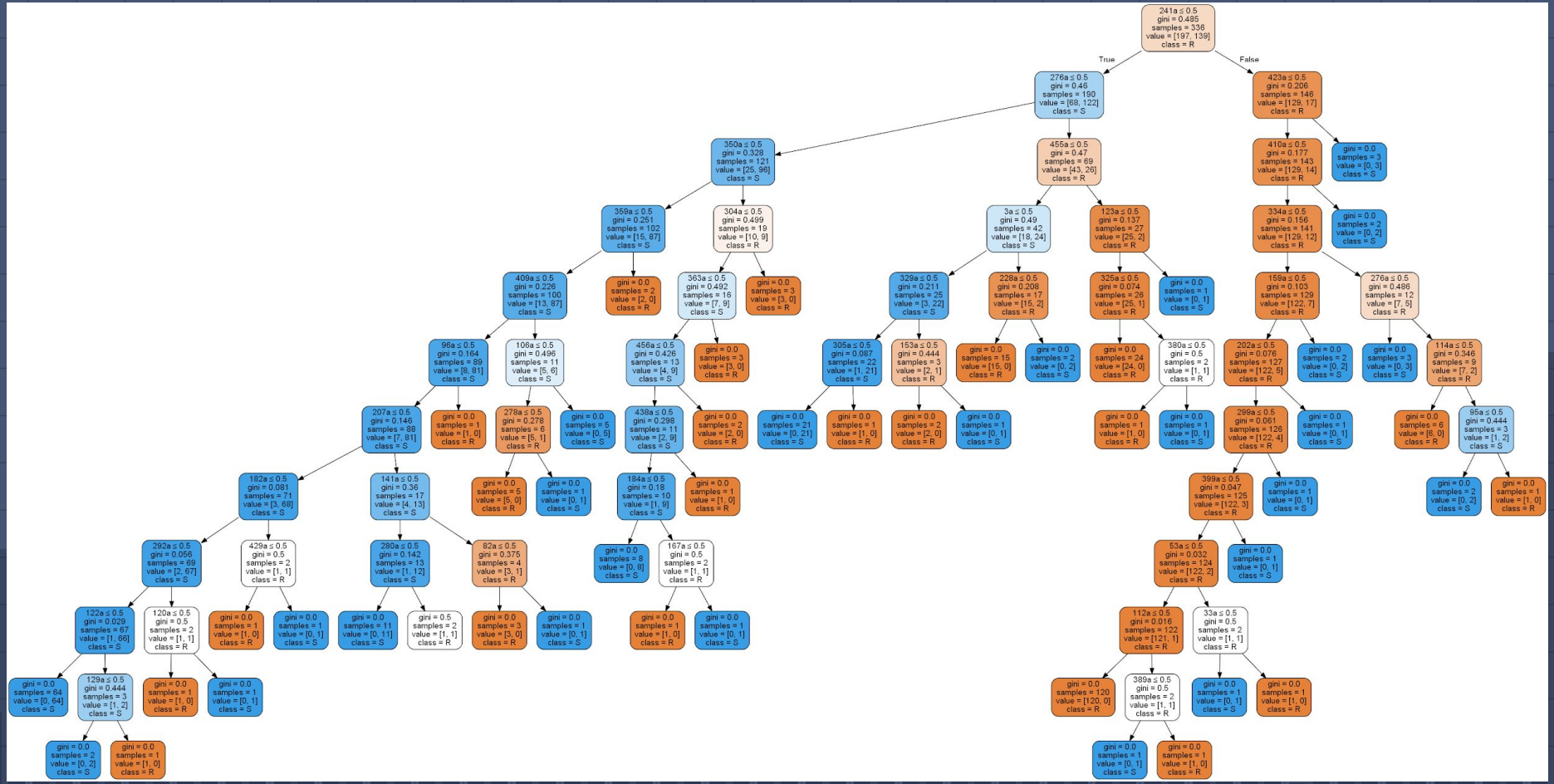
448a	449a	450a	451a	452a	453a	454a	455a	456a	457a	dst
0	0	0	0	0	0	0	1	0	0	S
0	0	1	0	0	0	0	0	1	0	R
0	0	0	0	1	0	0	0	0	0	R
0	0	0	0	0	0	0	0	0	0	R
0	0	0	0	0	0	0	0	0	0	S
0	0	0	0	0	0	0	0	0	0	R
0	0	0	0	1	0	0	0	0	0	R
0	0	0	0	0	0	0	0	0	0	R
0	0	0	0	1	0	0	0	0	0	R
0	0	0	0	0	0	0	0	0	0	R
0	0	0	0	0	0	0	0	0	0	R
0	0	0	1	0	0	0	0	0	0	R
0	0	0	0	0	0	0	0	1	0	R
0	0	1	0	0	0	0	1	0	0	S
0	0	0	0	0	0	0	0	0	0	R
0	1	0	1	1	0	0	0	0	0	R
0	0	0	0	0	0	0	0	0	0	S
1	0	0	0	0	0	0	0	0	0	S
0	0	0	0	0	0	0	0	0	0	S
0	0	1	0	1	0	0	0	0	0	S
0	1	0	1	1	0	0	0	1	0	S
0	0	0	0	0	0	0	1	0	0	S
0	0	0	0	0	0	0	0	0	0	R
0	0	0	0	0	0	0	0	0	0	R
0	0	0	1	0	1	0	0	0	0	S
0	0	0	0	0	0	0	0	0	0	S
0	0	0	0	0	0	0	0	0	0	S
0	0	0	0	0	0	0	0	0	0	S
0	0	0	1	0	0	0	0	0	0	S
0	0	0	0	1	0	0	1	0	0	S
0	0	0	0	1	1	0	0	0	0	S

Methods

3

CART (Classification and regression trees)

- Commonly known as decision trees
- Samples are divided based on certain characteristics using a binary tree
- Important genes are determined by looking at important leaf nodes

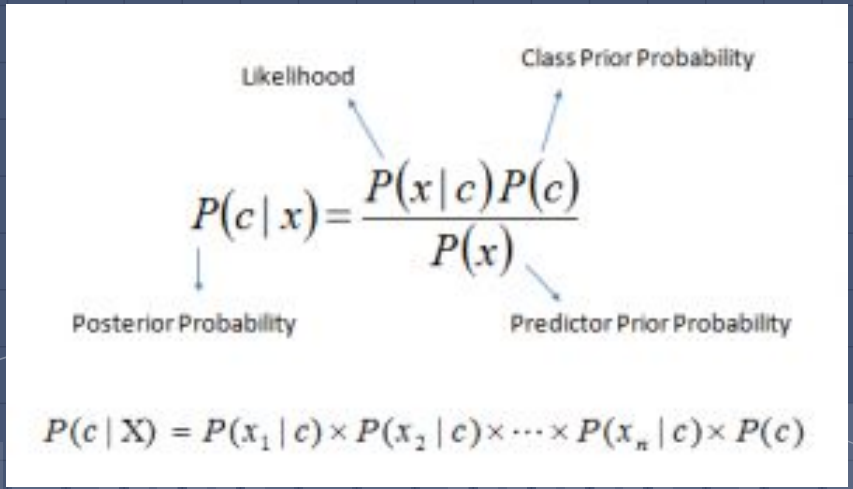

241a ≤ 0.5
gini = 0.485
samples = 336
value = [197, 139]
class = R
True
False
276a ≤ 0.5
gini = 0.46
samples = 190
value = [68, 122]
class = S
423a ≤ 0.5
gini = 0.206
samples = 146
value = [129, 17]
class = R
350a ≤ 0.5
gini = 0.328
samples = 121
value = [25, 96]
class = S
455a ≤ 0.5
gini = 0.47
samples = 69
value = [43, 26]
class = R
410a ≤ 0.5
gini = 0.177
samples = 143
value = [129, 14]
class = R
gini = 0.0
samples = 3
value = [0, 3]
class = S
359a ≤ 0.5
gini = 0.251
samples = 102
value = [15, 87]
class = S
304a ≤ 0.5
gini = 0.499
samples = 19
value = [10, 9]
class = R
3a ≤ 0.5
gini = 0.49
samples = 42
value = [18, 24]
class = S
123a ≤ 0.5
gini = 0.137
samples = 27
value = [25, 2]
class = R
334a ≤ 0.5
gini = 0.156
samples = 141
value = [129, 12]
class = R
gini = 0.0
samples = 2
value = [0, 2]
class = S
409a ≤ 0.5
gini = 0.226
samples = 100
value = [13, 87]
class = S
gini = 0.0
samples = 2
value = [2, 0]
class = R
363a ≤ 0.5
gini = 0.492
samples = 16
value = [7, 9]
class = S
gini = 0.0
samples = 3
value = [3, 0]
class = R
329a ≤ 0.5
gini = 0.211
samples = 25
value = [3, 22]
class = S
228a ≤ 0.5
gini = 0.208
samples = 17
value = [15, 2]
class = R
325a ≤ 0.5
gini = 0.074
samples = 26
value = [25, 1]
class = R
gini = 0.0
samples = 1
value = [0, 1]
class = S
159a ≤ 0.5
gini = 0.103
samples = 129
value = [122, 7]
class = R
276a ≤ 0.5
gini = 0.486
samples = 12
value = [7, 5]
class = R
96a ≤ 0.5
gini = 0.164
samples = 89
value = [8, 81]
class = S
106a ≤ 0.5
gini = 0.496
samples = 11
value = [5, 6]
class = S
456a ≤ 0.5
gini = 0.426
samples = 13
value = [4, 9]
class = S
gini = 0.0
samples = 3
value = [3, 0]
class = R
305a ≤ 0.5
gini = 0.087
samples = 22
value = [1, 21]
class = S
153a ≤ 0.5
gini = 0.444
samples = 3
value = [2, 1]
class = R
gini = 0.0
samples = 15
value = [15, 0]
class = R
gini = 0.0
samples = 2
value = [0, 2]
class = S
gini = 0.0
samples = 24
value = [24, 0]
class = R
380a ≤ 0.5
gini = 0.5
samples = 2
value = [1, 1]
class = R
202a ≤ 0.5
gini = 0.076
samples = 127
value = [122, 5]
class = R
gini = 0.0
samples = 2
value = [0, 2]
class = S
gini = 0.0
samples = 3
value = [0, 3]
class = S
114a ≤ 0.5
gini = 0.346
samples = 9
value = [7, 2]
class = R
207a ≤ 0.5
gini = 0.146
samples = 88
value = [7, 81]
class = S
gini = 0.0
samples = 1
value = [1, 0]
class = R
278a ≤ 0.5
gini = 0.278
samples = 6
value = [5, 1]
class = R
gini = 0.0
samples = 5
value = [0, 5]
class = S
438a ≤ 0.5
gini = 0.298
samples = 11
value = [2, 9]
class = S
gini = 0.0
samples = 2
value = [2, 0]
class = R
gini = 0.0
samples = 21
value = [0, 21]
class = S
gini = 0.0
samples = 1
value = [1, 0]
class = R
gini = 0.0
samples = 2
value = [2, 0]
class = R
gini = 0.0
samples = 1
value = [0, 1]
class = S
gini = 0.0
samples = 1
value = [1, 0]
class = R
gini = 0.0
samples = 1
value = [0, 1]
class = S
gini = 0.0
samples = 1
value = [0, 1]
class = S
299a ≤ 0.5
gini = 0.061
samples = 126
value = [122, 4]
class = R
gini = 0.0
samples = 1
value = [0, 1]
class = S
gini = 0.0
samples = 6
value = [6, 0]
class = R
95a ≤ 0.5
gini = 0.444
samples = 3
value = [1, 2]
class = S
182a ≤ 0.5
gini = 0.081
samples = 71
value = [3, 68]
class = S
141a ≤ 0.5
gini = 0.36
samples = 17
value = [4, 13]
class = S
gini = 0.0
samples = 5
value = [5, 0]
class = R
gini = 0.0
samples = 1
value = [0, 1]
class = S
184a ≤ 0.5
gini = 0.18
samples = 10
value = [1, 9]
class = S
gini = 0.0
samples = 1
value = [1, 0]
class = R
399a ≤ 0.5
gini = 0.047
samples = 125
value = [122, 3]
class = R
gini = 0.0
samples = 1
value = [0, 1]
class = S
gini = 0.0
samples = 2
value = [0, 2]
class = S
gini = 0.0
samples = 1
value = [1, 0]
class = R
292a ≤ 0.5
gini = 0.056
samples = 69
value = [2, 67]
class = S
429a ≤ 0.5
gini = 0.5
samples = 2
value = [1, 1]
class = R
280a ≤ 0.5
gini = 0.142
samples = 13
value = [1, 12]
class = S
82a ≤ 0.5
gini = 0.375
samples = 4
value = [3, 1]
class = R
gini = 0.0
samples = 8
value = [0, 8]
class = S
167a ≤ 0.5
gini = 0.5
samples = 2
value = [1, 1]
class = R
53a ≤ 0.5
gini = 0.032
samples = 124
value = [122, 2]
class = R
122a ≤ 0.5
gini = 0.029
samples = 67
value = [1, 66]
class = S
120a ≤ 0.5
gini = 0.5
samples = 2
value = [1, 1]
class = R
gini = 0.0
samples = 1
value = [1, 0]
class = R
gini = 0.0
samples = 1
value = [0, 1]
class = S
gini = 0.0
samples = 11
value = [0, 11]
class = S
gini = 0.5
samples = 2
value = [1, 1]
class = R
gini = 0.0
samples = 3
value = [3, 0]
class = R
gini = 0.0
samples = 1
value = [0, 1]
class = S
gini = 0.0
samples = 1
value = [1, 0]
class = R
gini = 0.0
samples = 1
value = [0, 1]
class = S
112a ≤ 0.5
gini = 0.016
samples = 122
value = [121, 1]
class = R
33a ≤ 0.5
gini = 0.5
samples = 2
value = [1, 1]
class = R
gini = 0.0
samples = 64
value = [0, 64]
class = S
129a ≤ 0.5
gini = 0.444
samples = 3
value = [1, 2]
class = S
gini = 0.0
samples = 1
value = [1, 0]
class = R
gini = 0.0
samples = 1
value = [0, 1]
class = S
gini = 0.0
samples = 120
value = [120, 0]
class = R
389a ≤ 0.5
gini = 0.5
samples = 2
value = [1, 1]
class = R
gini = 0.0
samples = 1
value = [0, 1]
class = S
gini = 0.0
samples = 1
value = [1, 0]
class = R
gini = 0.0
samples = 2
value = [0, 2]
class = S
gini = 0.0
samples = 1
value = [1, 0]
class = R
gini = 0.0
samples = 1
value = [0, 1]
class = S
gini = 0.0
samples = 1
value = [1, 0]
class = R

Random Forest

- Continuation of CART
- Creates many uncorrelated decision trees. Each tree decides on what the outcome should be, and the most commonly chosen outcome is returned

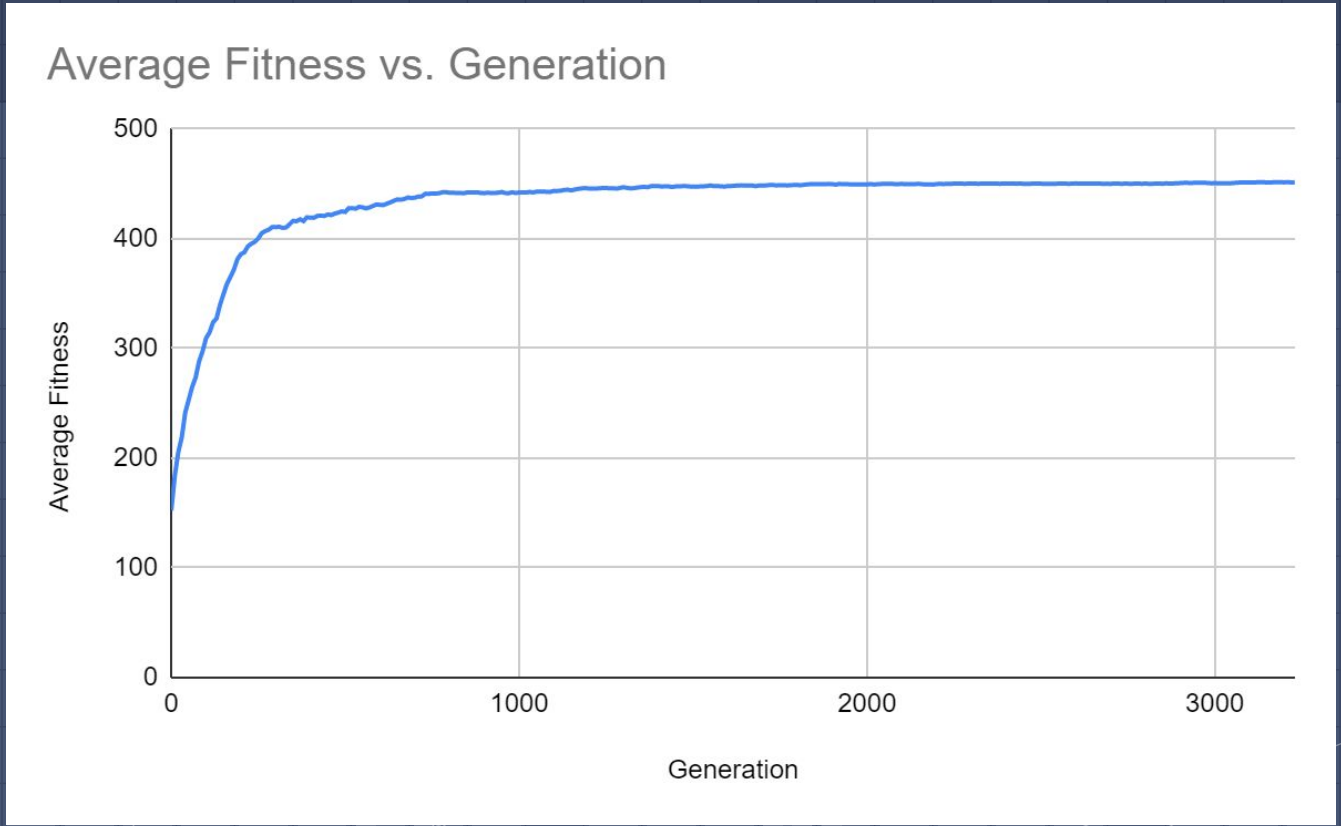
Naive Bayes

- Algorithm that estimates the probability of a strain being resistant/susceptible based on prior probabilities of predictors



Genetic Learning

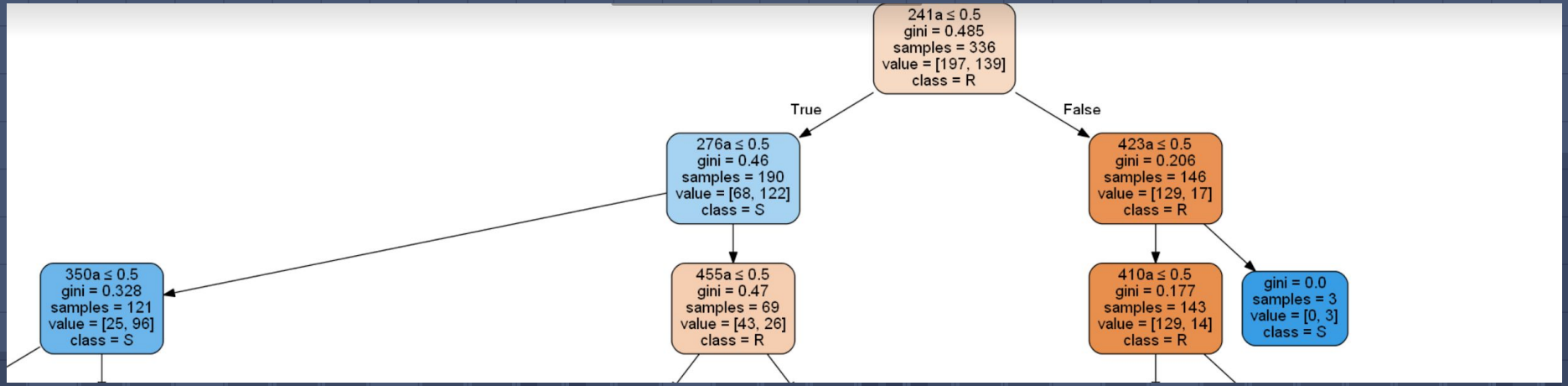
- Randomly generate 100 lists of weights. Each weight corresponds to a gene.
- For each strain, weights of mutated genes are added up. If the total exceeds 10, the strain is categorized as resistant, otherwise it is categorized as susceptible.
- The lists are rated on how many strains they categorize correctly.
- The best lists are used to produce new lists with a combination of their weights.
- After many generations, the lists become better at categorizing strains. We choose the first list that can correctly categorize 95% of the strains and analyze its weights.
- Higher weights represent more important genes.


Average Fitness vs. Generation
500
400
300
200
100
0
Average Fitness
0
1000
2000
3000
Generation

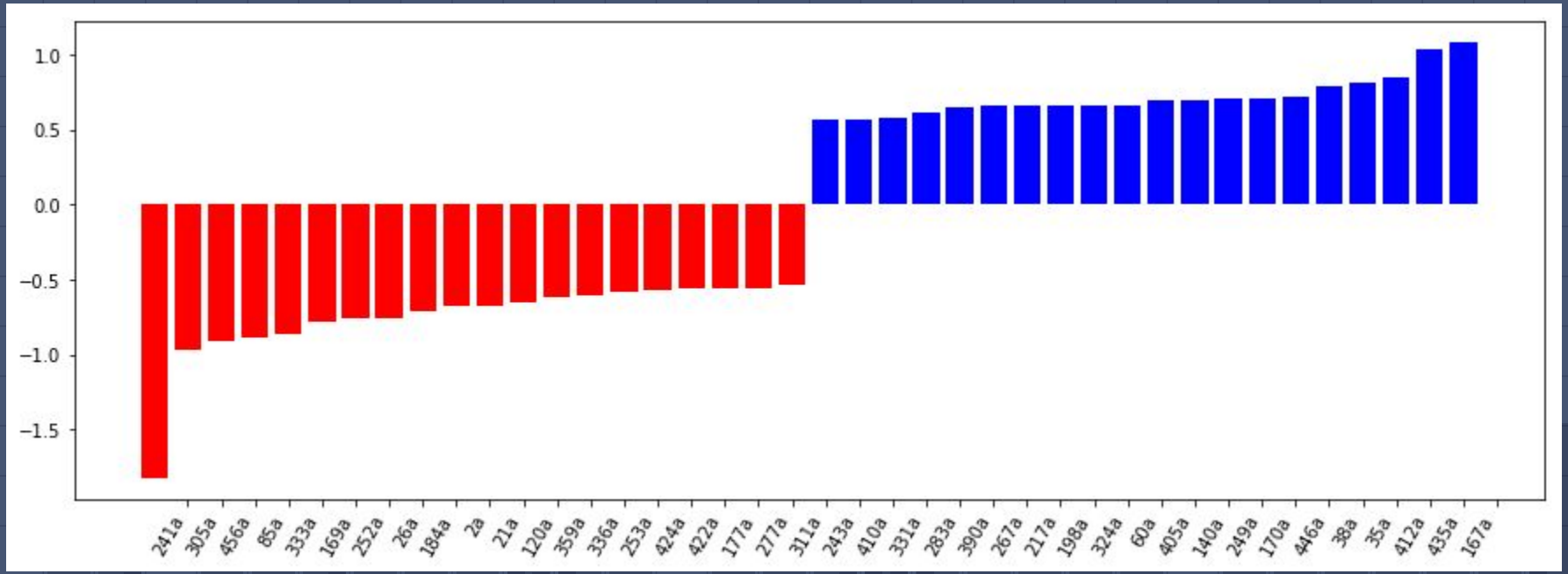
Results

4

Zoomed in on Decision Tree



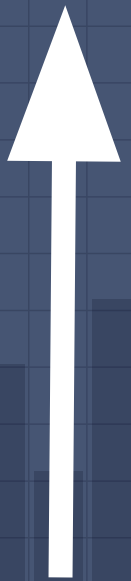
Naive Bayes Plot of Significant Features



Comparing Model Results side by side

Random Forest	Naive Bayes	Genetic Learning	CART
241a	241a	241a	241a
276a	305a	120a	423a
455a	456a	208a	410a
409a	85a	396a	334a
3a	333a	289a	159a
277a	169a	2a	202a
346a	252a	455a	299a
272a	26a	184a	399a
219a	184a	218a	53a
275a	2a	305a	112a

Most Significant



Gene 241a is pncA

Functional pncA gene converts *Pyrazinamide*, a common tuberculosis drug, into its active form, pyrazinoic acid, which accumulates inside tuberculosis cells kills them.

Mutated pncA gene is strongly correlated with Pyrazinamide resistant *Mycobacterium tuberculosis*, indicated by previous studies.

Thus, feature selection is successful for finding gene of interest.

Conclusions and Future Study

Feature Selection is successful for this dataset

Useful in combating emerging crisis of antibiotic resistance

Applicable to other datasets of similar nature with further research

Acknowledgments

- Dr. Alterovitz, Insung Na
- Ling Teng
- MIT PRIMES
- Our Parents