Predicting Conditions from Electronic Health Records

Jessica Cabrera, Chloe Jeon, Isha Karim, Maddie Shenkan

AI4ALL Project 4
7/26/19

What are EHRs?

- Electronic Health Records
- One large data from storing patient information
- Includes: patient information (demographics, age), conditions, medications, care plans, etc.
- Best source of large amounts of clinical data for analysis

Methodology

- data collection
  - public data (real)
  - Synthea (synthetic)
- data cleanup
- classifiers
  - decision trees
  - decision tree forests
  - accuracy
- confusion matrix

1

Preterm Birth

Maddie Shenkan

User Guide to the 2017 Natality Public Use File

CDC

CENTERS FOR DISEASE CONTROL AND PREVENTION™

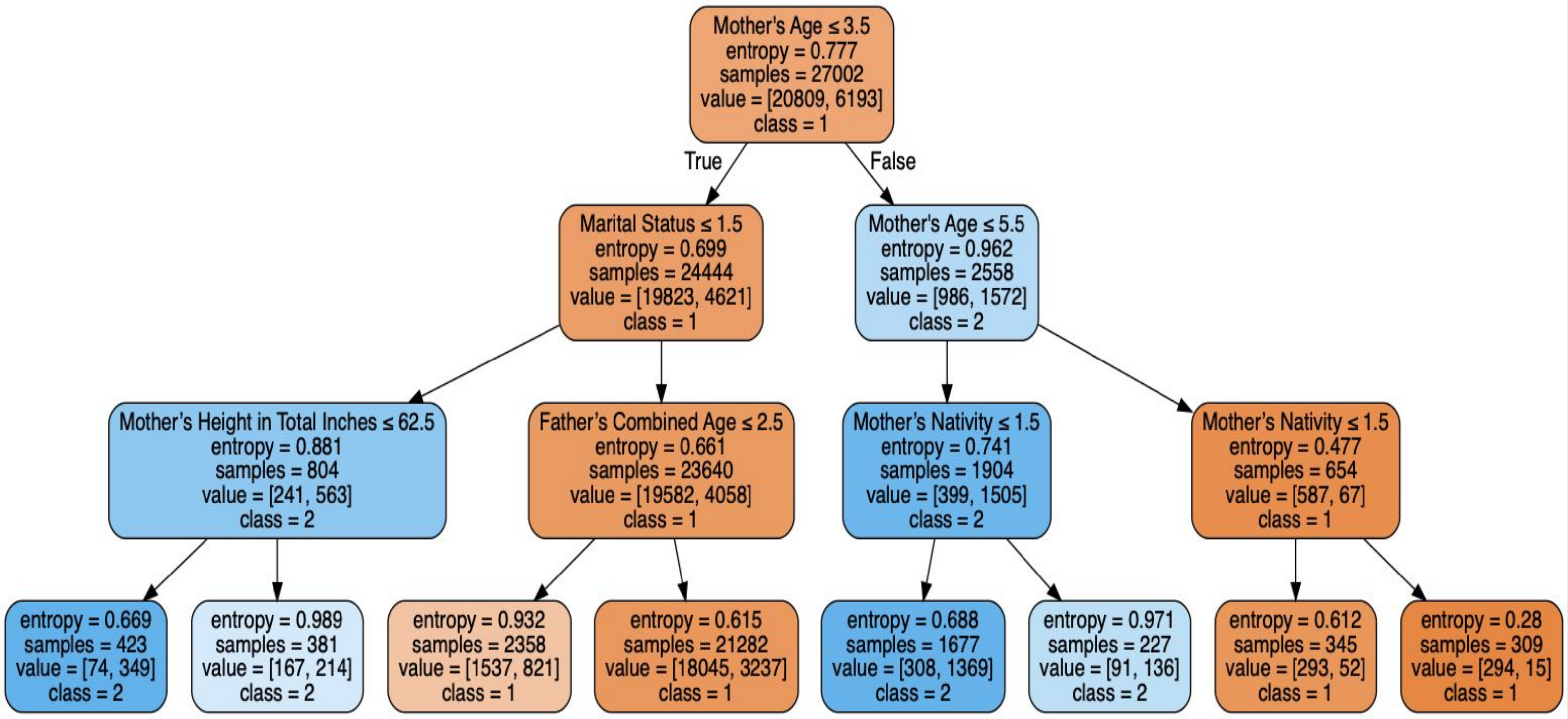
Fixed Width File

```
        201701    10052        11                1                       26094    2                   1055134    015 Y2  13 1                    220305513    0153 1     040000  5  5               02604
 88888   02604     0612          03  031      N10000000000001111N1        66124.52    1521   164 11221     NNNNNY111111NXX111Y03 110     NNNNN111111      NN 11                   NNNNNN1111111     14N11142
1     NNNNN 11111 1     1N1111     0941885   1    9               M 04  2016      39072    390723645 083     NNNNNN 111111 1     NNNNNN111111NNNNNN1111111     NYY1


        201701    18182        11                1                       19072    1                   2033033    013 X1  12 1                    270410615    0164 1     020000  3  3               99999
 88888   99999     0211          08  051      N12015150033301111Y1        72115.61    1151   125 11011     NNNNNN111111NXX111N00 111     NNNNN111111      NN 11                   NNNNNN1111111     14N11132
1     NNNNN 11111 1     1Y2211     0421083   1    9               F 07  2016      24021    250210860 021     YYYNNN 111111 0     NNNNNN111111NNNNNN1111111     NYY1


        201701    02043        11                1                       20083    1                   1011011    217 X1  13 1                    200301101    0113 1     000001  1  2               88888
 99999   99999     0412          09  061      N10000000000001111N1        68120.52    1351   154 11921     NNNNNN111111NXX111N00 111     NNNNN111111      NN 11                   NNNYNY1111110     11X11111
1     NNNNN 11111 1     1N2211     0521094   1    9               F 04  2016      37062    360512875 063     NNNNNN 111111 1     NNNNNN111111NNNNNN1111111     NYY1
```

DECISION TREE

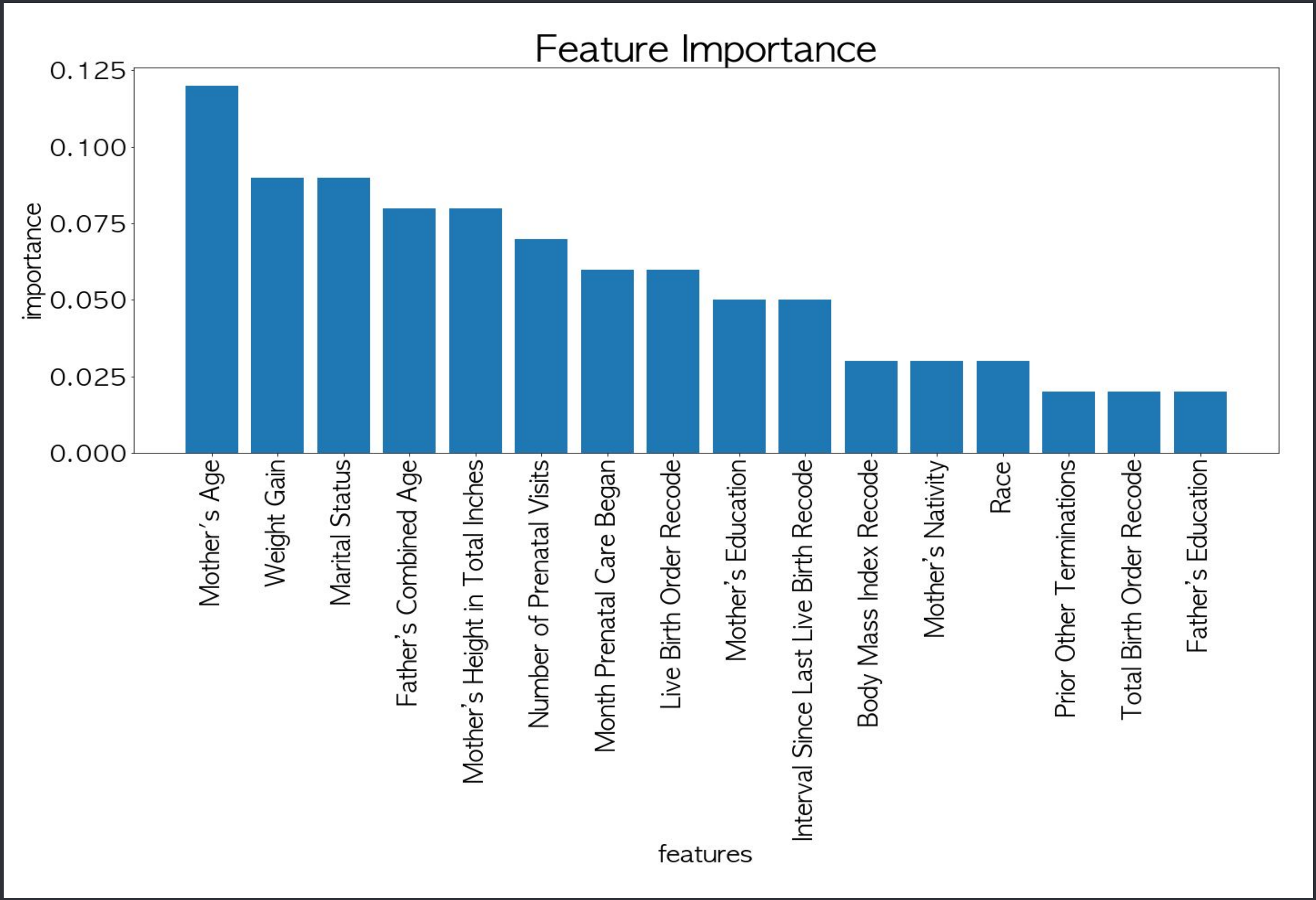


Accuracy : 82%

Class 1 = Pre-Term

Classe 2 = Full Term

Random Forest (Average of 100 Decision Trees)



Confusion Matrix

- True Positives = 8564

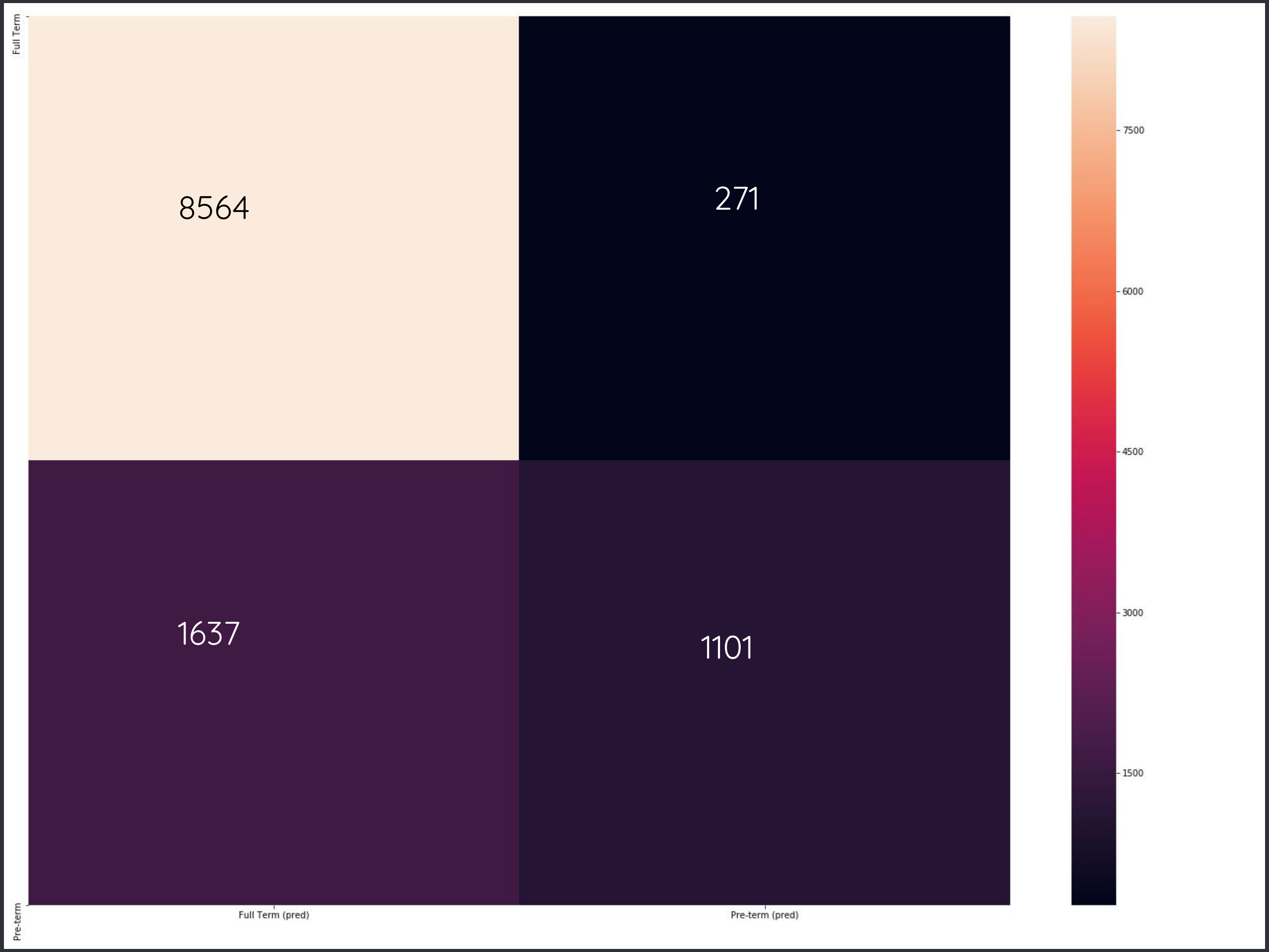
Ex: the woman is preterm and the algorithm predicted preterm birth

- True Negatives = 1161

Ex: the woman is full term and the algorithm did predict it

- False Positives = 1637
- False Negatives = 271

Confusion Matrix



2

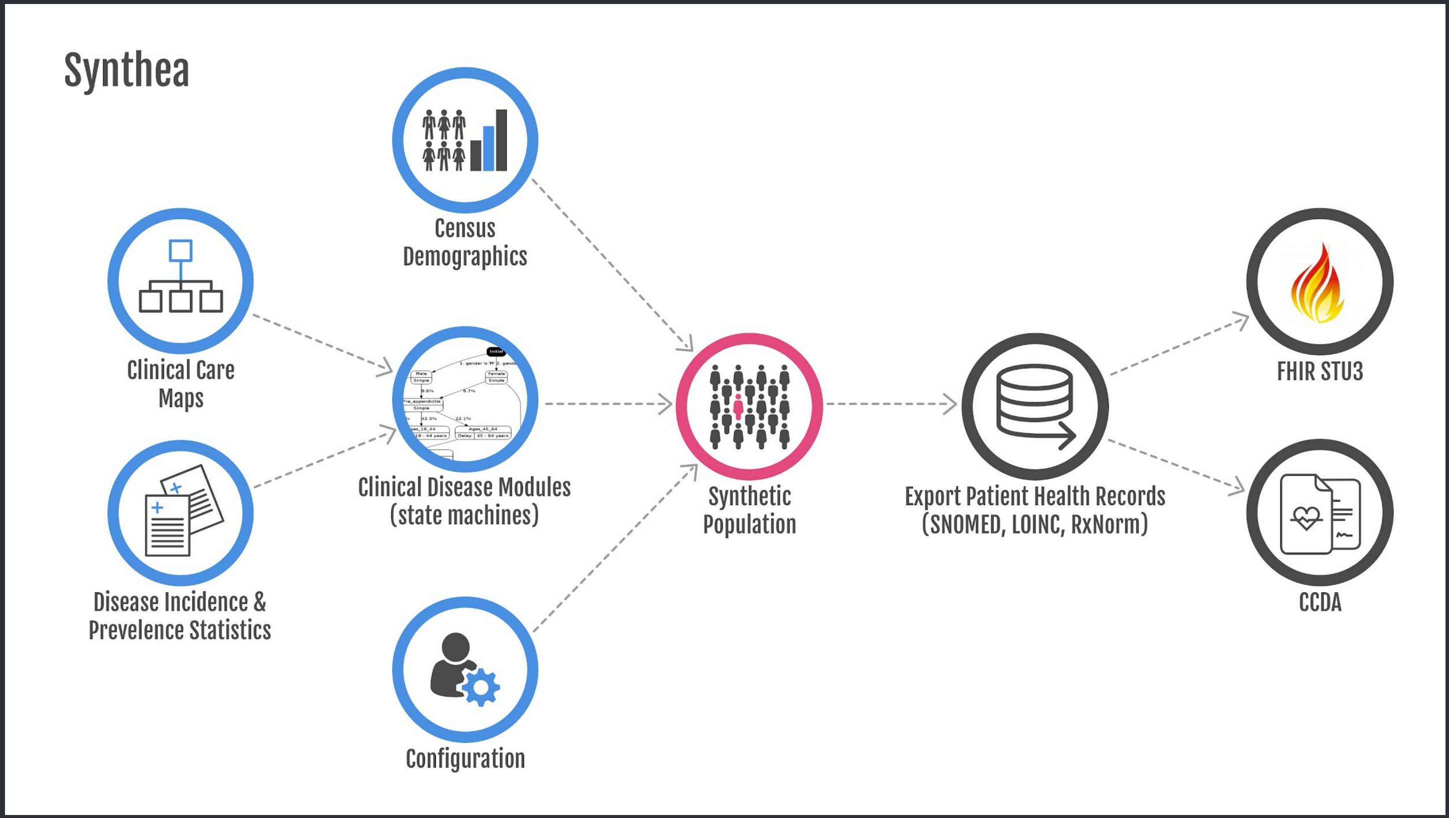
Coronary Heart Disease

Jessica Cabrera

SYNTHEA

- Uses real data to produce a realistic population and patient health record
- Well-organized, no missing or incorrect data
- Even if it’s as realistic as possible, it’s not real

SYNTHEA


Synthea
Census Demographics
Clinical Care Maps
Clinical Disease Modules (state machines)
Disease Incidence & Prevelence Statistics
Configuration
Synthetic Population
Export Patient Health Records (SNOMED, LOINC, RxNorm)
FHIR STU3
CCDA

Coronary Heart Disease

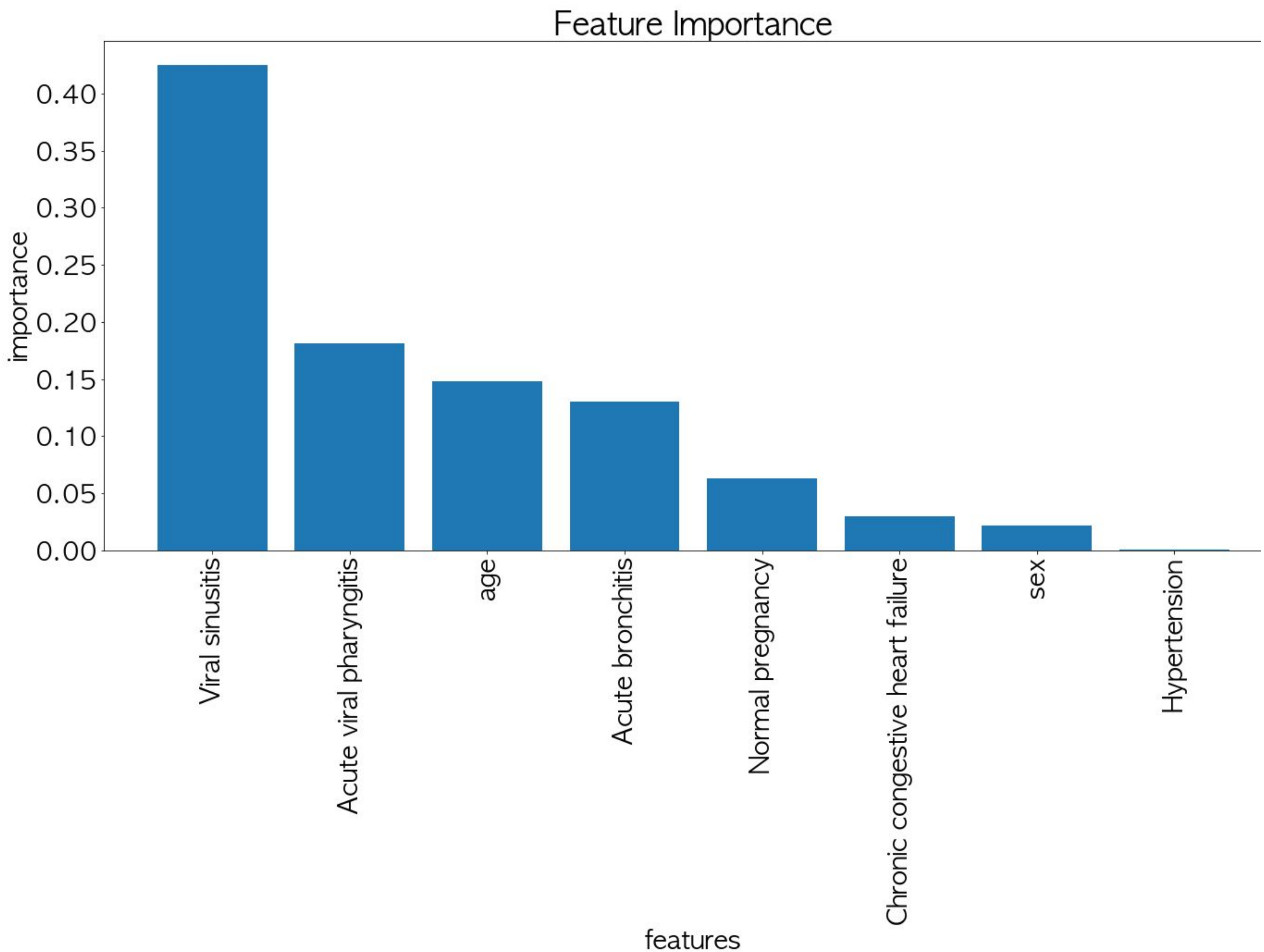
Background:

- CHD happens when the arteries harden, narrowing the blood supply to the heart

- This disease is very common
  - More than 3 million cases recorded per year

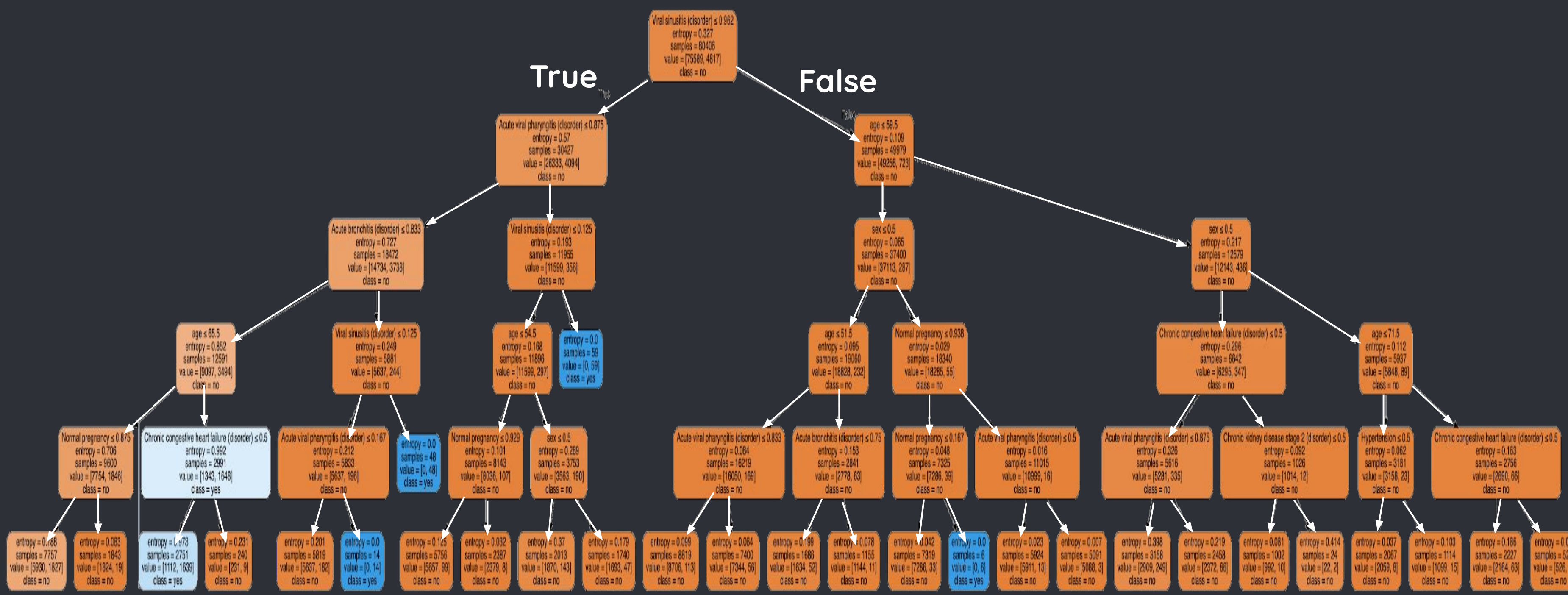
- 610,000 people die annually of heart disease alone

Bar Graph



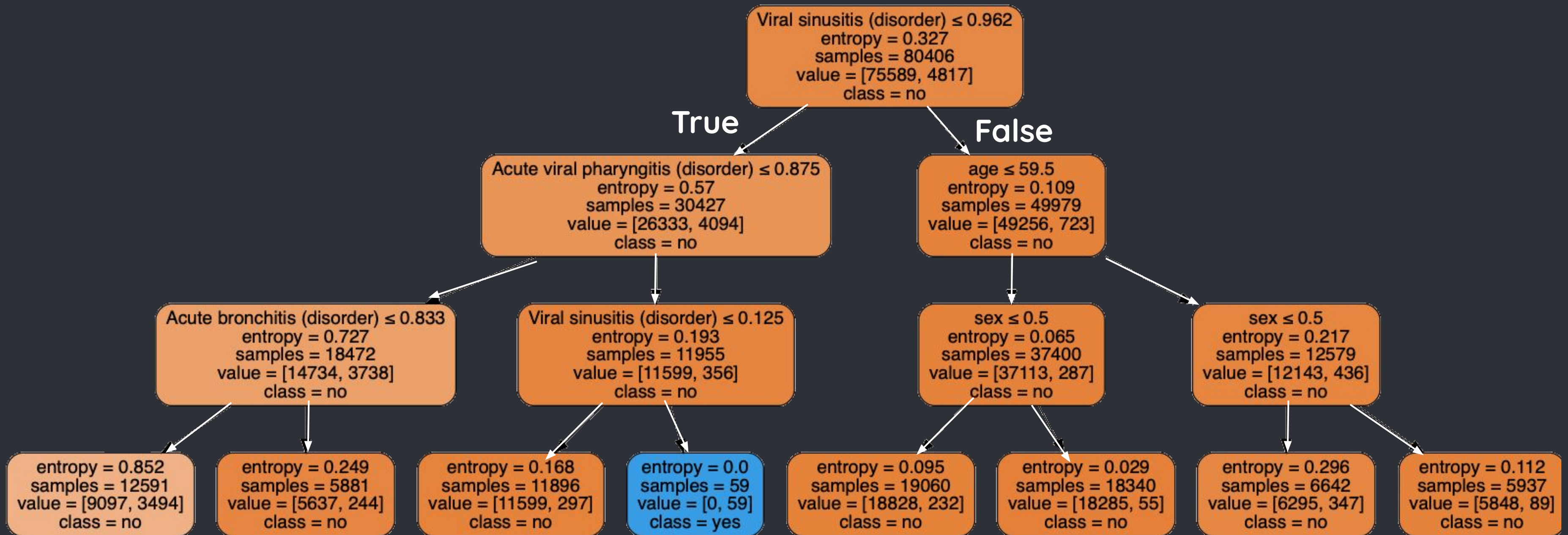
Feature Importance
importance
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Viral sinusitis
Acute viral pharyngitis
age
Acute bronchitis
Normal pregnancy
Chronic congestive heart failure
sex
Hypertension
features

Decision Tree - Complex



Accuracy: 95%

Decision Tree - Simplified



Accuracy: 94%

- 94% don't have CHD
- 6% do have CHD

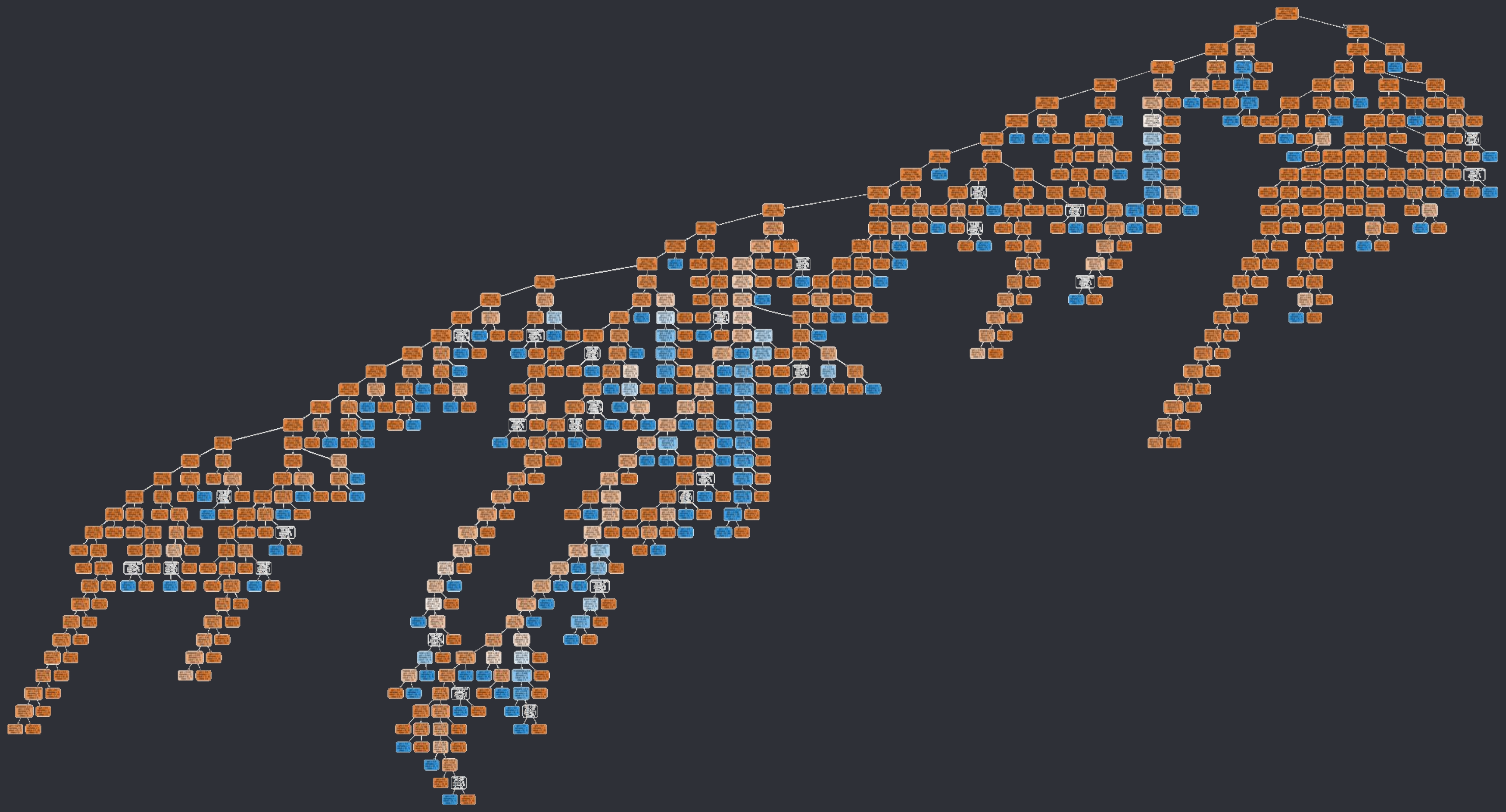
Future Implications

- Remove Viral Sinusitis
  - Correlation does not mean causation
- Investigate chronic congestive heart failure
- Consider other factors such as lifestyle, exercise, and environment
- Implementing a confusion matrix
  - To look into my original model if its always saying no

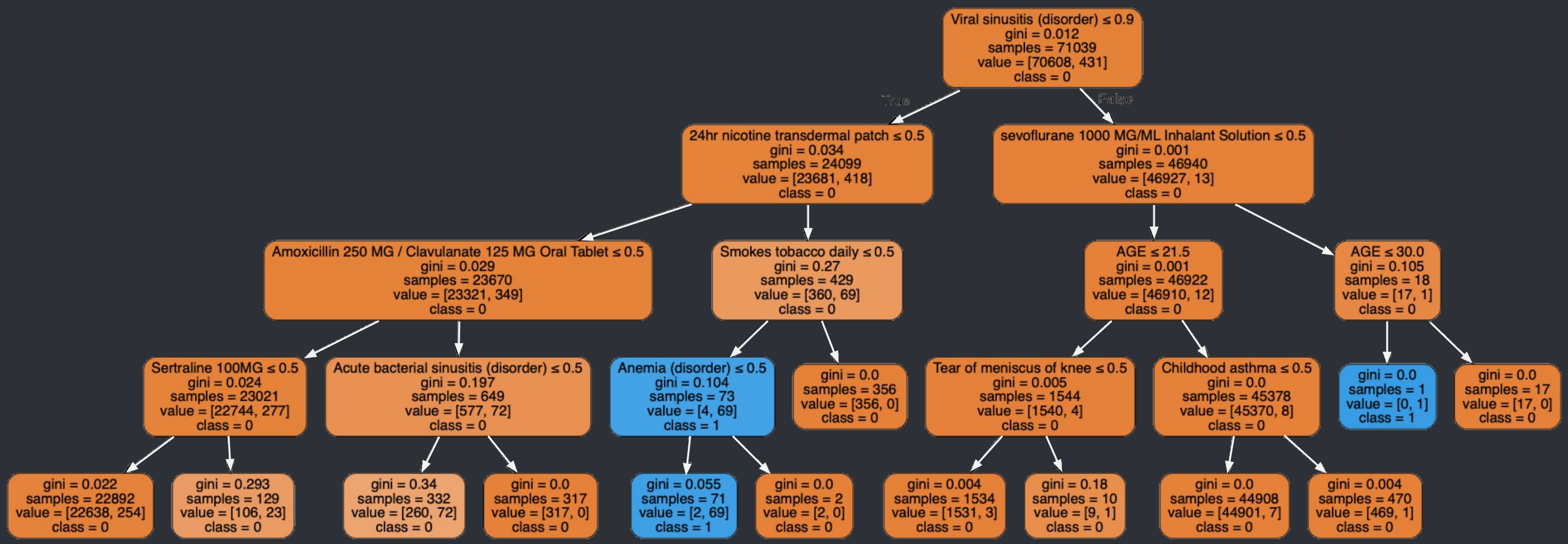
3

Depression

Chloe Jeon



Classifiers - Complicated Decision Tree



Classifiers - “Pruned” Decision Tree

99.63%

Complicated Tree

99.48%

Simplified Tree

99.73%

Complicated Forest

99.39%

Simplified Forest

Accuracies

ISSUES

- Too many controls (.5%)
  - removed: younger than 19, conditions after first depression-related diagnosis
  - depression is not rare; diagnosis is rare
  - Synthea doesn’t handle depression well?
- Decision tree tried to avoid false positives
  - tried to avoid saying “depressed” when they were actually not depressed
  - out of 200 depressed patients, only 2 were reported as depressed

SOLUTION?

- changed tree criteria
- improved results
- still only 100 of 200 depressed patients were reported as depressed

Before

```
[[30260      0]
 [  186      0]]
```

After

```
[[30255      5]
 [   77    109]]
```

Sinusitis (disorder)	0.132552
Smokes tobacco daily	0.127229
AGE	0.086067
Neoplasm of prostate	0.076878
Acetaminophen 21.7 MG/ML / Dextromethorphan Hydrobromide	0.051526
insulin human isophane 70 UNT/ML / Regular Insulin Huma	0.027877
Diabetes	0.027351
Acute viral pharyngitis (disorder)	0.026604
24hr nicotine transdermal patch	0.021142

Smokes tobacco daily	0.614428
24hr nicotine transdermal patch	0.102101
Acute bacterial sinusitis (disorder)	0.086585
Amoxicillin 250 MG / Clavulanate 125 MG Oral Tablet	0.070102
Viral sinusitis (disorder)	0.052665
Sertraline 100MG	0.040712
Anemia (disorder)	0.020857
AGE	0.010821
Tear of meniscus of knee	0.001084
sevoflurane 1000 MG/ML Inhalant Solution	0.000625
Childhood asthma	0.000021
Brain damage - traumatic	0.000000
History of myocardial infarction (situation)	0.000000
Preeclampsia	0.000000
Facial laceration	0.000000
Seasonal allergic rhinitis	0.000000
Laceration of forearm	0.000000
Metastasis from malignant tumor of prostate (disorder)	0.000000
Laceration of hand	0.000000

AGE	0.085819
Viral sinusitis (disorder)	0.043699
Body mass index 30+ - obesity (finding)	0.037186
insulin human isophane 70 UNT/ML / Regular Insulin Huma	0.034827
Carcinoma in situ of prostate (disorder)	0.028834
Acetaminophen 21.7 MG/ML / Dextromethorphan Hydrobromide	0.028733
Simvistatin 10 MG	0.026778
24hr nicotine transdermal patch	0.024916
Acute viral pharyngitis (disorder)	0.024844

Viral sinusitis (disorder)	0.101519
Acute viral pharyngitis (disorder)	0.080594
24hr nicotine transdermal patch	0.079947
Acetaminophen/Hydrocodone	0.058167
Prediabetes	0.045633
Sertraline 100MG	0.044584
Body mass index 30+ - obesity (finding)	0.043806
Acute bronchitis (disorder)	0.043121
Anemia (disorder)	0.038404
Carcinoma in situ of prostate (disorder)	0.034080
insulin human isophane 70 UNT/ML / Regular Insulin Huma	0.032856
0.25 ML Leuprolide Acetate 30 MG/ML Prefilled Syringe	0.030126
F	0.022643
Diabetes	0.022034
M_y	0.018456
Smokes tobacco daily	0.016040
1 ML DOCEtaxel 20 MG/ML Injection	0.015347
Suspected lung cancer (situation)	0.014684
Neoplasm of prostate	0.014627

Important Features

- Feature Importance
  - sinusitis (common cold)
  - smokes tobacco daily
  - 24hr nicotine transdermal patch
  - age (20–30)
  - acute viral pharyngitis (sore throat)
  - obesity
  - cancer
  - insulin (diabetes medication)
    - "Insulin-sensitizing drug relieves symptoms of chronic depression"

FUTURE IMPLICATIONS

- neural network
- synthetic data
- correlation doesn’t imply causation
  - common cold/sore throat are common
- depression factors not in EHRs
- study potential risk factors
  - inaccurate decision tree?
  - actual risk factor that was previously unknown?
- apply same methods to real data

4

Opioid Overdose

Isha Karim

BACKGROUND

- Opioid Overdose: toxicity due to excessive opioid consumption

- In 2017, over 47,000 Americans died as a result of opioid overdose.

MANAGING & SORTING DATA

- Public and synthetic data was merged to create a final dataframe.
  - Used a NLP matching algorithm
  - **NLTK library** classified data from each file (string matching) into dictionaries

Gender	Race	Ethnic Group	Manner of D	Manner Type	Manner Sub	Cause of Death	Cont
Male	White	White	Accident	Drug - Medic	Medication	Mixed fentanyl, alprazolam, and doxylami	None
Male	White	White	Accident	Drug - Medic	Drug and Me	Acute methadone, clonazazepam, gabape	None
Female	White	White	Accident	Drug - Medic	Medication	Acute fentanyl, oxycodone, alprazolam, ai	Rece
Male	White	White	Accident	Drug - Medic	Medication	Complications including anoxic encephalo	None
Male	White	White	Accident	Drug - Medic	Drugs of abu	Acute heroin intoxication	None
Male	White	White	Accident	Drug - Medic	Drug and Me	Oxycodone, alprazolam and methamphet	None
Male	White	White	Accident	Drug - Medic	Meds & Alco	Acute oxycodone, chlordiazepoxide and a	None
Female	White	White	Accident	Drug - Medic	Medication	Combined effects of fentanyl and morphi	Blunt
Male	White	White	Accident	Drug - Medic	Medication	Acute methadone intoxication	None
Female	White	White	Accident	Drug - Medic	Medication	Acute morphine, oxycodone, diphenhydra	None
Male	White	White	Accident	Drug - Medic	Drug and Me	Fentanyl and methamphetamine intoxicat	Athe
Male	White	White	Accident	Drug - Medic	Drug and Me	Fentanyl, cocaine, pseudoephedrine, and	None
Female	White	White	Accident	Drug - Medic	Medication	Mixed fentanyl, alprazolam, amphetamin	None
Male	White	White	Accident	Drug - Medic	Medication	Acute combined drug intoxication (amphe	Cong

ELIMINATING BIAS FROM THE PREDICTIVE MODEL

- RACE
- GENDER
- CONTROL

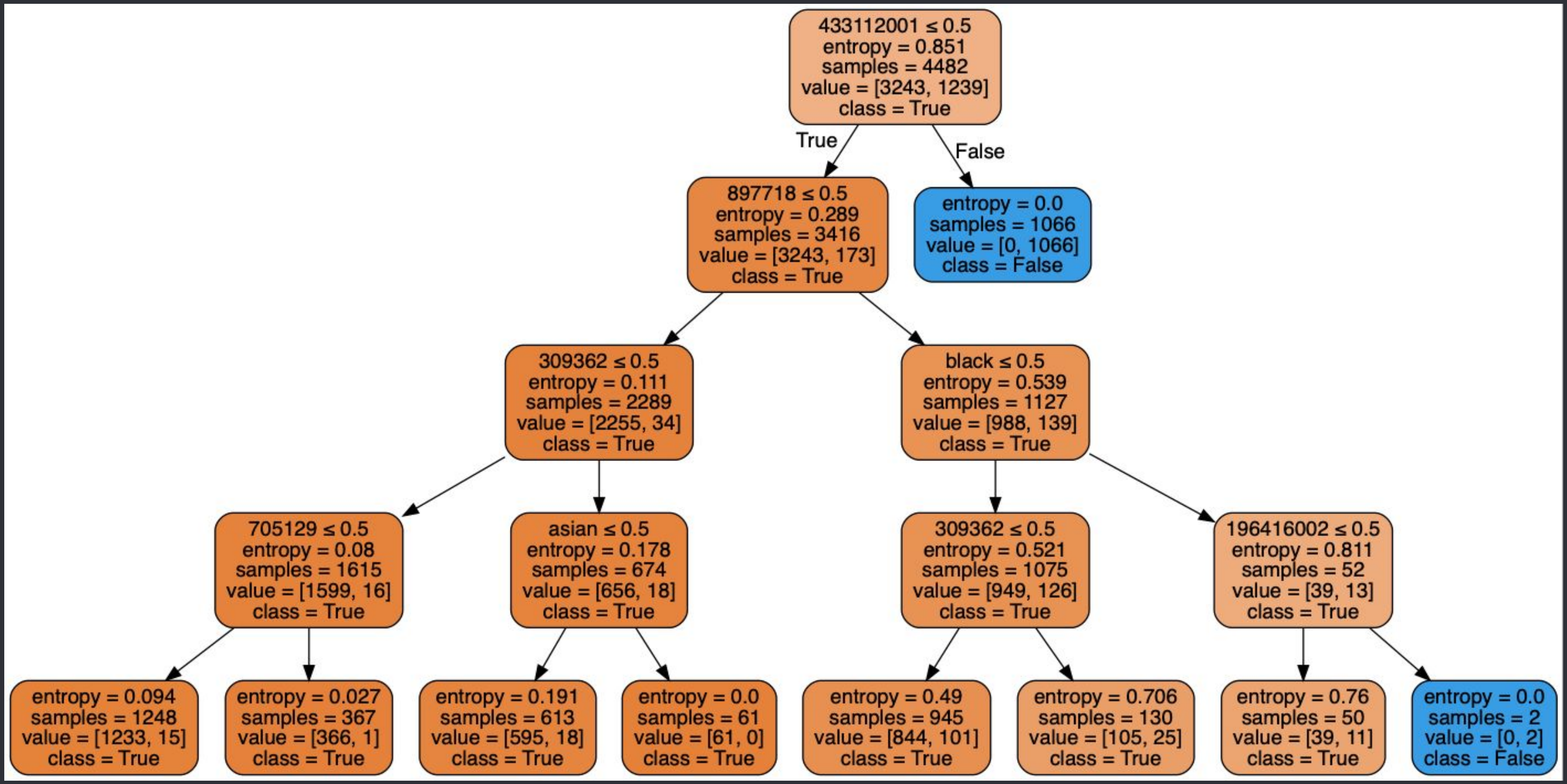
In the data set, an even breakdown of **all races and gender and control** pop. avoided overwhelming bias in the model.

- Randomly set the amount of controls equal to the overdose diagnoses

FINAL BINARY MATRIX

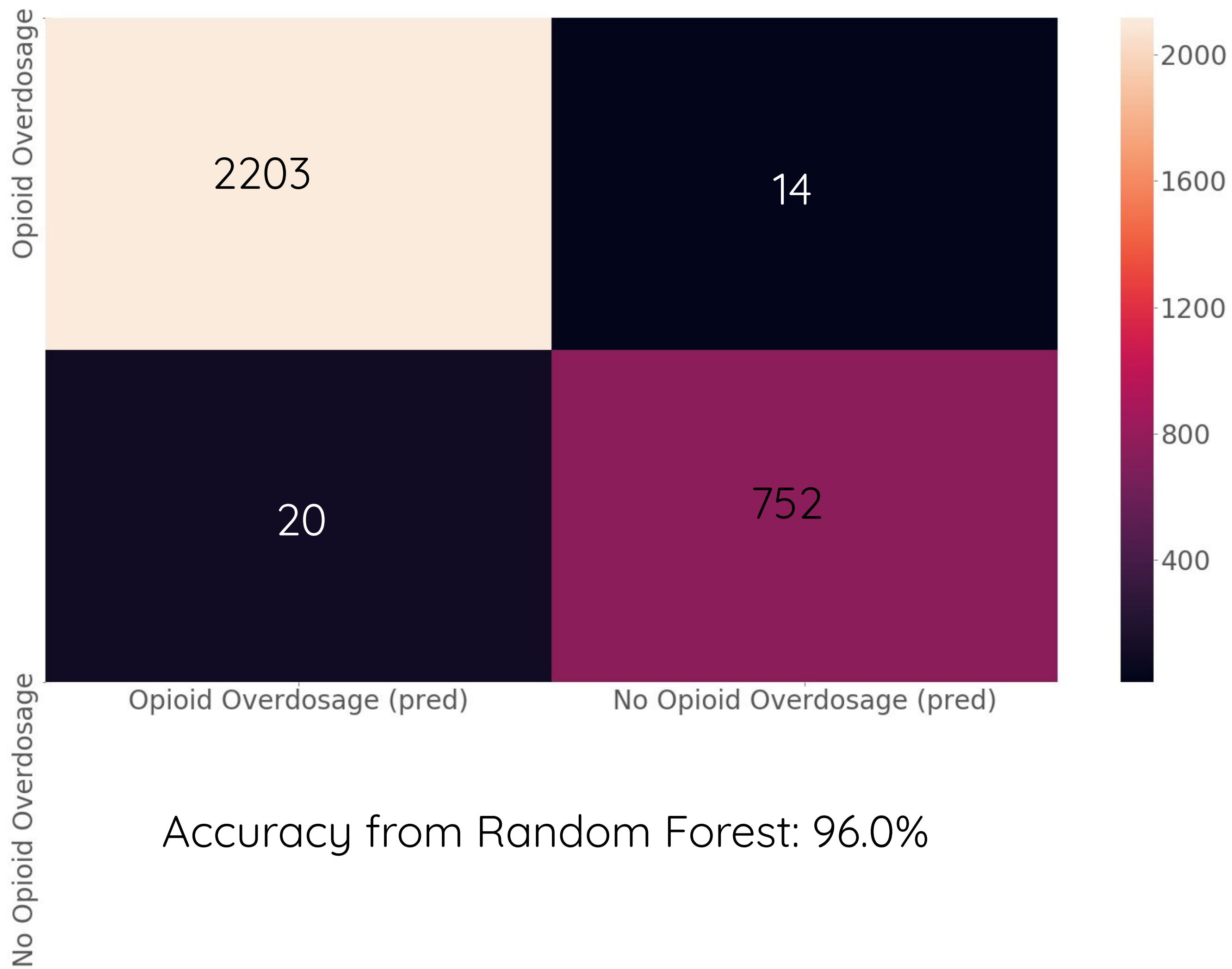
	22298006	49436004	53741008	55680006	82423001	196416002	230690007	399211009	410429000	429007001	...	433112001	447365002	F	M
0002b13d-b122-4912-b65f-d1b58b4cd4d6	1	0	1	0	0	0	0	1	0	0	...	0	0	0	1
00031e80-8438-4173-9a2c-deeda37af0c9	0	0	1	0	0	0	0	0	1	1	...	0	1	0	1
001ae9bc-bcd1-4012-9d3a-5c57c8b47721	0	1	0	0	0	0	0	0	0	0	...	0	0	1	0
00212701-01fa-4a12-b3e8-0f48e0e4c9d0	0	1	0	0	0	1	0	0	0	0	...	0	0	1	0
0033a913-4cdc-405e-b18a-1ee59f4ed78c	0	1	1	0	0	0	1	0	0	0	...	1	0	0	1

DECISION TREE



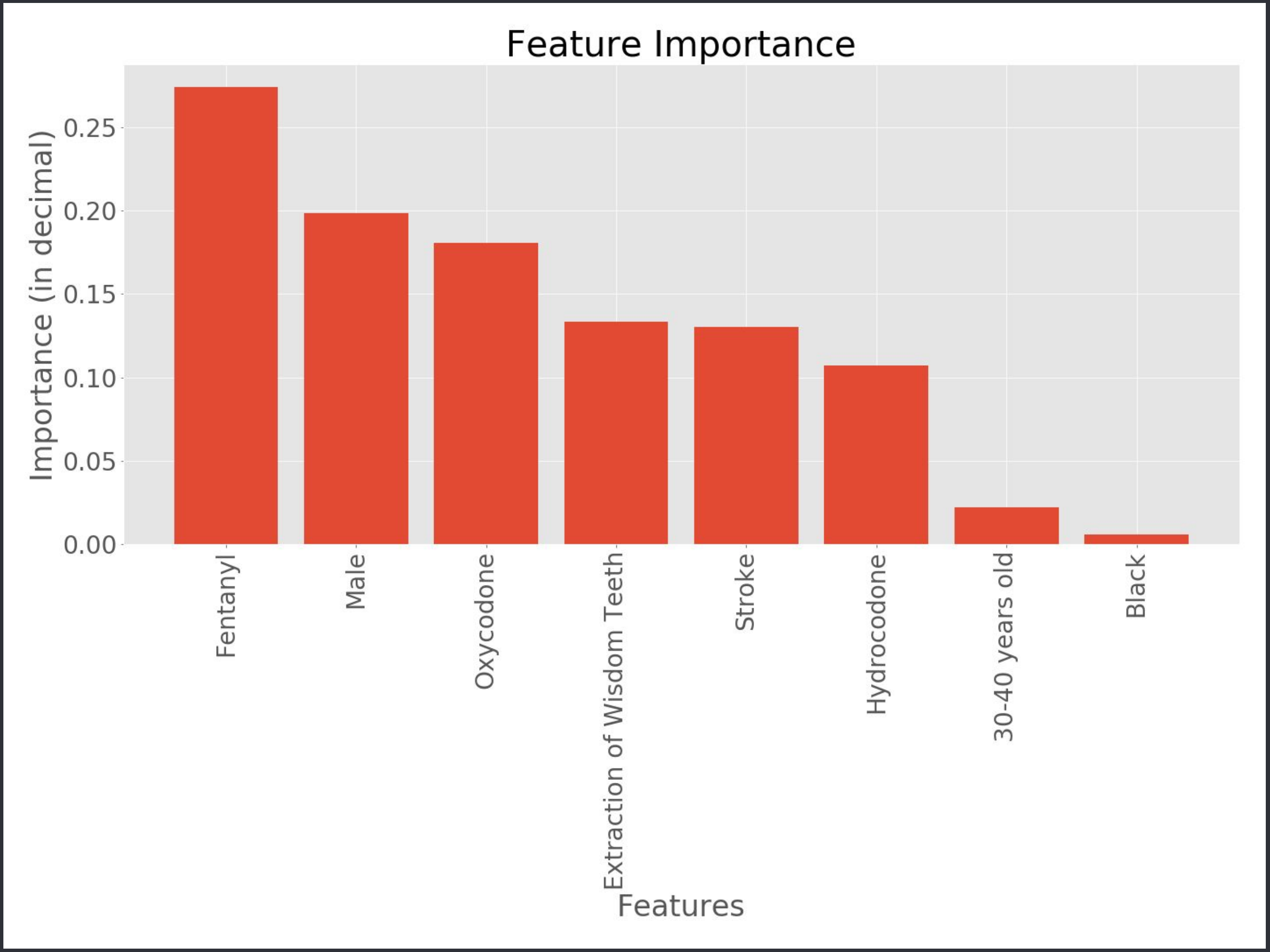
Accuracy : 94.42%

CONFUSION MATRIX



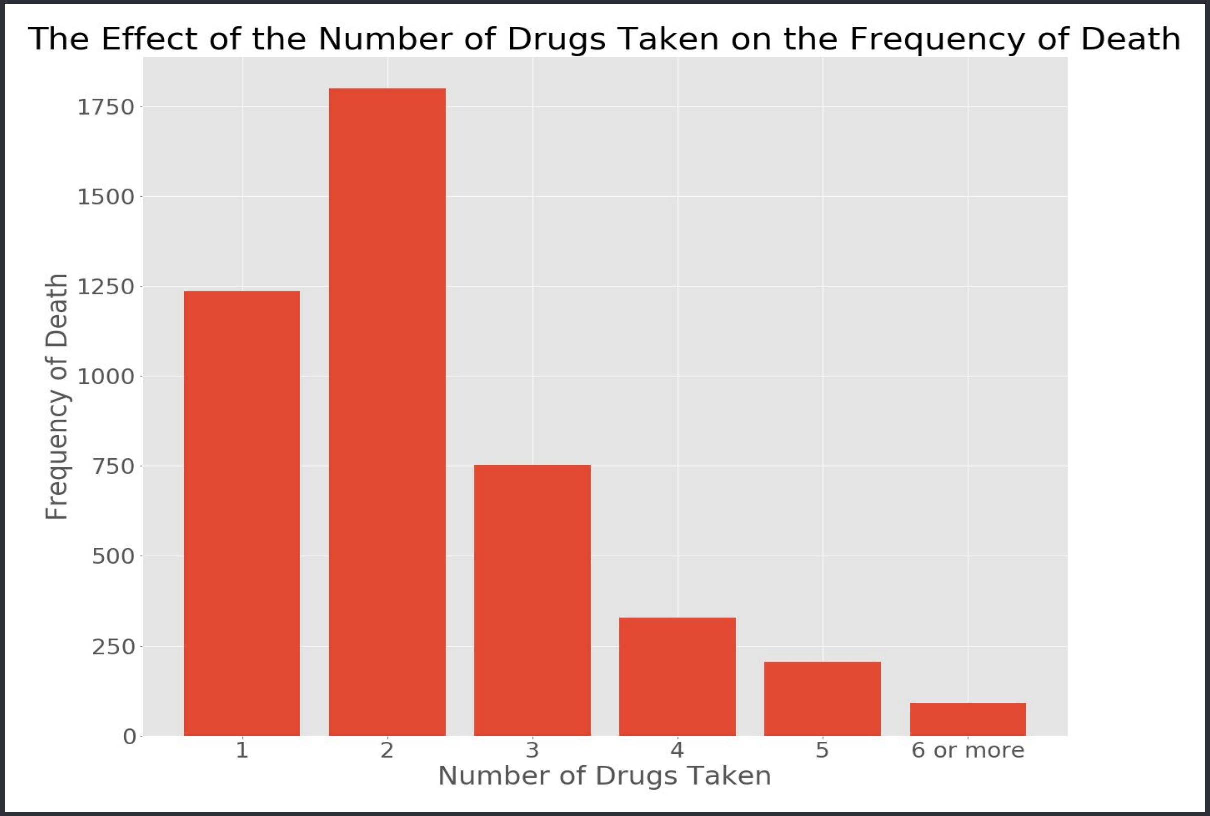
Accuracy from Random Forest: 96.0%

RISK FACTORS & FEATURE IMPORTANCE



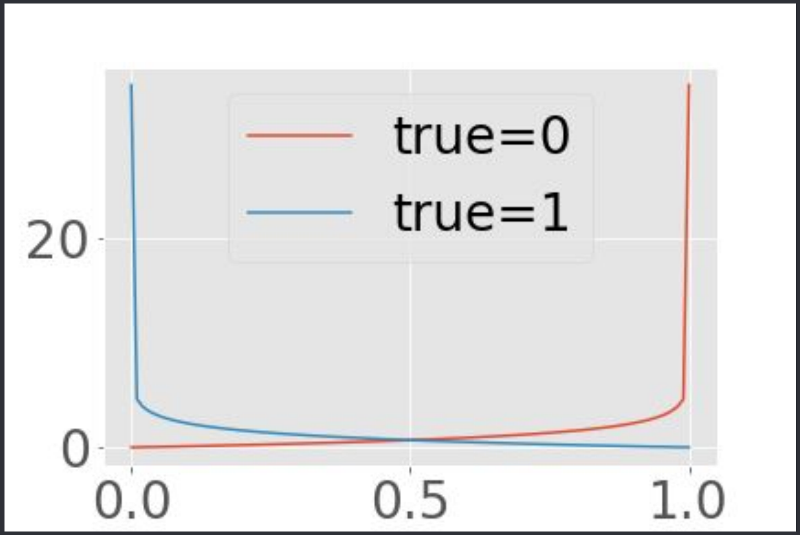
RISK FACTORS for HIGH RISK GROUPS

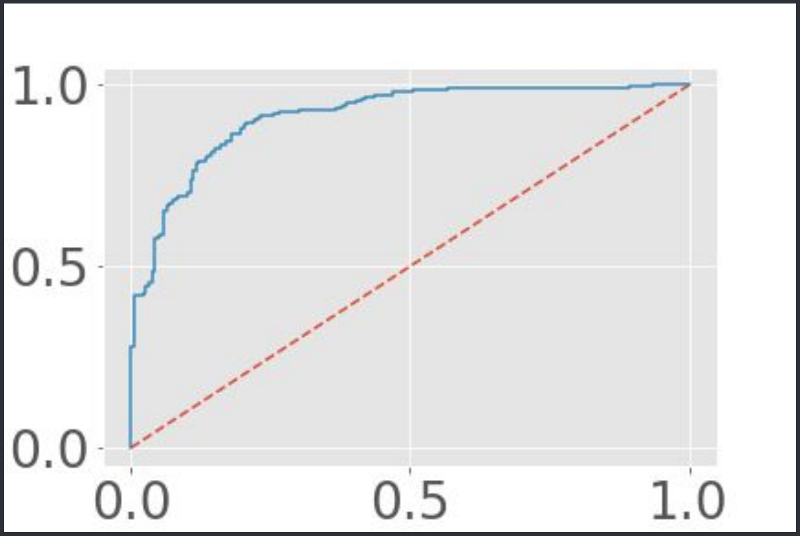
- Medications
  - Fentanyl, oxycodone, hydrocodone
  - Taking more than one at a time is worse
  - In combination with depressants, the risk is greater



ANALYZING METRICS

- Using LogLoss and ROC curves





CONCLUSIONS

- Using NLTK, merging of the public and synthetic datasets made the results more realistic and accurate.

- Even breakdown of features decreased **false negatives**.
  - Categorizing patients at risk for overdose as not at risk

CAVEATS AND FUTURE IMPROVEMENTS

- Subgroup Analysis by race
- Investigating how an opioid is taken
  - i.e. topically, orally, or injection-wise
- Delving into a deep neural network
  - Compare results with other predictive models
- Matching on more criteria

Acknowledgements

Special Thanks to:
Jean Costello, Brian Le, Sarah Tan
Pingyang Liu,
Maya Gonzalez, Jillian Burchard,
Marina Sirota,
Eva Kaye-Zweibel,
And
The AI4ALL Program