Healthcare Final Project:

A Case Study Using Predictive Analytics for Risk Profiling of COVID-19 Patients

Team 1:

Haina Yan/ Henry Gao/ Jemma Rong

Jimmy Hu/ Pengcheng Xu/ Ziyu Tang (Audit)

CONTENTS

PART | 01

Introduction

PART | 01 Introduction

Covid-19 has hit U.S. hard

a. Number of cases represent of 1/3 of total cases in the world
b. Evidence suggests community contagion much earlier than originally thought
c. Political polarization and incompetent administration exacerbate the situation
d. Not enough test kits are available

Covid-19 has ravaged American economy

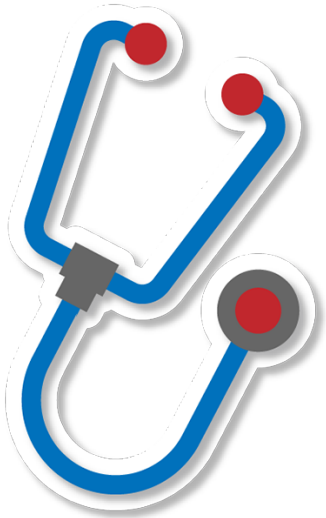
a. Major economic indices hit all time low since the Great Depression in 1930s
b. More than 30 million initial unemployment claims, suggesting an unemployment rate rises sharply to beyond 20%, worst since the Great Depression
c. Bankruptcy make thing worse

Critical moment: trade-off between economy re-opening and infection containment.

How do we address the trade-off without massive testing?

1. Identify those at high risk of succumbing to the virus using machine learning models
2. Let people stay at home until test kits and vaccines are widely available
3. Reopen economy partially

PART | 02



Descriptive Analysis

PART | 02 Descriptive Analysis

Data Preparation

- Map the age and sex variables into dummy variables respectively
- Get the upper categories (DGL_3_Extend) of ICD10 DX codes
- Map DGL_3_Extend into dummies by the following mechanism:

For each patient, if he/she was diagnosed by **at least** one ICD10 DX code of a specific DGL_3_Extend code, we will treat the code dummy variable

- Final dataset: 370633 records and 229 features

PART | 02 Descriptive Analysis (Cont)

Gender Difference and Mortality

- Perform **Fisher Exact Test** to explore relations between Gender difference & COVID-19 Mortality
- Odds ratio of 1.62 and **p-value < 0.05**
- Gender bias on COVID19 mortality is statistically significant
- Males might be more sensitive to COVID-19 and have a higher death rate than females

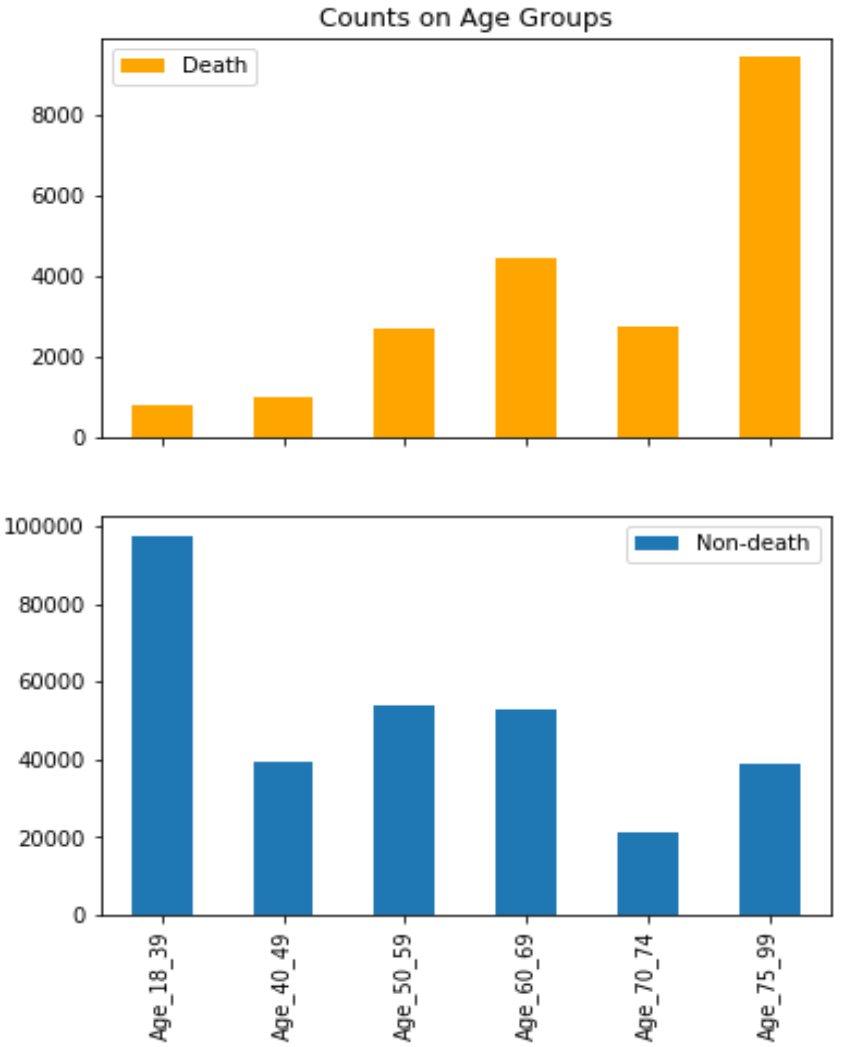
Cross Table of Fisher Exact Test

	Female (or unknown)	Male
Non-Death	193034	156329
Death	9197	12073

PART | 02 Descriptive Analysis (Cont)

Age and Mortality

- Plot bar charts on the number counts of different age groups on death & non-death patients
- Younger people are more tolerant on COVID-19
- Most death happened on elder group, especially aged over 75
- Death rate among people aged over 75 is the highest: **19.50%**



PART | 02 Descriptive Analysis (Cont)

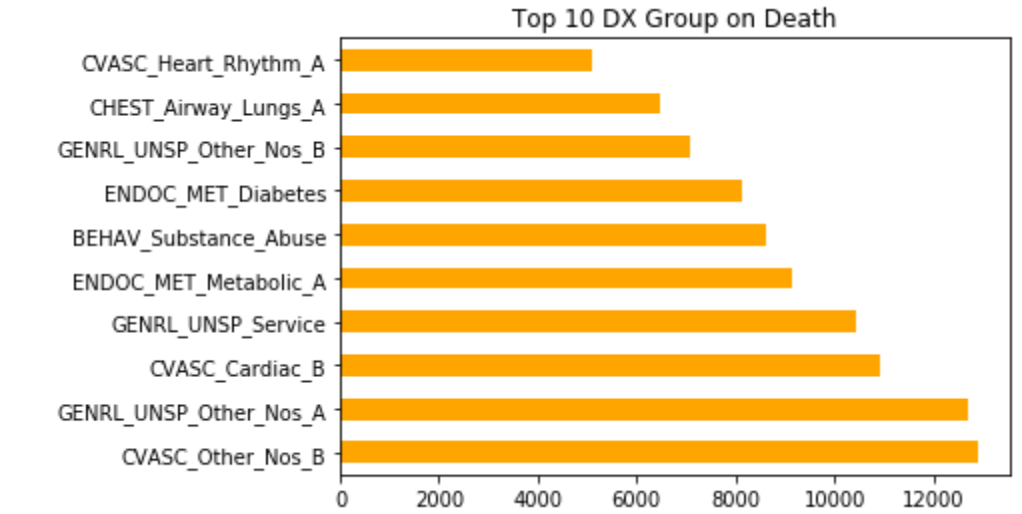
Dx Group and Mortality

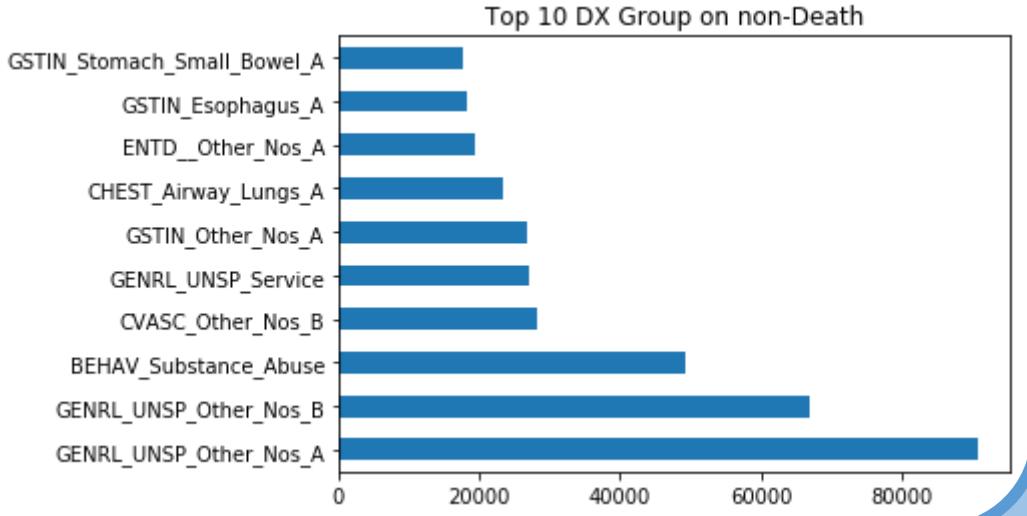
- Plot a bar chart on the number counts of different DX Group on death & non-death
- hard to directly discriminate between death and non-death due to some overlaps of DX groups
- Top DX groups on death patients:

CVAS (cerebrovascular accidents), General Unspecified group, Substance Abuse (e.g. alcohol abuse)

- Top DX groups on non-death patients:

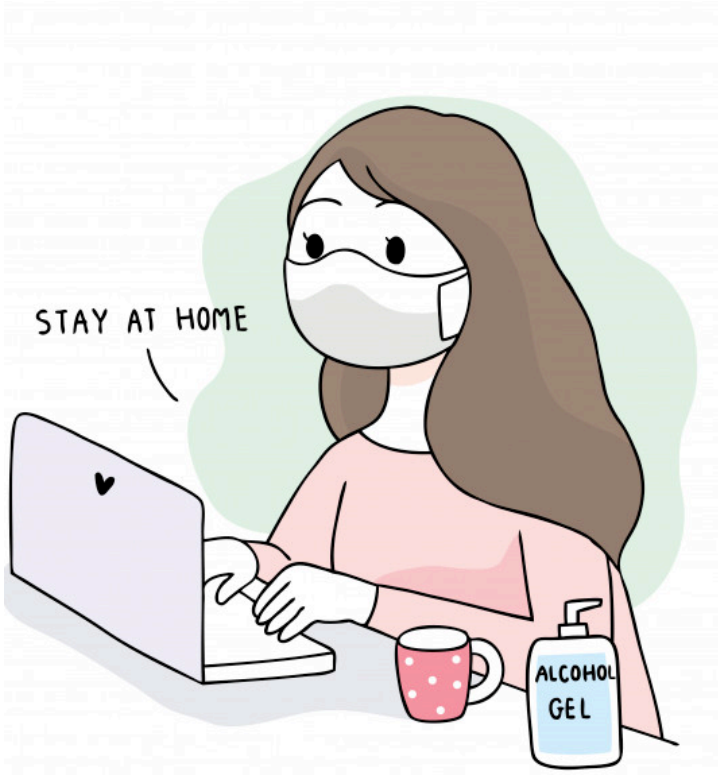
General Unspecified, GSTIN_Other (gastrointestinal related), Substance Abuse





PART | 03

Modeling Process

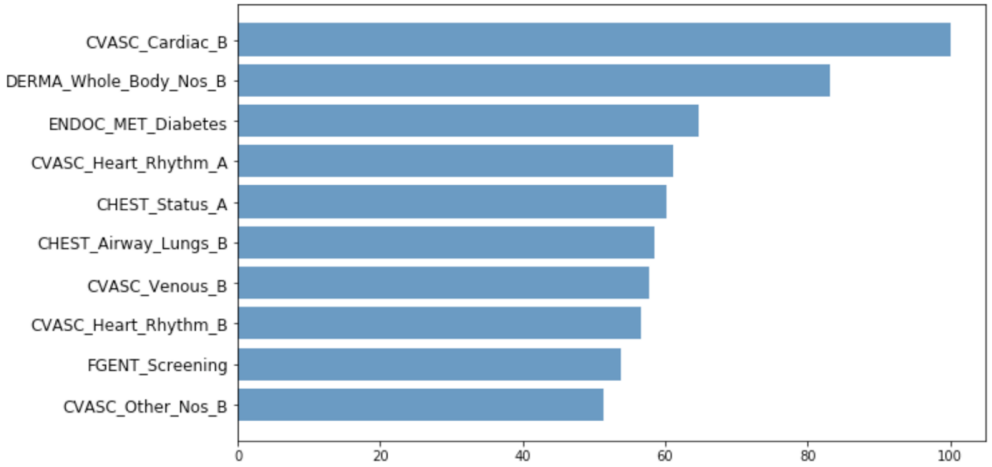


PART | 03 Logistics Regression

Model Introduction

- Used conditional probability in binary classification
- Needs a balanced dataset
- Provide coefficients of predictors for interpretation

Feature Importance of Logistics Model



Model Results

Confusion Matrix of Logistics Model

	Predicted non-death	Predicted death
True non-death	82436	4954
True death	352	4917

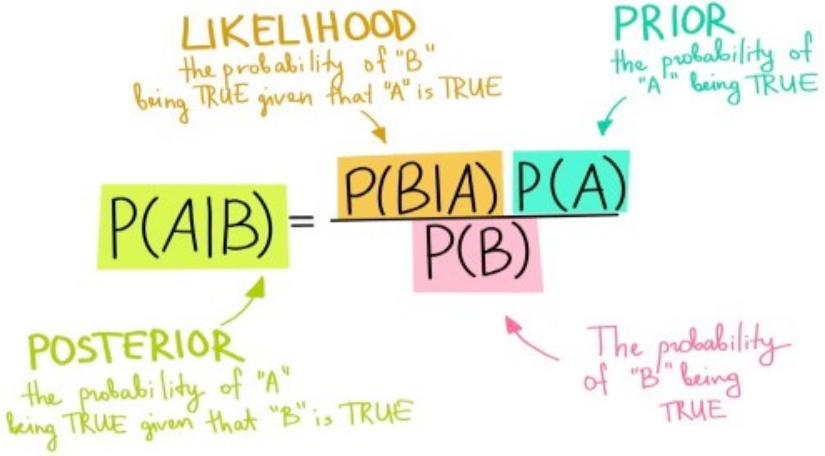
F1 Scores of Logistics Model

	precision	recall	f1-score	support
0	1.00	0.94	0.97	87390
1	0.50	0.93	**0.65**	5269
accuracy			0.94	92659

PART | 03 Bernoulli Naïve Bayes

Model Introduction

- Perform well on binary features (dummy variables)
- Not sensitive to unbalanced data
- Useful when the dataset is imbalanced



Model Results

Confusion Matrix of Bernoulli Naive Bayes

	Predicted non-death	Predicted death
True non-death	83878	3512
True death	688	4581

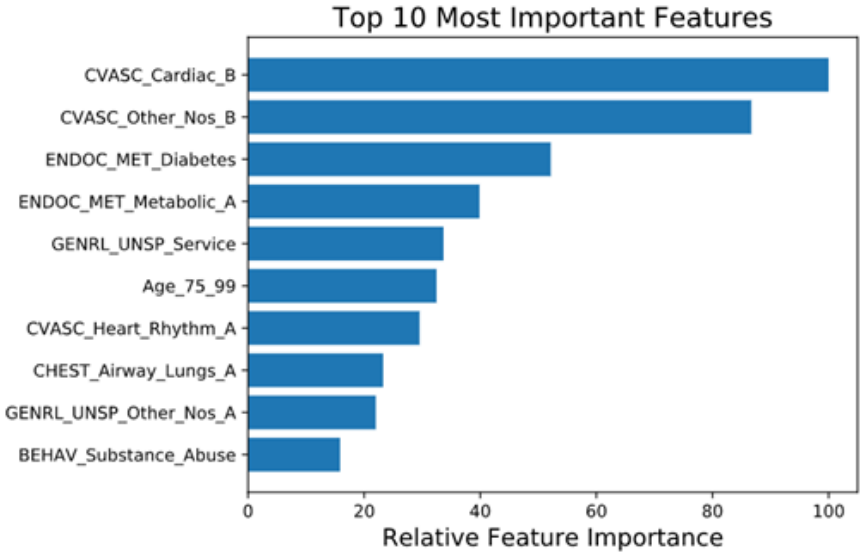
F1 Scores of Bernoulli Naive Bayes

	precision	recall	f1-score	support
0	0.992	0.960	0.976	87390
1	0.566	0.869	**0.686**	5269
accuracy			0.955	92659

PART | 03 ID3 Tree

Model Introduction

- Recursively searches the local best attribute with information gain
- Use bagging & bootstrap to train many ID3 Tree classifiers then aggregate into a powerful classifier



Model Results

Confusion Matrix of ID3 Tree

	Predicted non-death	Predicted death
True non-death	83659	3731
True death	518	4751

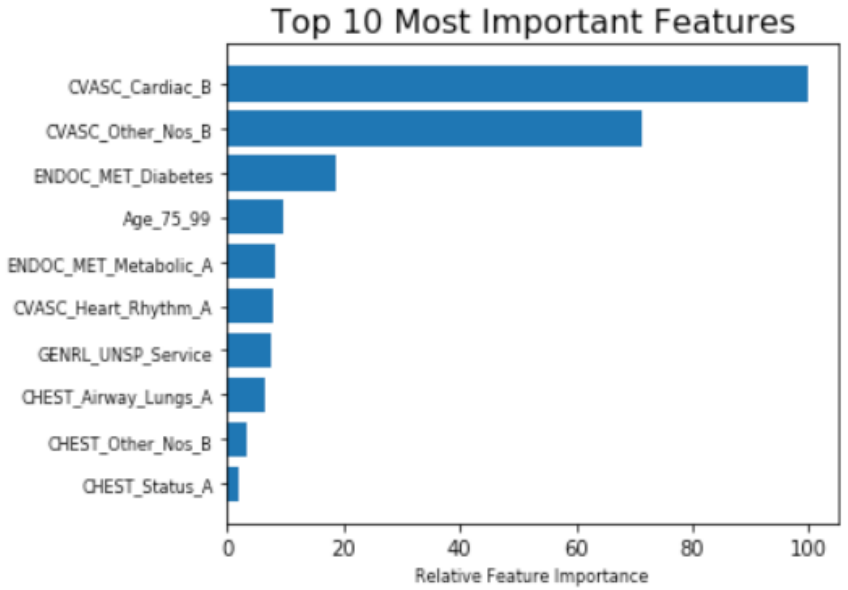
F1 Scores of ID3 Tree

	precision	recall	f1-score	support
0	0.99	0.96	0.98	87390
1	0.56	0.90	**0.69**	5269
accuracy			0.95	92659

PART | 03 Gradient Boosted Tree

Model Introduction

- Fits trees sequentially so that next tree focuses on misclassified records from the previous tree, improving accuracy
- Useful when the dataset is imbalanced



Model Results

Confusion Matrix of Gradient Boosted Tree

	Predicted non-death	Predicted death
True non-death	82624	4766
True death	374	4895

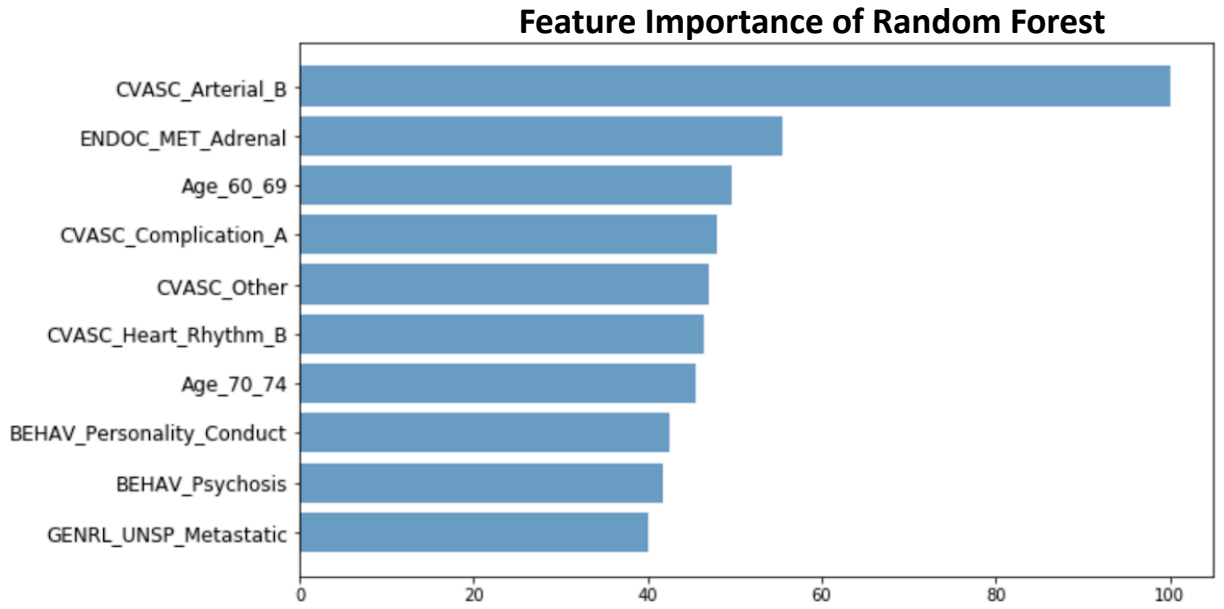
F1 Scores of Gradient Boosted Tree

	precision	recall	f1-score	support
0	1.00	0.95	0.97	87390
1	0.51	0.93	**0.66**	5269
accuracy			0.94	92659

PART | 03 Random Forest

Model Introduction

- Trains a bunch of trees in a parallel way o
  random subsets, and uses averaging to in
  the predictive accuracy and control overfi
- Needs a balanced dataset



Model Results

Confusion Matrix of Random Forest

	Predicted non-death	Predicted death
True non-death	86819	571
True death	1879	3390

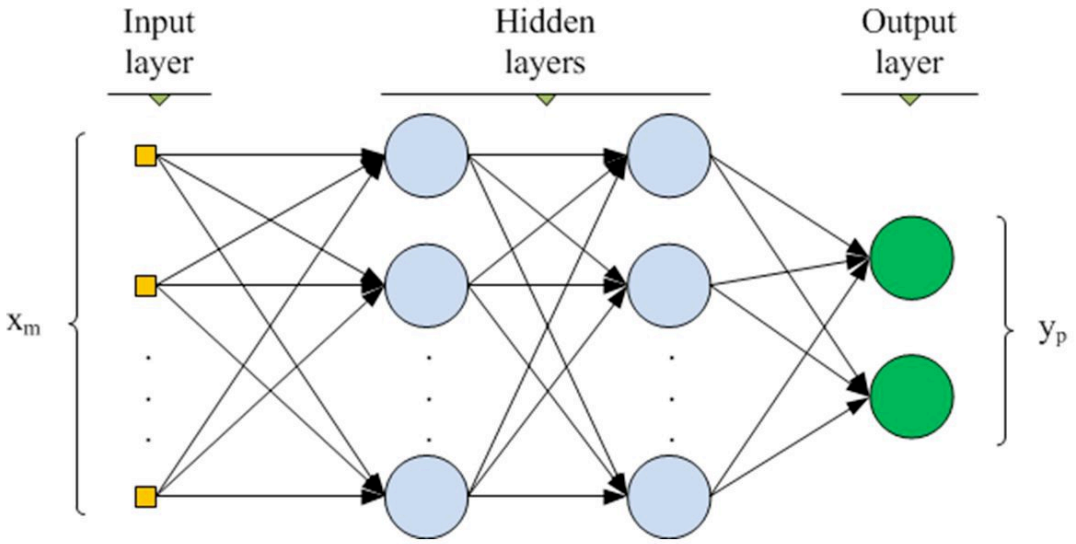
F1 Scores of Random Forest

	precision	recall	f1-score	support
0	0.98	0.99	0.99	87390
1	0.86	0.64	**0.73**	5269
accuracy			0.97	92659

PART | 03 Neural Network

Model Introduction

- Works as human brain to recognize patterns learn
- Good at modeling non-linear and complex relationships
- Needs a balanced dataset



Model Results

Confusion Matrix of Random Forest

	Predicted non-death	Predicted death
True non-death	84377	3013
True death	433	4836

F1 Scores of Random Forest

	precision	recall	f1-score	support
0	0.99	0.97	0.98	87390
1	0.62	0.92	**0.74**	5269
accuracy			0.96	92659

PART | 04

Model Comparison

PART | 04 Model Comparison

Table of Model Performance Comparison

	Precision	Recall	f1-score	Accuracy
Logistic Regression	0.5	**0.93**	0.65	0.94
Naïve Bayes	0.57	0.87	0.69	0.96
Bagging ID3 Tree	0.56	0.9	0.69	0.95
Random Forest	**0.86**	0.64	0.73	0.97
Boosting Tree	0.51	**0.93**	0.66	0.94
Neural Networks	0.62	0.92	**0.74**	0.96

Explanation

- Highest f1-score (balance in classifying both classes)

Neural Network (74%)

- Highest precision (performance in making classification in the *minority* class but more likely to misclassify the *majority* class)

Random Forest (86%)

- Highest Recall Rate (for the cases that are dead, the percent of correct predictions)

Logistic Regression & Boosting Tree (93%)

PART | 04 Model Comparison (Cont)

Table of Feature Importance

	Boosting Tree	Random Forest	Bagging ID3 Tree	Logistic Regression
Age_60_69		3		
Age_70_74		7		
Age_75_99	**7**		**6**	
BEHAV_Personality_Conduct		8		
BEHAV_Psychosis		9		
BEHAV_Substance_Abuse			10	
CHEST_Airway_Lungs_A	**3**		**8**	
CHEST_Airway_Lungs_B				6
CHEST_Other_Nos_B	**2**		**2**	
CHEST_Status_A	**1**		**1**	**5**
CVASC_Arterial_B		1		
CVASC_Cardiac_B	**10**			**1**
CVASC_Complication_A		4		

PART | 04 Model Comparison (Cont)

Table of Feature Importance (cont)

	Boosting Tree	Random Forest	Bagging ID3 Tree	Logistic Regression
CVASC_Heart_Rhythm_A	**5**		**7**	**4**
CVASC_Heart_Rhythm_B		**6**		**8**
CVASC_Other		5		
CVASC_Other_Nos_B				10
CVASC_Other_Nos_B	9			
CVASC_Venous_B				7
DERMA_Whole_Body_Nos_B				2
ENDOC_MET_Adrenal		2		
ENDOC_MET_Diabetes	**8**		**3**	**3**
ENDOC_MET_Metabolic_A	**6**		**4**	
FGENT_Screening				9
GENERL_UNSP_Metastatic		10		
GENRL_UNSP_Other_Nos_A			9	
GENRL_UNSP_Service	**4**		**5**	

PART | 05

Insights

PART | 05 Insights

Important Feature Description (From our models)

DGL_3_Extend	Description
CVASC_Cardiac_B	Cardiac diseases such as heart failure
CVASC_Other_Nos_b	Mixed diseases include lung, heart, blood pressure, blood vessel, etc.
ENDOC_MET_Diabetes	Diabetes and concomitant disease like diabetic chronic kidney disease
ENDOC_MET_Metabolic_A	Metabolic abnormality such as obesity
GENRL_UNSP_Service	Aftercare, counseling, and medical exam
Age_75_99	Elder population
CVASC_Heart_Rhythm_A	Irregular heartbeats such as tachycardia
CHEST_Airway_Lungs_A	Lung and respiratory abnormalities
GENRL_UNSP_Other_Nos_A	Mixed diseases include infection of liver, immunodeficiency, etc.
CHEST_Status_A	Lung and respiratory abnormalities

Patient Persona (From CDC)

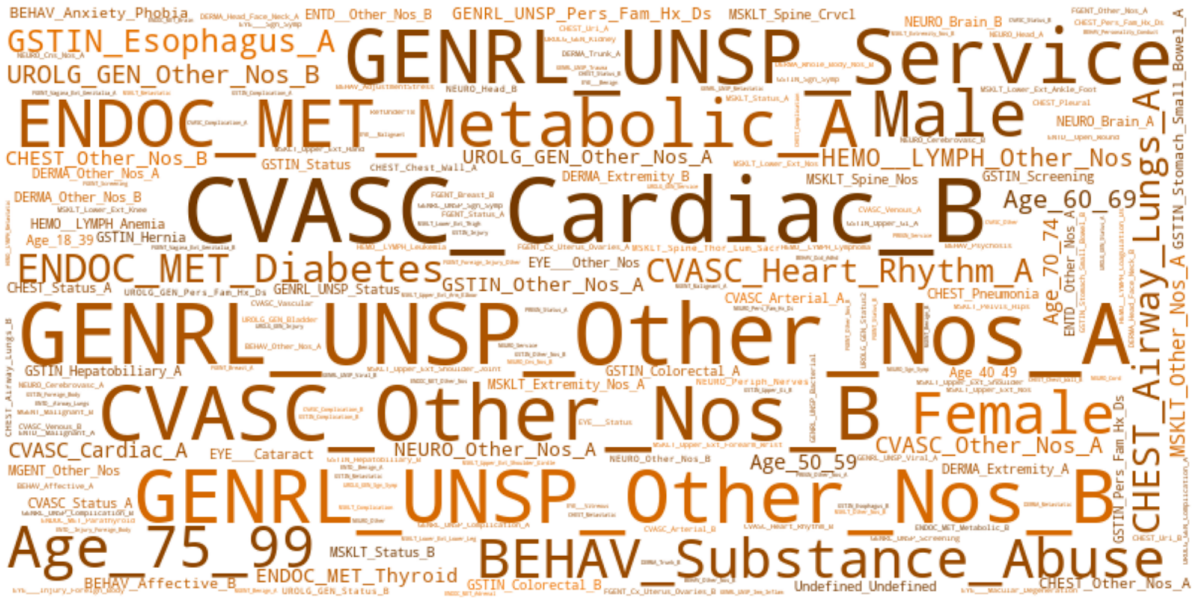
Cross Validation

- A 78 year-old male with COVID-19
- Had cardiac and heart rhythm diseases
- Contracted a severe fever for 4 days
- Died in the end 💔

- A 30 year-old female with COVID-19
- Had diabetes, chronic kidney disease, obesity issues and respiratory failure
- Relied on medical exam
- Had Died in the end 💔

PART | 05 Insights(Cont)

Text Analysis (Word Cloud)



Conducting Process

- Filter all of the patient information (Age, gender and DXs) from the death group and calculate frequency according to counts

Findings

- People with some heart disease, especially classified in Cardiac_B and Other_Nos_B, have high risk of death
- GENRL_UNSP_OTHER_Nos_A ranks second, which means the abnormal findings lab or exam, blood results
- ENDOC_MET_Metabolic is also a large category leading to death
- Age 75-99 is the majority of death population from COVID-19

PART | 06

Conclusion

PART | 06 Conclusion

What we did

A. Data manipulation

B. Machine learning models (Logistic Regression, Naive Bayes, ID3 Tree, Random Forest, Boosted Tree and Neural Networks)

C. Result interpretation

D. Text analysis

E. Cross Validation

What we got

A. Feature importance

B. Patient persona

- Demographics: 65 years and older
- Gender difference: Male in high-risk
- Disease Related: chronic lung disease, serious heart conditions, immunocompromised, severe obesity, diabetes, chronic kidney disease or liver disease

A. Insights and suggestions