Iowa Initiative for Artificial Intelligence

Final Report

Project title:	Develop examples of data visualizations that can be created using NVDRS data
Principal Investigator:	Cori Peek-Asa, Lisa Roth
Prepared by (IIAI):	Nam H. Le
Other investigators:	Ling Zhang
Date:	October 15, 2022
Were specific aims fulfilled:	YES
Readiness for extramural proposal?	
If yes … Planned submission date	
Funding agency	
Grant mechanism	
If no … Why not? What went wrong?	

Brief summary of accomplished results:

Created a method for data preprocessing for NVDRS data.

Evaluated and selected the best ML model to evaluate how circumstance variables can be used to predict firearm-related suicide incidents.

Provided various visualizations methods to explore feature rankings.

Research report:

- Aim:

Prior Aim: Using the IAVDRS, create a data platform that partners can use to visualize data

Updated Aim: Develop examples of data visualizations that can be created using NVDRS data. Visualizations will communicate how various circumstances and/or characteristics are related to types of suicide death (e.g., firearm related suicide).

- Data Description: (taken from the statement of intent)

The CDC/National Center for Injury Prevention and Control established the National Violent Death Reporting System (NVDRS) in 2002 (https://www.cdc.gov/violenceprevention/datasources/nvdrs/index.html). The objectives of the NVDRS are to establish a national surveillance system of information about violent deaths, collecting detailed information to better understand the causal risk and protective factors. At its inception in 2002, data from six states were included. States were phased in through 2019, when the system became national. Not all states are yet reporting statewide results; however, the system is now approximately 80% complete with weighting schemes for rate estimation. Violent deaths include all homicides and suicides identified through death certificates, autopsy reports, law enforcement investigation reports, and crime scene analysis, and detailed data about the violent event from each of these sources are available.

The data includes information at the event level (e.g. homicide/suicide/multiple; date/time), the victim(s) level and the perpetrator(s) (e.g. sociodemographic variables). Circumstances of the death include factors such as substance use, history of mental health issues, financial problems, relationship problems, work problems, legal issues, and health issues, among others. Circumstance variables identify if each issue was a problem, and furthermore identifies which were crises at the time of the event (noted in the investigation that the issue was a precipitating factor or present within two weeks of the event). Firearm information includes the make, type, and caliber of weapon, as well as the time of purchase and information about the registered owner.

AI/ML Approach:

The original task is "Using the NVDRS, use machine learning algorithms to ***identify differences in circumstances between firearm and non-firearm violent death.***" This project aims to demonstrate various methods of visualize the results taken from the ML task.

The first step was to prepare data for ML tasks, including:

- Variable selection criteria: 'Circumstance', background, and stressors variables that
  - Have more than 1 unique category
  - Unknown category accounts for <99% of total samples
- Data clean-up:
  - removed entries without age and sex information,
  - removed entries that contain unreadable categories
- Defining outcome variable
- Handling of missing and unknown values
- Renaming categories and variables for readability
- Conversion of categorical variables to nominal variables
- Regrouping of high-cardinality variables
- One-hot encoding of categorical variables and dropping redundant variable
- Creation of new variables
- Creation of data splits for stratified group 5-fold cross-validation

Experimental Methods, Validation Approach

I. Data Preprocessing

Please refer to Aim 1's report for detailed data preprocessing methods.

The overall data preprocessing pipeline includes:

- Variable selection criteria: 'Circumstance', background, and stressors variables that
  - Have more than 1 unique category
  - Unknown category accounts for <99% of total samples
- Data clean-up:
  - removed entries without age and sex information,
  - removed entries that contain unreadable categories
- Defining outcome variable
- Handling of missing and unknown values
- Renaming categories and variables for readability
- Conversion of categorical variables to nominal variables
- Regrouping of high-cardinality variables
- One-hot encoding of categorical variables and dropping redundant variable
- Creation of new variables
- Creation of data splits for stratified group 5-fold cross-validation

As a result of data preprocessing, the number of input variables, or 'features', is **141**.

Table. Input Variable List Selected for Study

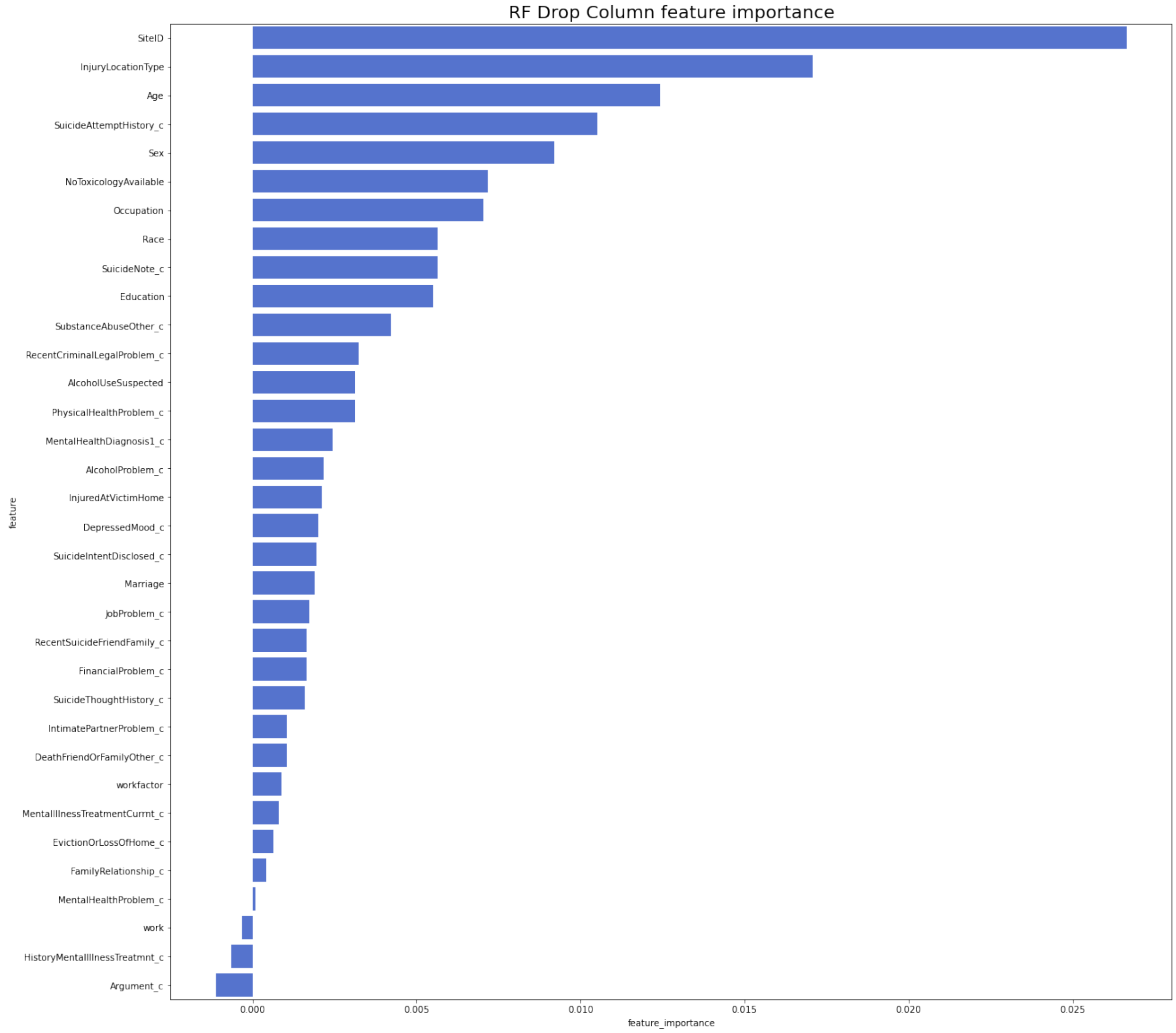
Variable Name	Type
1. Background	
Age Group	Numeric
Marital Status	Categorical
Education Level	Categorical
Occupation	
2. Disparity-prone category	
Sex (Male)	Binary
Race	Categorical
3. Time, events, locations of suicide (controversial as is it actually a circumstance)	
State	Categorical
Injury-related Location Type	Categorical
4. Mental health	
Mental Health Diagnosis	Categorical

Mental Health Problem	Binary
Depressed	Binary
Current Mental Illness Treatment	Binary
History of Mental Illness Treatment	Binary
5. Addictions	
Alcohol Problem	Binary
Substance Abuse Problem	Binary
Addiction Seriousness	Numeric
6. Relationships	
Intimate Partner Problem	Binary
Had Argument	Binary
7. Life stressors	
Eviction/Loss of Home	Binary
Recent Suicide Friend Family	Binary
Work Factor	Numeric
Work Factor (yes/no)	Binary
Job Problem	Binary
Family Problem	Binary
Financial Problem	Binary
Death Friend/Family	Binary
8. Heath problems	
Physical Health Problem	Binary
9. Suicide Intentions	
History of Attempted Suicide	Binary
History of Suicidal Thoughts	Binary
Suicide Intent Disclosed	Binary
Suicide Note	Binary
Suicide Intention Seriousness	Numeric
10. Specific Circumstances	Binary
Criminal Legal Problem	Binary
Injured at Home	Binary
Alcohol Use Suspected	Binary
11. Other	
Toxicology Information Availability	Binary

II. Visualization Methods

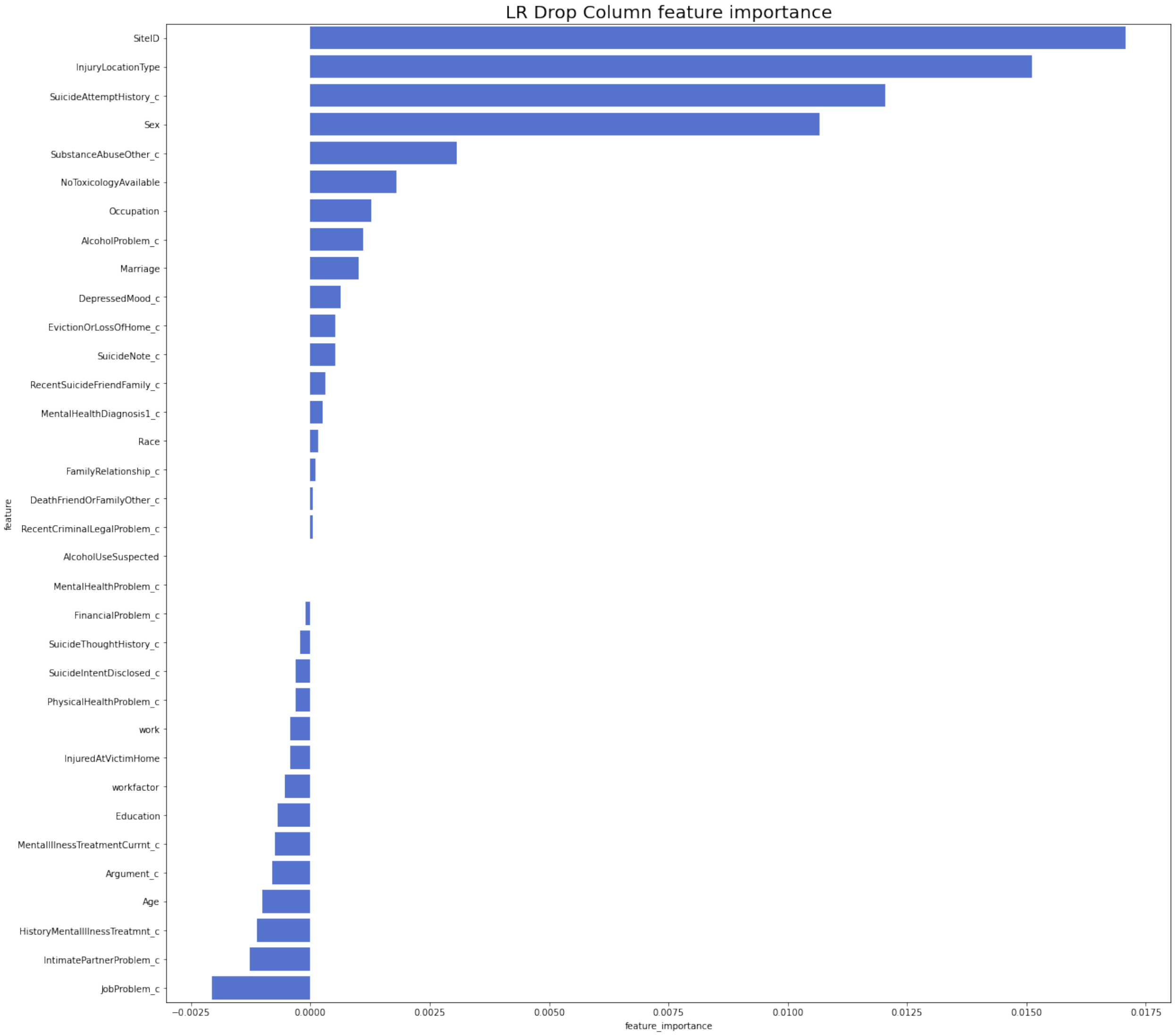
1. Feature rankings using model-based methods

Purpose: Feature importance of the random forest, using the mean decrease of gini impurity, is a mean to understand predictor powers of individual variables for the predictive task.



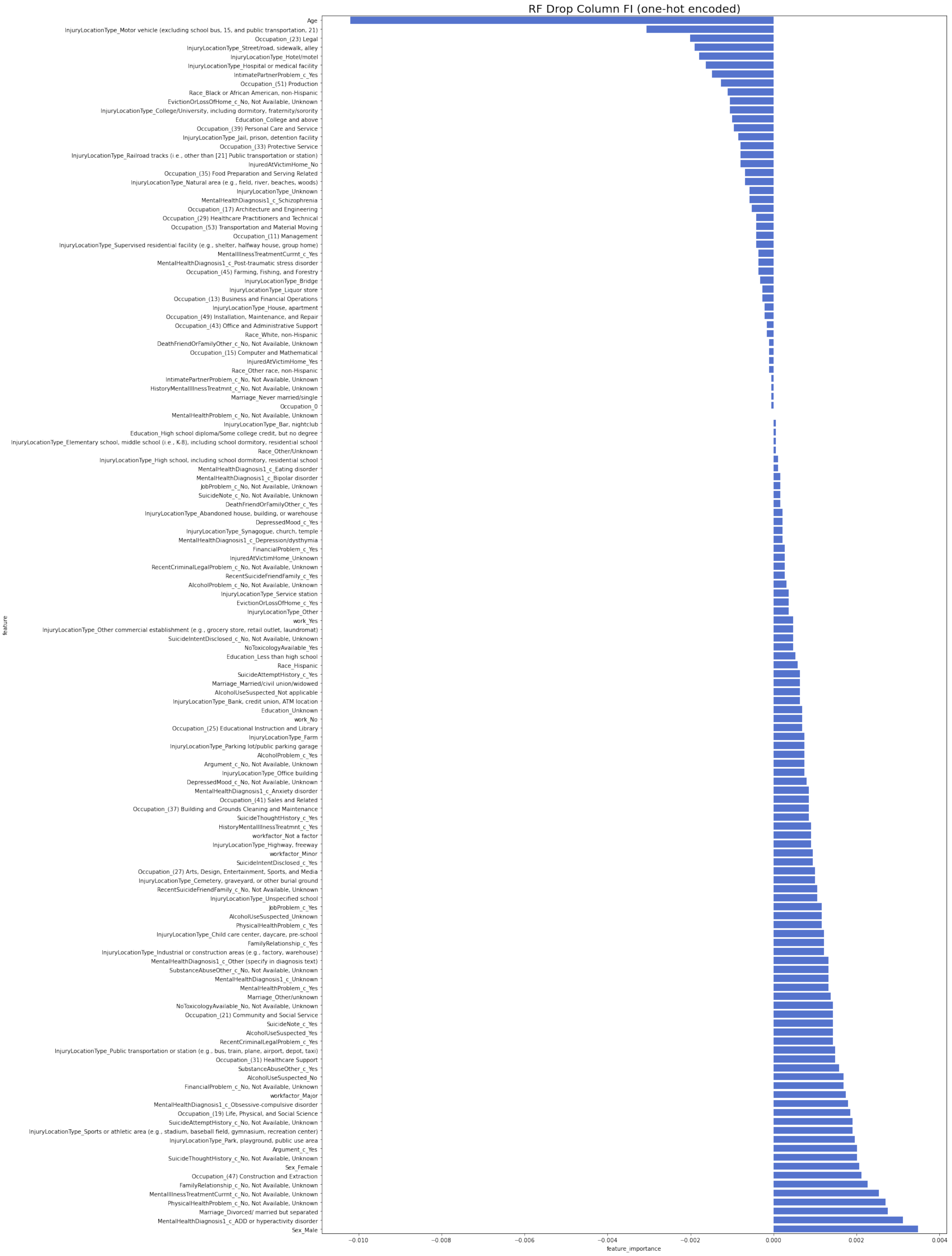
This method suffers from poor interpretability, the plot can be explained from the information-theory aspect but not from the clinical side.

On the other hands, the logistic regression coefficients produced by the logistic regression classifier can be explained this way: variables whose coefficients are positive are associated with the positive outcome (suicide using a firearm), variables whose coefficients are negative are associated with the negative outcome (suicide using another mean).


LR Drop Column feature importance
SiteID
InjuryLocationType
SuicideAttemptHistory_c
Sex
SubstanceAbuseOther_c
NoToxicologyAvailable
Occupation
AlcoholProblem_c
Marriage
DepressedMood_c
EvictionOrLossOfHome_c
SuicideNote_c
RecentSuicideFriendFamily_c
MentalHealthDiagnosis1_c
Race
FamilyRelationship_c
DeathFriendOrFamilyOther_c
RecentCriminalLegalProblem_c
AlcoholUseSuspected
MentalHealthProblem_c
FinancialProblem_c
SuicideThoughtHistory_c
SuicideIntentDisclosed_c
PhysicalHealthProblem_c
work
InjuredAtVictimHome
workfactor
Education
MentalIllnessTreatmentCurrnt_c
Argument_c
Age
HistoryMentalIllnessTreatmnt_c
IntimatePartnerProblem_c
JobProblem_c
feature
−0.0025
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
feature_importance

2. Model global-agnostic methods

a) Feature ranking using drop-one-column


RF Drop Column FI (one-hot encoded)
feature
Age
InjuryLocationType_Motor vehicle (excluding school bus, 15, and public transportation, 21)
Occupation_(23) Legal
InjuryLocationType_Street/road, sidewalk, alley
InjuryLocationType_Hotel/motel
InjuryLocationType_Hospital or medical facility
IntimatePartnerProblem_c_Yes
Occupation_(51) Production
Race_Black or African American, non-Hispanic
EvictionOrLossOfHome_c_No, Not Available, Unknown
InjuryLocationType_College/University, including dormitory, fraternity/sorority
Education_College and above
Occupation_(39) Personal Care and Service
InjuryLocationType_Jail, prison, detention facility
Occupation_(33) Protective Service
InjuryLocationType_Railroad tracks (i.e., other than [21] Public transportation or station)
InjuredAtVictimHome_No
Occupation_(35) Food Preparation and Serving Related
InjuryLocationType_Natural area (e.g., field, river, beaches, woods)
InjuryLocationType_Unknown
MentalHealthDiagnosis1_c_Schizophrenia
Occupation_(17) Architecture and Engineering
Occupation_(29) Healthcare Practitioners and Technical
Occupation_(53) Transportation and Material Moving
Occupation_(11) Management
InjuryLocationType_Supervised residential facility (e.g., shelter, halfway house, group home)
MentalIllnessTreatmentCurrnt_c_Yes
MentalHealthDiagnosis1_c_Post-traumatic stress disorder
Occupation_(45) Farming, Fishing, and Forestry
InjuryLocationType_Bridge
InjuryLocationType_Liquor store
Occupation_(13) Business and Financial Operations
InjuryLocationType_House, apartment
Occupation_(49) Installation, Maintenance, and Repair
Occupation_(43) Office and Administrative Support
Race_White, non-Hispanic
DeathFriendOrFamilyOther_c_No, Not Available, Unknown
Occupation_(15) Computer and Mathematical
InjuredAtVictimHome_Yes
Race_Other race, non-Hispanic
IntimatePartnerProblem_c_No, Not Available, Unknown
HistoryMentalIllnessTreatmnt_c_No, Not Available, Unknown
Marriage_Never married/single
Occupation_0
MentalHealthProblem_c_No, Not Available, Unknown
InjuryLocationType_Bar, nightclub
Education_High school diploma/Some college credit, but no degree
InjuryLocationType_Elementary school, middle school (i.e., K-8), including school dormitory, residential school
Race_Other/Unknown
InjuryLocationType_High school, including school dormitory, residential school
MentalHealthDiagnosis1_c_Eating disorder
MentalHealthDiagnosis1_c_Bipolar disorder
JobProblem_c_No, Not Available, Unknown
SuicideNote_c_No, Not Available, Unknown
DeathFriendOrFamilyOther_c_Yes
InjuryLocationType_Abandoned house, building, or warehouse
DepressedMood_c_Yes
InjuryLocationType_Synagogue, church, temple
MentalHealthDiagnosis1_c_Depression/dysthymia
FinancialProblem_c_Yes
InjuredAtVictimHome_Unknown
RecentCriminalLegalProblem_c_No, Not Available, Unknown
RecentSuicideFriendFamily_c_Yes
AlcoholProblem_c_No, Not Available, Unknown
InjuryLocationType_Service station
EvictionOrLossOfHome_c_Yes
InjuryLocationType_Other
work_Yes
InjuryLocationType_Other commercial establishment (e.g., grocery store, retail outlet, laundromat)
SuicideIntentDisclosed_c_No, Not Available, Unknown
NoToxicologyAvailable_Yes
Education_Less than high school
Race_Hispanic
SuicideAttemptHistory_c_Yes
Marriage_Married/civil union/widowed
AlcoholUseSuspected_Not applicable
InjuryLocationType_Bank, credit union, ATM location
Education_Unknown
work_No
Occupation_(25) Educational Instruction and Library
InjuryLocationType_Farm
InjuryLocationType_Parking lot/public parking garage
AlcoholProblem_c_Yes
Argument_c_No, Not Available, Unknown
InjuryLocationType_Office building
DepressedMood_c_No, Not Available, Unknown
MentalHealthDiagnosis1_c_Anxiety disorder
Occupation_(41) Sales and Related
Occupation_(37) Building and Grounds Cleaning and Maintenance
SuicideThoughtHistory_c_Yes
HistoryMentalIllnessTreatmnt_c_Yes
workfactor_Not a factor
InjuryLocationType_Highway, freeway
workfactor_Minor
SuicideIntentDisclosed_c_Yes
Occupation_(27) Arts, Design, Entertainment, Sports, and Media
InjuryLocationType_Cemetery, graveyard, or other burial ground
RecentSuicideFriendFamily_c_No, Not Available, Unknown
InjuryLocationType_Unspecified school
JobProblem_c_Yes
AlcoholUseSuspected_Unknown
PhysicalHealthProblem_c_Yes
InjuryLocationType_Child care center, daycare, pre-school
FamilyRelationship_c_Yes
InjuryLocationType_Industrial or construction areas (e.g., factory, warehouse)
MentalHealthDiagnosis1_c_Other (specify in diagnosis text)
SubstanceAbuseOther_c_No, Not Available, Unknown
MentalHealthDiagnosis1_c_Unknown
MentalHealthProblem_c_Yes
Marriage_Other/unknown
NoToxicologyAvailable_No, Not Available, Unknown
Occupation_(21) Community and Social Service
SuicideNote_c_Yes
AlcoholUseSuspected_Yes
RecentCriminalLegalProblem_c_Yes
InjuryLocationType_Public transportation or station (e.g., bus, train, plane, airport, depot, taxi)
Occupation_(31) Healthcare Support
SubstanceAbuseOther_c_Yes
AlcoholUseSuspected_No
FinancialProblem_c_No, Not Available, Unknown
workfactor_Major
MentalHealthDiagnosis1_c_Obsessive-compulsive disorder
Occupation_(19) Life, Physical, and Social Science
SuicideAttemptHistory_c_No, Not Available, Unknown
InjuryLocationType_Sports or athletic area (e.g., stadium, baseball field, gymnasium, recreation center)
InjuryLocationType_Park, playground, public use area
Argument_c_Yes
SuicideThoughtHistory_c_No, Not Available, Unknown
Sex_Female
Occupation_(47) Construction and Extraction
FamilyRelationship_c_No, Not Available, Unknown
MentalIllnessTreatmentCurrnt_c_No, Not Available, Unknown
PhysicalHealthProblem_c_No, Not Available, Unknown
Marriage_Divorced/ married but separated
MentalHealthDiagnosis1_c_ADD or hyperactivity disorder
Sex_Male
−0.010
−0.008
−0.006
−0.004
−0.002
0.000
0.002
0.004
feature_importance

The drop-one-column feature importance index gives the amount of the metric score (in this work, F-1), the classifier gains or lose when a variable is added at the training stage.

Table 1. Top 30 features using Feature Importance averaged from three best performing models based on F1 score (versus a baseline model)

	FEATURE	FI_XGBOOST	FI_LR	FI_LSVC	MEAN	SD	LOWER.Z	UPPER.Z	SIGNIFICANT.Z	LOWER.T	UPPER.T	SIGNIFICANT.T
1	Sex (Male)_Male	0.0105	0.0069	0.0057	0.0077	0.0025	0.0049	0.0105	Yes	0.0015	0.0139	Yes
2	Toxicology Information Availability_No	0.0068	0.0025	0.0012	0.0035	0.0029	0.0002	0.0068	Yes	-0.0037	0.0107	No
3	Injury-related Location Type_Detention facility	0.0030	0.0017	0.0017	0.0021	0.0008	0.0012	0.0030	Yes	0.0001	0.0041	Yes
4	History of Attempted Suicide	0.0059	-0.0001	-0.0001	0.0019	0.0035	-0.0021	0.0059	No	-0.0068	0.0106	No
5	Injury-related Location Type_Bridge	0.0026	0.0017	0.0010	0.0018	0.0008	0.0009	0.0027	Yes	-0.0002	0.0038	No
6	State_Massachusetts	0.0036	0.0000	-0.0005	0.0010	0.0022	-0.0015	0.0035	No	-0.0045	0.0065	No
7	Injury-related Location Type_Railroad tracks	0.0006	0.0016	0.0004	0.0009	0.0006	0.0002	0.0016	Yes	-0.0006	0.0024	No
8	Injured at Home	0.0012	0.0013	0.0000	0.0008	0.0007	0.0000	0.0016	Yes	-0.0009	0.0025	No
9	Injury-related Location Type_Street, sidewalk, alley	0.0009	0.0005	0.0007	0.0007	0.0002	0.0005	0.0009	Yes	0.0002	0.0012	Yes
10	Injury-related Location Type_Hotel/motel	0.0010	0.0005	0.0005	0.0007	0.0003	0.0004	0.0010	Yes	0.0000	0.0014	No
11	Mental Health Diagnosis_Post-traumatic stress disorder	0.0010	0.0009	0.0001	0.0007	0.0005	0.0001	0.0013	Yes	-0.0005	0.0019	No
12	Race_Other race, non-Hispanic	0.0001	0.0013	0.0003	0.0006	0.0006	-0.0001	0.0013	No	-0.0009	0.0021	No
13	Injury-related Location Type_Parking area	0.0002	0.0006	0.0008	0.0005	0.0003	0.0002	0.0008	Yes	-0.0002	0.0012	No
14	Injury-related Location Type_Motor vehicle	0.0015	0.0001	-0.0002	0.0005	0.0009	-0.0005	0.0015	No	-0.0017	0.0027	No
15	State_New Jersey	0.0021	-0.0001	-0.0006	0.0005	0.0014	-0.0011	0.0021	No	-0.0030	0.0040	No
16	Injury-related Location Type_House, apartment	0.0003	0.0006	0.0005	0.0005	0.0002	0.0003	0.0007	Yes	0.0000	0.0010	Yes
17	Race_Black	0.0014	-0.0001	0.0000	0.0004	0.0008	-0.0005	0.0013	No	-0.0016	0.0024	No
18	Injury-related Location Type_Highway, freeway	0.0004	0.0008	-0.0001	0.0004	0.0005	-0.0002	0.0010	No	-0.0008	0.0016	No
19	Mental Health Problem	0.0008	0.0006	-0.0004	0.0003	0.0006	-0.0004	0.0010	No	-0.0012	0.0018	No
20	Marital Status_Never	0.0002	0.0005	0.0001	0.0003	0.0002	0.0001	0.0005	Yes	-0.0002	0.0008	No
21	Intimate Partner Problem	0.0011	0.0001	-0.0003	0.0003	0.0007	-0.0005	0.0011	No	-0.0014	0.0020	No
22	Injury-related Location Type_Other commercial establishment	-0.0002	0.0009	0.0002	0.0003	0.0006	-0.0004	0.0010	No	-0.0012	0.0018	No
23	Injury-related Location Type_Public transportation/station	0.0002	0.0005	0.0000	0.0002	0.0003	-0.0001	0.0005	No	-0.0005	0.0009	No
24	Occupation_Retired, Students, Unemployed	0.0008	0.0000	0.0000	0.0003	0.0005	-0.0003	0.0009	No	-0.0009	0.0015	No
25	Death Friend/Family	-0.0002	0.0003	0.0006	0.0002	0.0004	-0.0003	0.0007	No	-0.0008	0.0012	No
26	Occupation_Installation, Maintenance, Repair	0.0007	0.0000	0.0000	0.0002	0.0004	-0.0003	0.0007	No	-0.0008	0.0012	No
27	Job Problem	0.0001	0.0005	0.0000	0.0002	0.0003	-0.0001	0.0005	No	-0.0005	0.0009	No
28	Injury-related Location Type_Supervised residential facility	-0.0001	0.0008	-0.0001	0.0002	0.0005	-0.0004	0.0008	No	-0.0010	0.0014	No
29	Physical Health Problem	-0.0001	0.0008	-0.0001	0.0002	0.0005	-0.0004	0.0008	No	-0.0010	0.0014	No
30	Occupation_Protective Service	0.0008	-0.0001	-0.0001	0.0002	0.0005	-0.0004	0.0008	No	-0.0010	0.0014	No

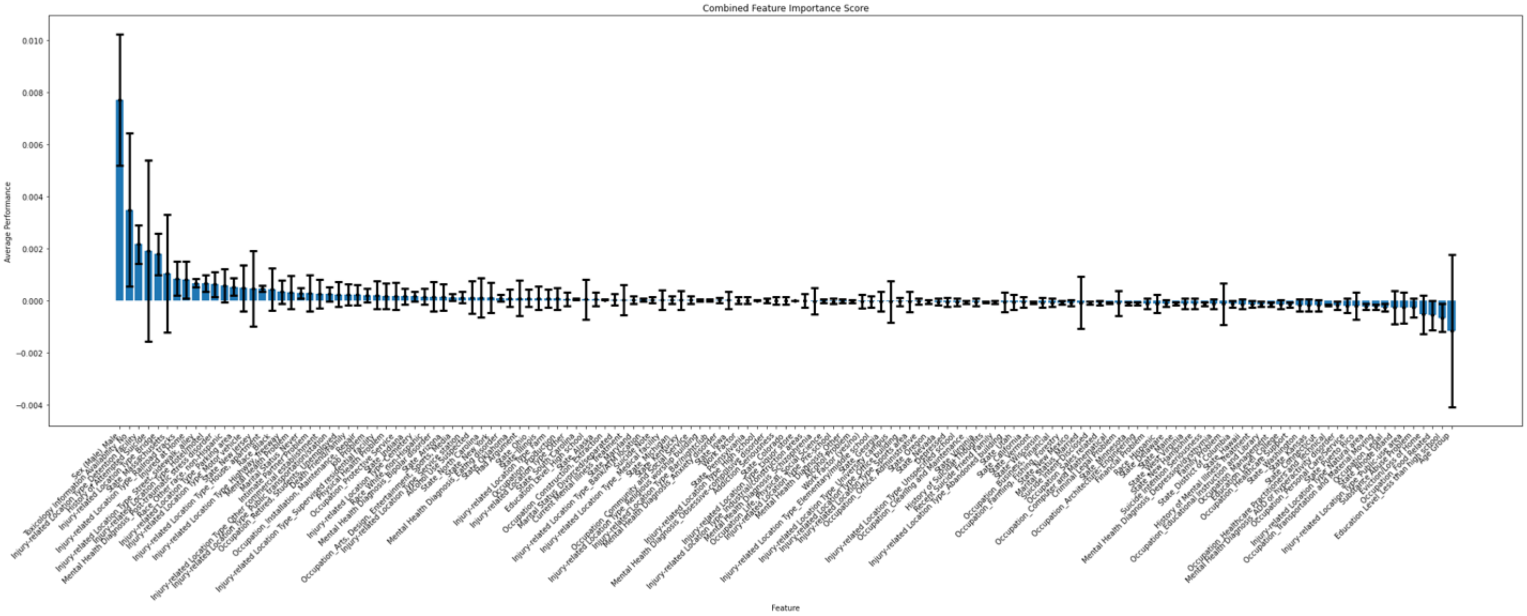
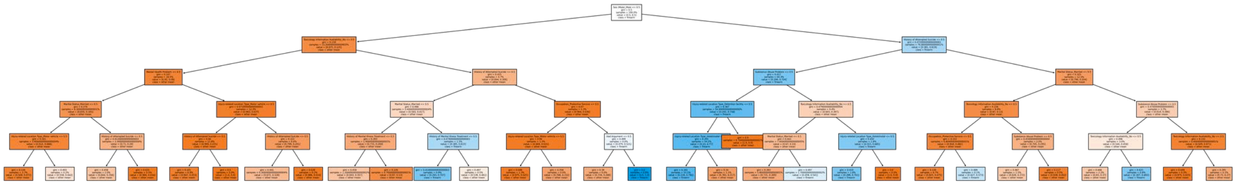


Figure. Feature Importance, including StateID

3. Knowledge distillation into a decision tree

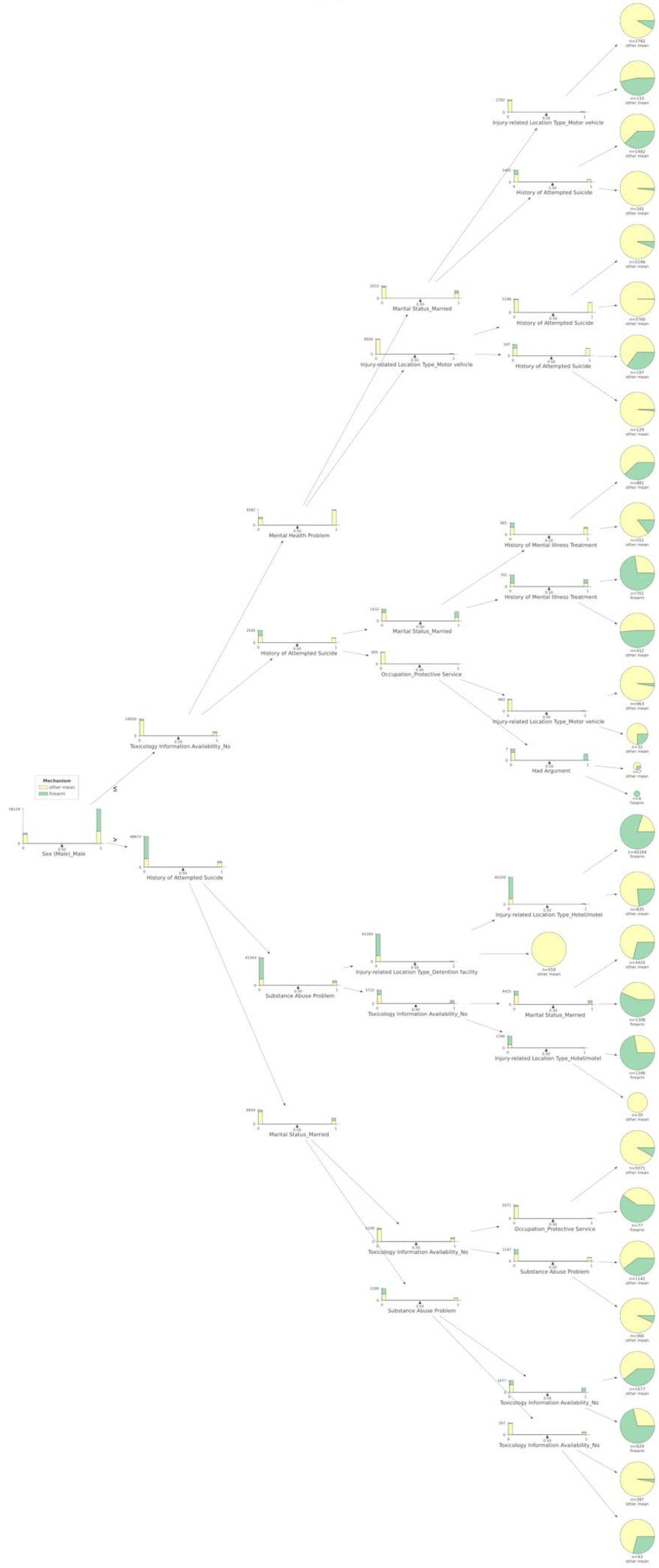
Purpose:

The random forest classifier use a collection of predictions made by decision trees to select the final prediction using majority voting. This method summary the decision paths of the random forest into a single tree that can be easily interpreted.



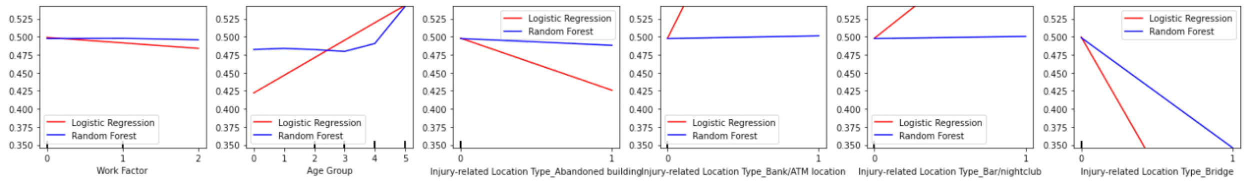
GINI-INDEX RANK	5-FACTOR CIRCUMSTANCE COMBINATION	GINI INDEX (LOWER IS BETTER)	POSITIVE CASES	% OF ALL SAMPLES
1	Sex: Female Toxicology Information Availability: No **History of Attempted Suicide: Yes** **Occupation_Protective Service: Yes** **Had Argument: Yes**	0	100%	0%
2	**Sex: Male** History of Attempted Suicide: No Substance Abuse Problem: No Injury-related Location Type_Detention facility: No Injury-related Location Type_Hotel/motel: No	0.341	78%	53%
3	Sex: Female Toxicology Information Availability: No History of Attempted Suicide: No **Marital Status_Married: Yes** History of Mental Illness Treatment: No	0.414	71%	1%
4	**Sex: Male** History of Attempted Suicide: No **Substance Abuse Problem: Yes** Toxicology Information Availability: No Injury-related Location Type_Hotel/motel: No	0.419	70%	2%
5	**Sex: Male** **History of Attempted Suicide: Yes** **Marital Status_Married: Yes** Substance Abuse Problem: No Toxicology Information Availability: No	0.425	69%	1%
6	**Sex: Male** **History of Attempted Suicide: Yes** Marital Status_Married: No Toxicology Information Availability: Yes/Unknown **Occupation_Protective Service: Yes**	0.489	57%	0%
7	**Sex: Male** History of Attempted Suicide: No **Substance Abuse Problem: Yes** **Toxicology Information Availability: Yes/Unknown** **Marital Status_Married: Yes**	0.497	54%	2%

Decision paths, all races



GINI-INDEX RANK	5-FACTOR CIRCUMSTANCE COMBINATION	GINI INDEX (LOWER IS BETTER)	POSITIVE CASES	% OF ALL TRAINING SAMPLES (N=ALL DATA SAMPLES)
1	Sex: Female Toxicology Information Availability: No **History of Attempted Suicide: Yes** **Occupation_Protective Service: Yes** **Had Argument: Yes**	0	100%	0%
2	**Sex: Male** History of Attempted Suicide: No Substance Abuse Problem: No Injury-related Location Type_Detention facility: No Injury-related Location Type_Hotel/motel: No	0.341	78%	53%
3	Sex: Female Toxicology Information Availability: No History of Attempted Suicide: No **Marital Status_Married: Yes** History of Mental Illness Treatment: No	0.414	71%	1%
4	**Sex: Male** History of Attempted Suicide: No **Substance Abuse Problem: Yes** Toxicology Information Availability: No Injury-related Location Type_Hotel/motel: No	0.419	70%	2%
5	**Sex: Male** **History of Attempted Suicide: Yes** **Marital Status_Married: Yes** Substance Abuse Problem: No Toxicology Information Availability: No	0.425	69%	1%
6	**Sex: Male** **History of Attempted Suicide: Yes** Marital Status_Married: No Toxicology Information Availability: Yes/Unknown **Occupation_Protective Service: Yes**	0.489	57%	0%
7	**Sex: Male** History of Attempted Suicide: No **Substance Abuse Problem: Yes** **Toxicology Information Availability: Yes/Unknown** **Marital Status_Married: Yes**	0.497	54%	2%

4. PDP plots



Purpose: the partial dependence plots show the relationship of input versus outcome.