Analysis of Death in the United States

Alex Ray, Bryan Brent, Nathan Lile, Austin Pearman

Project Description

This project is about gaining a better understanding of mortality, particularly with respect to the interrelationships between features such as race, sex, day and month of death, age, education, and manner of death in the United states from 2005 to 2015.

Research Questions

To what extent can we predict manner of death from day-of-death features such as age, education level, race, and marital status using decision trees?

- How is classification performance affected by choice of target class and number of classes?
- To what extent can feature importances from decision trees be used to extract interesting information or validate existing intuitions?

To what extent can we extract interesting information regarding mortality trends and patterns from a cluster analysis of mortality features?

Data Preprocessing

- Naturally binary features such as **gender**, and different types of **race** were re-coded into binary features.
- Other numerical features were extracted from the data, such as **age**.
- Also included categorical feature of **education** and **race**, which were each recoded into numerical value.
- Instances with NaN **manner of death** values were removed from the dataset.

e.g...

binary_male	binary_white	binary_black	binary_asian	binary_hispanic	detail_age	month_of_death	education	manner_of_death_cleaned
0	1	0	0	0	45	1	2	7.0
0	1	0	0	0	79	1	2	7.0
0	1	0	0	0	68	1	3	7.0

Tools

- Pandas
  - Data conglomeration, manipulation
  - Provides helper functions for data cleaning
- Scikit-learn
  - Models for ML algorithms (classification, clustering, etc)
- Matplotlib
  - For simple visualizations
  - Utilized by Pandas
- Jupyter Notebook
  - Used for interactive computing of data science languages, tools, and libraries.

Applied Data Mining Methods

Normal decision trees for:

- Classifying suicide vs non suicide
- Classifying homicide vs non homicide
- Classifying manner of death via One vs Rest

Random forest for:

- Multi-class manner of death prediction via One vs Rest

Clustering for:

- Pattern discovery and analysis

Knowledge Gained

Binary **suicide** and **murder** prediction:

- Reasonable classification task, even with limited attributes.

Multiclass manner of death prediction:

- Difficult classification task, even with random forests.

Decision tree **feature importances**:

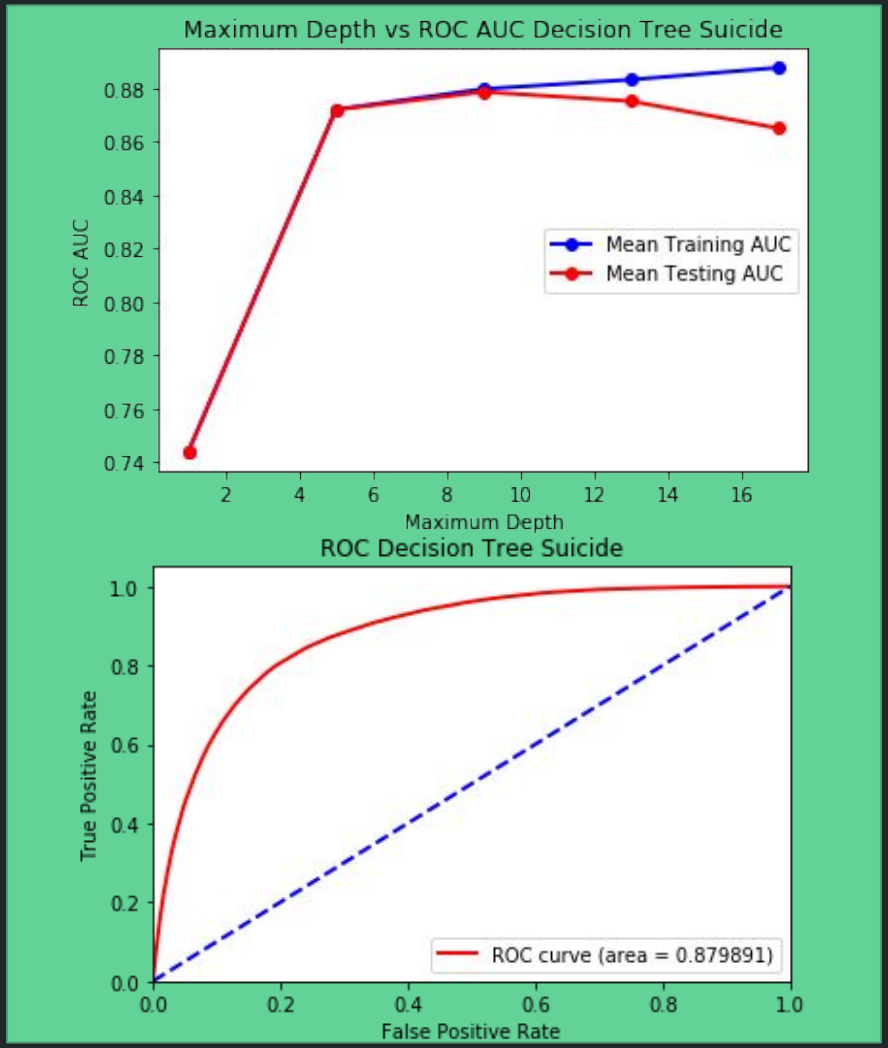
- Useful tool for interesting pattern discovery.
- Provides insight into classification results.

Unsupervised--Clustering

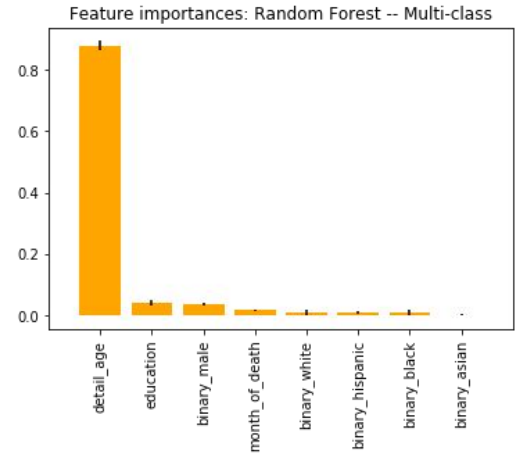
- Preprocessing for binary clustering inherently in opposition to K-Means clustering for optimal results.
- Clustering will require higher **variability** within attributes.

Binary **Suicide** Classifier

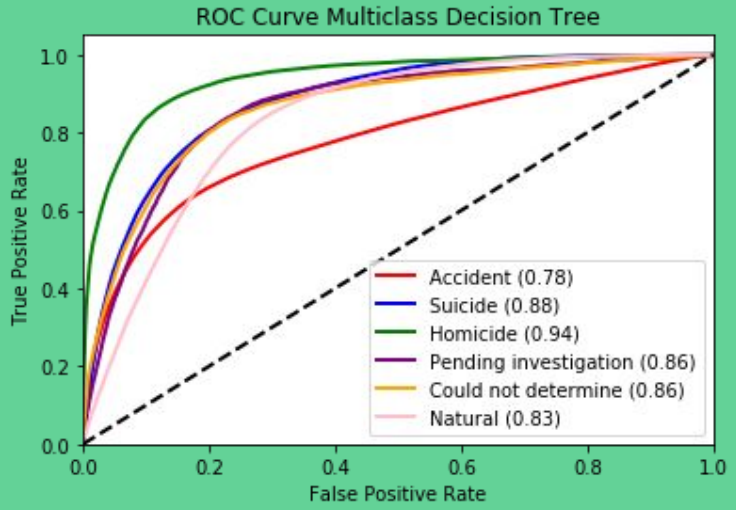
Gridsearch Decision Tree

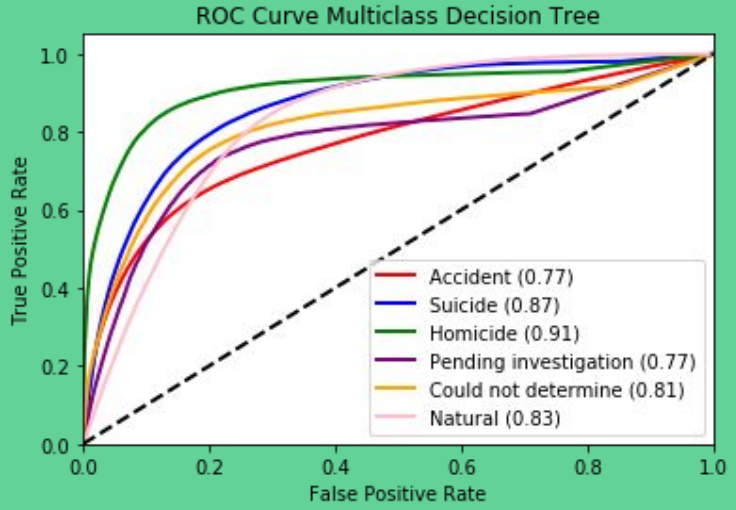


Multinomial **Manner of Death** Classifier



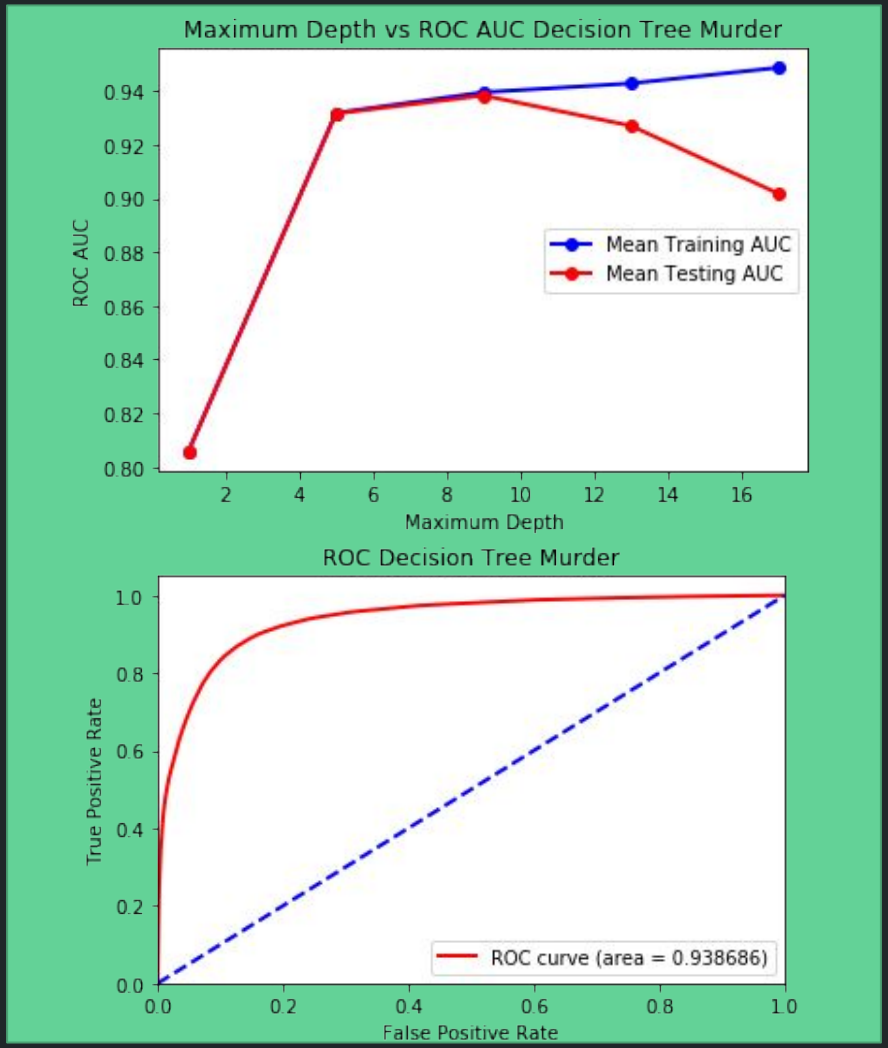
Decision Tree
and
Random Forest



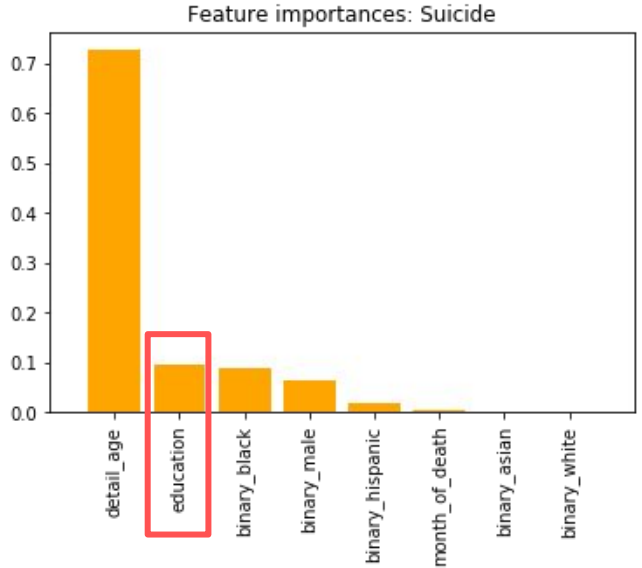


Binary **Homicide** Classifier

Gridsearch Decision Tree

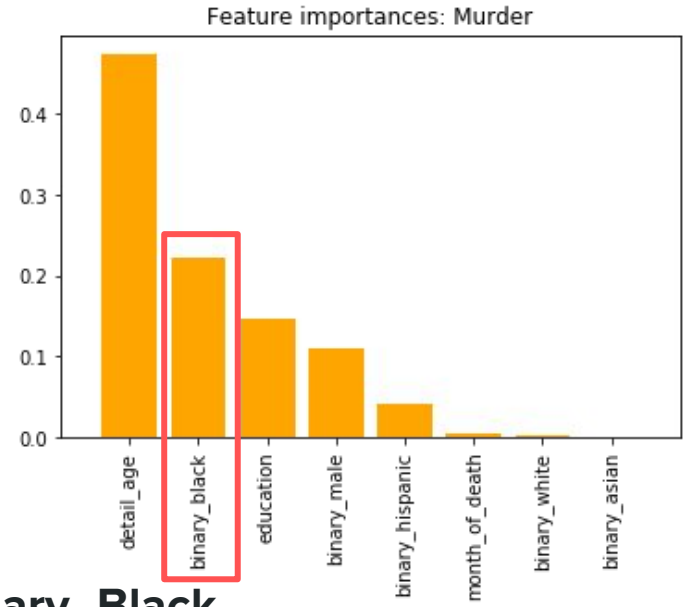


Feature Importances: *Suicide vs Homicide*



Education

- Second most important feature aside from age



Binary_Black

- Second most important feature aside from age

- **Both feature importance observations <u>validate existing intuition</u>**

Application of Knowledge

Suicide Classifier

- Present a proof of concept for help aiding social work

Multi-class Classifier

- Could explore avenues to identify at-risk groups for certain types of death
- Predict possible cause of death, if actual cause is unknown

Homicide Classifier

- Could be used to explore why individuals commit homicide as well as context into violent crime

Questions?

For references, refer to the final paper