Analysis of Death in the United States

Bryan Brent
Nathan Lile
Alex Ray
Austin Pearman

ABSTRACT

This project serves to provide a better understanding of mortality, particularly with respect to the interrelationships between day-of-death and historical features such as race, sex, date, age, education, and circumstance of death in the United States from 2005 to 2015. The US Centers for Disease Control and Prevention releases extensive mortality data every year to allow for a variety of census analyses for life expectancy and death statistics, among other uses.

Our project will serve as a study of classification techniques on mortality data. Prior work with similar datasets has achieved mixed results with a number of algorithms including random forests and naive bayes; we hope to perform a more comprehensive survey of classification techniques as well as better understand the affect of different attributes on classification performance.

Finally, this project will delve into utilizing unsupervised learning techniques such as K-means to potentially gain insight into interesting patterns and trends of mortality in the US.

ACM Reference Format:
Bryan Brent, Nathan Lile, Alex Ray, and Austin Pearman. 2018. Analysis of Death in the United States. In *Proceedings of Data Mining*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.475/123_4

1 PROBLEM STATEMENT/MOTIVATION

Given CDC data on mortality in the United States from 2005 to 2015, which provides day-of-death data such as cause, age, education level, race, and marital status, we will initially attempt to predict the manner of death. Manner of death includes categories such as suicide, homicide, accidental, etc. We will be assessing classification success given only day-of-death information as well as performance with historical features and specific details of the death such as activity (at the time of death), location of injury as well as their family and descendant status.

Depending on the success of these initial manner of death prediction tasks, this project will also attempt to predict cause of death using the same attributes. Due to the specificity of cause of death in the data, this classification task also requires a preprocessing step to "roll up" the cause of death into meaningful supergroups. In theory, this work would allow for a more in-depth analysis of characteristics of individuals "at risk" for different manners and causes of death; given this information, various forms of government programs, social work, and other support systems may be able to adapt their methodologies.

Finally, we will be attempting to cluster mortality features to recognize interesting patterns or trends. This exploratory analysis affords an opportunity to potentially uncover novel interrelationships between attributes in the dataset as well as assess the "completeness" of the existing feature-set. If notable "holes" in attributes are found, it is possible to integrate past temporal data into the existing dataset–market trends and weather data are examples of readily available data ready to be integrated.

Interesting patterns would–like the supervised learning analysis–provide further context regarding what constitutes an individual with notable risk of some manner of death. Even a relatively unsuccessful analysis as determined by quantitative evaluation metrics provides insight into what sorts of features or attributes are necessary to meaningfully predict or cluster mortality events.

2 LITERATURE SURVEY (PREVIOUS WORK)

2.1 Death in the United States Kaggle Page

The Kaggle page for the CDC mortality dataset being used in this paper is a good resource for research questions, discussion on relevant topics, and a repository of existing non-academic work. For example, the main overview page contains interesting ideas on expanding previous work through increasing granularity in age data.

The discussion page contains useful threads on dealing with cause of death recodes (a significant part of our project goals). Discussions also include interesting, specific prior work using this dataset including clustering and predictive analyses.

Finally, the kernel page contains a curated, ranked list of prior work with this dataset. These provide examples of previous work, examples of handling a variety of cleaning and preprocessing steps in Python, and examples of analyses with varying levels of success and interestingness. [2]

2.2 Lessons learned from data mining WHO mortality database (Paoin W)

Previous studies of the CDC's mortality database have provided a template for the most effective data mining methods. Past researchers have concluded that, due to the issues mentioned in our problem statement, classification was generally ineffective for predicting cause of death.

The study cites a lack of correlation between variables and death cause as the root of the issue. On the contrary, it was found that clustering as well as association produced the most interesting patterns. The exact quote from the study can be found below:

Data Mining, Spring 2018, Boulder,Colorado USA
© 2018 Copyright held by the owner/author(s).
ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

"Classification tools produced the poorest results in predicting cause of death. Given the inadequacy of variables in the WHO database, creation of a classification model to predict specific cause of death was impossible. Clustering and association tools yielded interesting results that could be used to identify new areas of interest in mortality data analysis. This can be used in data mining analysis to help solve some quality problems in mortality data." [5]. Our analysis hopes to elaborate on these classification findings by introducing algorithms beyond decision trees and naive bayes. Furthermore, we hope to emulate the prior success of unsupervised learning techniques on this sort of data.

2.3 Graphs of trends of deaths in the United states from 1900 until 2015 NCHS data visualization

The below graph displaying a century long longitudinal study showing trends in life expectancy and age adjusted death rates is indicative of the previous work that has been done in the study of mortality. An example graph is provided below. [3]

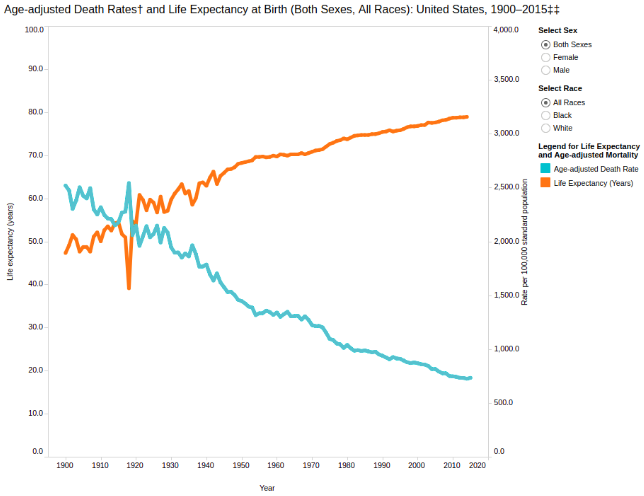

Figure 1: Age-adjusted Death Rates and Life Expectancy at Birth; United States, 1900-2015

2.4 CDC NVSS (National Vital Statistics System) Publication Page

This is the publication page for our database. It contains a list of studies that have been previously done with regards to the CDCs mortality data. We will be using this as a reference for selecting our clustering and association techniques. It has provided us with a roadmap for what does and doesn't work in terms of analyzing CDC mortality data. [4]

2.5 Applying data mining in medical data with focus on mortality related to accident in children

This study used data mining techniques and was seeking conclusions in line with our objectives. The researchers achieved results regarding classifying mortality rate with both decision trees as well as Bayes' theorem. We intend to use their methods as examples of possible techniques in our analysis. [6]

3 PROPOSED WORK

3.1 Data cleaning

(1) Several areas in the json files are of an incorrect syntax, essentially missing quotation marks. This will require a hands on reformatting of the database source.

(2) Converting "recodes" to a more usable format.

  (a) `age` is likely already at a suitable point without factoring in some of the more granular age recodes. This feature does not have missing values.

  (b) `race` has a large number of recodes in addition to the normal "race" feature. This race feature, while not missing any values, is much too granular for our purposes in and of itself. It separates into `white` and `black`, about 10 Asian races, and "other". As will be discussed later, we are also looking to limit the number of features to aid computation time; thus, we limit our analysis to considering `white`, `black`, `Asian`, and other (considered to be `Hispanic`).

  (c) `education` has a flag indicating which of two recode columns is relevant for the given instance, if any. This flag feature has no missing values and can therefore be used to combine these features into a single column.

  (d) `cause_of_death` has many, many different values. Converting this recode into a more usable format likely requires an auxiliary processing step such as clustering.

(3) Ensuring continuity of feature representation and format across the entire database.

  (a) Attributes such as `sex` are converted from binary string values ('M' and 'F') into a single numerical binary feature `binary_male` to allow for a traditional analysis without extra vectorization.

  (b) `manner_of_death` has empty values throughout that must be converted to a numerical 0.

3.2 Data preprocessing

(1) Identifying extraneous attributes that we can drop from the study before we begin analysis. Particularly for initial classification tasks, the analyses begin with a very limited number of attributes to avoid excessive computation time.

(2) Joining features to decrease dataset complexity while maintaining continuity and information integrity.

  (a) Examples of this were introduced in section 3.1–combining many race options as well as multiple recodes into four meaningful categories vastly simplifies the analysis.

(3) "Rolling up" approximately 500 causes of death to create meaningful supergroups that are large enough to yield interesting results.

(a) Note that this step is only needed for a subset of our final analysis tasks described below. It may require more advanced topic extraction methods if we attempt to autonomously perform this step.

3.3 Data integration

We may use other databases to explore the possibility that the database is overlooking features which contribute to the cause of death. Particularly if the classification and/or clustering analyses do not achieve the intended success given our evaluation metrics, integration of external datasets may provide opportunities for more successful analyses.

External datasets may include (but are not limited to) the following:

1. Weather
2. Sporting events
3. Political events
4. Financial markets

4 DATASETS

We will be using the CDC mortality database (URL below). The dataset contains data from deaths in the United States from 2005-2015. The data is compiled yearly by the CDC in the National Vital Statistics system. All deaths are accompanied by a wide array of attributes including age, race, cause of death, descendant status, education level, marital status, month, and day, among many others.

The database has been downloaded by Bryan Brent on his personal machine.

URL: `https://www.kaggle.com/cdc/mortality`

4.1 Temporality

Temporality in the CDC dataset opens the door for integration of other data to add additional features (weather, stock market, political events, etc).

5 TOOLS

5.1 Pandas

Will be utilized for all data analysis and manipulation unless need arises for more powerful tools.

Pandas provides intuitive and easy to use helper functions for data manipulation, analysis, and cleaning. Pandas dataframes afford powerful summarization and correlation tools as well as useful preprocessing functionality such as interpolation.

5.2 Scikit-learn

Scikit-learn provides a wide array of machine learning models and tools. These tools include but are not limited to classification and clustering algorithms, vectorization algorithms, and preprocessing algorithms like principal component analysis.

1. `SGDClassifier` for performing analyses with logistic regression and linear support vector machines. Training is done with stochastic gradient descent (SGD).
2. `LogisticRegression` for performing multiclass logisitic regression.
3. `DecisionTreeClassifier` for performing analyses with decision trees.

5.3 Matplotlib

For simple visualizations such as scatter plots, box and whisker plots, and histograms. Note that Pandas integrates Matplotlib to provide a simple plotting interface for dataframes.

5.4 D3.js

D3.js provides more advanced plotting functionality that allows for the creation of interesting infographics and designs to display results.

6 EVALUATION METHODS

Both the supervised and unsupervised analyses have traditional concrete quantitative evaluation metrics.

6.1 Classification

Evaluation of results can be performed using a number of metrics including accuracy, precision, recall, and F-score (a conglomeration of precision and recall). The specific metric to be used is highly dependent on the goal of the classification, as they each evaluate using a false positive-false negative tradeoff.

For the most part, the set of analyses discussed in this paper will be performed using the accuracy metric, as the concern is mostly with the percent of cases correct out of total, instead of the percent of positives labeled correctly.

6.2 Clustering

Clustering allows a more distance-based quantitative evaluation metric. This can come in the form of using different distance formulas to assess inter-cluster distance (Manhattan distance, Euclidean distance, etc as well as more refined "closeness" measures.

Evaluating clustering analysis results also includes a notable qualitative aspect that returns to the idea of "interesting" patterns. The success of a cluster analysis for this dataset certainly depends on the novelty of the findings as well as the relevance to real-world questions.

7 MILESTONES

1. Combine each annual dataset into single complete dataset. ***March 12***
2. Clean and preprocess data for first manner-of-death classification task. ***March 12***
   (a) Convert `education` related revision attributes into reasonable "rolled up" education summary.
   (b) Convert `age` related revision attributes into reasonable age summary with necessary granularity.
   (c) Convert `race` and `hispanic` related revision attributes into race summary with necessary granularity.
   (d) Split dataset into training and test sets, possibly using a random sample of the total number of instances to allow for more computationally reasonable training times.
      (i) Possibly apply principle component analysis (PCA) or another dimensionality reduction technique to help training time.

(3) Perform classification analysis to evaluate ability to determine manner of death from day-of-death attributes. ***March 19***
   (a) Perform and evaluate binary classification for a suicide vs non-suicide manner of death using logisitic regression, support vector machine, and decision trees, among others.
   (b) Perform and evaluate multiclass classification on all `manner_of_death` options with multinomial logisitic regression.
(4) Clean and preprocess data for unsupervised learning analysis. ***March 23***
   (a) Summarize and process yet-unprocessed attributes such as place of death, injury at work, method of disposition, autopsy, and activity.
(5) Perform unsupervised learning analysis to evaluate ability to extract interesting information from mortality trends and patterns. ***March 30***
   (a) Perform and evaluate the results of clustering techniques such as K-means. This includes an evaluation using different hyperparameters such as number of clusters and distance formula used.
   (b) Assess need to extend analysis to include further techniques or algorithms.
(6) Roll-up cause of death attribute recodes into meaningful summary supersets. ***Loosely March 30***
(7) Given successfully preprocessing cause of death attributes, repeat previous classification analysis for cause of death prediction. ***April 13***
   (a) Include non-day of death attributes such as place of death, activity, and descendant status.

7.1 Milestones Completed

(1) Combine each annual dataset into single complete dataset. We successfully downloaded all years of data, combined, and loaded into Pandas. This process includes taking all comma separated value files, loading into separate Pandas dataframes, and concatenating the dataframes together to make the final dataframe.
   Note that due to the size of the dataset, it's often prudent to subsample each of the CSV files prior to concatenation when considering reasonable computation time. At the time of writing, each CSV pulled fully into the final dataframe and any partitioning or sub-sampling occurs before running individual algorithms.
(2) Clean and preprocess data for first manner-of-death classification task. At the time of writing, we have successfully loaded relevant features for initial classification tasks into a Pandas dataframe. This includes preprocessing most relevant "time-of-death" type features such as education, age, and race of the deceased.
   (a) `education` recodes were combined into a single feature. As discussed previously, education had an `education _reporting_flag` feature to indicate which of two recodes (if any) have relevant values to use. The two recodes each worked on a different level of granularity–one included individual years of each level of education, while the other worked at a less granular level.
   We compromised and combined the two recodes into a single feature indicating a person has one of the following levels of education:
      (i) Unreported/unknown
      (ii) Some amount of education through the end of primary school
      (iii) Some amount of secondary school
      (iv) Some amount of post-secondary education up to a Bachelor's degree
      (v) Some amount of post-Bachelor's education
   (b) `race` recodes were combined into a single feature, with the general granularity of a typical job application. Many of these recodes go very in-depth on the specific nationality or subgroup of the deceased; however, our analysis is only looking for general demographic information. Thus, the final `race` attribute consists of these four different options represented with one-hot encoding:
      (i) White
      (ii) Black
      (iii) Asian
      (iv) Hispanic
   (c) Other small preprocessing steps, outlined in section 3.1, were performed to ensure continuity of data types.
   (d) Once the final dataset has been created (given the preprocessing steps mentioned previously) it can be separated into training and test datasets. Furthermore, to aid computation time–particularly given the size of the dataset–Pandas' `sample` method can be utilized to randomly sample some fraction of the data.
(3) Perform classification analysis to evaluate ability to determine manner of death from day-of-death attributes. We've begun attaining preliminary results with binary classification using logistic regression, linear SVM, and decision tree. Through the "feature importance" result of the decision tree module, we understand that the `age` feature is by far the most important for binary suicide vs non-suicide classification. We have also gotten results with multiclass classification using multinomial linear regression.
   These results are elaborated on in section 8.

7.2 Milestones To-do

The milestones laid out in the original document that have not yet begun include the following:

(1) Clean and preprocess data for unsupervised learning analysis.
   Note that this step is mostly complete given the existing work on preprocessing for classification tasks.
(2) Perform unsupervised learning analysis to evaluate ability to extract interesting information from mortality trends and patterns.
(3) Roll-up cause of death attribute recodes into meaningful summary supersets.
   This step may include techniques from the unsupervised learning milestone. Specifically, clustering techniques may allow for a naive clustering of cause of death options given

their short names and descriptions. This may involve recoding causes into broader categories prior to clustering. The current database offers recodes of 39 to 348 codes, but these tend to be either have far too many causes for us to find interesting patterns or are too obviously geared towards detecting a specific set of causes. Beyond recoding, we are open to other manual options that will allow for more meaningful clusters.

(4) Given successfully preprocessing cause of death attributes, repeat previous classification analysis for cause of death prediction. These techniques will look extremely similar to the multiclass classification done with `manner_of_death`.

(5) Potentially refactor races to consider for more finite ethic background. This means making unrolling the general Hispanic race into whether or not they're Spaniard, Mexican, or from the Southern American continents.

8 RESULTS SO FAR

8.1 Decision Tree Binary Classification

Our first classification task is to attempt to classify whether or not a death was a suicide or not. We begin our analysis with decision trees using Scikit-learn's `DecisionTreeClassifier`. Decision trees are computationally efficient, relatively performant, and extremely interpretable which make them great choices for initial stabs at classifying data. Furthermore, performance on decision trees affords the ability to make an informed decision regarding whether or not to delve into more complex binary classification algorithms such as support vector machines.

Note that for this analysis, only instances with reported manner of death were included. Also note that this is a relatively unbalanced binary classification task, as the positive class (suicide) is only 17.3% of the data with reported manner of death.

8.1.1 Feature Importances. To get a better "feel" for the data as well as better understand the preprocessed features, `DecisionTreeClassifier` has a `feature_importances_` attribute. From the official Scikit-learn documentation, "The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as the Gini importance." A larger number corresponds to a more important feature when making the final classification. This table was generated with a decision tree with a maximum depth of 16 on a suicide vs non-suicide binary classification problem. `detail_age`–the numeric age attribute–is obviously the most important feature when classifying suicide vs. non-suicide deaths. Furthermore, `education` is the next most important feature, followed closely by binary "black" and "male" attributes. All other features are significantly less important.

Feature Name	Importance
detail_age	0.67948668
education	0.11023995
binary_black	0.09935316
binary_male	0.07111632
binary_hispanic	0.02666306
month_of_death	0.00912728
binary_white	0.00279102
binary_asian	0.00122251

Table 1: Sorted feature names and their corresponding importances for suicide vs non-suicide classification

8.1.2 Binary Classification Suicide vs Non-Suicide. Results from classifying suicide vs non-suicide deaths using the features listed above are as follows. As adjusting the maximum depth (known as "pruning" the decision tree) is the main way to adjust overfitting with decision trees, the plot shows training and test accuracy across multiple settings for maximum depth. All other hyperparameters were set as the Scikit-learn default values.

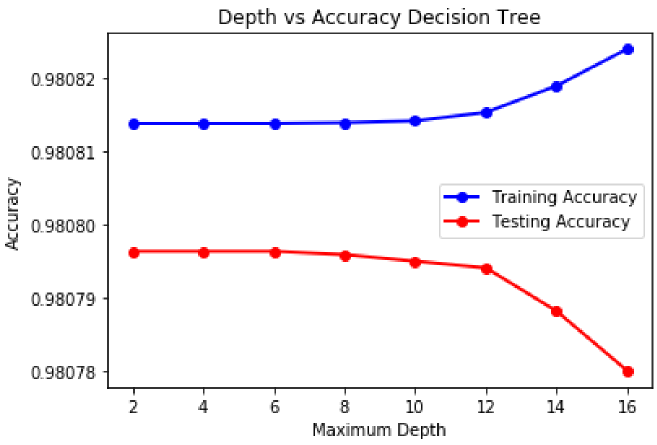

Figure 2: Accuracy vs maximum depth of decision tree when classifying deaths as suicide vs non-suicide.

From these results we can see that deciding whether or not a given death was a suicide is not a particularly difficult classification task. Mid-90s testing accuracy is extremely accurate.

While it normally makes little-to-no sense to have test accuracy higher than training accuracy in machine learning, both test and training accuracies are high enough and similar enough in absolute value that this is not particularly concerning. Furthermore, it makes sense that as we "unprune" the tree, testing accuracy gets worse and training accuracy gets better.

However, due to the skewed data, we can look at the performance of the decision tree with a receiver operating characteristic curve (ROC curve), which is affected less by class distribution.

This ROC curve also shows very good classification performance on the test data. Because increasing or decreasing the proportion of positives in the set would proportionally increase both the false positive and the true positive rate equally, this metric is more class distribution-agnostic than accuracy alone.

8.2 Decision Tree Multiclass Classification

Having seen good performance classifying suicide vs non-suicide using decision trees, we can now extend our analysis to attempt to classify all manners of death in our dataset. Note that the recode

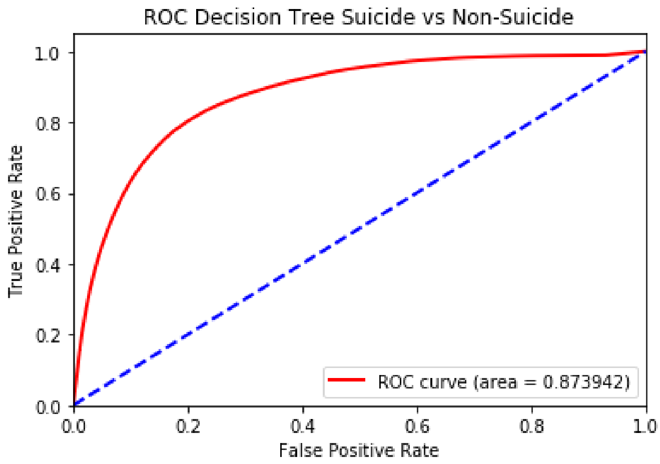

Figure 3: ROC curve for binary decision tree with maximum depth of 16.

defines more than are listed, but the additional ones do not show up in practice.

Manner of Death	Number of Instances
Natural	19914162
Accident	1350268
Suicide	423361
Homicide	200528
Could not determine	119756
Pending investigation	55808

Table 2: Manner of death sorted by number of instances

Now, instead of making all non-suicide values 0 and suicide values 1 we can leave the original encoding and run the same analysis as before using a `DecisionTreeClassifier`.

Feature importances for the multiclass decision tree look very similar to binary, though they're even more skewed towards age.

Just as we did for binary classification, we can explore the performance of this classifier with ROC curves. Scikit-learn's ROC functions do not accept "real" multiclass problems; however, this problem can be avoided by using a One-vs-Rest classification scheme.

Note that the ROC curves for multiclass–because One-vs-All was necessary–is not showing the same information as the accuracy for the actual multiclass decision tree. In general, the further away from the main diagonal the better performance of the classifier. While interpreting results of multiclass classification is non-trivial, it makes sense that the two largest classes in terms of number of instances had the smallest area under the curve (AUC). Being the broadest categories, intuitively it would be hardest to distinguish instances of those classes from all others.

8.3 Homicide and Decision Trees

An interesting note from the One-vs-All ROC curve is the relative performance of the "Homicide" class against all others. Homicide doesn't have the lowest number of instances of all classes, so there must something else causing such good results. We can repeat our previous binary classification analysis steps with Homicide instead of suicide.

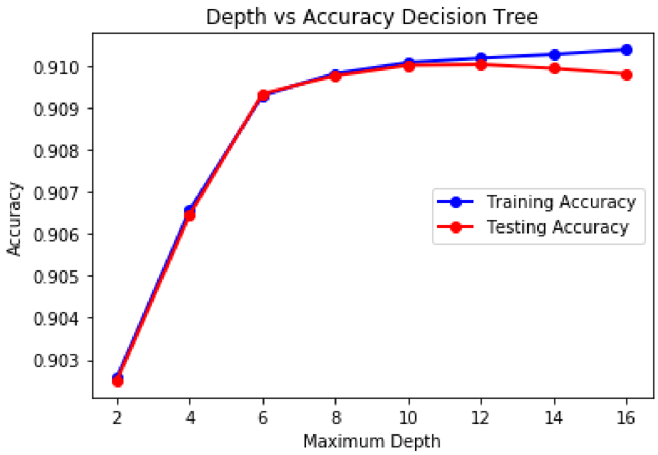

Figure 4: Depth vs accuracy for multiclass decision tree with maximum depth of 16.

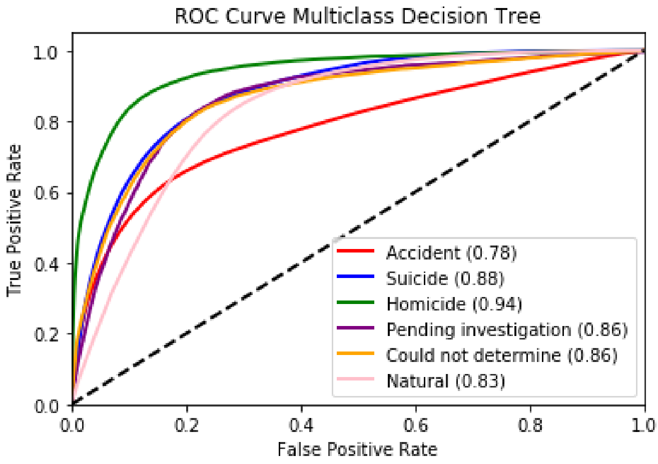

Figure 5: ROC curves for multiclass decision tree with maximum depth of 16.[1]

Clearly, the same classifier performs significantly better when classifying homicides vs non-homicides when compared to classifying suicides vs non-suicides. Looking back to feature importances, we can see that a plausible explanation for this performance increase may lie in the notable increase in importance in the `binary_black` feature compared to our previous analysis.

REFERENCES

[1] [n. d.]. ROC Curve Example Scikit-learn. ([n. d.]). Retrieved April 8, 2018 from http://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html#sphx-glr-auto-examples-model-selection-plot-roc-py
[2] kaggleMortality 2018. Death in the United States. (2018). Retrieved March 5, 2018 from https://www.kaggle.com/cdc/mortality/kernels

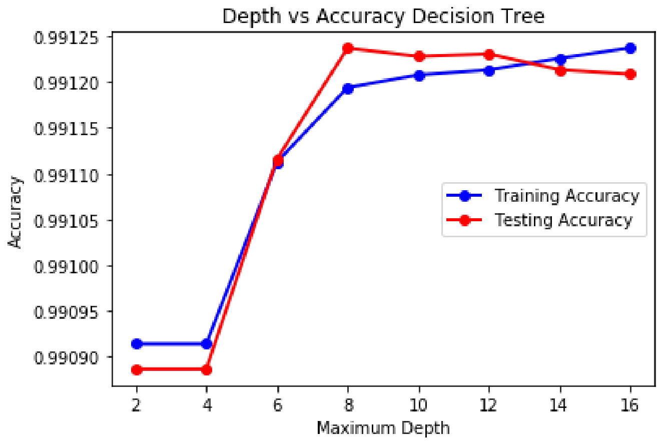

Figure 6: Depth vs accuracy for homicide vs non-homicide binary decision tree with maximum depth of 16.

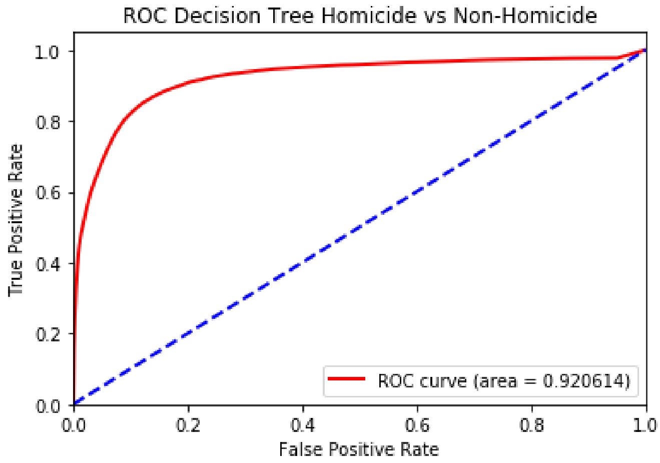

Figure 7: ROC curve for homicide vs non-homicide decision tree classification with maximum depth of 16.

Feature Name	Importance
detail_age	0.46771064
binary_black	0.2143926
education	0.14302406
binary_male	0.10711167
binary_hispanic	0.04252411
month_of_death	0.0214145
binary_white	0.00310923
binary_asian	0.00071318

Table 3: Sorted feature names and their corresponding importances for homicide vs non-homicide classification

[3] nchsVisTool 2017. Mortality Trends in the United States, 1900-2015. (2017). Retrieved March 5, 2018 from https://www.cdc.gov/nchs/data-visualization/mortality-trends/

[4] nvss 2015. NVSS - Mortality Data. (2015). Retrieved March 2, 2018 from https://www.cdc.gov/nchs/nvss/deaths.htm

[5] W. Paoin. 2011. Lessons Learned from Data Mining of WHO Mortality Database. *Methods of Information in Medicine* 50, 4 (jun 2011), 380–385. https://doi.org/10.3414/me10-02-0019

[6] M. H. Saraee, Z. Ehghaghi, H. Meamarzadeh, and B. Zibanezhad. 2008. Applying data mining in medical data with focus on mortality related to accident in children. In *2008 IEEE International Multitopic Conference*. 160–164. https://doi.org/10.1109/INMIC.2008.4777728