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 binary and multiclass classification techniques on mortality data, primarily using decision trees. Prior work with similar datasets has achieved mixed results with a number of algorithms including random forests and naive bayes; we perform a survey of classification techniques for manner of death using a variety of day-of-death attributes of the deceased. In addition to evaluating the performance of decision trees and random forests with binary and multiclass classification, we to utilize the interpretability of the decision tree classifier to gain insight into attribute importances in mortality-related classification tasks.

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

This work utilizes common computational methods and attributes to gain context into the manner of death. Using the concept of feature importance, we seek to connect classification results to existing knowledge on the demographics of suicide and murder and by doing so add to the general body of literature on the subject. Furthermore, results and relics of computational techniques can then be used in a number of practical applications.

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, 11 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Using the CDC data on mortality in the United States from 2005 to 2015, we will initially attempt to predict whether or not an instance was a suicide or not a suicide via the manner of death attribute. Manner of death includes categories such as suicide, homicide, accidental, etc. This analysis will use day-of-death features such as age, race, and education status of the deceased. After initial binary classification, we perform both multiclass classification as well as binary classification of homicide vs non-homicide instances.

In the final part of the analysis we attempt 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.

Using the supervised and unsupervised learning techniques described above, we seek to answer a number of research questions to better understand the utility of common computational methods on this widely-spanning dataset. By evaluating a number of different binary and multiclass classification tasks on the dataset, we can then investigate how classification performance is affected by the choice of target class as well as the number of target classes; in this case, the classes are chosen from the set of manner of death options in the dataset. Furthermore, we seek to better understand the extent to which relics from supervised learning techniques can be used as investigative tools with practical applications in the real world.

In the unsupervised learning section, we seek to understand the potential role of clustering analyses in uncovering "interesting" information with regards to mortality.

The knowledge gained through these common computational techniques may have a wide array of potential applications, both theoretical and practical. Firstly, classification analysis using such a limited feature set can potentially serve as a proof of concept for the utilization of computational techniques in social work. While the CDC dataset is likely too general for any meaningful out-of-the-box solution to be created, success at prediction of specific manners of death can potentially inspire future applied work by domain experts. Furthermore, we seek to contribute to an existing relevant body of literature on mortality and at-risk groups. Finally, successful classification results could provide insight into potential data preprocessing tools for filling in missing data.

2 RELATED 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.

Data Mining, Spring 2018, Boulder,Colorado USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

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. [6]

2.2 Lessons learned from data mining WHO mortality database

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:

"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." [8]. 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. [5]

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. [3]

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. [9]

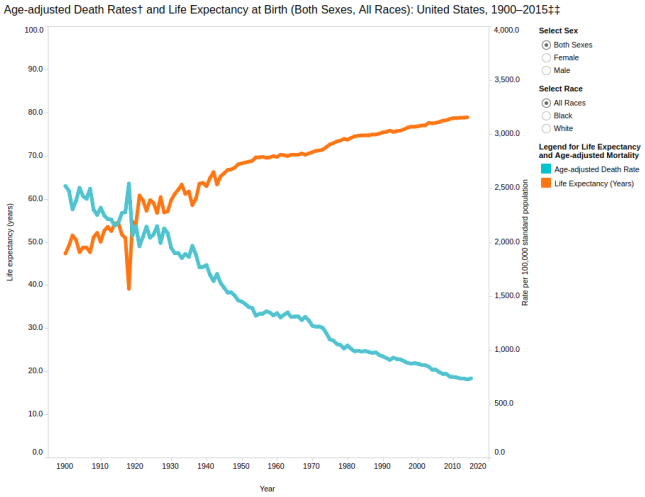

Figure 1: Age-adjusted Death Rates and Life Expectancy at Birth; United States, 1900-2015

3 DATASET

We will be using the CDC mortality dataset [6]. The dataset contains data extensive data about deaths in the United States from 2005-2015, on the order of 27 million instances of the deceased. 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 both day-of-death attributes as well as historical and background features about the individual. In this analysis, we limit our scope to the following attributes:

(1) `age`, given as an integer year
(2) `race`, given as a complex set of recodes. These will be discussed further in the techniques section.
(3) `education level`, given as one of two recodes. To be elaborated on further in techniques.
(4) `month of death`, given as an integer between 1 and 12.
(5) `manner of death`, given as one of:
  (a) Accident
  (b) Suicide
  (c) Homicide
  (d) Pending investigation
  (e) Could not determine
  (f) Self-Inflicted
  (g) Natural

Limiting the analysis to this subset of features is first and foremost necessary given the size of the dataset. We do not have the time or computational power to run many of the models we describe with many more features included. Note that due to limited time, we consider dimensionality reduction techniques such as Principle Component Analysis (PCA) to be out of scope of this project.

4 TECHNIQUES

4.1 Tools

4.1.1 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.

4.1.2 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) `GridSearchCV` We leveraged GridsearchCV in all of our binary classification tasks, a library that methodically evaluates every combination of algorithm parameters specified by the user and also performs k-fold cross validation.
(2) ROC tools in Scikit-learn were utilized to calculate ROC AUC values and product ROC curves [1].
(3) `DecisionTreeClassifier` for performing analyses with decision trees.
(4) `RandomForestClassifier` for preforming analyses with random forests.

4.1.3 Matplotlib. For simple visualizations such as line charts. Note that Pandas integrates Matplotlib to provide a simple plotting interface for dataframes.

4.2 Data cleaning

(1) Several areas in the json files are of an incorrect syntax, essentially missing quotation marks.
(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 `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) `manner_of_death` has many missing values. Due to the size of the dataset, we can remove all instances without reported manner of death. Furthermore, we need not consider some listed manner of death values as they do not appear in the dataset. Finally, depending on the classification task, we binarize the manner of death values.
(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.

4.3 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. Examples of this were introduced in section 4.2–combining many race options as well as multiple recodes into four meaningful categories vastly simplifies the analysis.

4.4 Suicide Binary Classification with Decision Trees

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. As discussed previously, a recurring consideration throughout this analysis is the computational efficiency of our models due to the number of instances in the dataset. 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 a reported manner of death were included. Also note that even with this caveat, 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.

To get a better "feel" for the data as well as better understand the preprocessed features, `Decision TreeClassifier` 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 10 on a suicide vs non-suicide binary classification problem.

Feature Name	Importance
`detail_age`	0.727560
`education`	0.096528
`binary_black`	0.086486
`binary_male`	0.064593
`binary_hispanic`	0.019309
`month_of_death`	0.003177
`binary_asian`	0.001343
`binary_white`	0.001004

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

This relationship can be represented as a bar chart as follows.

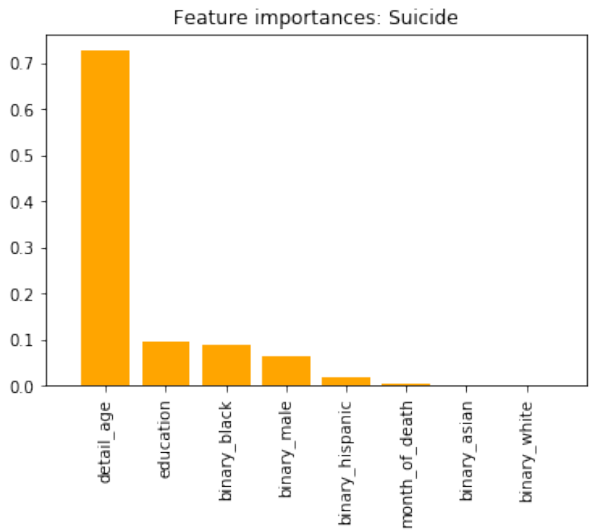

Figure 2: Sorted feature names and their corresponding importances for suicide vs non-suicide classification.

`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.

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, we chose this as the main hyperparameter to adjust in `GridSearchCV` to find the most performant model.

Once the grid search of parameter combinations is complete, we can then fit the a model with the highest scoring combination of hyperparameters based on any one of a variety of scoring metrics, for our purposes we chose `roc_auc` (Reciever Operating Characteristics Area Under the Curve).

Accuracy as a scoring metric does not lend itself to our classification task due to the unbalanced class distribution in the dataset (as discussed previously, instances with suicide as the manner of death are a significant minority); this skewed distribution leads to artificially inflated accuracy scores. For example, if 98% of all instances are the positive label in a binary classification task, guessing the positive label every time will result in 98% accuracy.

`roc_auc` is a better model scoring metric as it is known as a class distribution independent metric. This is because the metric is simply the area under the true-positive-rate vs false-positive-rate curve; these values remain stable regardless of skewed distributions. Furthermore, the Scikit-learn documentation states that "this implementation is restricted to the binary classification task or multilabel classification task in label indicator format." which is exactly what we were looking to accomplish.

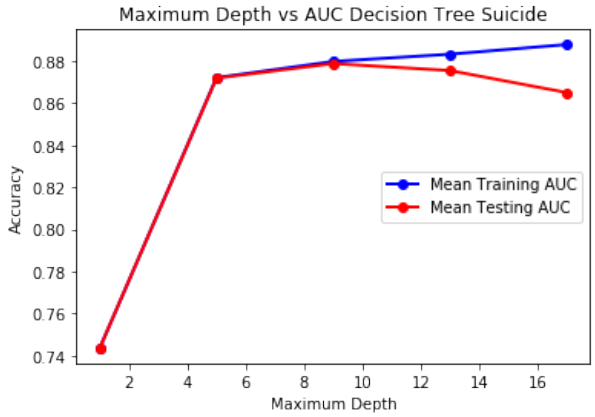

Figure 3: Suicide ROC AUC vs Depth using `GridSearchCV`.

This figure presents the mean training `roc_auc` and the mean testing `roc_auc` as a function of the prescribed maximum depth of the decision tree. These averages were calculated using 3-fold cross validation with `GridSearchCV`, on 80% of the total instances.

Furthermore, it makes sense that as we "unprune" the tree, testing accuracy gets worse and training accuracy gets better–the model is able to fit better to the training data, which makes it less generalizable to new data.

We can now present the test accuracy with the `GridSearchCV` "best estimator" on the held-out 20% of the dataset.

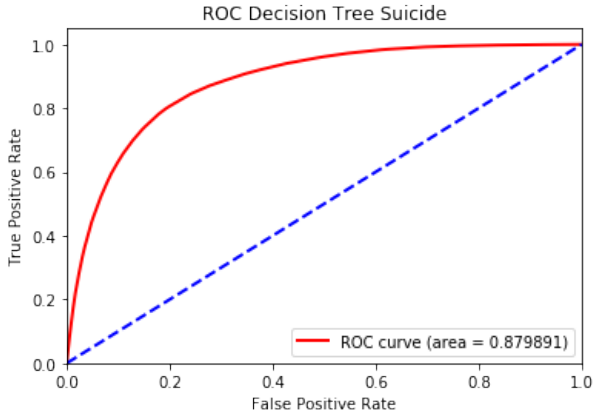

Figure 4: ROC curve for suicide on test dataset, using `GridSearchCV` best estimator.

This ROC curve also shows very good classification performance on the test data. As mentioned previously, 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.

4.5 Multiclass Classification with Decision Trees

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 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. See appendix A for the relevant figure.

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 where n binary classifiers are trained to classify each class against all others. Note that in the interest of computational runtime, this One-vs-Rest analysis does not utilize `GridSearchCV` for hyperparameter tuning or cross validation; all classifiers are trained with a maximum depth of 10, due to the results from the suicide vs non suicide classification task in section 5.1.

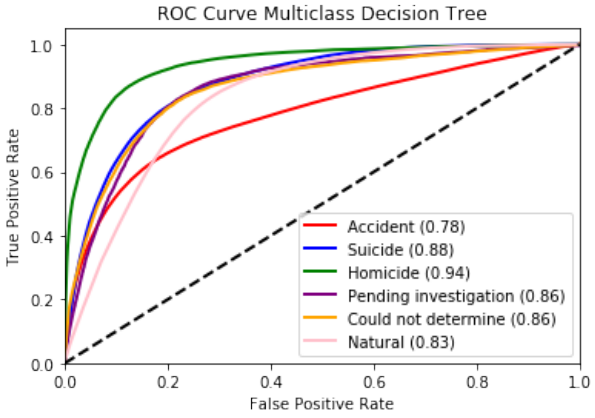

Figure 5: ROC curve for One vs Rest decision tree with maximum depth of 10.

Note that the ROC curves for multiclass – because One-vs-Rest was necessary – is not showing the same information as results for a single multiclass decision tree. This is essentially 6 separate binary classification tasks, and the AUC score for suicide reflects that as it is the same as the AUC score for binary suicide vs non suicide with a maximum depth of 10.

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.

Another avenue of approach for this multiclassifiaction task is to utilize random forest conjoined with the one-vs-rest approach in an attempt to see the performance differences observed when using a ensemble classification method, specifically Scikit-Learn's `RandomForestClassifier` when compared to the One-vs-Rest approach with a single decision tree.

As we can see from Figure 7, the multiclass random forest estimator performed similarly to the One vs Rest approach in Appendix A, with some oddities with the 'pending investigation' cause of death. Interestingly enough, it appears that the random forest preformed ever so slightly worse against the same test-set when compared to the decision tree classifier, however these results are only marginal. These preliminary results seem to confirm previously cited challenges with multiclass classification models for mortality specific datasets.

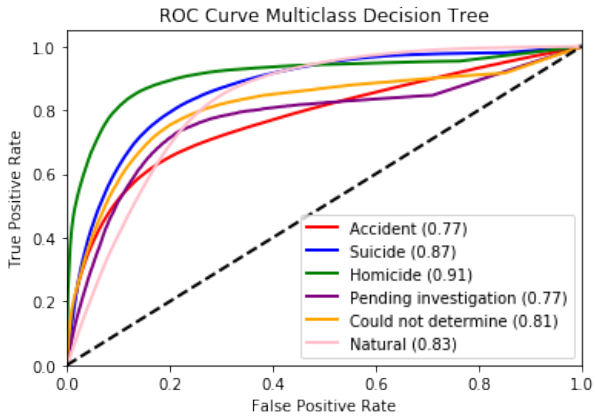

Figure 6: ROC curve for multiclass random forest with maximum depth of 10.

An additional benefit of utilizing a random forest instead of many binary classifiers for multiclass classification is the ability to take a look at feature importances from a statistical viewpoint, namely looking at variance. Figure 8 shows the feature importances along with error bars for the random forest multiclass classifier.

Clearly, the two multiclass models both effectively assert that age is the most important feature when considering splitting criteria for manner of death. Furthermore, taking the feature importances from all binary classification tasks reinforces the relative insignificance of the other attributes, on the whole.

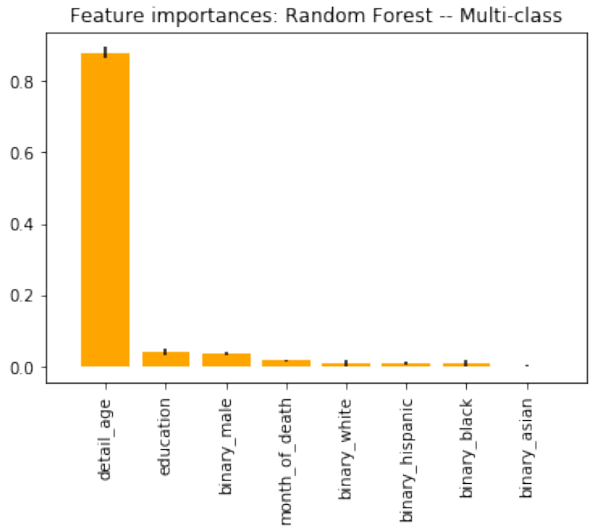

Figure 7: Feature Importances for multiclass random forest. Error bars represent Gaussian distribution over the various individual trees inside the forest.

4.6 Homicide and Decision Trees

An interesting note from both of the multiclass ROC curves is the relative performance of the "Homicide" class against all others. Homicide doesn't have the lowest or highest number of instances of all classes and `roc_auc` works well for skewed data, so there must something else causing such good results. We can repeat our previous binary classification analysis steps with Homicide instead of suicide.

First, we again run `GridSearchCV` with a decision tree classifier using 3-fold cross validation. As before, the grid search is looking for optimal hyperparameter combinations, and in this case we present results while only varying maximum depth of the decision tree.

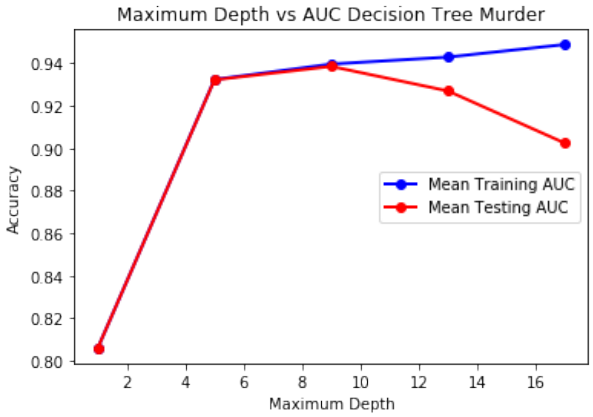

Figure 8: Homicide ROC AUC vs Depth using `GridSearchCV`.

Clearly, the same classifier performs significantly better when classifying homicides vs non-homicides when compared to classifying suicides vs non-suicides, based on the mean training and testing accuracy curves produced by `GridSearchCV`. This result is corroborated when producing an ROC curve using the best estimator from figure 10.

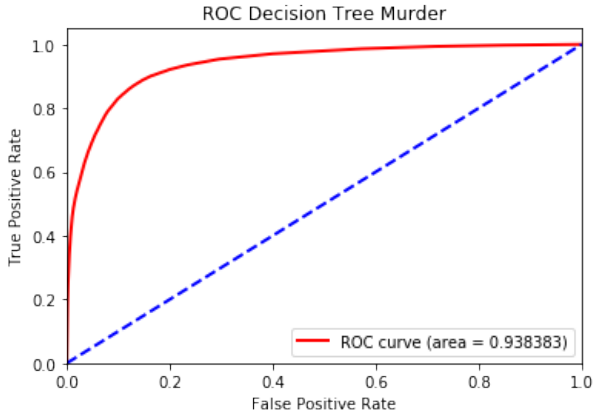

Figure 9: ROC curve for decision tree with maximum depth of 16.

The ROC area under the curve value for murder vs non-murder on the held-out test data is 0.938, whereas the area under the curve value for the same data looking at suicide vs non-suicide is 0.880. This is a relatively large difference in classification performance, particularly as both experiments end up choosing the same maximum depth value.

The question then becomes why is the binary murder classification task so much more successful than others? 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.

Feature Name	Importance
`detail_age`	0.473739
`binary_black`	0.221837
`education`	0.146484
`binary_male`	0.110524
`binary_hispanic`	0.040780
`month_of_death`	0.003605
`binary_white`	0.002740
`binary_asian`	0.000292

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

When presented as a bar chart, the differences between feature importances with suicide and feature importances with homicide become more clear.

As will be discussed more in later sections, the importance of `binary_black` with the murder classification task is more than double that of the suicide task. This is interesting as it confirms existing

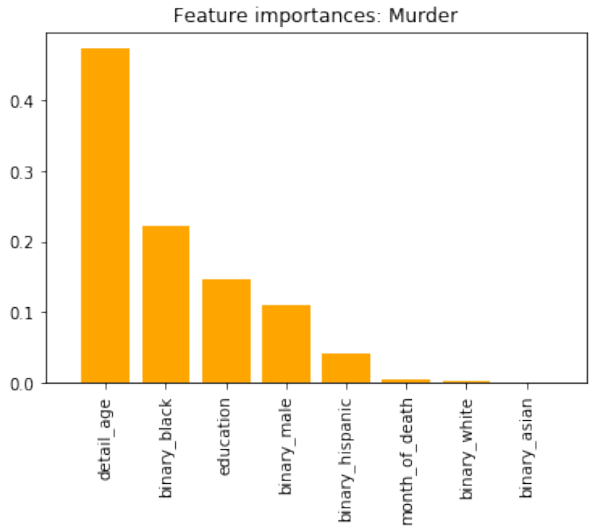

Figure 10: Feature importances for murder vs non-murder classification using decision tree with maximum depth of 10.

knowledge on the demographics of violent crime and also speaks to the utility of feature importance as a metric for information extraction.

We can look more into the relative feature importances of murder against suicide and see that age plays a lesser role in classification, which also matches existing intuition.

4.7 K-Means Cluster Analysis

As we began to delve into the realm of unsupervised learning, we took note of the previous literature on this issue. As opposed to classification, which had been previously proven to be ineffective, clustering was said to have yielded interesting results.

The recodes that have been done previous to our preprocessing as well as those done by our team have made replicating those results difficult. Recodes, while fantastic for classication, made for unvaried data that led to hard to interpret clusters. This was especially true when trying to cluster on binary variables, which often led to clusters with centroids that lied exactly at the 0 or 1 mark for the attribute in question.

It was this challenge that led to our rationale on which attributes would be the most interesting to cluster on. The first obvious choice was age due to the fact that we had mortality data from infancy to beyond the 100 year mark. The next was month of death. Month, while significantly more varied than our binary recodes, still only contained twelve codes as opposed to the 100+ in the age attribute.

After finding interesting attributes to cluster on, we had to decide how many clusters we would need in order to produce interesting results. This was relatively easy using Scikit-learn's score tool, which assigns a score based on the variance as indicated by the number of clusters. Note that the score on the y-axis is scaled by 1e18, indicating a remarkably high distance between cluster members and the centroid and centroid:

Figure 11: Score as indicated by variance for differing numbers of clusters

The performance of this score calculation limited us to testing up to 6 clusters. That being said, it is clear that the variance levels off at three clusters, with little increase in score with additional clusters, after a sharp leveling at 2 centroids. For that reason, we began clustering between Age and Month of Death with three centroids:

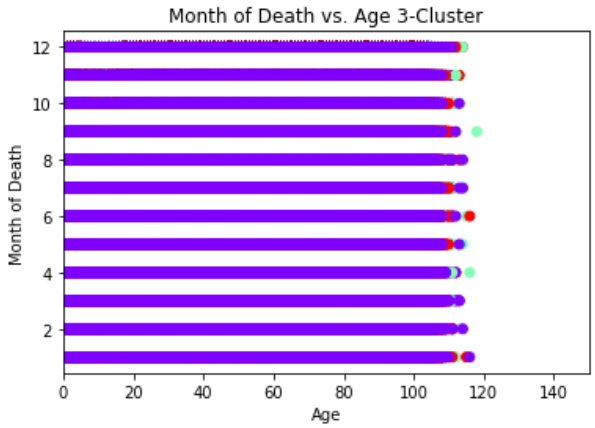

Figure 12: K-means cluster comparing age to Month of Death

Our goal was to identify distinct groups in the dataset. We would then be able to use these groups to analyze patterns and perhaps draw conclusions about new objects based on which cluster they belong to.

The above scatter plot demonstrates some of the issues we had while finding suitable attributes to cluster on. Ideally, we would see three distinct clusters, with high inter-class dissimilarity and high intra-class similarity between clusters. These clusters appear to have little significance with low inter-class dissimilarity and low intra-class similarity. This is unsurprising given our poor scores in Figure 12.

The lack of distinct clusters in our dataset made our scatter plots difficult to interpret and therefore unsuitable for pattern analysis. What we do learn from this graph is that age and month of death

have no noticeable correlation, implying that age cannot be used as an accurate predictor of time of death to any more of a specific degree than the year.

Given our difficulties with clustering attributes, potentially due to the lack of variety in the recodes, we decided we may need to increase the dimensionality of our analysis in order to draw interesting conclusions. To this end, we added the Manner of Death attribute to our cluster analysis:

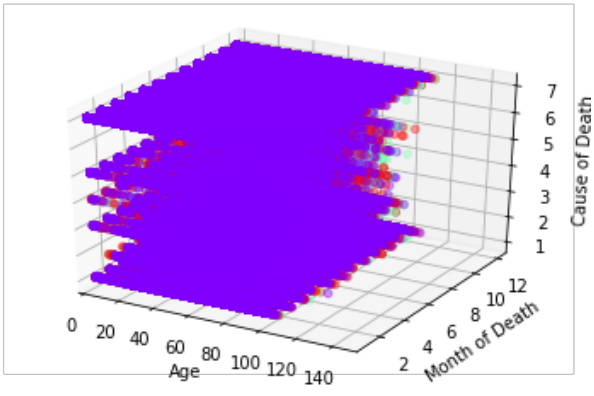

Figure 13: K-means cluster comparing age, month of death, and cause of death

As before, we see little to no correlation between any of the three attributes and little distinction between clusters. Although, scatter plot analysis brought to light an interesting lack in correlation between our attributes, making R-squared analysis a worthwhile research topic.

4.8 R-squared analysis

Due to the lack of correlation seen in scatter plots, it led to the assumption that the attributes in our dataset are poorly correlated. To verify this assumption, we generated a table showing the R-squared values between what we believed to be our most interesting attributes:

	BinaryWhite	BinaryMale	Month	Manner	Age
BinaryWhite	1.000	0.000081	0.000017	0.000303	0.015901
BinaryMale	0.000081	1.000	0.000018	0.010546	0.021243
Month	0.000017	0.000018	1.000	0.000236	0.000052
Manner	0.000303	0.010546	0.000236	1.000	0.085205
Age	0.015901	0.021243	0.000052	0.085205	1.000

Table 4: R-squared values between select attributes

From the above, it is obvious that the majority of our attributes are only loosely correlated. Even in the case of our strongest correlation, that of manner of death and age, only 8.5% of the variance can be explained by the model.

The clustered scatter plot between the two most highly correlated attributes can be seen below:

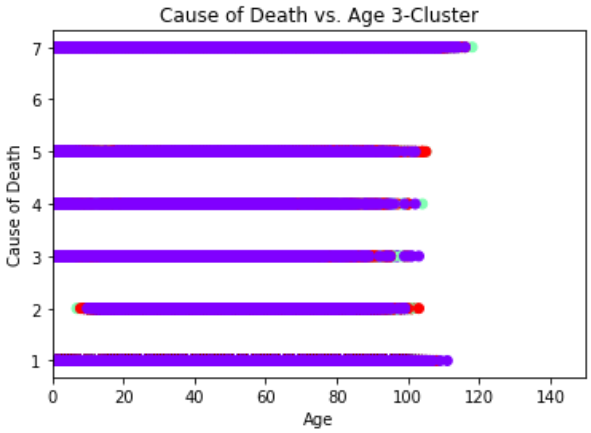

Figure 14: Clustered scatter plot between cause of death and age

5 RESULTS

5.1 ROC AUC and Feature Importance Discussion

In general, our binary classifiers managed to score fairly well and produce good ROC curves, the two being suicide, scoring (0.88) and homicide, with scores ranging from (0.93) to (0.94). As is evidenced by homicide, it appears that classifier performance drastically improves as additional important features get in the mix.

Upon examining feature importance with suicide as the cause of death (in our findings, the number 3 reported cause of death after Accident and Natural), we observed that the age of the individual, (appx. 0.7275) scored more than seven times as high as the next highest feature, education (appx. 0.0965).

Similarly, when we drilled down and isolated homicide as the cause of death, just after suicide in terms of total number of reported deaths, we again observed that age scored highest (appx. 0.4677) and binary_black second (appx. 0.2143) in terms of feature importance.

We were surprised to learn that according to the Caspian Journal of Internal Medicine, that it is generally accepted in the field of medicine in order to be considered medically plausible, a test must achieve an ROC score of (0.95) or higher. This helped to put the precision of our results into better perspective, informing us that our results appear to be of high quality. [2]

5.1.1 Verifying challenges with classification. One of the most notable pieces of information gathered from our results are general validations of previous findings. Many of our results corroborated previously stated challenges in using classification make meaningful predictions from attributes of the deceased.

The other side of success of the suicide and homicide binary classifications can be seen with classes such as accident in the multiclass classification. This class had poor ROC AUC (0.77) for with both decision trees and random forests; intuitively this makes sense, as accidents would seem like they would be more evenly distributed across demographics than manners of death such as murder. Age would be the only feature that would intuitively have a significant correlation to accident as a manner of death.

5.1.2 Decision Tree vs Random Forest Multiclass. In an effort to cross validate the Binary classification findings of our Decision tree, our next models incorporated a multiclass, One vs Rest classification approach to the data using decision trees and random forests. Note that we found that our multiclass decision tree classifier found `detail_age` the feature of highest scoring importance in general across all causes of death (0.88); it is of interest to note that this score exceeds any feature importance score from our binary classifier. Another observation of interest from the multiclass decision tree's ROC score is that homicide starts out as far and away the best scoring feature, peaks the soonest and has a very shallow plateau-decline, remaining better scoring the other features nearly across the board.

In an effort to optimize our results from our One vs Rest decision tree scheme, we implemented the same setup using random forests with 10 estimators. We were surprised to find that nearly across the board, ROC scores either fell or remained constant. Many of the features remained in the same order when organized by highest ROC score compared with the decision tree classifiers, with homicide and suicide receiving the top scores (0.91) and (0.87) respectively. The pending investigation class fluctuated most drastically, falling (0.09) points in the random forest classifier (0.77) compared to the decision tree classifier (0.86). Additionally, the pending investigation class was the only class in to suffer a sharp sudden decline in ROC score, as seen in Figure 6 before rising sharply to rejoin the other features at the end of the ROC curve.

In terms of feature importance, our random forest multiclass classifier corroborated previous classifiers valuation of `detail_age` as the most highly scored feature of the dataset, scoring this feature in the mid to high (0.80's). When graphing the feature importance, we provided error bars based on the Gaussian distribution over the various individual trees inside the forest. We can, therefore, offer an accurate range and likely median value for the score of the feature importance of our random forest multiclass classifier.

Given our limited time, we were unable to come to any firm conclusions as to why the random forest multiclass classifier reacted to drastically different with this particular feature over other classifier models, although it may warrant further in-depth analysis. We can propose the hypothesis that 10 estimators was insufficient to fully capture a wide range of hyperparameter settings. We limited our analysis to 10 estimators to keep computation time in check, so future work may wish to expand on the utility of random forests in manner of death prediction.

5.2 K-means Clustering

While K-Means clustering failed to produce easy-to-analyze clusters, we were able to analyze why this method may have been ineffective for us and offer some potential next steps if this method were to be attempted again on this dataset. As indicated by our unacceptably high variance within the clusters, the lack of distinctness between clusters, as well as our difficult-to-interpret plots, our method and dataset (post pre-processing) were unsuitable. Future work in regards to clustering on this dataset will more than likely require a different preprocessing process than the one that was able to produce intersting results in classification. Binary attributes are not ideal for this type of cluster analysis. Furthermore, the rest of the attributes as a whole will require a higher degree of variety in order to produce interesting clusters.

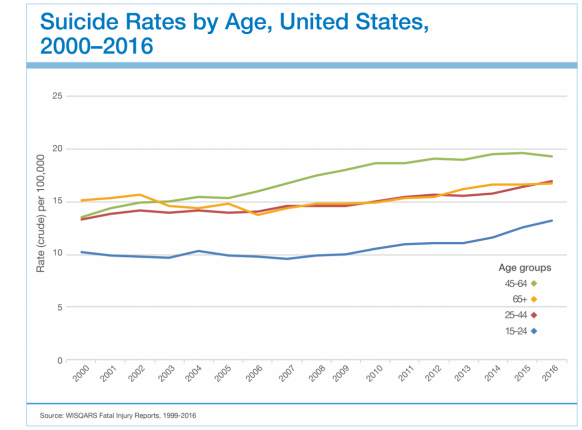

Figure 15: Suicide trends by age from 2000 to 2016

Note that other clustering methods may be more suited for this type of clustering task. While outside of the scope of this analysis, density-based clustering methods (such as DBSCAN) could potentially be utilized in an effort to extract meaningful trends from the data.

Another potential avenue for improving cluster analysis may be to utilize sampling. The high number of objects resulted in plots loaded with datapoints. This made it difficult to draw meaningful conclusions. Effective sampling as well as increasing the variety in our attributes could very well be the next step in producing interesting results.

5.3 Comparing findings

5.3.1 Suicide. After discovering the apparent importance of distinctive features in both suicide and homicide, we wanted to see if other researchers had observed similar results. Research conducted by the Suicide Prevention Resource Center over a very similar time period and geography to our own study (Suicide rates with respect to age in the USA from 2000 to 2016) seem to come to verify the feature importance of age in respect to suicide as the cause of death. In figures 15 and 16, we see both the reported suicide rates of Americans per 100,000 in various age groups from 2000 to 2016 and the ranking of suicide, in respect to the top 10 leading cause of death in the same age groups over the same time period. As illustrated particularly clearly in figure 16,the ranking of suicide as a top 10 leading cause of death, age appears to be of high importance for a majority of the age ranges. [4]

5.3.2 Homicide. Looking into work done surrounding the importance of age and race with homicide as the cause of death, a 28 year longitudinal study by the USA Bureau of Justice looking at homicide trends in the USA from 1980 to 2008 offered similar results to our study. As seen in figure 17, both age and race were prominent features with respect to the victims of homicide also supporting our finding that race coded as black_binary presented

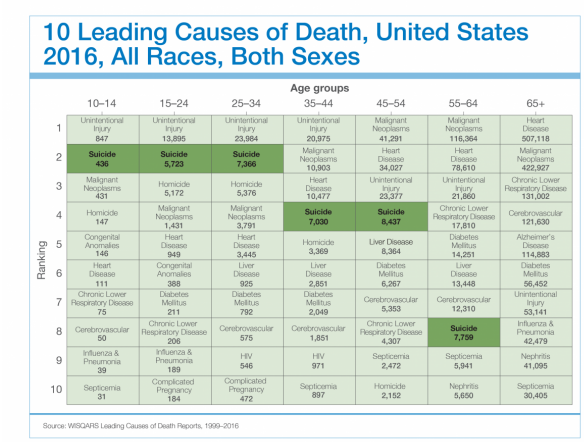

Figure 16: Leading cause of Death in the USA 2000 to 2016

TABLE 1
Victims and offenders, by demographic group, 1980–2008

	Percent of—			Rate per 100,000	
	Victims	Offenders	Population	Victims	Offenders
Total	100%	100%	100%	7.4	8.3
Age					
Under 14	4.8%	0.5%	20.0%	1.8	0.2
14-17	5.2	10.6	5.8	6.6	15.0
18-24	24.4	37.5	10.6	17.1	29.3
25-34	28.7	28.0	15.6	13.7	14.9
35-49	22.8	17.1	21.1	8.0	6.7
50-64	8.9	4.9	14.7	4.5	2.7
65 or older	5.1	1.6	12.3	3.1	1.1
Sex					
Male	76.8%	89.5%	48.9%	11.6	15.1
Female	23.2	10.5	51.1	3.4	1.7
Race					
White	50.3%	45.3%	82.9%	4.5	4.5
Black	47.4	52.5	12.6	27.8	34.4
Other*	2.3	2.2	4.4	3.8	4.1

*Other race includes American Indians, Native Alaskans, Asians, Native Hawaiians, and other Pacific Islanders.

Figure 17: Homicide trends in the USA from 1980 to 2008 from the Bureau of Justice

itself as the highest scoring race identifier with respect to victims of homicide.

6 APPLICATION OF KNOWLEDGE

6.1 Contribution to Existing Literature

While we have been unable to provide strong evidence or meaningful correlation between the features we explored and cause of death, our finding do suggest strong feature importance of age in both suicide and homicide (as well as the other manners of death), with the addition of race scoring high as an important feature with respect to homicide. To further qualify and support our results, and given that our findings seem to agree with independent studies conducted on similar trends, we can firstly conclude that these are well documented trends and are generally widely accepted prior to our work. Therefore, in lieu of specific actionable steps, we offer additional broad support for these findings.

Since there are clear and persistent findings of age being of high importance with respect to suicide and both age and race exhibiting high feature importance when homicide is the cause of death, we believe that greater research must be done to drill into both these areas in an attempt to explore deeper correlations and trends. The role of corroborating work in academia is well established extremely important throughout the scientific community.

6.2 Proof of Concept for Social Work Applications

Aside from the utility of feature importances in validating existing intuition and research on mortality, the relative success of decision trees and random forests in the classification of suicide and homicide suggest potential uses for similar models in a social work setting. While it is clear our models are far too generalized and have a limited feature set, the ease and interpretability of our chosen classifiers could lend themselves to an application in social work, potentially as a method for targeting support information to relevant demographics.

An example of such a system may be one in a university mental health center–employing a model to assess individual's risk of a variety of negative outcomes, both lethal and non-lethal. This type of targeting can ensure more relevant information to specific sets of people, and allow a more personalized support system. Note, again, that universities or other institutions implementing such a system would presumably have a much greater amount of context-specific information on the constituent base. Furthermore, it is imperative that such a system would undergo a thorough examination by domain experts in social science to ensure the validity of the models and the maintenance of privacy and safety standards.

6.3 Data Preprocessing Tool

Our classification results present the opportunity to return to our initial preprocessing step of removing a significant number of instances from our dataset with missing manner of death values. While this was acceptable, in our case, due to the large size of the original dataset, future analyses of this data may wish to use manner of death as a feature. If future work is able to improve on the multiclass classification performance for manner of death, these models could be utilized to fill in missing (NA or NaN) values.

6.4 Demographic Risk Publicization

Understanding that both our findings as well as existing literature acknowledge the fact that age seems to be the highest weighted feature in suicide deaths–and suicide is a high cause of death in the United States–indicates that this information should be more widely publicized throughout general media. Relating back to the contribution of these findings to existing literature, our work can serve to provide additional scientific backing to efforts to acknowledge at-risk demographics. Secondly, individuals in positions of trust, parents, teachers, caretakers and legal guardians, should be made aware of the seeming importance age plays with respect to suicide related deaths.[7]

A ONE VS REST DECISION TREE

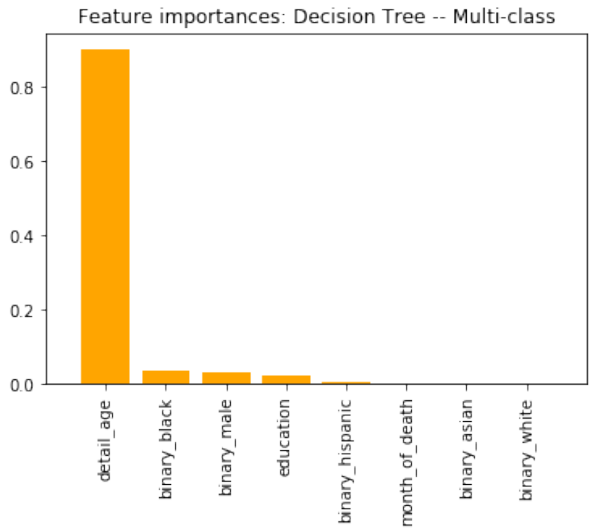

Figure 18: Feature importance for One vs Rest decision tree classifier with maximum depth of 10.

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] 2013. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. (2013). Retrieved March 27, 2018 from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824/

[3] 2015. NVSS - Mortality Data. (2015). Retrieved March 2, 2018 from https://www.cdc.gov/nchs/nvss/deaths.htm

[4] 2016. Suicide Rates by age, USA 2000 to 2016. (2016). Retrieved March 22, 2018 from https://www.sprc.org/scope/age/

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

[6] 2018. Death in the United States. (2018). Retrieved March 5, 2018 from https://www.kaggle.com/cdc/mortality/kernels

[7] Bureau of Justice Statistics 2008. Homicide Trends in the United States, 1980-2008. (2008). Retrieved March 22, 2018 from https://www.bjs.gov/content/pub/pdf/htus8008/

[8] 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

[9] 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