CPE695 Final Report

1st Junyan Yang
Department of Computer Science
Stevens Institute of technology
Hoboken, USA
jyang68@stevens.edu

2nd Xiaohan Hou
Department of Computer Science
Stevens Institute of technology
Hoboken, USA
xhou9@stevens.edu

3rd Hongyi Zhang
Department of Computer Science
Stevens Institute of technology
Hoboken, USA
hzhan29@stevens.edu

***Abstract*—In this project, we are going to look into the The Big Cities Health Inventory dataset. We are going to analyze the dataset, and perform some machine learning technology to see if the result can be accurately predicted.**

***Index Terms*—Machine Learning, Big Cities Health Inventory (BCHI) data platform**

I. Introduction

The Big Cities Health Inventory dataset includes 569 health datasheet records including more than 50 indicators that describe health status, death rates, and other socio-economic and demographic factors. The purpose of the dataset was to identify the relationship between leading causes of morbidity and mortality in the United States. The 9 key leading causes of morbidity included: Behavioral Health and Substance Abuse, Cancer, Chronic Disease, Environmental Health, Food Safety, HIV/AIDs, Infectious Disease, Injury and Violence, and Maternal and Child Health.

	Indicator Category	Indicator	Year	Sex	Race/Ethnicity	Value	Place	BCHC Requested Methodology	Source	Methods	Notes	90% Confidence Level - Low	90% Confidence Level - High	90% Confidence Level - High	95% Confidence Level - Low	95% Confidence Level - High
0	Behavioral Health/Substance Abuse	Opioid-Related Unintentional Drug Overdose Mor...	2010	Both	Black	1.8	U.S, Total, U.S, Total	Age-Adjusted rate of opioid-related mortality ...	CDC WONDER	NaN	This indicator is not exclusive of other drugs...	NaN	NaN	NaN	1.7	1.9
1	Behavioral Health/Substance Abuse	Opioid-Related Unintentional Drug Overdose Mor...	2010	Both	American Indian/Alaska Native	6.8	U.S, Total, U.S, Total	Age-Adjusted rate of opioid-related mortality ...	CDC WONDER	NaN	This indicator is not exclusive of other drugs...	NaN	NaN	NaN	5.8	7.9
2	Behavioral Health/Substance Abuse	Opioid-Related Unintentional Drug Overdose Mor...	2010	Both	All	4.4	U.S, Total, U.S, Total	Age-Adjusted rate of opioid-related mortality ...	CDC WONDER	NaN	This indicator is not exclusive of other drugs...	NaN	NaN	NaN	4.4	4.5
3	Behavioral Health/Substance Abuse	Opioid-Related Unintentional Drug Overdose Mor...	2010	Both	White	3.3	Fort Worth (Tarrant County), TX	Age-adjusted rate of opioid-related mortality ...	National Center for Health Statistics	NaN	This indicator is not exclusive of other drugs...	NaN	NaN	NaN	2.4	4.9
4	Behavioral Health/Substance Abuse	Opioid-Related Unintentional Drug Overdose Mor...	2010	Both	Other	NaN	Fort Worth (Tarrant County), TX	Age-adjusted rate of opioid-related mortality ...	National Center for Health Statistics	NaN	Records where the value is blank in the data f...	NaN	NaN	NaN	NaN	NaN

After reviewing the data set, we decided to define the attribute "Value" as the dependent variable y as it represented the number of deaths per 100,000 people. We found this to be appropriate given that we were interested in understanding what factors and attributes drive the death rate in our population. After defining "Value" as the dependent variable y, we did some exploration to find and define the independent variables which would have a possible correlation to y. The goals and our next steps would revolve around finding these variables, creating a model based on algorithms we learnt, which then would allow us to predict "Value" (dependent variable) given inputs we enter.

Health and well-being has been a great focus for the public in recent years. By being able to predict the Death Rate for a specific group of people given their characteristics, the government may be able to better forecast the mortality rate for people of all types. They would also be able to forecast mortality rates if special circumstances arise and change status quo (e.g. COVID-19). As a result, we found further investigation into this issue to be intriguing and whether we'd be able to build a model to better estimate death rates if we gave a model a set of characteristics as input.

II. Existing Solution to This Problem

Reviewing the internet, we've noticed the Big Cities Health Data Platform is used quite extensively for various research and case studies regarding different aspects of health among people in the USA. The latest report/analysis performed in January 2019 by students to analyze the data and present their findings regarding the relationship between obesity and income. Additionally, another analysis was performed looking at the connection between drug overdose and mortality.

While the data set was rather considerable, sharing many health related attributes across 30 cities, we realized all the analysis were performed focusing on a specific health condition, for example:

- Opiod overdose
- Smoking
- Food safety
- HIV
- Violence

However, we noticed no analyses were performed on the entire data set using AI techniques to understand the correlation between all the health attributes and the death rate. We found it would be interesting to see if we could find trends and develop a model that could predict a mortality rate given a person's attributes. The one paper we have reviewed in detail is from Michael Benusic, in his paper, his focus is on relationship between obesity and health. He gave out an vivid image on the how wealthy families have an impact on obesity.

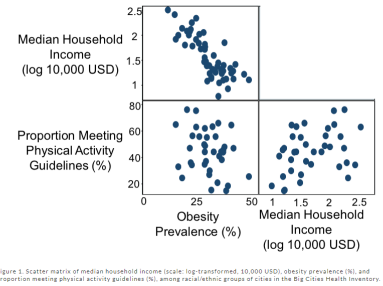


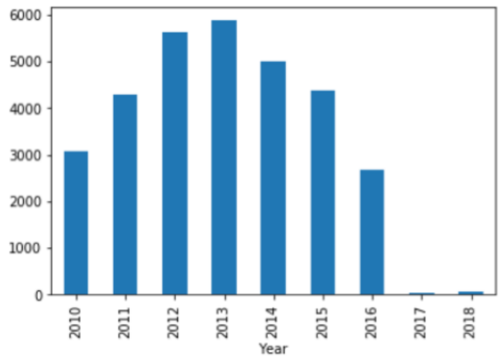
Figure 1. Scatter matrix of median household income (scale: log-transformed, 10,000 USD), obesity prevalence (%), and proportion meeting physical activity guidelines (%), among racial/ethnic groups of cities in the Big Cities Health Inventory.

This is a great example on how our topic can be researched further.

III. Data Description / Analysis

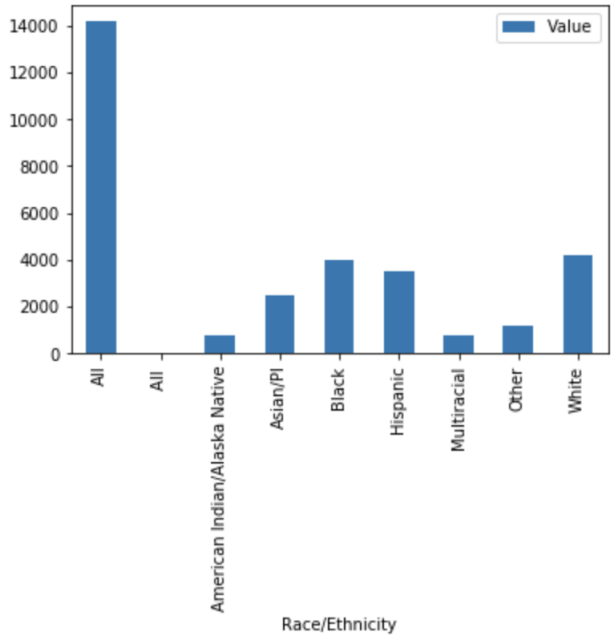
The below summarizes our results from our analysis of the data parameters in the dataset and how they provide meaningful information or potential trends. We utilized bar charts and graphs to plot out various aspects of our data to identify trends and to support the inclusion of certain parameters as independent variables and for our algorithms.

A. Year vs. Number of recorded values

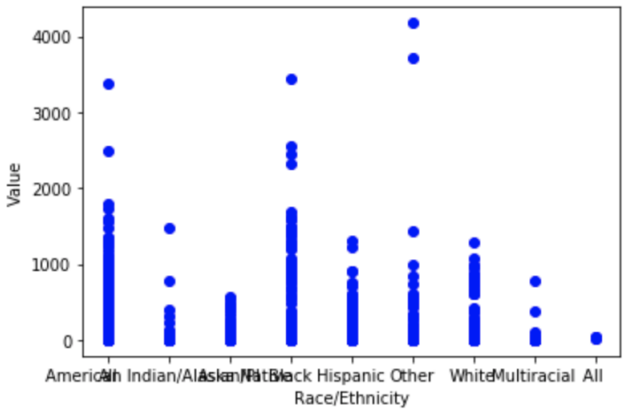


Plotting the total number of recorded samples against the parameter year from 2010 to 2018 allowed us to identify a high level trend, that the number of recorded values was increasing rapidly from 2010 to 2013, peaking in 2013. Beginning 2014, the number of recorded values began to steadily decrease to 2016. The drop off noted in 2017 and 2018 is likely due to a lack of available data.

B. Race/Ethnicity

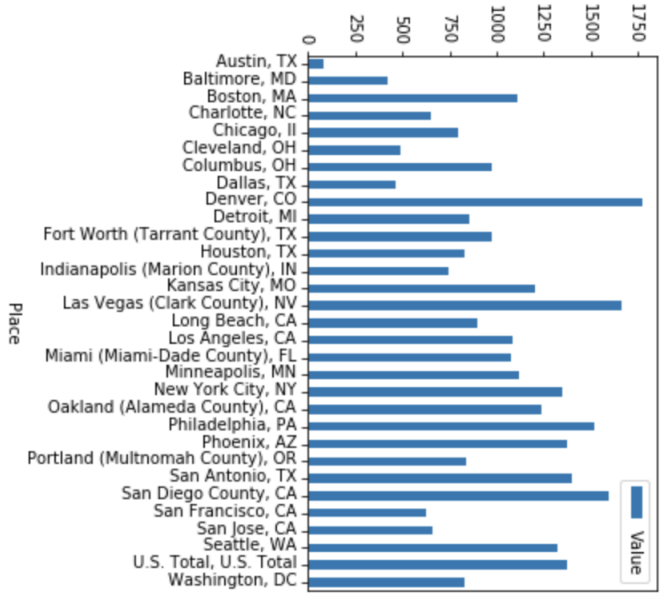


This chart displays the number of recorded values across each race/ethnicity. Reviewing the graph, the results are rather inconclusive as approximately 14,000 of the data points assigned a race/ethnicity value of “All” indicating that the samples were in majority not race specific. The remaining split was relatively even.



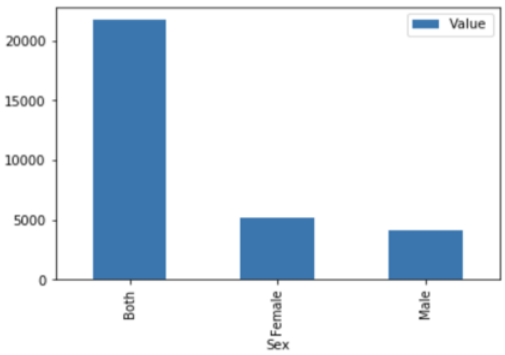
To dig this variable further, we tried to see if there’s any correlation that we can find in the scatter plot. As the scatter plot shown above, we put up the plot with all independent variables within Race and dependent variable ‘Value’, however, by the plot itself, we cannot discover any relationships between our independent variables and the dependent variable.

C. Place/Location

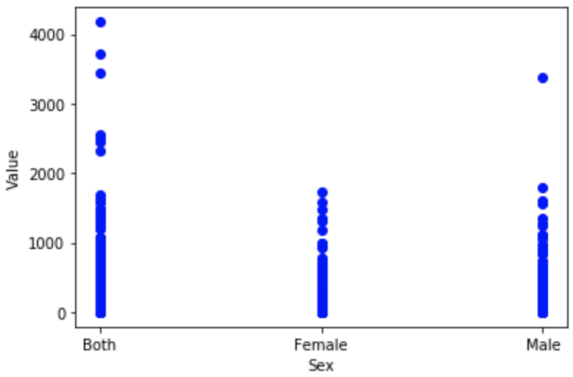


Using a bar chart to plot the number of recorded values against the city/location we are able to identify where the majority of recorded values occur and whether living in a specific location may increase the likelihood of morbidity. The leaders of the recorded values from our dataset were Denver, CO, followed by Las Vegas and San Diego County. By contrast, the recorded values in the location with the fewest recorded values was Austin, TX with a count of 80 drastically less compared with Denver.

D. Gender



The following chart shows the recorded values distribution for the different genders. Similar with our findings from our analysis of Race, many data entries did not assign a specific gender (Both). As a result, given the fairly equal distribution, we are unable to conclude how gender currently impacts the number of recorded values.

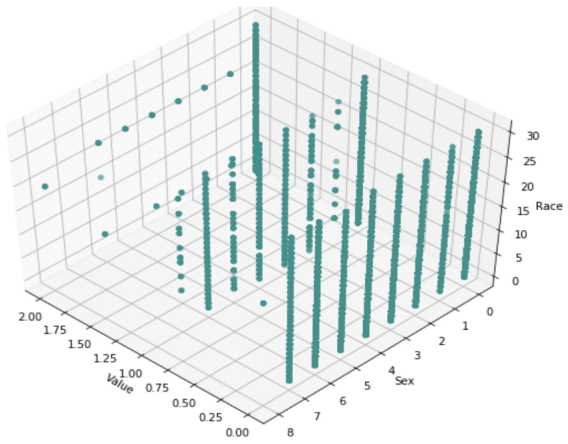


Here, we also try to plot the scatter plot of 'Sex' and 'Value'. As you can see in the plot, there are no clustering or linear relationship between these two variables.

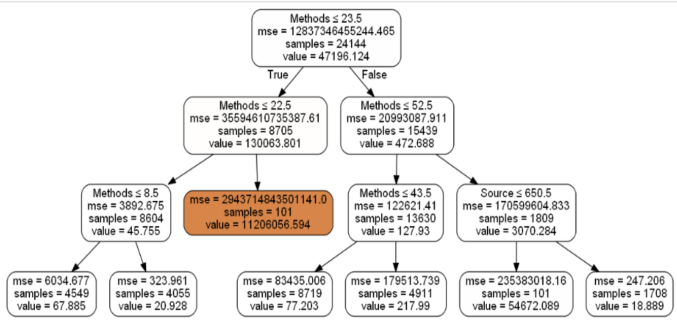
IV. Implementation Process

A. Early Stage- Exploring The Data

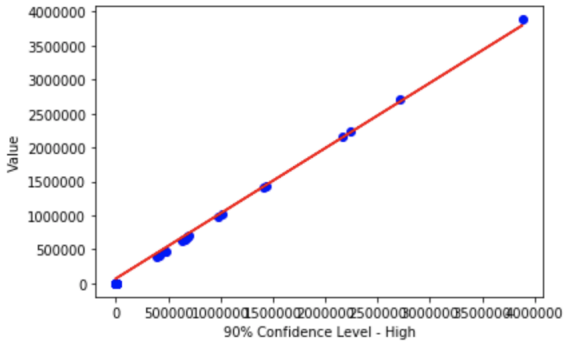
During our early stage of our project, we decided to try out some supervised and unsupervised learning algorithms. However, this data set contained a lot of categorical variables. Normally, these categorical variables were handled by defactorized them into numerical variables, but after our first try out of modeling. The models don't seem to work the way we desire. Below is the graph for k-means clustering. As shown, we can notice that there are no obvious clusters found by using 'Race', 'Sex', 'Value'.



We also use a regression tree model during our early modeling stage. The result is just as bad. As shown in the graph below, we can tell that our tree doesn't lead us to any decision at all.



Next, we did some exploration of our only numerical variables: '90% confidence level-low' to '95% confidence level-low'. We tried to do simple linear regression with respect to each one of these confidence level columns to 'Value'. However, the scatter points and the regression line is perfectly fitted, which means the confidence level columns cannot be used as independent variables in our future modeling process. Below is an example.



In conclusion of our early stage process, we need to explore more of our data information and try doing feature engineering to select the best independent variables. Also, we need to consider a different way to transform our data in order to train our models. It's vital to do data transformation due to inefficiency of the categorical variables.

B. Defining Key Independent Variables

First, we need to deal with missing values. Below is the information of the missing values of our data set.

```
data.isnull().sum()
```

Indicator Category	0
Indicator	0
Year	0
Sex	0
Race/Ethnicity	0
Value	3444
Place	0
BCHC Requested Methodology	0
Source	4964
Methods	26968
Notes	19851
90% Confidence Level - Low	31655
90% Confidence Level - High	31655
95% Confidence Level - Low	25657
95% Confidence Level - High	25616
dtype: int64	

From 'Methods' to '95% Confidence Level- High', the missing values percentage is larger than 50%. So we decided to just drop these six columns since the missing values percentage is huge. Also we fill the missing value in the "Value" column with Average Imputation: Use the average value of the responses from the other participants to fill in the missing value. If the average of the 30 responses on the question is a 4.1, use a 4.1 as the imputed value. This choice is not always recommended because it can artificially reduce the variability of your data but in some cases makes sense.

Having reviewed the list of indicators in the data set and analyzed the plot diagrams, we decided to pinpoint the key indicators to focus on and further assess, while dropping other attributes we did not deem material or data that was unavailable(missing values).

Indicator Category	Indicator	Year	Sex	Race/Ethnicity	Value	Place	BCHC Requested Methodology	Source	Methods	Notes	90% Confidence Level - Low	90% Confidence Level - High	90% Confidence Level - High	95% Confidence Level - Low	95% Confidence Level - High

Our team designated Race, Area, and Sex to be the key drivers for determining the death rate given the following:

- Race, location and sex are attributes that are more readily available and obtainable in comparison to other indicators such as indicator category (Reason for death)
- The indicator category (Reason for death) was not selected as a key indicator as we believe the correlation between this independent variable and the dependent variable (number of deaths) would be too high and prevalent and might outweigh the impact of other variables.
- Additionally, when attempting to apply machine learning and create a model for estimating the number of deaths, the reason or cause of death might not be available.

C. Pre-processing Data

As many of the indicator values in the data set were not numerical but strings, we were unable to directly analyze or quantify the correlation between the independent variables and dependent variable. As a result, with each key indicator noted

above, we transformed their categorical features into numerical features by implementing one-hot coding illustrated below.

This would allow our data set to be analyzed deeper.

```
In [17]: df_categorical_col = data.select_dtypes(exclude=np.number).columns
         df_categorical_col

Out[17]: Index(['Sex', 'Race/Ethnicity', 'Place'], dtype='object')

In [18]: df_numeric_col = data.select_dtypes(include=np.number).columns
         df_numeric_col

Out[18]: Index(['Value'], dtype='object')

In [19]: df_onehot = pd.get_dummies(data[df_categorical_col])
         df_onehot.head(5)

Out[19]:
```

	Sex_Both	Sex_Female	Sex_Male	Race/Ethnicity_All	Race/Ethnicity_All	Race/Ethnicity_American Indian/Alaska Native	Race/Ethnicity_Asian/PI	Race/Ethnicity_Black	Race/
0	1	0	0	1	0	0	0	0	
1	1	0	0	1	0	0	0	0	
2	1	0	0	1	0	0	0	0	
3	1	0	0	1	0	0	0	0	
4	1	0	0	1	0	0	0	0	

5 rows × 43 columns

D. Divided training data vs test data

First of all, we tried cross validation to find the best way to divide our data into training data set and test data set, but the results after each cross validation step are very different. Therefore we just apply the simple train and test split to divide our data set. Next, we divided our population data set into two categories,
a) Training Data - Data used to train our model and
b) Test Data - Remaining data population to later assess the accuracy of our model and the residual error.

E. Algorithms

We implemented various algorithm methodologies to assess which would provide us the lowest level of errors and highest levels of correlation between our variables. Below details our results from applying the following methods.

- linear regression
- k Nearest Neighbor (kNN Model)
- Regression Tree
- SVM Algorithm

1) Multi-Linear Regression: Applying Multi-linear regressions first allowed us to identify whether linear relationships existed between the individual independent variables and the dependent variable Value (Death rate).

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n \tag{1}$$

We then have our weights as can be shown in the following image

```
array([[ 2.59874034e+17,  2.59874034e+17,  2.59874034e+17,
        -9.56348050e+16, -9.56348050e+16, -9.56348050e+16,
        -9.56348050e+16, -9.56348050e+16, -9.56348050e+16,
        -9.56348050e+16, -9.56348050e+16, -9.56348050e+16,
        -3.21955357e+17, -3.21955357e+17, -3.21955357e+17,
        -3.21955357e+17, -3.21955357e+17, -3.21955357e+17,
        -3.21955357e+17, -3.21955357e+17, -3.21955357e+17,
        -3.21955357e+17, -3.21955357e+17, -3.21955357e+17,
        -3.21955357e+17, -3.21955357e+17, -3.21955357e+17,
        -3.21955357e+17, -3.21955357e+17, -3.21955357e+17,
        -3.21955357e+17, -3.21955357e+17, -3.21955357e+17,
        -3.21955357e+17, -3.21955357e+17, -3.21955357e+17,
        -3.21955357e+17, -3.21955357e+17, -3.21955357e+17,
        -3.21955357e+17, -3.21955357e+17, -3.21955357e+17,
        -3.21955357e+17]])
```

After getting the parameters from the training data, we then applied the model from above to the test data which provided the result of $RMSE$ (Root-mean-square deviation)=3331346.43963654.

2) kNN Model: kNN is the simplest supervised algorithm mostly used for classifying data points based on how it's neighbors are classified. In our earlier steps, we transformed all categorical features into numerical discrete attributes 0,1 for the instances, so that we can apply the kNN algorithms here. In order to determine the best value for "k", we assessed the RMSE as we increased the factor of k. The best k is the one that minimizes the prediction error RMSE. Referring to the chart below, we were able to conclude that utilizing a k factor of 3 resulted in the lowest RMSE.

```
Rmse for k = 1 is 3334819.233243835
Rmse for k = 2 is 3334332.7992320075
Rmse for k = 3 is 3334185.2569772894
Rmse for k = 4 is 3334972.892561651
Rmse for k = 5 is 3334744.266393955
Rmse for k = 6 is 3334494.7269203314
Rmse for k = 7 is 3334352.040725512
Rmse for k = 8 is 3334467.092127072
Rmse for k = 9 is 3334915.375326055
Rmse for k = 10 is 3334689.428503498
```

The RMSE for test data when k=3 is the minimum given only 10 iterations (Root-mean-square deviation)=3334185.2569772894, we will do further iterations of the neighbor number to maybe find a smaller RMSE for the model.

3) Regression Tree: Our next method applied involved decision tree learning. Regression tree analysis is when the predicted outcome can be considered a real number(wiki). The goal was to create a model that predicts the value of our target variable based on the several input variables/parameters noted earlier. The predicted outcome can be considered a real number, which in this case would represent the 'value' given our input. We will apply Reduced Error Pruning for the tree in later experiments. The RMSE for test data when k=4 is the minimal given only 10 iterations

```
Rmse for k = 1 is 3331917.708214744
Rmse for k = 2 is 3329179.579061839
Rmse for k = 3 is 3315780.0760172135
Rmse for k = 4 is 3315742.1185617554
Rmse for k = 5 is 3315751.480798816
Rmse for k = 6 is 3315762.327509618
Rmse for k = 7 is 3315762.207724649
Rmse for k = 8 is 3315753.450087217
Rmse for k = 9 is 3315753.2626337833
Rmse for k = 10 is 3315757.476472194
```

(Root-mean-square deviation)=3315742.1185617554, we will do further iterations of the max-node number to maybe find a smaller RMSE for the model.

4) SVM (Regression): For our last model, we used a support vector machine for regression.SVR gives us the flexibility to define how much error is acceptable in our model and will find an appropriate line (or hyperplane in higher dimensions) to fit the data. Our predicted 'Values' seems very strange, as you can see in the graph below. The rmse is a little bit better than the rmse value of our KNN model.

```
svmpred=svm.predict(test_x)
svmpred
```

```
array([17.60197907, 17.66959576, 17.65887972, ..., 17.60762637,
       17.66446414, 17.61329883])
```

```
test_aberror=metrics.mean_squared_error(test_y,svmpred)
rmse=np.sqrt(test_aberror)
print('svm regression Rmse is',rmse)
```

```
svm regression Rmse is 3334043.0159499
```

F. Tuning parameters

The process is mentioned under KNN model. Basically, we trained the model with different K values and compared which k has the least RMSE value.

V. Conclusion

Having run through the four different algorithm approaches, to determine which model was the best fit for our data set, we reviewed the model output, accuracy results (RMSE), computational costs, and advantages/disadvantages for each model.
HIGH LEVEL FINDING: which model is the most practical and provides something tangible to the user

A. RMSE and Accuracy

Our first goal was to identify which model would provide us the lowest level of RMSE from when applying our model/technique to the test data. From our results noted above, to minimize the RMSE, the model using multi-Linear Regression was by far our best approach with a RMSE =3331346.43963654. Given our current input, the model using kNN had the worst performance with the RMSE 3334185.2569772894 with k=3.

B. Computational Cost

We then assessed the computational cost for each of the models and techniques applied. Our focus was to identify the efficiency and speed when running each which is important in the circumstance we had to increase the scale and size of our dataset. Below is a summary of the times and computational costs for each model applied:

The model with the lowest computational cost was the Multi-linear Regression model while the model with the highest computational cost was the KNN model. It is noted however, that although the cost was relatively high, this weakness was offset with a higher RMSE.

C. Advantages vs. Disadvantages for Each Approach

Each model applied on our datasets provided us a different perspective and view. The below summarizes the strengths and weaknesses for each of the models applied and our final recommendation on which model provides the greatest value.
1) Multi-Linear regression and Polynomial Regression: The linear regression was extremely useful as a visual guide to understand at a base level the level of interaction or correlation between each of the independent variables and the dependent variable we were trying to understand and predict (Death rate). We could understand which of the parameters within the entire data set were the most relevant to the dependent variable.

However beyond understanding correlation, we were unable to fully see the full picture as there were multiple variables, and each variable played a part in being able to predict the death rate.
2) k Nearest Neighbor (kNN Model): The KNN model is useful when we face a classification problem, but it can also be applied for a regression problem. Its advantages include the following 1. No Training Period: kNN is called Lazy Learner (Instance based learning). It does not learn anything in the training period. It does not derive any discriminative function from the training data. In other words, there is no training period for it. It stores the training data set and learns from it only at the time of making real time predictions. This makes the kNN algorithm much faster than other algorithms that require training e.g. SVM, Linear Regression etc. 2. Since the kNN algorithm requires no training before making predictions, new data can be added seamlessly which will not impact the accuracy of the algorithm. 3. kNN is very easy to implement. There are only two parameters required to implement kNN i.e. the value of K and the distance function (e.g. Euclidean or Manhattan etc.)

It also has disadvantages: . Does not work well with large dataset: In large datasets, the cost of calculating the distance between the new point and each existing point is huge which degrades the performance of the algorithm. 2. Does not work well with high dimensions: The KNN algorithm doesn't work

well with high dimensional data because with large numbers of dimensions, it becomes difficult for the algorithm to calculate the distance in each dimension. 3. Need feature scaling: We need to do feature scaling (standardization and normalization) before applying kNN algorithm to any dataset. If we don't do so, kNN may generate wrong predictions. 4. Sensitive to noisy data, missing values and outliers: KNN is sensitive to noise in the dataset. We need to manually impute missing values and remove outliers.

3) Regression Tree: The advantages of the Regression Tree model are: A decision tree does not require normalization of data. A decision tree does not require scaling of data as well. Missing values in the data also does NOT affect the process of building decision trees to any considerable extent. A Decision trees model is very intuitive and easy to explain to technical teams as well as stakeholders.

The disadvantages of using the Regression Tree model are: A small change in the data can cause a large change in the structure of the decision tree causing instability. For a Decision tree sometimes calculation can go far more complex compared to other algorithms. Decision tree often involve higher time to train the model. Decision tree training is relatively expensive as complexity and time taken is more. The Decision Tree algorithm is inadequate for applying regression and predicting continuous values.

4) SVM Algorithm: The SVM model also has outstanding advantages: SVM works relatively well when there is a clear margin of separation between classes. SVM is more effective in high dimensional spaces. SVM is effective in cases where the number of dimensions is greater than the number of samples. SVM is relatively memory efficient.

However, the disadvantages can not be ignored:SVM algorithm is not suitable for large data sets. SVM does not perform very well, when the data set has more noise i.e. target classes are overlapping. In cases where the number of features for each data point exceeds the number of training data samples , the SVM will underperform. As the support vector classifier works by putting data points, above and below the classifying hyper plane there is no probabilistic explanation for the classification.

VI. Future Research Directions to Improve Algorithm

Understanding our dataset further and which parameters played the greatest role in impacting mortality rates in the united states, we would like to further improve our algorithm by:

a)Increasing the population size by adding additional years until the present

b)Add other parameters such as reason for death as a factor, although modifying it from a category to numerical

c)Increasing the sample size to build a better model.

d)May need to search for a different approach to a dataset with such a huge number of categorical variables.

e)In order to increase the accuracy of our models, we also need to try a different method to handle the missing values.

References

https://www.bigcitieshealth.org/methodology
https://measuringu.com/handle-missing-data/
http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-knn.html
https://medium.com/@dhiraj8899/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a https://medium.com/@dhiraj8899/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107
https://www.bigcitieshealth.org/michaelbenusicgrant