Classifying COVID Data Based on Conditions

ABSTRACT

The world is currently facing a pandemic known as COVID-19. This report aims to analyse data pertaining to this novel virus and observe the preexisting conditions and the age groups of people who have died from this virus. After preprocessing the data, it will be used in multiple algorithms to make models that will predict which condition and/or age group that person belongs to. The following algorithms were used: Logistic Regression, Spectral Clustering, Support Vector Machines, and Decision Trees. The results from all these algorithms suggest that COVID-19 and respiratory and circulation conditions have similar classifications regardless of age.

INTRODUCTION

The goal of this analysis is to build models that can make meaningful inferences about what kind of pre-existing condition group and age group a certain number of people who died of COVID-19 may have belonged to. These models are valuable because even as vaccine rollouts occur across the country and around the world, many people are still at risk of death from COVID-19, and understanding which groups of people are in need of protection can help decrease loss of human life.

I. DATASET DESCRIPTION

The dataset used for this project is from the Center for Disease Control. It provides pre-existing health conditions and contributed causes mentioned in conjunction with deaths involving COVID-19 by age group. The dataset was last updated on June 2, 2021; and was created on May 8, 2020. The dataset is updated weekly, and the number of conditions provided are tabulated from deaths received and do not represent all deaths that occurred in that period. For this project, the variables used were the name of the condition group (e.g. Respiratory Conditions, Alzheimer Disease, Diabetes, etc.), age group, and corresponding number of deaths associated with COVID-19.

METHODOLOGY

I. DATA PREPROCESSING

In order to prepare the data for the models, twelve datasets were created for each condition group, where the row value for the variable “condition_group” was converted to binary labels-- 1 was assigned to the condition group in question, and 0 was assigned to all other conditions. The same method was used to create eight datasets for each of the eight possible age groups. The predictor variable, the number of people who died of COVID-19 in each group, was isolated from the dataset and converted into a numpy array. Then, the function “train_test_split” from the sklearn package was used to split each dataset and corresponding number of people who died of COVID-19 into training and test groups. Due to the smaller size of this dataset, a test size of 40% and a training size of 60% were used. These training and testing sizes may have negatively impacted classification rates.

II. LOGISTIC REGRESSION

In order to implement the logistic regression algorithm, the “LogisticRegression” function from the sklearn package was used. Logistic regression is an algorithm that performs binary classification-- it’s a generalized linear model that predicts the probability that an event will occur. Specifically, logistic regression uses regular linear regression to model the ‘logit’ function: $l = log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_i x_i ...$

III. SPECTRAL CLUSTERING

The spectral clustering of the data was performed by first creating a function in which the data was randomly sampled with 10k entries. This was done due to the kernel crashing as it couldn’t handle datasets with more than 10k entries. However, there were conditional datasets with less than 10k entries, so the sampling is controlled by an if-statement where it checks the length of the dataset. Datasets with less than 10k entries were not sampled. A Spectral Clustering model was made, which stems from Sklearn’s cluster class. The method also included the StandardScaler and normalize function from the preprocessing class of said package to normalize the data, the PCA function from the decomposition class was used for creating the plot, and the silhouette score was implemented from the metrics class. The matplotlib package was also used to plot the clusters. The model was made using two clusters in the arguments and the affinity was set at “rbf.” THis was all encased in a function to efficiently plot each dataset, as well as a title parameter to gives the plots the correct title.

IV. SUPPORT VECTOR MACHINES

The SVC() function from the sklearn package was used for the support vector machine algorithm. Support vector machines control the trade-off between classifying trading points correctly and having a smooth decision boundary. In summary, in support vector algorithms, the classifier is the separating hyperplane and the most important training points are support vectors that define the hyperplane.

V. DECISION TREE

For “Decision Tree” we use classification and regression tree (CART). The idea is at each time we split the data by using a single feature *t* and a threshold *v*. The head node(cell) is called *the root node.* We split the data at the root node into two subsets: the left part and the right part. Searching for the pair *t, v* so that the lost function can decrease the most.

Loss Function : $$L(t, v) = \frac{n_{left}}{n} C_{left} + \frac{n_{right}}{n} C_{right}$$

Supported criteria are "gini" for the Gini index and "entropy" for the information gain. In this case we are using the Gini index.

TABLES AND FIGURES

I. LOGISTIC REGRESSION

The table below provides the classification rate for the logistic regression models for each condition group.

Condition Group	Accuracy Score	Condition Group	Accuracy Score
Respiratory	74.17%	Alzheimer Disease	95.10%
Circulatory	69.57%	Vascular/Unspecified Dementia	95.17%
Sepsis	95.87%	Renal Failure	95.75%
Malignant Neoplasms	95.75%	Intentional/Unintentional Injury/Poisoning	96.00%
Diabetes	95.71%	All other conditions	95.57%
Obesity	96.10%	COVID-19	95.29%

The table below provides the classification rate for the logistic regression models for each age group.

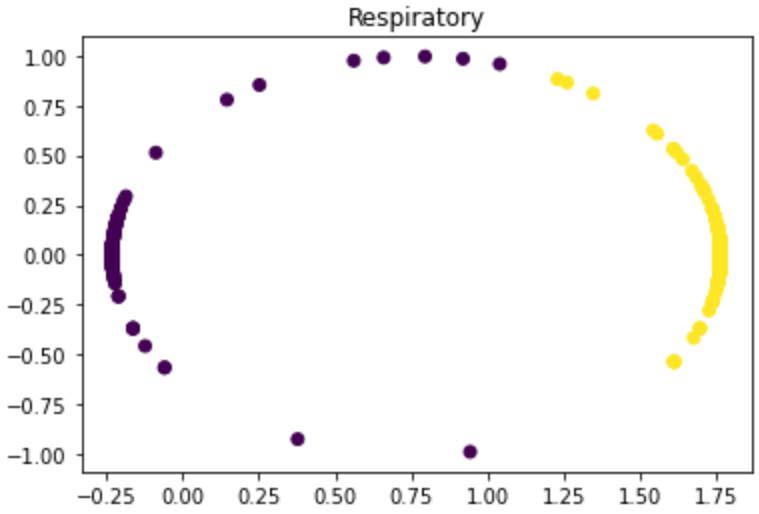
Age Group	Accuracy Score	Age Group	Accuracy Score
Ages 0 - 24	84.76%	Ages 55 - 64	88.58%
Ages 25 - 34	86.39%	Ages 65 - 74	88.07%
Ages 35 - 44	87.61%	Ages 75 - 84	88.04%
Ages 45 - 54	88.77%	Ages 85+	87.74%

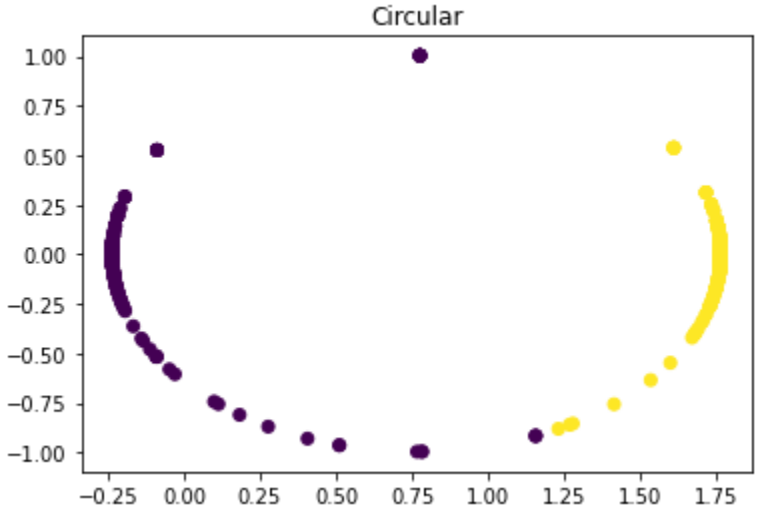
II. SPECTRAL CLUSTERING

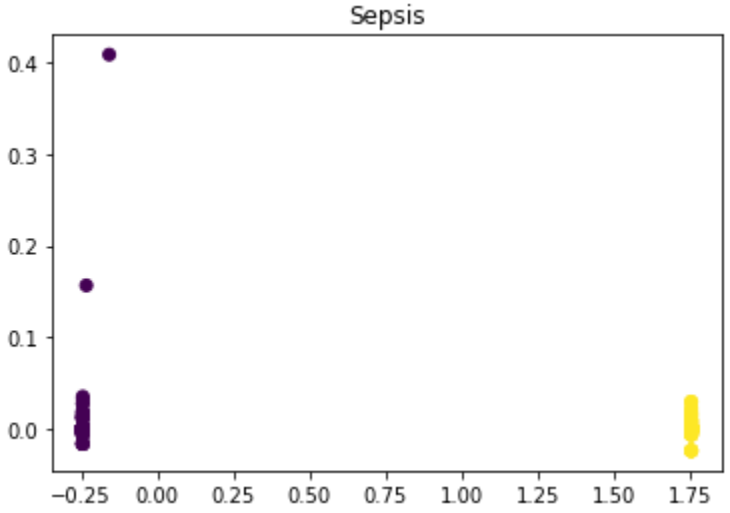
This section contains the visualization for the clusters, along with the silhouette scores for each condition group and each age group.

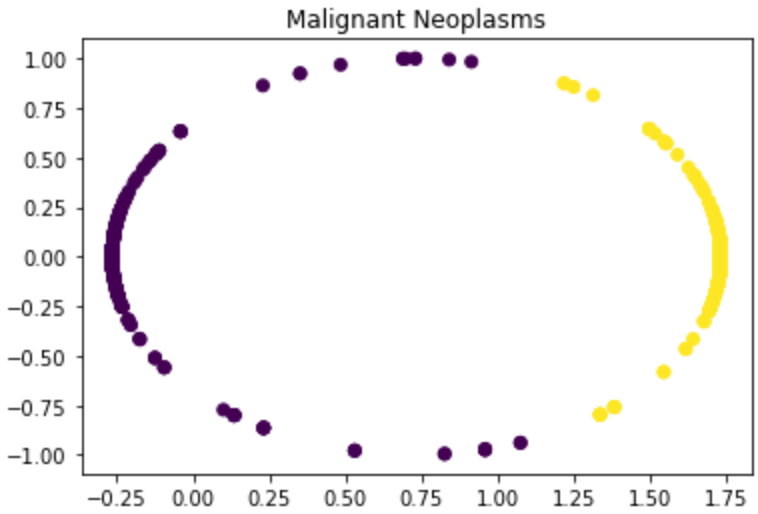
Condition Group, Visualization:

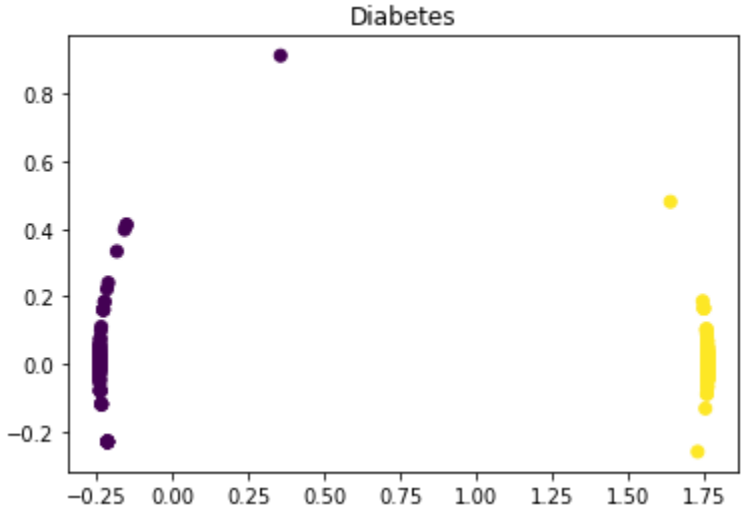
(on the next few pages)

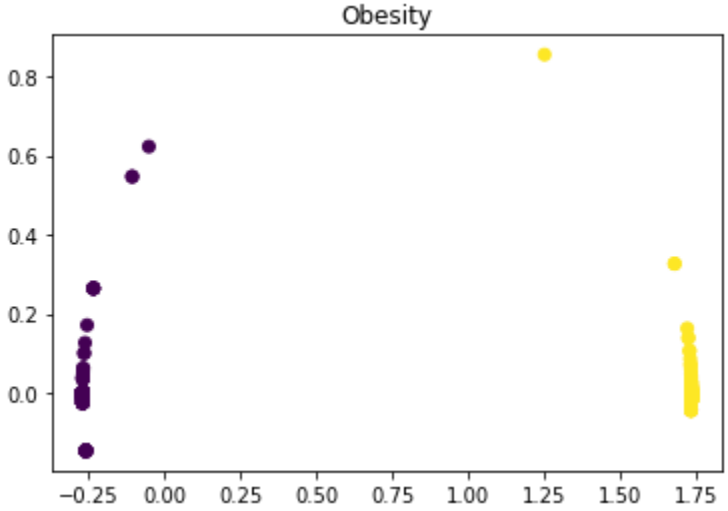

Respiratory
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75

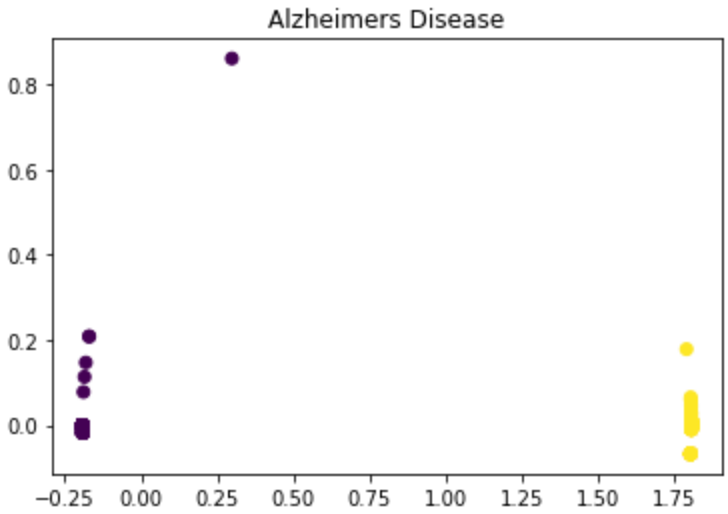

Circular
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75

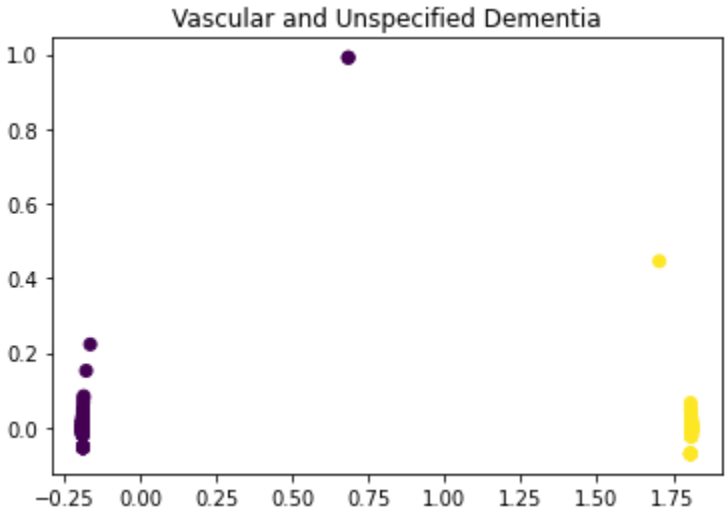

Sepsis
0.4
0.3
0.2
0.1
0.0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75

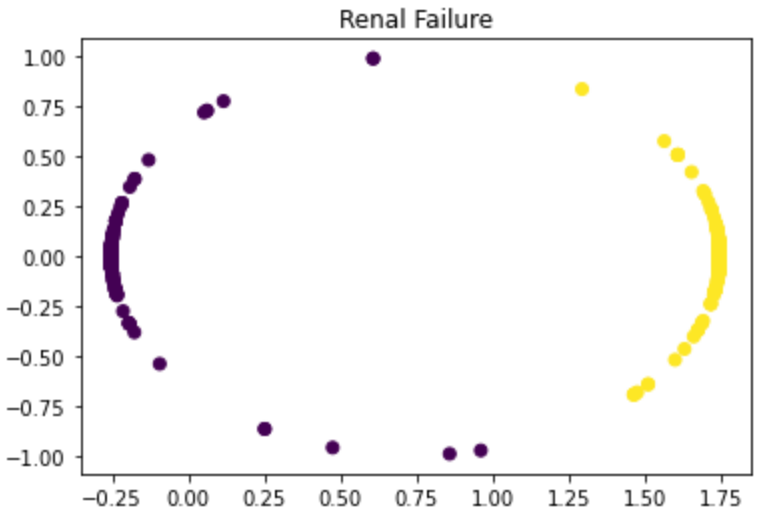

Malignant Neoplasms
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75

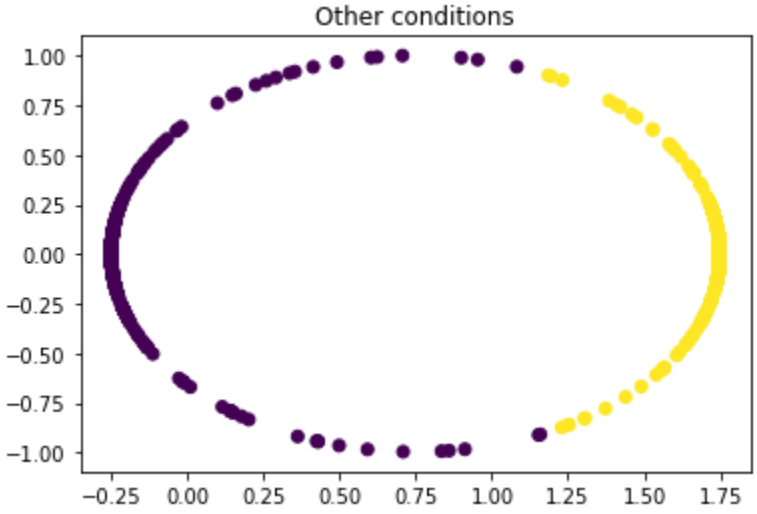

Diabetes
0.8
0.6
0.4
0.2
0.0
−0.2
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75

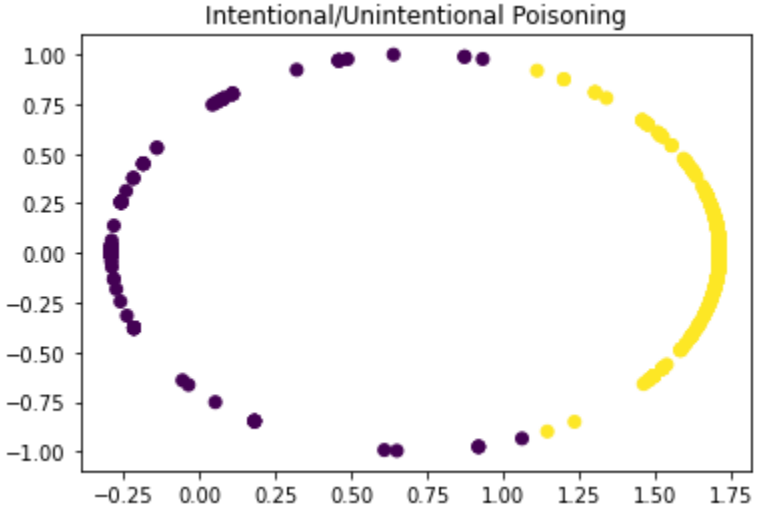

Obesity
0.8
0.6
0.4
0.2
0.0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75

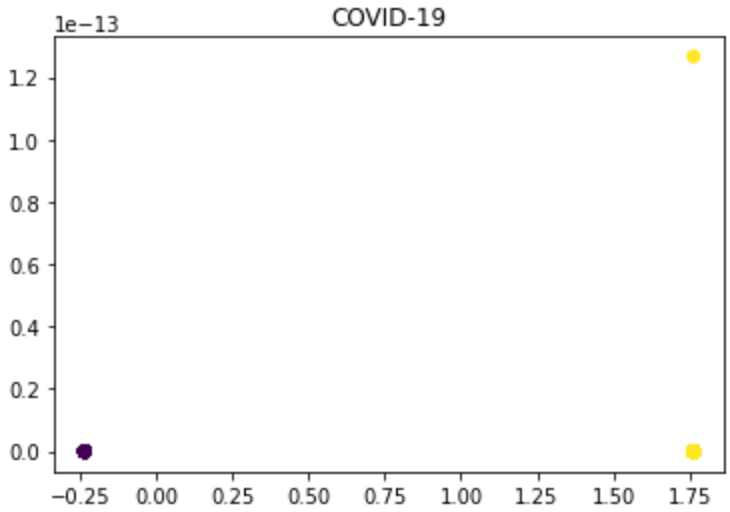

Alzheimers Disease
0.8
0.6
0.4
0.2
0.0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75


Vascular and Unspecified Dementia
1.0
0.8
0.6
0.4
0.2
0.0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75


Renal Failure
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75


Other conditions
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75

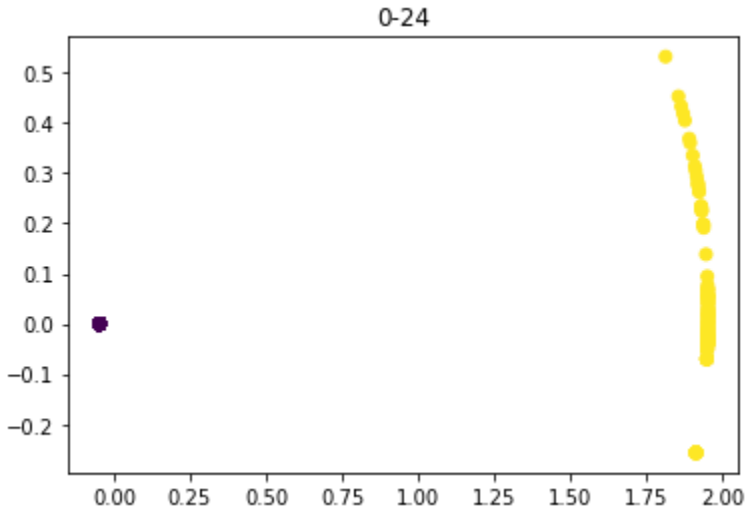

Intentional/Unintentional Poisoning
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75

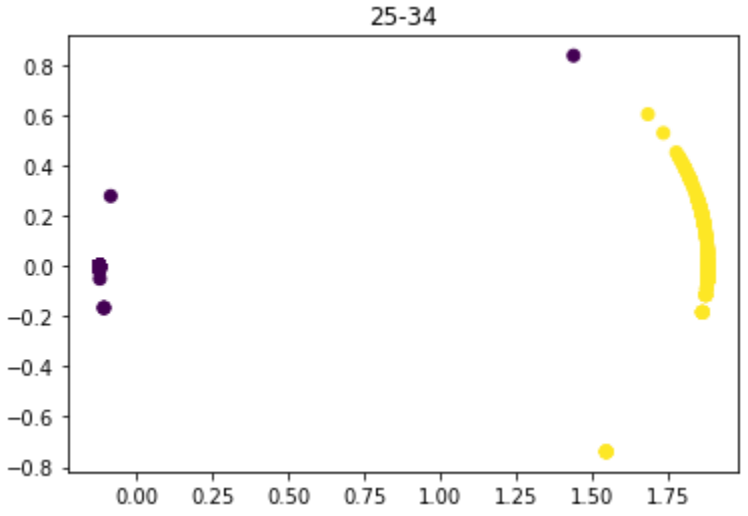


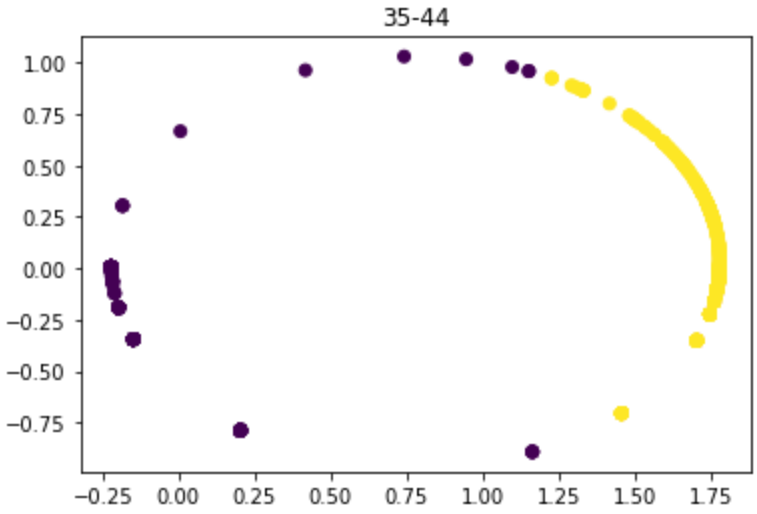
Condition Group, Scores:

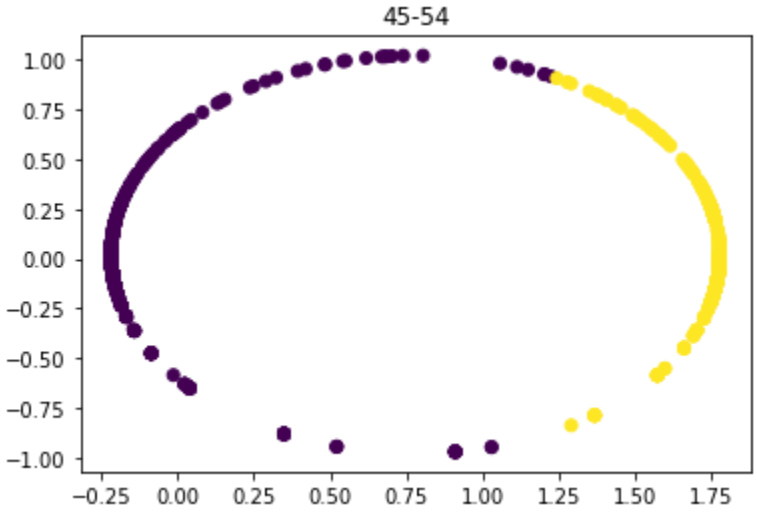
Condition Group	**Score**
Respiratory Diseases	80.75%
Circulatory Diseases	80.54%
Sepsis	73.95%
Malignant Neoplasms	78.63%
Diabetes	79.72%
Obesity	78.17%
Alzheimers	84.45%
Vascular and Unspecified Dementia	84.22%
Renal Failure	78.79%
Intentional and unintentional injury poisoning and other adverse events	78.46%
All other conditions and causes (residual)	79%
COVID-19	80.09%

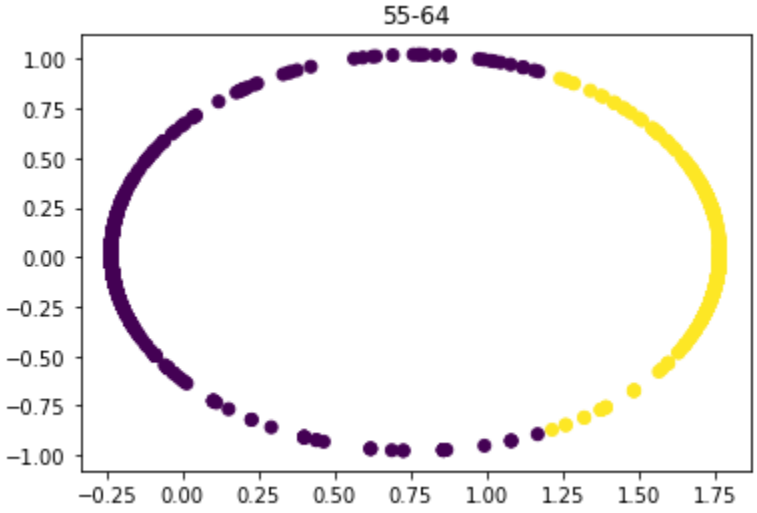
Age Group, Visualization:

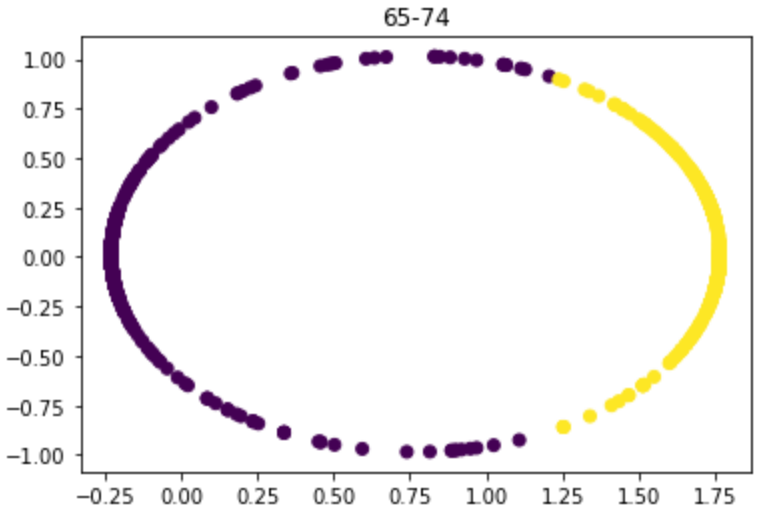


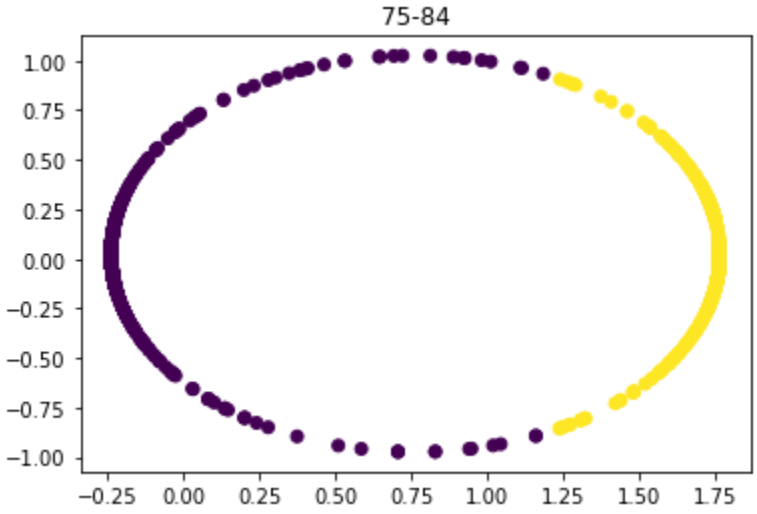


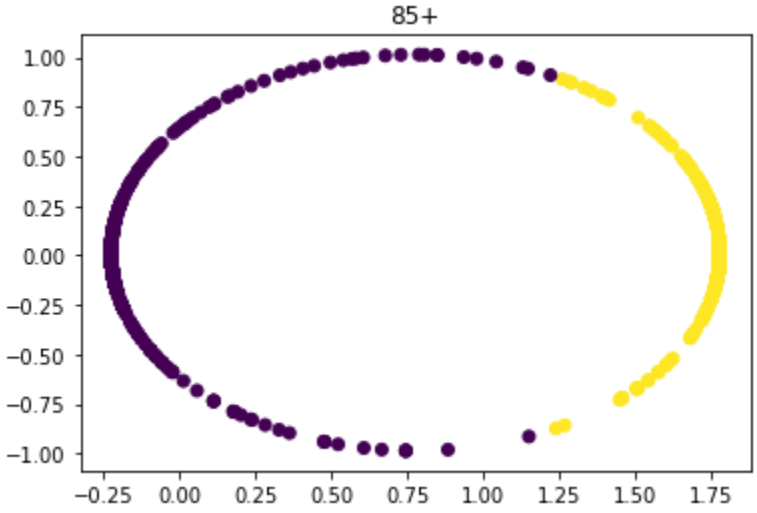

35-44
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75


45-54
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75


55-64
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75


65-74
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75


75-84
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75


85+
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75

Age Group, Scores:

Age Group	**Scores**
0-24	96.49%
25-34	91.01%
35-44	82.5%
45-54	81.63%
55-64	80.02%
65-74	80.2%
75-84	79.55%
85+	80.74%

III. SUPPORT VECTOR MACHINES

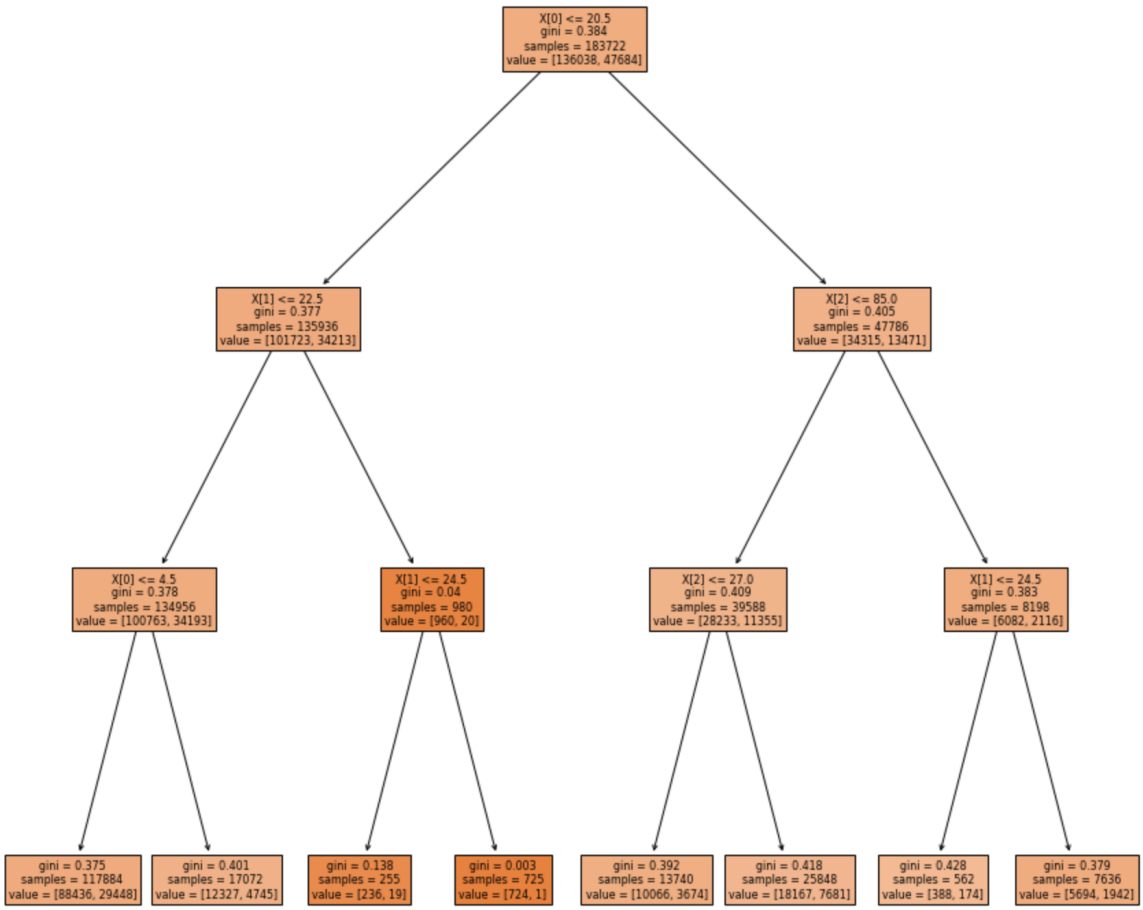
The Support Vector Algorithm was performed in a similar fashion to Logistic Regression. Below is the table of results for the SVC models for each condition group:

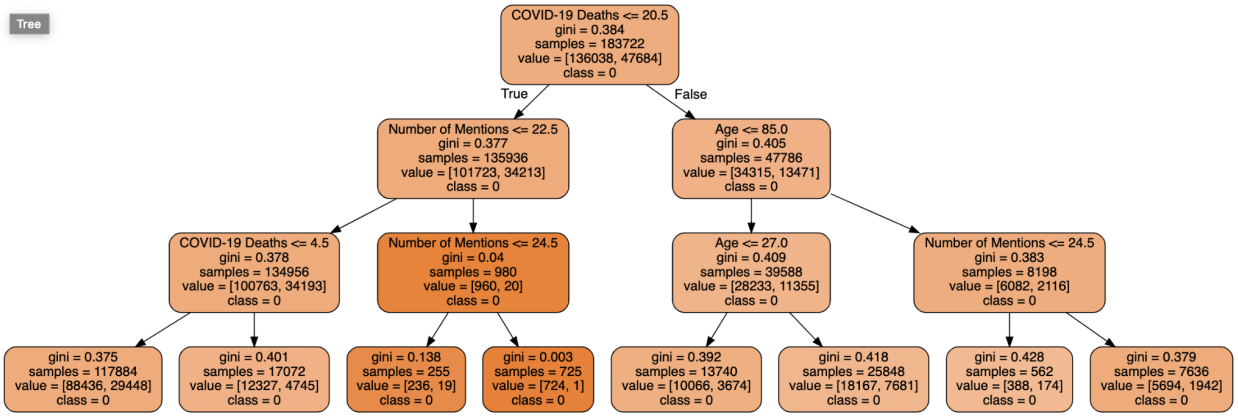
Condition Group	**Accuracy Score**	**Condition Group**	**Accuracy Score**
Respiratory	73.65%	Alzheimer Disease	95.55%
Circulatory	69.97%	Vascular/ Unspecified Dementia	95.30%
Sepsis	95.60%	Renal Failure	95.70%
Malignant Neoplasms	95.96%	Intentional / Unintentional Injury/ Poisoning	96.46%
Diabetes	95.37%	All other conditions	95.35%
Obesity	95.96%	COVID-19	94.93%

Similarly, below is the table of results for the SVC models for each age group:

Age Group	Accuracy Score	Age Group	Accuracy Score
0-24	89.31%	55-64	86.26%
25-34	91.02%	65-74	84.65%
35-44	90.66%	75-84	84.91%
45-54	88.24%	85+	84.94%

IV. DECISION TREE





DISCUSSION OF RESULTS

I. LOGISTIC REGRESSION

Logistic regression performed well for the condition groups, with the exception of respiratory and circulatory conditions. This could have happened for a number of reasons-- one such reason is that circulatory and respiratory conditions present very similar risks when an individual with one of these conditions contracts COVID-19. As a result, the number of people in each of these two groups who died after contracting COVID-19 were very similar, and so the logistic regression model had trouble predicting which condition group the death toll belonged to. This problem could potentially be solved by parameter tuning or more data for each group. Logistic regression performed slightly less well for the age groups. This problem could also potentially be solved by parameter tuning, but it's worth mentioning that the age of an individual who contracts COVID-19 is a much less reliable indicator of how likely death is to occur when compared to pre-existing conditions, especially pre-existing conditions that have to do with the parts of the human system most impacted by COVID-19.

II. SPECTRAL CLUSTERING

As we can see here, Alzheimers, Dementia, Respiratory Diseases, and Circulatory Diseases have the most accurate scores. It can be also observed that the groups' cluster points that form a perfect oval are usually correlated with scores between the 2nd and 3rd quartile. COVID-19 patients from newborn to 54 years have the most accurate scores. Notice how the first two cluster plots have a nonexistent circle and have a higher score.

This indicates that having little to no clusters means that the silhouette score for the data is bound to be completely accurate, while data that will form a complete oval from the points will have a less accurate silhouette score. From this model, we can conclude that old patients, injured patients, COVID-19, etc. have led to inaccurate model rates.

The results from this model alone were more inaccurate from the other models, which means that Spectral Clustering is a model/method that should be avoided with this dataset.

III. SUPPORT VECTOR MACHINES

The Support Vector Machine algorithm yielded the same conclusion as Logistic Regression. Each Support Vector Machine model had a high accuracy score with the only exception being the respiratory and circulatory condition, scoring around 74% and 70% for accuracy, respectively. The remaining condition group and age group model performed really well, scoring between 85%-96% for accuracy. As explained above, the reason for this is most likely because respiratory and circulatory diseases have a lot of overlapping symptoms with COVID-19. In fact, COVID-19 is categorized as a respiratory illness. The condition model that performed the best is the Intentional / Unintentional Injury/ Poisoning model and this is most likely because this condition is very different from the rest of the conditions listed and would be hard to misclassify. Overall, the support vector machine algorithm was an efficient model to use as it performed pretty well across the board.

IV. DECISION TREE

The decision tree algorithm was difficult to implement due to the type of the data set chosen. Most of the variables were non binary categorical, so to implement the given algorithm numerical variables such as “Number of Death”, “Covid Mentions”, and “Age” were used as X’s, in addition one binary - “Respiratory” was used with 1 being patients with some kind of respiratory illness and 0 being patients without any respiratory illnesses. The idea was to run the algorithm based on numerical inputs and let it classify people with respiratory diseases and without.

CONCLUSION

Support vector machines and logistic regression performed quite similarly on this dataset-- they both returned high classification rates for the condition groups, with the exception of respiratory and circulatory conditions; and returned high classification rates for the age groups as well. From the analysis so far, it seems that support vector machines and logistic regression could be used interchangeably for the same results on this specific dataset. Spectral Clustering churned out poor accuracy rates for both the condition groups and the age groups, and Decision Tree produced widespread results, where the lowest accuracy score was at the 60s and the highest accuracy score was at the 90s. The spectral clustering visualization plots showed mostly a few points that formed an oval, and some points that were scattered. The decision tree also showed that the amount of mentions had the most importance.

RESOURCES

Dataset

https://data.cdc.gov/NCHS/Conditions-Contributing-to-COVID-19-Deaths-by-Stat/hk9y-quqm?fbclid=IwAR2G1mgWw6GwU5Uw14duM2i6rT40gf6H9z8twuPFgyG35RofGfhkoU5pdkM