COVID-19 Mortality Prediction

Aria Kieras
Dept. of Computer Science
University of Massachusetts
Lowell, MA 01854, USA
aria_keiras@student.uml.edu

Akshita Pachauri
Dept. of Computer Science
University of Massachusetts
Lowell, MA 01854, USA
akshita_pachauri@student.uml.edu

Bibi Hajira M
Dept. of Computer Science
University of Massachusetts
Lowell, MA 01854, USA
bibihajira_mahammada@student.uml.edu

Shamsun Nahar Edib
Dept. of Electrical and Computer Engineering
University of Massachusetts
Lowell, MA 01854, USA
shamsunnahar_edib@student.uml.edu

Abstract—**The coronavirus disease 2019 (COVID- 19), caused by the SARS-Cov2 virus, has a huge impact on the mortality rate as well as different clinical manifestations such as, lungs diseases, cardiovascular diseases, renal diseases, etc. Predicting the morality risks at the early stage of a pandemic can help the policy maker to decide the efficient resource allocation and treatment planning which may result in a reduction in the number of deaths. The aim of this paper is to predict the mortality risk of a county due to COVID- 19 based on four supervised classification techniques and compare the efficacy of each model with one another. After selecting the most important features based on factors that have been confirmed by previous studies to have impact on COVID- 19 mortality, the data for 29 features out of 227 features were extracted and aggregated county wise for the counties in the U.S. Neural Network (NN), Support Vector Machine (SVM), Random Forest, and Decision Tree classification based models considering the 29 features were used as baseline models to predict the mortality rate. After obtaining the feature importance graph from the baseline models, the four models were trained again based on the top 20 features to obtain the final model. The efficacy of the models was evaluated based on f1-score and accuracy. Based on the accuracy, the Random Forest (91.71%) was the best model in predicting the mortality risk followed by SVM, NN, and Decision Tree with 90.92%, 88.85%, and 88.69% accuracy, respectively. By performing principle component analysis (PCA) on the Random Forest Model, it was possible to obtain 6 features which have the most impact on the prediction of mortality rate with a cost of approximately 0.09% less accuracy.**

Keywords—COVID- 19, mortality rate, death, Neural Network, Support Vector Machine, Random Forest, Decision Tree.

I. Introduction

The COVID- 19 epidemic was at first thought of as nothing more than a news story, but then quickly became an everyday reality. With the disease spreading ever more rapidly throughout the United States and the world, with little information as to what causes its effects to worsen from case to case, and the infection rate and death rate only going up, work and research to understand what factors cause the SARS-Cov2 virus to have a more severe effect on someone to lead to death became the main focus for many people. There has been a large amount of research conducted into discovering what factors will cause a more severe case of COVID- 19, and most resolve the following as factors of a severe case; obesity, diabetes, smoking, and old age [1]. From these features, we can gather that in general, persons with already present health issues such as the aforementioned, or similar health problems, will have a higher likelihood of a severe case of COVID- 19. However, a severe case of the disease does not mean absolute death. While these factors have been proven to cause an increase in the probability of severity, they have not quite been proven to increase the chance of death. Other research has been conducted in order to determine what factors lead to a higher mortality rate of COVID- 19, and it was shown that while old age did have a large impact on mortality, obesity, diabetes, or other underlying health conditions did not have much of an impact on the mortality rate as they did for the severity of the infection [1]. What was found was that factors such as overcrowding, being uninsured, and being below the poverty line had more of an impact on determining the mortality rate. In this paper, we will be looking at such features, including the percentage of people at or above the age of 65, overcrowding, population density, ethnicity, etc. With this data we will look to determine if there is any relation between these features and the mortality rate of COVID- 19.

II. Background Work

A. Background Work

In the past two years, the COVID- 19 disease has become the greatest health concern for worldwide nations as it has a huge impact on the mortality and morbidity throughout the world population. COVID- 19 mainly has negative impacts on the lungs along with proof of causing acute respiratory distress syndrome (ARDS) [2]. Nonetheless, it is reported to also have effect on patient's cardiovascular, neurological, vascular, and renal systems [2].

Predicting the mortality of critically ill and hospitalized COVID- 19 patients is important to provide the necessary clinical help to the most effected areas. In order to facilitate the clinical decision making and resource allocation, an in-hospital mortality prediction model based on modification of partial least square (SIMPLS) based method is proposed in [3]. Additionally, to classify the hospitalized patients based on high- and low-risk, a Latent class analysis (LCA) is performed for clustering. However, the model was built and tested on a single-center small scale dataset without validating the performance on a larger set and multi-center settings.

A mortality rate prediction model based on Decision Tree, Multi-layer Perceptron (MLP), k Nearest Neighbor (kNN), Random Forest, and Support Vector Machine (SVM) data mining techniques are presented in [4], which showed that dyspnea was the most effective factor for predicting the death of COVID- 19 patients.

Ref. [5] predicted the mortality risk in patients with COVID-19 based on SVM, Neural Network (NN), Random Forest, Decision Tree, Logistic Regression, and kNN based on symptoms, pre-existing conditions, and demographics. The results show that NN using 10-fold cross-validation was able to predict the mortality rate with an 89.98% accuracy followed by kNN and SVM with 89.83% and 89.02% accuracy, respectively.

Although, predicting mortality for hospitalized patients is important, identifying the geographic region where people are at high risk of mortality at an early stage of a pandemic is also an important issue. Ref [6] studied mortality risk considering sociodemographic and socio-environmental factors at state- and county-level across the U.S. It is seen that the areas with high population and high pollution, areas containing air hub and race minorities (non-white population) are at a higher risk of COVID- 19 related deaths at the first stage of the pandemic.

County-level mortality counts due to COVID- 19 is used to find the factors that have the most impact on the mortality rate using clustering analysis and Kruskal-Wallis tests [7]. Poor health status is founded as the factor mostly affecting the COVID- 19 mortality rate along with sex, race/ethnicity and outdoor environment.

B. *Problem Statement*

As with any new disease that is spreading, the main problem is in understanding what is causing the disease to spread, affect people more severely, and lead to death. In understanding the causes of these problems, the spread and mortality rate of COVID- 19, and any disease, can be mitigated. In order to determine the features that do have an impact on the mortality rate at the county level, four machine learning models have been trained against the selected dataset in order to determine separately what the most important features were, and to then be compared to one another to see if all models came to the same conclusion. From this, the accuracy of each model is then compared to check if one model is better suited to a problem such as this.

III. METHODOLOGY

In this project, we took the approach of implementing four different models in order to determine if one model would have a better accuracy rating for the selected features in order to determine the mortality rate of persons infected with COVID-19. These four models are; a NN, a SVM, a Decision Tree, and a Random Forest. By taking the approach of using these different models against the data, we can see if one model is better suited to discovering relations between the features and the mortality rate of COVID- 19, and we can see if the models determine the same relationships. In other words, we will be able to see if all the models come to the same conclusion, and thusly will be able to prove through multiple means that these

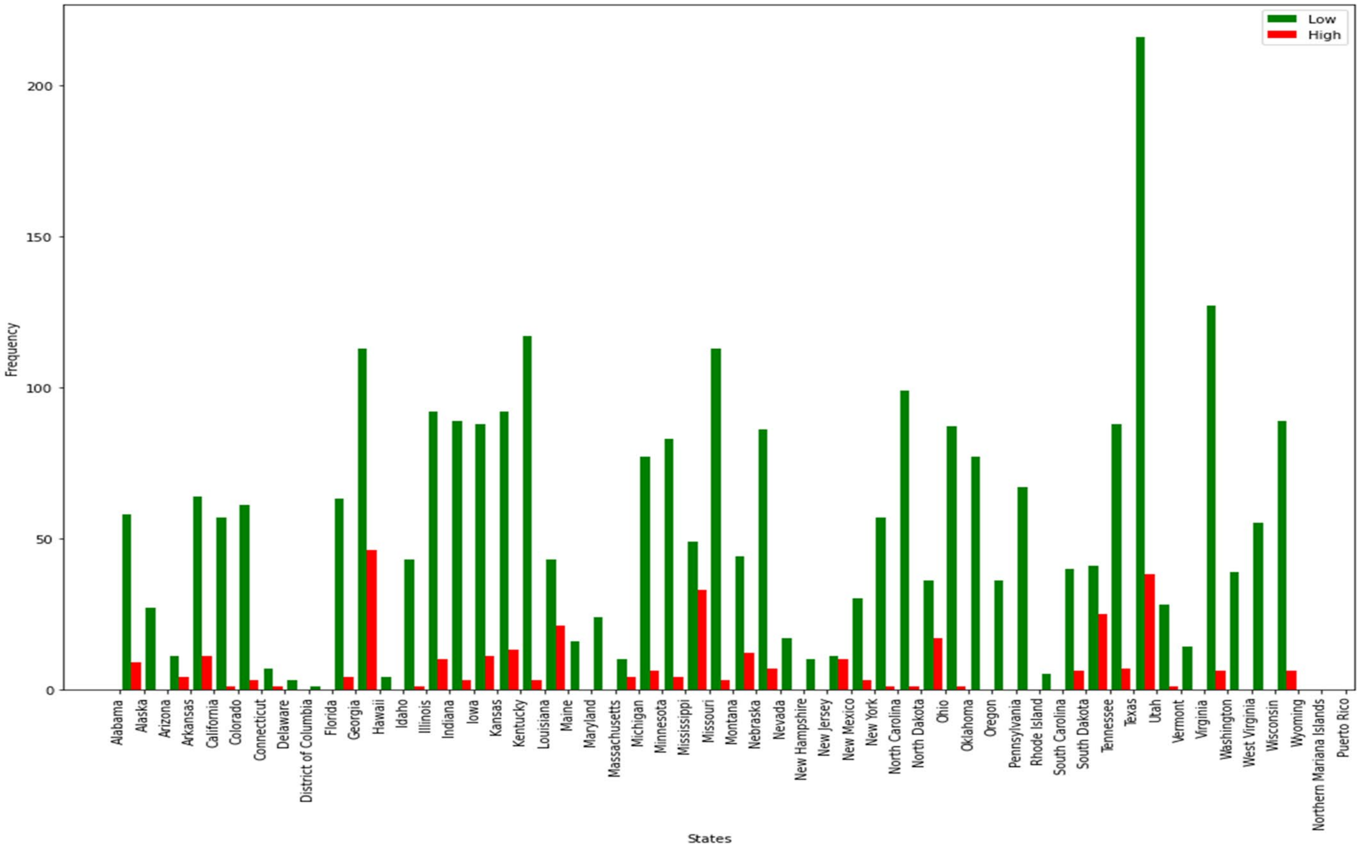

Fig. 1. A histogram representing the frequency of counties in a state that are at low and high risk of mortality due to COVID- 19.

features do in fact cause a higher mortality rate. All the models were set up as classification models, implemented to classify from the features selected into one of two categories, if a county has a high mortality rate or if a county has a low mortality rate from COVID- 19.

A. Data

The primary data that was used for this project was enhanced by adding features to the original data that contained three features, Age Groups, Ethnicity and Sex considered at the state level. After completing the data pre-processing on the initial data and running that through the different models, it became apparent that more features were needed, as using just three features, the models were unable to classify the counties based on mortality rate. So the new challenge became finding and compiling additional features to use along with the previous features to improve county classification. After looking through data available at the United States Census Bureau, which had data relating to population totals for each county in the United States, and other sources that were able to provide the total number of deaths per county over a period of a few months, the data from these sources was compiled together along with the previous features already being used. As adding these features was improving the accuracy, more features were added, and from this data a better understanding of what features might be needed to further increase the accuracy of the models was attained, and further research was done to compile the rest of the features being used now.

The finalized dataset was built by enhancing the original dataset with additional features [8] resulting in 227 features and 790,331 rows, for 3,139 counties out of 3,143 counties in the U.S. spanning over March - April 2020. It is important to note that not all counties appear in the dataset because not all counties reported a COVID- 19 case in the specified timeframe. Exploratory data analysis was performed to explore the dataset visually and statistically, in order to describe the features. From this data, the features that were extracted were the features that were proven to impact the mortality rate of COVID- 19 [2-5]. Since the data comprised of daily entries spanning over two

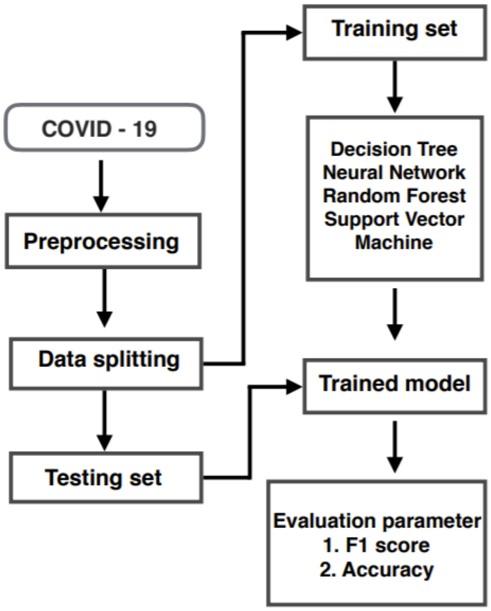

Fig. 2. Flowchart of the final model.

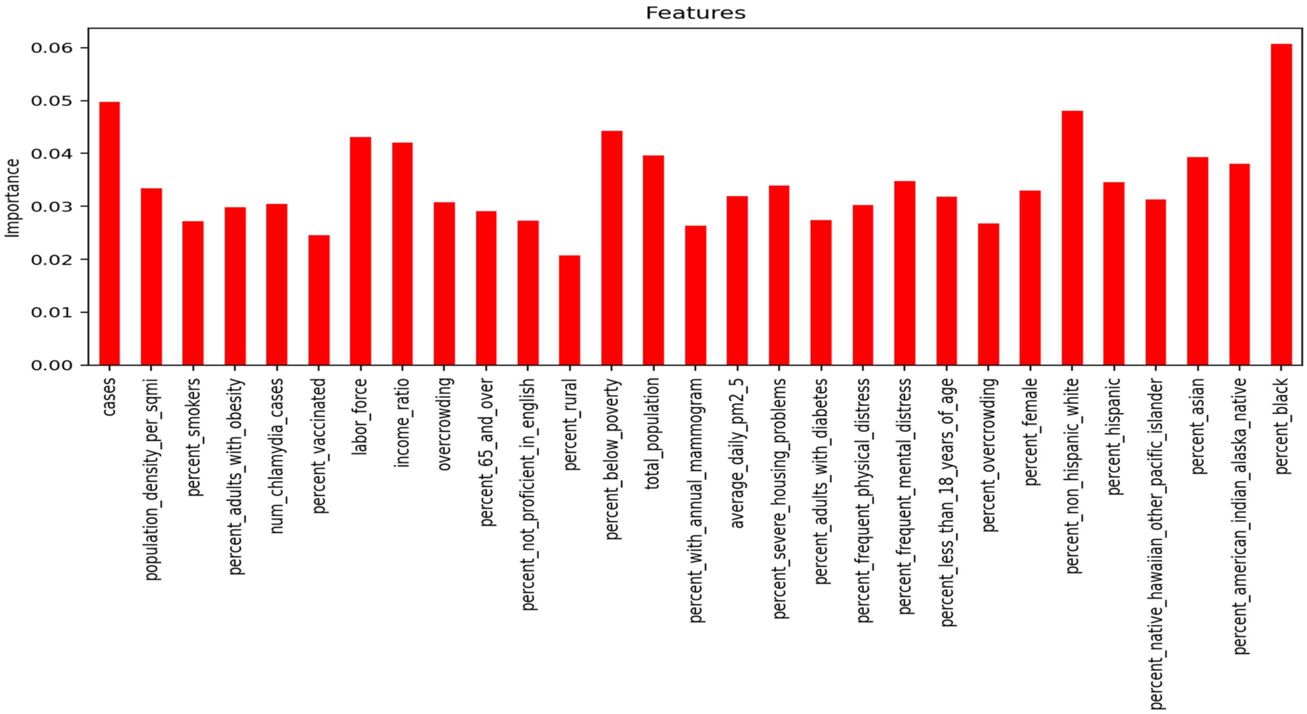

Fig. 3. Feature importance graph.

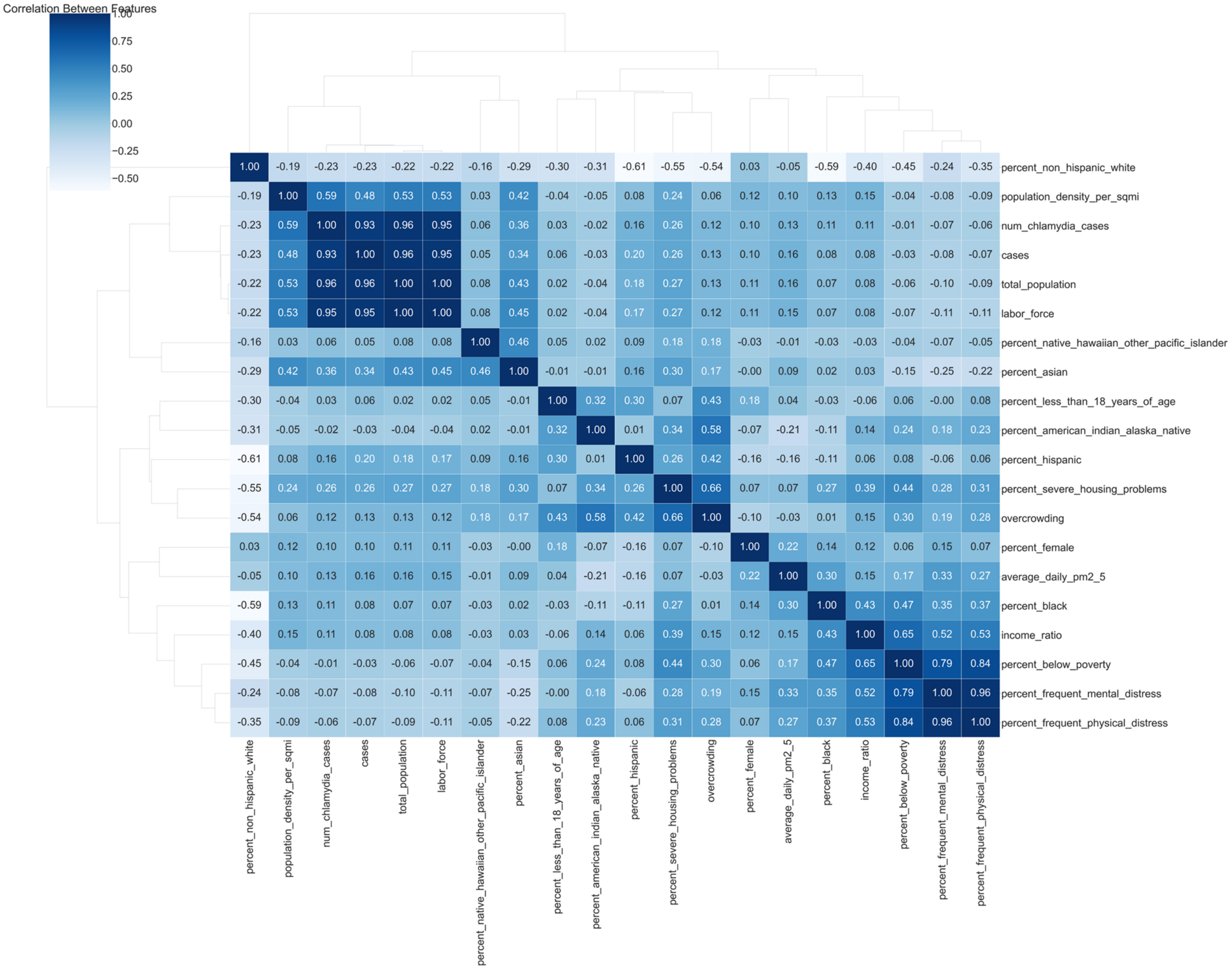

Fig. 4. Heat map of the extracted subset of 20 features obtained from feature importance graph.

months it was necessary to aggregate it to the county level for the data to be consistent with the objective of this project. The data transformation from daily to county wise data was done by grouping the data over the column 'fips', since it is the unique id for each county, performing aggregation for each column individually, and all NaN values were replaced with the average of the respective features. Finally, a feature was added, ′deaths_class′, which takes the total deaths caused by COVID-19 in a county and divides that by the total population to attain the cause specific (COVID- 19) mortality rate [9] for each county, and if the mortality rate for that county is above a set threshold percentage, that county will be classified as having high mortality, otherwise the county will be classified as having low mortality. From the dataset, it was seen that the highest mortality rate of a county is 0.75%. For choosing the threshold value for creating two classes, 0.17% mortality rate was considered as the threshold value as it is around 20% of the maximum mortality rate. As it can be seen in Fig. 1 which depicts the green bar as the number of counties in a particular state with a mortality rate below the threshold value, while the red bar depicts the number of counties in that state with a mortality rate higher than the threshold value.

TABLE I
EXTRACTED SUBSET OF 20 FEATURES OBTAINED FROM FEATURE IMPORTANCE GRAPH FOR PREDICTING MORTALITY DUE TO COVID-19

Feature Name	Feature Name
Cases	Percentage frequent physical distress
Population density per sqmi	Percentage frequent mental distress
Number of chlamydia cases	Percentage <18 years old
Labor force	Percentage female
Income ratio	Percentage non-Hispanic white
Overcrowding	Percentage Hispanic
Percentage below poverty level	Percent native Hawaiian and other pacific islander
Total population	Percentage Asian
Average daily pm2_5	Percentage American Indian, Alaska native
Percentage severe housing problems	Percentage Black

B. Models

As is previously mentioned, four models were implemented to run against the dataset in order to classify the U.S. counties into either high or low mortality rate. After selecting the feature sub-set, the models used for classification are; a NN, a SVM, and Random Forest, and a Decision Tree. These models were selected as Random Forest being an ensemble takes the advantage of the algorithmic crowd, Decision Trees algorithms are efficient for large datasets, NNs are well suited for finding patterns in data over large datasets, and SVMs are effective with high dimensional data. The train-test split for all the models were chosen to be 80% and 20%, respectively. Parameter turning, grid search, and feature engineering were used to improve the accuracy of the models. The flow how all the models were trained and tested can be visualized from Fig. 2.

IV. Experiments

The main experimentation that was carried out was with the data being used, as was previously mentioned in the section III the dataset went through many alterations in order to compile the best feature set to use for the different models. After the dataset had been compiled however, there was the question of how to calculate the mortality rate, as there are multiple ways to do so. The case mortality rate for COVID-19 can be used to calculate the proportion of individuals who die due to COVID-19 against the number of positive cases, and it can be defined as below [9]:

$$Case\ Mortality\ Rate = \frac{Total\ Number\ of\ COVID{-}19\ Deaths}{Total\ Number\ of\ COVID{-}19\ Cases} \qquad (1)$$

On the other hand, the cause specific mortality rate due to COVID- 19 shows the proportion of individuals that died from COVID- 19 against the total population size, and it can be represented as follows :

$$Cause\ Specific\ Mortality\ Rate = \frac{Total\ Number\ of\ COVID{-}19\ Deaths}{Total\ Population} \ (2)$$

When considering which way of calculating the mortality rate there are reasons to use either option, however, when considering the main goal of the problem statement, using the cause specific mortality rate calculation was the better choice.

A. Feature Engineering and Parameter Tuning

In order to reduce and extract the most important features, feature engineering was conducted on the dataset containing 227 features. Initially, 29 features were pulled from this dataset that were shown from other research to impact the mortality rate of COVID- 19 [2, 4, 10]. After running this sub-set of 29 features, further feature selection was conducted after compiling a feature importance graph as shown in Fig. 3 to create a new sub-set of 20 features, using a threshold of 0.03 to pull out the most important features. The extracted 20 features are listed in Table I, and the heat map for these features which displays the correlation among different features in this grouping is shown in Fig. 4. Again, this subset listed in Table I was run against the models and feature importance was again determined and resulted in the last sub-set of 6 most important features as depicted in Fig. 5. Feature scaling was performed by normalizing the features on the scale of 0 to 1, which is clearly evident from the boxplot showing in Fig. 5. For parameter tuning 3-fold cross validation with randomized search, as well as 3-fold cross validation with grid search was used.

TABLE II
Comparison of performance of the four models when the number of features is 29 and 20

No. of feat.	Metrics	NN	SVM	Random Forest	Decision Tree
29	f1-score	0.9313	0.9461	0.9467	0.9238
	Accuracy	87.26%	89.81%	89.97%	86.31%
20	f1-score	0.9396	0.9525	0.9487	0.8843
	Accuracy	88.85%	90.92%	91.71%	88.69%

TABLE III
CONFUSION MATRIX OF NN AND SVM

		NN		SVM	
		Predicted		Predicted	
		Positive	Negative	Positive	Negative
Actual	Positive	546	26	571	1
	Negative	44	12	56	0

TABLE IV
CONFUSION MATRIX OF RANDOM FOREST AND DECISION TREE

		Random Forest		Decision Tree	
		Predicted		Predicted	
		Positive	Negative	Positive	Negative
Actual	Positive	565	7	493	79
	Negative	54	2	50	6

B. Results

For evaluating the performance of the four models, f1-score and accuracy metrics were used. The f1-score shows how precise the classifier is, and can be defined as below [11]:

$$F_1 = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \qquad (3)$$

where TP, FP, and FN represent true positive, false positive, and false negative. Accuracy takes the properly classified cases against the total cases, and it is defined as [11]:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

where TN represents true negative. Therefore, the larger value of these metrics means that the model was able to predict the mortality risk of the counties more accurately. The comparison of performance of the four models based on f1-score and accuracy is presented in Table II considering the number of features as 29 and 20. From the table, it is clearly visible that the performance improves when the number of features is reduced to 20 for all the models. This may be due in part to having an excess of features that do not directly have any relation to the mortality rate, and therefore, removing the unneeded features positively impacts the accuracy. Comparing the metrics when the number of features is 20, it can be seen that the Random Forest is performing the best among all the models with an accuracy of 91.71%. The SVM model is performing the second best with an accuracy of 90.92% followed by NN with an accuracy of 88.85%. It is found that the Decision Tree model performed the weakest with an accuracy of 88.69%.

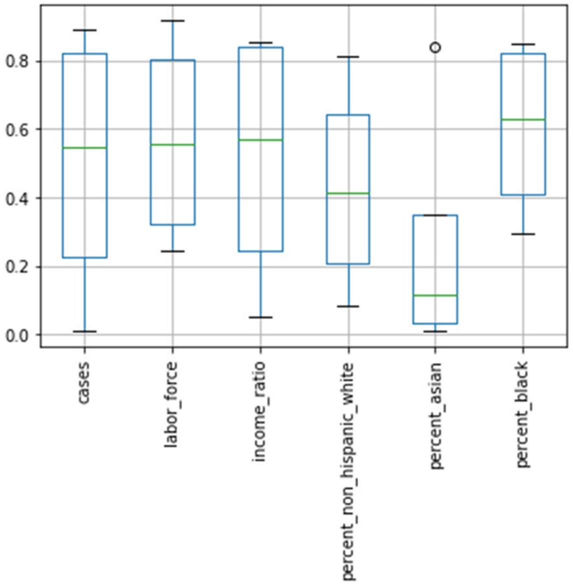

Fig. 5. Boxplot for 6 features that has most significant impact on the mortality rate of COVID-19.

The confusion matrices for the NN and the SVM are listed in Table III, and for the Random Forest and the Decision Tree are listed in Table IV. The true positive values of the NN, the SVM, and the Random Forest are better than that of the Decision Tree, which correlates with the information shown in Table II.

After testing the models on the 20 features sub-set, principle component analysis was performed to further reduce the size of the feature subset and extract the most informative features that can explain the mortality rate of COVID– 19 at the county level. By training the Random Forest model with the extracted subset containing 6 features as can be referenced in Fig. 5, it was seen that although the accuracy reduced slightly, these 6 features were clearly the most important factors affecting the mortality rate of COVID– 19 at the county level.

C. Application

As this project focused on the early months of the COVID-19 epidemic, March and April 2020, to determine what contributing factors at the county level, i.e., old age, obesity, overcrowding etc., may have a real impact on if a county is more at risk for a higher mortality rate for COVID- 19. And so with an understanding of what factors do impact a disease at the beginning stages, the work done for this project can be applied to any other new disease that may develop in the U.S. to help identify which counties may need more assistance and resources in order to better combat the spreading disease and to hopefully, reduce their mortality rate.

IV. Conclusion

This project was intended to implement four different machine learning algorithms to determine the effectiveness of the models in predicting the likelihood of a high risk of mortality per county. Following the algorithm implementations, each resulting model was compared against one another to determine the most effective model. The results from the implemented algorithms provided the information on how different factors such as; ethnicity, cases, population, age, underlying health conditions, overcrowding, air quality, income, and occupation largely affect the COVID- 19 mortality rate. Moreover, the most significant factors were: ethnicity, occupation, income, and cases. The experiments run against the four models for each sub-set of data used consistently proved that the Random Forest model to be the best model for this particular problem and dataset. However, the three models performed adequately having only at most 3% difference in accuracy.

Future work on this project would include using data from other widespread diseases across the U.S. for the same features previously determined in this paper, and determine if these features do hold true in predicting high mortality rates at the county level for diseases. Furthermore, this project can be expanded to take into account worldwide data. By knowing this, and taking into account any further features from other diseases, this project can assist in future to identify the most at risk counties in the U.S. as well as other geographical locations in the world so that resources and aid can be supplied to those areas to help reduce the mortality rate.

References

[1] S. Lam *et al.*, "Social determinates of health and COVID-19 mortality rates at the county level," *2020 Fourth International Conference on Multimedia Computing, Networking and Applications (MCNA)*, Oct. 2020, pp. 159-165.

[2] W.-j. Guan *et al.*, "Clinical characteristics of coronavirus disease 2019 in China," *New England Journal of Medicine,* vol. 382, no. 18, pp. 1708-1720, Apr. 2020.

[3] M. Banoei, R. Dinparastisaleh, A. Vaeli Zadeh, and M. Mirsaeidi, "Machine-learning-based COVID-19 mortality prediction model and identification of patients at low and high risk of dying," *Critical Care,* vol. 25, 2021.

[4] K. Moulaei, F. Ghasemian, K. Bahaadinbeigy, R. Ershad Sarbi, and Z. Mohamadi Taghiabad, "Predicting mortality of COVID-19 patients based on data mining techniques," (in eng), *J Biomed Phys Eng,* vol. 11, no. 5, pp. 653-662, Oct. 2021.

[5] M. Pourhomayoun and M. Shakibi, "Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making," *Smart Health,* vol. 20, p. 100178, Apr. 2021.

[6] E. Correa-Agudelo, T. B. Mersha, A. J. Branscum, N. J. MacKinnon, and D. F. Cuadros, "Identification of vulnerable populations and areas at higher risk of COVID-19-related mortality during the early stage of the epidemic in the United States," (in eng), *Int J Environ Res Public Health,* vol. 18, no. 8, p. 4021, 2021.

[7] T. Tian *et al.*, "Risk factors associated with mortality of COVID-19 in 3125 counties of the United States," *Infectious Diseases of Poverty,* vol. 10, Dec. 2021.

[8] kaggle.com [Online]. Available : https://www.kaggle.com/johnjdavisiv/us-counties-weather-health-hospitals-covid19-data/data?select=US_counties_COVID19_health_weather_data.csv.

[9] World Health Organization [Online]. Available: https://www.who.int/news-room/commentaries/detail/estimating-mortality-from-covid-19#:~:text=Calculating%20CFR%20Case%20fatality%20ratio,severity%20among%20detected%20cases%3A.

[10] USA FACTS [Online]. Available : https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/.

[11] Towards Data Science [Online]. Available : https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234.

APPENDIX

A. Work Distribution

Name	Task
Aria	Implemented NN, lead report writing
Shamsun	Implemented SVM, created class split for covid deaths, worked on report.
Akshita	Implemented Random Forest, finalized all models, performed data transformation, parameter tuning.
Hajira	Implemented Decision Tree, implemented plots for data visualization.
Group	- Presentation.