COVID-19 Risk Model on Macroscopic Data in the European Population

CS6140 Project Group 08

Introduction

The prediction of Covid-19 cases, severity, and outcomes has been a topic of interest since the initial outbreak of the disease in late 2019 and early 2020. In just three years, numerous research groups have published novel machine learning models to predict case severity and exposure likelihood based on various health factors, such as presented symptoms and vital measurements [1,2]. This data has been made publicly available through public health databases [1,2,3], making the development of new models and improvement of existing ones highly feasible. While some of these models achieve relatively high accuracy in predicting Covid-19 severity and exposure, many heavily rely on data acquired in a hospital setting, which may disproportionately enable their application to areas with high access to hospital consultations and treatments. Additionally, patient data may only become widely available once disease has already been spread, so these models may not be useful at the early and preventative stages of an outbreak.

Here, we examine Covid-19 data from a different angle to predict region-specific mortality rate based on various non-health parameters, such as population density, age groups, land environmental variables. We want to determine which variables are likely predictors of Covid-19 outcomes. Through this model, we will enable mortality rate predictions for future outcomes, without the need for hospital visit data, to inform early public policy interventions.

Data

Dataset:

We turn to Mendeley Data for a csv dataset that compiles regional socioeconomic and environmental demographics for towns in the European Union [3]. The dataset consists of 1471 rows where each row contains data for an individual town in Europe. Each town is described by a unique town code (column label CODE_NUTS3). The continuous features in the raw dataset include population demographic variables, economic variables, comorbidities, health outcomes, economic variables, and environmental variables. The only categorical variable is lockdown severity, designated by a value 0 through 3 for different levels of restrictions.

Preprocessing:

For our models, we focus on population demographics and environmental factors to isolate non-health variables that may contribute to Covid-19 outcomes. We created our dataset by uploading the raw data file into a Pandas dataframe and dropping unwanted columns (i.e. unused features). We drop 12 rows for which NO2 data was empty. We did not include female population proportion as it is relatively constant around 50% for all regions. We were interested in including unemployment rate, but a significant portion of the dataset did not have this data available, so we excluded it as well. We explore different models that are trained with different features.

Training and test split:

To ensure consistent comparison between models, we first split the entire dataset into training and test sets using sklearn, reserving 33% of the data for testing. Then, we create 'population' and 'population + environment' training and test subsets from the split sets which are used for the different models. This allows us to compare models trained with the same data points, but with different features.

Implementation

We generate two models. One model uses a dataset that contains data for population demographics and environmental factors. The other model uses data for population demographics only. This is motivated by the question of whether adding environmental data to population data builds a stronger predictive model. We use the model outputs to predict possible determinants of COVID-19 infection and outcomes.

We explore the following implementations for each model: SVM, Perceptron, Linear Regression, and Ridge Regression. All are implemented using sklearn.

Implementations:

SVM: We implemented an SVM model with Support Vector Regression (SVR) with a non-linear kernel. SVM applies kernel tricks to determine boundaries between classes, or, in our case of regression with continuous variables, between continuous outputs. We applied a grid search cross validation to determine the optimal combination of c values and gamma values for the model.

Perceptron*:* We chose a Multi-layer perceptron (MLP) regressor to implement a perceptron model. We chose MLP since it is commonly used to approximate continuous and non-linearly separable outputs. MLP was optimized using grid search to determine the optimal hidden layer size.

Linear Regression: We performed a linear regression task using the LinearRegression class function from sklearn. We did not expect accurate predictions with linear regression, but included it as a comparison to the MLP and SVM implementations as a regression control.

Ridge Regression: As a continuation to Linear Regression, we performed Ridge Regression with built-in cross validation, which uses a leave-one-out CV approach. We selected 3 values of alpha for the model to try and the optimal model was used to fit the training data.

Random Forest: Our Random Forest Regression implementation was the only model that utilized ensemble learning. It combines multiple algorithms to improve predictions. Its use in predicting Covid-19 outcomes and cases has been previously demonstrated in literature [4,5]. It was recently utilized to determine important clinical features for predicting Covid-19 mortality rates [6], which motivated our utilization of it in this project.

Results & Discussion

To analyze and compare the predictive power for each implementation, we calculated its coefficient of determination. This measurement informs us how good the fit of the model is, with 1.0 meaning a perfect fit. Models with coefficient of determination values closer to 1.0 perform better at predicting mortality rate and are thus preferred in this application.

For MLP, SVM, and Random Forest, we summarize in a heat map the resulting coefficient of determination from a grid search using *training data* to demonstrate how model parameters were selected (Figures 1 and 2).

For all implementations, we report the final coefficient of determination resulting from test data (Table 1). These values result from predicting mortality rates for *test data* using the best estimator of each implementation.

Implementation	Population Coefficient of Determination	Population + Environment Coefficient of Determination
Linear Regression	-4.058	0.207
Ridge Regression	0.049	0.207
MLP	-0.429	-0.0001
SVM	0.058	0.061
Random Forest	0.113	0.281

Table 1: Summary of results from each model with four implementations: SVM, MLP, Linear Regression, Ridge Regression, and Random Forest. Test data is used here.

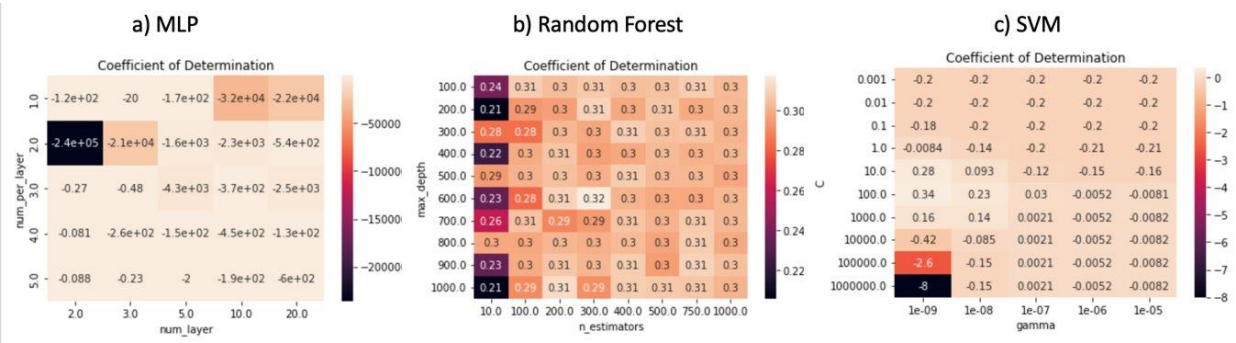

Figure 1: Coefficient of determination heatmap for MLP, Random Forest, and SVM implementations trained and tested with population data. Training data is used here.

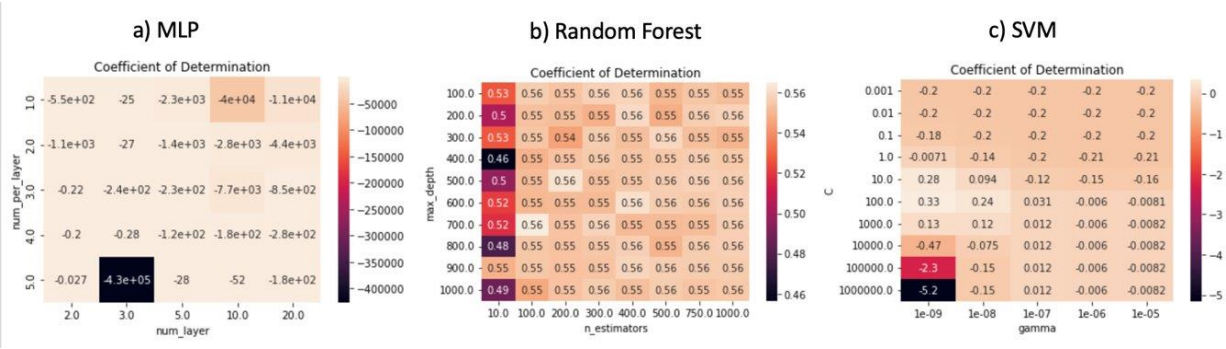

Figure 2: Coefficient of determination heatmap for MLP, Random Forest, and SVM implementations trained and tested with population and environmental data. Training data is used here.

Linear Regression & Ridge Regression:

In the cases of Linear Regression and Ridge Regression implementations, Model 1 (population data only) performed significantly worse than Model 2 (population and environment data combined) at predicting mortality rate for test data. This suggests that for these regression models, adding more features may make predictions more accurate; that is, training the models with population data alone does not lead to accurate predictions.

MLP & SVM:

For the MLP models, Model 1 and Model 2 both performed poorly, yielding coefficients of determination less than zero for test data predictions. As the MLP training grid search heat maps suggest (Figure 1a, Figure 2a), all parameter combinations in MLP models lead to low fit accuracy of training data. Both SVM grid search heat maps show poor fits for all parameter combinations in training as well, but still better than MLP. The test coefficient of determination for SVM is higher than MLP in both cases as well, but still insufficient to provide reliable prediction. Therefore, the data suggest that MLP and SVM are not appropriate with our datasets.

Random Forest Regression:

The Random Forest Regression model resulted in the highest coefficient of determination among all implementations in case of population only data as well as population and environmental data. From the grid search heatmap, one observes that aside from models with 10 estimators, the remaining parameter combinations perform rather well on training data compared to the other regression models used. The coefficient of determination more than doubled after we added the environmental data, which informs us that the environmental data may hold key determinants of Covid-19 outcomes and that Random Forest Regression is an appropriate model choice for our data.

Conclusion

We trained and tested 5 regression models to predict Covid-19 mortality rates based on population data with or without environmental data. The implementations are Linear Regression, Ridge Regression, Multilayer Perceptron, SVM, and Random Forest Regression. Of these models, Random Forest Regression trained with population and environmental data resulted in the best coefficient of determination for testing data with a value of 0.281. This translates to the highest accuracy in prediction.

Our suggestion moving forward would be to implement Random Forest Regression with more hyperparameter finetuning and with more data. We could also explore other target outcomes, such as infection rate instead of mortality rate. Finally, we would be interested in applying these models to other diseases, such as influenza, to see if environmental results similarly improve prediction of mortality rates.

Data Availability

All data and code are available at https://github.com/willchoprojects/cs-6140-final-project

References

[1] Gao Y, Cai GY, Fang W, et al. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. Nat Commun. 2020;11(1):5033. Published 2020 Oct 6. doi:10.1038/s41467-020-18684-2

[2] Guhathakurata S, Kundu S, Chakraborty A, Banerjee JS. A novel approach to predict COVID-19 using support vector machine. Data Science for COVID-19. 2021:351–64. doi: 10.1016/B978-0-12-824536-1.00014-9. Epub 2021 May 21. PMCID: PMC8137961.

[3] Omrani H, Modroiu M, Lenzi J, et al. COVID-19 in Europe: Dataset at a sub-national level. Data Brief. 2021;35:106939. doi:10.1016/j.dib.2021.106939

[4] Galasso J, Cao DM, Hochberg R. A random forest model for forecasting regional COVID-19 cases utilizing reproduction number estimates and demographic data. Chaos Solitons Fractals. 2022 Mar;156:111779. doi: 10.1016/j.chaos.2021.111779. Epub 2022 Jan 5. PMID: 35013654; PMCID: PMC8731233.

[5] Alali, Y., Harrou, F. & Sun, Y. A proficient approach to forecast COVID-19 spread via optimized dynamic machine learning models. *Sci Rep* **12**, 2467 (2022). https://doi.org/10.1038/s41598-022-06218-3

[6] Moslehi, S., Rabiei, N., Soltanian, A.R. *et al.* Application of machine learning models based on decision trees in classifying the factors affecting mortality of COVID-19 patients in Hamadan, Iran. *BMC Med Inform Decis Mak* **22**, 192 (2022). https://doi.org/10.1186/s12911-022-01939-x