Comparison of two selected ML models for predicting deaths due to COVID-19 outbreak.

Michał Bogacz
Warsaw University of Technology
Poland

***Abstract*— Virus COVID-19 had significant impact on the human life and economy of the whole world at the beginning of 2020. Many people died and have been infected. Based on that world deaths statistics, two Machine Learning models for predicting are compared: SVM and LSTM.**

Keywords—COVID-19, Python, SVM, LSTM, Data analysis, prediction algorithms

I. Introduction

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate[6]. SVM algorithms use a set of mathematical functions that are defined as the kernel. The function of kernel is to take data as input and transform it into the required form. Function can be different types: linear, polynomial, radial (RBF) . RBF is most used type, because it has localized and finite response along the entire x-axis.

Long Short-Term Memory network (LSTM) is used to time series forecasting. Model solve problems compromised of a single series of observations and a model is required to learn from the series of past observations to predict the next value in the sequence.

In this paper, these two models (SVM and LSTM) will be compared in predicting deaths due to COVID-19 outbreak.

II. Data Preparation

A. Clean-up

Downloaded database has plenty of information such as County, Province State, County Region, Population etc. Some of them are not needed and some of them must be checked for null values. First thing to prepare data was to delete cells with null values and columns with useless information. After cleaning, only four parameter has left: Country_Region, Date, Target and TargetValue. The name "Country_Region" was changed into shorter "Country".

B. Changing names of countries into ISO 3166 codes

ISO 3166 standard maintains codes for the representation of names of countries and their subdivisions. In this paper is used ISO 3166-1 alpha-3 set, which contains three-letter country codes which may allow a better visual association [7]. The standard compromises 249 countries. Extended package *pycountry* providing conversion functions.

Several countries (e.g. Burma) from database created problems due to name change. Most of them was from the African region which is why they were deleted because of the differences in climate and extremely different statistics which disrupt learning process of neuron network. Other valuable countries (e.g. US) had names changed to fit the ISO 3166 norm.

C. Split data

Better, easier analysis of the problem ensure separation data into two datasets: Confirmed cases and Fatalities. After that, data needed to be divided into train data and test data. Data set to train the neural network range from the first day in database. Last week from 2020-05-20 to the end was devoted to the train data. This method allow to check effectivity of the ML model.

III. Displaying Data

A. World map

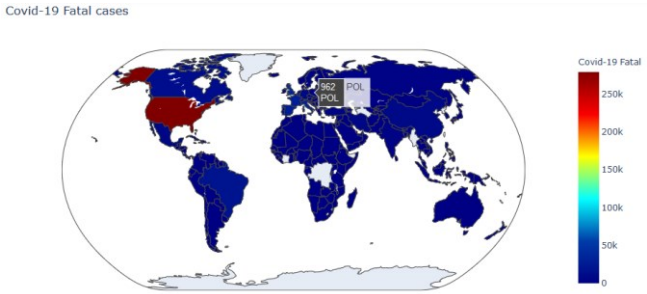

Fig. 1 World map

Showing sum of deaths and average deaths per day can help in finding countries similar to Poland.

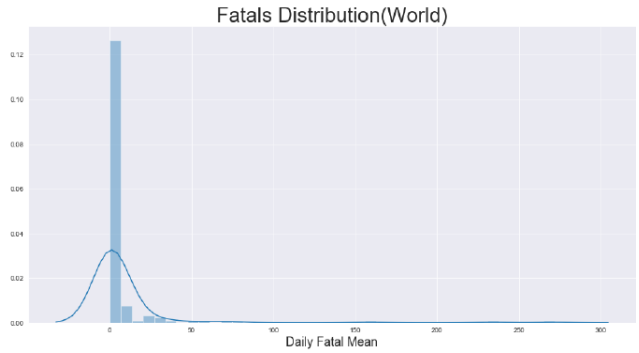

Fig. 2 Fatalities distribution in the World

B. EU map

Displaying whole world is missing the point, USA is changing the scale too much. Taking it into consideration, only European countries was considered not only due to scale problems, but also due to the political and cultural differences. Created models considering the closest countries may lead to the best results in prediction. The general approach to epidemy through all Europe is quite

similar, so teaching model on data from European Union may yield better results.

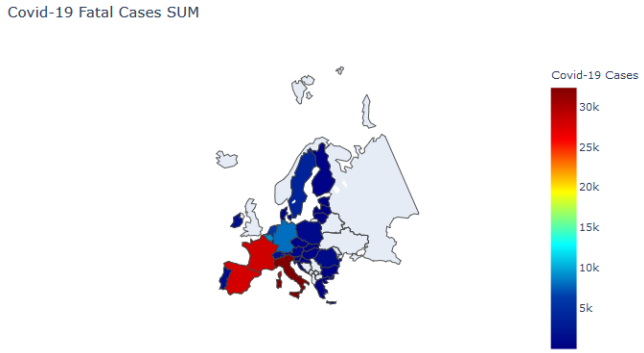

Fig. 3 Europe map

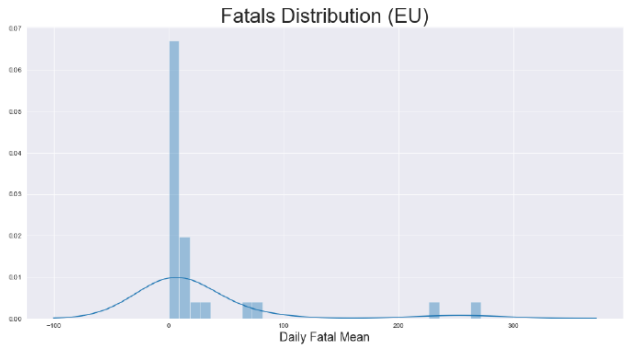

Fig. 4 Fatalities distribution in Europe

One can seem that Poland is settled inside the highest spike, on every distribution plot, so it might be possible for neural network to find some numerical similarities between Poland and other countries, so definitely the correct approach is to train network on all possible data and later propagate this onto Polish case. The algorithm should be able to differentiate between countries similar or not.

IV. SOLUTION

Two selected machine learning models: SVM and LSTM will be tested in a various approaches. Firstly, two models will be tested only on the data from Poland. Secondly, model with better precision will be used on all available data on four different sets: Polish, Similar_PL (countries politically and culturally similar to Poland), EU (all countries from Europe) and data from all over the world.

A. Polish way

The way our country approached the pandemic vary from any other. Without classifying the countries, using specialistic knowledge it is very hard to use available data to learning. That is why in this case it is used only data set from Poland and omitted information from abroad. This approach is similar to Single Stock Price Predictions.

Sets are divided into train and test. Test is representing number of days to predict. Later in the modelling process the train set will be divided into training and validation sets.

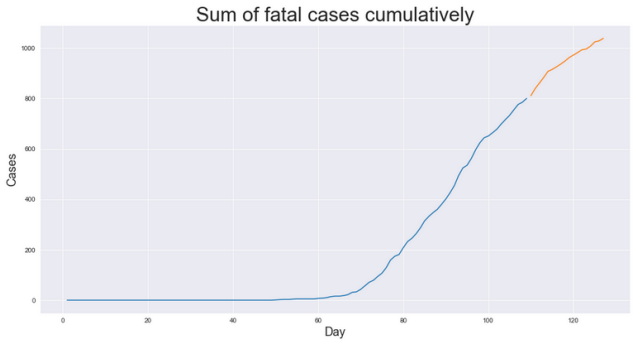

Fig. 5 Cumulative chart of fatal cases

Cumulative chart in case of data from Poland is very simple. One can be easily try to predict oncoming days just by linear regression.

B. SVM modeling

Function will compare three SVM systems with different kernels:

- linear
- polynomial
- radial

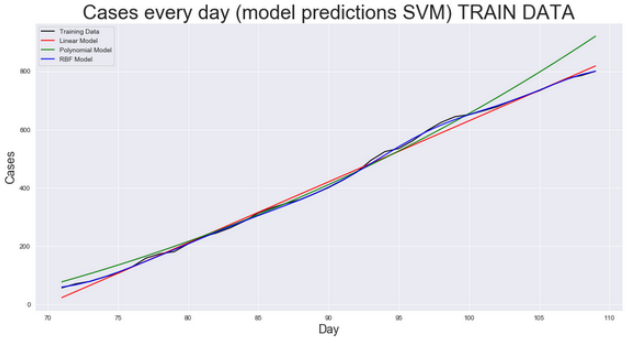

Fig. 6 Cumulative fatal cases - train data

Table 1 Values of error - train data

	Linear	Polynomial	RBF
RMSE	15,7	36,8	5,2
MAE	12,2	24,3	3,8

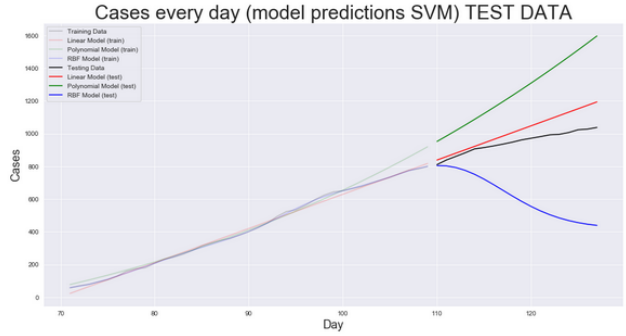

Fig. 7 Cumulative fatal cases - test data

Table 2 Values of error - test data

	Linear	Polynomial	RBF
RMSE	84,3	341,8	383,8
MAE	70,6	314,4	330,3

Created models with polynomial and linear kernels are very simple due to the type of the neuron used in this prediction. The research shows that it is impossible to predict any type of changes in the future.

Radial based function is great in the field it was training it in, due to the characteristic of RBF kernel, when we move out of the gamma field, the reductions are getting useless. The function of RBF neuron is a local Gaussian function, so the parameter gamma will decide how far from the centre it is activated. Due to no training in area of testing data, no neuron was train to predict there. Taking it into consideration, training metrics are great but testing metrics are unsatisfactory.

C. *Long Short-Term Memory*

LSTM model prediction is based on previous values. It use input from past tries to predict upcoming days. Preparation of data is different. Here database is divided on the basis of factor train_size (0-1). In this case it equals 0,8. 20% is a test data set, 80% is train set.

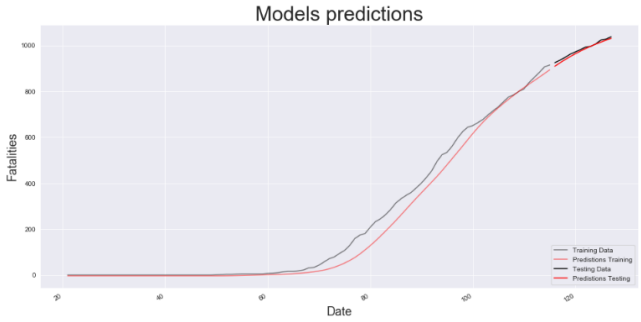

Fig. 8 Cumulative death cases - Poland

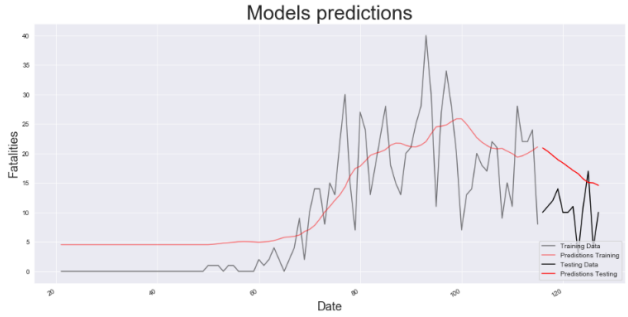

Fig. 9 Everyday death cases - Poland

Using daily fatal cases is useless, there is no direction that can be foreseen. Due to this fact there is used only cumulative sum of deaths.

Summing up, results from both models, it is shown, that LSTM model gave better outcome in prediction.

D. *Similar to Poland*

In this approach countries with similar statistics to Poland was selected (e.g. Sweden, Finland). The data was combine in order to predict the future fatal cases.

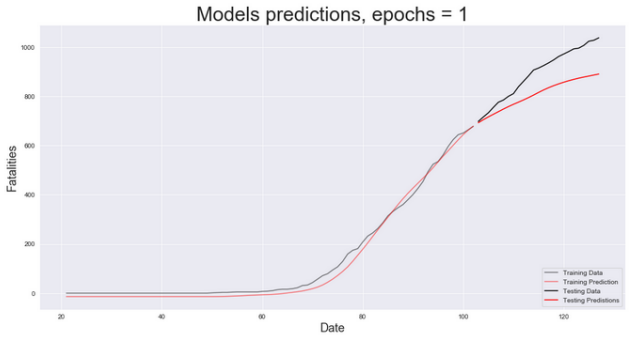

Fig. 10 Cumulative death cases, 1 epoch

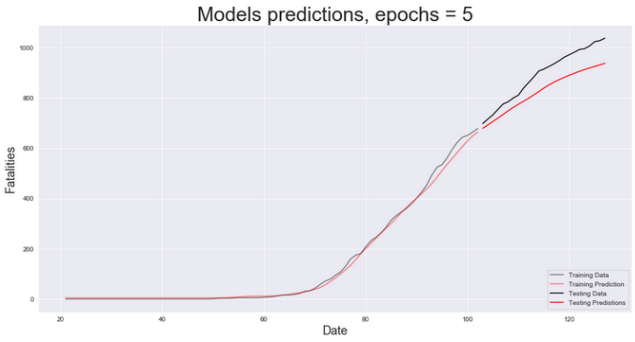

Fig. 11 Cumulative death cases, 5 epochs

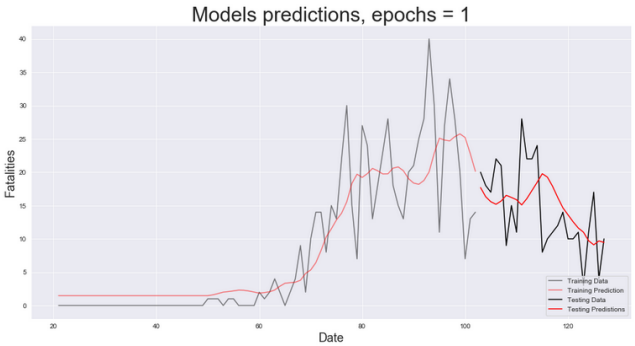

Fig. 12 Everyday death cases, 1 epoch

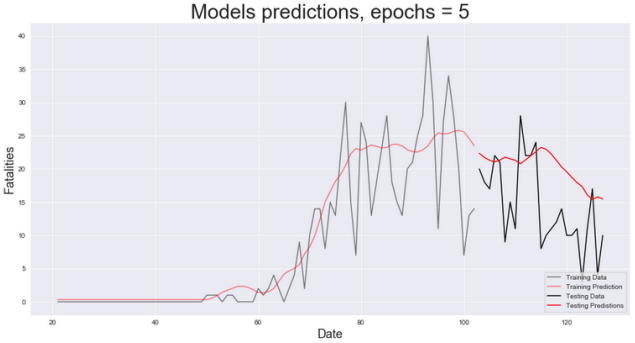

Fig. 13 Everyday death cases, 5 epoch

E. *Europe way*

Thanks to the generalization properties of neural networks it might be beneficial to teach the network on as much data as it is possible. It should manage to choose the most important parts by itself.

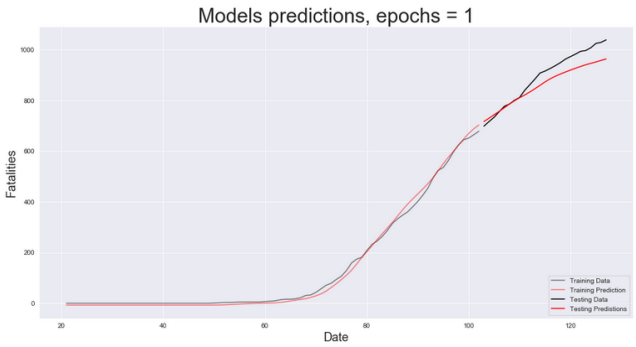

Fig. 14 Cumulative death cases, 1 epoch

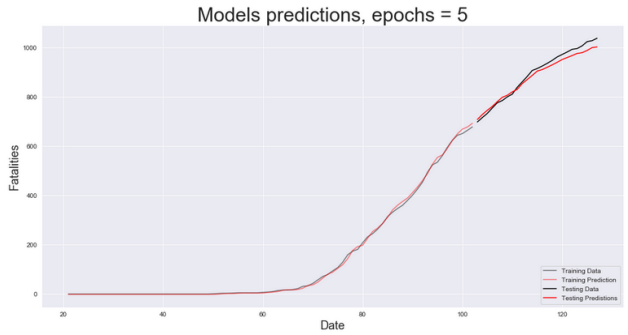

Fig. 15 Cumulative death cases, 5 epochs

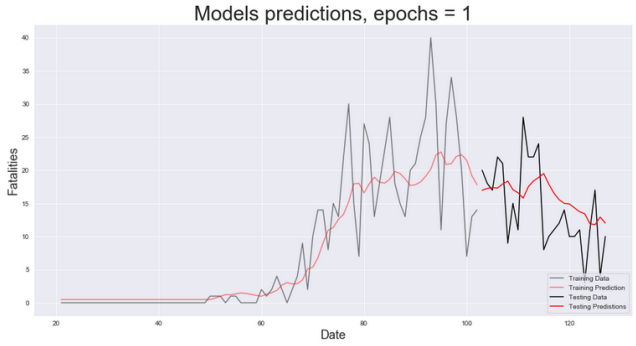

Fig. 16 Everyday death cases, 1 epoch

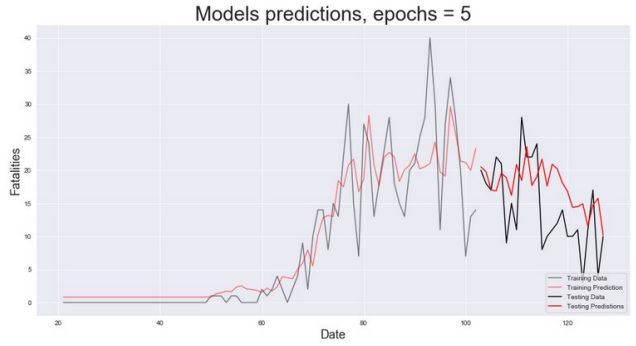

Fig. 17 Everyday death cases, 5 epoch

F. *World way*

This approach takes into consideration all available data from all over the world.

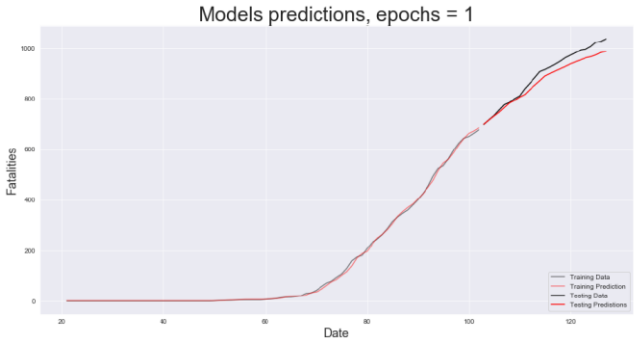

Fig. 18 Cumulative death cases

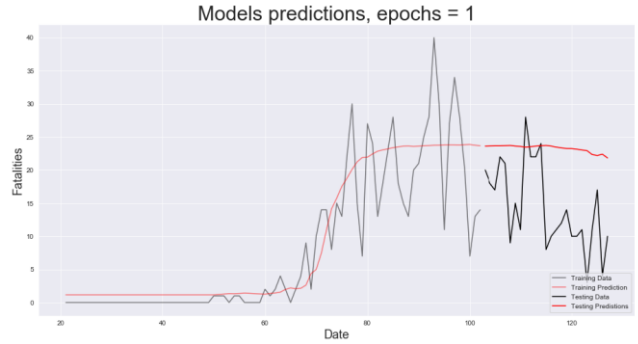

Fig. 19 Everyday death cases

CONCLUSION

Due to the nature of both algorithms it was immanent that results of LSMT prediction will be much better. SVR modelling is very useful in close proximity of training data, classification and regression is much better when identified data lies in between training data, which is very easy to notice after first predictions. The RBF kernel is very useful, but in our case, where the extrapolation is supposed to spread far out of training data one can easily see the influence of the restricted reach of generalization in this topology of neural network. LSTM on other hand is much more flexible, the only constraint to be considered is more time and resource consuming data preparation process. Based on last *n* samples one can easily extend the prediction as far as necessarily. Due to this fact this method was titled as best model, for problem presented in the article.

The work will be surely continued in the future, presented challenge is definitely very interesting and could be developed infinitely. Next steps should be adding more geographical information into the model, such as WHO districts. Additionally information about border closures, national movements and societies activity might introduce important insight into regression. Due to quickly growing dataset size, one might consider using deep neural network in analysis connected with right transfer learning method. Some opensource stock price algorithms might yield very good results in such prediction.

REFERENCES

[1] S. Osowski, “Sieci neuronowe do przetwarzania informacji”, wyd. 3, popr., Oficyna Wydawnicza Politechniki Warsawskiej, Warszawa, 2013

[2] R. Szmurło, Python programming and data analysis - lecture materials, 2020

[3] G. James, D. Witten, T. Hastie, R. Tibshirani , “An Introduction to Statistical Learning” , Springer, New York, 2013

[4] M. Dawson, “Python dla każdego”, wyd. 3, Helion, Gliwice, 2010

[5] J. Portilla “Python for Data Science and Machine Learning Bootcamp,” online course

[6] https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[7] https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes

[8] http://www.healthdata.org/covid