STATISTICS IN TRANSITION new series, March 2023
Vol. 24, No. 2, pp. 59–83, DOI 10.59170/stattrans-2023-020
Received – 30.05.2021; accepted – 25.11.2022

The prediction of new Covid-19 cases in Poland with machine learning models

Adam Chwila[1]

Abstract

The COVID-19 pandemic has had a huge impact both on the global economy and on everyday life in all countries all over the world. In this paper, we propose several possible machine learning approaches to forecasting new confirmed COVID-19 cases, including the LASSO regression, Gradient Boosted (GB) regression trees, Support Vector Regression (SVR), and Long-Short Term Memory (LSTM) neural network. The above methods are applied in two variants: to the data prepared for the whole Poland and to the data prepared separately for each of the 16 voivodeships (NUTS 2 regions). The learning of all the models has been performed in two variants: with the 5-fold time-series cross-validation as well as with the split into the single train and test subsets. The computations in the study used official statistics from government reports from the period of April 2020 to March 2022. We propose a setup of 16 scenarios of the model selection to detect the model characterized by the best ex-post prediction accuracy. The scenarios differ from each other by the following features: the machine learning model, the method for the hyperparameters selection and the data setup. The most accurate scenario for the LASSO and SVR machine learning approaches is the single train/test dataset split with data for the whole Poland, while in case of the LSTM and GB trees it is the cross validation with data for whole Poland. Among the best scenarios for each model, the most accurate ex-post RMSE is obtained for the SVR. For the model performing best in terms of the ex-post RMSE, the interpretation of the outcome is conducted with the Shapley values. The Shapley values make it possible to present the impact of auxiliary variables in the machine learning model on the actual predicted value. The knowledge regarding factors that have the strongest impact on the number of new infections can help companies to plan their economic activity during turbulent times of pandemics. We propose to identify and compare the most important variables that affect both the train and test datasets of the model.

Key words: machine learning, time series, COVID-19, forecasting, economic activity.

[1] University of Economics in Katowice, Katowice, Poland. E-mail: achwila@gmail.com.
ORCID: https://orcid.org/ 0000-0003-4671-4298.

1. Introduction and literature review

As of the start of the second quarter of 2022, the world is still struggling with the outbreak of the Covid-19 pandemic. The first official case of Covid-19 in Poland was registered on March 4, 2020, and as of 17 December 2020 the sum of all confirmed cases since March 2020 was equal to 1.17 million (Ministry of Health Republic of Poland, 2022). The predictions of the daily new infections can be very helpful in several different areas: the preparation of hospitals and medical services, the introduction of new restrictions that potentially can reduce the dynamic of the pandemic, and the plans regarding the future economic activity of the companies. They can also influence the development of vaccination programs.

Overall, the dynamic of the pandemic occurred to be a non-trivial issue due to many factors that can potentially influence the number of new infections. It also causes the forecasting of new confirmed cases much more challenging. Therefore, the application of the machine learning models that can potentially deliver accurate predictions based on non-linear dependencies is worth researching. This paper is structured as follows: Section 2 discusses the applied models, Section 3 describes the considered datasets as well as the concept of time series cross-validation, Section 4 discusses the results, Section 5 discusses limitations and the proposition of the future research, Section 6 summarizes the research.

Many researchers have successfully applied different forecasting approaches at different stages of pandemic development. There were several attempts to forecast the dynamic of the Covid-19 outbreak with the compartmental epidemiology models: in Italy (Giordano et al., 2020), China, Italy, and France (Fanelli, Piazza, 2020), Japan, Singapore, South Korea, and Italy (Chen, Lu, 2020), China, South Korea, Australia, USA and Italy (Cooper et al., 2020), Nigeria (Okuonghae, Omame, 2020). The different variants of compartmental models are mainly the modifications of the SIR susceptible-infected-removed model, which based on different parameters predicts the curves of pandemic dynamics (Kermack, McKendrick, 1927). However, these models focus mainly on the prediction of the whole pandemic dynamics, rather than on the daily changes based on the most recent data. Some studies regarding Covid-19 SIR models concluded that these models can be very sensitive to the assumed parameter describing the fraction of asymptomatic cases (Arino, Portet, 2020). The number of daily new confirmed cases was predicted with the ARIMA model with application to the data from January to February 2020 (Benvenuto et al., 2020).

Various classes of machine learning models have been applied to the Covid-19 data: Support Vector Machines regression, based on the lagged values of new daily confirmed cases (Peng, Nagata, 2020), logistic model (Wang et al., 2020), the long-short term neural network model for data in Canada (Chimmula, Zhang, 2020) and in India (Tomar, Gupta, 2020). The studies that involved finding the countries with similar

dynamics of Covid-19 outbreak with k-means and hierarchical analyses have been also conducted (Aydin, Yurdakul, 2020). The dependence on the mortality rate associated with the Covid-19 outbreak and the weather conditions with machine learning models have been studied for the data for Italy (Malki et al., 2020). Some studies regarding forecasting the new Covid-19 infections pointed out the following challenges associated with the complex machine learning models: the small datasets of historical data regarding the Covid-19 pandemic (less than 150 observations) and the inclusion of the variables connected to the government restrictions (Ahmad et al., 2020). In this paper, we try to refer to both issues, by the usage of the dataset with more than 250 daily records of data as well as the introduction of the variables associated with government restrictions. There were also studies regarding the performance of different machine learning methods (i.e. neural networks and Support Vector Machines) on small Covid-19 datasets (Fong et al., 2020). The cubic regression was also applied to the Covid-19 data from China (Gu et al., 2020). In the case of the new confirmed cases in Greece, there was a suggested network-defined splines model (Demertzis et al., 2020). The attempt to estimate the unobserved Covid-19 infections with an unbiased hierarchical Bayesian estimator with the auxiliary variable of current fatalities has been conducted for North American data (Vaid et al., 2020). Besides the application of machine learning methods in the case of pandemic forecasting, there has been a lot of research that compared the performance of machine learning methods with classic approaches. In the case of medical applications, there is a study comparing the performance of Support Vector Machines and neural networks with logistic regression for the problem of a number of oocytes retrieved, where the accuracy of machine learning models was higher than for the logistic regression (Barnett-Itzhaki et al., 2020). A study focused on ozone concentration prediction (Jumin et al., 2020) showed that Gradient Boosted trees outperformed the performance of linear regression and neural network models. In the case of air pollution concentration (Chen et al., 2019) the comparative study of different algorithms showed that generalized boosted model, random forest, and bagging outperformed backward stepwise linear regression, Support Vector Regression, and neural networks. There was also a study that analyzed results from 14 different articles based on the Covid-19 modelling with supervised and unsupervised methods (Kwekha-Rashid et al., 2021). The authors concluded that machine learning can produce an important role in COVID-19 investigations, prediction, and discrimination. Additionally, it can be involved in the health provider programs and plans to assess and triage the COVID-19 cases (Kwekha-Rashid et al., 2021).

To sum up and compare our work with different approaches taken by the researchers in the above studies over Covid-19 we can differentiate the following:

- The studies focused on compartmental epidemiology models, mainly the modifications of the SIR susceptible-infected-removed model. These methods aim

to deliver the long-term scenarios of the pandemic (i.e. 2 year horizon), based on strict assumptions i.e. that people who recovered from the disease are not going to get infected again. The goal of these models is quite different than the aim of the author's study, which is short-term prediction and explanation of the auxiliary variables actual impact on the new daily cases. The authors of SIR models aim to produce realistic predictions in the long horizon. Our goal is to accurately predict new daily cases in a horizon of 1-7 days and point out the variables that have the highest impact on the actual predictions.

- Autoregressive models taking into account solely lagged values of the new confirmed cases. In the following study, we include the component of lagged values of the new confirmed cases. Also, the additional auxiliary variables are considered to obtain more accurate results.
- Unsupervised machine learning models, i.e. to find the countries with similar dynamics of Covid-19 outbreak. A study with similar applications could be conducted with the NUTS-2 regions for the whole Poland. Nevertheless, it is out of the scope of the proposed research, which is focused on short-term predictions and the explainability of different factors considered in the modelling process. The unsupervised methods are designed to solve problems of a different nature than the considered supervised machine learning models (LASSO, SVR, LSTM, and GB trees).
- Supervised machine learning methods, focused on the short-term predictions of the new cases or similar statistics (i.e. fatalities). The research focuses on the whole spectrum of machine learning methods, either one or several different for comparison purpose. In our study we also focus on the choice of several machine learning methods:
    - the considered LASSO model is the linear regression with only one additional hyperparameter. It is the simplest among the considered methods, present to evaluate if the more complex methodologies significantly improve predictions,
    - the GB trees and SVR are the models that in different ways aim to take into account nonlinear relationships between variables,
    - the LSTM neural network is the most complex among the considered methods – an interesting aspect of the study is the comparison of the LSTM with GB trees/SVR and the LASSO (modified linear regression) in terms of stability and prediction power.

Although we consider 4 machine learning methods, the arbitrary choice of the considered methods is one of the limitations of our study. Hence, additionally we decided to choose models from 3 different levels of complexity to obtain a satisfying

range of results and evaluate if more complex approaches are better than the simpler ones. In this paper we propose a comparison of several different machine learning approaches with the setup, which based on our best knowledge is not presented in the literature:

- taking into consideration the complexity of the time series of new confirmed cases we propose different scenarios for the application of machine learning models: the models trained on the single train and test subsets as well as with 5-fold time series cross validation. The methods of different train and test data splits result in different hyperparameters chosen for the final model form,
- the study aims to compare the models trained on the times series data collected for the whole country, with the models trained on the data collected separately for each of the 16 voivodeships (NUTS 2 regions),
- studies presented in the current literature rarely compare the different machine learning approaches with each other when it comes to modelling Covid-19 data. Even if several machine learning models are proposed, the Long-Short Term Memory (LSTM) neural networks are not compared with other, less complex techniques. In this study, LSTM networks are in the scope with other methods,

Additionally, the study aims to deliver some insight into the impact of different factors on the actual new confirmed Covid-19 cases. The proposed models consider 38 variables collected from different sources. In different studies, it was analysed whether restrictions of movements can significantly influence the transmission of Covid-19 (Nouvellet et al., 2021). Therefore, the considered factors include:

- daily weather data,
- Covid-19 daily statistics (new confirmed cases, fatalities, tests, etc.)
- vaccinations data,
- Covid-19 Variants of Concern and Variants of interest data,
- place and time indicators,
- general policy and government restrictions,
- Covid-19 international indexes for Poland (i.e. containment health index),
- people mobility data.

The impact of the different considered factors on the actual number of new confirmed cases is an important aspect of the studies. We aim to use explanatory methods of complex machine learning models to detect the most important variables. The explanatory method is considered for the model with the best predictive power among the considered scenarios. We use the idea of the Shapley values (Shapley, 1953), successfully applied by Lundberg and Lee (2017), to explain the impact of factors on the model.

2. Models and methods

This study aims to predict the daily number of new infections in Poland based on the data from the two previous days. The compared machine learning models are Least Absolute Shrinkage and Selection Operator (LASSO) regression, Gradient Boosted (GB) regression trees, Support Vector Regression (SVR), and Long-Short Term Memory (LSTM) recurrent neural network. Each of these models is estimated for the data collected for the whole Poland as well as separately for 16 voivodeships (NUTS 2 regions). In the case of models estimated for the voivodeships, the sum of 16 predictions gives the prediction for new infections in Poland. In addition, each of the models is estimated in 2 variants of establishing hyperparameters, which gives 16 models in total (assuming that the set of hyperparameters for the dataset division variants is different for each model). All the models are estimated with code written in Python.

The first of the considered models is the LASSO Regression (Ranstam, Cook, 2018). The LASSO regression can be perceived as the modification of standard linear regression, which is made to reduce overfitting (the poor performance of the model on the dataset on which the model parameters are not trained). It also addresses the problem of the automated feature selection. The parameters of the LASSO regression are obtained by minimizing the modified error function (in this paper RMSE). The modification increases the value of the computed error function by as much as the sum of absolute values of model parameters multiplied by the hyperparameter α (Ranstam, Cook, 2018).

The exact value of α is established during the model training process. In this paper the LASSO regression is performed with Scikit-learn Python library (Pedregosa et al., 2011) with a function *sklearn.linear_model.Lasso*, where the LASSO regression parameters $\widehat{\boldsymbol{\beta}}$ are estimated by minimization of the equation:

$$\frac{1}{2n}||\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Y}||_2^2 + \alpha|\widehat{\boldsymbol{\beta}}|_1, \tag{1}$$

where n is the number of samples based on which the LASSO parameters are estimated, $\widehat{\boldsymbol{\beta}}$ is the vector of estimated parameters, $\boldsymbol{X}$ is the matrix of auxiliary variables from the sample, $\boldsymbol{Y}$ is the vector of the modelled variable from the sample, α is a regularization hyperparameter with a value chosen by the researcher.

Because of the model error modification, all the considered auxiliary variables need to be normalized or standardized. The standardization and normalization min-max of the auxiliary feature is made with the following equations, respectively:

$$x_{is} = \frac{x_i - \bar{x}}{\sigma}, \tag{2}$$

$$x_{in} = \frac{x_i - \min(x)}{\max(x) - \min(x)}, \tag{3}$$

where x_i is ith instance of the considered variable, x_{is} is the standardized ith instance of the considered variable, $\bar{x}$ is the mean of all instances of the considered dataset, σ is

the standard deviation of the considered dataset, $\min(x)$ is the minimum value of the considered dataset, $\max(x)$ is the maximum value of the considered dataset. The normalization min-max rescales the feature range to be [0, 1]. The mean, standard deviation as well as minimum and maximum values are always computed for the training subset. Then the values computed for the training subset are applied to the testing subset, which is done in order to avoid information leak between subsets. In the case of the LASSO regression, models with standardization, normalization min-max and no preprocessing of the data are tested.

The second considered model is Support Vector Regression (SVR) (Vapnik et al., 1994). It introduces the nonlinear relationships between the auxiliary features with the usage of Kernel functions (Sato et al., 2008). In this paper, the radial basis function kernel is considered, which can generalize the infinite-degree polynomial with the single hyperparameter $\gamma > 0$ which controls the influence of a single learning sample (Peng, Nagata, 2020), given by:

$$\kappa(x_i, x_j) = e^{-\gamma||x_i - x_j||^2}, \tag{4}$$

where $\kappa(x_i, x_j)$ is the radial basis kernel of samples x_i and x_j.

Another feature of the SVR is the specific error function minimized during the algorithm learning. The differences between the fitted and real values of the model are accounted for in the computed error only if they are higher than a certain value of hyperparameter ε. Therefore, the minimized function is given by (Peng, Nagata, 2020):

$$L_\varepsilon(y_i, \hat{y}_i) = \begin{cases} |y_i - \hat{y}_i| - \varepsilon, & if\ |y_i - \hat{y}_i| > \varepsilon \\ 0, & if\ |y_i - \hat{y}_i| \leq \varepsilon \end{cases}. \tag{5}$$

The next SVR hyperparameter is a penalty, which works similar to the α in the LASSO regression, but instead of summing the absolute values of model parameters, it takes the squares of model parameters (Hastie et al., 2008). The variables for each of the considered SVR scenarios are standardized. In this paper, the SVR is performed with the Scikit-learn Python library (Pedregosa et al., 2011) with the function *sklearn.svm.SVR*.

The third considered model is Gradient Boosted (GB) regression trees (Friedman, 2001). Decision tree is a very popular machine learning algorithm, which in its basic structure divides data many times into segments (leaves). After dividing the data into segments, the arithmetic mean of the response variable is determined for each leaf (Hastie et al., 2008). The GB algorithm is an enhanced version of the decision tree model. GB trees with standardization or with no preprocessing of the data are considered.

In this paper, the GB tree algorithm is performed with the XGBoost Python library (Chen, Guestrin, 2016) with the function *xgboost.XGBRegressor*. The applied GB trees algorithm is as follows.

1. The subsample of the observations is randomly drawn from the train dataset, the size of the subsample is a hyperparameter defined by the researcher (i.e. 70%).
2. For the subsample drawn in the previous step the decision tree is fitted with the CART algorithm (Breiman et al., 1984). During each following split of the single segment into two separate leaves, the different, randomly drawn subsample of the auxiliary variables is considered (i.e. 80% of variables). The subsample size is a hyperparameter, defined by the researcher.
3. After the creation of a tree the fitted values $\widehat{\boldsymbol{Y}}$ for each observation in the train dataset are calculated.
4. The fitted values are multiplied by learning rate hyperparameter η from range $[0,1]$, i.e. by 0.01. The residuals of the model are calculated with the equation:

$$\boldsymbol{r_b} = \boldsymbol{Y} - \eta\widehat{\boldsymbol{Y}}_{\boldsymbol{b}}. \tag{6}$$

5. Vector $\boldsymbol{Y}$ is replaced by the residuals obtained in the previous step: $\boldsymbol{Y} = \boldsymbol{r_b}$.
6. The algorithm goes back to the first step. Steps 1-5 are repeated B times, where B is a defined hyperparameter, i.e. 500.
7. The final form of GB trees is given by:

$$\widehat{\boldsymbol{Y}}_{boost} = \sum_{b=1}^{B} \eta\widehat{\boldsymbol{Y}}_b. \tag{7}$$

The fourth considered model is Long-Short Term Memory (LSTM) recurrent neural network (Hochreiter, Schmidhuber, 1997). Recurrent neural network is a method widely used in sequential data modelling (Toharudin et al., 2021). The recurrent neural network is an iterative method that in each iteration estimates the fitted values and additionally takes into consideration the values obtained with the previous iterations of the model. The LSTM is a modified recurrent neural network addressing some issues regarding the learning of the network with the backpropagation algorithm: the vanishing or exploding gradient (Hochreiter, Schmidhuber, 1997). The neural networks introduce complex, nonlinear relationships between variables by the usage of multiple neurons (the number of neurons is a hyperparameter) with nonlinear activation functions (the type of activation function is a hyperparameter). In this paper one-layered LSTMs are considered. In the LSTM network, two types of activation functions are used. The first one is typically a sigmoid function (Chimmula, Zhang, 2020), which gives an output in the range [0, 1] and is used for example to properly scale the output from the previous LSTM iteration. The sigmoid function is given by an equation:

$$Sigmoid(x) = \frac{1}{1+e^{-x}}. \tag{8}$$

The second one is similar as in the case of ordinary neural network and introduces nonlinearity to the structure. In this paper the Rectified Linear Unit (ReLU) function is chosen during the learning process, given by (He et al., 2015):

$$ReLU(x) = \max(0, x). \tag{9}$$

Other considered hyperparameters are: the distribution from which the weights are initialized (weights are the parameters of the network), the number of epochs (number of iterations based on which the backpropagation algorithm adjusts the weights), and the regularization hyperparameter. The variables for each of the considered LSTM scenarios are normalized with min-max normalization. The LSTM networks are built with the Keras Python library (Gulli, Pal, 2017) with *tf.keras.wrappers.scikit_learn.KerasRegressor* function, which is the implementation of the Scikit-learn (Pedregosa et al., 2011) regressor API for Keras Python library. The optimal weights of the LSTM networks are obtained with the usage of the adam (adaptive moment estimation) algorithm (Kingma, Ba, 2015).

For all 4 models the hyperparameters are established based on the 692 daily observations. In the case of models built for voivodeships the data are extended with 15 zero one variables indicating the voivodeship (NUTS 2 region) affiliation. In the case of the LSTM network, the auxiliary data from all the voivodeships for each day are accumulated in one data row and the output of the model for each row is a vector of 16 fitted/forecasted values of new infections (one for each voivodeship). The modified structure of the data for neural network correctly reflects the time dependencies between the variables, which is important in the LSTM concept.

3. Data and the split into subsets

The modelling of the new daily official confirmed cases of Covid-19 is a challenging issue. During the 24 considered months of the data (since the beginning of the pandemic in Poland) several trends connected to the different conditions have been observed, which can be observed in Figure 1.

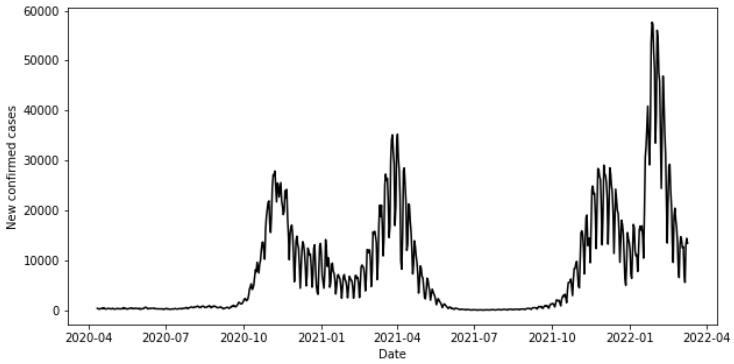

Figure 1: The dynamic of new official Covid-19 infections in Poland, data from 11th of April 2020 to 10th of March 2022

Table A1 in the appendix describes all the relevant data sources and Table A2 in the appendix explains each of the auxiliary variables. The incidental lack of data for some variables for single days are linearly interpolated. In the case of a lack of data on the borders of the considered time series, the nearest observation is assigned to the record with the missing data. The data used for modelling are lagged by 1 or 7 days. The Covid-19 data (new infections, new confirmed cases, Covid-19 variants, vaccinations) and weather data are lagged by one day. The general Covid-19 policy and mobility data are the variables lagged by 7 days because these variables are considered as additional conditions that affect the spread of the virus. The new infections are commonly noticed by affected people after a few days. For example, more strict gatherings restrictions are not supposed to influence the detected new infections the next day after the shift, but rather after at least a few days. The categorical variables are replaced by new 0-1 dummy variables. The official sources of Covid-19 data are connected to some limitations. Not all of the actual new infections of Covid-19 disease are recorded in the official statistics (Vaid et al., 2020). In this paper, the forecast of new confirmed cases is based solely on the official data from the government records. The concept of machine learning models is based on the division of all available data into train and test datasets. The train dataset is used for estimation of the model parameters and establishing the values of hyperparameters during the learning process. The test dataset contains samples which were not used by the researcher in any form during the learning process. Therefore, the predictions made on the test dataset allow assessing the model accuracy (Xu, Goodacre, 2018). In order to correctly compare the different machine learning methods with each other, the same division into train and test datasets should be applied for each method. The split that is often applied for common machine learning tasks in the literature is the train dataset equal to 80% of the available data and the test set equal to 20% (Hastie et al., 2008).

The whole dataset contains 699 records: the daily data of new confirmed cases from 11th April 2020 to 10th March 2022 and the data for the auxiliary variables from 4th April 2020 to 9th March 2022. The last 7 days of data are excluded from the dataset based on which the hyperparameters of the models are established (4th to 10th March 2022). The last 7 days of data are used for the evaluation of the predictive power of models. Therefore, the new confirmed cases based on which the model hyperparameters and parameters are established, are based on the period 11th April 2020 – 3rd March 2022 (692 observations). In machine learning, the hyperparameter is a value that affects the learning process of the given method (Hutter et al., 2014). The range of tested hyperparameters is defined by the researcher before the start of the whole process. The values of the parameters of each model are obtained with the training process.

The hyperparameters and parameters of the models are evaluated in two stages (Hastie et al., 2008). Firstly, the model is learned on the given training subset: for each combination of the hyperparameters, the parameters of the model are established, and then the prediction accuracy RMSE (Root Mean Squared Error) of the forecasts is calculated. The applied RMSEs are given by the following equations:

$$RMSE_{country\ level}^{goodness-of-fit} = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}}, \tag{10}$$

$$RMSE_{country\ level}^{ex\ post\ accuracy} = \sqrt{\frac{\sum_{i=n}^{n+m}(y_i-\tilde{y}_i)^2}{m}}, \tag{11}$$

$$RMSE_{NUTS\ 2\ \ level}^{goodness-of-fit} = \sqrt{\frac{\sum_{i=1}^{n}(\sum_{j=1}^{k}y_{ij}-\sum_{j=1}^{k}\hat{y}_{ij})^2}{n}}, \tag{12}$$

$$RMSE_{NUTS\ 2\ level}^{ex\ post\ accuracy} = \sqrt{\frac{\sum_{i=n}^{n+m}(\sum_{j=1}^{k}y_{ij}-\sum_{j=1}^{k}\tilde{y}_{ij})^2}{m}}, \tag{13}$$

where n is the number of observations based on which models are estimated, m is the number of observations for which predictions are made, k is the number of NUTS 2 regions, $\hat{y}_i$ is modelled fitted value, $\tilde{y}_i$ is a forecast, y_i is the real value. The RMSE indicates how the modelled values deviate from the real values on average.

In the next step, the hyperparameters of the model with the lowest value of prediction accuracy RMSE calculated for the testing subset are remembered and the parameters of the model with the remembered hyperparameters are estimated based on the whole considered dataset for the model creation purpose. Then, the performance of the model is established based on the 7 last observations of the dataset – the observations that are not involved in the model creation procedure. If the hyperparameters of the model are chosen based on the procedure of the k-fold cross-validation, then the dataset based on which the set of hyperparameters is established is divided into training and testing subsets k times. For each set of hyperparameters, the prediction accuracy RMSE is calculated for each of the k testing subsets (Hastie et al. 2008). Then for each set of hyperparameters, the average prediction accuracy RMSE calculated on testing subsets is computed and the hyperparameters with the lowest average prediction accuracy RMSE are chosen for parameters estimation based on the whole considered dataset for the model creation purpose. There are 2 scenarios of the division of the dataset into training and testing subsets. The first one is the single division of the dataset into training and testing subsets (where 90% of the observations is in the training subset) and the second one is the 5-fold cross validation adapted for the time series problem (Bergmeir, Benítez, 2012). Given the nature of the dependence of the following observations in the time series, the division of the dataset into training

and testing subsets should not be random, but rather established in a way that the order of the observations is not disrupted. The two ways of the division of the dataset are presented in Figure 2 and Figure 3.

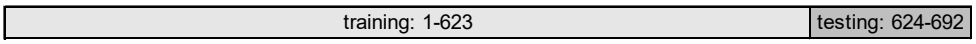

Figure 2: The division of the dataset into single training and testing subsets

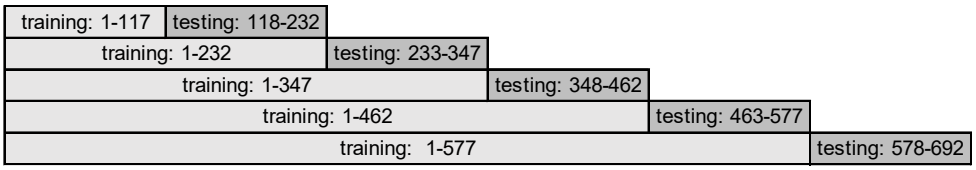

Figure 3: The division of the dataset with 5-fold time series cross validation Method

4. Results

Table 1 presents the RMSEs computed for the whole learning subset of 692 observations and the validation subset of 7 observations.

The model characterized by the best prediction power in terms of RMSE among the considered models is SVR trained with a single split of data into training and testing subsets and trained on the data for the whole Poland. The comparison of the real and predicted values for the validation subset is presented in Figure 4.

Table 1: Results of the models learned in different scenarios

Model	Training method: cross validation (cv) or single training/testing split (s)	Data: Poland (pl) or voivodeships (voi)	RMSE for learning subset	RMSE for validation
GB	cv	pl	9.1	1276.8
	s	pl	8.2	1401.3
	cv	voi	465.1	1928.2
	s	voi	455.5	1926.5
LASSO	cv	pl	1688.1	1793.0
	s	pl	1662.7	1502.7
	cv	voi	2816.8	3199.0
	s	voi	2444.6	2366.7
LSTM	cv	pl	215.3	2894.6
	s	pl	1519.8	3710.8
	cv	voi	847.8	4767.6
	s	voi	902.1	4957.0
SVR	cv	pl	147.0	1301.9
	s	pl	780.6	**1003.7**
	cv	voi	617.3	1270.8
	s	voi	679.0	2522.4

To maintain the consistency between the approaches considering the predictions for the whole Poland and for the voivodeships, the final predictions made for NUTS-2 regions are summed-up and compared to the results obtained on the country level (as indicated in equations 10-13). Because the data for NUTS-2 regions are collected into one dataset with the voivodeship indicator, it means that models automatically tend to focus on the days and NUTS-2 regions with the higher number of new observed cases. For example, the correct prediction for an instance with 10 000 actual new confirmed cases for NUTS-2 region A is more important than a very close prediction for an instance with 50 new cases for NUTS-2 region B. Therefore, the consistency of the approach for analysis of the NUTS-2 regions with data on the country level is maintained. We can observe that the predictions made for the whole Poland and on the level of NUTS-2 regions do not deviate significantly from each other for the RMSE for the validation dataset. The hyperparameter ranges can seriously influence model performance.

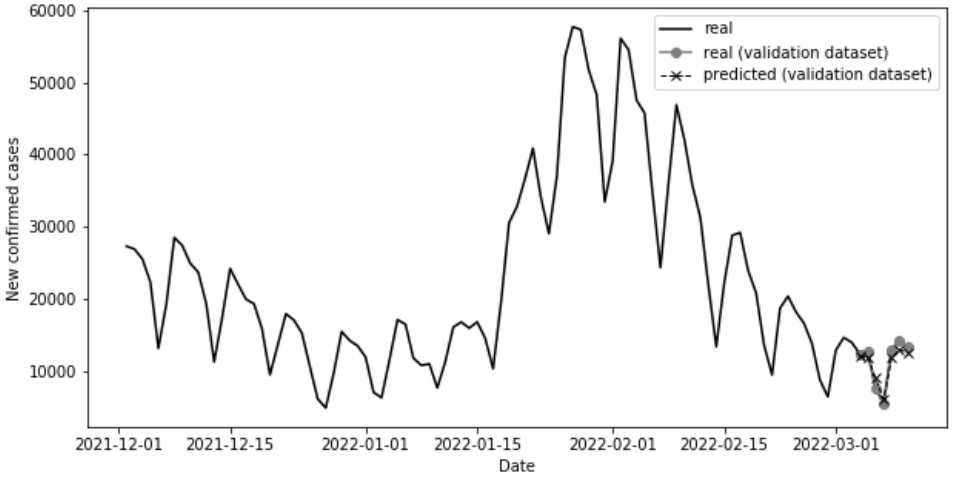

Figure 4: The prediction for 7 observations from the validation dataset by the Support Vector Regression with the lowest ex-post prediction accuracy RMSE

The mean absolute percentage error (MAPE) is given by the following equation:

$$MAPE = \frac{1}{m}\sum_{i=n+1}^{n+m}\frac{|y_i - \tilde{y}_i|}{|y_i|}, \tag{14}$$

where n is the number of observations based on which models are estimated, m is the number of observations for which predictions are made, $\tilde{y}_i$ is a forecast, y_i is the real value. The MAPEs for all considered scenarios are presented in Figure 5.

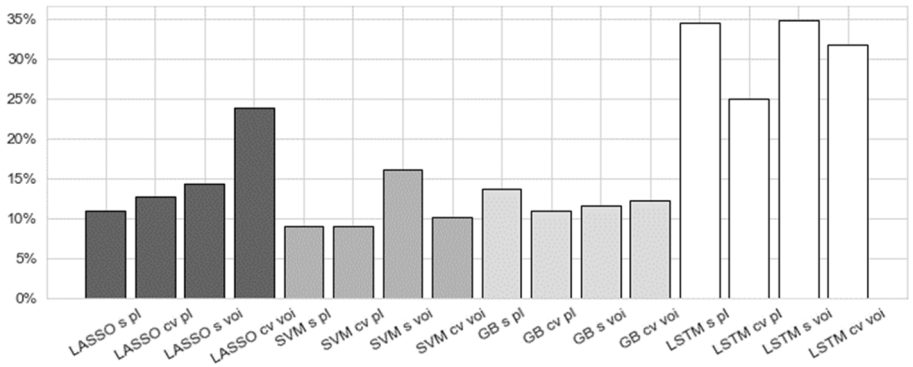

Figure 5: The MAPEs computed for validation dataset for all the considered scenarios

The boxplots of the ex-post predictions errors for 7 values from the validation dataset are presented in Figure 6. They indicate if the given method is characterized by the overfitting or underfitting of the predicted values and also indicate if the value of mean error measures is caused by consistent error values or rather single outlier observations.

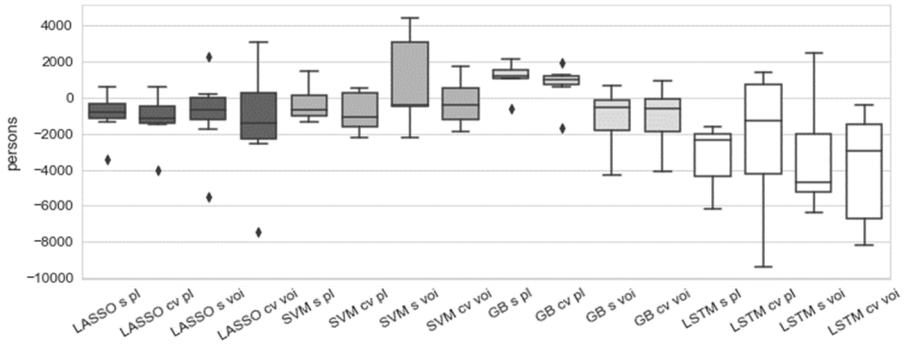

Figure 6: The boxplots of ex-post prediction errors for the validation dataset for all the considered scenarios

The errors produced by SVR and GB models are the most consistent, while in the case of LASSO and LSTM there are several outlier observations.

In addition to the choice of the model characterized by the lowest prediction RMSE we want to establish the impact of the considered factors on the actual predictions made by the model. We use the Shapley values to detect the variables that have the highest impact on the final numbers of new confirmed cases made by the SVR model established on the whole Poland data (with the hyperparameters choice based on the single train-test split of data). The idea of the Shapley values was originally proposed as

the concept of players' contribution in cooperative game theory (Shapley, 1953). In the context of machine learning models, we assume that each auxiliary variable is a "player" in a cooperative game, which contribute in a certain way to the final model prediction (Molnar, 2022). With the Shapley values, we want to estimate how much one of the concrete features impacts the deviation from the average prediction. The Shapley value is the average marginal contribution of a feature value across all possible feature subsets. To compute the exact Shapley value of the $j-th$ feature for a given instance of data, all possible sets of feature values have to be considered (Molnar, 2022). If the overall number of features is relatively high, the computation of the exact Shapley value is very time-consuming. Therefore, we use the following method to estimate the Shapley value for a single feature (Štrumbelj, Kononenko, 2014):

1. We choose the number of iterations M, model f, the dataset X, single instance x, and the $j-th$ feature for which the Shapley value is estimated.
2. For all $m = 1, \dots, M$:
    a. we draw a random instance z from dataset X, other than x,
    b. choose a random permutation p of all the considered features, which includes the $j-th$ feature,
    c. generate random order of the features in a permutation p,
    d. the vectors of auxiliary features for instances x and z are as follow: $x_p = (x_{(1)}, \dots, x_{(j)}, \dots x_{(k)})$, $z_p = (z_{(1)}, \dots, z_{(j)}, \dots z_{(k)})$, where k is the total number of features in permutation p,
    e. re-train model f on all the instances from the original dataset on the permutation of features p,
    f. we construct the two new instances of data by combining the instances x and z: we replace the features placed to the right of $j-th$ feature in instance x, including or excluding the $j-th$ feature from the replacement,
    g. the created instances are:
    $$x_{p\,new} = (x_{(1)}, \dots, x_{(j-1)}, x_{(j)}, z_{(j+1)}, \dots z_{(k)}) \text{ and}$$
    $$z_{p\,new} = (x_{(1)}, \dots, x_{(j-1)}, z_{(j)}, z_{(j+1)}, \dots z_{(k)}),$$
    h. we compute the marginal contribution of the $j-th$ feature on the prediction: $\phi_j^m = f(x_{p\,new}) - f(z_{p\,new})$.
3. We compute the Shapley value for instance x as the average marginal contribution: $\phi_j(x) = \frac{1}{M}\sum_{m=1}^{M} \phi_j^m$.

The above algorithm is repeated for all the features. In this study estimation of the Shapley values is performed with the Shap Python library (Lundberg, Lee, 2017).

The importance of the given feature is computed as the average of the absolute Shapley values for all the considered instances. The distributions of the Shapley values in a form of violin charts for the ten most important features for all the instances from

the training dataset are presented in Figure 7. The respective information for the testing dataset is presented in Figure 8.

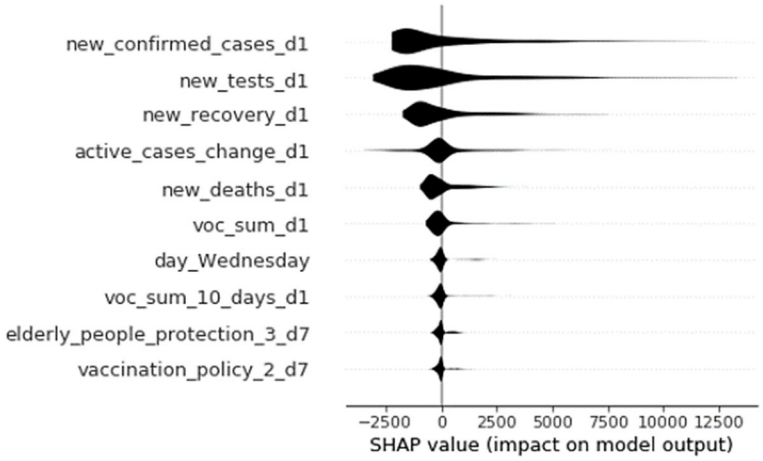

Figure 7: The distributions of the training dataset (692 instances) for the 10 features with the highest average of the absolute Shapley values

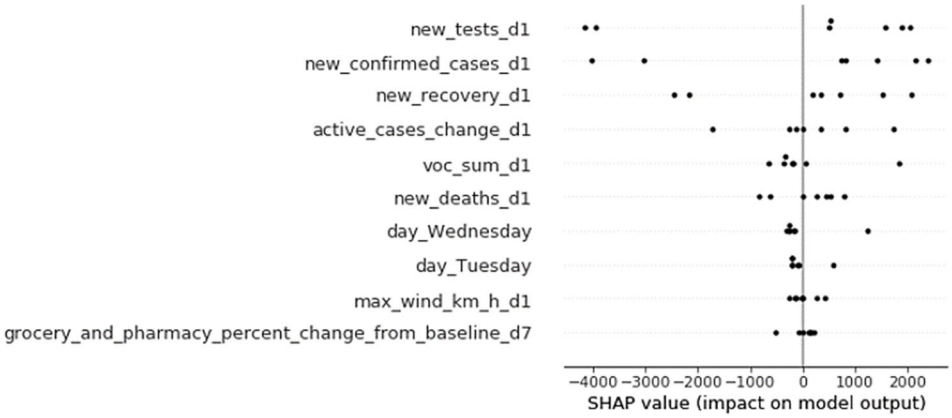

Figure 8. The distributions of the testing dataset (7 instances) for the 10 features with the highest average of the absolute Shapley values

The vertical line in Figure 7 and Figure 8 is the baseline (mean of the predictions of new daily confirmed cases). In the case of the training dataset the variables that have the highest impact on the final prediction for a given day are: new confirmed cases from

the previous day, new conducted tests in the previous day, new recoveries from the previous day, the change of active cases from the previous day, the new deaths from the previous day, new VOC and VOI cases reported in the previous day, time indicator for Wednesday, the sum of VOC and VOI occurrences from the last 10 following days (sum from the previous day), the indicator of elderly people protection from 7 days ago with level 3 (which means extensive restrictions for isolation and hygiene) , the indicator of vaccination policy from 7 days ago with level 2 (which means availability of vaccination for medical key workers and clinically vulnerable groups).

In the case of the testing dataset the list of the 10 most important features is quite similar, however, there are some new variables: time indicator for Tuesday, maximum wind speed from the previous day, the change from baseline for traffic congestion in groceries and pharmacies from 7 days ago.

There are some interesting aspects of the study:

- The time indicators for Tuesday and Wednesday are quite important for the final model results – the indicators for other days are less important.
- The high importance of the number of newly conducted tests may indicate that unfortunately, the true number of infections is much higher than officially reported. With the increasing number of tests, the new infections also increase.
- The overall number of vaccinated people is less important for the model than the overall vaccination policy, which may mean that the availability of vaccinations may change the people's behavior, which influenced the mobility.
- The wind speed may be important due to the indirect relationship of less frequent going out from home in the case of high wind speed. Also, the bad weather may indirectly affect the willingness to go out for a Covid-19 test.

The above interpretations of the results are only several of the possible ones.

5. Limitations and future work

One of the limitations of the studies is that it is based on the data from official government reports and there are a certain number of unobserved new infections. Another limitation is that the model can produce the predictions only for the next day, due to the consideration of variables lagged by one day. In future work, the application of variables lagged by more than one can be considered. Another limitation is the arbitrary choice of the concrete period based on which the model results are evaluated (the last 7 days of data). The important limitation of the study is the arbitrary choice of the number of days by which the auxiliary variables are lagged (1 or 7 days). Another limitation is the inclusion of auxiliary variables: the overall number of 38 variables is considered, but other indicators might be also included. The next limitation is the arbitrary choice of the searched ranges of hyperparameters, due to time-consuming

learning process. In future studies, the usage of data from other countries can be considered. Another limitation is the validity of the data. For example, the number of variants of concerns is dependent on the forwarding of the results to the public database GISAID from the different labs, which are not required to send the data. An additional limitation is connected to the methods proposed for the comparative studies (4 different machine learning models), which was an arbitrary choice.

6. Conclusions

We conclude that we propose the setup of 16 scenarios of model selection to detect the model with the best predictive power. The scenarios differ from each other by: machine learning model, the way of hyperparameters selection and the data setup (data for the whole Poland or for each of 16 Polish NUTS-2 regions). The model that produces the lowest error predictions for Covid-19 new daily infections in Poland is the Support Vector Regression model, with ex-post RMSE equal to 1003.7 cases. Ex-post RMSE is an average difference between the actual number of the new cases and the predictions for 7 days. The training process of the model is based on a single split of data into training and testing datasets. For the scenario characterized by the best predictive power, the impact of the auxiliary variables on the final results has been estimated with the Shapley values. Among the factors that have the highest impact on the final results are: Covid-19 statistics (confirmed cases, deaths, recoveries, active cases) from the previous day, Variants of Concern, time indicator for Wednesday, elderly people protection and the general vaccination policy. The machine learning models can help not only successfully predict the different Covid-19 characteristics in the short term periods, but also explain the factors that have the highest impact on the predictions for considered datasets.

Acknowledgment

The author is thankful to Michał Rogalski, who collected the daily Covid-19 data from the government reports, based on which most of the data were prepared: bit.ly/covid19-poland.

References

Ahmad, A., Garhwal, S., Ray, S., K., Kumar, G., Malebary, S., J., Barukab, O., M., (2020). The number of confirmed cases of Covid-19 by using machine learning: methods and challenges, *Archives of Computational Methods in Engineering*, https://doi.org/10.1007/s11831-020-09472-8.

Arino, J., Portet, S., (2020). A simple model for Covid-19, *Infectious Disease Modelling*, 5, pp. 309–315, https://doi.org/10.1016/j.idm.2020.04.002.

Aydin, N., Yurdakul, G., (2020). Assessing countries' performances against Covid-19 via WSIDEA and machine learning algorithms, *Applied Soft Computing Journal*, 97, p. 106792, https://doi.org/10.1016/j.asoc.2020.106792.

Barnett-Itzhaki, Z., Elbaz, M., Butterman, R., Amar, D., Amitay, M., Racowskyc, C., Orvieto, R., Hauser, R., Baccarelli, A., Machtinger, R., (2020). Machine learning vs. classic statistics for the prediction of IVF outcomes, *Journal of Assisted Reproduction and Genetics*, 37, pp. 2405–2412, https://doi.org/10.1007/s10815-020-01908-1.

Benvenuto, D., Giovanetti, M., Vasallo, L., Angeletti, S., Ciccozzi, M., (2020). Application of the ARIMA model on the Covid-2019 epidemic dataset, *Data in brief*, 29, p. 105340, https://doi.org/10.1016/j.dib.2020.105340.

Bergmeir, C., Benítez, J. M., (2012). On the use of cross-validation for time series predictor evaluation, *Information Sciences*, 191, pp. 192–213, https://doi.org/10.1016/j.ins.2011.12.028.

Blavatnik School of Government, University of Oxford, (2022). [online] Available at <https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker> [Accessed April 30, (2022)].

Breiman, L., Friedman, J., Olshen, R., Stone, C., (1984). *Classification and regression trees*, Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Chen, J., Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelick, M., Donkelar, A., Hividtfeldt, U., Katsouyanni, K., Et Al., (2019). A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide, *Environment International*, 130, p. 104934, DOI: 10.1016/j.envint.2019.104934.

Chen, T., Guestrin, T., (2016). XGBoost: a scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, DOI: 10.1145/2939672.2939785.

Chen, Y., Lu, P., (2020). A time-dependent SIR model for Covid-19 with undetectable infected persons, *IEEE Transactions on Network Science and Engineering*, 7(4), pp. 3279–3294, DOI: 10.1109/TNSE.2020.3024723.

Chimmula V., K., R., Zhang, L., (2020). Time series forecasting of Covid-19 transmission in Canada using LSTM networks, *Chaos, Solitons & Fractals*, 135, p. 109864, https://doi.org/10.1016/j.chaos.2020.109864.

Cooper, I., Mondal A., Antonopoulos C. G., (2020). A SIR model assumption for the spread of Covid-19 in different communities, *Chaos, Solitons* & *Fractals*, 139, p. 110057, https://doi.org/10.1016/j.chaos.2020.110057.

Daily Temperature In Capital Cities Of Voivodeships In Poland, (2022). [online] Available at: <https://freemeteo.pl.> [Accessed March 20, (2022)].

Demertzis, K., Tsiotas, D., Magafas, L., (2020). Modeling and forecasting the covid-19 temporal spread in Greece: an exploratory approach based on complex network defined splines, *International Journal of Environmental Research and Public Health*, 17, p. 4693, doi:10.3390/ijerph17134693.

Fanelli, D., Piazza, F., (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France, *Chaos, Solitons* & *Fractals*, 134, p. 109761, https://doi.org/ 10.1016/ j.chaos.2020.109761.

Fong, S., J., Li, N., D., G., Crespo R., G., Herrera-Viedma, E., (2020). Finding an accurate early forecasting model from small dataset: a case of 2019-ncov novel coronavirus outbreak, *International Journal of Interactive Multimedia and Artificial Intelligence,* 6, pp. 132–139, DOI: 10.9781/ijimai.2020.02.002.

Friedman, J., H., (2001). Greedy function approximation: a gradient boosting machine, *Annals of Statistics*, 29 (5), pp. 1189–1232, DOI: 10.1214/aos/1013203451.

Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., Colaneri, M., (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy, *Nature Medicine*, 26(6), pp. 855–860, https://doi.org/10.1038/s41591-020-0883-7.

GISAID database, (2022). [online] Available at <https://www.gisaid.org/> [Accessed April 30, (2022)].

Google Covid-19 Community Mobility Reports, (2022) [online] Available at: <https://www.google.com/covid19/mobility/> [Accessed March 20, (2022)].

Gu, C., Zhu, J., Sun, Y., Zhou, K., Gu, J., (2020). The inflection point about covid-19 may have passed, *Science Bulletin*, 65(11), pp. 865–867, DOI: 10.1016/ j.scib.2020.02.025.

Gulli, A., Pal, S., (2017). *Deep learning with Keras*, Packt Publishing Ltd.

Hastie, T., Tibshirani, R., Friedman, J., (2008). *The Elements of Statistical Learning*, Springer Science + Business Media LLC, New York.

He, K., Zhang, X., Ren, S., Sun, J., (2015). Delving deep into rectifiers: surpassing human-level performance on ImageNet classification, *Proceedings of the 2015 IEEE*

International Conference on Computer Vision (ICCV), Santiago, Chile, pp. 1026–1034, https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.123.

Hochreiter, S., Schmidhuber, J., (1997). Long Short-Term Memory, *Neural Computation*, 9(8), p. 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.

Hutter, F., Hoos, H., Leyton-Brown K., (2014). An efficient approach for assessing hyperparameter importance, *Proceedings of the 31st International Conference on Machine Learning*, 32(1), pp. 754–762.

Jumin, E., Zaini, N., Ahmed A., Abdullah, S., Ismail, M., Sherif, M., (2020). Machine learning versus linear regression modelling approach for accurate ozone concentrations prediction, *Engineering Applications of Computational Fluid Mechanics*, 14(1), pp. 713–725, https://doi.org/10.1080/19942060.2020.1758792.

Kermack, W. O., Mckendrick, A., G., (1927). A contribution to the mathematical theory of epidemics, *Proceedings of the Royal Society A*, 115(772), pp. 700–721, https://doi.org/10.1098/rspa.1927.0118.

Kingma, D., Ba, J., (2015). ADAM: a method for stochastic optimization, *International Conference on Learning Representations 2015*, San Diego, USA, https://arxiv.org/abs/1412.6980.

Kwekha-Rashid, A.S., Abduljabbar, H.N., Alhayani, B., (2021). Coronavirus disease (COVID-19) cases analysis using machine-learning applications, *Applied Nanoscience*, https://doi.org/10.1007/s13204-021-01868-7.

Lundberg, S., M., Lee, S. I., (2017). A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 4768–4777.

Malki, Z., Atlam, E., Hassanien, A., E., Dagnew, G., Elhosseini M., A., Gad, I., (2020). Association between weather data and Covid-19 pandemic predicting mortality rate: machine learning approaches, *Chaos, Solitons* & *Fractals*, 138, p. 110137, https://doi.org/10.1016/j.chaos.2020.110137.

Molnar C. (2022). *Interpretable machine learning. A guide for making black box models explainable*, Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, [online] Available at: <https://christophm.github.io/interpretable-ml-book/> [Accessed April 30, (2022)].

Ministry of Health Republic of Poland, (2022). [online] Available at: <https://www.gov.pl/web/coronavirus> [Accessed March 20, (2022)].

Nouvellet, P. Et Al., (2021). Reduction in mobility and COVID-19 transmission, *Nature Communications*, 12(1090), ttps://doi.org/10.1038/s41467-021-21358-2.

Okuonghae, D., Omame, A., (2020). Analysis of a mathematical model for Covid-19 population dynamics in Lagos, Nigeria, *Chaos, Solitons & Fractals*, 139, p. 110032, https://doi.org/10.1016/j.chaos.2020.110032.

Pedregosa, F. Et Al., (2011). Scikit-learn: machine learning in Python, *Journal of Machine Learning Research*, 12, pp. 2825–2830.

Peng, Y., Nagata, M. H., (2020). An empirical overview of nonlinearity and overfitting in machine learning using Covid-19 data, *Chaos, Solitons* & *Fractals*, 139, p. 110055, https://doi.org/10.1016/j.chaos.2020.110055.

Ranstam, J., Cook, J., A., (2018). LASSO regression, *British Journal of Surgery*, 105, p. 1348, https://doi.org/10.1002/bjs.10895.

R Interface To Covid-19 Data Hub, (2022). [online] Available at <https://cran.r-project.org/web/packages/COVID19/index.html> [Accessed March 20, (2022)].

Shapley, L. S., (1953). A value for n-person games, *Contributions to the Theory of Games*, 2 (28), pp. 307–317.

Štrumbelj, E., Kononenko I., (2014). Explaining prediction models and individual predictions with feature contributions, Knowledge and information systems, 41(3), pp. 647–665.

Toharudin, T., Pontoh, R. S., Caraka, R. E., Zahroh, S., Lee, Y., Chen, R., C., (2021). Employing long short-term memory and Facebook prophet model in air temperature forecasting, *Communications in Statistics – Simulation and Computation*, pp. 1–24, DOI: 10.1080/03610918.2020.1854302.

Tomar, A., Gupta, N., (2020). Prediction for the spread of covid-19 in India and effectiveness of preventive measures, *Science of The Total Environment*, 728, p. 138762, https://doi.org/10.1016/j.scitotenv.2020.138762.

Sato, J., R., Costafreda, S., Morettin, P., A., Brammer, M., J., (2008). Measuring time series predictability using Support Vector Regression, *Communications in Statistics – Simulation and Computation*, 37(6), pp. 1183–1197, https://doi.org/10.1080/03610910801942422.

Vaid, S., Cakan, C., Bhandari, M., (2020). Using machine learning to estimate unobserved Covid-19 infections in North America, *The Journal of Bone and Joint Surgery Incorporated*, 102 (70), pp. 1–5, http://dx.doi.org/10.2106/JBJS.20.00715.

Vapnik, V., Levin E., Cun Y. L., (1994). Measuring the vc-dimension of a learning machine, *Neural Computation*, 6(5), pp. 851–76, DOI: 10.1162/neco.1994.6.5.851.

Wang, P., Zheng, X., Li, J., Zhu, B., (2020). Prediction of epidemic trends in Covid-19 with logistic model and machine learning technics, *Chaos, Solitons & Fractals,* 139, p. 110058, DOI: 10.1016/j.chaos.2020.110058.

Xu, Y., Goodacre R., (2018). On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning, *Journal of Analysis and Testing*, 2, pp. 249–262.

Appendix

Table A1: Description of data source

Data	Description	Source
Weather	Daily temperature and wind speed data for the capital of voivodeship (in the case of data for whole Poland temperature for Warsaw).	https://freemeteo.pl (Daily temperature in capital cities of voivodeships in Poland, 2022).
Covid-19 data	Data for new Covid-19 confirmed cases, new fatalities, new recoveries, change of active cases, new tests, and new cases/tests ratio. The sum of people fully vaccinated (with two doses or with one of the Johnson & Johnson vaccination) and the sum of people vaccinated with the third dosage. Data are available for voivodeships and the whole Poland.	Data collected based on reports provided by the Ministry of Health, data from the WSSE, PSEZ, Voivodeship Offices, and those obtained in requests for access to public information: https://www.gov.pl/web/coronavirus (Ministry of Health Republic of Poland, 2022).
Covid-19 variants of concern and variants of interest data	The number of Covid-19 variants of concern (VOC) and variants of interest (VOI) reported by different labs analyzing Covid-19 tests in Poland.	The GISAID database: https://www.gisaid.org (Gisaid database, 2022).
Mobility data	The reports for movement trends of people in different places (change from baseline days). The baseline day is always the same day of the week. The value of movement trend in baseline day is the median value from the 5 weeks Jan 3 – Feb 6, 2020. Data are available for voivodeships and the whole Poland.	Google Covid-19 community mobility reports: https://www.google.com/covid19/mobility (Google Covid-19 community mobility reports, 2022).
General Covid-19 policy and government restrictions data	Available variables: closing of schools, closing of workplaces, cancelation of events, gatherings restrictions, closing of transport, stay-at-home restrictions, internal movement restrictions, international movement restrictions, information campaigns, testing policy, contact tracking, facial coverings, vaccination policy, elderly people protection, government response index, stringency index, containment health index, economic support index.	The policy measures from the R package are provided by Oxford Covid-19 Government Response Tracker (Blavatnik School of Government, University of Oxford, 2022): R Interface to COVID-19 Data Hub, 'Covid19' R package: https://cran.r-project.org/web/packages/COVID19/index.html (R interface to Covid-19 data hub, 2022).

Table A2: Description of auxiliary variables

Predictor
Covid-19 data:
- new confirmed cases from the previous day
- new fatalities from the previous day
- new recoveries from the previous day
- change of active cases - state for the previous day
- the number of new conducted tests - data from the previous day
- ratio of new confirmed cases to the number of conducted tests (from the previous day)
Vaccination data:
- the sum of people fully vaccinated (with two doses or with one of Johnson & Johnson vaccination) - state from the previous day
- the sum of people vaccinated with the third dosage - state from the previous day
Place and time indicators:
- weekday indicator: separate zero-one variable for weekday from the previous day
- voivodeship indicator: separate zero-one variable
Covid-19 variants of concern and variants of interest data:
- the number of VOC and VOI: VOC Omicron, VOC Alpha, VOC Delta, VOC Beta, VOC Gamma, VOI Eta, and VOI Lambda. Considered variables: the number of new VOC and VOI cases reported in the previous day and the sum of VOC and VOI occurrences from the last 10 following days
General Covid-19 policy and government restrictions data:
- school closing indicator from 7 days ago (4 levels)
- workplace closing from 7 days ago (4 levels)
- cancelation of events from 7 days ago (3 levels)
- gatherings restrictions from 7 days ago (5 levels)
- transport closing from 7 days ago (3 levels)
- stay home restrictions from 7 days ago (4 levels)
- internal movement restrictions from 7 days ago (3 levels)
- international movement restrictions from 7 days ago (5 levels)
- information campaigns from 7 days ago (3 levels)
- testing policy from 7 days ago (4 levels)
- contact tracking from 7 days ago (3 levels)
- facial coverings from 7 days ago (5 levels)
- vaccination policy from 7 days ago (6 levels)
- elderly people protection from 7 days ago (from no measure to extensive restrictions)
- government response index from 7 days ago
- stringency index from 7 days ago
- containment health index from 7 days ago
- economic support index from 7 days ago
Mobility:
The reports for movement trends of people in different places (change from baseline days). The baseline day is always the same day of the week. The value of movement trend in the baseline day is the median value from the 5 weeks Jan 3 – Feb 6, 2020. The indicator from 7 days ago is considered. The variables are: retail and recreations places, grocery and pharmacy, parks, transit stations, workplaces, residential.
Weather:
- 3 variables: maximum daily temperature, minimum daily temperature and a maximum speed of wind for the capital of voivodeship (in the case of data for whole Poland temperature for Warsaw) from the previous day