Machine Learning Approach to COVID-19 Epidemic Process Simulation using Polynomial Regression Model

Darina Kapusta, Alireza Mohammadi, Dmytro Chumachenko

National Aerospace University "Kharkiv Aviation Institute", Chkalow str., 17, Kharkiv, Ukraine

Abstract

The article presents an approach to modeling epidemic processes based on machine learning. A model is built based on the polynomial regression method. The simulation results allow us to calculate the predicted incidence of coronavirus infection in a certain area. The model has been shown to be accurate enough for use in public health policy-making settings. The disadvantage of using machine learning methods is the impossibility of identifying factors affecting the dynamics of the epidemic process. But, due to their high accuracy, such models can be used in an ensemble with agent-based and compartment models.

Keywords [1]

Epidemic model, polynomial regression, COVID-19 simulation, machine learning, artificial intelligence.

1. Introduction

An epidemic of a previously unknown coronavirus that causes pneumonia broke out in January 2020 in the Chinese province of Hubei [1]. Within two weeks, the virus spread to other countries.

The first cases of infection with the new coronavirus were recorded at the end of December 2019. Its appearance is associated with the seafood market in the city of Wuhan in China (Hubei province). Until the market closed on January 1, 2020, marine mammals, bats, chickens, rabbits and snakes were sold here [2]. Chinese virologists suggested that one of these animal species could become a source of infection.

During the first month, almost 6,000 people were infected with the coronavirus, more than 130 died from pneumonia caused by the virus. China has restricted communications with a number of metropolitan areas, quarantining 56 million people in 17 cities in Hubei. Later it turned out that the new coronavirus almost 90% coincides with the SARS-CoV virus, which appeared in China in the early 2000s and claimed about 800 lives. The new coronavirus was first assigned the code 2019-nCoV, and from February 11 it was renamed SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) [3].

As the epidemic was contained, the authorities of individual countries began to gradually ease restrictive measures in order to minimize damage to the economy and prevent social problems [4]. In the fall of 2020, the second and third waves of the epidemic began in many countries.

The death toll from coronavirus in the United States as of August 2021 has reached almost 700 thousand people [5]. This made the pandemic the deadliest in the 20th century, ahead of the Spanish flu pandemic.

Recently, an average of 10,000 people per day have been dying from coronavirus [6]. This is the highest figure since the beginning of March this year.

Scientists associate this with the emergence of strains "delta" and "iota", characterized by increased infectiousness and lethality [7].

International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2021), September 20-21, 2021, Kharkiv, Ukraine

EMAIL: kapusta.darina.yurievna@gmail.com (D. Kapusta); alireza.mohammadi9207@gmail.com (A. Mohammadi); dichumachenko@gmail.com (D. Chumachenko).
ORCID: 0000-0002-8168-8411 (D. Kapusta); 0000-0002-4964-4494 (A. Mohammadi); 0000- 0003-2623-3294 (D. Chumachenko).

The authorities of many countries have introduced compulsory vaccination for certain groups of the population (doctors, teachers, government officials, etc.) [8]. However, the number of vaccinated people is not enough to reduce the dynamics of the spread of the pandemic [9].

Mathematical and simulation modeling is an effective tool for identifying the rules for the spread of a pandemic. With the help of the simulation results, it is possible to form a scientifically grounded policy of countering the epidemic and the introduction of anti-epidemic measures to reduce the incidence.

The **aim of the paper** is to develop machine learning model of COVID-19 epidemic process dynamics and investigate it's results.

2. Analysis of epidemic process simulation approaches

To date, research teams from around the world have built many models for the spread of COVID-19. And morbidity modeling goes back centuries, when the breakthrough model of SIR, proposed by Kermak and McKendrick [10].

Models useful for studying infectious diseases on a population scale can generally be classified into two types: deterministic and stochastic.

In deterministic models, a large population is divided into smaller groups called classes, where each group represents a specific stage of the epidemic. Such models are often formulated as a set of differential equations (in continuous time) or difference equations (in discrete time) that help explain what, on average, happens on a population scale [11]. The decision of a deterministic model is a function of time or space and, as a rule, uniquely depends on the input data.

The stochastic model is formulated in terms of a stochastic process, which, in turn, is a set of random variables, *X (t,ω) ≡ X(t)*, defined as:

$$\{X(t,\omega)\mid t\in T\ \&\ \omega\in\Omega\}, \tag{1}$$

where T and Ω represent the time and total space for the sample.

The solution to the stochastic model is the probability distribution for each of the random variables [12]. Such models capture the inherent variability of demographic and environmental variability and are useful in small populations. More specifically, they allow observation of every person in a population at random.

Deterministic models are used to address questions such as “What proportion of people would have been infected during an epidemic outbreak?”, “What conditions must be satisfied to prevent and control an epidemic?”, etc. Deterministic models are the best when studying a large population, and models of stochastic epidemics are useful for a small population and provide answers to questions such as: “How long can a disease last?”, “What is the probability of a large fire”, etc.

Unlike deterministic models, stochastic models can be time consuming to create and require many simulations run to generate useful predictions. They can become very complex mathematically and lead to misperceptions of dynamics.

Different modeling approaches are suitable for investigating different problems. For example, simple deterministic models can be useful for understanding the underlying dynamics of an infection, but they are of limited use as a forecasting tool because any epidemic is unique and unlikely to follow the “average” pattern. Stochastic models are difficult to construct, but are especially useful for assessing risks and can be used to investigate the likelihood of different outcomes.

As for August 2021 wore than 250 thousand sources have been published on issues relating to COVID-19 in various fields, starting from medicine and biology, and finishing with computer science and mathematics. A lot of researches dedicated to COVID-19 modeling from different perspectives, such as COVID-19 characteristics [13], epidemiology [14], general Artificial Intelligence [15], machine learning [16], etc. They are solving different tasks, such as virus detection [17], contact tracing [18], forecasting [19], vaccine development [20], etc.

Main limitations of developed models and approaches are complex data, low quality of data, limited data, heterogeneity of population, interactions between multi-source data, disclosing unknown attributes, etc. All these limitations lead to a decrease in the accuracy of the forecast obtained with the help of modeling. The machine learning approach to epidemic process simulation can eliminate that drawback and shows high accuracy in simulation dynamics.

3. Polynomial regression model

Polynomial Regression is a supervised regression learning algorithm. The regression algorithm establishes a regression model between variables and obtains the correlation between variables and dependent variables in the learning process [21]. Regression analysis can be used for predictive or classification models. Common regression algorithms include: linear regression, nonlinear regression, logistic regression, polynomial regression, comb regression, lasso regression) and ElasticNet regression. Among them, the most commonly used are linear regression, nonlinear regression, and logistic regression.

In many cases, the linear model may not fit well with the target data curve, which requires the introduction of a non-linear regression model [22]. There are several strategies for nonlinear regression: the first strategy is to convert nonlinear regression to linear regression, and the second strategy is to convert nonlinear regression to polynomial regression. Polynomial regression adds a higher cardinality of an element (such as a square or cubic term), which is equivalent to increasing the model's degree of freedom to capture non-linear changes in the data.

The goal of regression analysis is to model the expected value of the dependent variable y in terms of the value of the independent variable (or vector of independent variables) x. In simple linear regression, the model

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{2}$$

where ε is the unobservable random error with mean zero due to the scalar variable x. In this model, for each unit of increase in the value of x, the conditional expectation of y is increased by β_1 units.

In many cases, such a linear relationship may not be observed. For example, if we model the number of deaths from COVID-19 depending on the percentage of population vaccinated in a certain area, then the yield increases also due to the decrease in the availability of places in hospitals. In this case, you can use a quadratic model of the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon. \tag{3}$$

In general terms, we can model the expected value of y as a polynomial of the nth degree, obtaining a general polynomial regression model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots + \beta_n x^n \ \varepsilon. \tag{4}$$

It is convenient that all these models are linear in terms of estimation, since the regression function is linear in terms of unknown parameters β_0, β_1, ... Therefore, for least squares analysis, computational and logical problems polynomial regression can be completely solved using the methods multiple regression. For this, x, x^2, ... are treated as separate explanatory variables in a multiple regression model.

From the formula (4), it can be concluded that to obtain a polynomial regression model that fits the target dataset perfectly, the key is to solve the value of the weight of each property-independent variable. Linear regression first constructs a convex function optimization function (such as: the minimum sum of squares of the difference between a given function value and the model's prediction value) and uses least squares [23] and gradient descent [24] to compute the final fit parameters.

Although polynomial regression is technically a special case of multiple linear regression, the interpretation of a fitted polynomial regression model requires a slightly different perspective. It is often difficult to interpret the individual coefficients when fitting a polynomial regression because the underlying monomials can be highly correlated. For example, x and x^2 have a correlation of more than 0,9 when x is evenly distributed over the interval $(0,1)$. Although the correlation can be reduced by using orthogonal polynomials, it is usually more informative to look at the fitted regression function as a whole. Point or simultaneous confidence intervals can then be used to determine the uncertainty in the estimate of the regression function.

4. Results

To simulate COVID-19 epidemic process in Ukraine we have used data of new cases and deaths, provided by the Center for Public Health of the Ministry of Health of Ukraine. Polynomial regression model program implementation has been made with Python programming language.

An important part of developing a software product is designing it. The process approach is the main element of management in organizations that manage public health in Ukraine. At the same time, one of the key aspects of this approach is to ensure the visibility ("transparency") of the management object (organization or system) through its accurate, sufficient, concise, easy-to-understand and analyze description.

Obviously, for complex systems, which include all institutions of the public health system in Ukraine at all levels (from regional laboratory centers to the Public Health Center under the Ministry of Health of Ukraine), it is almost impossible to obtain a single description suitable for any case, with faced by managers. Being multifaceted in the form and content of presentation, an organization (complex system) as a set of interrelated components can be represented by independent, complete "projections", the number of which is determined by the needs and tasks of management. To model business processes, the IDEF0 and DFD methodologies were used.

Functional model consists of four main elements:

- Process (Eng. Process), i.e. a function or sequence of actions that must be taken in order for the data to be processed. This can be creating an order, registering a customer, etc. It is customary to use verbs in the names of processes, i.e. Create customer (not create customer) or process order (not place an order). There is no strict system of requirements, as, for example, in IDEF0 or BPMN, where notations have a hard-coded syntax, since they can be executable. But still, certain rules should be adhered to so as not to confuse other people when reading the DFD.
- External entities. These are any objects that are not included in the system itself, but are for it a source of information or recipients of any information from the system after data processing. It can be a person, an external system, any storage media and data storage.
- Data store. Internal data storage for processes in the system. The received data before processing and the result after processing, as well as intermediate values must be stored somewhere. These are databases, tables, or any other option for organizing and storing data. It will store customer data, customer requests, invoices and any other data that entered the system or is the result of processing processes.
- Data flow. In the notation, it is displayed in the form of arrows that show what information is included and what comes from a particular block in the diagram.

The functional model of the program complex is shown in Figure 1.

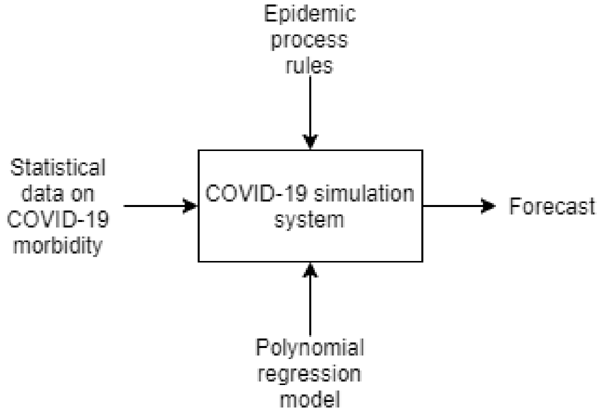

Epidemic process rules
Statistical data on COVID-19 morbidity
COVID-19 simulation system
Forecast
Polynomial regression model

Figure 1: Functional model of system.

Decomposition of functional diagram is presented in Figure 2.

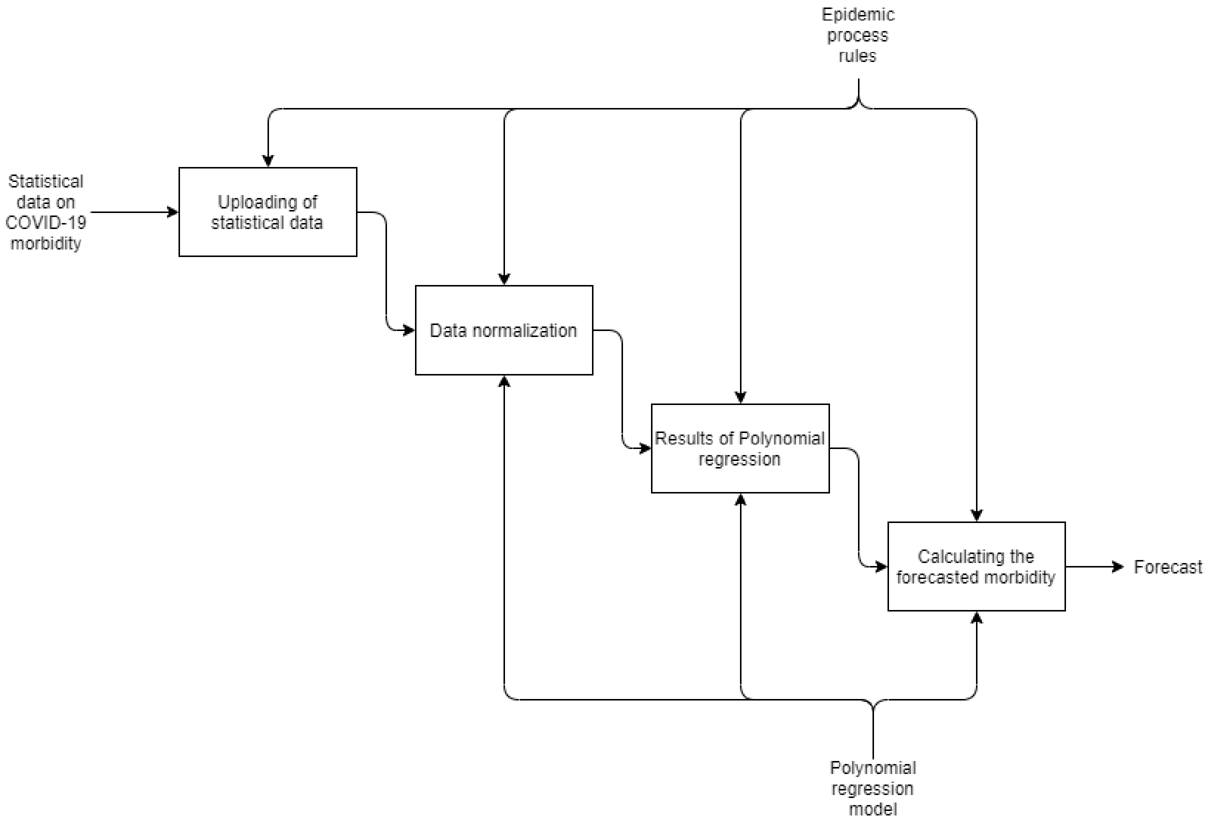

Epidemic process rules
Statistical data on COVID-19 morbidity
Uploading of statistical data
Data normalization
Results of Polynomial regression
Calculating the forecasted morbidity
Forecast
Polynomial regression model

Figure 2: Decomposition of functional model of system.

Implementing a particular use case requires the participation and interaction of specific instances of actors and classes. The most suitable tool for describing this interaction is sequence and communication diagrams, which essentially represent the same information. For the information system for forecasting the epidemic process COVID-19, use case diagrams are built, presented in Figures 3-5.

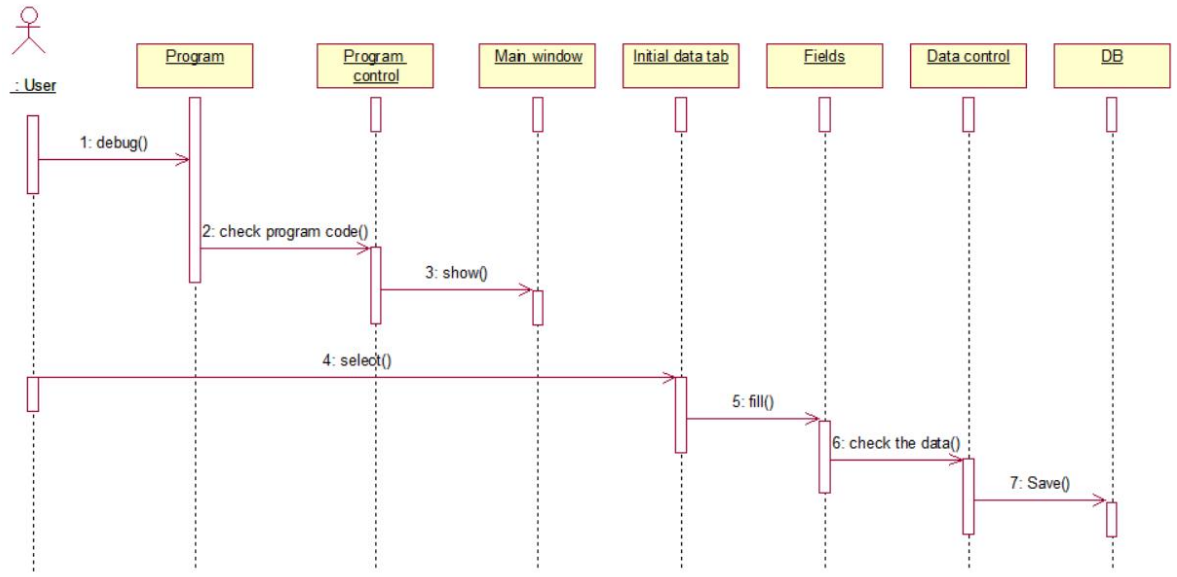


Figure 3: Case diagram “Initial data tab”.

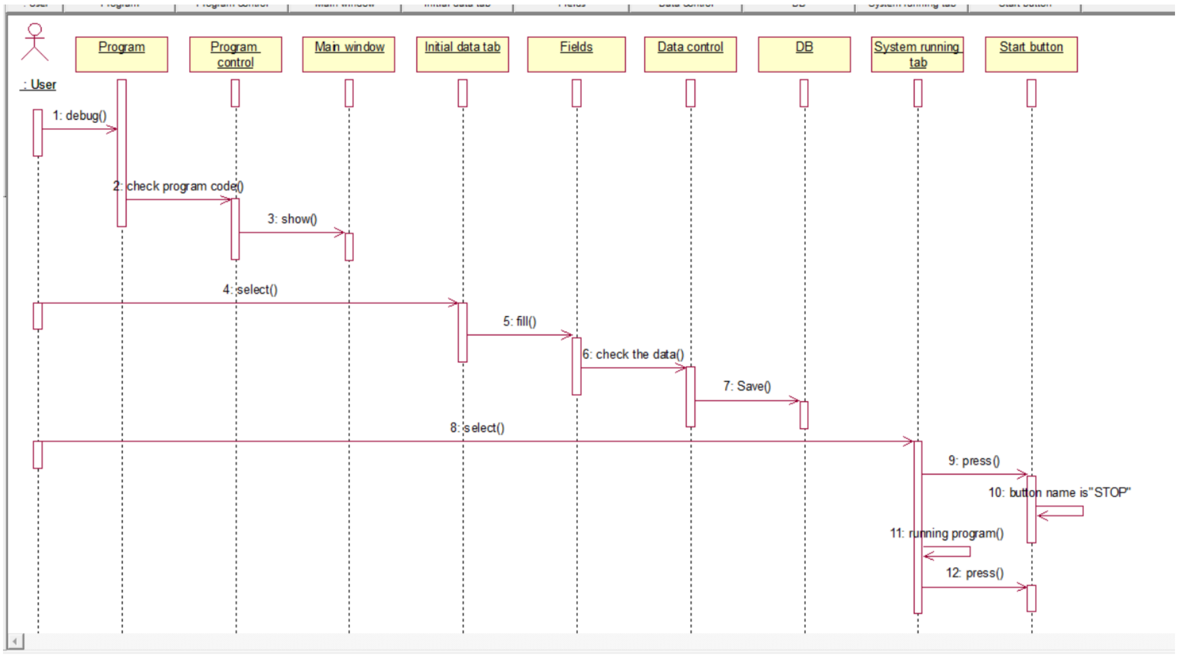


Figure 4: Case diagram “System running tab”.

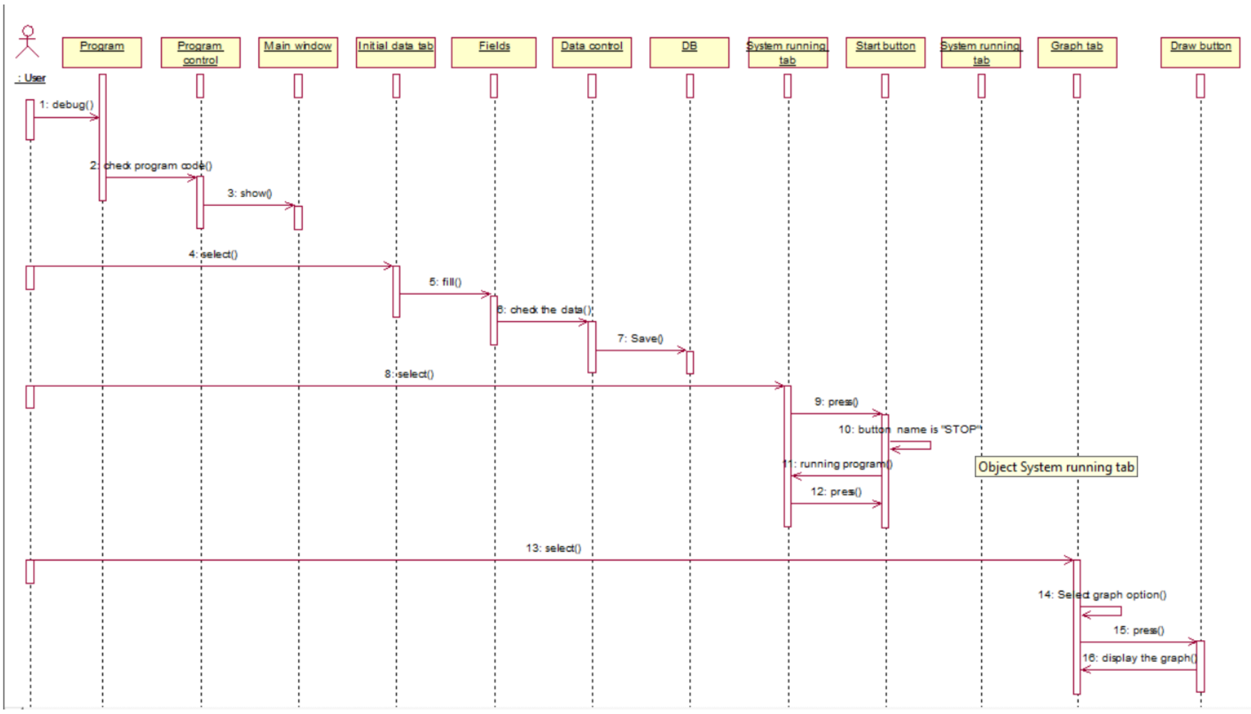
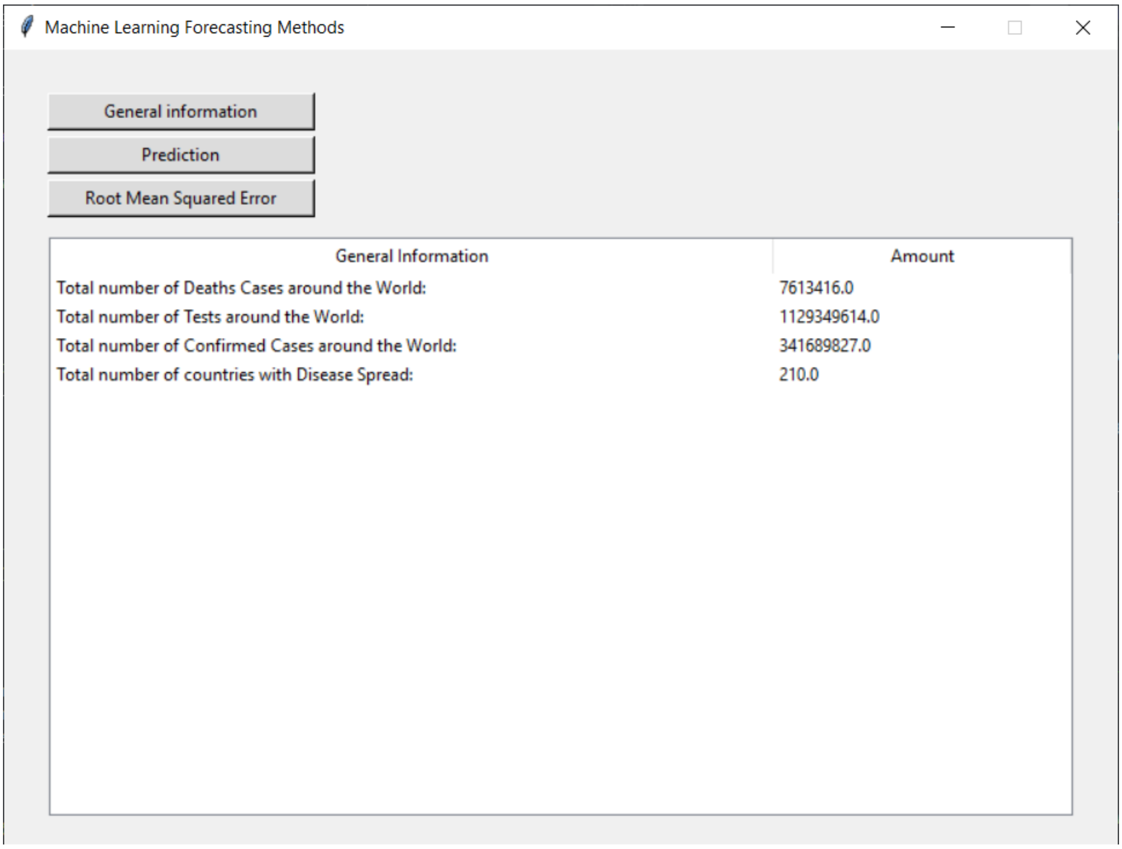


Figure 5: Case diagram “Graph tab”.

The developed information system includes both general data on COVID-19 morbidity in the world and detailed data on COVID-19 morbidity in Ukraine. Worldwide data is automatically loaded from the John Hopkins University database, and detailed data on Ukraine provided by the Center for Public Health of the Ministry of Health of Ukraine (fig. 6).

Results of COVID-19 epidemic process in Ukraine using Polynomial regression model are shown in Figure 7. The results are calculated for 10, 20 and 30 days. Analysis of experimental study shows that the most accurate result is provided with 10-days forecast. Still, other forecasts shows enough accuracy to use that results in Public Health institutions to provide anti-epidemic measures to combat to COVID-19 pandemic.



General Information	Amount
Total number of Deaths Cases around the World:	7613416.0
Total number of Tests around the World:	1129349614.0
Total number of Confirmed Cases around the World:	341689827.0
Total number of countries with Disease Spread:	210.0

Figure 6: Interface of information system of COVID-19 morbidity forecasting.

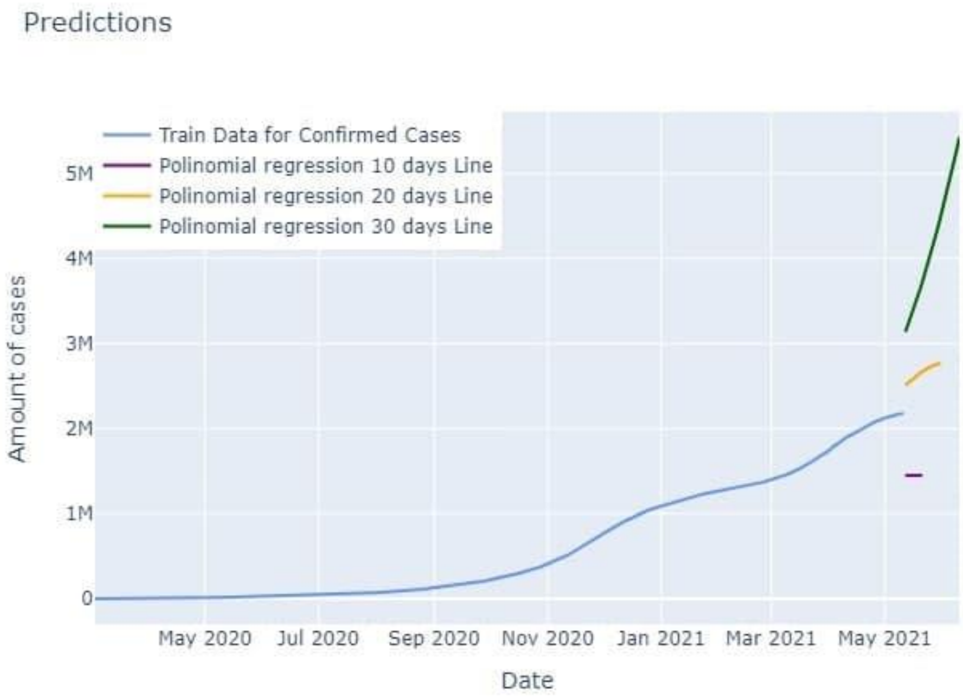


Figure 7: Results of COVID-19 simulation with Polynomial regression.

The errors of forecasting are presented in table 1.

Table 1
Forecasting errors

Simulation period	Root Mean Squared Error
10	65684.09170587435
20	196632.83601238678
30	480749.771966806

5. Conclusions

Within the framework of the study, a model for predicting the dynamics of the epidemic process COVID-19 was built on the basis of the polynomial regression method. Based on the model, an information system has been developed that can be implemented in public health institutions to make decisions on the implementation of anti-epidemic measures to reduce the dynamics of the incidence of COVID-19 in Ukraine.

The highest accuracy is shown by a forecast built for 10 days. However, the accuracy of the model allows the use of forecasts built for a longer period. Taking into account the incubation period of COVID-19, which averages 14 days, we recommend using a forecast for 20 days, which will also include new contacts with the source of infection.

The advantage of the developed model is the high, in comparison with other approaches, the accuracy of constructing the predicted morbidity. The disadvantage of the model is the impossibility of identifying factors influencing the dynamics of morbidity. Therefore, it is recommended to use the proposed model in an ensemble with other models that allow analyzing the informativeness of factors, for example, with multi-agent or compartment ones. A machine learning model can be used to verify predictions from other approaches, increasing their accuracy.

Acknowledgements

The study was funded by the National Research Foundation of Ukraine in the framework of the research project 2020.02/0404 on the topic “Development of intelligent technologies for assessing the epidemic situation to support decision-making within the population biosafety management” [25].

References

[1] T. Asselah, D. Durantel, E. Pasmant, G. Lau, R.F. Schinazi: COVID-19: Discovery, diagnostics and drug development. Journal of Hepatology 74 (1) (2021) 168-184. doi: 10.1016/j.jhep.2020.09.031.

[2] X. Lu, Y. Xing, G.W. Wong: COVID-19: lessons to date from China. Archives of Disease in Children 105 (12) (2020) 1146-1150. doi: 10.1136/archdischild-2020-319261.

[3] Z. Xu, et. al.: China shares experience during the COVID-19 outbreak. Burns: journal of the International Society for Burn Injuries 47 (1) (2021) 249-250. doi: 10.1016/j.burns.2020.05.014.

[4] L. Webb: COVID-19 lockdown: A perfect storm for older people's mental health. Journal of Psychiatric and Mental Health Nursing 28 (2) (2021) 300. doi: 10.1111/jpm.12644.

[5] O.O. Woolcott, R.N. Bergman: Mortality Attributed to COVID-19 in High-Altitude Populations. High Altitude Medicine and Biology 21 (4) (2020) 409-416. doi: 10.1089/ham.2020.0098.

[6] M.V. Blagosklonny: From causes of aging to death from COVID-19. Aging (Albany NY) 12 (11) (2020) 10004-10021. doi: 10.18632/aging.103493.

[7] I. Torjesen: Covid-19: Delta variant is now UK's most dominant strain and spreading through schools. BMJ (Clinical research) 373 (2021) n1445. doi: 10.1136/bmj.n1445.

[8] I. Ali: Impact of COVID-19 on vaccination programs: adverse or positive? Human Vaccines and Immunotherapeutics 16 (11) (2020) 2594-2600. doi: 10.1080/21645515.2020.1787065.

[9] F.M. Russell, B. Greenwood: Who should be prioritised for COVID-19 vaccination? Human Vaccines and Immunotherapeitocs 17 (5) (2021) 1317-1321. doi: 10.1080/21645515.2020.1827882.

[10] F. Brauer: The Kermack–McKendrick epidemic model revisited. Mathematical Biosciences 198 (2) (2005) 119-131. doi: 10.1016/j.mbs.2005.07.006.

[11] M.B. Trawocki: Deterministic Seirs Epidemic Model for Modeling Vital Dynamics, Vaccinations, and Temporary Immunity. Mathematics 5 (7) (2017) doi: 10.3390/math5010007

[12] L.J.S. Allen: A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. Infectious Disease Modelling 2 (2) (2017) 128-142. doi: 10.1016/j.idm.2017.03.001

[13] H. Esakandari, M. Nabi-Afjadi, J. Fakkari-Afjadi, N. Farahmandian, S.M. Miresmaeili, E. Bahreini: A comprehensive review of COVID-19 characteristics. Biological Procedures Online 22 (2020) 1–10.

[14] M. Park, A.R. Cook, J.T. Lim, Y. Sun, B.L. Dickens: A systematic review of COVID-19 epidemiology based on current evidence. Journal of Clinical Medicine 9 (4) (2020) 967.

[15] M.N. Islam, T.T. Inan, S. Rafi, S.S. Akter, I.H. Sarker, A.K.M.N. Islam: A Survey on the Use of AI and ML for Fighting the COVID-19 Pandemic. arXiv e-prints (2020), arXiv–2008.

[16] T.T. Nguyen: Artificial intelligence in the battle against coronavirus (COVID-19): a survey and future research directions. (2020) arXiv:2008.07343

[17] I. Izonin, et. al.: Predictive modeling based on small data in clinical medicine: RBF-based additive input-doubling method. Mathematical Biosciences and Engineering 18 (3) (2021) 2599-2613. doi: 10.3934/mbe.2021132

[18] Y. Mao, S. Jiang, D. Nametz: Data-driven Analytical Models of COVID-2019 for Epidemic Prediction, Clinical Diagnosis, Policy Effectiveness and Contact Tracing: A Survey. (2020).

[19] O. Byambasuren, et. al.: Estimating the Extent of True Asymptomatic COVID-19 and Its Potential for Community Transmission: Systematic Review and Meta-Analysis. Journal of the Association of Medical Microbiology and Infectious Disease Canada 5 (4) (2020) 223–234.

[20] A.K. Arshadi, et. al.: Artificial Intelligence for COVID-19 Drug Discovery and Vaccine Development. Frontiers in Artificial Intelligence 3 (2020) 65.

[21] H. Li, S. Yamamoto: Polynomial regression based model-free predictive control for nonlinear systems. 2016 55th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE) (2016) 578-582, doi: 10.1109/SICE.2016.7749264.

[22] S. Kavitha, S. Varuna, R. Ramya: A comparative analysis on linear regression and support vector regression. 2016 Online International Conference on Green Engineering and Technologies (IC-GET) (2016) 1-5, doi: 10.1109/GET.2016.7916627.

[23] V.P. Mashtalir, et. al.: Group structures on quotient sets in classification problems. Cybernetics and Systems Analysis 50 (4) (2014) 507-518. doi: 10.1007/s10559-014-9639-z

[24] S.N. Gerasin, et. al.: Set coverings and tolerance relations. Cybernetics and Systems Analysis 44 (3) (2008) 333-340. doi: 10.1007/s10559-008-9007-y

[25] S. Yakovlev, et. al., The concept of developing a decision support system for the epidemic morbidity control, CEUR Workshop Proceedings 2753 (2020) 265–274.