УДК 004.89 **doi:** https://doi.org/10.20998/2522-9052.2022.2.03

Г. М. Хорошун, О. І. Рязанцев, М. О. Коверга, С. А. Покришка

Східноукраїнський національний університет імені Володимира Даля, Сєвєродонецьк, Україна

МОДЕЛІ МАШИННОГО НАВЧАННЯ ДЛЯ ПЕРЕДБАЧЕННЯ КІЛЬКОСТІ ЗАХВОРІЛИХ НА COVID-19 В УКРАЇНІ ТА ІНДІЇ

Анотація. Побудовані моделі передбачення кількості захворілих на COVID-19 з використанням методів машинного навчання. Побудовані моделі навчались на даних зібраних з різних офіційних джерел, включаючи Всесвітню Організацію Здоров'я, з початку епідемії до теперішнього часу. Для навчання моделей передбачення кількості захворілих на COVID-19 обрано Україну та Індію. Методами, що надали високу точність прогнозу для існуючих даних, виявились алгоритми лінійної регресії для України та градієнтного бустингу для Індії. Аналіз даних проводився за допомогою мови програмування Python, з використанням бібліотеки Sklearn, яка побудована на основі SciPy (Scientific Python). Крім того, використовувалась бібліотека алгоритму градієнтного бустингу XGboost. Для розробки моделі обрано багатофакторне прогнозування часових рядів з використанням у якості предикаторів запізнення часового ряду. Визначено характеристики, що враховуються при навчанні моделі, а саме: дата початку події, день тижня, номер тижня, місяць та інші. Проведено аналіз щодо визначення впливу цих параметрів на якість навчання моделі. Оцінені похибки моделей та точність прогнозу з найкращими показниками 0.83 для України та 0.75 для Індії. Побудовані моделі дозволяють передбачати епідеміологічну ситуацію в майбутньому, координувати дії у різних галузях охорони здоров'я та проводити обґрунтовані превентивні заходи на державному рівні.

Ключові слова: машинне навчання; моделі прогнозування; метод лінійної регресії; метод градієнтного прискорення.

Постановка та аналіз проблеми

Коронавірусна інфекція 2019 року також відома як COVID-19 (абревіатура від англ. COronaVIrus Disease 2019) є епідемією, що охопила країни всього світу. Важкий перебіг захворювання, постковідний синдром та значна кількість летальних випадків [1] призвели до необхідності детального вивчення цього питання та прогнозування.

Цінність і значущість зібраних даних щодо кількості захворювань за певний проміжок часу можна усвідомити, побудувавши моделі, які дозволяють передбачати епідеміологічну ситуацію в майбутньому, координувати дії у різних галузях охорони здоров'я та проводити обґрунтовані превентивні заходи на державному рівні.

Для побудови передбачуваних моделей використовуються набори методів та алгоритмів машинного навчання з області штучного інтелекту, які забезпечують ефективне самонавчання моделі. Вперше визначення поняття машинного навчання, як області досліджень, що дає комп'ютерам здатність вчитися без того, щоб їх явно програмували надане в 1952 році А. Самюелем.

Наразі результати працюючих моделей щодо передбачення ринку, людської поведінки та вирішення задач збільшення прибутку використовують відомі компанії по всьому світу. Основні тренди у штучному інтелекті та методах машинного навчання в різних галузях на сучасному етапі розвитку наведені в роботі [2].

Отже, існує необхідність побудови передбачуваних моделей з використанням методів машинного навчання, які будуть навчатися на існуючій статистиці по кількості захворілих на COVID-19, починаючи з 2019 року.

В ході вирішення множини подібних завдань навчена модель з визначеною точністю буде надавати прогноз по кількості захворілих на обраний період часу. Актуальність роботи забезпечується сьогоденним епідеміологічним становищем по COVID-19 та зацікавленістю в результатах прогнозу у всіх сферах виробництва та наданні послуг в різних країнах світу.

Загальна задача створення передбачуваних моделей

В якості алгоритмів для навчання моделі, що може передбачити кількість захворілих на COVID-19 в майбутньому пропонується використовувати лінійну регресію [3] та XGBoost [4].

Модель лінійної регресії може бути формалізована наступним чином. Кількість хворих, визначених в певний день є залежною змінною y. Набір характеристик, які на нашу думку впливають на величину y є незалежні змінні $x = (x_1, \ldots, x_n)$, де n – це кількість параметрів. Вважаємо, що залежність між x та y лінійна і описується рівнянням регресії:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon \qquad (1)$$

з коефіцієнтами регресії $\beta_0, \beta_1, \ldots, \beta_n$ та випадковою помилкою ε. Лінійна регресія обчислює прогнозовані ваги виміру, що позначаються як $b_0, b_1, \ldots, b_n$. Вони визначають оцінена функцію регресії

$$f(x) = b_0 + b_1 x_1 + \cdots + b_n x_n, \qquad (2)$$

яка презентує залежність між входами та виходом. Для кожного результату спостереження $i = 1, \ldots, n$, оцінена або передбачена відповідь $f(x_i)$ має бути якомога ближчою до відповідної фактичної відповіді y_i. Різниці $y_i - f(x_i)$ всім результатів спостережень називаються залишками. Регресія визначає найкращі прогнозовані ваги виміру, які відповідають найменшим залишкам.

Градієнтний бустинг це метод машинного навчання, який будує модель прогнозування у вигляді ансамблю слабких прогностичних моделей - дерев рішень та використовує фреймворк градієнтного бустингу.

Введемо модель виваженого голосування:

$$h(x) = \sum_{i=0}^{n} c_i a_i, x \in X, b_i \in R, \quad (3)$$

де X — це простір, в якому знаходяться об'єкти, c_i, a_i - це коефіцієнти моделі та дерева рішень. Припустимо, що на якомусь кроці з допомогою описаних правил вдалося додати у композицію T-1 слабкий алгоритм. Щоб навчитися розуміти, який саме має бути алгоритм на кроці T, введемо функцію помилки:

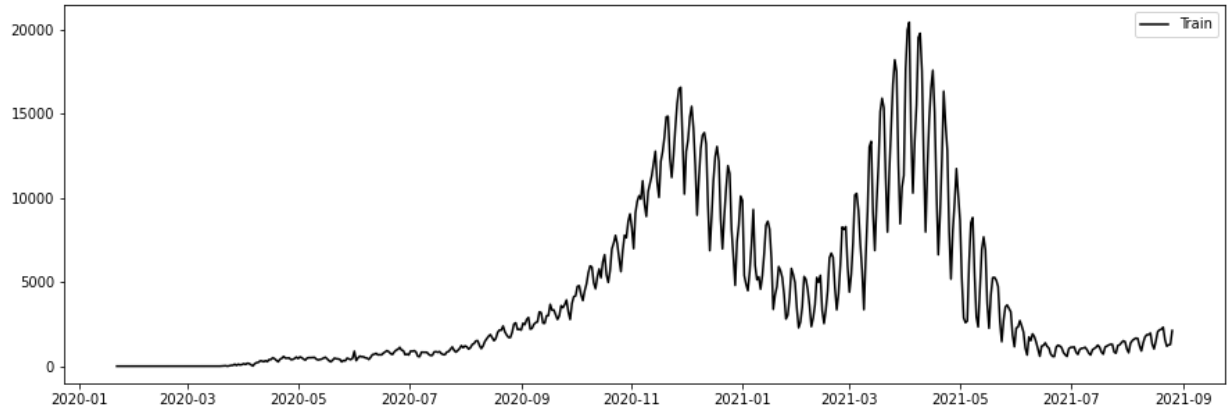
$$err(h) = \sum_{j=1}^{N} c_i L(\sum_{i=1}^{T-1} a_i c_i(x_j) + c_\tau a_\tau(x_j)) \to \min_{a_\tau c_\tau}. \quad (4)$$

Найкращим алгоритмом буде той, який зможе максимально зменшувати помилку, отриману на попередніх ітераціях. Градієнтний бустинг дозволяє визначати мінімальну помилку.

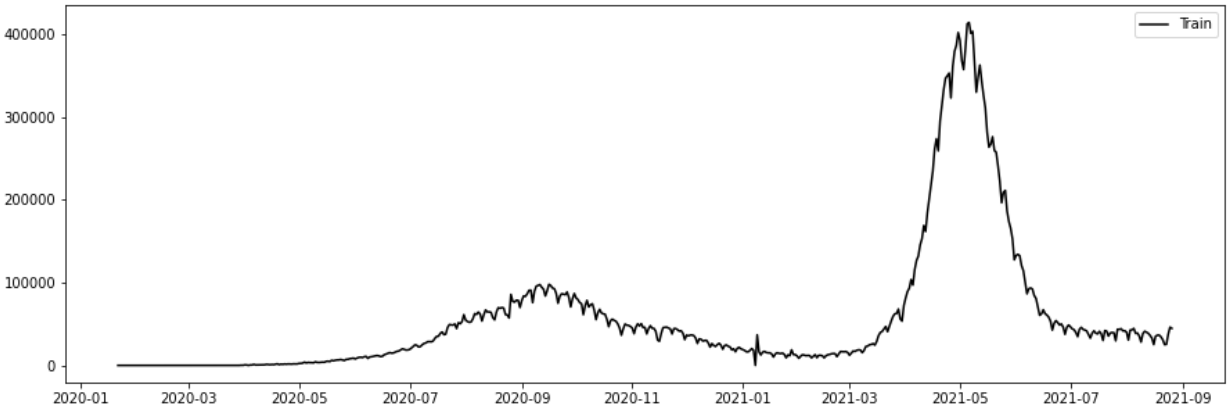
Аналіз даних проводився за допомогою мови програмування python, з використанням бібліотеки Sklearn - Scikit-learn (Sklearn) з реалізацією цілого ряду алгоритмів для навчання з учителем та без учителя, яка побудована на основі SciPy (Scientific Python). Крім того, використовувалась бібліотека алгоритму градієнтного бустингу XGboost.

Результати роботи моделей

Для проведення порівняльного аналізу можливостей двох методів розглянемо дві країни – Україну та Індію, які відрізняються за кількістю населення та рівнем медичного обслуговування. Набори даних про кількість випадків захворювань на COVID-19 в Україні та Індії завантажено з ресурсу [5] та презентовано на рис. 1. Дані часових рядів фіксують серію точок даних по кількості хворих, записаних через регулярний інтервал, в нашому випадку щодня. Досліджуваний набір даних вміщує наступні змінні: країна, широта, довгота, кількість захворювань по днях починаючи з 22 січня 2020 року до 2021-12-22. На момент дослідження даних, набір вміщує дані по кількості захворювань за 702 дні.



а



б

Рис. 1. Часовий ряд кількості випадків захворювань на COVID-19 в Україні (а) та Індії (б)
(**Fig. 1.** Time series of the number of COVID-19 cases in Ukraine (a) and India (b))

Прогнозування – це визначення величини наступного кроку часового ряду, на якому необхідно передбачити майбутнє значення, яке прийме ряд. Розглянуто два методи навчання моделей: лінійна регресія та градієнтний бустинг.

Для розробки моделі обрано багатофакторне прогнозування часових рядів з використанням у якості предикаторів запізнення часового ряду. Визначено характеристики, що враховуються при навчанні моделі, а саме дата початку події, день тижня, номер

тижня, місяць та інші. Проведено аналіз щодо визначення впливу цих параметрів на якість навчання моделі. Підготовлені дані, згідно з загальними рекомендаціями, поділені у відношенні 5:1 для тренувального набору та тестового, що становить 583 та 119 днів відповідно. Тестовий сет містить дані за період від 31.08.2021 до 27.12.2021.

Прогнози побудованих моделей щодо кількості захворілих на коронавірус з використанням лінійної регресії наведено на рис. 2.

Тренувальний сет візуалізований чорною лінією з колами, а прогнозовані значення моделлю - сірою лінією з трикутниками.

Графіки ілюструють високе співпадіння прогнозованих та реальних даних для України (рис.2, а) та слабке для Індії (рис. 2, б).

Розглянемо результати роботи моделі прогнозу побудованої за допомогою алгоритму градієнтного бустингу. Часові ряди кількості випадків захворювань на COVID-19 в Україні та Індії наведені на рис. 3. Тренувальний сет показаний чорною лінією з колами, а прогнозовані значення за допомогою моделі - сірою лінією з трикутниками. З графіків на рис. 3, а для України можна зробити висновок, що данні моделі суттєво відрізняються від фактичних. Натомість для Індії співпадіння даних тестового набору та модельних значень є значно кращим.

Оцінемо похибки побудованих моделей за допомогою відомих формул для відносних абсолютної та квадратичної похибок відповідно [6] та [7]. Ще одна величина, що дозволяє швидко оцінити якість передбачуваної моделі це точність прогнозу, яку можна розрахувати за формулою:

$$PAC \equiv r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right]\cdot\left[n\sum y_i^2 - \left(\sum y_i\right)^2\right]}}. \quad (5)$$

Розраховані величини похибок та точності прогнозів наданих за допомогою методів лінійної регресії та градієнтного бустингу наведені для обох країн в табл. 1. З отриманих даних можна зробити висновок, що для України краще підходить побудована модель з використанням лінійної регресії, а для Індії – градієнтного бустингу.

Таблиця 1 – **Метрики оцінки похибки та точності побудованих моделей для передбачення кількості захворілих по дням в Україні та Індії**

Країна	Україна		Індія	
Метод	Лінійна регресія	Градієнтний бустинг	Лінійна регресія	Градієнтний бустинг
Відносна абсолютна похибка	0.375505	0.852222	0.625125	0.429781
Відносна квадратична похибка	0.161849	0.887345	0.401214	0.242251
Точність прогнозу	0.838150	0.112655	0.598785	0.757748

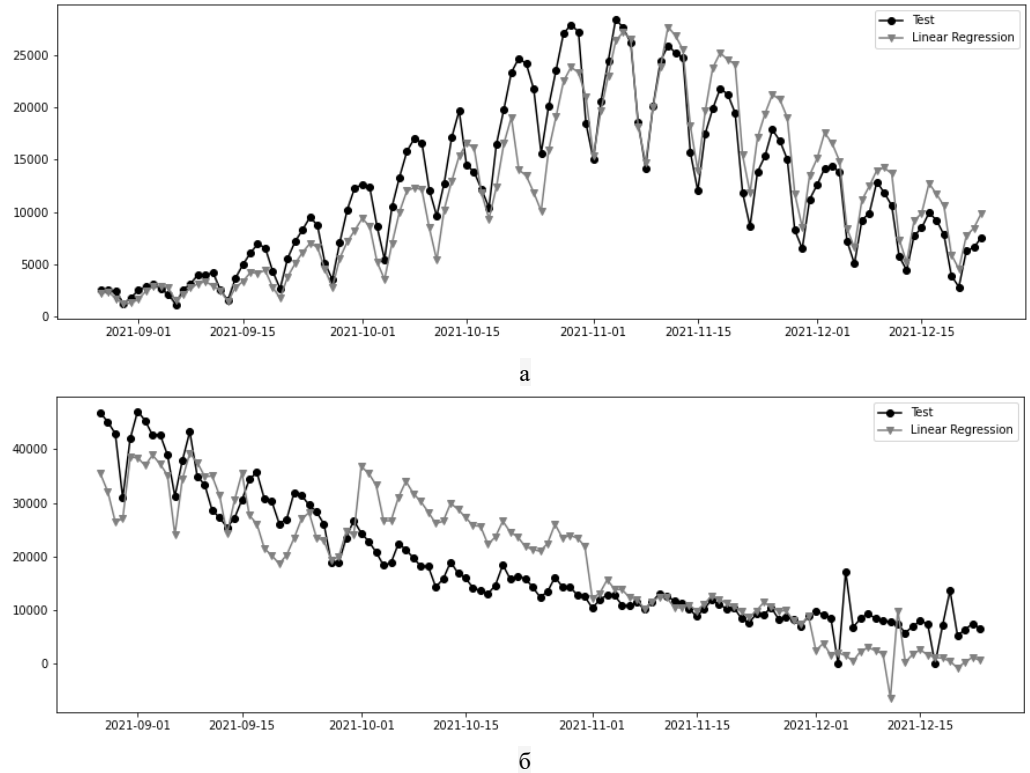
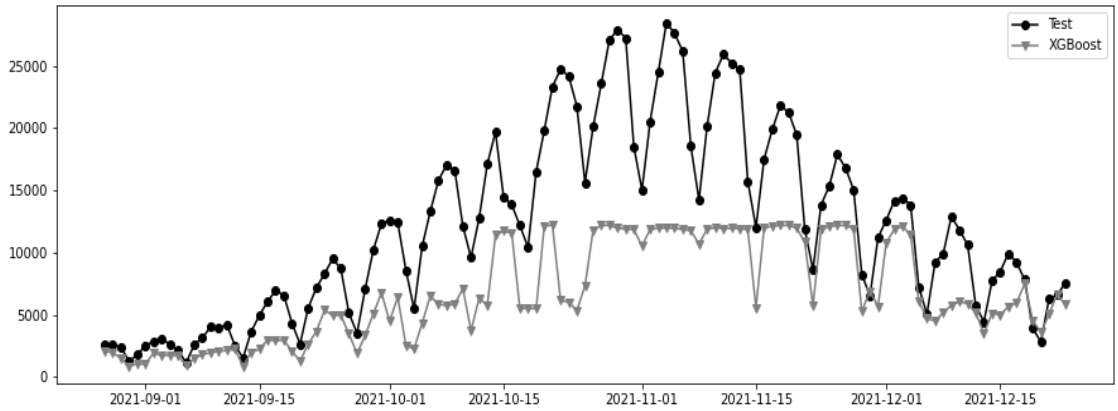
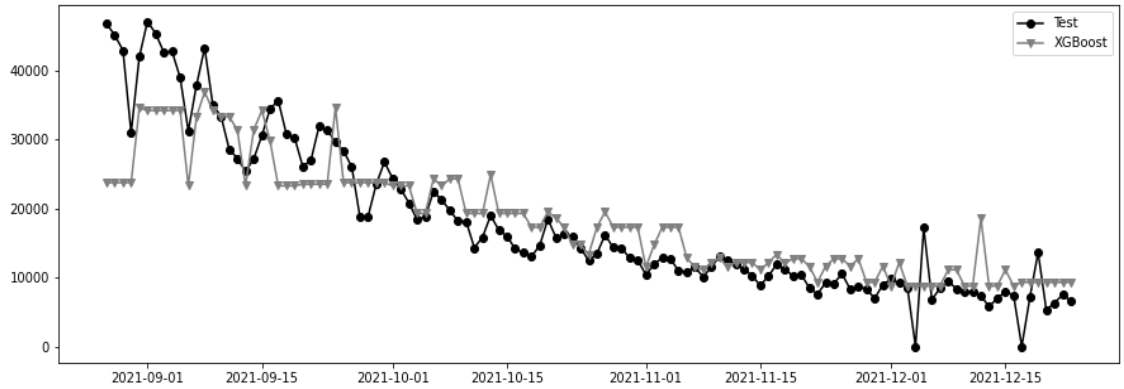


Рис. 2. Часовий ряд кількості випадків захворювань на COVID-19 в Україні (а) та Індії (б), отриманий за допомогою моделі, побудованої з використанням методу лінійної регресії
(**Fig. 2.** A time series of the number of COVID-19 cases in Ukraine (a) and India (b) was obtained using a linear regression model)



а



б

Рис. 3. Часовий ряд кількості випадків захворювань на COVID-19 в Україні (а) та Індії (б), отриманий за допомогою моделі, побудованої з використанням методу градієнтного бустингу
(**Fig. 3.** A time series of the number of COVID-19 cases in Ukraine (a) and India (b) was obtained using a XGBoost model)

Передбачувані моделі для кількості захворілих по дням в Україні та Індії на проміжок часу з 28 грудня 2021 року до 26квітня 2022 року вказують на те, що в обох країнах невдовзі почнеться спад кількості захворілих та невеликі коливання навколо стабільних значень з часом.

Висновки

Побудовані моделі передбачення кількості захворілих на COVID-19 з використанням методів машинного навчання.

Побудовані моделі навчались на даних зібраних з початку епідемії для України та Індії.

Для навчання моделей використані алгоритми лінійної регресії та градієнтного бустингу.

Досліджуваний набір даних вміщує кількість захворювань по днях починаючи з 22 січня 2020 року до 2021-12-22 для різних країн. На момент дослідження даних, набір вміщував дані по кількості захворювань за 702 дні. Визначено характеристики, що враховуються при навчанні моделі: дата початку події, день тижня, номер тижня, місяць та інші.

Проведено аналіз щодо визначення впливу цих параметрів на якість навчання моделі.

Встановлено, що для прогнозування часових рядів з епідеміологічної ситуації в Україні краще підходить алгоритм лінійної регресії, натомість в Індії – градієнтного бустингу.

Оцінені похибки моделей та точність прогнозу з найкращими показниками 83% для України та 75% для Індії.

Побудовані моделі, дозволяють передбачати епідеміологічну ситуацію в майбутньому, координувати дії у різних галузях охорони здоров'я та проводити обґрунтовані превентивні заходи на державному рівні.

Список літератури (References)

1. (2019), *Coronavirus disease 2019 COVID-19* (wikipedia.org), available at: https://en.wikipedia.org/wiki/COVID-19.
2. (2020), *Artificial Intelligence and Machine Learning Trends in 2020*, available at: https://www.dataversity.net/artificial-intelligence-and-machine-learning-trends-in-2020/Artificial Intelligence and Machine Learning Trends in 2020 - DATAVERSITY
3. Kononova, K.Yu. (2020), *Machine learning: methods and models*, V.N. Karazin KhNU Kharkiv, 301 p.
4. Goodfellow, J., Curville, A. and Bengio I. (2018), *Deep Learning*, 654 p.

5. (2021), CSSEGIS and Data/COVID-19**,** available at: https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv
6. (2021), *GeneXproTools 5.0 – Relative Absolute Error*, available at: https://www.gepsoft.com/GeneXproTools/AnalysesAndComputations/MeasuresOfFit/RelativeAbsoluteError.htm
7. (2021), *GeneXproTools 5.0 – Relative Squared Error*, available at: https://www.gepsoft.com/GeneXproTools/AnalysesAndComputations/MeasuresOfFit/RelativeSquaredError.htm

Received (Надійшла) 27.12.2021
Accepted for publication (Прийнята до друку) 13.04.2022

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Хорошун Ганна Миколаївна – кандидат фізико-математичних наук, доцент, доцент кафедри комп'ютерних наук та інженерії, Східноукраїнський національний університет імені Володимира Даля, Сєвєродонецьк, Україна;
Ganna Khoroshun – PhD (Optics and Laser Physics), Associated Professor, Associated Professor of Computer Science and Engineering Department, Volodymyr Dahl East Ukrainian National University, Severodonetsk, Ukraine;
e-mail: an_khor@i.ua; ORCID ID: https://orcid.org/0000-0002-1272-1222.

Рязанцев Олександр Іванович – доктор технічних наук, професор, завідувач кафедри комп'ютерних наук та інженерії, Східноукраїнський національний університет імені Володимира Даля, Сєвєродонецьк, Україна;
Oleksandr Ryazantsev – Doctor of technical sciences, Professor, Head of Computer Science and Engineering Department, Volodymyr Dahl East Ukrainian National University, Severodonetsk, Ukraine;
e-mail: a_ryazantsev@ukr.net; ORCID ID: https://orcid.org/0000-0002-3740-3132.

Коверга Марк Олександрович – аспірант кафедри комп'ютерних наук та інженерії, Східноукраїнський національний університет імені Володимира Даля, Сєвєродонецьк, Україна;
Mark Koverha – PhD student of Computer Science and Engineering Department, Volodymyr Dahl East Ukrainian National University, Severodonetsk, Ukraine;
e-mail: markkoverga@gmail.com; ORCID ID: https://orcid.org/0000-0001-9906-4845.

Покришка Сергій Анатолійович – аспірант кафедри комп'ютерних наук та інженерії, Східноукраїнський національний університет імені Володимира Даля, Сєвєродонецьк, Україна;
Sergey Pokryshka – PhD student of Computer Science and Engineering Department, Volodymyr Dahl East Ukrainian National University, Severodonetsk, Ukraine;
e-mail: xakermans@gmail.com; ORCID ID: https://orcid.org/0000-0002-4092-1221.

Модели машинного обучения для предвидения количества заболевших на COVID-19 в Украине и Индии

Г. Н. Хорошун, А. И. Рязанцев, М. А. Коверга, С. А. Покрышка

Аннотация. Построены модели предсказания количества заболевших COVID-19 с использованием методов машинного обучения. Построенные модели обучались на данных, собранных из разных официальных источников, включая Всемирную Организацию Здоровья, с начала эпидемии до настоящего времени. Для обучения моделей предсказания количества заболевших COVID-19 выбраны Украина и Индия. Методами, которые придали высокую точность прогноза для существующих данных, оказались алгоритмы линейной регрессии для Украины и градиентного бустинга для Индии. Анализ данных проводился с помощью языка программирования Python, с использованием библиотеки Sklearn, построенной на основе SciPy (Scientific Python). Кроме того, использовалась библиотека алгоритма градиентного бустинга XGboost. Для разработки модели выбрано многофакторное прогнозирование временных рядов с использованием в качестве предикаторов запоздания временного ряда. Определены характеристики, влияющие на обучение модели: дата начала события, день недели, номер недели, месяц и другие. Проведен анализ определения влияния этих параметров на качество обучения модели. Оценены ошибки моделей и точность прогноза с лучшими показателями 0.83 для Украины и 0.75 для Индии. Построенные модели позволяют предсказывать эпидемиологическую ситуацию в будущем, координировать действия в различных отраслях здравоохранения и проводить обоснованные превентивные мероприятия на государственном уровне.

Ключевые слова: машинное обучение; модели прогнозирования; метод линейной регрессии; метод градиентного бустинга.

Machine learning models for predicting the number of COVID-19 patients in Ukraine and India

Ganna Khoroshun, Oleksandr Ryazantsev, Mark Koverha, Sergey Pokryshka

Abstract. Models for predicting the number of patients with COVID-19 using machine learning methods have been built. The data for models are obtained from various official sources, including the World Health Organization, from the beginning of the epidemic to the present time. The data in Ukraine and India were selected to teach models for predicting the number of patients with COVID-19. Algorithms of linear regression for Ukraine and gradient boosting for India proved to be the methods that provided high accuracy of the forecast for the existing data. Data analysis was performed using the Python programming language with Sklearn library which is based on SciPy (Scientific Python). In addition, the XGboost gradient boost algorithm library was used. To develop the model, multifactor prediction of time series with the delays as predictors was chosen. It is established that the such characteristics as the date of the event, day of the week, week number, month affect to the model. Model errors are smallest and forecast accuracy were estimated with the best values of 0.83 for Ukraine and 0.75 for India. The built models allow to predict the epidemiological situation in the future, to coordinate actions in different areas of health care and to carry out reasonable preventive measures at the state level.

Keywords: machine learning, prediction models, linear regression method, gradient boosting method.