Informatics in Medicine Unlocked 28 (2022) 100857

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

COVID and nutrition: A machine learning perspective

Nafiseh Jafari [a], Mohammad Reza Besharati [b,*], Mohammad Izadi [b], Alireza Talebpour [c]

[a] Engineering Department, University of Qom, Qom, Iran
[b] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
[c] Computer Science and Engineering Department, Shahid Beheshti University, Tehran, Iran

ARTICLE INFO

Keywords:
Machine learning
Random forest
Multilayer perceptron
COVID-19
Nutrition
Diet
Big data

ABSTRACT

A self-report questionnaire survey was conducted online to collect big data from over 16000 Iranian families (who were the residents of 1000 urban and rural areas of Iran). The resulting data storage contained over 1 M records of data and over 1G records of automatically inferred information. Based on this data storage, a series of machine learning experiments was conducted to investigate the relationship between nutrition and the risk of contracting COVID-19. With highly accurate scores, the findings strongly suggest that foods and water sources containing certain natural bioactive and phytochemical agents may help to reduce the risk of apparent COVID-19 infection.

1. Introduction

The Sars-Cov-2 pandemic (COVID-19) is a global crisis that has caused widespread devastation. Numerous researchers have attempted to address its various facets since it first surfaced. In computer engineering, machine learning is a prominent method of providing data-driven insights into newly emerging diseases such as the COVID-19.

Various aspects of this pandemic are data-driven, including infection diagnosis based on CT scans of patients [1,2] or other symptoms [3], infection diagnosis based on metabolomics [4] and serologic data [5,6], epidemiologic analysis [7,8] and predictions [9], viral genetics [10] and host epigenetics studies [11], evolutionary path discovery [12], contact tracing [13] and quarantine enforcing [13], and numerous other aspects [14].

An observational study was conducted to ascertain the relationship between families' dietary nutrition regimens and their risk of contracting COVID-19 [15]. To this end, an online self-report questionnaire survey was conducted to collect data from over 16000 Iranian families (residents of 1000 urban and rural areas of Iran). The resulting data storage contained over 1 M records of data and over 1G records of automatically inferred information. Based on this data storage, a series of machine learning experiments was conducted to investigate the relationship between nutrition and the risk of contracting COVID-19.

2. Data collection

The resulting data storage includes some records regarding the effects of lifestyle factors (e.g., nutrition, water consumption sources, physical activity, smoking, age, gender, ethnic origin, health and disease factors, and a variety of other factors) on COVID-19 infection status in families (i.e., the residents of a home). These items combine to form a collection of 125 features (84 features for the nutrition state of the family). Phase 1 collected 11K completed questionnaires until the end of Mordad (July–August). Following that, an additional 5K completed questionnaires were added until Day (December), bringing the total to over 16K completed questionnaires in Phase 2. A subset of the research data is available in Ref. [16].

3. Data preprocessing

All incomplete or blank records were discarded (less than 3% of the total data). An object-oriented model for data processing was designed and implemented in Java. This Java code generated the required CSV tables for machine learning experiments.

4. Hyperparameter optimization

A greedy parameter optimization algorithm was used to calculate the best window size for running averages (Fig. 1). Running averages let us

* Corresponding author.
E-mail addresses: nafisehjafari75@gmail.com (N. Jafari), besharati@ce.sharif.edu (M.R. Besharati), izadi@sharif.edu (M. Izadi), talebpour@sbu.ac.ir (A. Talebpour).

https://doi.org/10.1016/j.imu.2022.100857
Received 2 October 2021; Received in revised form 12 January 2022; Accepted 12 January 2022
Available online 19 January 2022
2352-9148/

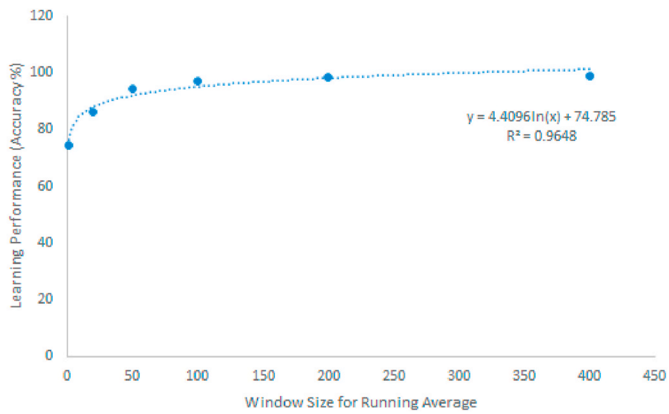

Fig. 1. Learning Performance vs. Window Size for the Running Average (an averaging filter for inputs).

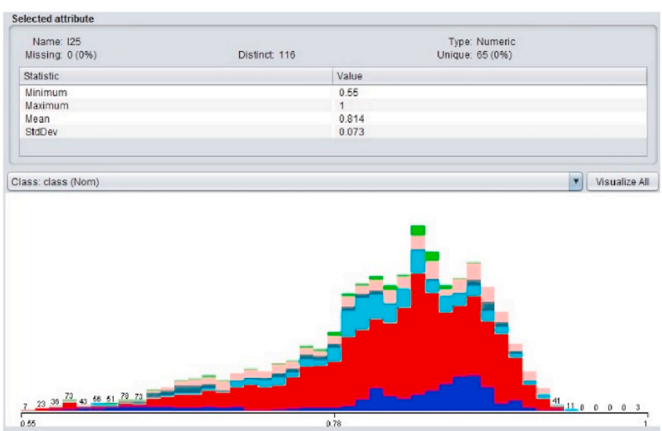

Fig. 2. Histogram for Feature 25 with Class-Tag Coloring (Daily Tea Drinking as a habit in lifestyle). The greater value indicates micro-communities with a higher prevalence of tea consumption.

transform discrete data to continuous space data for micro-communities [24] (Fig. 2).

5. Experiments and results

Weka was used as the primary platform, running on a Corei7-equipped PC. The results of twenty experiments (Tables I–II indicated that the accuracy rate was acceptable. Numerous classification algorithms have been evaluated. The random forest algorithm [17] and the multilayer perceptron algorithm [18] both performed better in terms of accuracy. According to calculations on billions of permutations of nutrition conditions and dietary regime items using data from people's diets and infection status, many dietary conditions significantly reduced the risk of apparent COVID-19 infection by 90%. In comparison, certain dietary factors increased risk by a factor of three or more. The findings indicate that certain diets may have a protective effect against COVID-19-related death (Fig. 3) (see Table 3).

An ID3 algorithm [19] (with 2540 instances of data and 9 features) was executed on Colab, and a decision tree was developed for several essential features with a Gini coefficient of 0.5 (Fig. 4).

The Appendix contains some of the observed results (for Phase 1 until Mordad for 11000 families). The researchers could obtain additional information about the data [16] or submit a request.

6. Metabolites experiments

Nutrition and lifestyle factors can affect the blood serum metabolite profile. Thus, metabolite analysis is a technique for examining the relationship between nutrition and the COVID-19. This section analyzed metabolomics data from a Chinese study (in Wuhan) [20], which included 430 metabolite features for 96 blood tests on 44 samples (including healthy, moderate, severe, and fatal COVID-19 cases). As a result, 96 instances with 430 features were available to analyze the relationship between blood metabolites and the status and severity of COVID-19 infection. Additionally, five data experiments were conducted in this section (with 10-fold cross-validation). The results indicated that precision and accuracy were nearly 90%, and the ROC was approximately 0.99 (see Table 3).

The J48 algorithm's decision tree indicated that the key control variables "death" and "survival" in severe COVID-19 cases were the blood level of T3 thyroid hormone (see Fig. 5). This finding corroborates the research results of several previous studies [21,22].

Table 1
Results of random forest with 10-fold cross-validation.

Random Forest	Window Size For Running Average (Averaging Filter)	# Of Features	# Of Instances	# Of Classes	Accuracy %	Time (Computational Complexity)
EXP-1	1	9	2540	4	67	20 seconds
EXP-2	20	9	2540	4	47	20 seconds
EXP-3	20	83	16227	4	85.17	2 minutes
EXP-4	20	122	16227	4	86.31	5 minutes
EXP-5	1	125	16227	2	87.39	5 minutes
EXP-6	1	125	16227	4	74.35	5 minutes
EXP-7	20	125	16227	4	86.40	5 minutes
EXP-8	50	125	16227	4	94.33	5 minutes
EXP-9	100	125	16227	4	96.96	5 minutes
EXP-10	200	125	16227	4	98.18	5 minutes
EXP-11	400	125	16227	4	99.04	5 minutes

Table 2
Results of multilayered perceptron with 10-fold cross-validation.

Multilayer Perceptron	Window Size For Running Average (Averaging Filter)	# Of Features	# Of Instances	# Of Classes	Accuracy %	Time (Computational Complexity)
EXP-12	1	9	2540	4	71*	1 minutes**
EXP-13	20	9	2540	4	37	10 minutes
EXP-14	20	83	16227	4	81.26	2 hours
EXP-15	20	122	16227	4	76.00	3 hours
EXP-16	1	125	16227	2	84.51	3 hours
EXP-17	1	125	16227	4	67.25	3 hours
EXP-18	20	125	16227	4	76.43	3 hours
EXP-19	50	125	16227	4	92.22	3 hours
EXP-20	100	125	16227	4	94.99	3 hours

* Deep Neural Network.
** Using Colab.research.google.com.

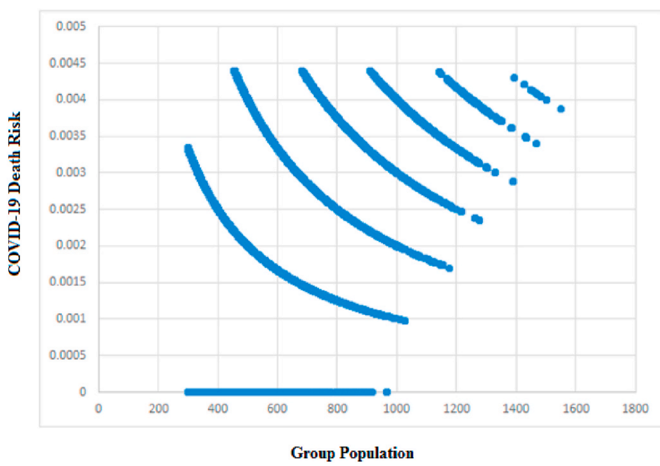

Fig. 3. The above diagram was plotted for the citizens of Tehran in the research dataset for 330K dietary conditions associated with a reduction in the risk of COVID-19. Each point represents a distinct group of dietary conditions, and each condition is further subdivided into four subparts (e.g., daily coffee consumption, daily dairy consumption, weekly consumption of fish, and high consumption of fast foods).

7. Dietary experiments of countries

On a broader scale, differences exist between countries regarding nutrition diets and COVID-19 statistics. This study conducted some classification experiments using the dataset provided by Ref. [23]. The first 99 countries with a high COVID prevalence were classified into 46

Table 3
Results of metabolites data experiments.

	Task	Classification Algorithm	Precision %	Recall %	ROC
EXP-M1	COVID-19 Fatality Prediction	J48	85	78	0.84
EXP-M2	COVID-19 Fatality Prediction	Dl4jMlp (Deep Neural Network)	86	77	0.88
EXP-M3	COVID-19 Fatality Prediction	Multilayer Perceptron	90	97	0.989
EXP-M4	COVID-19 Fatality Prediction	Logistic Regression	90	97	0.994
EXP-M5	COVID-19 Fatality Prediction	Random Forest	82	100	0.98

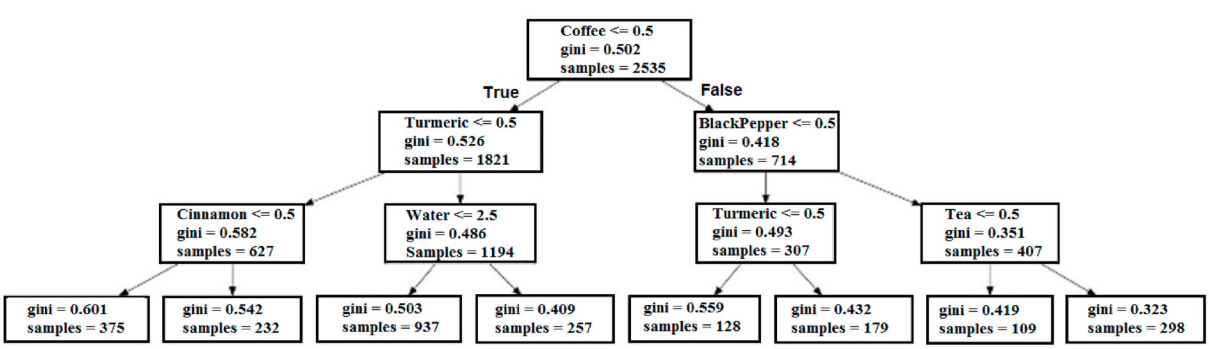

Fig. 4. ID3 for 2540 instances of data with 9 features.

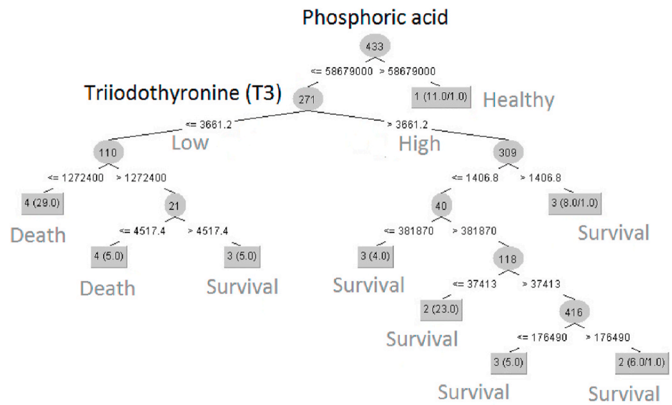

Fig. 5. The J48 algorithm's decision tree suggests that the key control variables for "death" and "survival" in severe COVID-19 cases were the level of T3 thyroid hormone in the blood.

Table 4
Results of dietary data experiments results.

	Task	Classification Algorithm	Window Size for Running Average	Accuracy %
EXP-C1	COVID-19 Mortality Rate Prediction	Random Forest	1	64.65
EXP-C2	COVID-19 Mortality Rate Prediction	Random Forest	10	92.3

countries with a high COVID-19 mortality rate and 53 countries with a low COVID-19 mortality rate. The classification algorithms were validated with 10-fold cross-validations using 31 nutritional and dietary features. The reported findings show a strong correlation between countries' nutritional/dietary states and their COVID-19 mortality rates (see Table 4).

8. Conclusion

A comprehensive questionnaire survey was conducted with over 16000 Iranian families to collect data (the residents of more than 1000 different urban cities and rural areas of Iran). The survey resulted in the creation of big data of COVID-19 and lifestyle (with more than 1 M of data records and more than 1G of items collected by acquiring semantic entailment rules- for a digest report, see Table 5). The resulting big data set included records about the effect of lifestyle factors (nutrition, water sources, physical activity, smoking, age, gender, health and disease factors, and a variety of other factors) on COVID-19 infection status in families (i.e., the residents of a home). The findings strongly indicated that foods and water sources containing several naturally occurring hypomethylating agents significantly reduced the risk of apparent COVID-19 infection. Overall, the experimental data indicated an acceptable level of accuracy for the relationship between nutrition and Sars-Cov-2 infection. Moreover, computations on billions of combinations of nutrition conditions and dietary regime items indicated that several dietary conditions mitigated the risk of apparent COVID-19 infection.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We would like to express our utmost gratitude to Prof. Reza Malekzadeh, Dr. Ehsan Mostafavi, Dr. Alireza Ghasempour, Dr. Hossein Shojaei, Dr. Zahra Tarokh, Dr. Navid Rabiee, Dr. Ehsan Shabani, Dr. Hossein Moradi, Dr. Seyyed Hossein Pourhoseini, Dr. Maryam Hourali, Dr. Zahra Besharati, Faculty Staff of QM Research Institue of Shahid Beheshti University, DisysLab members of Sharif University of Technology, and 16000 contributing Iranian families in the online questionnaire survey.

Appendix

Table 5
a digest of results.

Factor in Family LifeStyle	Observed COVID-19 apparent Infection Risk Change (in %)	Relative Risk[1] (RR)	Statistical Significance due to 99.9% Confidence Interval	Statistical Significance due to 95% Confidence Interval	Number of Questionnaires
Turmeric	-87	0.45	Yes	Yes	More Than 11K
Black pepper	-65	0.51	Yes	Yes	More Than 11K
Islamic Fasting for entire Ramadan	-61	0.55	Yes	Yes	About 10K
Cinnamon	-59	0.55	Yes	Yes	More Than 11K
Legume and chickpea	-52	0.55	Yes	Yes	About 5K
Dark chocolate, dark cocoa	-50	0.54	Yes	Yes	About 3K
Bell pepper	-48	0.59	Yes	Yes	More Than 11K
Tea	-48	0.66	Yes	Yes	More Than 11K
Sea salt	-48	0.57	Yes	Yes	About 10K
Vitamin D or Multivitamin Tablets	-46	0.62	Yes	Yes	More Than 11K
Walnuts or Nuts	-46	0.62	Yes	Yes	More Than 11K
Consuming Rose Water Once a Few Days In Food Or Drink	-46	0.60	No	Yes	About 5K
High consumption of apple juice or apple	-14	0.86	No	No	About 5K
Home water purification devices	21	1.23	No	Yes	More Than 11K
High consumption of deep frying or fried foods	24	1.25	No	Yes	More Than 11K
Soft drinks and soda	24	1.25	No	No	About 3K
High consumption of sugar	26	1.27	No	No	About 3K
Monthly consumption of fish meat or seafood	27	1.31	Yes	Yes	More Than 11K
Weekly consumption of fish meat or seafood	28	1.30	Yes	Yes	More Than 11K
High consumption of sweet pepper (not bell pepper, excluding bell pepper)	37	1.37	No	Yes	About 10K
Eat fish meat or seafood once every two or three days	42	1.44	Yes	Yes	More Than 11K
High consumption of fast food	45	1.46	Yes	Yes	More Than 11K
Pumpkin	49	1.5	No	Yes	About 3K
Sugar substitute, artificial sweeteners	60	1.62	Yes	Yes	More Than 11K
Fruits natural products: Grape Syrup	-45	0.61	Yes	Yes	About 5K
Daily yogurt consumption	-45	0.64	Yes	Yes	About 10K
Tahini and natural products like it	-44	0.61	Yes	Yes	About 10K
Low or controlled consumption of oil	-44	0.62	Yes	Yes	More Than 11K
Consume Courgette once every ten days	-44	0.59	Yes	Yes	about 5K
Garlic	-42	0.65	Yes	Yes	More Than 11K
Consume Eggplant once every ten days	-42	0.64	Yes	Yes	about 5K
High consumption of fruits and vegetables	-41	0.67	Yes	Yes	More Than 11K
Natural Honey	-41	0.67	Yes	Yes	More Than 11K
Green pea	-38	0.63	No	No	About 5K
Ginger	-36	0.68	Yes	Yes	More Than 11K
Fruits natural products: fruit-roll	-35	0.67	Yes	Yes	About 10K
Local dairy products	-33	0.71	Yes	Yes	More Than 11K
Daily coffee consumption	-33	0.69	Yes	Yes	More Than 11K
Traditional breads (whole-wheat flour)	-32	0.71	Yes	Yes	More Than 11K
Soybean and its products	-32	0.70	No	Yes	About 5K
Head Cabbage	-31	0.69	No	Yes	About 10K
Islamic fasting, once a week	-31	0.69	No	No	About 10K
Vegetarian diet	-29	0.71	No	No	About 10K
Probiotic dairy products	-25	0.76	No	Yes	More Than 11K
Slimming weight loss diet or low-calorie diet	-24	0.77	No	No	About 10K
Non-alcoholic beer	-21	0.79	No	No	About 5K
Physical exercise and walking	-19	0.82	No	Yes	More Than 11K
Tobacco and smoking	-15	0.86	No	No	More Than 11K

References

[1] Ghavami Rassa, Hamidi Mehrab, Masoudian Saeed, Mohseni Amir, Lotfalinezhad Hamzeh, Ali Kazemi Mohammad, Moradi Behnaz, et al. Accurate and rapid diagnosis of COVID-19 pneumonia with batch effect removal of chest CT-scans and interpretable Artificial intelligence. 2020. p. 11736. arXiv preprint arXiv: 2011.

[2] Cai Wenli, Liu Tianyu, Xue Xing, Luo Guibo, Wang Xiaoli, Shen Yihong, Fang Qiang, Sheng Jifang, Chen Feng, Liang Tingbo. CT quantification and machine-learning models for assessment of disease severity and prognosis of COVID-19 patients. Acad Radiol 2020;27(12):1665–78.

[3] Ahamad Md Martuza, Aktar Sakifa, Rashed-Al-Mahfuz Md, Uddin Shahadat, Pietro Liò, Xu Haoming, Summers Matthew A, Quinn Julian MW, Moni Mohammad Ali. A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients. Expert Syst Appl 2020;160:113661. K. Elissa, "Title of paper if known," unpublished.

[4] Halamka John, Paul Cerrato, Adam Perlman. Redesigning COVID 19 care with network medicine and machine learning: a review. Mayo Clin Proc: Innovat Qual Outcome 2020.

[5] Wu Jiangpeng, Zhang Pengyi, Zhang Liting, Meng Wenbo, Li Junfeng, Tong Chongxiang, Li Yonghong, et al. Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results. medRxiv; 2020.

[6] Subudhi Sonu, Verma Ashish, Patel Ankit B, Hardin Charles C, Khandekar Melin J, Lee Hang, Stylianopoulos Triantafyllos, Munn Lance L, Dutta Sayon, Jain Rakesh K. Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. medRxiv; 2020.

[7] García-Ordás, Teresa María, Arias Natalia, Benavides Carmen, García-Olalla Oscar, Benítez-Andrades José Alberto. Evaluation of country dietary habits using machine learning techniques in relation to deaths from COVID-19. In: Healthcare. 8. Multidisciplinary Digital Publishing Institute; 2020. p. 371. 4.

[8] Dun Chen, Walsh Christi, Bae Sunjae, Adalja Amesh, Toner Eric, Lash Timothy A, Hashim Farah, Joseph Paturzo, Segev Dorry L, Makary Martin A. A machine learning study of 534,023 medicare beneficiaries with COVID-19: implications for personalized risk prediction. medRxiv; 2020.

[9] Willette Auriel A, Willette Sara A, Wang Qian, Pappas Colleen, Klinedinst Brandon S, Scott Le, Larsen Brittany, Amy Pollpeter, Brenner Nicole, Tim Waterboer. Using machine learning to predict COVID-19 infection and severity risk among 4,510 aged adults: a UK Biobank cohort study. medRxiv; 2020. 06.

[10] Esmail Sally, Danter Wayne R. DeepNEU: a machine learning stem cell simulation platform for evaluating the impact of loss of function and gain of function mutations in the SARS-CoV-2 genome. 2020.

[11] Glasscock Jarret. RNA and machine learning: a rational design for multidimensional biomarkers. Drug Target Rev 2020.

[12] Derecichei Iulian, Atikukke Govindaraja. Machine learning model to track SARS-CoV-2 viral mutation evolution and speciation using next-generation sequencing data. In: Proceedings of the 11th ACM International Conference on bioinformatics. Computational Biology and Health Informatics; 2020. 1-1.

[13] Narzullaev Anvar, Muminov Zahriddin, Narzullaev Mavlutdin. Contact tracing of infectious diseases using wi-fi signals and machine learning classification. In: 2020 IEEE 2nd International Conference on artificial Intelligence in engineering and Technology (IICAIET). IEEE; 2020. p. 1–5.

[14] Lalmuanawma Samuel, Hussain Jamal, Chhakchhuak Lalrinfela. Applications of machine learning and artificial intelligence for COVID-19 (SARS-CoV-2) pandemic: a review. Chaos, Solitons & Fractals; 2020. p. 110059.

[15] Besharati Mohammad Reza. Nutrition and COVID-19, risk factors and relative risks (version 3). August 15. Zenodo; 2020. https://doi.org/10.5281/zenodo.3986834.

[16] Besharati Mohammad Reza. COVID-19 risk and LifeStyle (Part1-Nutritions). V2. Mendeley Data; 2020. https://doi.org/10.17632/y37mz23vyv.2.

[17] Breiman Leo. Random forests. Mach Learn 2001;45(no. 1):5–32.

[18] Ruck Dennis W, Rogers Steven K, Kabrisky Matthew. Feature selection using a multilayer perceptron. J Neural Network Comput 1990;2(2):40–8.

[19] Quinlan JRoss. Induction of decision trees. Mach Learn 1986;1(1):81–106.

[20] Wu Di, Shu Ting, Yang Xiaobo, Song Jian-Xin, Zhang Mingliang, Yao Chengye, Wen Liu, et al. Plasma metabolomic and lipidomic alterations associated with COVID-19. medRxiv; 2020.

[21] Gao W, Guo W, Guo Y, Shi M, Dong G, Wang G, Ge Q, Zhu J, Zhou X. Thyroid hormone concentrations in severely or critically ill patients with COVID-19. J Endocrinol Invest 2020:1–10.

[22] Agarwal Shubham, Agarwal Sanjeev Kumar. Endocrine changes in SARS-CoV-2 patients and lessons from SARS-CoV. Postgrad Med 2020.

[23] García-Ordás, Teresa María, Arias Natalia, Benavides Carmen, García-Olalla Oscar, José Alberto Benítez-Andrades. Evaluation of country dietary habits using machine learning techniques in relation to deaths from COVID-19. In: Healthcare, 8. Multidisciplinary Digital Publishing Institute; 2020. p. 371. 4.

[24] Besharati, M.R.; Izadi, M. SimulaD: A novel feature selection heuristics for discrete data. Preprints 2021, 2021020260 (doi: 10.20944/preprints202102.0260.v3).