IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/46AI/2025/4637825/
Vol. 2025 (34), Article ID 4637825

Identifying Harmful Food Products Using Machine Learning Methods*

Mateusz Szmyt and Joanna WIŚNIEWSKA

Institute of Information Systems, Faculty of Cybernetics,
Military University of Technology, Gen. S. Kaliskiego 2, 00-908 Warsaw, Poland

Correspondence should be addressed to: Mateusz Szmyt, mtusz1999@gmail.com

* Presented at the 46th IBIMA International Conference, 26-27 November 2025, Ronda, Spain

Abstract

The growing prevalence of diet-related diseases such as obesity, diabetes, and cardiovascular disorders is closely linked to the excessive consumption of unhealthy food products and overeating. Identifying harmful food items based on their chemical and nutritional composition has therefore become an important research direction in public health and artificial intelligence (AI). This study presents an AI-based approach to recognizing unhealthy food products by analyzing their ingredient lists and nutritional profiles. The dataset used in the research was collected from various online sources, including publicly available food databases and manufacturer websites. Several machine learning models were trained and evaluated to classify food products according to their health impact. The algorithms included linear regression, ridge regression, logistic regression, stochastic gradient descent (SGD), multilayer perceptron (MLP) neural networks, passive-aggressive classifier, support vector machines (SVC) with Gaussian and linear kernels, k-nearest neighbors, naive Bayes, decision trees, random forests, gradient boosting, extremely randomized trees (Extra Trees), AdaBoost, and neural network classifiers. Experimental results demonstrated that ensemble and neural-based methods achieved the highest classification accuracy, confirming the effectiveness of AI in supporting nutritional assessment and promoting healthier dietary choices.

Keywords: machine learning models, classification, regression, food products

Introduction

Artificial intelligence is developing at a dizzying pace. When used appropriately, it can facilitate work in many professions, for example, for beginner programmers by generating and translating code, for companies using virtual assistants by enabling them to answer frequently asked questions, or even for editorial staff by creating engaging content. One subfield of artificial intelligence is machine learning, which enables computers to learn from data sets and the patterns within them. Using the information provided, such as user activity history, colors, or other characteristic features, the program can predict and classify certain objects as similar to others. These machine learning properties were used in this work to classify harmful foods based on their composition. This is crucial because unhealthy eating can lead to diseases such as obesity, diabetes, heart and circulatory system disease, osteoporosis, tooth decay, cancer, and mental disorders (see: Żarnowski at al. (2022)).

In this paper, we describe the negative impact of food on human health and analyze how machine learning methods can help identify foods that may be harmful. We consider potentially useful data sources and then conduct a computational experiment to train models. Finally, we compare the obtained results.

Cite this Article as: Mateusz Szmyt and Joanna WIŚNIEWSKA, Vol. 2025 (34) "Identifying Harmful Food Products Using Machine Learning Methods " Communications of International Proceedings, Vol. 2025 (34), Article ID 4637825, https://doi.org/10.5171/2025.4637825

The impact of food on human health

Food manufacturers often save money on their products by using cheaper, though not entirely healthy, substitutes. An example of such a substitute is palm oil instead of sunflower or rapeseed oil. At first glance, the name doesn't seem particularly dangerous, but after reviewing the harmful effects of this ingredient, including those found in an article by Robb-Nicholson (2024), one might conclude that excessive consumption leads to increased cholesterol levels, obesity, and an increased risk of cardiovascular disease. This is just one example, but there are many more similar unhealthy ingredients added to food products. Flavor enhancers, colorings, preservatives, and sweeteners often have a negative impact on human health and well-being.

However, society is increasingly paying attention to this issue and is trying to raise awareness of it. Despite this, the market is flooded with food products with varying compositions, making it significantly more difficult for consumers to make informed purchasing decisions..

Products advertised as healthy alternatives to unhealthy alternatives can often be worse than those considered unhealthy. Diet drinks, for example, are often sweetened with other sweeteners. Instead of sugar, the ingredient list often lists aspartame (known as E951) according to WHO (2023). In 2006, a study conducted on rats by Soffritti et al. (2006) revealed that the sweetener in question is carcinogenic.

One of the most common diseases of the 21st century caused by unhealthy eating habits is obesity. This occurs when body weight is not directly proportional to height. This leads to numerous problems, including impaired body function, increased body fat, and often problems with physical fitness or minor activities. Over the years, the percentage of obesity cases has been rising, and in each case, the increase from 1990 to 2022 has almost doubled. In Poland, the average annual increase is 0.42%, in Europe only 0.28%, while globally it is 0.29%. In Poland, the increase is galloping compared to Europe and the rest of the world. Data from 2022 also shows that one in three residents of the Americas suffers from obesity according to WHO (2024). Obesity leads to cardiovascular diseases, including hypertension, atherosclerosis, coronary artery disease, heart attack, and even stroke. According to the WHO, over 2.5 million people worldwide die each year from diseases related to overweight and obesity. We treat stuffing ourselves with sweets as a stress reliever. We often choose the computer or TV over walks outdoors. Instead of burning calories, we store them, and those calories quickly turn into round numbers on the scale. Even just 150 minutes of moderate activity a week is enough to feel better about ourselves, lose weight, and lower blood pressure.

Diabetes is another deadly disease that, if left untreated, leads to serious complications, including heart, kidney, eye, nerve, and leg disease, and increases the risk of premature death. One cause of diabetes is an unhealthy lifestyle, as is the case with obesity. Once it develops, a healthy diet should be strictly followed to effectively lower blood sugar levels. According to data from diabetesatlas.org, approximately 537 million adults (aged 20-79) worldwide had diabetes in 2021, representing about 10% of the total population. More than 75% of diagnosed diabetes cases were reported in low- and middle-income countries. The number of people with diabetes is much lower in high-income countries. Diabetes caused 6.7 million deaths, or one death every five seconds.

Besides diabetes and obesity, there are other diseases associated with unhealthy eating. Heart and circulatory system diseases are the leading cause of death worldwide. They are characterized by shortness of breath, chest pain, palpitations, fainting, headaches and dizziness, and increased fatigue. Osteoporosis leads to weakening of bone structure and ultimately fracture, and one of its causes is the poor diet discussed in this article.

Dental caries, according to the WHO definition, "is a pathological, localized process of extracorporeal origin, leading to decalcification and proteolytic breakdown of hard tooth tissues. Carious lesions are caused by the action of acids produced from sugars by bacteria present in the oral cavity. In children, the demineralization process in primary teeth, but also in immature permanent teeth, is much faster than in adults" by Hallas et al. (2011) and ProHEALTH Dental (2023). People who consume foods, especially those containing sugars, are at risk. Therefore, to avoid dental caries, you should visit your dentist regularly and brush your teeth at least twice a day, preferably after each meal. Even better, try to avoid foods containing sugar or high amounts of it.

There's no way to completely eliminate the problem of diseases caused by poor nutrition, but you can try to eliminate foods containing harmful ingredients from your daily diet, which have a negative impact on the

functioning of your body. To ensure you're eating properly throughout the day (eliminating foods with unhealthy ingredients from your diet), you could use an app that checks the ingredients of food products and analyzes whether they contain the aforementioned harmful ingredients.

Due to the large number of substances in food products, it's difficult to determine which of them are toxic or have an adverse effect on humans. Furthermore, some ingredients may only be toxic when combined with others. Therefore, it's difficult to write an application that can recognize thousands of ingredients and predict the healthiness of food products based on their ingredients. However, using machine learning methods, one can attempt to divide food into those that are safe to eat (healthy) and those that should be strictly avoided (unhealthy). The result of machine learning methods would be a model that could assign the appropriate label with a certain degree of probability.

The aforementioned app could use the model mentioned above to recognize the healthiness of a food product based on its ingredients. Ideally, the mobile app would have the ability to take a photo of the ingredients, allowing the image-based text recognition algorithm to read the ingredients as an array of ingredients. This input, fed to the app's embedded machine learning model, would then enable it to determine whether the scanned product is safe.

Searching appropriate data sets

To conduct the machine learning process, it is first necessary to obtain a dataset, which, after appropriate preparation, will be treated as input data (see: Cielen et al. (2016)). Ready-made datasets can be found online, but after searching numerous websites, we were unable to find a ready-made data file for the discussed problem. It was necessary to prepare our own dataset based on information found on one of the websites. Data that has not yet been processed is called raw data due to the need for further processing and preparation for machine learning.

The analysis was carried out on products available in Poland. The website czytamyetykiety.pl contains a comprehensive list of products along with ingredients and a health impact index. It is ideal for machine learning, as it provides data based on which supervised learning can be performed. The health index can be used as a label and it assumes three states:

- Green – when the product is fully edible and healthy.
- Yellow – when the product has not been labeled as healthy due to trace amounts of harmful substances, the appropriate amounts of which are not harmful to health.
- Red – when the products are strictly discouraged due to their harmfulness, even in trace amounts.

On the website skladnikisklep.com.pl, you can find a list of products and their ingredients. Unlike the previous one, the health effects are unknown and the product list is small. The products on this site tend to fall into exotic categories. In this case, machine learning would require manually labeling harmful and healthy products (which would be time-consuming) or attempting unsupervised learning due to the lack of explicit labeling.

The pl.openfoodfacts.org website is similar to the first source, but the products found there come from all over the world. Consequently, the ingredient list is often not translated into Polish from the original language, such as German, which significantly complicates data acquisition. This would require an additional step of translating the acquired data, which would require recognizing the original language to translate it into the target language. A significant advantage over the first solution is the availability of a triple health rating scale. These are the Nutri-Score (an indicator of the product's nutritional value using a five-color scale), NOVA (an indicator of the degree and method of food processing), and Green-Score (an indicator of negative impact on the ecosystem). It's worth noting that the products come from all over the world, and the production methods for the same food can vary significantly in different locations. Another advantage of this version is the detailed ingredient list, which highlights allergens and, in some cases, provides the amounts of ingredients in the products (percentages). However, in the case of information processing, it may prove unnecessary, as this information may complicate the analysis of components, requiring additional steps required for data normalization.

Taking into account various sources, the data used in this work was obtained from the website czytamyetykiety.pl due to the predominance of the number and variety of products, the marking of healthy and harmful products, and a simple list of ingredients without taking into account the percentage share in the food product, as well as the unified Polish language on all pages.

Finding data online doesn't end the data preparation process, as the data must be downloaded and then properly prepared to feed the machine learning process. Popular data acquisition methods include:

- Manually transcribing or copying information from the screen into a text editor. This is very tedious work if there is a huge amount of data to be retrieved, but manual transcribing requires accuracy because the transcriber has the information in front of them at all times. If an error occurs in the information, the user can note the discrepancy and decide to correct, delete, or discard it.
- Screen scraping, a technique for taking photos and reading the text from them using additional software that uses OCR technology (recognition of characters contained in the captured image). This method is faster than manual transcribing, but character errors often occur depending on the OCR technology used. Mistakes are common, for example, when distinguishing between the lowercase letter "L" and the number 1, and not all OCR systems support Polish characters.
- Some websites provide APIs (application programming interfaces), which allow data to be retrieved without having to load the entire page. Only by sending the appropriate request to the appropriate address does the server respond with data in the form of a data packet, e.g., JSON. It's important to remember that not every website uses an API, so this method rarely works.
- Web scraping (see: Broucke and Beasens (2018), Mitchell (2018)) is a technique similar to screen scraping, but without the need to take images and extract information from them. Text data is copied directly from the page and saved, for example, to a text file with a .CSV extension. A web scraping script directly references the page from which the data is to be retrieved and requests the return of the page in HTML code, which it then parses without the need to render the entire page. HTML tags, their types, the classes they belong to, and the identifiers they are assigned, are analyzed. Initially, the programmer should know what the script should be looking for. The content of such tags is extracted, which can then be saved to a file or subjected to further checks (e.g., whether an element with a specified identifier or class is found in the array).

Analyzing the data from the websiteczytaetykiety.pl, the best strategy seems to be using the web scraping method due to the lack of API and the disadvantages of the other two methods mentioned above.

Computational experiment

With a ready-made dataset for machine learning, the next step was to select machine learning methods. All supervised machine learning methods available from scikit-learn.org were considered (see also: Géron (2019)).

However, not all algorithms were suitable due to the type and amount of data. Methods that were used to build models include (see: Hearty (2016), Vasilev et al. (2019)):

- linear regression,
- ridge regression,
- logistic regression,
- stochastig gradient descent (SGD),
- perceptron (neural network),
- passive-aggressive classifier,
- support vector machines (SVC) with Gaussian and linear kernels,
- k-nearest neighbors classifier,
- naive Bayes classifier,
- decision trees,
- random forests,
- gradient boosting,
- extremely randomized trees (so-called Extra Trees) by Scikit-learn (2025),
- AdaBoost,
- multilayer perceptron (MLP) neural network.

All algorithms were placed in a structure called a dictionary so they could be easily managed by using the appropriate method name as a key. The machine learning process took about an hour because some methods took longer to build the model than others. The results are presented in table 1.

Table 1: Obtained values of quality indicators of machine learning models

Algorithms	Quality indicators of machine learning models							Learning time [min:sec]
	Accuracy	Precision	Sensitivity	F1-score	R^2	MSE	RMSE	
Linear regression	It does not occur for linear regression				0,329	0,424	0,651	5:47
Ridge regression	0,786	0,793	0,774	0,782	0,535	0,219	0,468	0:22
Logistic regression	0,816	0,819	0,808	0,813	0,597	0,189	0,435	0:15
SGD	0,808	0,809	0,803	0,805	0,590	0,195	0,442	0:14
Perceptron	0,785	0,791	0,771	0,780	0,538	0,219	0,468	0:10
Passive-aggressive classifier	0,781	0,772	0,785	0,778	0,526	0,224	0,473	0:16
SVC with Gaussian kernel	0,665	0,749	0,577	0,604	0,275	0,342	0,584	9:19
SVC with linear kernel	0,777	0,800	0,750	0,768	0,503	0,230	0,480	7:03
k-nearest neighbors classifier	0,770	0,766	0,765	0,764	0,447	0,243	0,493	0:01
Naive Bayes classifier	0,568	0,605	0,639	0,570	0,237	0,516	0,718	0:02
Decision trees	0,777	0,772	0,770	0,771	0,507	0,230	0,479	0:06
Random forests	0,826	0,834	0,810	0,820	0,626	0,177	0,421	0:33
Gradient boosting	0,758	0,773	0,734	0,750	0,464	0,249	0,499	1:11
Extra trees	0,750	0,750	0,750	0,746	0,437	0,259	0,509	0:01
AdaBoost	0,632	0,665	0,600	0,617	0,126	0,388	0,623	0:29
MLP	0,791	0,783	0,796	0,789	0,545	0,214	0,463	7:28

Without ready-made methods from the sklearn library, it would be necessary to create an error matrix for each model and use it to calculate metrics such as accuracy, precision, sensitivity, and F1-score. The MSE error would have to be calculated as the average of the squared sum of the differences between the predicted values and the actual values, and the RMSE would also have to be calculated by taking the square root of the obtained MSE error.

Fig. 1 summarizes the model development time; the lower the value, the better. The SVC model – a support vector classifier with a Gaussian kernel – devoted the most training time. Training this model took almost 10 minutes. The other models were created faster. It's worth noting models such as extremely random trees, k-nearest neighbors, and the naive Bayes classifier, which took a while to train. However, time is not the most important factor if the classifier or regression makes errors when predicting values or categorizing them.

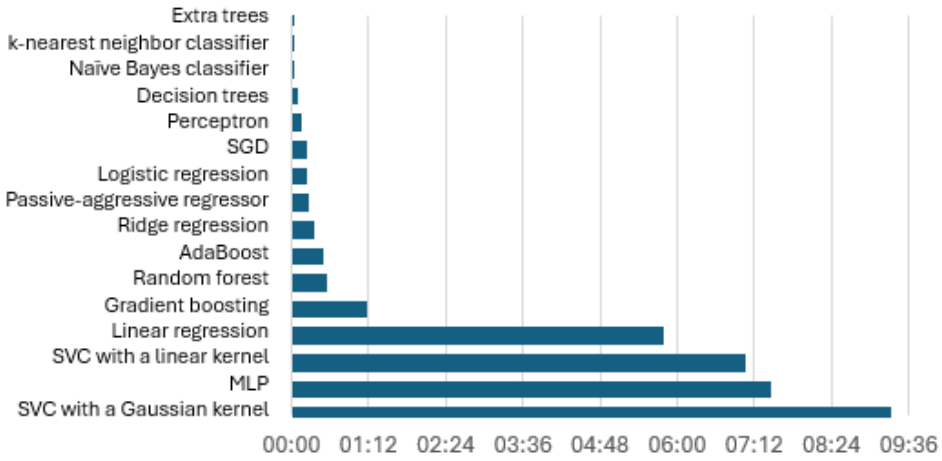

Fig 1. Machine learning model creation times in minutes

Fig. 2 presents a summary of classification quality indicators excluding linear regression, for which it was not possible to calculate the indicators due to prediction, not label classification. The higher the indicator value (bar), the better. For each indicator, the best classifier is the random forest, as the percentage value of each indicator exceeds 80%. Accuracy of 82.6%, precision of 83.4%, and sensitivity of 81% are very good results. The naive Bayes classifier, the support vector machine with a Gaussian kernel, and the AdaBoost method performed the worst. Logistic regression and linear regression with stochastic gradient descent achieved results close to the random forest.

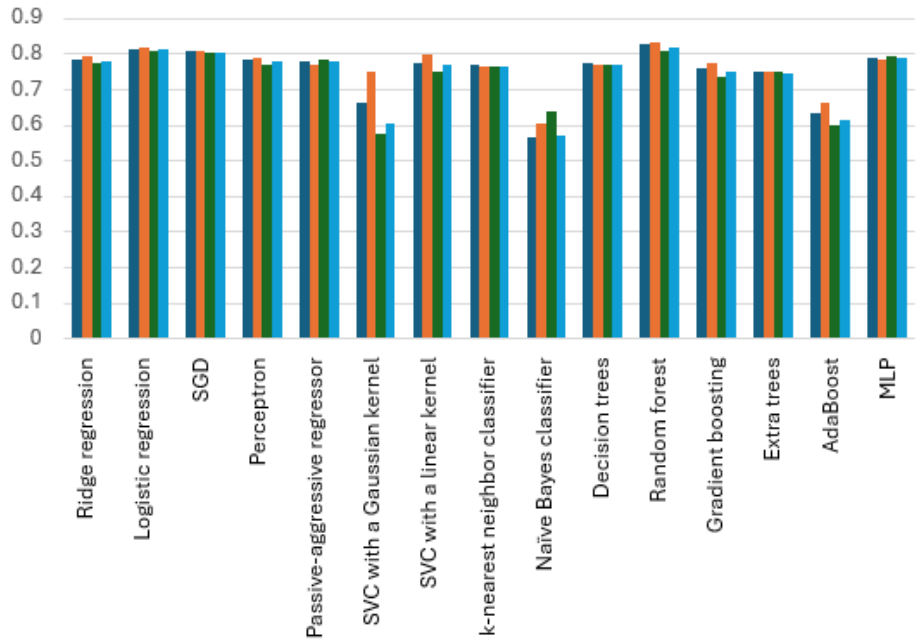

Fig 2. Machine learning classification metrics: Accuracy, Precision, Sensitivity, and F1-score

The graph in Fig. 3 shows the R^2 coefficient of determination values, confirming that the best model was obtained using the random forest method, with a coefficient value of 62.6%. This is a value above half, but a value around 70-80% was expected, given the previous values. In the case of R^2 analysis, the Ada Boost model performed the worst, assigning the explained variables correctly only 12.6% of the time.

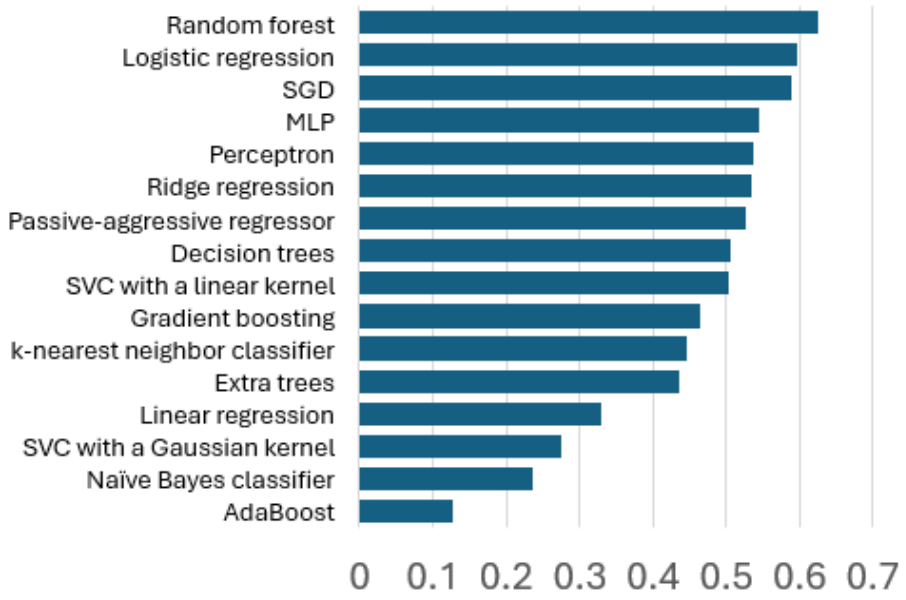

Fig 3. Wartości współczynnika R^2 dla wyuczonych modeli

The graph in Fig. 4 shows that the Random Forest model had the smallest mean error, while the Naive Bayes classifier had the largest mean error.

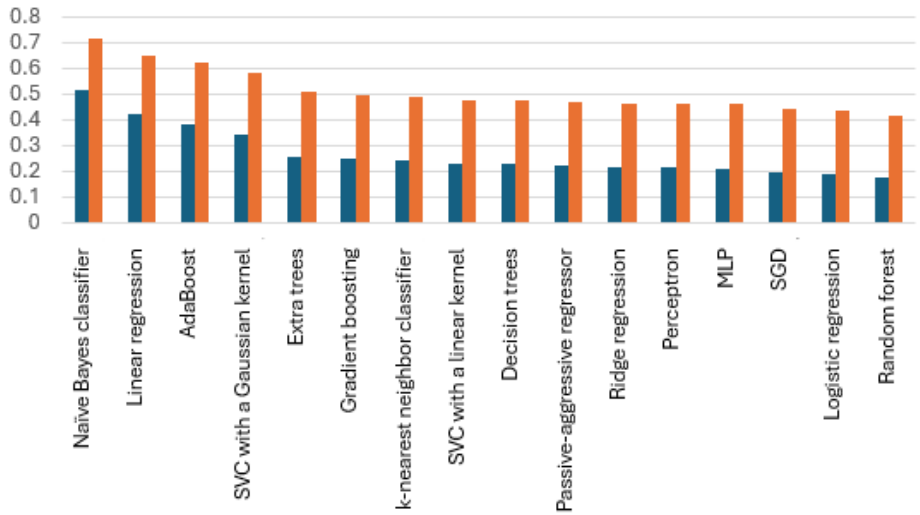

Fig 4. MSE error (blue) and its root RMSE (orange) for used models

Summarizing all the obtained graphs, it can be concluded that the best classification or regression models for the data from the analyzed dataset are random forest, logistic regression and linear regression with stochastic gradient descent (SGD).

Summary and further works

We successfully built and trained multiple machine learning models based on manually collected data from the website. The model obtained using the random forest method was selected, with satisfactory quality indicators. Work on this project took a long time due to the extensive learning of individual machine learning methods, data acquisition and mining methods, and the exploration of various technologies supporting machine learning.

It's possible that the model would be more accurate if additional factors were taken into account, such as the calorie content, sugar, salt, and fat content of the analyzed food products. Also, if the model had precise data on the amount of ingredients (even percentages) contained in the food products, its quality could improve or deteriorate; in such a case, this data would be omitted, as in this case.

The resulting model could be used as part of an application for identifying harmful food based on, for example, a photo of food ingredients. In this case, a new application would need to be written and expanded with a text recognition module for the photo. The input data received, in the form of a list of ingredients, would then serve as input for the model built in this work, while the model would return (with some accuracy) information about the harmfulness or healthfulness of the scanned product.

Another idea is to expand the model with additional data, as a database of 16,613 items was used for training, even though there were initially more items. This is because some ingredients may only appear in a single product, and therefore the model may not recognize the harmfulness of that ingredient when it encounters one during classification.

It would also be necessary to check the behavior of the collected classification and regression models for the population drawn using the train_test_split function using a different value for the random seed than the value used in this work.

Acknowledgment

The work was financed by the Military University of Technology in Warsaw, Poland as part of the project No. UGB 531-000023-W500-22.

References

- Broucke, S. V. and Beasens, B. (2018) ‘Practical Web Scraping for Data Science’. APRESS, New York.
- Cielen, D., Meysman, A. D. B. and Ali, M. (2016) ‘Introducing Data Science’. Manning Publications, New York.
- Géron, A. (2019) ‘Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow’. O’Reilly Media, California.
- Hallas, D., Fernandez, J., Lim, L., Carobene, M. (2011) ‘Nursing strategies to reduce the incidence of early childhood caries in culturally diverse populations’. J Pediatr Nurs., 26: 248–256.
- Hearty, J. (2016) ‘Advanced Machine Learning with Python’. Packt Publishing, Birmingham.
- Mitchell, R. (2018) ‘Web Scraping with Python’. O’Reilly Media, California.
- ProHEALTH Dental (2023) ‘How Poor Nutrition Affects Your Oral Health’, [Online, Retrieved November 20, 2025], https://www.phdental.com/oral-health-news/2023/april/how-poor-nutrition-affects-your-oral-health
- Robb-Nicholson, C. (2024) ‘By the way, doctor: Is palm oil good for you?’, [Online], Harvard Health Publishing, [Retrieved October 22, 2025], https://www.health.harvard.edu/staying-healthy/by-the-way-doctor-is-palm-oil-good-for-you
- Scikit-learn (2025) ‘ExtraTreesClassifier Documentation.’ [Online]. [Retrieved October 22, 2025]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html
- Soffritti, M., Belpoggi, F., Degli Esposti, D., Lambertini, L., Tibaldi, E., Rigano A. (2006) ‘First experimental demonstration of the multipotential carcinogenic effects of aspartame administered in the feed to Sprague-Dawley rats’. Environ Health Perspect. 2006 Mar;114(3):379-85. doi: 10.1289/ehp.8711. PMID: 16507461; PMCID: PMC1392232.
- Vasilev, I., Slater, D., Spacagna, G., Roelans, P. and Zocca, V. (2019) ‘Python Deep Learning’, 2nd ed. Packt Publishing, Birmingham.
- WHO (2023) ‘Aspartame hazard and risk assessment results released’, [Online], World Health Organization, [Retrieved October 22, 2025], https://www.who.int/news/item/14-07-2023-aspartame-hazard-and-risk-assessment-results-released
- WHO (2024) ‘Age-standardized prevalence of obesity among adults (18+ years)’, [Online], World Health Organization, [Retrieved October 23, 2025], https://data.who.int/indicators/i/C6262EC/BEFA58B

- Żarnowski, A., Jankowski, M., and Gujski M. (2022) ‘Public Awareness of Diet-Related Diseases and Dietary Risk Factors: A 2022 Nationwide Cross-Sectional Survey among Adults in Poland’, Nutrients. 2022 Aug 11;14(16):3285. doi: 10.3390/nu14163285. PMID: 36014795; PMCID: PMC9416498.