Multivariate Analysis and Comparison of Machine Learning Algorithms: A Case Study of Cereals of America

Rashika Gupta[1](✉), E. Lavanya[1], Nonita Sharma[2], and Monika Mangla[3]

[1] Electronics and Communication Engineering (AI) Department, Indira Gandhi Delhi Technical University for Women, Delhi, India
gupta.rashika@gmail.com

[2] Information Technology Department, Indira Gandhi Delhi Technical University for Women, Delhi, India
nonitasharma@igdtuw.ac.in

[3] Department of Information Technology, Dwarkadas J. Sanghvi College of Engineering, Mumbai, India

Abstract. This research work aims to analyze the nutritional value of different cereals available in the market through various machine learning models. This analysis is supplemented with the visualization of data also for enhanced understanding. This understanding enables users to devise market strategies as they are competent to evaluate quality of each product and thus its reception in the market. The works starts with statistical analysis through of the data through various plots which provides insight of the data. Further authors perform a comparative analysis of different cereals based on various parameters. This analysis helps to determine the best cereal according to our requirements. The authors have implemented machine learning models on the data to predict the vitamins of any cereal based on their nutritional value. The implementation of various models viz. Linear regression, decision tree, logistic regression, random forest, and KNN advocates the efficacy of various machine learning models to the given problem.

Keywords: Feature selection · linear regression · KNN · random forest · logistic regression · decision tree

1 Introduction

All living organisms have been relying on various sources of energy for their life. Among the different sources of energy, whole grain cereals are one of the most sought-after sources of energy and nutrition. These cereals provide various nutrients viz. Protein, vitamins, fiber, and antioxidants. As there are plenty of cereals available, it is important to analyze the nutritional value of each cereal to select the one as per users' requirements. In this paper, authors have attempted to analyze the nutritional value of various cereals through different machine learning models. For the same, authors have considered the

S. Nandan Mohanty et al. (Eds.): ICISML 2022, LNICST 471, pp. 262–271, 2023.
https://doi.org/10.1007/978-3-031-35081-8_21

online available dataset. Ahead of application of machine learning models, authors have performed data preprocessing which extracts the principal features of the considered dataset. The preprocessing involves the detection of null values if any and then dropping those values [1, 21, 23]. This may also include data normalization to bring the entire data in same range so that it becomes meaningful to compare that data. Further, it involves detection of outliers as outliers negatively affect the statistical and mathematical analysis of a machine learning algorithm leading to lower precision and accuracy. Thereafter, various statistical tools and scatter plots may be used to illustrate the data in a graphical manner.

Further, various machine learning models namely linear regression, K-Nearest Neighbors, Random Forest, Decision Trees, and Logistic Regression are implemented [2, 3, 24]. Current work is significant as it analyzes the cereals manufactured by different companies. The obtained information can be used by consumers to make an informed decision about their choice of cereal according to their dietary requirements [4]. It can also be used for strategizing the marketing and researchers' policies of the manufacturers to lead in this competitive scenario [5, 6]. Hence, the implementation of machine learning can also be used in this domain of nutrition and agriculture [7, 22] in addition to several other problems.

Here, the main contribution of the authors is that authors strongly recommend that empirical evaluation and statistical analysis of the data must be performed ahead of application of machine learning models. Employment of rigorous statistical analysis enables to efficiently select the appropriate machine learning model which is cost effective in terms of time and resources.

Current research work is organized in various sections. Section 1 gives a brief introduction to the cereals' nutritional value. The related work by different authors has been presented in Sect. 2. Section 3 is dedicated to the methodology and the results are discussed in Sect. 4. Finally, the conclusion is given in Sect. 5.

2 Related Work

Several researchers have attempted to work in this direction to determine the nutritional value of different cereals as there are plenty of cereals available viz. Oats, barley, millet, sorghum, triticale, etc. [8]. All these cereals provide a huge range of vital nutrients namely protein, fiber, minerals, carbohydrates, etc. Among various crops, wheat and rice constitute around 50% of the total production. All cereals consist of an embryo containing the material for production in the new plant. Here, the authors claim that it is very important to store cereals in the required conditions to maintain their nutritional value. Additionally, storage time also impacts the nutritional value i.e. storage for a shorter period leads to less change in nutritional value and vice versa. Further, the nutritional value will be influenced by the milling operation. Here, it depends on the extent to which covering bran and aleurone are taken off. As per the authors in [8], intensive research is required in this field as cereals are the most common source of daily needs. It is also known that regular consumption of cereals may prevent chronic diseases. Although the mechanism that helps to prevent such diseases is not clear.

Further, authors in [9] also believe that cereals have a significant contribution by declaring that although cereals provide a widerange of nutrients, processing may impact

their nutritional value. Authors in [9] primarily review the literature that discusses the impact of processing and their mixing with other food items on nutritional value. This study may be helpful to a health professional to give guided advice to their patients. Authors in [10] have presented the significance of barley, known as the fourth floor. Authors claim that barley has high benefits for obesity, diabetes hypertension, etc. owing to its dietary fibers. Additionally, barley is also a good source of starch, minerals, and vitamins and thus can be considered a complete food. Here, the authors present the macro constituents and micro constituents of barley. Its prebiotic effects and nutritional value is also discussed. Further, the authors also discuss required cultivars to preserve their nutritional value.

Authors in [11] perform a comparative analysis of barley with rye, rye wheat, wheat, and oats. For the same, the authors perform a feeding test with mice. Further, a dyebinding (DBC) method is also used to estimate lysine. The focus of the study is to develop high-lysine barleys and high-temperature drying processes for livestock feeds of cereals. DBC method was successful to recover a high-lysine barley line. Further, the authors also discussed the issue of grain deterioration and measures to prevent it. The effect of heat on lysine during various stages namely storage, baking, drying, etc. are also discussed. The study achieves to develop an effective high-temperature drying method that has been adopted in Sweden.

The research is carried forward by authors in [12] by studying the Kodo millet, an ancient grain that grows in arid areas. Kodo millet is very rich in fiber and various minerals. However, processing it may lead to a reduction in nutritional value. In India, several traditional food items are prepared from Kodo millet by blending it with other cereals to enhance its nutritional value.

Additionally, as per the authors in [13], there is a consistent pressure to increase food production in response to population growth. It can be achieved by increasing production capacity or changing consumption habits. For instance, it becomes very difficult to cater to the rising demand for staple cereals like wheat, rice, maize, etc. If the consumption pattern shifts towards cereals like millet and sorghum, it will be highly satisfying as these cereals can be grown in adverse conditions also. Further, these cereals also have a high nutritional value similar to popular staple foods.

Thus, it becomes evident from the research by various researchers that we need to take conscious efforts to cater to the rising demand for cereals without compromising the nutritional demands of the persons.

3 Methodology

This section demonstrates the process and various methods used for the analysis of cereals using various machine learning models and data visualization [21]. For implementation, dataset is collected from Kaggle which is compiled by Mr. Chris Craw-ford. The considered dataset has 77 rows and 17 columns. Figure 1 demonstrates the complete methodology followed in the research work. During preprocessing, all null values are handled. The data normalization handles the significant diversity among range of data.

It also handles outliers to remove the noise if any [14]. Further, to find the correlation among various attributes correlation matrix can be obtained that is illustrated in form of heat map. This correlation enables to determine the principal components. Through statistical analysis, authors illustrates the range of data, the maximum, minimum, and average of each attribute, and a comparative analysis of different attributes [15, 23].

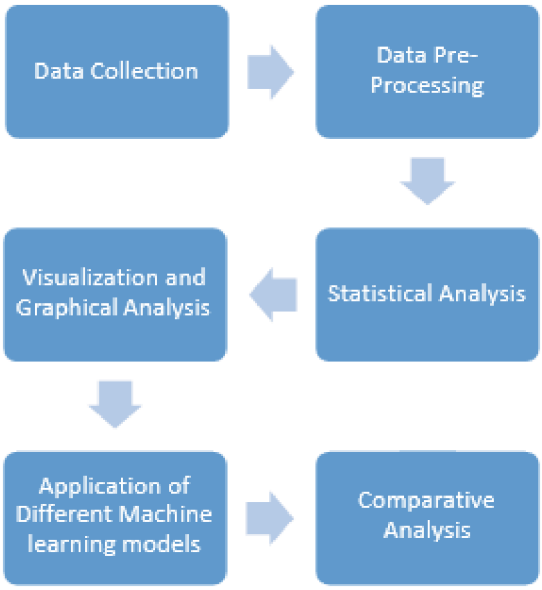

Fig. 1. Methodology of the project.

Here, vitamins is the dependent variable. The split of training and testing data is 80:20. Further, 5 machine learning algorithms are applied namely Linear regression [17], Logistic regression, KNN, Decision trees, and Random Forest. In logistic regression, the maximum and maximum iterations are chosen as 12000 [18]. During kNN, the authors have applied the criteria of Minkowski distance of order 2 [19]. Random forest can be considered as a model containing multiple decision trees, which are applied to numerous subsets of the dataset and the average value is taken thereby improving the predictive precision and accuracy of the respective dataset, to produce a more efficient model and analysis [19]. Instead of entire operation being carried by one decision tree, this model uses the prediction from all of the random trees and based on the majority votes of predictions, the final output is predicted. The parameter used in this model is entropy and the number of fragments of decision trees in this model is taken as 20.

Further, decision tree splits the data according to a certain parameter till maximum precision and accuracy are obtained. The tree mainly comprises two constituents, leaves (which are the decision or the outcome) and decision nodes (which are the joints where data is being split), according to the chosen parameter**.** Two types of criteria can be used-entropy and Gini, and the authors have used the parameter entropy with maximum depth (number of nodes) as 5. This is used to reduce the randomness of the prediction in each step, till we get a minimum level of randomness while preventing overfitting [19].

3.1 Comparative Analysis

The authors have analyzed the different models and found the precision, accuracy, recall, F1 score, and support of each model. The mathematical formulation for these metrics is as follows [20]:

$$Accuracy = \frac{TP + TN}{TP + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1\,Score = \frac{2X\ Recall \times Precision}{Recall + Precision} \tag{4}$$

4 Results and Discussion

The authors have performed multivariate analysis of cereals using diverse methods of statistical and visualization analysis. It is followed by application of machine learning models, advanced predictive analysis.

4.1 Data Analysis and Visualization

Authors have used various tools for data analysis which gives insight about the data. The correlation among various attributes is illustrated in Fig. 2. Correlation is the extent or degree to which any two random variables are linearly related to each other. The correlation coefficient is denoted by r which always lies between -1 to 1 that indicates maximum negative correlation and maximum positive correlation respectively. From Fig. 2, it is evident that Sugar and rating are most negatively correlated while weight and calories are most positively correlated.

Figure 3 depicts a multivariate pair plot. Four different attributes-protein, vitamins, carbohydrates, and fiber have been chosen and plotted against each other, to determine the calories. When an attribute is plotted against itself, we obtain a sinusoidal waveform graph of the normal distribution, while a scatter plot is obtained in all other cases. The variation in the range of the calories has been color-coded, and the scale has been provided. The graph concludes that when carbohydrates are plotted on the dependent axis and the vitamins are plotted on the independent axis, then the calories are minimum when carbohydrates are 10 and vitamins are 0, and the maximum when the carbohydrates are 15 and vitamins are 100. Similarly, in the graph between fiber across the dependent axis, and the protein on the independent axis, the maximum value of calories is obtained when both protein and fiber lie between 2.5 and 5.0, while it obtains its minimum value when protein lies between 2.5 and 5, while fiber is greater than 12.5.

Figure 4 depicts a joint plot of fat and protein with fat on the dependent axis and protein on the independent axis. It is a combination of the scatter plot and bar graphs in the samegraph. The scatter plot is used for finding the extent of correlation and the distribution of the data, while the bar plot gives the variation in the quantity of the attribute. The graph gives the interpretation that protein and fat are positively correlated. Further, Fig. 5 shows a boxplot of the potassium content of the cereals. This is used for detecting outliers, which is a record that lies at an abnormal distance from the remaining data set.

Fig. 2. Correlation matrix for the data

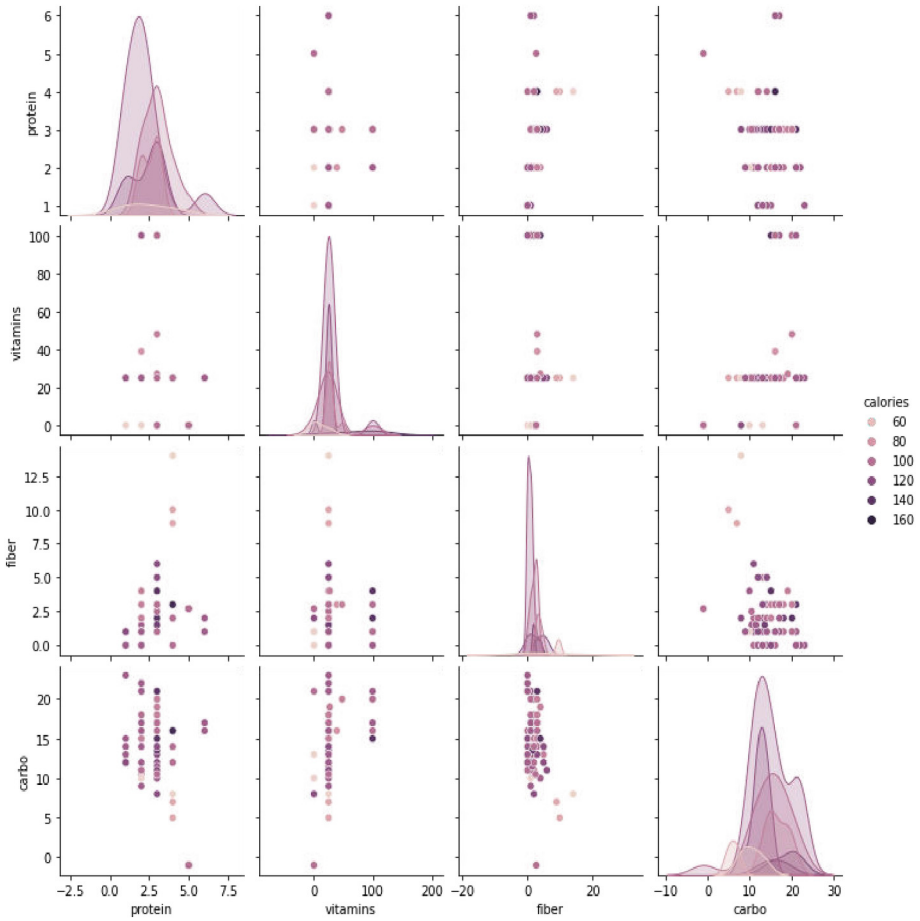

Fig. 3. Multivariate analysis for cereals

Authors have found which quantity of fat is present in the highest count of the cereals, and can conclude that 1 mg of fat is the amount of fat present in maximum cereals. Through the scatter plot, frequency distribution and the scattered plot distribution of protein versus fat, and type versus fat can be analyzed. Displayed pair grids and cross grids for the visual representation of the entire dataset. This visualization greatly reduces the complexity of studying and understanding the data. Authors have also made relational plots to show the variation between the manufacturer and fat content in two different types of cereals as shown in Fig. 4. Authors have also made a joint plot along with the regression, which is useful for machine learning. The frequency distribution is analyzed cereals using a histogram Further, outliers can be detected using the boxplot as illustrated in Fig. 5.

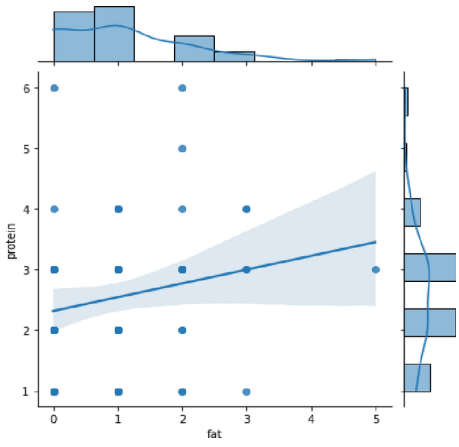

Fig. 4. Joint plot of fat vs protein

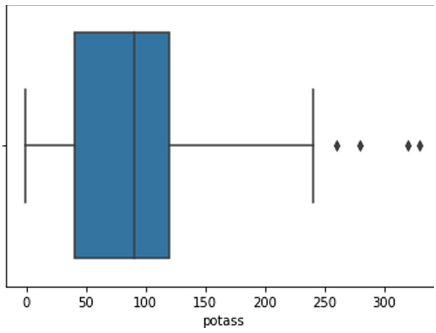

Fig. 5. Boxplot for detecting outliers

4.2 Comparative Analysis of Machine Learning Models

The authors have evaluated the effectiveness of all classifiers in terms of precision, accuracy, and F1 score. The Comparative Analysis has been made through the graph of different Machine Learning models used in our project – Linear regression, Logistic regression, K-NN, random forest, and decision trees on different parameters. Each model provides high accuracy of 0.9375 other than the linear regression giving an accuracy of 0.8355865644. The authors have established a comparative analysis of the different models by comparing the value of the training score of each model. The model with the highest train score, Random Forest, is chosen as the best model for the data prediction from the chosen dataset. The details of various metrics obtained are presented in Table 1.

Table 1. Comparative Analysis of different models

Model	Training score
Linear Regression	0.9495
Logistic regression	0.9672
K-NN	0.8196
Random forest	0.9836
Decision trees	0.9672

4.3 Challenges

During implementation of various machine learning models, authors experienced various challenges as follows:

- Selection of an appropriate visualization technique for the accurate representation of the data can become a tedious job.
- Data should be appropriately chosen or the precision and accuracy may differ distinctly.
- Wrong data pre-processing and feature selection may lead to imprecise and incorrect predictions in the machine learning models.
- The parameters chosen, and the number of maximum iterations specified should be selected appropriately to ensure maximum accuracy of prediction.

5 Conclusion

In this work, the authors have attempted to determine the nutritional value of various bowls of cereal through various methods. The various bowls of cereal like wheat, rice, maize, barley, oats, millets, etc. are discussed. It is clarified from the study by various researchers that although wholesome cereals have high nutritional value; processing may lead to a reduction in the same. Hence conscious efforts must be taken during the processing of the cereals. Also, the need to increase the growth of cereals is briefly discussed given the rising population. For the same, initiatives like the inclusion of non-staple cereals must be taken. The requirement to understand the nutritional value of cereals is important and hence continuous research must prevail in this direction.

References

1. Sharma, N., Yadav, S., Mangla, M., Mohanty, A., Mohanty, S.N.: Multivariate analysis of COVID-19 on stock, commodity & purchase manager indices: a global perspective (2020)
2. Sharma, N., et al.: Geospatial multivariate analysis of COVID-19: a global perspective. Geo J., 1–15 (2021)
3. Callahan, S.P., Freire, J., Santos, E., Scheidegger, C.E., Silva, C.T., Vo, H.T.: VisTrails: visualization meets data management. In: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, pp. 745–747 (2005)
4. Sadiku, M., Share, A.E., Musa, S.M., Akujuobi, C.M., Perry, R.: Data visualization. Int. J. Eng. Res. Adv. Technol. (IJERAT) **2**(12), 11–16 (2016)
5. Meyer, R.D., Cook, D.: Visualization of data. Curr. Opin. Biotechnol. **11**(1), 89–96 (2000)
6. Mangla, M., Sharma, N., Mehta, V., Mohanty, S.N., Saxena, K.: Statistical analysis for air quality assessment and evaluation: a data mining approach. In: 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 1–5. IEEE (2021)
7. Mangla, M., Shinde, S.K., Mehta, V., Sharma, N., Mohanty, S.N. (eds.) Handbook of Research on Machine Learning: Foundations and Applications. CRC Press (2022)
8. McKevittith, B.: Nutritional aspects of cereals. Nutr. Bull. **29**(2), 111–142 (2004)
9. Dewettinck, K., Van Bockstaele, F., Kühne, B., Van de Walle, D., Courtens, T.M., Gellynck, X.: Nutritional value of bread: influence of processing, food interaction and consumer perception. J. Cereal Sci. **48**(2), 243–257(2008)

10. Farag, M.A., Xiao, J., Abdallah, H.M.: Nutritional value of barley cereal and better opportunities for its processing as a value-added food: a comprehensive review. Crit. Rev. Food Sci. Nutr. **62**(4), 1092–1104 (2022)
11. Munck, L.: Improvement of nutritional value in cereals. Hereditas **72**(1), 1–128 (1972)
12. Deshpande, S.S., Mohapatra, D., Tripathi, M.K., Sadvatha, R.H.: Kodo milletnutritional value and utilization in Indian foods. J. Grain Process. Storage **2**(2), 16–23 (2015)
13. Vila-Real, C., Pimenta-Martins, A., Maina, N., Gomes, A., Pinto, E.: Nutritional value of indigenous whole grain cereals millet and sorghum. Nutr. Food Sci. Int. J. **4**(1) (2017)
14. Sharma, N.: XGBoost. The extreme gradient boosting for mining applications. GRINVer-lag (2018)
15. Mitchell, T., Buchanan, B., DeJong, G., Dietterich, T., Rosenbloom, P., Waibel, A.: Machine learning. Ann. Rev. Comput. Sci. **4**(1), 417–433(1990)
16. Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives, and prospects. Science **349**(6245), 255–260 (2015)
17. Sharma, N., Juneja, A.: Combining random forest estimates using LSboost for stock market index prediction. In: 2017 2nd International conference for convergence in Technology (I2CT), pp. 1199–1202. IEEE (2017)
18. Sharma, N., Juneja, A.: Extreme gradient boosting with a squared logistic loss function. In: Tanveer, M., Pachori, R. (eds.) Machine Intelligence and Signal Analysis. pp. 313–322. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-0923-6_27
19. Oduntan, O.E., Hammed, M.: A predictive model for improving cereals crop productivity using supervised machine learning algorithm, pp. 1–11 (2018)
20. Jensen, S.M., Akhter, M.J., Azim, S., Rasmussen, J.: The predictive power of regression models to determine grass weed infestations in cereals based on drone imagery—statistical and practical aspects. Agronomy **11**(11), 2277 (2021)
21. Arora, A., Gupta, P.K.: Data science and its relation to big data and machine learning. Int. Res. J. Modernization Eng. Technol. Sci. **3**(5), 61–65 (2021)
22. Gupta, P.K., Rishi, R, Biswas, R.: A comparative analysis of temporal data models. Int. J. Adv. Comput. Eng. Network. **1**(8), 34–38 (2013)
23. Gupta, P.K., Singh, J.P., Kaliraman, J.: Master data management emerging issues. Int. J. Eng. Technol. Sci. Res. **4**(6), 268–272 (2017)
24. Gupta, P.K., et al.: Deep learning architecture and algorithms. IN: Proceedings of Techbyte (A National Symposium, held at JIMS, New Delhi, India, pp. 42–47 (2019)