International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Prediction of Accuracy of Datasets by Using Machine Learning Algorithms on Food Habits among College Going Students in Eastern India

Kamalika Chatterjee[1], Pranabesh Ghosh[2], Soumendra Nath Talapatra[3]

[1]Ph. D Scholar, Department of Food and Nutrition, School of Life Sciences, Seacom Skills University, Kendradangal, Birbhum, West Bengal, India
Corresponding Author Email: *cr.kamalika[at]gmail.com*
Phone: +91 – 8240239051

[2, 3]School of Life Sciences, Seacom Skills University, Kendradangal, Birbhum, West Bengal, India

Abstract: ***The eating habits of fast food may lead to various disorders such as obesity, diabetes, etc. among adults. The present study was evaluated to predict accuracy performance of dataset for food habits viz. The datasets were used as Food_habit, Number_of_meals/day, Omit_any_meal, Take_any_special_food_or_not, Type_of_food_normally_taken, Type_of_meal_preferred, Frequency_of_eating_outside, Habit_of_skipping_breakfast and class effects viz. Normal and abnormal category among college going students (35 nos.) of eastern India. In this study, the prediction accuracy was obtained through 7 machine learning (ML) algorithms especially BayesNet (BN), NaiveBayes (NB), logistic regression (LR), Stochastic Gradient Descent (SGD), Sequential minimal optimization of Support Vector Machine (SMO), K - nearest neighbour (IBK), and Lazy. KStar (K*), by using ML tool (WEKA, version 3.8.5) as per cross - validation (CV) test for above - mentioned classes. As per the statistical summary results, the prediction accuracy through precision recall curve (PRC) were obtained as highest values (98%) for these algorithms. Future study should be performed with other big dataset with ML algorithms especially Tree algorithms.***

Keywords: Machine learning algorithms, Prediction accuracy, Food habits dataset, Habits of fast food eating, College students

1. Introduction

In recent trends, the junk food and fast food consumption become a trend among college going students as they lead a sedentary lifestyle. Different factor like single parents' family, working mother, longer study time, demographic profile is responsible for this kind of food habit. Along with the food habit socioeconomic status are also related with the good health. [1]

Daily meal planning along with nutritious food is a weapon to combat any kid of health disorder as well as it creates a positive impact on one's well - being. To know the nutritional type and meal pattern among college going students and educate them on the proper nutrition during college times is very much helpful tool because they spend most of the time in the college. Through this tool the college going students will be educated enough for their healthy lifestyle. [2]

The outstanding progress has been noticed over the last few decades to understand the nutrition pattern is interacted with health. But despite the wealth of knowledge, health habits related to food are widespread and sometimes on the rising trend. [3] In precision nutrition study, four machine learning (ML) assignments are important in the form of regression, classification, recommendation and clustering in which most of these employ a supervised approach. [3]

The characteristics of ML made it appropriate for such analysis and thus offered itself as an unconventional tool to deal with data of this kind. ML has already been applied in important problem areas in nutrition, such as obesity, metabolic health, and malnutrition. [4] In this context, the prediction of accuracy performance of dataset for food habit among college students of eastern India is lacking.

The present study was evaluated to predict accuracy performance of dataset for food habits by using ML algorithms.

2. Materials and methods

The data mining through ML modelling algorithm was performed by using WEKA (Waikato Environment for Knowledge Analysis) tool (version, 3.8.5) developed by Frank et al. [5] The WEKA explorer was developed with data pre - processing, classification, regression, and association rules. [6]

In this study, the prediction accuracy was obtained through 7 machine learning (ML) algorithms especially BayesNet (BN), NaiveBayes (NB), logistic regression (LR), Stochastic Gradient Descent (SGD), Sequential minimal optimization of Support Vector Machine (SMO), K - nearest neighbour (IBK), and Lazy. KStar (K*), by using ML tool (WEKA, version 3.8.5) as per cross - validation (CV) test for normal and abnormal class. All the statistical summary results were retrieved for "F - measure, Matthew's correlation coefficient (MCC), receiver operating characteristic (ROC) curve and Precision - recall curve (PRC) ".

The datasets were used as Mother_tongue, Food_habit, Number_of_meals/day, Omit_any_meal, Take_any_special_food_or_not, Type_of_food_normally_taken, Type_of_meal_preferred, Frequency_of_eating_outside, Habit_of_skipping_breakfast

and class effects viz. Normal and abnormal category of college going students (35 nos.).

3. Results

Table 1 evaluates the summary results of correctly and incorrectly classified instances of studied models. In the case of algorithm classification, the highest values were observed in BN, NB, LB and SGD (97.14), followed by IBK and K* (94.28) and lower value on SMO (88.57) as per 10 - fold CV.

Table 1: Summary results of different models (correctly and incorrectly classified instances)

Classifier models	Correctly classified instances	Incorrectly classified instances
BN	97.14	2.86
NB	97.14	2.86
LR	97.14	2.86
SGD	97.14	2.86
SMO	88.57	11.43
IBK	94.28	5.71
K*	94.28	5.71

BN = BayesNet; NB = NaiveBayes; LR = Logistic regression; SGD = Stochastic Gradient Descent; SMO = Sequential minimal optimization of Support Vector Machine; IBK = K - nearest neighbour; K* = Lazy. KStar

Table 2 evaluates the summary results of Kappa statistic (KS), mean absolute error (MAE) and root mean squared error (RMSE) of studied models related to 10 - fold CV. In case of prediction accuracy of the class of KS values, the highest values were observed in BN, NB, LR and SGD followed by IBK and K* while lower values of MAE and RMSE in the same manner.

Table 2: Model summary (Kappa statistic, mean absolute error and root mean squared error) results

Classifier models	KS	MAE	RMSE
BN	0.94	0.07	0.20
NB	0.94	0.07	0.20
LR	0.94	0.03	0.17
SGD	0.94	0.03	0.17
SMO	0.77	0.11	0.34
IBK	0.88	0.06	0.16
K*	0.88	0.06	0.17

The statistical data based on F - measure, MCC, ROC area and PRC area of studied models as per 10 - fold CV. In case of prediction accuracy of the algorithms, the MCC, ROC area and PRC ranges between 77% - 94%, 88% - 99% and 81% - 99%, respectively observed. The graphical representation for statistical result of each algorithm is exhibited in Fig 1 to Fig 7.

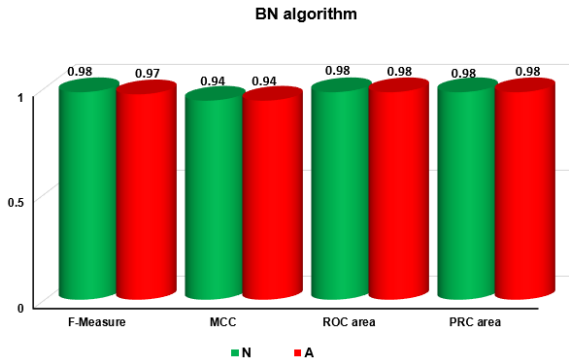

Figure 1: Graphical representation of prediction accuracy of BN algorithm as per dataset

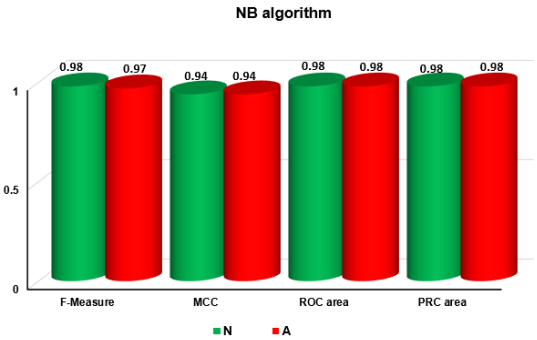

Figure 2: Graphical representation of prediction accuracy of NB algorithm as per dataset

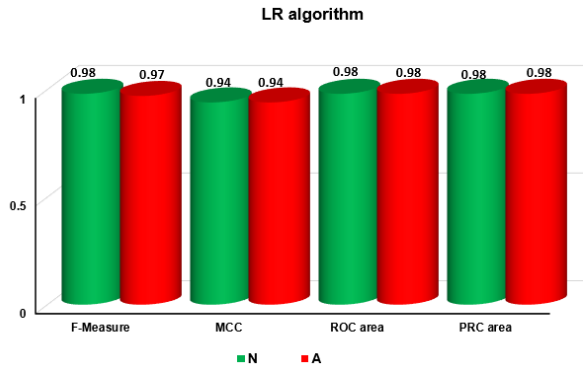

Figure 3: Graphical representation of prediction accuracy of LR algorithm as per dataset

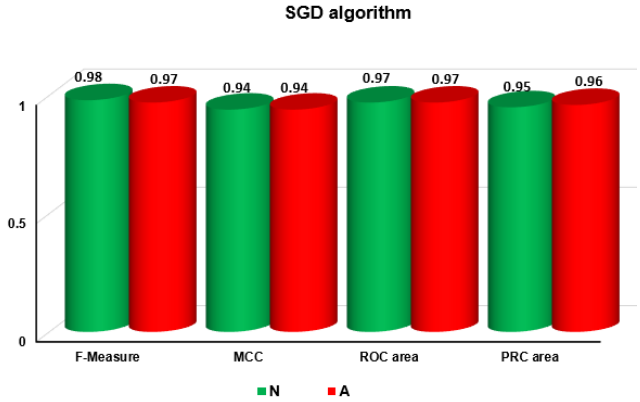

Figure 4: Graphical representation of prediction accuracy of SGD algorithm as per dataset

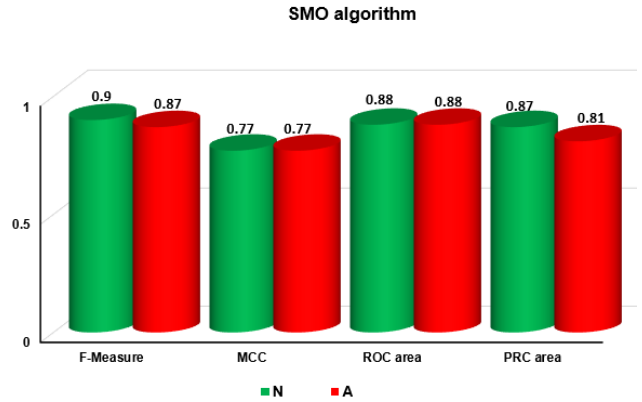

Figure 5: Graphical representation of prediction accuracy of SMO algorithm as per dataset

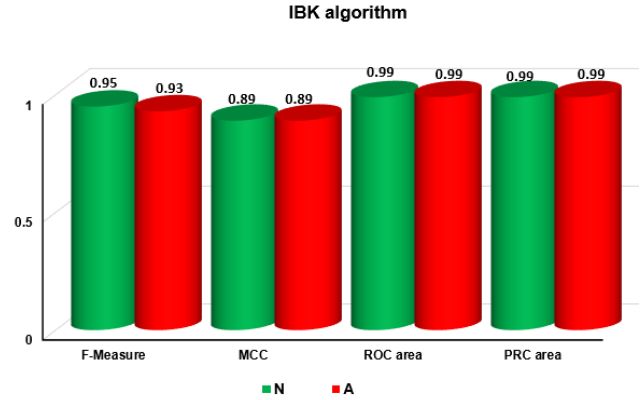

Figure 6: Graphical representation of prediction accuracy of IBK algorithm as per dataset

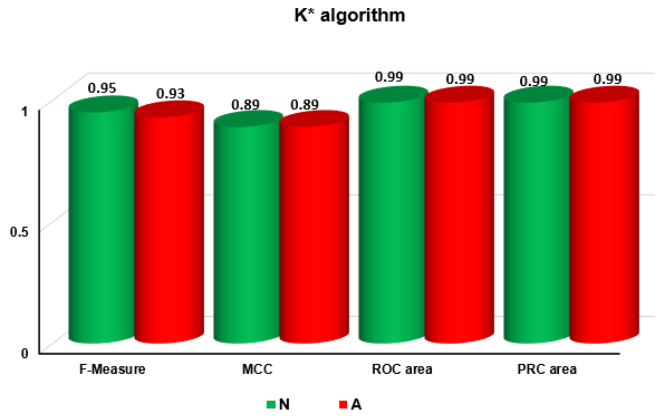

Figure 7: Graphical representation of prediction accuracy of K* algorithm as per dataset

4. Discussion

A similar result was obtained by Qasrawi et al. [7] related to DT algorithm in which accuracy rate of about 82.1%. While RamyaSri et al. [8] revealed better accuracy of SVM (74.88%) in comparison with naïve - Bayes algorithm (73.02%). Ratra and Gulia [9] evaluated a heart disease dataset as per metabolic syndrome. The comparative analysis revealed that the values of Precision Metric for Naïve bays (83.7% and 82.4 %), Random Forest (81.8% and 77.9%) and k - nearest (75.3% and 58.0%), respectively were recorded for WEKA and Orange tool separately. Talapatra et al. [10] predicted through ML algorithm models viz. BayesNet (BN), NaiveBayes (NB), logistic regression (LR), Lazy. KStar (K*), decision tree (DT) J48, Random forest (RF) and Random tree (RT) to know the predictive accuracy of the dataset created from an image of fish erythrocytes as per mutation It was observed the better predictive accuracy of specific algorithms viz. RF and RT followed by K*, LR, BN andDTJ48 and lowest in NB as per training and testing dataset. Mondal et al. [11] predicted the accuracy of dataset related to the shape of normal and abnormal cellular and nuclear features of fish erythrocytes through 11 ML algorithms. The prediction accuracy of a class of different statistical values viz. TP, FP, precision, recall, MCC, ROC and PRC were obtained, and the higher values were recorded in RF, RT, IBK, K*, SC, BN, DTJ48 and LR while lower values in NB, SGD, and SMO algorithms according to the studied dataset. Moreover, the higher PRC (>95%) values for 8 algorithms viz. RF, RT, IBK, K*, SC, BN, DTJ48 and LR in their dataset.

5. Conclusion

ML algorithms viz. BN, NB, LR, SGD, IBK and K* performed accurately from the dataset and obtained rich information with statistical validation by using WEKA tool. It was easily be identified that the dataset of food habit predicted the classifier accuracy among the studied algorithms. In the present study, PRC values were recorded the ranged between 81% to 99% for the prediction of the dataset accuracy on food habit among college going students. It was recorded that the food habit was better in this dataset.

Conflict of interest
No conflict of interest

References

[1] Mallikarjuna T, Janakiramaiah G, Naidu RV. Changing food habits of the adolescence girls in urban areas: A sociological evaluation in Tirupati city. Social Work Review.2015; 51 (1).
[2] Gupta Y, Chhabra S. Food choices and dietary habits of college going girls with special emphasis on breakfast consumption practices. International Journal of Home Science.2022; 8 (1): 169 - 178.
[3] Kirk D, Catal C, Tekinerdogan B. Precision nutrition: A systematic literature review. Computers in Biology and Medicine.2021; 133: 104365.
[4] Kirk D, Kok E, Tufano M, Tekinerdogan B, Feskens EJM, Camps G. Machine learning in nutrition research. Advances in Nutrition (Bethesda, Md.).2022; 13 (6): 2573 - 89.
[5] Frank E, Hall MA, Witten I H. The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2016.
[6] Witten IH, Frank E, Hall MA. Data Mining: Practical machine learning tools and techniques.3rd edn, Morgan Kaufmann, Burlington, MA, 2011.
[7] Qasrawi R, Badrasawi M, Al - Halawa DA, Polo SV, Khader RA, Al - Taweel H, et al. (2024). Identification and prediction of association patterns between nutrient intake and anemia using machine learning techniques: results from a cross - sectional study with university female students from Palestine. European Journal of Nutrition.2024; 63: 1635 - 49.

[8] RamyaSri R, Susmitha G, SaiSurya K, Bhavani V, Venkata Raju K. (2019). Consumption of food by college students using Ml algorithms. International Journal of Recent Technology and Engineering.2019; 8 (4): 10339 - 45.
[9] Ratra R, Gulia P. Experimental evaluation of open source data mining tools (WEKA and Orange). International Journal of Engineering Trends and Technology.2020; 68 (8): 30 - 5.
[10] Talapatra SN, Chaudhuri R, Ghosh S. CellProfiler and WEKA tools: Image analysis for fish erythrocytes shape and machine learning model algorithm accuracy prediction of dataset. World Scientific News.2021; 154: 101 - 116.
[11] Mondal B, Bhattacharya K, Talapatra SN. Image analysis to obtain dataset and machine learning for prediction accuracy on abnormal features of peripheral erythrocytes of fish specimen. Journal Of Electronics Information Technology Science and Management.2023; 13 (1): 125 - 135.