Alcohol Consumption Rate Prediction using Machine Learning Algorithms

Advait Singh
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar 751024, Odisha
advaitsingh27@gmail.com

Vinayak Singh
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar 751024, Odisha
vinayaksooryavanshi@gmail.com

Mahendra Kumar Gourisaria
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar 751024, Odisha
mkgourisaria2010@gmail.com

Ashish Sharma
Department of Computer Engineering and Applications
GLA University
Mathura 281406, UP
ashishs.sharma@gla.ac.in

2022 OITS International Conference on Information Technology (OCIT) | 978-1-6654-9348-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/OCIT56763.2022.00026

***Abstract*—Consumption of alcohol among students, mainly college or university students, has risen immensely over the past couple of years. It has been determined that students experiment with alcohol during their college years and around 80% of students consume alcohol in some manner or degree and 50% are involved in binge drinking. This is mainly due to students wanting to explore their newfound independence and freedom which they didn't have during their school years. In this paper, we have analyzed students belonging to two courses of a Secondary School-Maths and Portuguese Language Course. We have applied Feature Scaling along with various machine learning classification models to determine higher alcohol consumption where the Random Forest Model outperformed all other models that have been applied such as Linear, Ridge, and Lasso Regression, Decision Tree, k-NN, XG Boost, Support Vector Machine, ADA Boosting Regressor and Gradient Boosting Regressor for analysis of alcohol consumption among secondary school students.**

Keywords—Machine Learning, Deep Learning, Alcohol Consumption, Random Forest, Math Course, Portuguese Language Course.

I. Introduction

Alcohol, chemically referred to as ethanol, can be termed an intoxicating substance found in drinks such as beer, wine, etc. It can be termed as a drug that leads to erratic behavior and speech and disturbed senses. Alcohol consumption can be described as the drinking or consuming of drinkables containing ethanol. These beverages are drunk due to the intoxicating effects they have and are often consumed at social events. According to a survey conducted on a national level, it was observed that 53 percent of students from ages 18 to 22 who are full-time in college drink alcohol once a month and that 33 percent are involved or participate in binge drinking during that same time [1]. Excessive alcohol consumption can even lead to various mental health problems namely depression, specific phobia, social anxiety, and persistent fear [2].

According to research on alcohol consumption among undergraduate students in the UK over the past 25 years, male students tend to drink more than is reasonable, and one in two of them engage in binge drinking, compared to female students, who made up 40% of the sample who drink more than the sensible level of consumption and those women were observed more heavy drinkers than men or their male counterparts [3]. According to the day of the week, it is seen that Thursday is the day with the highest overall alcohol consumption because of weekend-like behaviors and the impact of Friday class attendance and time [4].

Various machine learning algorithms that have been applied in this study can be utilized for various other analysis and research purposes such as Early Detection of Fetal Disease [5], Medical Imaging [6], and Prediction of Heart Diseases [7]. More medical topics, such as Lung Pattern Classification [8], Mycobacterium Tuberculosis Detection [9], Alzheimer's diagnosis [10], and Maize Crop Disease [11], can be addressed and explored by the application of deep learning techniques like CNN along with RNN and MLP to name a few.

In this paper, we will analyze the various machine learning algorithms applied to the student alcohol consumption dataset where Feature Scaling was applied for the best possible accuracy. The rest of the paper is organized into sections where Section II illustrates the related work, Section III analyzes the methods and materials applied in this research. Section IV interprets the implementation and results and Section V deals with the conclusion and future work. After the application of the models, the model with the best results was selected, providing the best and high accuracy for classifying alcohol consumption based on student courses.

II. Related Work

Rao *et al.* (2021) [12] claimed that a QCM Sensor dataset with different chemical compounds of alcohol was classified using Quartz Crystal Microbalance (QCM) sensors. The objective or aim was to obtain the aptest classification for the QCM Sensor. Gradient boosting is the most successful technique compared to other classification methods such as Logistic Regression, Decision Tree, and Linear discriminant Analysis where Logistic Regression and Linear Regression resulted in an overall accuracy of 90% and multi-layer perceptron with a value of 79%. A. Pisutaporn *et al.* (2018) [13] studied alcohol consumption in a public dataset containing student attributes and grades. Classification models, including the Decision Tree algorithm and the random forest algorithm,

978-1-6654-9348-2/22/$31.00 ©2022 IEEE
DOI 10.1109/OCIT56763.2022.00026

were used. It was concluded that random forest was the most favorable as it provided accuracy of 88.07% and 79.43%.

Shukla *et al.* (2018) [14] investigated the compatible attributes from the performances of Secondary School students to obtain important facts to overcome the limitations of the existing data mining models. Correlation-based Feature Scaling (CFS) was performed along with Information Gain (IG), Relief-F, and Chi-Square (CS) using various classifiers. An accuracy of 71.39% was achieved as the highest along with 71.34% sensitivity, and 66.86% precision with ROC-AUC at 85.89% respectively. The National Institute of Health reported in 2016 [15] that 64% of 12 graders, 47% of 10 graders, and 26% of 8 graders have consumed alcoholic drinks. To investigate this, the study makes predictions about the behavior associated with alcohol use. Classification accuracy was compared between the variations of the neural network algorithm and it was found that a self-tuning multilayer perceptron classifier (AutoMLP) produced an accuracy of 64.54% which was greater compared to the accuracy produced by the standard MLP neural network of 61.78%.

Jain *et al.* (2022) [16] used machine learning algorithms and models along with data mining solutions such as Random Forest, Decision Tree, and so on to predict campus placement for undergraduate students and whether or not the cell responsible for training and placement's workload will be impacted. It was observed that the Voting classifier achieved the highest accuracy and outperformed SVM, k-NN, and other classifiers as mentioned above. Sisodia *et al.* (2019) [17] analyzed the rise in alcohol consumption among students belonging to Secondary Schools as it is a matter of concern. Random Trees, Simple Logistic, Random Forest, etc. were the classifiers employed. The results after evaluation of all the classifiers suggested that Simple Logistic performed best as an individual classifier with an accuracy of 87.02% whereas Random Forest performed best as an ensemble classifier with an accuracy of 93.45%. Jinbo Bi *et al.* (2013) [18] worked on a model to set a plausible basis and to offer acumen for a better and more adequate college-level alcohol mediation. First, a correlated support vector machine was applied to compose a classifier for different attributes for individual students. Secondly, a combination of Feature Scaling and cluster analysis was done to identify drinking patterns and risk factors associated with each of those patterns. Onur Sapci *et al.* (2021) [19] adapted a theoretical model of social interactions in alcohol consumption using three machine-learning algorithms to overcome the limitations of existing research using machine learning based on student alcohol consumption. It was observed that the algorithm that performs the best at forecasting alcohol intake is extreme gradient boosting with an accuracy of 73.49%.

Dutta *et al.* (2022) [20] used dimension reduction and data mining techniques to predict early-stage coronary ailment and develop a cost-effective treatment using AdaBoost, Decision Tree. The dimensionality reduction models applied were Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to increase and better the performance of the machine learning algorithms by reducing or minimizing the number of attributes. It was concluded that Artificial Neural Network was the best algorithm obtaining an accuracy of 88.52% with PCA and 85.24% with LDA. Pal *et al.* (2017) [21] studied alcohol consumption in higher education institutes or among Secondary School students referring to students of age lesser than 18. They applied four data mining techniques namely Decision table, Sequential minimal optimization (SMO), Bagging, and REP Tree to improve the efficiency of academic performance. It was concluded that the Bagging classifier was the best performer with an accuracy of 80.25%.

III. Dataset Preparation And Technology Used

This section provides insight into the dataset used for the study, its preparation, and the technology used. It is comprised of two tables belonging to students of different courses. For better evaluation, different scaling methods were used and then the classifiers were applied. The classifier with the best accuracy was selected as result. The section is subdivided into the following sections - *A. Dataset Used, B. Data Exploration C. Technology Used.*

A. Dataset Used

Dataset was issued by Data Society in 2016 and it contains records of Secondary School students belonging to two courses-Math and Portuguese Language Course along with their social, gender, and study data. It was obtained through a survey and contains data from 395 students of Math and 649 students of Portuguese Language Courses. Fig. 1 and Fig. 2 show the correlation matrices for both courses.

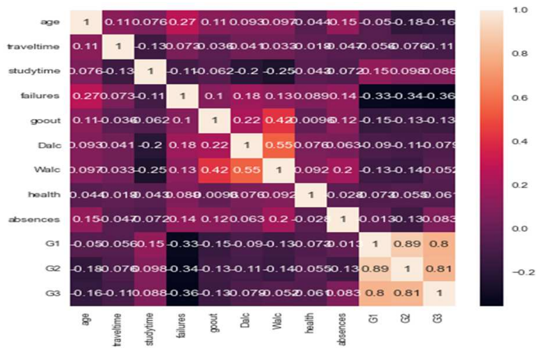

Fig. 1. Math Course Correlation

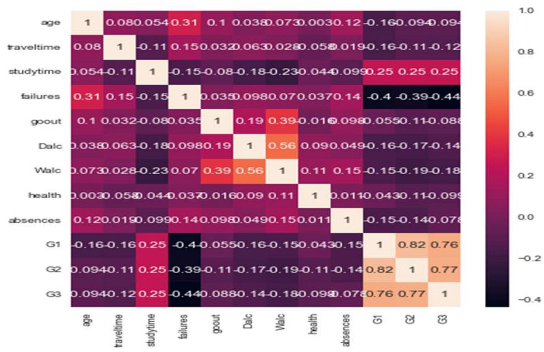

Fig. 2. Portuguese Language Course Correlation

B. Data Exploration

The most crucial component of a successful machine-learning model is Data Exploration. It is a stage of data analysis, used to examine and display data to find insights right away or pinpoint regions or patterns to further investigate. In our case, scaling methods such as Standard Scaler, Min-Max Scaler, and Robust Scaler were applied to both sets of students to eliminate irrelevant and redundant features as both Math and Portuguese Language Courses contained the same attributes such as Age, Traveltime, Freetime, Study time, Absences, etc. Standard Scaler proved to be most fruitful after the application of various Classification algorithms such as Linear Regression, Lasso Regression, Random Forest, Decision Tree, XG Boost, and SVM post-training and testing as it uniformly scaled the input dataset's functionality and provided the best accuracy along with Mean Squared Error, R Squared Error, and Root Mean Squared Error. The workflow for analyzing the given dataset is below in Figure 3.

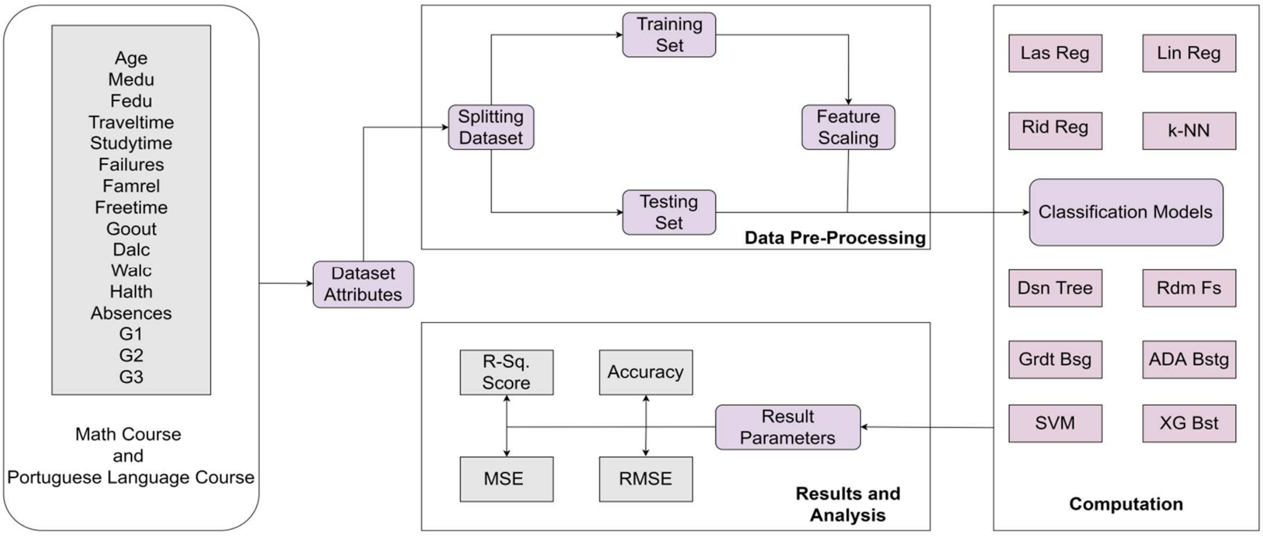

Fig. 3. Workflow for analyzing the dataset

C. Technology Used

Different machine learning algorithms were used and then compared. Algorithms can be used along with data mining models for further research purposes such as Surface Crack Detection [22], Corner Detection [23], and in medical cases, Prediction of Liver Disease [24]. The table given below explains all the applied classifier algorithm models along with their advantages and disadvantages.

TABLE I. TABULAR COMPARISON OF CLASSIFIER MODELS APPLIED

Classifier	Advantage	Disadvantage
Linear Regression	For linearly separable data, it works remarkably well.	The presumption is that dependent and independent variables are linear.
Ridge Regression	When there are many predictors compared to observations in a big multivariate dataset, they continue to perform well (n).	Incapable of Feature Scaling.
Lasso Regression	For improved prediction and model interpretation, Lasso performs shrinkage and variable selection simultaneously.	Group selection is not possible with Lasso Regression.
k-NN	k-NN is fairly simple to use. The distance function and the value of K are the only two factors needed to implement k-NN.	The k-NN method struggles with big dimensional data and requires Feature Scaling.
Decision Tree	Decision Trees take less work to prepare the data during pre-processing than other methods do.	Decision Trees are ineffective for applying regression and forecasting continuous values and frequently take more time to train the model.
Random Forest	Regression and classification issues can be resolved using Random Forest.	In contrast to Decision Trees, which only build one tree, Random Forest creates many trees and integrates their results.
Gradient Boosting Regressor	Increased training effectiveness and speed	It leads to overfitting and an overemphasis on outliers.
ADA Boosting Regressor	By applying AdaBoost, weak classifiers' accuracy can be increased.	Before implementing an AdaBoost algorithm, noisy data and outliers must be avoided.
SVM	Support vector machines perform well when the margin of class dissociation is measured.	SVM does not perform very well when target classes overlap.
XG Boost	In XG Boost, several hyper-parameters can be tuned.	On unstructured and sparse data, XG Boost does not perform as well.

IV. IMPLEMENTATION AND RESULTS

This section deals with the information and result obtained after the implementation of Feature Scaling along with various classification algorithms to obtain the best results. The section has been divided into the following sub-sections as given below: *A. Hardware and Software, B. Experimental Information, and C. Results*.

A. Hardware and Software

With the aid of the Jupyter notebook and Python 3, all machine-learning implementations and analyses were conducted. The hardware workstation runs Windows 11 and has an Intel i7 10th generation @1.60 GHz CPU and 16 GB RAM.

B. Experimental Information

All the machine learning models for categorization presented above were applied using different scaling methods. The initial approach used a Feature Scaling mechanism called Standard Scaler that normalizes the data within a specific range, removes the mean, and scales each feature/variable to unit variance. Min-Max Scaler was the second method in which smaller standard deviations were obtained which will suppress the effect of outliers, and then Robust Scaler is the third method which will remove the median and scale the data according to the quantile range. After all the methods, classification models were applied through which we determined the various measures for the evaluation sections like Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error for both tables i.e., for Math Course and Portuguese Language Course students given in Table 2 and Table 3 below.

The pair plot and bar plot of the respective course students is also given in Fig. 4 and Fig. 5 for Math Course students and Fig. 6 and Fig. 7 for Portuguese Language Course students. The result shown is based on the Feature Scaling method that demonstrated and achieved the best results after the application of classification methods. Standard Scaler was selected as it reduced all the characteristics down to a similar scale without distorting the variations in the value ranges. Min-Max Scaler was not selected due to its inability to handle the outliers and Robust Scaler was denied due to its resistance to outliers. When Random Forest was applied, the number of trees or *n_estimators* was taken as 1500 before finding the average of predictions for both the Math Course and Portuguese Language Course.

C. Results

After applying classification models, Random Forest was most effective since it provided the best accuracy for both Math and Portuguese Language Courses with an accuracy of 0.90 and 0.82. All models produced good results, but owing to scaling methodology, Math results were higher than those from the table of the Portuguese Language Course for instance XG Boost gave an accuracy of 0.89 and 0.80 which are very close to the highest accuracy achieved whereas the lowest accuracy of 0.5912 for Math Course was given by k-NN and 0.6056 for Portuguese Language Course was given by Lasso Regression.

Random Forest yielded the best results as the overfitting issue was avoided because the final result is based on average or majority rating meaning on various samples, it constructs Decision Trees and uses their majority vote to classify data. XG Boost also performed well due to its ability to parallel processing and regularization on small to medium datasets. Lasso Regression did not perform well due to its fitting algorithm whereas k-NN also performed poorly because of its inability to in analyzing large datasets or a high number of dimensions.

TABLE II. PERFORMANCE EVALUATIONS FOR MATH COURSE

Model	Accuracy	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Linear Regression	0.7918	-1.3216	-3.5147	-1.8676
Ridge Regression	0.7909	-1.3208	-3.5175	-1.8701
Lasso Regression	0.7169	-1.4844	-4.7791	-2.1858
k-NN	0.5912	-1.6967	-5.2132	-2.2816
Decision Tree	0.8558	-1.0417	-2.6269	-1.5910
Random Forest	0.9005	-0.9087	-1.6821	-1.2917
Gradient Boosting Regressor	0.8648	-1.0851	-2.2765	-1.5071
ADA Boosting Regressor	0.8818	-1.0379	-2.1334	-1.4147
SVM	0.7575	-1.1503	-4.0183	-2.0037
XG Boost	0.8938	-0.9396	-1.7977	-1.3383

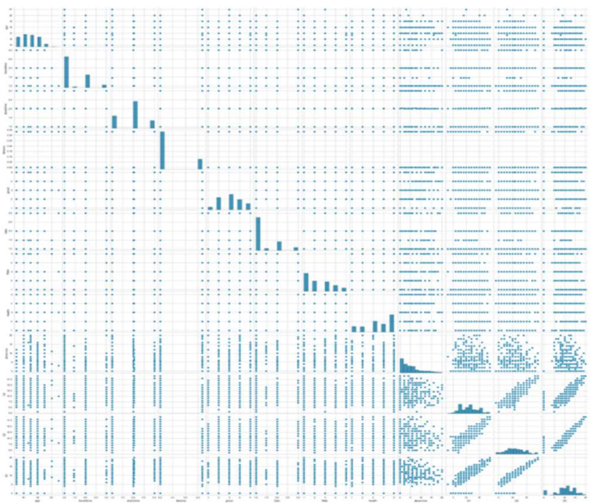

Fig. 4. Pair plot of Math Course

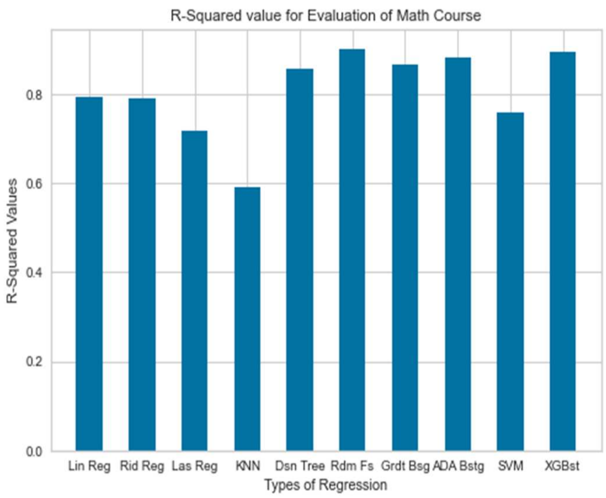

Fig. 5. Bar plot of Math Course

TABLE III. PERFORMANCE EVALUATIONS FOR THE PORTUGUESE LANGUAGE COURSE

Model	Accuracy	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Linear Regression	0.7522	-0.9053	-1.9829	-1.4029
Ridge Regression	0.7512	-0.9066	-1.9862	-1.4076
Lasso Regression	0.6056	-1.2169	-3.1487	-1.7737
k-NN	0.6584	-1.2261	-2.7292	-1.6456
Decision Tree	0.7013	-0.9711	-2.3603	-1.5068
Random Forest	0.8211	-0.8006	-1.4363	-1.1957
Gradient Boosting Regressor	0.7977	-0.9298	-1.6007	-1.2673
ADA Boosting Regressor	0.7155	-1.1917	-2.3401	-1.5028
SVM	0.7414	-0.8701	-2.0622	-1.4341
XG Boost	0.8027	-0.8808	-1.5746	-1.2531

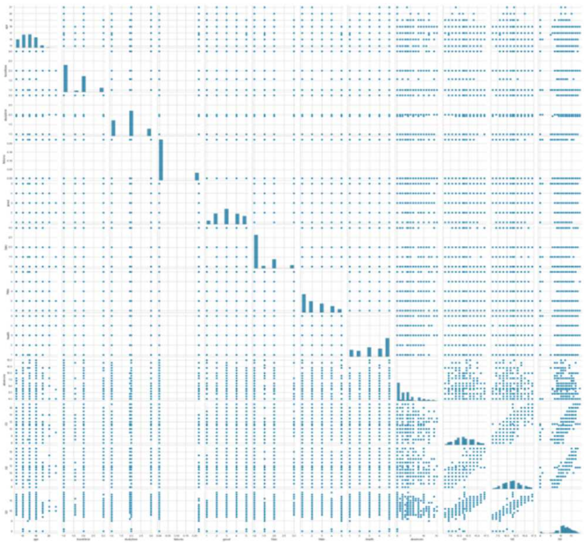

Fig. 6. Pair plot of Portuguese Language Course

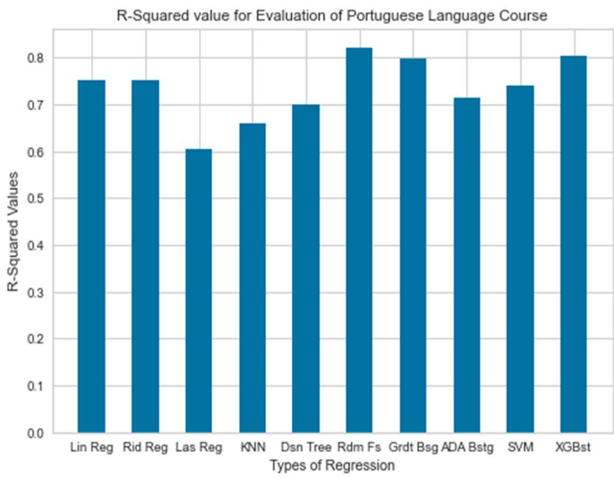

Fig. 7. Bar plot of Portuguese Language Course

V. CONCLUSION AND FUTURE WORK

In this study, we used machine learning and deep learning to identify whether or not a student drinks alcohol. If so, based on the courses they are in, how many of them do so? We applied various machine learning methods and models to the same dataset, including Decision Trees, random forests, lasso, and ridge regression. We employed Feature Scaling to facilitate understanding and accelerate model training. Our main objective was to effectively answer this categorization problem while precisely verifying each piece of information that belonged to each row and column. Additionally, we found that XG Boost and Random Forest are strong competitors. After much testing and analysis, we concluded that Random Forest produced the best results with the highest accuracy of 0.90 and 0.82 for the Math and Portuguese Language Course i.e., more students in the Portuguese Language Course engage in alcohol use than those in the Math Course.

For various outcomes in future work, we can employ additional Feature Scaling and machine learning techniques. Additionally, sophisticated machine learning and deep learning approaches can be applied to categorize student alcohol consumption. Convolutional Neural Network (CNN) is one method known to automatically detect essential features without any supervision from the user. Dimensionality Reduction techniques like Principal Component Analysis (PCA) which eliminate correlated variables that play no role in decision making and Linear Discriminant Analysis (LDA) minimize the within-class variance and maximize the between-class variance. More Deep Learning methods can be applied such as Recurrent Neural Networks (RNN) which can help in achieving short-term memory in a network, Self-Organizing Maps (SOMs) enable data visualization to reduce the dimensions of data through self-organizing artificial neural networks and choose a vector at random from the training data. Also, some data transformation techniques like an autoencoder transformation and various scaling methods like YeoJohnson, Boxcox, standard scaling, and min-max scaling can be used for scaling up the dataset.

REFERENCES

[1] College Drinking. (2022). National Institute on Alcohol Abuse and Alcoholism(NIAAA).https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/college-drinking

[2] Nor Safika Mohd Shafiee, Sofianita Mutalib. " Prediction of Mental Health Problems among Higher Education Students Using Machine Learning ", International Journal of Education and Management Engineering (IJEME), Vol.10, No.6, pp.1-9, 2020. DOI: 10.5815/ijeme.2020.06.01

[3] J. S. Gill (2002). Reported levels of alcohol consumption and binge drinking within the UK undergraduate student population over the last 25 years. Alcohol and Alcoholism, 37(2), 109-120.

[4] P. K. Wood, K. J. Sher, and P. C. RutledgeWiley Online, "College Student Alcohol Consumption, Day of the Week, and Class Schedule," Wiley Online Library, 19-Apr-2007. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1530-0277.2007.00449.x.

[5] V. Singh, R. Agrawal, M. K. Gourisaria, P. Kumar Singh and H. Das, "Comparative Analysis of Machine Learning Models For Early Detection of Fetal Disease using Feature Extraction," 2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT), 2022, pp. 169-175, DOI: 10.1109/CSNT54456.2022.9787635.

[6] M. N. Wernick, Y. Yang, J. G. Brankov, G. Yourganov, and S. C. Strother, "Machine Learning in Medical Imaging," in IEEE Signal Processing Magazine, vol. 27, no. 4, pp. 25-38, July 2010, DOI: 10.1109/MSP.2010.936730.

[7] S. Sarah, M. K. Gourisaria, S. Khare, and H. Das, "Heart disease prediction using Core Machine Learning Techniques-A Comparative Study," SpringerLink, 08-Feb-2022. Available: https://link.springer.com/chapter/10.1007/978-981-16-5689-7_22.

[8] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network," in IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1207-1216, May 2016, DOI: 10.1109/TMI.2016.2535865.

[9] V. Singh, M. K. Gourisaria, G. M. Harshvardhan, and V. Singh, "Mycobacterium tuberculosis detection using CNN ranking approach," SpringerLink, 07-Dec-2021. Available: https://link.springer.com/chapter/10.1007/978-981-16-4369-9_56.

[10] Y. Huang, J. Xu, Y. Zhou, T. Tong, and X. Zhuang, "Diagnosis of Alzheimer's disease via multi-modality 3D Convolutional Neural Network," Frontiers, 31-May-2019. Available: https://www.frontiersin.org/articles/10.3389/fnins.2019.00509/full.

[11] R. Agrawal, V. Singh, M. K. Gourisaria, A. Sharma and H. Das, "Comparative Analysis of CNN Architectures for Maize Crop Disease," 2022 10th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-22), 2022, pp. 1-7, DOI: 10.1109/ICETET-SIP-2254415.2022.9791628.

[12] B. K. Rao, P. S. Kumar, D. K. K. Reddy, J. Nayak, and B. Naik, "QCM sensor-based alcohol classification by Advance Machine Learning Approach," SpringerLink,01-Jan-1970.Available: https://link.springer.com/chapter/10.1007/978-981-15-8439-8_25

[13] A. Pisutaporn, B. Chonvirachkul, and D. Sutivong, "Relevant factors and classification of student alcohol consumption," 2018 IEEE International Conference on Innovative Research and Development (ICIRD), 2018, pp. 1-6, DOI: 10.1109/ICIRD.2018.8376297.

[14] A. K. Shukla, P. Singh, and M. Vardhan, "Predicting alcohol consumption behaviors of the secondary level students," SSRN, 07-May-2018. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3170173.

[15] S. Palaniappan, N. A. Hameed, A. Mustapha, and N. A. Samsudin, "Classification of alcohol consumption among Secondary School students," JOIV,2017.Available:http://joiv.org/index.php/joiv/article/view/64.

[16] P. Jain, S. Khare, and M. K. Gourisaria, "A Data Mining Solution to Predict Campus Placement," 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), 2021, pp. 1-7, DOI: 10.1109/GUCON50781.2021.9573551.

[17] D. S. Sisodia, R. Agrawal, and D. Sisodia, "A comparative performance of classification algorithms in predicting alcohol consumption among Secondary School students," SpringerLink, 01-Jan-1970. Available: https://link.springer.com/chapter/10.1007/978-981-13-0923-6_45.

[18] J. Bi, J. Sun, Y. Wu, H. Tennen, and S. Armeli, "A machine learning approach to college drinking prediction and risk factor identification," ACM Transactions on Intelligent Systems and Technology, 01-Sep-2013. Available: https://dl.acm.org/doi/abs/10.1145/2508037.2508053.

[19] A. Amialchuk, O. Sapci & J. D. Elhai, "Applying machine learning methods to model social interactions in alcohol consumption among adolescents", Taylor and Francis, 22-Feb-2021, Available: https://www.tandfonline.com/doi/full/10.1080/16066359.2021.1887147 DOI: 10.1080/16066359.2021.1887147

[20] K. Dutta, S. Chandra, and M. K. Gourisaria, "Early-stage coronary ailment prediction using dimensionality reduction and data mining techniques," SpringerLink, 01-Jan-1970. Available: https://link.springer.com/chapter/10.1007/978-981-16-3346-1_58.

[21] S. Pal and V. Chaurasia, "Performance analysis of students consuming alcohol using data mining techniques," SSRN, 26-Jun-2017. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2991748.

[22] A. Chordia, S. Sarah, M. K. Gourisaria, R. Agrawal and P. Adhikary, "Surface Crack Detection Using Data Mining and Feature Engineering Techniques," 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), 2021, pp. 1-7, DOI: 10.1109/GUCON50781.2021.9574002.

[23] E. Rosten, R. Porter and T. Drummond, "Faster and Better: A Machine Learning Approach to Corner Detection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 1, pp. 105-119, Jan. 2010, DOI: 10.1109/TPAMI.2008.275.

[24] V. Singh, M. K. Gourisaria, H. GM, and V. Singh, "Mycobacterium tuberculosis detection using CNN ranking approach," SpringerLink, 07-Dec-2021.Available: https://link.springer.com/chapter/10.1007/978-981-16-4369-9_56.