Predicting Student's Alcohol Drinking Habits Using Machine Learning Techniques

Sanjan R[1]
Student,
Department of MCA, Jawaharlal Nehru
New College of Engineering, Shimoga, India

Hemanth Kumar[2]
Associate Professor,
Department of MCA, Jawaharlal Nehru
New College of Engineering, Shimoga, India

Abstract:- Alcohol drinking among college student's is common problem that can leads to low academic performance, health issues and risky behavior. When student's in college, they often experience more freedom, which can result to increased consumption of alcohol. Therefore, this work explores the prediction of student's alcohol drinking habits utilizing Machine Learning techniques. Data set is utilized in this work contains 25 parameters collected by various college students and using this dataset is employed for Machine Learning algorithms, such as Random Forest Classifier, Decision Tree, and Logistic Regression, are employed. This experiment aims to classify students into categories of alcoholic, non-alcoholic and maybe alcoholic based on various influencing factors. The results obtained are contrasted with various Machine Learning techniques.

Keywords:- *Machine Learning, Random Forest Classifier, Decision Tree, Logistic Regression.*

I. INTRODUCTION

In today's world, we are using advanced technology to understand people better. One important area we are studying is how students behave regarding to drinking alcohol. Student addict's alcohol due to various factors, including stress related to academic performance, increased freedom and independence, peer pressure and absence of awareness about responsible drinking habits. Excessive drinking alcohol leads to various negative outcomes, such as poor academic performance, health issues, and social problems. Therefore, it is essential to identify students who are at potential of developing harmful drinking habits at early stages of life. Drinking alcohol is common among college students. Many students drink at parties, social events or just to relax, drinking lot of alcohol leads to serious problems like accidents and injuries. It can affect students mental health, lead to addiction and cause academic issue. To find out students might have drinking problems, we usually ask about how much they drink and do survey on that. But these ways can take longtime. This work is done by employing Machine Learning techniques to make predictions. To implement this, various data is collected from college students that includes their age, how much influence their friends have on them, their family history with alcohol, how interested they are in new things and other factors that could affect their drinking habits. Employing Machine Learning algorithms to categorize students into various groups those who are drinkers, those who maybe drink sometimes, and those who are unlikely to drink. It is implemented by first cleaning up the information to make sure it is accurate and then using it to teach the models how to make predictions. This helps to gain insight into the factors that have an impact on students to drink alcohol and how accurate Machine Learning can occur in predicting their behavior. This information can would be beneficial for parental awareness and college awareness. This experiment helps to know what factors are that making addicting alcohol habits. Knowing these factors, one can help students to stay away from the harmful habits and improve in academics.

II. LITERATURE SURVEY

Based on Predicting Student's drinking alcohol habits using Machine Learning methods, various works has been completed from different authors like Dilip Singh Sisodia et. al. in [1] projected their work by comparing classification algorithms for predicting consumption of alcohol in secondary school students, they used different algorithms in that Logistic Regression and Random Forest models performs well among individual and ensemble classifiers, respectively. Rijad Saric et. al. in [2] proposed a decision tree algorithm to identify addiction of alcohol consumption among high school students and achieve approximately 93% accuracy in predicting alcohol addiction. Shuhaida Ismail et. al. in [3] aimed to predict consumption of alcohol usage among secondary school student's using data mining. They used Decision Tree, k-Nearest Neighbor, Random Forest and Naive Bayes classification algorithms, in that Decision Tree had the best accuracy. Advait Singh et. al. in [4] proposed to predict consumption of alcohol among student's using various Machine Learning(ML) algorithms. They used Linear Regression method, Ridge Regression technique, Lasso Regression, Decision Tree, k-Nearest Neighbor, Random Forest, XGBoost, SVM, ADA Boosting and Gradient Boosting algorithms, in this Random Forest Classifier Algorithm perform well accuracy. Work in [5] study they predicted student alcohol consumption applying educational data mining approaches like k-Nearest Neighbors(K-NN), J48, and Random Forest Classifier, with K-NN and Random Forest Classifier shows best accuracy. Ali Ibrahim et. al. in [6] reviews Machine Learning(ML) methods for forecasting alcohol use disorder from 2010 to 2021. A systematic review was conducted, In ML model Support Vector Machines were most used gives best result. Hind Almayan et. al. in [7] predicts alcohol consumption of student by employing data analysis techniques and machine learning classifiers such as Neural Networks, Random Forest. Results show improved accuracy in predicting consumption of alcohol with fewer features. Santhiya and Nancy Jasmine Goldena in [8] projected two machine learning methods, K-NN and Generalized Linear Model (GLM), to forecast student alcohol use from a dataset. The output show GLM is more accurate with 82% accuracy, while K-NN has 76% accuracy. Ashish Shrestha in [9] used Machine Learning (ML), specifically decision tree classification, to forecast student alcohol abuse built on a dataset of 1033 high school student's. The model achieved a maximum accuracy of 87.93%, highlighting factors affecting alcohol misuse among students. Navdeep Kaur and Williamjeet Singh in [10] proposed study uses J48 and Random Tree algorithms to predict alcoholism based on risk factors in various age groups and achieving higher accuracy with Random Tree 75.94% over J48 71.26%. Saurabh Pal and Vikas Chaurasia in [11] paper examines alcohol's impact on student performance utilizing Decision Tree algorithms BFTree, J48, Simple Cart on data from MCA student's, highlighting BFTree's perform well in this study. Mendoza-Palechor Fabio et. al. in [12] details data mining approaches like Support Vector Machine (SVM) and decision trees, and Bayesian networks to forecast alcohol consumption in Portuguese teenagers using the dataset. In that SVM gives best result. Faruk Bulut et. al. in [13] proposed their work focuses on an urgent precaution system to detect students subjected to substance abuse using classification procedures. They used k-Nearest Neighbors (K-NN), Naive Bayes and decision tree in that K-NN Performs well. Md. Ariful Islam Arif et. al. in [14] explored many machine learning algorithms for alcohol addiction risk, in those logistic regression performers well with 97.91% accuracy. Wendy Wagster et. al. in [15] explored forecasting consumption of alcohol among college student using Artificial Neural Network, they explored Multilayer Perceptron's, Generalized Feedforward Networks and Radial Basis Function (RBF) Networks technique for predicting college student's alcohol consumption based on behavioral factors. Gizem Kapansahin et. al. in [16] examined factors influencing university student's alcohol consumption, finding correlations with age, gender, parental and peer alcohol use, and attitudes toward alcohol. The model achieved a 87.3% correct classification rate.

III. METHODOLOGY

In predicting student's alcohol drinking habits utilizing Machine Learning (ML) technique, experiments involves preprocessing of data, feature engineering, model training, evaluation and prediction. It includes techniques like Random Forest Classifier Method, Decision Tree Classification technique and Logistic Regression. Fig. 1 shows the Machine Learning model.

- **Training Data**: Training data is the initial data set used train the ML model. The data should be well defined and labeled
- **Train ML Algorithm**: Using training data ML algorithm is trained. During this phase, the algorithm learns the pattern and relationship with data. This step involves selecting suitable ML model.
- **Model Input data**: In this process, data gets validate and evaluate the trained model. It is separate from training data.

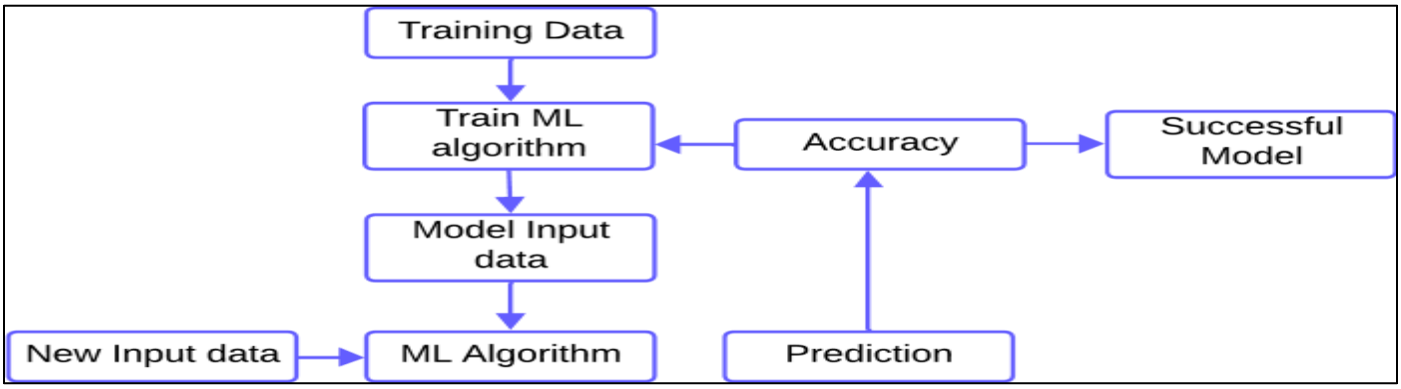


Fig. 1 Machine Learning Model Diagram

- **Training Data**: Training data is the initial data set used train the ML model. The data should be well- defined and labeled
- **Train ML Algorithm:** Using training data ML algorithm is trained. During this phase, the algorithm learns the pattern and relationship with data. This step involves selecting suitable ML model.
- **Model Input data**: In this process, data gets validate and evaluate the trained model. It is separate from training data.
- **ML Algorithm**: The ML Algorithm is selected based on our problem statement. There are many algorithms in ML like Logistic Regression model, Random Forest Classifier approach, Decision Tree etc.,
- **New Input Data**: This is new data that ML model has not seen before. The new data is given for predicting ML model.
- **Prediction**: Predictions are the results and outputs generated by ML algorithm when new input data is given.
- **Accuracy**: Model accuracy is evaluated through various methods including precision, recall, F1-score and confusion matrix. High accuracy means the ML model is good at predicting the outcomes.

- **Successful Model**: If the model achieves satisfactory level of accuracy, it is regarded as successful model.

Three classification algorithms are utilized for predicting student's alcohol drinking habits employing Machine Learning technique, in this work. They are:

- **Random Forest Classifier**: Random Forest Classifier method is ensemble Machine Learning algorithm, it randomly select subset for training data to create multiple bootstrap samples and build a decision tree for every bootstrap sample. For classification problem it gives the result based on majority aggregation (voting) of decision tree results.
- **Decision Tree**: Decision Tree method is supervised Machine Learning approach deployed for both regression and classification problem. It consist root node that represents entire dataset, which spilt into two or more node. Decision node represent where the data spilt based on an attribute. Leaf node represents final decision. Splitting is process of dividing node into two or more node.
- **Logistic Regression**: Logistic Regression algorithm is a supervised method used for binary categorization, this method employed for classification problem. Multinomial Logistic Regression method is a method extension of logistic regression, in logistic regression only binary classification problems are achieved, but in multinomial logistic regression has more number of classifications can be achieved.

IV. PROPOSED MODEL

➢ *Fig. 2 Depicts the Working of Proposed System Model. The Model Includes*

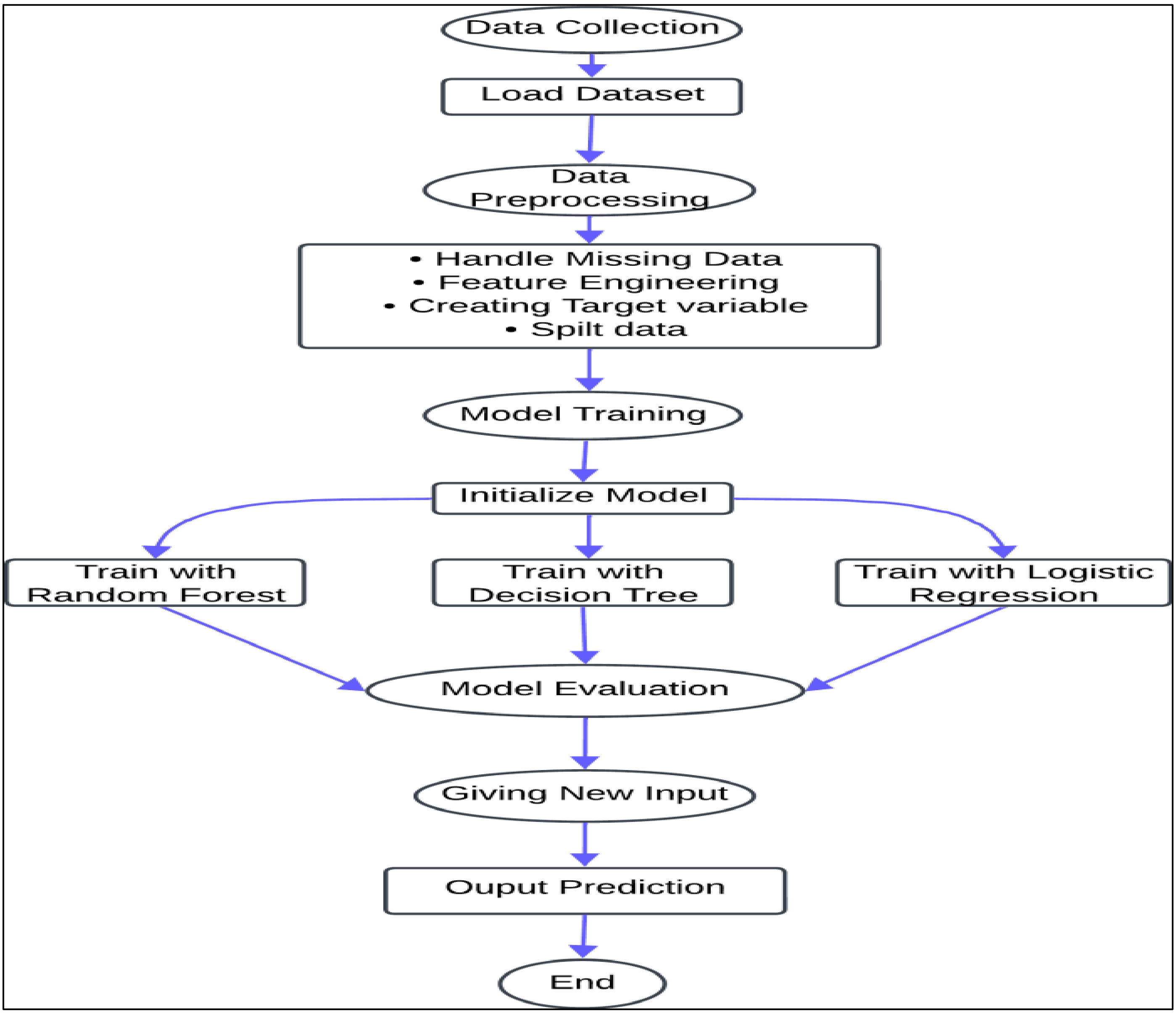


Fig 2 Proposed System Diagram

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1287

- **Data Collection**: The dataset applied in this experiment is collected by various college student, that include Age, Gender, Friend influence, Family Background, Academic Pressure, curiosity, Depression including these and other 25 factors.
- **Data Preprocessing:** This step involves clean and preprocess the data by handling missing values, replace the empty string with 'Nan' (Not a Number), dropping unnecessary columns, converting categorical variables to numerical, handling missing values, splitting data.
- **Data Split**: This step involves the data splitting. Where data is divided into training and testing sets. Split is done on 80% for training set and 20 % for testing set.
- **Model Initialization**: Different models of ML are initialized featuring Random Forest Classifier approach, Decision Tree Classifier algorithm and Logistic Regression techniques.
- **Training Models**: Random Forest Classifier, Decision Tree and Logistic Regression algorithm are trained using training datasets
- **Model Evaluation:** Predictions are made on the test dataset using trained model, evaluating the performance of model is done using accuracy score and classification reports.
- **Predicting New Student Data:** New student's data is given as input to the model for predicting the drinking habits of the student. Output includes that the new student drinks alcohol, may drink alcohol or student is a nondrinker.

V. RESULT

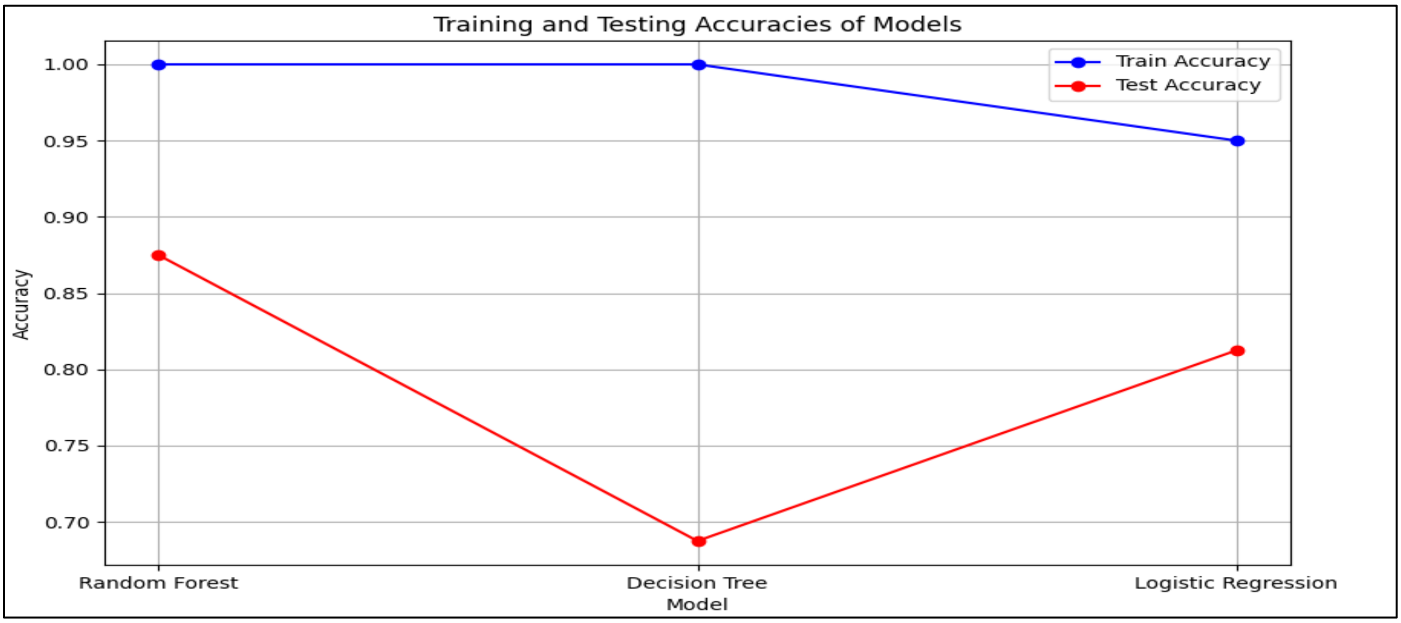
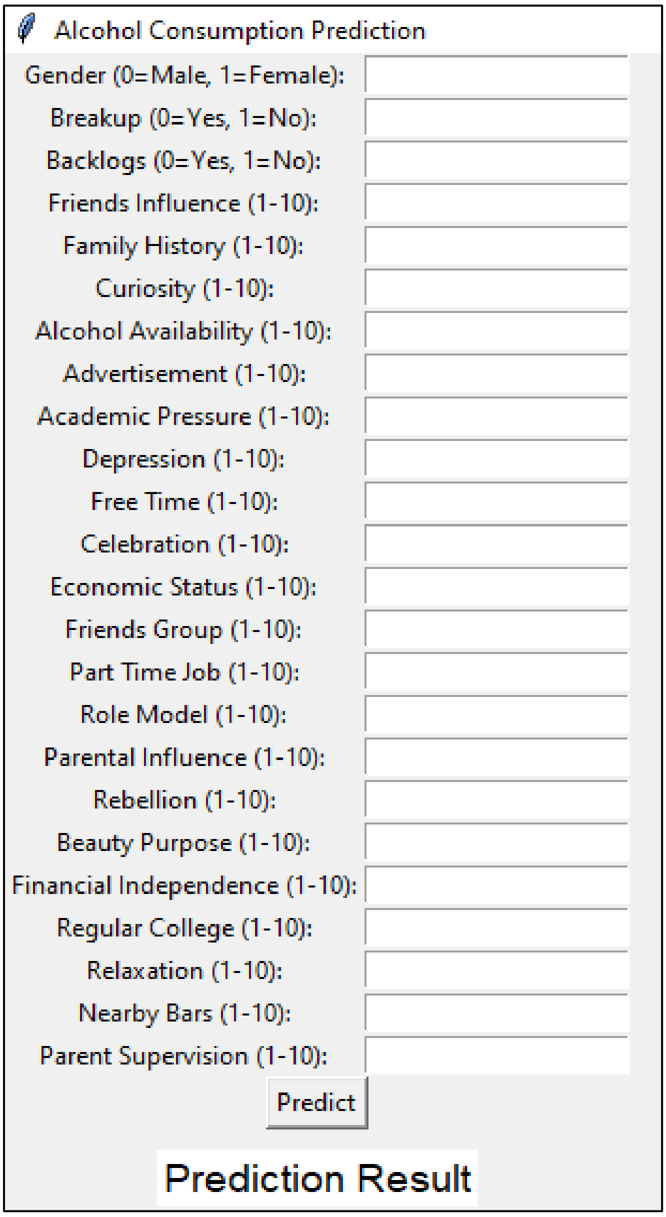


Fig 3 Test and Train Accuracy for Various Algorithm

Fig. 3 gives the detail of test and train accuracy of Random Forest Classifier method, Decision Tree Classifier technique and Logistic Regression, in this result Random Forest classifier technique achieves 88% accuracy, Decision Tree Classifier technique gives 69% accuracy and Logistic Regression gives the 81% accuracy.

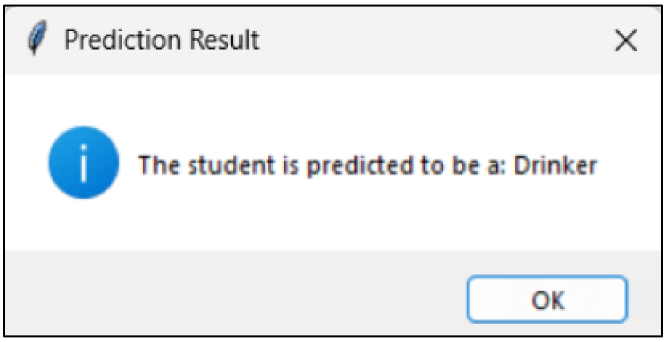


Gender (0=Male, 1=Female):
Breakup (0=Yes, 1=No):
Backlogs (0=Yes, 1=No):
Friends Influence (1-10):
Family History (1-10):
Curiosity (1-10):
Alcohol Availability (1-10):
Advertisement (1-10):
Academic Pressure (1-10):
Depression (1-10):
Free Time (1-10):
Celebration (1-10):
Economic Status (1-10):
Friends Group (1-10):
Part Time Job (1-10):
Role Model (1-10):
Parental Influence (1-10):
Rebellion (1-10):
Beauty Purpose (1-10):
Financial Independence (1-10):
Regular College (1-10):
Relaxation (1-10):
Nearby Bars (1-10):
Parent Supervision (1-10):

Fig 4 User Interface for Predicting New Input Data

Fig. 4 shows, the user interface for predicting new input data, it contains various factors for predicting whether the student is drinker or maybe drinker or not a drinker.

Fig. 5 shows, the result of prediction, the student is predicted to be an alcoholic.



The student is predicted to be a: Drinker

Fig 5 Result of Prediction

VI. CONCLUSION

In this study, implemented Machine Learning approach to predict whether college students are inclined to drink alcohol or not based on various factors. Three Machine Learning approach employed Viz. Random Forest Classifier technique, Decision Tree Classifier technique and Logistic Regression technique. In these three algorithms, Random Forest Classifier gives the best accuracy in predicting student's alcohol drinking habits.

REFERENCES

[1]. Dilip Singh Sisodia, Reenu Agrawal and Deepti Sisodia, "A Comparative Performance Of Classification Algorithms in Predicting Alcohol Consumption among Secondary School Student's", Springer Nature, 2019

[2]. Rijad Saric, Dejan Jokic, and Edhem Custovic, "Identification of Alcohol Addicts among High School Student's Using Decision Tree Based Algorithm", Springer Nature, 2020. Doi: https:10.1007/978-3-030-17971-7_69.

[3]. Shuhaida Ismai, Nik Intan Areena and Nik Azlan, Aida Mustaph, "Prediction of Alcohol Consumption among Portuguese Secondary School Student's: A Data Mining Approach", IEEE 2018.

[4]. Advait Singh, Mahendra Kumar Gourisaria, Vinayak Singh, Ashish Sharma, "Alcohol Consumption Rate Prediction using Machine Learning Algorithms", OITS International Conference on Information Technology (OCIT), 2022. DOI: 10.1109/OCIT56763.2022.00026.

[5]. Tincymol M T and Grace Joseph, "Predicting Alcohol Consumption in Student's Using Data Mining Tool", Proceedings of the National Conference on Emerging Computer Applications (NCECA), Vol.3, Issue.1, 2021.

[6]. Ali Ebrahimi, Uffe Kock Will, Thomas Schmidt, Amin Naemi and Anette Sogaard Nielse and Marjan "Predicting the Risk of Alcohol Use Disorder Using Machine Learning: A Systematic Literature Review", IEEE Access, Vol. 9, November 16, 2021.

[7]. Hind Almayyan and Waheeda Almayyan, "Student Alcohol Consumption Prediction: Data Mining Approach", International Journal of Computer Science and Information Security (IJCSIS), Vol. 16, No. 4, 2018.

[8]. Santhiya and Nancy Jasmine Goldena, "Comparative Analysis Of Alcohol Consumption Prediction by Using Machine Learning Algorithms", International Journal of Creative Research Thoughts (IJCRT) Vol. 10, Issue 12 December 2022.

[9]. Ashish Shrestha, "Predicting Student Alcohol Consumption using Machine Learning", http://aasys.io

[10]. Navdeep Kaur and Williamjeet Singh, "Alcoholic Behavior Prediction through Comparative Analysis of J48 and Random Tree Classification Algorithms using WEKA", Indian Journal of Science and Technology, Vol. 9, August 2016. DOI: 10.17485/ijst/2016/v9i32/100716.

[11]. Saurabh Pal and Vikas Chaurasia, "Is Alcohol Affect Higher Education Student's Performance: Searching and Predicting pattern using Data Mining Algorithms", International Journal of Innovations & Advancement in Computer Science IJIACS, Vol. 6, Issue 4, April 2017.

[12]. MendozaPalechor Fabio, De la HozManotas Alexis, MoralesOrtega Roberto, MartinezPalacio Ubaldo, Diaz Martinez Jorge and Combita Nino Harold, "Designing A Method for Alcohol Consumption Prediction Based on Clustering and Support Vector Machines", Research Journal of Applied Sciences, Engineering and Technology , 2017.

[13]. Faruk Bulut and Ihsan Om Bucak, "An urgent precaution system to detect student's at risk of substance abuse through classification algorithms", Turkish Journal of Electrical Engineering & Computer Sciences, 2014.

[14]. Md. Ariful Islam Arif, Saiful Islam Sany, Farah Sharmin, Md. Sadekur Rahman and Md. Tarek Habib, "Prediction of addiction to drugs and alcohol using machine learning: A case study on Bangladeshi population", International Journal of Electrical and Computer Engineering (IJECE), Vol. 11, No. 5, October 2021.

[15]. Wendy Wagster et al., "Forecasting Alcohol Consumption Trends Among College Student's Using Artificial Neural Network (ANN)", Proceedings ASEE Gulf Southwest Annual Conference, 2005.

[16]. Gizem Kapansahin and Taner Ersoz, "Investigation of the Alcohol Usage Habits of University Student's by Logistic Regression Analysis", Karabuk University Journal of Institute of Social Sciences, Vol. 9, Issue 1, 2019.