Influential Factors Identification in Alcohol Addiction of Adolescents using RELIEF

Chim Tingshing

Faculty of Computer Science and Information Technology, Universiti Malaya, 50603 Kuala Lumpur, Malaysia
u2005412@siswa.um.edu.my

***Abstract*—Alcohol addiction is a prevalent problem worldwide, even among underaged adolescents. Thus, this paper aims to obtain the influential factors that cause alcohol addiction among adolescents from the dataset using RELIEF based algorithms. Three feature selection algorithms, Relief, ReliefF and RReliefF was used on a dataset from Portuguese secondary schools. Four machine learning models were then trained on the selected attributes of each algorithm. A baseline comparison is also conducted by comparing the performance of the models with the traditional supervised machine learning methods without utilizing any feature selection techniques. From the results, RReliefF performed the best. The most influential factors are number of past failed classes, travel time to school, family educational support, age, home area type, mother's highest education, amount of free time, whether the student wants to take higher education or not, family relationship quality, father's highest education and sex. The best model was random forest with tuned hyperparameters which was trained on attributes selected by RReliefF, with F1-score = 0.706, balanced accuracy = 0.794 and ROC Area Under Curve (AUC) = 0.850.**

Keywords—Alcohol addiction; Machine learning; Adolescents; RELIEF.

I. Introduction

Alcohol is a psychoactive substance that affects the functioning of the brain after consumption. Excessive use of alcohol in the long-term are associated with health risks such as high blood pressure, stroke, liver diseases and brain damage (NHS, 2022). Harmful use of alcohol also contributes 3 million deaths worldwide every year, representing 5.3% of all deaths (WHO, 2022). Alcohol addiction, or alcohol use disorder (AUD) is a medical condition characterized by a strong craving of alcohol and lack of control over alcohol consumption (NIAAA, 2021). People who have alcohol addiction are more susceptible to the risks associated with excessive alcohol use, which will bring harm to them.

Alcohol addiction is a prevalent problem worldwide. In fact, according to the NSDUH (2019), there are approximately 1 million people aged 18 or above who received treatment for alcohol use disorder in 2019 itself. In Malaysia, according to the National Health Morbidity Survey 2019, 11.8% of adults aged 18 and above drink alcohol, and 1 in 10 of those adults practice heavy episodic drinking, which is consuming 6 or more standard alcohol drinks in one sitting weekly. Even though binge drinking does not equate alcohol addiction, people who have higher rates of binge drinking are at risk of developing alcohol addiction (Gowin et al., 2017). However, alcohol addiction is not restricted to adults and can also be found in adolescents who are not of legal age to consume alcohol. For example, a cross-sectional study done on 327 secondary school students in Miri, Sarawak showed that 42.2% of adolescents had problematic alcohol use (Chhoa et al., 2019).

Currently, alcohol addiction is identified using AUDIT, a screening tool developed by WHO for alcohol addiction. However, formal medical diagnosis of alcohol addiction follows DSM-5, which is an authoritative guide for medical professionals to diagnose mental disorders. Both AUDIT and DSM-5 processes require discussing and answering questions regarding symptoms for diagnosis. The diagnosis may be a quick process for a single patient, but when the number of patients increases, the process is time-consuming. Regarding this matter, classifying alcohol addiction by using machine learning can overcome the problem mentioned as machine learning algorithms can process large amounts of data in a short time, allowing medical professionals to give immediate treatment to patients classified with alcohol addiction. Classification in machine learning aims to classify each instance of data into predefined categories. It has already been implemented to classify several other types of mental disorders such as depression (Qiao, 2020), bipolar (Jadhav et al., 2019; Sau & Bhakta, 2019) and anxiety (Priya et al., 2020; Sau & Bhakta, 2019).

As of today, multiple studies have already been done to develop machine learning models to classify alcohol addiction among underage adolescents. Ismail et al. (2018) developed multiple machine learning models including decision tree, random forest, k-nearest-neighbor and naive bayes model to classify alcohol addiction among secondary students. On the other hand, Palaniappan et al. (2017) also tried to classify alcohol addiction among secondary students by using deep learning techniques such as MLP and AutoMLP.

Although the studies above have successfully achieved their objectives, there are a few limitations that are present. For example, Ismail et al. (2018) trained the machine learning models on the chosen dataset without any feature selection techniques (known as flat classification), in order to predict alcohol consumption behaviour. Moreover, Palaniappan et al. (2017) manually preselected 15 features from 33 features as inputs, again without using any feature selection techniques.

To address the gaps mentioned above, this study proposes as follows:

1) What are the influential factors to alcohol addiction among adolescents?
   - To obtain the influential factors that cause alcohol addiction among adolescents from the dataset using RELIEF based algorithms

2) What machine learning algorithms can be used to classify alcohol addiction?
   - To develop classification models to classify alcohol addiction among adolescents using the obtained influential factors

3) What are the different metrics that can be used to evaluate the classification models?
   - To assess and report the results of the classification models using evaluation metrics.

The scope of this study includes data collection, data preprocessing, training, testing, and deployment of the machine learning model with the best evaluation score to a web application. The feature selection techniques that will be used are RELIEF based algorithms such as Relief, ReliefF and RReliefF. RELIEF algorithms are classified as a filter model as it selects features based on the characteristics of the data without using any learning algorithms (Wang S. et al., 2017). In addition, the machine learning models that are used include logistic regression (LR), support vector machine (SVM), k-nearest neighbor (KNN), and random forest (RF).

Lastly, this paper is organized in the following manner: Section 2 presents the literature review done; Section 3 discusses the methodology followed throughout the project; Section 4 presents the results obtained; Section 5 discusses findings from results and limitations; Section 6 presents conclusions and future works.

II. Literature Review

During recent years, many studies have been done to research the influential factors that affect alcohol addiction among adolescents. Besides that, similar studies have also been done to investigate the factors associated with binge drinking and alcohol consumption use among adolescents. These studies are relevant to alcohol addiction because as mentioned just now, higher rates of binge drinking equate to higher risk of developing alcohol addiction. Alcohol addicts are also more prone to consume more alcohol because of their addiction.

Maharjan and Mahar (2018) conducted a cross sectional study to investigate the prevalence of alcohol consumption and factors associated with alcohol use. In their study, chi-square test was used to test the statistical significance of various variables with alcohol use. From their tests, they found a significant relationship between monthly pocket money and binge drinking (p-value=0.011). Besides that, a highly significant relationship between alcohol consumption and smoking habit (p-value≤0.0001) was found. In comparison to non-smokers, smokers were more likely to drink alcohol, with a likelihood that is 13.52 times greater.

Arshad et al. (2015) explored the contributing factors to alcohol addiction among youths aged 15 to 24, with the mode age of 17 among participants of their study. Various parametric tests and correlation analysis were done to understand the relationship of the factors and alcohol addiction. They observed that 39.3% of respondents drank alcohol due to their curiosity and desire to try something new. In addition, the weightage of respondents that drank alcohol due to the influence of their family and friends are 14% and 10% respectively. Moving on to alcohol addiction, it was concluded that there was a significant positive relationship between the two factors, interpersonal problems, and availability of alcohol with alcohol addiction among youths. This means that respondents who had more interpersonal problems and more readily access to alcohol had higher chances of having an alcohol addiction.

Wang et al. (2018) performed a cross-sectional study on 442 different schools and gathered a total of 23543 survey responses to investigate the factors affecting binge drinking school students. Multivariable logistic regression was used as a statistical test to test the association between different factors and binge drinking. The statistical tests conducted showed that loneliness, longer durations of exposure to electronic screens, insomnia, previous suicide attempts, poor academic performance, older age, smoking cigarettes along with fighting, bullying and sexual experience all had positive associations with binge drinking. The results of this study further supports the study done by Maharjan and Mahar (2018) that smokers had a greater likelihood to drink alcohol compared to non-smokers.

Mattick et al. (2018) did a cohort study on 1927 adolescents and their parents with the aim of studying the relationship between the supply of alcohol from parental and non-parental figures with the 5 different outcomes, which were binge drinking, alcohol-related harm, alcohol abuse, alcohol dependence, and alcohol addiction. The data used was data spanning 6 years and logistic regression and multinomial logistic regression in the statistical analysis of the data. From the tests, they found that adolescents who received their supply of alcohol from their parents had higher probabilities of binge drinking (p-value<0.0001), alcohol-related harm (p-value<0.0001) and alcohol addiction symptoms (p-value=0.0008) compared to adolescents who did not have supply of alcohol.

Curtis et al. (2018) engaged in a meta-analysis on already published literature to investigate the relationship between engagement of alcohol-related content on social media along with alcohol consumption and alcohol-related problems among adolescents. A total of 19 different literatures were used after screening was done to subset the articles. Through their analysis, they discovered that increased engagement in social media with alcohol-related content has correlations with increased alcohol consumption and alcohol-related problems.

Brunborg et al. (2021) did a study to examine the association between disposable income and binge-drinking among adolescents. They obtained a nation-wide survey sample consisting of 1845 adolescents. Their results showed that adolescents who had a higher income also had higher chances of binge drinking. This may be due to several reasons such as having enough money to purchase alcohol and being able to afford huge amounts of alcohol to binge drink. However, family financial problems were shown to be independent to binge drinking. Thus, the adolescent's own disposable income is what affects binge drinking, not their families.

Similarly to finding the factors that cause influential factors, there are also equally many studies that were conducted in which machine learning models were developed to predict alcohol consumption and alcohol addiction. For instance, Pagnotta et al. (2018) utilized machine learning models to discover the most relevant features affecting alcohol consumption among secondary school students. Survey data from Portuguese secondary schools were used in their study, along with feature selection techniques such as filtering with linear correlation and backwards elimination. The machine learning models used were random forest and decision tree. Results showed that the most influential features that were calculated from the trees of the random forest model were being male, going out more frequently and having more free time with an impact percentage of 25.35%, 21.13% and 9.39% respectively.

Likewise, Pisapatorn et al. (2018) also researched the relevant factors of alcohol consumption using information gain from decision trees and random forest on the student's weekday and weekend alcohol consumption levels. The top 3 most influential factors were the student's gender, their frequency of going out and the amount of study time, having 2 out of 3 same features with the previous study by Pagnotta et al. (2018). Next, they proceeded to use the decision tree and random forest models to predict the weekday and weekend alcohol consumption. The models were tested using 10-fold cross validation. For both weekday and weekend alcohol consumption, random forest model performed better with accuracy of 79.43% and 88.07% respectively.

Palaniappan et al. (2017) used the same dataset as the previous studies, but by using a different method. Instead, 15 features were preselected as input features, while deep learning techniques such as multilayer perceptron (MLP) and AutoMLP were used instead of traditional machine learning models. According to E. Kavlakoglu (2020), deep learning is a subfield of machine learning using neural networks. An accuracy score of 61.78% was achieved using MLP while AutoMLP had a higher score of 64.54%

In addition, Ismail et al. (2018) developed 4 different machine learning models which are decision tree, random forest, KNN and naïve bayes to predict alcohol consumption behaviors students using the same Portuguese secondary school dataset. The 4 machine learning models were fed different types of data using different types of preprocessing techniques. Their results concluded that random forest was the best model, showing the highest weighted mean accuracy of 99.77%, weighted mean recall of 47.07% and weighted mean precision of 58.99%.

Lastly, Afzali et al. (2019) compared seven machine learning algorithms to predict alcohol use among adolescents. The data used consisted of two cohorts of Canadian and Australian adolescents, in total consisting of 6016 adolescents. Instead of traditional feature selection techniques, different clusters of predictors were made and used to train the machine learning models, to evaluate the models in the presence and absence of each cluster of predictors. Elastic-net machine learning model showed the best performance initially and was trained on both cohorts. Next, elastic-net successfully achieved an F1-score of 0.885 and 0.876 for Canada and Australia cohorts respectively.

In every study, there are limitations and gaps present that can be improved in the future. For the literature mentioned above, there are a few reoccurring gaps such as the high dependency on statistical methods, where only statistical methods were utilized for influential factors identification without any classification using machine learning models (Maharjan and Mahar, 2018; Arshad et al., 2015; Wang et al., 2018, Mattick et al., 2018; Curtis et al., 2018; Brunborg et al., 2021). Moving on, even in the studies where machine learning models were developed, a few gaps were present such as the limitation on the evaluation of the machine learning models (Pagnotta et al., 2016), no feature selection techniques or manual preselection of features (Pisataporn et al., 2018; Palaniappan et al., 2017, Ismail et al., 2018) and different measurements of alcohol in data collection from two different cohorts (Afzali et al., 2019).

III. Methodolology

The methodology of the study includes data acquisition, data preprocessing, modelling, model evaluation and deployment. The overview of the methodology is shown in Figure 1.

A. Data Acquisition

The dataset was obtained from Kaggle but was originally composed by P. Cortez and A. Silva (Cortez & Silva, 2008). It contains a total of 33 attributes and consists of questionnaire data from secondary school students who were taking mathematics and Portuguese language courses. There was a total of 395 students from the mathematics class and 649 students from the Portuguese language class. The description of each attribute of the dataset is shown in Table 1.

B. Data Preprocessing

Since the data consisted of students from two classes with shared attributes the data was merged to obtain a total of 1044 students.

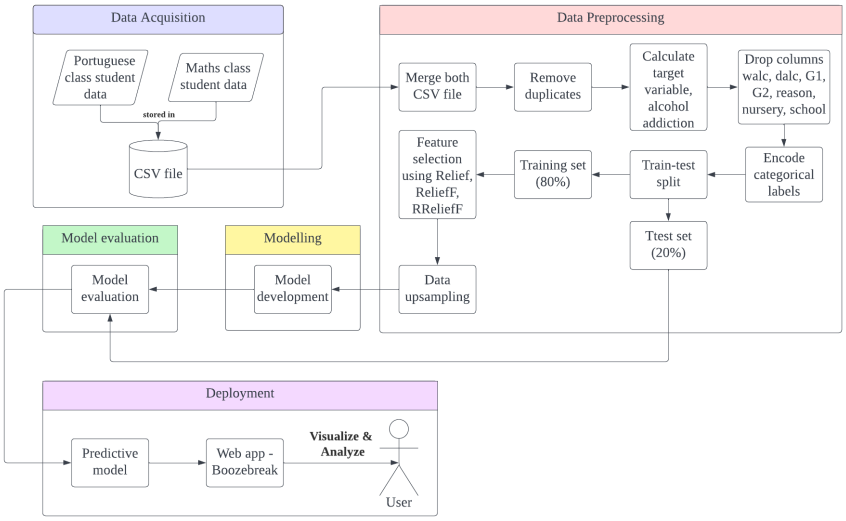


Figure 1: Overview of Methodology

Table 1: Dataset Description

Attribute	Description
school	student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
sex	student's sex (binary: 'F' - female or 'M' - male)
age	student's age (numeric: from 15 to 22)
address	student's home address type (binary: 'U' - urban or 'R' - rural)
famsize	family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus	parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Fedu	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Mjob	mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Fjob	father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
reason	reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
guardian	student's guardian (nominal: 'mother', 'father' or 'other')
traveltime	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures	number of past class failures (numeric: n if 1<=n<3, else 4)
schoolsup	extra educational support (binary: yes or no)
famsup	family educational support (binary: yes or no)
paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
internet	Internet access at home (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	free time after school (numeric: from 1 - very low to 5 - very high)
goout	going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc	weekday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

Next, the data consisted of duplicated rows because Portuguese secondary students were eligible to take both mathematics and Portuguese language courses at the same time. Thus, duplicated records were removed by identifying duplicates using all attributes except for G1, G2, G3 and paid, since those 4 attributes are unique to the mathematics and Portuguese language course themselves.

Since the data did not contain the target variable alcohol addiction, a new variable named addiction was derived from existing attributes using (1) which was proposed by Shukla et al. (2018):

$$addiction = [(Walc \times 2) + (Dalc \times 5)] / 7 \qquad (1)$$

The new target variable, addiction, was then mapped to a binary outcome as shown in (2) following the method proposed by Shukla et al. (2018). This is to measure alcohol addiction, where 1 means having alcohol addiction and 0 means not having an alcohol addiction.

$$\begin{gathered} addiction > 3 = 1 \\ addiction \leq 3 = 0 \end{gathered} \qquad (2)$$

Then, the attributes Dalc and Walc were removed as they were used to calculate the target variable addiction. Besides that, the attributes G1, G2 were dropped due to the attributes having high correlation with G3, as shown in Figure 2. Also, nursery, reason and school were dropped, due to the attributes being irrelevant to the study, as they are very specific and offer very limited generalization.

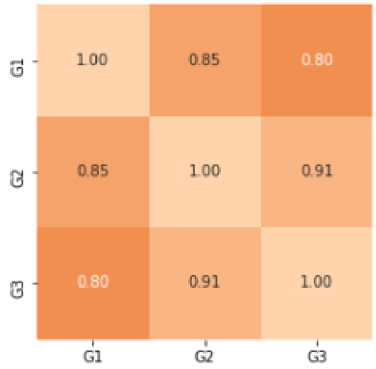


Figure 2: Correlation plot for attributes G1, G2 and G3

The transformer, OrdinalEncoder from the scikit-learn library in Python was used to encode all categorical attributes. This step was conducted as most machine learning models cannot handle categorical data, thus the attributes have to be encoded into numerical format.

The dataset was then divided into 2 sets, which are the training set and the testing set with a ratio of 8:2. The split was done in a stratified manner, ensuring the same ratio of people with addiction and people without addiction in both training and testing set. This is because the original dataset had an unbalanced number of instances for the target variable addiction, as shown in Figure 3. Stratifying will ensure that no sampling errors are present, as both labels of the target variable, addiction will have equal representation in model training and model evaluation steps.

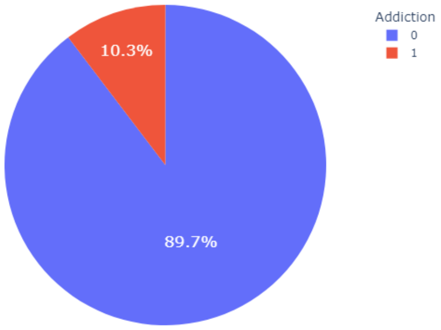


Figure 3: Distribution of Target Variable, Alcohol Addiction

In addition, recursive feature elimination was performed on a decision tree model evaluated using cross validation, with F1-score being utilized as the scoring method. This was performed to determine the number of features to retain from the dataset after feature selection. The results are shown in Figure 4. From the results, the F1-score reaches a maximum when number of features is 11. Thus, 11 attributes were to be selected when conducting feature selection.

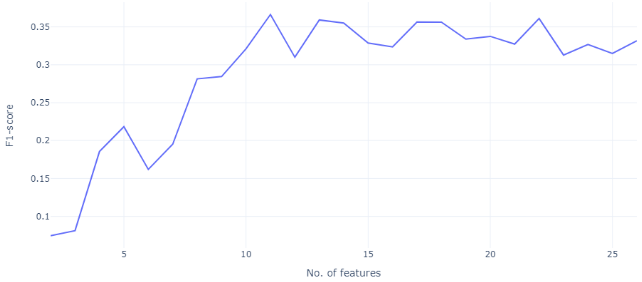


Figure 4: Graph of F1-score against number of features

After that, Relief and RReliefF was implemented using sklearn_relief library in Python while ReliefF was implemented using RapidMiner. After feature selection was done, the training set and test set was further split into training and test sets for each feature selection technique as different sets of attributes were selected by all the algorithms. A training set and test set was also included for the dataset without any feature selection, which acts as a baseline comparison for evaluation. Thus, a total of 4 training sets were present.

Lastly, as mentioned just now, the training and test sets have imbalanced data. Thus, SMOTE oversampling was utilized to upsample the minority class, which are those with alcohol addiction. The minority class was upsampled in all 4 training sets.

C. Modelling

Four machine learning algorithms were utilized in the study, which are logistic regression, SVM, random forest and KNN. The models were trained using scikit-learn library in Python. Initially, the models were trained on the 4 sets of data using default hyperparameters. Then, optimal hyperparameters were searched using GridSearchCV which was implemented using scikit-learn. The parameters for the machine learning models are shown in Table 2, 3, 4 and 5.

D.Evaluation

The evaluation metrics that were used to evaluate the machine learning models include F1-score, balanced accuracy, and ROC Area Under Curve (AUC). F1-score is the mean between precision and recall. Balanced accuracy is the mean between specificity and sensitivity. ROC plots the true positive rate (TPR) against the false positive rate (FPR). The ROC AUC summarizes the curve with a single number, and generally, a higher ROC AUC indicates a better ROC curve and a better evaluated model.

E. Deployment

The best performing model which was random forest with RReliefF as the feature selection technique, along with hyperparameter tuning is used as the predictive model in the web application.

The web application was built using Streamlit, an open-source Python web application framework for data applications. It was primarily chosen due to it being a low-code solution that requires no prior front-end and back-end development experience.

Table 2: Default Hyperparameters of Algoritms

Algorithms	Parameters
LR	penalty='l2', C=1, solver='lbfgs'
SVM	C=1, kernel='rbf', gamma='scale'
RF	n_estimator=100, criterion='gini', max_depth='None', bootstrap=True, max_features='sqrt'
KNN	n_neighbors=5, weights='uniform', algorithm= 'auto', p=2

Table 3: Tuned Hyperparameters for Relief

Algorithms	Parameters
LR	penalty='l1', C=0.1, solver='liblinear'
SVM	C=10, kernel='rbf', gamma=0.1
RF	n_estimator=100, criterion='entropy', max_depth='None', bootstrap=False, max_features='sqrt'
KNN	n_neighbors=2, weights='uniform', algorithm= 'brute', p=1

Table 4: Tuned Hyperparameters for ReliefF

Algorithms	Parameters
LR	penalty='l2', C=1, solver='newton-cg'
SVM	C=10, kernel='rbf', gamma='1'
RF	n_estimator=100, criterion='entropy', max_depth='None', bootstrap=True, max_features='sqrt'
KNN	n_neighbors=3, weights='distance', algorithm= 'brute', p=1

Table 5: Tuned Hyperparameters for RReliefF

Algorithms	Parameters
LR	penalty='l2', C=0.1, solver='saga'
SVM	C=10, kernel='rbf', gamma='1'
RF	n_estimator=300, criterion='gini', max_depth='None', bootstrap=True, max_features='sqrt'
KNN	n_neighbors=3, weights='distance', algorithm= 'brute', p=1

IV. Results

A. Feature Selection

The attributes that were selected using Relief, ReliefF and RReliefF and their weightage are shown in Table 6.

Table 6: Attributes and Weightage for each Feature Selection Algorithm

Relief		ReliefF		RReliefF	
G3	1.065	sex	0.096	failures	0.389
age	0.78	higher	0.075	traveltime	0.337
absences	0.69	guardian	0.072	famsup	0.32
goout	0.635	address	0.065	age	0.291
health	0.585	failures	0.062	address	0.286
Fedu	0.58	famsup	0.036	Medu	0.28
Medu	0.575	paid	0.035	freetime	0.263
freetime	0.56	famsize	0.031	higher	0.257
Mjob	0.52	romantic	0.027	famrel	0.251
famrel	0.475	Pstatus	0.026	Fedu	0.194
Fjob	0.465	famrel	0.025	sex	0.149
studytime	0.46	Fjob	0.025	studytime	0.143
sex	0.43	traveltime	0.02	goout	0.091
famsup	0.43	activities	0.018	guardian	0.074
traveltime	0.405	goout	0.018	internet	0.074
activities	0.31	Medu	0.015	Mjob	0.074
guardian	0.295	Fedu	0.014	health	0.034
romantic	0.27	health	0.014	famsize	0.023
famsize	0.26	age	0.013	schoolsup	0.006
address	0.255	freetime	0.003	romantic	-0.029
paid	0.255	studytime	0.003	paid	-0.04
failures	0.215	G3	0.001	Pstatus	-0.04
internet	0.195	absences	0	activities	-0.057
schoolsup	0.16	internet	-0.011	Fjob	-0.097
Pstatus	0.16	Mjob	-0.019	G3	-0.297
higher	0.08	schoolsup	-0.044	absences	-1.143

NOTE: Highlighted attributes were selected for model training.

B. Model Performance

The evaluation of the models using F1-score, balanced accuracy and ROC AUC is shown in Table 7, 8 and 9.

Table 7: Model Evaluation using F1-score

Algorithm / Type	LR	SVM	RF	KNN
Baseline (D)	0.419	0.364	0.600	0.292
Relief (D)	0.293	0.253	0.588	0.250
ReliefF (D)	0.409	0.380	0.500	0.369
RReliefF (D)	0.386	0.388	0.647	0.358
Relief (T)	0.304	0.222	0.514	0.327
ReliefF (T)	0.409	0.512	0.452	0.457
RReliefF (T)	0.395	0.706	0.706	0.650

NOTE: D – using default hyperparameters; T – using tuned hyperparameters

Table 8: Model Evaluation using balanced accuracy

Algorithm / Type	LR	SVM	RF	KNN
Baseline (D)	0.741	0.725	0.722	0.662
Relief (D)	0.655	0.612	0.738	0.609
ReliefF (D)	0.818	0.798	0.688	0.704
RReliefF (D)	0.777	0.817	0.766	0.698
Relief (T)	0.663	0.565	0.708	0.639
ReliefF (T)	0.818	0.740	0.663	0.680
RReliefF (T)	0.783	0.794	0.794	0.805

NOTE: D – using default hyperparameters; T – using tuned hyperparameters

Table 9: Model Evaluation using ROC AUC

Algorithm / Type	LR	SVM	RF	KNN
Baseline (D)	0.805	0.790	0.944	0.674
Relief (D)	0.653	0.692	0.787	0.654
ReliefF (D)	0.828	0.866	0.855	0.784
RReliefF (D)	0.839	0.849	0.848	0.738
Relief (T)	0.658	0.600	0.794	0.695
ReliefF (T)	0.828	0.788	0.845	0.753
RReliefF (T)	0.836	0.834	0.850	0.824

NOTE: D – using default hyperparameters; T – using tuned hyperparameters

According to Table 7, without any hyperparameter tuning, the best performing model is random forest using RReliefF, with an F1-score of 0.647. With hyperparameter tuning, the best performing models are SVM and random forest with an F1-score of 0.706 for both models.

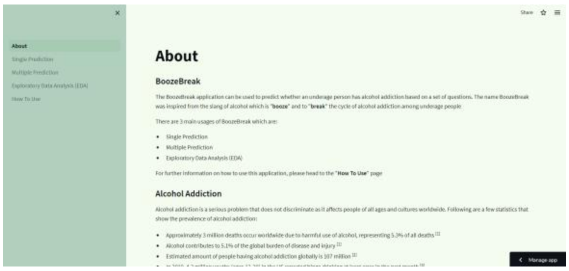
According to Table 8, SVM using default hyperparameters and using RReliefF as the feature selection model had the highest balanced accuracy of 0.817.

According to Table 9, Table 9, without any hyperparameter tuning, the baseline random forest model had the highest ROC AUC of 0.944. However, when using feature selection techniques along with hyperparameter tuning, RReliefF had the highest ROC AUC of 0.850.

By comparing the results from Table 7, 8 and 9, the model with the best performance is random forest using RReliefF as feature selection along with hyperparameter tuning. It is the best performing model due to it having the best F1-score of 0.706, along with good balanced accuracy of 0.794 when compared with other models which have very poor F1-score. Although SVM using RReliefF and hyperparameter tuning has same the F1-score and balanced accuracy with random forest, the ROC AUC for random forest is higher with an ROC AUC of 0.850, thus showing that it is the best model.

C. Web Application

About Page: The landing page for the application is shown in Figure 5. The page consists of a brief explanation on the application's functionality, while also containing brief explanation on alcohol addiction and machine learning along with relevant sources.



About

BoozeBreak

The BoozeBreak application can be used to predict whether an underage person has alcohol addiction based on a set of questions. The name BoozeBreak was inspired from the slang of alcohol which is "**booze**" and to "**break**" the cycle of alcohol addiction among underage people

There are 3 main usages of BoozeBreak which are:

- Single Prediction
- Multiple Prediction
- Exploratory Data Analysis (EDA)

For further information on how to use this application, please head to the "**How To Use**" page

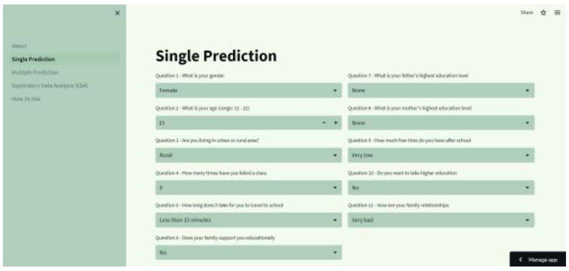
Alcohol Addiction

Alcohol addiction is a serious problem that does not discriminate as it affects people of all ages and cultures worldwide. Following are a few statistics that show the prevalence of alcohol addiction:

- Approximately 3 million deaths occur worldwide due to harmful use of alcohol, representing 5.3% of all deaths [1]
- Alcohol contributes to 5.1% of the global burden of disease and injury [1]
- Estimated amount of people having alcohol addiction globally is 107 million [2]

Figure 5: About Page

Single Prediction Page: Allows users to make prediction of alcohol addiction of adolescents using the best performing model. The threshold to predict an instance as positive can be set by the user. Besides that, the prediction will also show probabilities of each class label, alcohol addiction and no alcohol addiction. The page is shown in Figure 6.



Single Prediction

Question 1 - What is your gender: Female

Question 2 - What is your age (range: 12 - 22): 15

Question 3 - Are you living in urban or rural area?: Rural

Question 4 - How many times have you failed a class: 0

Question 5 - How long does it take for you to travel to school: Less than 15 minutes

Question 6 - Does your family support you educationally: No

Question 7 - What is your father's highest education level: None

Question 8 - What is your mother's highest education level: None

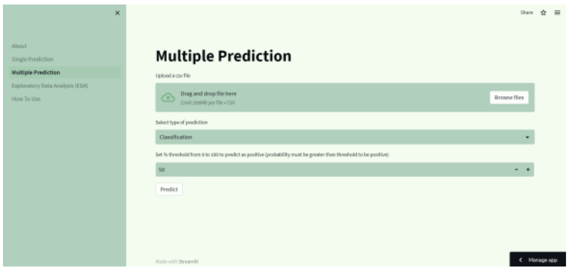
Question 9 - How much free time do you have after school: Very low

Question 10 - Do you want to take higher education: No

Question 11 - How are your family relationships: Very bad

Figure 6: Single Prediction Page

Multiple Prediction Page: Allows users to make multiple predictions by uploading a CSV file. Users are then able to download the new CSV file containing the prediction results. The page is shown in Figure 7.



Multiple Prediction

Upload a csv file

Select type of prediction: Classification

Set % threshold from 0 to 100 to predict as positive (probability must be greater than threshold to be positive): 50

Figure 7: Multiple Prediction Page

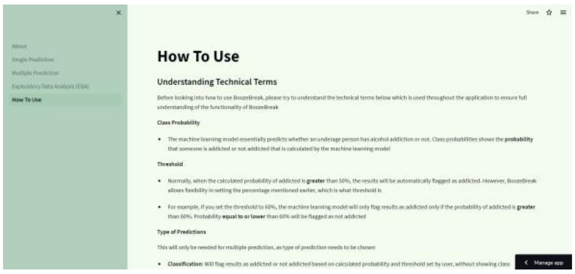
Exploratory Data Analysis (EDA) Page: Allows users to perform EDA on existing data by uploading CSV file. The page is shown in Figure 8.

Exploratory Data Analysis (EDA)

Upload a csv file

Figure 8: Exploratory Data Analysis (EDA) Page

How To Use Page: Gives definition on the terminologies used in the application. Besides that, the steps, prerequisites, and limitations on how to use the web application are also included. The page is shown in Figure 9.



How To Use

Understanding Technical Terms

Before looking into how to use BoozeBreak, please try to understand the technical terms below which is used throughout the application to ensure full understanding of the functionality of BoozeBreak

Class Probability

- The machine learning model essentially predicts whether an underage person has alcohol addiction or not. Class probabilities shows the **probability** that someone is addicted or not addicted that is calculated by the machine learning model

Threshold

- Normally, when the calculated probability of addicted is **greater** than 50%, the results will be automatically flagged as addicted. However, BoozeBreak allows flexibility in setting the percentage mentioned earlier, which is what threshold is
- For example, if you set the threshold to 60%, the machine learning model will only flag results as addicted only if the probability of addicted is **greater** than 60%. Probability **equal to or lower** than 60% will be flagged as not addicted

Type of Predictions

This will only be needed for multiple prediction, as type of prediction needs to be chosen

- **Classification:** Will flag results as addicted or not addicted based on calculated probability and threshold set by user, without showing class

Figure 9: How To Use Page

V. Discussion

The study implemented and compared three different feature selection techniques with a baseline comparison using no feature selection. Hyperparameter tuning was also done after feature selection to obtain the best possible predictive model.

From the results shown earlier, for default parameters, RReliefF performed the best as models with the highest F1-score, balanced accuracy and ROC AUC were trained on attributes selected by RReliefF. As for the influential factors selected by RReliefF, the attributes are shown in Table 6. The most influential factors includes failures, traveltime, famsup, age, address, Medu, freetime, higher, famrel, Fedu and sex.

RReliefF performed better compared to the baseline comparison without using any feature selection because it subsets the most influential attributes, thus less noisy data is present which allows for better predictions by the machine learning models. However, models trained on the attributes selected by Relief performed the worst, even compared to the baseline models. Given that Relief only updates weights based on nearest hit and nearest miss instead of updating weights based on k-nearest-neighbors like ReliefF and RReliefF, it is more susceptible to noisy data which may make its estimations unreliable (Kononenko, 1994). In terms of machine learning models, KNN generally performed worse compared to other models.

The results and findings obtained should be taken in light of the following limitations. First, the dataset used only consisted of students who are attending secondary schools. Adolescents who were not attending school did not participate in in the survey. They may be more susceptible to alcohol addiction, while also having distinct characteristics that might influence the results obtained from the study. Next, the dataset is very imbalanced, with a ratio of 9:1 for those labelled without alcohol addiction and those who were labelled with alcohol addiction. With a more numerically balanced dataset, the quality of predictions will improve, as currently, there are very less samples to train the machine learning models. Even though SMOTE was utilized, having more data is still a better option as SMOTE only creates synthetic examples based on k-nearest-neighbors (Chawla et al., 2002). Lastly, generalizability with the influential factors identified along with the machine learning models on adolescents worldwide might be limited due to the dataset only consisting of students from Portuguese secondary schools.

VI. Conclusion

In conclusion, feature selection techniques successfully extracted the influential factors of alcohol addiction from the given dataset, with RReliefF performing the best. Besides that, four different machine learning algorithms were trained to predict alcohol addiction among adolescents. The best performing model was random forest with tuned hyperparameters that was trained on attributes selected by RReliefF, having an F1-score of 0.706, balanced accuracy of 0.794 and ROC AUC of 0.850. In terms of future works, the dataset for adolescents could include those who are not attending school too. Besides that, a more balanced dataset with more adolescents who have alcohol addiction could be included to improve the quality of the models, and to better identify the influential factors. The dataset could also include data of adolescents from multiple countries and continents, to further improve generalization.

Acknowledgement

I am grateful to all of those who have helped me throughout the course of this project, especially my supervisor, Associate Prof. Ts. Dr. Sri Devi A/P Ravana. Also, I want to thank Myat Noe Win for the guidance and support he has given me in completing this study.

References

Afzali, M. H., Sunderland, M., Stewart, S., Masse, B., Seguin, J., Newton, N., Teesson, M., & Conrod, P. (2019). Machine-learning prediction of adolescent alcohol use: A cross-study, cross-cultural validation. *Addiction, 114*(4), 662-671.

Arshad, M., Omar, M., & Shahdan, N. A. (2015). Alcoholism among youth: A case study in Kuala Lumpur, Malaysia. *International Journal of Culture and History, 1*(1), 21-28.

Brunborg, G. S., Von Soest, T., & Andreas, J. B. (2021). Adolescent income and binge drinking initiation: prospective evidence from the MyLife study. *Addiction, 116*(6), 1389-1398.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321-357.

Chhoa, K. H., Zakaria, H., & Abd Rahman, F. N. (2019). Problematic alcohol use and depression in secondary school students in Miri, Malaysia. *Pediatrics international, 61*(3), 284-292.

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance.

Curtis, B. L., Lookatch, S. J., Ramo, D. E., McKay, J. R., Feinn, R. S., & Kranzler, H. R. (2018). Meta-analysis of the association of alcohol-related social media use with alcohol consumption and alcohol-related problems in adolescents and young adults. *Alcoholism: Clinical and Experimental Research, 42*(6), 978-986.

Gowin, J. L., Sloan, M. E., Stangl, B. L., Vatsalya, V., & Ramchandani, V. A. (2017). Vulnerability for alcohol use disorder and rate of alcohol consumption. *American Journal of Psychiatry, 174*(11), 1094-1101.

Institute for Public Health Malaysia. (2019). NationalHeal and Morbidity Survey (NHMS)2019: Non-Communicable Diseases, Healthcare Demand and Health Literacy. (Publication No. NMRR-18-3085-44207).

Ismail, S., Azlan, N. I. A. N., & Mustapha, A. (2018). Prediction of alcohol consumption among Portuguese secondary school students: A data mining approach. In *2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)* (pp. 383-387). IEEE.

Jadhav, R., Chellwani, V., Deshmukh, S., & Sachdev, H. (2019). Mental disorder detection: Bipolar disorder scrutinization using machine learning. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 304-308). IEEE.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. In *European conference on machine learning* (pp. 171-182). Springer, Berlin, Heidelberg.

Maharjan, P., & Magar, K. (2017). Prevalence of alcohol consumption and factors associated with the alcohol use among the youth of suryabinayak Municipality, Bhaktapur. *J Pharma Care Health Sys, 4*(1), 68.

Mattick, R. P., Clare, P. J., Aiken, A., Wadolowski, M., Hutchinson, D., Najman, J., Slade, T., Bruno, R., McBride, N., & Kypri, K. (2018). Association of parental supply of alcohol with adolescent drinking, alcohol-related harms, and alcohol use disorder symptoms: a prospective cohort study. *The Lancet Public Health, 3*(2), e64-e71.

National Institute on Alcohol Abuse and Alcoholism. (2021). *The cycle of alcohol addiction*. Retrieved January 9, 2023, from https://www.niaaa.nih.gov/publications/cycle-alcohol-addiction

NHS (2022, October 4). *Alcohol misuse – Risks*. Retrieved January 9, 2023 from https://www.nhs.uk/conditions/alcohol-misuse/risks/

Pagnotta, F., & Amran, H. (2016). Using data mining to predict secondary school student alcohol consumption. *Department of Computer Science, University of Camerino*, 1-9.

Palaniappan, S., Hameed, N. A., Mustapha, A., & Samsudin, N. A. (2017). Classification of alcohol consumption among secondary school students. *JOIV: International Journal on Informatics Visualization, 1*(4-2), 224-226.

Pisutaporn, A., Chonvirachkul, B., & Sutivong, D. (2018). Relevant factors and classification of student alcohol consumption. In *2018 IEEE international conference on innovative research and development (ICIRD)* (pp. 1-6). IEEE.

Priya, A., Garg, S., & Tigga, N. P. (2020). Predicting anxiety, depression and stress in modern life using machine learning algorithms. *Procedia Computer Science, 167*, 1258-1267.

Qiao, J. (2020). A systematic review of machine learning approaches for mental disorder prediction on social media. 2020 International Conference on Computing and Data Science (CDS), (pp. 433-438). IEEE

QSAMHSA, CBHSQ. (2019). 2019 National Survey on Drug Use and Health. Table 7.56A—Received Alcohol Use Treatment at Any Location in Past Year among Persons Aged 12 or Older with Past Year Alcohol Use Disorder, by Age Group: Numbers in Thousands, 2015-2019. Retrieved January 9, 2023 from https://www.samhsa.gov/data/sites/default/files/reports/rpt29394/NSDUHDetailedTabs2019/NSDUHDetTabsSect7pe2019.htm

Sau, A., & Bhakta, I. (2019). Screening of anxiety and depression among the seafarers using machine learning technology. *Informatics in Medicine Unlocked, 16*, 100149.

Shukla, A. K., Singh, P., & Vardhan, M. (2018). Predicting alcohol consumption behaviours of the secondary level students. In *Proceedings of 3rd international conference on internet of things and connected technologies (ICIoTCT)* (pp. 26-27).

Wang, H., Hu, R., Zhong, J., Du, H., Fiona, B., Wang, M., & Yu, M. (2018). Binge drinking and associated factors among school students: a cross-sectional study in Zhejiang Province, China. *BMJ open, 8*(4), e021077.

Wang, S., Tang, J., & Liu, H. (2017). Feature Selection. *Encyclopedia of Machine Learning and Data Mining* (2017): 503-511.

World Health Organization. (2022, May 9). *Alcohol*. Retrieved January 9, 2023 from https://www.who.int/news-room/fact-sheets/detail/alcohol