WIH3001

Data Science Project

Project Proposal

Influential Factors Identification In Alcohol Addiction Using RELIEF

Prepared By

Chim Tingshing

U2005412

Semester 1

2022/2023

Supervisor

Associate Prof. Ts. Dr. Sri Devi Ravana

Table of Contents

1.0 Project Title 3
2.0 Introduction 3
3.0 Problem statement 4
4.0 Research questions 5
5.0 Objectives 6
6.0 DS Methodology 6
6.1 Problem Understanding 6
6.2 Data Acquisition 6
6.3 Data Cleaning 7
6.4 Calculating Target Variable 7
6.5 Exploratory Data Analysis (EDA) 8
6.6 Data Preprocessing 8
6.7 Modelling and Evaluation 9
6.8 Deployment 10
7.0 References 11

1.0 Project Title

The title of the project is “Influential factors identification in alcohol addiction using RELIEF”. I plan to use the influential features obtained from a group of students undertaking math and Portuguese language courses in secondary school and find the influential factors using RELIEF to train a classification model which can predict alcohol addiction.

2.0 Introduction

Alcohol is a substance that has been widely used worldwide in many different cultures throughout multiple centuries. Alcohol has remained prevalent throughout the test of time because it acts as a social binder. Alcohol has remained prevalent for such a long time despite being classified as a psychoactive substance as it may influence how our brains work after consumption. According to the National Institute on Alcohol Abuse and Alcoholism (NIAAA), examples of the aftereffects of consumption of alcohol include changes in mood, behaviour and disrupted coordination. Alcohol acts as a social binder as it can positively affect someone’s mood when they consume it together with a platonic friend in a non-stress environment (Molly A. Bowdring & Michael A. Sayette, 2021). However, harmfully abusing alcohol has the potential to cause diseases and have significant social and economic consequence. Thus, the classification of alcohol addiction can help an individual as they are able to diagnose their addiction and find ways to help themselves overcome it.

It is important to understand what alcohol addiction is to further understand the project and its initiatives. Addiction can be defined as having no self-control over what you’re doing and consuming something to a point where it may harm you (National Health Service, 2021). Having an alcohol addiction essentially means that you have no control over your alcohol intake and you abuse the consumption of alcohol, causing dangerous effects on your body and its health. Alcohol addiction is associated with the compulsive drinking of alcohol, lack of control over its intake and also the negative emotional reaction when alcohol is not available, hence a withdrawal reaction (NIAAA, 2021). In 2021, the NIAAA mentioned that there are 3 stages to alcohol addiction. The first stage is named the intoxication stage and this stage is when a person enjoys the positive effects of alcohol to

the body such as euphoria and anxiety reduction during social interactions. The following stage is the withdrawal stage where one experiences physical and emotional withdrawal symptoms when they stop drinking alcohol. The final stage is the anticipation stage where an individual seeks alcohol consumption after abstaining for a period of time. Therefore, even though alcohol may positively affect someone's mood as mentioned just now, abusing alcohol to the point of addiction is a serious problem as both an individual's physical and emotional state will be affected by it.

Besides classifying alcohol addiction in an individual, RELIEF will be used as the feature selection method in this project. Feature selection is a technique which reduces the number of input variables and allows us to find the best subset of features for a predictive model. RELIEF was first introduced in 1992 by Kira and Rendell in their paper titled "The Feature Selection Problem: Traditional Methods and a New Algorithm". In their paper, Kira and Rendall (1992) mentioned that RELIEF is a feature weight-based algorithm which chooses features that are statistically relevant to the target variable. RELIEF can be considered a filter model as filters models select features based on the characteristics of the data (Wang S. et al., 2017). By reducing the number of features through feature selection techniques such as RELIEF, increased accuracy of the machine learning can be achieved along with decreased computational time as lesser number of features are used (M. Ramaswami & R. Bhaskaran, 2009).

3.0 Problem statement

Throughout the globe, there is an estimated of 3 million every year which are the result of harmful use of alcohol, which in total represents 5.3% of all deaths (WHO, 2022). Next, WHO (2022) also mentions that alcohol contributes to 5.1% of the global burden of disease and injury, while also being attributable to 13.5% of deaths among people aged 20 – 39 years. All the statistics mentioned has shown the potential harm that alcohol can cause. Alcohol addiction is also a very big problem as victims tend to consume a lot of alcohol, bringing harmful effects to themselves and potentially the people around them. It is estimated that around 107 million people globally have an alcohol addiction (Ritchie H. & Roser M., 2022). As seen, the dangers of alcohol accompanied by the addiction to it is a prevalent problem worldwide.

Moving on, even though Malaysia's official religion is Islam and the majority of the population practices Islam, alcohol addiction is still a prevalent problem in Malaysia. In 2019, a survey done by the National Institutes of Health Malaysia discovered that the prevalence of heavy episodic drinkers among the population was about 1%. Heavy episodic drinkers were defined as people who consumed 6 or more standard alcohol drinks at one sitting at least once per week. To explain in terms of numbers, 1% of the population means that an estimated population of 212,144 adults aged 18 years and above in Malaysia were heavy episodic drinkers. Abusing alcohol does not only affect an individual's health, but it may also affect the lives of innocent people. For instance, a study done on 16698 inmates showed that the more intoxicated the inmates, the higher the chances of them committing violent crimes and burglary (Richard B. & Jeremy S., 2010). The same study suggests that alcohol plays a role in homicide and physical assault even though the offenders are only drinking a moderate amount. It is indisputable that alcohol intoxication may cause violence, thus people who are addicted to alcohol are also more susceptible to this problem as they are highly likely to consume more alcohol.

In a nutshell, we know that alcohol can both positively and negatively affect someone, depending on the circumstances during consumption it and whether they are addicted to it which was explained in the previous sections. Also, the prevalence of alcohol addiction still exists in Malaysia despite most of the population practicing Islam. Hence, for the scope of the project, to help Malaysian people identify whether they're addicted to alcohol, I plan to take a machine learning approach. I will use RELIEF to find the influential factors that affect addiction to alcohol, and I will use those features to train multiple classification models while comparing the obtained results.

4.0 Research questions

- What are the influential factors to alcohol addiction?
- What machine learning algorithms can be used to classify alcohol addiction?
- What are the different metrics that can be used to evaluate the classification model?

5.0 Objectives

- To obtain the influential factors that cause alcohol addiction from the dataset using RELIEF
- To develop classification models to classify alcohol addiction using the obtained influential factors
- To assess and report the results of the classification models using evaluation metrics

6.0 DS Methodology

6.1 Problem Understanding

Before jumping into exploring the data and modelling it, it is important to first understand the problem that is going to be solved. In this project, the aim is to investigate the influential factors of alcohol addiction using RELIEF. Next, supervised machine learning models will be trained to predict whether someone is addicted to alcohol based on certain characteristics.

6.2 Data Acquisition

The dataset acquired was obtained through Kaggle, which is an online community platform which contains datasets published by other users. The dataset is named “Student Alcohol Consumption” and it was made publicly available by the University of California Irvine. A study was first done on the dataset in 2008 by P. Cortez and A. Silva in their research titled “Using data mining to predict secondary school student performance”. From the study, it was found that a student’s academic achievements are largely affected by their previous performances. Nonetheless, the dataset consists of data from a questionnaire answered by students from a secondary school who were taking mathematics and Portuguese language courses. It contains a total of 33 columns with characteristics such as age, parents’ job, internet access and their alcohol consumption. However, the dataset does not consist of the target variable needed, which is alcohol addiction. This will be dealt with in the later steps as a target variable is needed to train the supervised machine learning models.

6.3 Data Cleaning

Data cleaning can be defined as the process of converting raw data into clean data. In the process, the fixing and removing of inaccurate, duplicate and incomplete data is done within the dataset. There are many important reasons to why data cleaning is essential in the methodology of a data science project. The quality of a predictive model is as good as the data that is fed to it (J. Brownlee, 2020). This fact is normally known as garbage in, garbage out. Garbage is used to describe the quality of the data. If low quality data is used to train the predictive model, the quality of the predictive model will be low as well.

Moving on, if duplicate data is found in the dataset, it will be removed. As for empty data fields, there are multiple ways to deal with this type of data. The common ways include:

- Deletion
  - The record that contains missing data is completely deleted

- Imputation
  - The missing data is substituted with a value. There are several types of imputation techniques, such as mean imputation, hot deck imputation, cold deck imputation and regression imputation

6.4 Calculating Target Variable

Since the dataset does not contain the target variable that is needed which is alcohol addiction, it is important that we populate this column through calculations of other columns in the dataset. The formula that is going to be used to calculate alcohol addiction is as below:

$$addiction = [(walc \times 2) + (dalc \times 5)] \div 7$$

$where$:

$walc = weekend\ alcohol\ consumption\ (1 - very\ low\ to\ 5 - very\ high)$

$dalc = weekday\ alcohol\ consumption\ (1 - very\ low\ to\ 5 - very\ high)$

The walc column and dalc column will then be removed from the dataset as the columns are not relevant anymore due to the presence of addiction which was derived from them.

6.5 Exploratory Data Analysis (EDA)

The following step is exploratory data analysis. In this step, the focus is on further understanding the data, finding trends or insights through visualizations. Data visualization is also a very important step because it helps by highlighting the trends and outliers of the data, showing it in a pleasant way that is easier to understand compared to a table of values. Besides that, correlation analysis can also be done to identify the strength of the correlation between features and also the target variable.

6.6 Data Preprocessing

- Feature Selection

  As mentioned previously, RELIEF will be used as the find the influential features that cause alcohol addiction, allowing the size of the dataset to be smaller subsequentially.

- Feature Scaling

  Feature scaling is normally used on datasets that have columns with different ranges to help the machine learning algorithm converge faster. There are a few basic scaling techniques that can be applied such as:

  i. Min-Max Scaling

  Min-max scaling is normally used to transform features to have a similar scale of range [0, 1]. The formula for min-max scaling is as below:

$$X_{new} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

$$where$$

$$\min(x) = minimum\ value\ of\ feature\ x$$

$$\max(x) = maximum\ value\ of\ feature\ x$$

  A disadvantage of using min-max scaling is that it is very susceptible to outliers. A very big outlier will have a huge effect on the data. For example, if there are 9 values of X which have a range of [10, 40], and the tenth value is 1000, then the rest of the 9 values will be transformed to a value between 0.01 to 0.04.

ii. Z-Score Normalization

Z-score normalization on the other hand deals with the outlier issue. The formula for Z-score normalization is as below:

$$X_{new} = \frac{x - \min(x)}{\sigma}$$

$where$

$\min(x) = minimum\ value\ of\ feature\ x$

$\sigma = standard\ deviation\ of\ feature\ x$

However, although Z-score normalization can deal with outliers, there is still a disadvantage as the features are not on the exact same scale. Thus, there is no perfect solution and the technique used will depend on the dataset.

6.7 Modelling and Evaluation

Firstly, before modelling the data, the data will be split into training and test data. After that, the training data will be used to train the machine learning model while the test data will be used to evaluate the model. There are a few machine learning algorithms that are going to be used to train the classification model which are as follows:

- Logistic Regression
- Naive Bayes
- Support Vector Machine
- K-Nearest Neighbor

After training has been completed, the test data is used to evaluate the classification models. There are multiple ways to evaluate classification models such as the following (non-exhaustive):

- Confusion matrix
  - A confusion matrix is a table with 4 different combinations of predicted and actual values. The confusion matrix contains the true positive, false positive, true negative and false negative which are important for other evaluation metrics.

- Recall

- Recall is used to quantify how many positive classes were correctly predicted as positive from all the positive classes that exist in the dataset. The formula for recall is as follows:

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

$$where$$

$$TP = true\ positive$$

$$FN = false\ negative$$

Besides what was mentioned in the above list, there are also other evaluation metrics that may be used to evaluate and assess the classification models.

6.8 Deployment

The model that shows the best results will be deployed onto a web application. The web application will be built using Streamlit. Streamlit is an open-source Python web application framework which is used for data science and machine learning to share data apps. Streamlit was chosen mainly due to the time constraints as it is a low-code solution. Other web application frameworks such as Django, ASP.NET and Angular are able to build better large-scale web applications, but the time needed is longer, thus Streamlit as an alternative was chosen.

7.0 References

Addiction: what is it? (2021) *NHS choices*. NHS. Available at: https://www.nhs.uk/live-well/addiction-support/addiction-what-is-it/ (Accessed: November 6, 2022).

Alcohol (2022) *World Health Organization*. World Health Organization. Available at: https://www.who.int/news-room/fact-sheets/detail/alcohol (Accessed: November 6, 2022).

Alcohol's effects on the body (no date) *National Institute on Alcohol Abuse and Alcoholism*. U.S. Department of Health and Human Services. Available at: https://www.niaaa.nih.gov/alcohols-effects-health/alcohols-effects-body (Accessed: November 6, 2022).

Bowdring, M. A., & Sayette, M. A. (2021). The effect of alcohol on mood among males drinking with a platonic friend. *Alcoholism, clinical and experimental research*, *45*(10), 2160–2166. https://doi.org/10.1111/acer.14682

Brownlee, J. (2020). *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*: Machine Learning Mastery.

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

Felson, R.B. and Staff, J. (2010) "The effects of alcohol intoxication on violent versus other offending," *Criminal Justice and Behavior*, 37(12), pp. 1343–1360. Available at: https://doi.org/10.1177/0093854810382003.

Institute for Public Health (IPH) 2019. *National Health and Morbidity Survey (NHMS) 2019: Non-Communicable Diseases: Risk Factors and other Health Problems*. (Publication No. NMRR-18-3085- 44207). Federal Territories of Kuala Lumpur, Malaysia.

Kira, K., & Rendell, L. A. (1992). *The Feature Selection Problem: Traditional Methods and a New Algorithm.* Paper presented at the AAAI.

Ramaswami, M., & Bhaskaran, R. (2009). A Study on Feature Selection Techniques in Educational Data Mining. *Journal of Computing, 1*.

Ritchie, H. and Roser, M. (2018) *Alcohol consumption, Our World in Data*. Available at: https://ourworldindata.org/alcohol-consumption (Accessed: November 6, 2022).

The cycle of alcohol addiction (2021) *National Institute on Alcohol Abuse and Alcoholism*. U.S. Department of Health and Human Services. Available at: https://www.niaaa.nih.gov/publications/cycle-alcohol-addiction (Accessed: November 6, 2022).

Wang, S., Tang, J., & Liu, H. (2016). Feature Selection. In (pp. 1-9).