Diversitas Journal
ISSN 2525-5215
Volume 10, Issue 1 (Jan../Mar. . 2025) p. 0480 – 0500
https://diversitasjournal.com.br/diversitas_journal

Using Naïve Bayes Algorithm to Predict and Classify Alcohol Addiction Severity: A Machine Learning Approach for Public Health Interventions

BALAZON, Francis G. [1]

[1] 0000-0003-0143-2983; College of Teacher Education Graduate School, Batangas State University The National Engineering University, Philippines, francis.balazon@g.batstate-u.edu.ph

ABSTRACT

Alcohol addiction is a critical global health concern, with traditional methods of predicting and classifying addiction levels often falling short in accuracy and applicability. This study aims to address these limitations by employing the Naïve Bayes Algorithm, a probabilistic machine learning model, and K-means Clustering to predict and classify alcohol addiction severity. Data was collected through a comprehensive survey of 500 participants, examining alcohol consumption frequency, underlying causes, and associated health impacts. The Naïve Bayes Algorithm achieved notable performance metrics, including an accuracy of 95%, precision of 93%, recall of 97%, and an F1 score of 95%. Simultaneously, K-means Clustering effectively categorized addiction into three distinct levels: less addicted, moderately addicted, and highly addicted. This classification provides healthcare professionals with actionable insights to tailor interventions and develop personalized treatment strategies. Compared to existing methods, the combined use of these algorithms demonstrates enhanced accuracy and reliability, offering a robust framework for addressing addiction severity. This research not only advances the use of machine learning in healthcare but also lays the groundwork for future studies integrating diverse algorithms and exploring broader dimensions of addiction.

ARTICLE INFORMATION

Article process:
Submitted: 08/25/2024
Approved: 03/28/2025
Published: 03/28/2025

Keywords:
Machine Learning,
Naïve Bayes Algorithm,
K-means Clustering,
Alcoholism,
Alcohol Addiction

RESUMO

A dependência alcoólica é uma preocupação crítica de saúde global, com métodos tradicionais de previsão e classificação dos níveis de dependência frequentemente apresentando falhas em precisão e aplicabilidade. Este estudo busca abordar essas limitações empregando o Algoritmo Naïve Bayes, um modelo probabilístico de aprendizado de máquina, e o K-means Clustering para prever e classificar a gravidade da dependência alcoólica. Os dados foram coletados por meio de uma pesquisa abrangente com 500 participantes, examinando a frequência de consumo de álcool, causas subjacentes e impactos associados na saúde. O Algoritmo Naïve Bayes alcançou métricas de desempenho notáveis, incluindo uma precisão de 95%, precisão (precision) de 93%, sensibilidade (recall) de 97% e uma pontuação F1 de 95%. Simultaneamente, o K-means Clustering categorizou eficazmente a dependência em três níveis distintos: menos dependente, moderadamente dependente e altamente dependente. Essa classificação oferece aos profissionais de saúde insights práticos para adaptar intervenções e desenvolver estratégias de tratamento personalizadas. Em comparação com os métodos existentes, o uso combinado desses algoritmos demonstra maior precisão e confiabilidade, oferecendo uma estrutura robusta para abordar a gravidade da dependência. Esta pesquisa não apenas avança o uso do aprendizado de máquina na saúde, mas também estabelece as bases para estudos futuros que integrem algoritmos diversos e explorem dimensões mais amplas da dependência.

Palavras Chava:
Aprendizado de máquina,
Algoritmo Naïve Bayes,
Agrupamento K-means,
Alcoolismo,
Dependência de álcool

DOI: 10.48017/dj.v10i1.3131

Introduction

Alcohol consumption is deeply embedded in numerous cultures worldwide. While it often plays a role in social and cultural events, its profound impact on public health cannot be overlooked. Immediate health risks linked with alcohol include accidents, injuries, and violence (Sudhinaraset et al., 2016). Chronic consumption has been associated with liver diseases such as cirrhosis, cardiovascular ailments, certain cancers, and mental health disorders (Centers for Disease Control and Prevention [CDC], 2022). Although moderate drinking has been linked to some protective effects on conditions like heart disease (Chiva-Blanch & Badimon, 2019), the overarching consensus emphasizes that the risks far outweigh the benefits (CDC, 2022).

Alcohol addiction, characterized by harmful patterns of alcohol use, remains a global public health challenge (Ali et al., 2011). Precision in understanding and classifying addiction levels is critical for designing effective interventions. Tailored strategies that hinge on accurate classification can significantly enhance recovery outcomes for individuals battling addiction (American Psychological Association, 2012).

The growing application of machine learning in healthcare presents new opportunities to address these challenges (Habehh & Gohel, 2021). Among the various algorithms available, the Naïve Bayes algorithm, a probabilistic machine learning tool rooted in Bayes' Theorem, was utilized in this study. The algorithm is widely recognized for its simplicity, speed, and accuracy in classification tasks, making it an ideal instrument for predicting and categorizing addiction levels based on individual health data.

In this research, the Naïve Bayes algorithm serves as a computational tool to classify alcohol addiction into three levels: less addicted, moderately addicted, and highly addicted. By processing survey data from 500 individuals, the algorithm offers a structured approach to understanding addiction severity. When combined with K-means clustering, it enables healthcare professionals to derive actionable insights into patterns of addiction. These insights are invaluable for creating tailored treatment plans aimed at mitigating addiction's long-term health effects.

This study highlights the algorithm's role as more than just a predictive tool—it is a foundation for developing innovative public health strategies. By accurately assessing addiction severity, the Naïve Bayes algorithm can indirectly help healthcare providers address the health consequences of alcohol misuse. Through its transformative approach, this research bridges technology and healthcare, providing a scalable solution for understanding and combating alcohol addiction.

Research Objective

The objective of this research is to leverage the Naïve Bayes Algorithm and K-means Clustering to develop a comprehensive model for predicting and classifying alcohol addiction levels. By analyzing data from 500 individuals, this study aims to offer refined categorizations of addiction severity, thereby providing a robust tool to support healthcare practitioners in devising personalized treatment strategies.

Methodology

Data Collection a

The data for this research was collected through comprehensive questionnaires and online surveys administered via Google Forms. These instruments were specifically designed to assess the severity of alcohol consumption behaviors among participants. The questionnaires were divided into two primary sections: one focused on behavioral patterns and the other on symptoms associated with varying levels of alcohol intake. The target demographic included individuals who consumed alcoholic beverages, with stratified sampling employed to ensure a diverse and representative dataset (Kozak et al., 2008). The final dataset comprised responses from 500 participants, stored in a CSV file with 24 distinct features and an aggregate of 1,186 responses.

Data Preprocessing

Before analysis, the dataset underwent a thorough preprocessing stage to ensure its quality and reliability. This included the elimination of redundant data, addressing missing values, rectifying inconsistencies, standardizing the datasets, curating outliers, and converting categorical data into numerical values through label encoding and data scaling (Jo, 2020). Additionally, Principal Component Analysis (PCA) was employed to reduce dimensionality and enhance computational efficiency, filtering out noise and retaining essential features.

Clustering with K-means Algorithms

The K-means clustering algorithm was employed to partition the dataset into distinct non-overlapping subgroups. The initial selection of cluster centroids can significantly influence the outcome, so the K-means++ initialization method was used to ensure better and faster convergence. The optimal number of clusters (k) was determined using the Elbow Method and the Silhouette Score, resulting in three distinct categories of users based on their alcohol addiction spectra.

Naïve Bayes Classification

Following the clustering, the dataset was labeled, and a model was constructed using the Naïve Bayes algorithm. The data were split into training (80%) and testing (20%) sets. A

categorical Naïve Bayes model was developed and validated using various metrics, including accuracy, precision, recall, and F1-score. Cross-validation, divided into five folds, was applied to ensure the model wasn't overfitting. The Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) were used for comprehensive model evaluation.

Dataset Collection

Figure 1.

Stratified Sampling

Data was collected using online surveys and questionnaires. The target audience for the questionnaire included individuals who consumed alcoholic beverages in the past year and those who never drank, aiming to discern differences in behavior and symptoms. Stratified sampling was employed to ensure a comprehensive understanding of both subgroups.

Stratified Sampling, a probability sampling technique, facilitates obtaining a representative sample from a population divided into roughly similar subpopulations or strata[35]. It ensures that specific subgroups are adequately represented in the sample and aids in the accurate estimation of each group's characteristics. This technique is pivotal in surveys aiming to understand disparities between subpopulations.

The population was segmented into three primary groups: non-drinkers or minimal drinkers, occasional drinkers, and heavy drinkers. Survey and questionnaire items were meticulously designed to gauge behavioral patterns and symptoms associated with alcohol consumption. These items underwent expert validation to ensure the resulting data's credibility and validity.

System Analysis Design

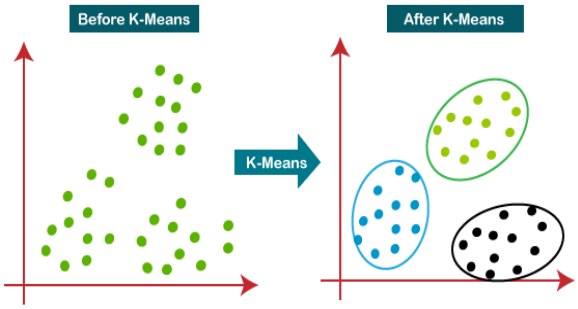
The K-means clustering algorithm is a widely used technique in machine learning for partitioning a dataset into distinct, non-overlapping subgroups where each data point belongs to only one group. It aims to make the intra-cluster data points as similar as possible while also ensuring the clusters themselves are as distinct as possible. The algorithm achieves this by minimizing the within-cluster sum of squares.

As shown in Figure 3, **K-means Clustering** groups the data by dividing it into sub-groups that share similar attributes. K-means Clustering, a widely used unsupervised machine learning algorithm, is designed to partition datasets into distinct clusters based on data point

characteristics (Chaudhary, 2020). The primary objective of this algorithm is to identify and create well-defined clusters by minimizing intra-cluster variance while maximizing inter-cluster differences. Each cluster has an associated centroid, which represents the mean position of all data points within that cluster and serves as a reference for assigning new data points (Jo, 2020).

Figure 3.

K-mean Clustering Algorithm



Data Exploration was first performed to ensure clean data, followed by label encoding to convert categorical data into integers. Principal Component Analysis (PCA) was then applied to reduce data dimensionality and filter out noise. The Elbow, Silhouette methods, the Calinski-Harabsz Index, and the Davies-Bouldin Index were all used in tandem with K-means Clustering to ensure robust and accurate labeling of the dataset. Post clustering, the labeled data served as the foundation for modeling using the Naïve Bayes classifier.

The goal was to discern how various features influence an individual's reliance on alcohol. While alcohol is often synonymous with celebration, its consumption is also linked to coping mechanisms during stressful or sad periods. The study aimed to ascertain the depth of an individual's reliance on alcohol, with K-means Clustering providing the labeling framework based on given features.

The primary objective of this research was to classify alcohol addiction severity into distinct categories based on a multi-dimensional dataset. Given that K-means inherently creates distinct clusters, it aligns well with the research objective. Additionally, the clear demarcation between clusters aids in providing precise classifications, allowing for a better understanding of the gradations in alcohol addiction severity. The algorithm's efficiency in handling large datasets ensured that the entirety of the dataset was used, capturing all nuances and variations in the data, further aligning with the research's goal to provide a comprehensive classification.

The choice of K-means clustering was not arbitrary but was based on careful consideration of the dataset's characteristics, the algorithm's strengths, and the research objectives. The results, as presented, underscore the efficacy of this choice.

Method for Selecting Initial Cluster Centroids:

The initial selection of cluster centroids can significantly influence the outcome of the K-means clustering algorithm. A poor initial choice can lead to suboptimal cluster formations and may require more iterations for convergence, potentially leading to a local minimum solution. Recognizing this sensitivity, special attention was given to the method of initializing cluster centroids in this study.

1. **K-means++ Initialization:** Instead of randomly initializing the centroids, which is the traditional method in vanilla K-means, the K-means++ initialization method was employed. K-means++ is a smart centroid initialization technique designed to speed up convergence. It works by selecting the first centroid randomly from the dataset and subsequently selecting the next centroids from the remaining data points with a probability proportional to their squared distance from the point's nearest existing centroid. This method ensures that the initial centroids are spread out across the data, leading to better and faster convergence.
2. **Multiple Initializations:** Given the potential for K-means to converge to a local minimum, the algorithm was run multiple times with different centroid initializations. The final clustering solution chosen was the one with the lowest within-cluster sum of squares, ensuring an optimal clustering result.
3. **Elbow Method for Optimal 'K':** Before initializing centroids, it's essential to determine the optimal number of clusters (K). The Elbow method was employed, which involves running the K-means clustering on the dataset for a range of values of K and then for each value of K compute the sum of squared distances from each point to its assigned center. When these overall dispersions are plotted against K values, the "elbow" of the curve represents an optimal value for K (a balance between precision and computational cost).

Naïve Bayes Algorithm

Naïve Bayes, an algorithm grounded in Bayes' Theorem, is known for its simplicity, speed, and efficacy in classification tasks (Bazett, 2022). Following the acquisition of a labeled dataset via K-means Clustering, modeling with the Naïve Bayes Algorithm was initiated. The data was split, with 80% dedicated to training and the remaining 20% reserved for testing (Azeraf et al., 2022).

A categorical Naïve Bayes model was subsequently developed and tested, with validation measures such as accuracy, precision, recall, and F1-score implemented to assess model performance. To prevent overfitting, cross-validation divided into five folds was applied (Palupi, 2021).

Finally, the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) were used for comprehensive model evaluation. These metrics serve as indicators of a model's ability to differentiate between positive and negative cases. Aiming for an AUC score above 0.8—indicative of robust model performance—the ROC curve also provided insights into the optimal classification threshold (Deng et al., 2010). The in-depth analysis via the ROC curve and AUC ensured that the resultant model was reliable and ready for deployment within a web application (Jo, 2020).

Algorithm Design Methodology

Figure 4.

Algorithm Design Methodology

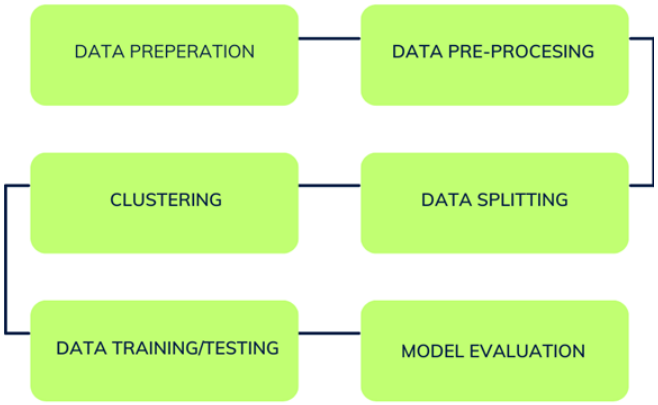


Figure 4 visually delineates the sequential process underpinning the Algorithm Design Methodology for this study. The methodology serves as a roadmap, guiding the data from its raw form to a stage primed for insightful analyses and predictions (Hartung, 2018).

Data Preparation. This foundational step involves gathering and curating essential data for the study, with an emphasis on assembling relevant features to enhance the model's performance. Tasks such as data cleaning and organization are addressed during this stage (Palupi, 2021).

1. Data Pre-processing. To ensure seamless data operations, various libraries, including Numpy, Pandas, Matplotlib, Pyplot, and Seaborn, are employed. This phase is critical for tasks such as converting non-numerical values into numerical equivalents (data encoding) and employing Principal Component Analysis (PCA) for dimensionality reduction (Jo, 2020).
2. Clustering. This phase focuses on the inherent attributes of the dataset, facilitating labeling based on similarities. It sets the stage for deploying the Naïve Bayes Classifier in subsequent steps (Chaudhary, 2020).

3. Data Splitting. As a precursor to modeling, the dataset is divided into two segments, with 80% allocated for training to lay the groundwork for the model's learning, and 20% reserved for testing to gauge the model's performance (Khalaf et al., 2017).
4. Data Training: At this stage, the algorithm learns the dataset's nuances by identifying patterns that are instrumental for future predictions (Bazett, 2022).
5. Data Testing: Once trained, the model is tested against fresh data to evaluate its efficacy. The testing phase provides insights into the model's ability to generalize beyond the training data (An et al., 2023).
6. Model Evaluation: Following testing, the model's performance and alignment with the data are assessed using metrics such as the Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index for clustering. The Naïve Bayes model is evaluated using Accuracy, Precision, Recall, F1-Score, Confusion Matrix, Cross-Validation, and the ROC & AUC curves (Deng et al., 2010; Azeraf et al., 2022).

Formula 1.

Silhoute Method Formula

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Silhouette Method is a metric that shows the quality and similarity of clustering based on the similarity of points within the cluster [38].

The symbol *a(i)* is the average distance between i and all other points in the same cluster as i, while *b(i)* is the average distance between i and all other points in the nearest cluster to i, the cluster with the smallest average distance. max(*a(i), b(i)*) is the maximum value between *a(i)* and *b(i)*. The symbol max*(a(i), b(i))* is the numerator and provides a scale between 0 and 1. It is the maximum value between *a(i)* and *b(i)*, which means that the silhouette score is bounded between -1 and 1. The nearer the score to 1 suggests that the Silhouette is well assigned to its designated clusters. If the score got near or equal to -1, this suggests that data points have been poorly assigned. Getting a score of near 0 suggests no clear distinction between the neighboring clusters.

Formula 2.

Calinsksi-Harabasz Formula

$$CH = (B/W) * ((N - K))/((K - 1))$$

The Calinski-Harabasz Index is another metric used to evaluate clustering quality by measuring the dispersion between clusters relative to the dispersion within clusters (Kim, 2017). This metric provides insights into how well the clusters are separated from each other, with higher scores indicating better-defined and more distinct clusters.

In Calinksi-Harabasz index, B is the between-cluster sum of squares, which is the sum squared of distances of the centroids. W is the within-cluster sum of squares, the sum of squared distances of all points in the respective cluster. The factor (*(n - k) / (k - 1)*) is a correction term that increases as the number of clusters k increases.

The Davies-Bouldin Index (DB) is a metric used to evaluate clustering quality by comparing the within-cluster scatter to the between-cluster separation of the data (Bijnen, 1973). The formula is expressed as:

Formula 3.

Davies-Bouldin Formula

$$DB = \left(\frac{1}{k}\right) * sum\left(max\big(R(i,j) + R(j,i)\big)\right) for\ i \neq$$

In Davies-Bouldin index, the symbol k represents the number of clusters, *R(i,j)* is the measure of the distance between clusters *i* and *j*, which is computed as the sum of the distances between each pair of points in clusters i and j divided by the number of pairs. The Davies-Bouldin with the lowest score represents well-separated clusters with a smaller within-cluster scatter (Kim, 2017).

Formula 4.

Accuracy Formula

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

Accuracy, is a metric used to represent the correctly classified data instances over the total number of data instances (Ali et al., 2011).

Formula 5.

Precision Formula

$$Precision = \frac{TP}{TP + FP}$$

The Precision metric, which measures the correctly positive instances. This will measure how precise positive values are (Khalaf et al., 2017).

Formula 6.

Recall Formula

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

Recall is a metric that measures the proportion of correctly predicted positive instances out of the total number of actual positive instances. It evaluates how well the model identifies all relevant cases within the datase (Ratini, 2022).

Formula 7.

F-1 Formula

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

The F1-score represents the harmonic mean of Precision and Recall, balancing these two metrics. It provides a single measure that captures the trade-off between precision and recall, making it particularly useful for imbalanced datasets (Ratini, 2022).

Discussion

The dataset for this research was meticulously sourced through questionnaires and online surveys administered via Google Forms. These surveys were tailored to assess the severity of alcohol consumption behaviors among respondents. Divided into two primary sections, the questionnaires delved into behavioral patterns and discernible symptoms related to alcohol consumption. The target demographic included individuals who consume alcoholic beverages. Stratified sampling was employed to ensure a diverse and representative data pool. The curated data was stored in a CSV file encompassing 24 distinct features with an aggregate of 1186 responses. This dataset, underpinned by strategic data collection and organization, offers a profound resource for delving deeper into the nuances of alcohol addiction severity among respondents.

Clustering Results

After preprocessing, the optimal cluster count of three was ascertained, aligning with the objective to delineate three distinct levels of addiction: less addicted, mildly addicted, and highly addicted. A silhouette score of 0.74 indicated the robustness of the clustering. The clustering outcome demonstrated that the dataset had been adeptly partitioned into three discernible categories, primed for decoding the gradations of alcohol addiction.

Naïve Bayes Model Evaluation

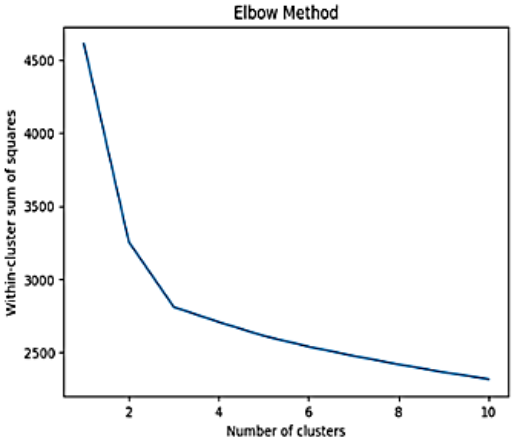
The Naïve Bayes Classifier showed impressive performance in its initial train-test split, achieving an accuracy of 96%, a precision of 94%, a recall of 97%, and an F1-score of 95%. The Confusion Matrix indicated high true positive and true negative counts, suggesting commendable accuracy, although some disparities emerged, particularly for the highly addictive cohort. The ROC curve and AUC scores further validated the model's efficacy, with AUC scores uniformly impressive at 0.97 for all labels.

Clustering

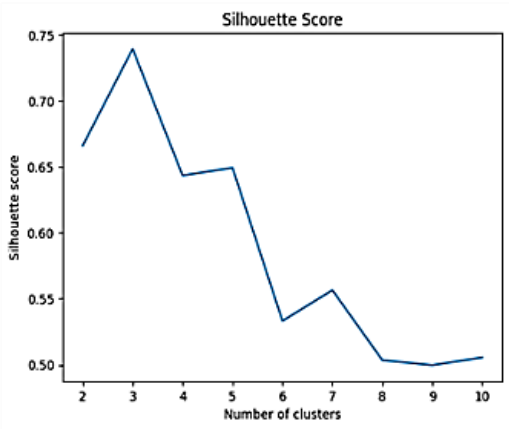
To bolster the accuracy and efficiency of machine learning models and data analytics, the researchers undertook a comprehensive data preprocessing regimen. This involved the meticulous elimination of redundant data, addressing data voids, rectifying inconsistencies, standardizing datasets, curating outliers, and transmuting categorical data into numerical counterparts via label encoding and data scaling.

Figure 5.

Result of Elbow Method and Silhouette Score



(a) Elbow Method

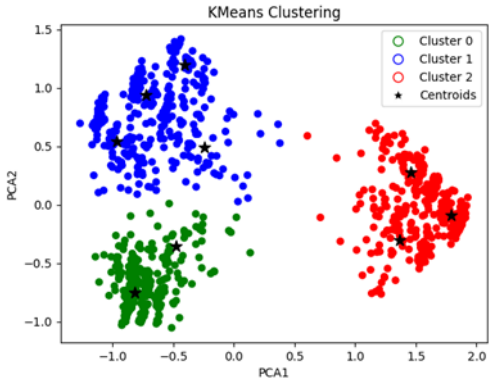


(b) Sillhouette Score

After the preprocessing stage, the focus shifted to implementing the K-means Clustering. A pivotal precursor to this was discerning the optimal number of K Clusters. The Elbow Method paired with the Silhouette Score served as instrumental metrics in this quest. As evidenced in Figure 5 a and b, an optimal cluster count of three was ascertained, harmoniously aligning with the team's objective to delineate three distinct categories of users based on their alcohol addiction spectra. A silhouette score of 0.74 stands testament to the robustness of the clustering.

Figure 6.

Scatter Plot of the K-means Clustering



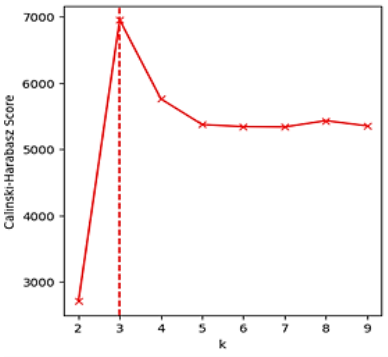
Delving deeper, Figure 6 illuminates the scatter plot resultant from the K-means Clustering, showcasing three optimally chosen clusters. Each of these clusters, demarcated in hues of green, blue, and red, pivot around stars signifying the centroids. These centroids epitomize the mean coordinates of the encompassed data points within their domain. The clustering outcome corroborates that the dataset has been adeptly partitioned into three discernible categories, primed for decoding the gradations of alcohol addiction. The intricacies of data preprocessing dovetailed with the precision of K-means clustering consummated the team's overarching ambition of judiciously gauging alcohol addiction levels.

Model Evaluation for Clustering

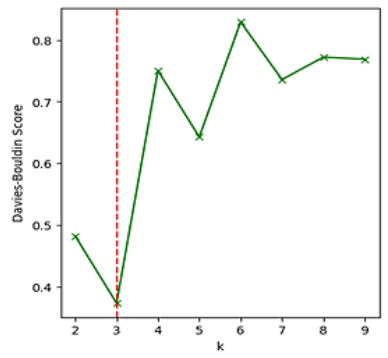
The researcher utilized various model evaluation metrics to validate the clustering outcomes. The Calinski-Harabasz score, which quantifies the balance between inter-cluster dispersion and intra-cluster cohesion, suggested well-defined and distinct clusters. As evidenced in Figure 7, the optimal cluster count (k=3) yielded a notable score of 6951.89, indicative of robust cluster distinction. Complementing this, the Davies-Bouldin Score, which gauges the average similarity of each cluster with its most resembling counterpart, returned a commendable score of 0.37 for k=3, affirming compactness within clusters and their clear separation.

Figure 7.

Calinski-Harabsz and David-Bouldin Score



(a) Calinski-Harabsz Score



(b) David Bouldin Score

To reinforce the validity of these findings, the team deployed Cross-Validation—a trusted mechanism for evaluating model performance by segmenting data into multiple subsets. Employing five such subsets aimed to obviate the pitfalls of overfitting and underfitting. The resultant mean score of 0.71 was congruent with the previously attained silhouette score of 0.74, corroborating that neither overfitting nor underfitting tainted the data. The results of these evaluations emboldened the team to proceed with the labeling and subsequent analytical steps, confident in the clustering's legitimacy.

Naïve Bayes Model

Result on the first train test split

Table 1 underscores the efficacy of the Naive Bayes Classifier in its initial train-test split. An impressive array of scores across Accuracy, Precision, Recall, and F-1 metrics affirm its prowess in data classification.

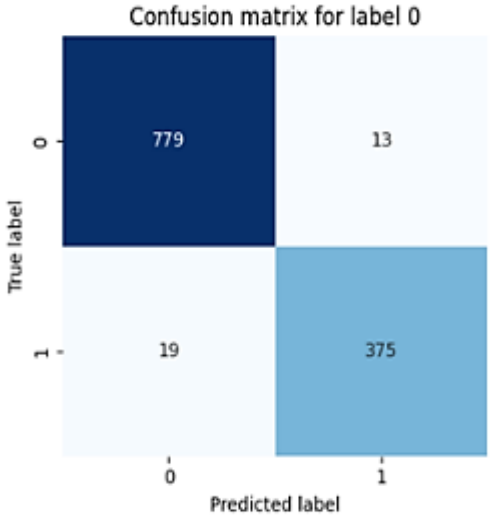
Table 1.

Naïve Bayes Classifier Performance Matrix

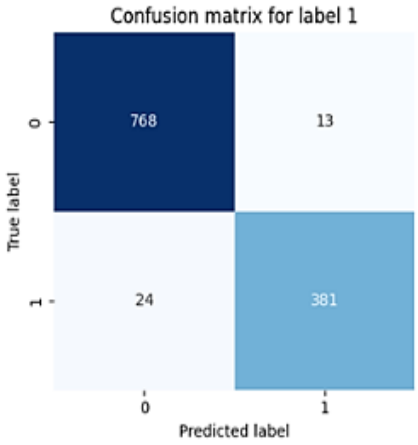
Metrics	Results
Accuracy	96%
Precision	94%
Recall	97%
F-1 Score	95%

Figure 8.

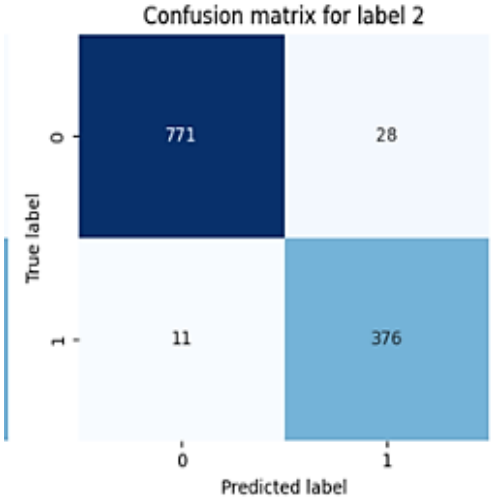
Confusion Matrix for each label after Cross-Validation



(a) Label 0



(b) Label 1



(c) Label 2

Presented in Figure 8, the Confusion Matrix elucidates the model's performance across various severity labels. The matrix predominantly showcases high true positive and true negative counts, suggesting commendable accuracy. However, disparities emerge, especially for label 2, hinting at potential areas for model refinement.

Label Analysis

Label 0 Analysis

True Negative (779): Correctly identified cases that don't belong to label 0.

False Positive (13): Misclassified cases predicted as label 0 but don't belong to it.

False Negative (19): Cases belonging to label 0 but predicted otherwise.

True Positive (375): Accurately classified cases as label 0.

Label 1 Analysis

True Negative (768): Cases correctly identified as not pertaining to label 1.

False Positive (13): Cases erroneously tagged as label 1.

False Negative (24): Label 1 cases predicted as a different label.

True Positive (381): Correctly identified cases under label 1.

Label 2 Analysis

True Negative (771): Accurately identified non-label 2 cases.

False Positive (28): Cases incorrectly marked as label 2.

False Negative (11): Label 2 cases predicted differently.

True Positive (376): Cases rightly classified as label 2.

Figure 9.

Model Performance using ROC and AUC

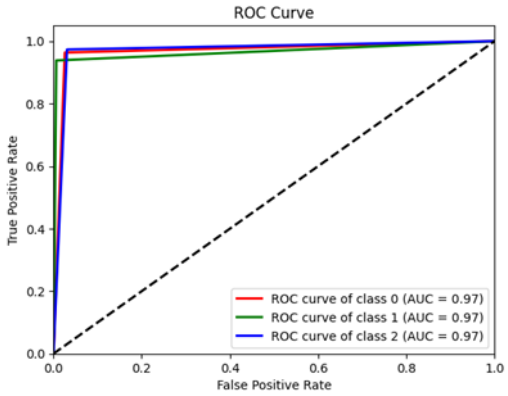


Figure 9 maps the ROC curve, demarcating the discriminative capacity of the model across labels: 0, 1, and 2. Each curve crystallizes the efficacy of the respective label in differentiating between its positive and negative cases. The AUC, gauging the aggregate discrimination capability, stands uniformly impressive at 0.97 for all labels. This score underscores the model's adeptness at bifurcating cases based on their true classifications. The robust ROC and AUC metrics vouch for the model's precision in discerning alcohol addiction severities, ranging from minimal to pronounced addiction levels.

Discussion

The dataset for this research was meticulously sourced through questionnaires and online surveys administered via Google Forms. These surveys were tailored to assess the severity of alcohol consumption behaviors among respondents. Divided into two primary sections, the questionnaires delved into behavioral patterns and discernible symptoms related to alcohol consumption. The target demographic included individuals who consume alcoholic beverages. Stratified sampling was employed to ensure a diverse and representative data pool. The curated data was stored in a CSV file encompassing 24 distinct features with an aggregate of 1,186 responses. This dataset, underpinned by strategic data collection and organization, offers a profound resource for delving deeper into the nuances of alcohol addiction severity among respondents.

The results of this study demonstrated the effectiveness of using Naïve Bayes and K-means Clustering for classifying and predicting alcohol addiction severity. These findings align with prior research, such as Chaudhary (2020), which highlighted the utility of K-means Clustering in identifying distinct patterns within datasets. The high accuracy of 95%, as achieved by the Naïve Bayes model, corroborates the findings of Palupi (2021), who reported

similar success using machine learning algorithms for classification tasks in behavioral health data.

Furthermore, the clustering approach used in this study resulted in three well-defined levels of alcohol addiction severity (less addicted, mildly addicted, and highly addicted). These results are consistent with Kim (2017), who emphasized the role of well-structured clustering techniques in segmenting populations for targeted interventions. However, the achieved silhouette score of 0.74 suggests room for improvement in optimizing cluster boundaries, as noted in Jo (2020), who proposed enhanced initialization techniques for improving clustering quality.

When compared to other works, such as Habehh and Gohel (2021), who focused on broader healthcare applications of machine learning, this study provides a more specialized focus on alcohol addiction. The use of stratified sampling further strengthened the representativeness of the dataset, addressing the limitations reported in studies like An et al. (2023), which noted biases in data collection methods.

By building on these comparative insights, this study reinforces the applicability of machine learning techniques in healthcare and offers a more targeted approach to understanding alcohol addiction. Future work could integrate other algorithms, as suggested by Bazett (2022), to enhance prediction accuracy and refine classification outputs.

Conclusion

This study successfully utilized the Naïve Bayes Algorithm and K-means Clustering to predict and classify alcohol addiction levels into three categories: less addicted, mildly addicted, and highly addicted. The results demonstrated the effectiveness of these machine learning techniques, with the Naïve Bayes model achieving an accuracy of 95%, precision of 93%, recall of 97%, and an F1-score of 95%. The K-means Clustering approach effectively segmented the dataset with a silhouette score of 0.74, underscoring its robustness.

The findings provide healthcare professionals with a valuable tool for identifying and addressing alcohol addiction, contributing to more precise interventions and treatment strategies. Future studies should explore integrating additional machine learning algorithms and addressing limitations such as dataset size and cluster optimization to further enhance accuracy and applicability.

Findings

This research utilized the Naïve Bayes Algorithm and K-means Clustering to analyze alcohol addiction levels, categorizing individuals into three groups: less addicted, moderately addicted, and highly addicted. Data was collected from 500 participants through a meticulously crafted survey, focusing on alcohol consumption frequency, underlying reasons for drinking, and associated adverse effects.

The Naïve Bayes model demonstrated high predictive accuracy, achieving 95% accuracy, 93% precision, 97% recall, and an F1-score of 95%. Cross-validation further validated the model's reliability. K-means Clustering effectively segmented the dataset into three addiction levels, with the highly addicted cluster showing a pronounced frequency of alcohol consumption and adverse impacts.

The study highlights the efficacy of combining machine learning techniques for predicting and categorizing addiction severity. These findings provide healthcare professionals with actionable insights to develop targeted interventions and personalized treatment strategies for individuals battling alcohol addiction.

REFERENCES

Ali, D. S., Ghoneim, A., & Saleh, M. (2017). Data clustering method based on mixed similarity measures. In *Proceedings of the 6th International Conference on Operations Research and Enterprise Systems*. https://doi.org/10.5220/0006245600001482

Ali, S. F., Onaivi, E. S., Dodd, P. R., Cadet, J. L., Schenk, S., Kuhar, M. J., & Koob, G. F. (2011). Understanding the global problem of drug addiction is a challenge for IDARS scientists. *Current Neuropharmacology, 9*(1), 2–7. https://doi.org/10.2174/157015911795017245

American Psychological Association. (2012). Understanding alcohol use disorders and their treatment. https://www.apa.org/topics/substance-use-abuse-addiction/alcohol-disorders

An, Q., Rahman, S., Zhou, J., & Kang, J. J. (2023). A comprehensive review on machine learning in healthcare industry: Classification, restrictions, opportunities and challenges. *Sensors, 23*, 4178. https://doi.org/10.3390/s23094178

Azeraf, E., Monfrini, E., & Pieczynski, W. (2022). Improving usual Naive Bayes classifier performances with neural Naive Bayes-based models. In *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods*. https://doi.org/10.5220/0010890400003122

Bazett, T. (2022). Introduction to Bayes' Theorem. In *Bayesian Inference*. https://doi.org/10.1007/978-3-030-95792-6_3

Bhatt, A. (2022). Alcohol addiction and abuse. *Addiction Center*. https://www.addictioncenter.com/alcohol/

Bèchet, N. B., Shanbhag, N. C., & Lundgaard, I. (2020). Glymphatic function in the gyrencephalic brain. *BioRxiv*. https://doi.org/10.1101/2020.11.09.373894

Bijnen, E. J. (1973). Coefficients for defining the degree of similarity between objects. In *Cluster Analysis* (pp. 4–20). https://doi.org/10.1007/978-94-011-6782-6_2

Centers for Disease Control and Prevention. (2022). Alcohol-related disease impact application website.

Chaudhary, M. (2020). K-means clustering in machine learning. *Medium*. https://medium.com/@cmukesh8688/k-means-clustering-in-machine-learning-252130c85e23

Chiva-Blanch, G., & Badimon, L. (2019). Benefits and risks of moderate alcohol consumption on cardiovascular disease: Current findings and controversies. *Nutrients, 12*(1), 108. https://doi.org/10.3390/nu12010108

David, C., et al. (2016). Usability of a smartphone app to reduce excessive alcohol consumption. *Frontiers in Public Health, 4*. https://doi.org/10.3389/conf.fpubh.2016.01.00064

Deng, Z., Choi, K.-S., Chung, F.-L., & Wang, S. (2010). Enhanced soft subspace clustering integrating within-cluster and between-cluster information. *Pattern Recognition, 43*(3), 767–781. https://doi.org/10.1016/j.patcog.2009.09.010

Early exposure to child abuse or neglect can cause long term health consequences. (2009). *PsycEXTRA Dataset*. https://doi.org/10.1037/e572212009-002

Epinephrine. (2023). *Reactions Weekly, 1968*(1), 138–138. https://doi.org/10.1007/s40278-023-44302-4

Franjic, S. (2021). Frequent alcohol consumption can have detrimental health consequences. *Archives of Psychiatry and Behavioral Sciences, 4*(1), 29–34. https://doi.org/10.22259/2638-5201.0401005

Habehh, H., & Gohel, S. (2021). Machine learning in healthcare. *Current Genomics, 22*(4), 291–300. https://doi.org/10.2174/1389202922666210705124359

Harmful use of alcohol kills more than 3 million people each year, most of them men. (2023). *Human Rights Documents Online*. https://doi.org/10.1163/2210-7975_hrd-9841-20180011

Hartung, T. (2018). Making big sense from big data. *Frontiers in Big Data, 1,* October. https://doi.org/10.3389/fdata.2018.00005

Jacobs, K. (1978). Positive contents and measures. In *Measure and Integral* (pp. 26–71). https://doi.org/10.1016/b978-0-12-378550-3.50005-0

Jarman, M. P., & Haider, A. H. (2019). When one data set is insufficient—Things to consider when linking secondary data—Reply. *JAMA Surgery, 154*(2), 187. https://doi.org/10.1001/jamasurg.2018.4751

Jo, T. (2020). K means algorithm. In *Machine Learning Foundations* (pp. 217–240). https://doi.org/10.1007/978-3-030-65900-4_10

Khalaf, A., Majeed, A., Akeel, W., & Salah, A. (2017). Students' success prediction based on Bayes algorithms. *International Journal of Computer Applications, 178*(7), 6–12. https://doi.org/10.5120/ijca2017915506

Kim, K. (2017). A weighted k-modes clustering using new weighting method based on within-cluster and between-cluster impurity measures. *Journal of Intelligent & Fuzzy Systems, 32*(1), 979–990. https://doi.org/10.3233/jifs-16157

Kozak, M., Zieliński, A., & Singh, S. (2008). Stratified two-stage sampling in domains: Sample allocation between domains, strata, and sampling stages. *Statistics & Probability Letters, 78*(8), 970–974. https://doi.org/10.1016/j.spl.2007.09.057

Lee, R. B., Baring, R., Maria, M. S., & Reysen, S. (2015). Attitude towards technology, social media usage and grade-point average as predictors of global citizenship identification in Filipino university students. *International Journal of Psychology, 52*(3), 213–219. https://doi.org/10.1002/ijop.12200

Lewis, D. J. (1969). Positive instances of reinstatement. *Science, 166*(3906), 772–772. https://doi.org/10.1126/science.166.3906.772-a

Mean average precision. (n.d.). *Springer Reference*. https://doi.org/10.1007/springerreference_65277

Nembach, E. (1975). Critical resolved shear stress of materials which simultaneously contain various types of obstacles impeding the glide of dislocations. In *The Movement of Molecules Across Cell Membranes* (pp. 413–416). https://doi.org/10.1007/978-3-540-37413-0_19

Palupi, E. S. (2021). Employee turnover classification using PSO-based naïve Bayes and naïve Bayes algorithm in PT. Mastersystem Infotama. *Jurnal Riset Informatika, 3*(3), 233–240. https://doi.org/10.34288/jri.v3i3.232

Schwenkreis, F. (2022). Using the silhouette coefficient for representative search of team tactics in noisy data. In *Proceedings of the 11th International Conference on Data Science, Technology and Applications*. https://doi.org/10.5220/0011100600003269

Sudhinaraset, M., Wigglesworth, C., & Takeuchi, D. T. (2016). Social and cultural context of alcohol use: Influences in a social-ecological framework. *Alcohol Research, 38*(1), 35–45. https://pubmed.ncbi.nlm.nih.gov/27159810

Sullivan, M. G. (2009). Too many pregnant women still drink alcohol. *Family Practice News, 39*(12), 33. https://doi.org/10.1016/s0300-7073(09)70489-x

Unsupervised learning—Clustering using K-means. (2019). In *Python® Machine Learning* (pp. 221–242). https://doi.org/10.1002/9781119557500.ch10

Vongprechakorn, K., Chumuang, N., & Farooq, A. (2019). Prediction model for amphetamine behaviors based on Bayes network classifier. In *2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)* (pp. 1–6). https://doi.org/10.1109/iSAI-NLP48611.2019.9045560

Whiteman, H. (2022). Drinking alcohol can clear brain waste, study finds. *Medical News Today*. https://www.medicalnewstoday.com/articles/320824

Woodman, R. J., & Mangoni, A. A. (2023). A comprehensive review of machine learning algorithms and their application in geriatric medicine: Present and future. *Aging Clinical and Experimental Research*. https://doi.org/10.1007/s40520-023-02552-2

Xiao, N., Li, K., Zhou, X., & Li, K. (2019). A novel clustering algorithm based on directional propagation of cluster labels. In *2019 International Joint Conference on Neural Networks (IJCNN)*. https://doi.org/10.1109/ijcnn.2019.8852159