JBE Journal of Biostatistics and Epidemiology
J Biostat Epidemiol. 2025;11(2): 197-213

ORIGINAL ARTICLE
DOI: *https://doi.org/10.18502/jbe.v11i2.20558*
*Received **March** 26, 2025; Accepted **April** 11, 2025;*
*Published online **June** 15, 2025.*

Classifying Substance Abuse Tendencies Using the Naive Bayes Algorithm

Esin **Avci***

Department of Statistics, Faculty of Arts and Sciences, Giresun University, Türkiye.

ABSTRACT

Introduction: Uncertainty in human life often arises from a lack of knowledge based on past events or unrealized circumstances. The Naive Bayes classification technique, rooted in conditional probability, offers a hypothesis-driven approach to linking two random occurrences and calculating posterior probabilities. Substance addiction remains a critical issue, particularly in patients hospitalized in community mental health centers, necessitating effective predictive methods for early identification and intervention.

Methods: This study employed the Naive Bayes algorithm to classify substance addiction tendencies in patients. Data of all 205 patients registered at the Giresun Province Prof. Dr. A. Ilhan Özdemir State Hospital Community Mental Health Center was obtained from the database. To enhance prediction accuracy, feature selection was conducted using the Information Value (IV) method. Ten patient attributes were analyzed, including gender, education level, marital status, income status, urban status, living alone, family disease, relation with family and environment, activity status, and age. Features with strong or medium predictive power were prioritized for the model. Accuracy, recall, precision, and F1 score were used as evaluation metrics of the model.

Results: Based on the strong or medium predictive power of IV, four features: gender, education level, income status, and relationship status with family and environment (respectively 0.45, 0.2, 0.17, and 0.17) were related to substance abuse. The Naive Bayes algorithm revealed that males (78%) are approximately four times more likely than females (22%) to develop substance addiction. Patients with education levels ranging from primary to high school were more prone than those with college-level education or higher. Additionally, those under state protection exhibited a higher likelihood (39%) of substance abuse compared to other income statuses. Finally, individuals with poor or neutral relationships with family and their environment were more susceptible to addiction (30%). Respectively, recall, precision, F1 score, and accuracy were obtained as 75%, 65%, 70%, and 76%, indicating the proper classification rate.

Conclusion: The Naive Bayes algorithm effectively classified substance addiction tendencies in hospitalized patients, emphasizing key predictive factors such as gender, education level, income status, and relational dynamics. These findings highlight the importance of targeted interventions tailored to at-risk populations, improving early detection and management strategies in community mental health settings.

Key words: Machine learning; Naive bayes algorithm; İnformation value; Classification; Substance abuse

*Corresponding Author:
esinavci@hotmail.com

INTRODUCTION

In our era, the constant advancement of information technologies and their convenient accessibility lead to an increase in their frequency of use. For this reason, a lot of data is obtained every day. Machine learning covers various analysis methods that reveal information from this collected data. Machine learning, first mentioned by Arthur Lee Samuel, is defined as computers gaining the ability to learn.[1] Machine learning is a branch of artificial intelligence that uses statistical and computational power to identify complex patterns and make rational decisions.[2] Machine learning uses algorithms independently to learn about data. For this reason, algorithms are developed in machine learning, allowing systems to communicate with people, create autonomous cars, write and publish match reports, etc.

When a machine improves its performance with experience, it is assumed that the machine is retaining data and learning by requiring algorithms and programs that reveal interesting or useful patterns.[3] Machine learning algorithms make predictions about future situations by examining data from past experiences based on a mathematical theory.[4] Machine learning is the simplest way to predict the future from past experiences.[5,2] Machine learning is used to solve many different problems, including optical character detection, facial recognition, spam email filtering, spoken language understanding, medical diagnosis, customer segmentation, fraud detection, and weather forecasting.[6] In the literature, learning strategies used in machine learning are classified in different ways.[7]

Compared to other classifiers like logistic regression, decision trees, and support vector machines (SVMs), the Naive Bayes algorithm has a number of advantages, especially when dealing with high-dimensional data and sparse training sets. Naive Bayes is computationally efficient, needing less training time but yet achieving good results in text classification, spam filtering, and sentiment analysis because of its strong independence assumption.[8] In contrast to SVMs, which may need a lot of fine-tuning, or decision trees, which have the tendency to overfit complicated datasets, Naïve Bayes performs well with noisy data and offers reliable probabilistic interpretations.[9] Furthermore, it works well even with tiny datasets, which makes it a sensible option for real-time and categorical data application.[10]

Variables with low predictive information are occasionally employed while creating a classification model. This may make classification more biased. A solid theoretical basis for investigating, filtering, and modifying variables in binary classification is offered by Weight of Evidence (WoE) and Information Value (IV). The predictive ability of a variable to distinguish between binary classes can be gauged with the aid of its IV value.

In this study, the Naive Bayes algorithm, which is one of the machine learning classification algorithms that provides more unbiased classification with the Information Value (IV) feature selection method, was examined in terms of revealing the substance abuse tendency status of patients hospitalized in a community mental health center.

The use of a drug in quantities or in ways that are detrimental to the user or others is referred to as substance abuse, or drug abuse. It is a type of disorder linked to substances. Drug abuse is defined differently in the fields of criminal justice, medicine, and public health. When someone is under the influence of drugs, they may occasionally act criminally or antisocially, and they may also have long-term personality changes.[11] Certain drug usage may result in criminal consequences in addition to potential bodily, social, and psychological harm, however, these might differ greatly depending on the local jurisdiction.[12]

In recent years, numerous researchers have used machine learning to analyze substance addiction data. Miotto et al. (2018) used deep learning techniques, such as neural networks, in healthcare, providing instances of their usage in predicting substance misuse and related outcomes.[13] Shatte et al. (2019) applied machine learning to mental health issues, including substance misuse, and discussed the problems and prospects in this sector.[14] Han et al. (2020) used machine learning to evaluate Prescription Drug Monitoring Programs (PDMP) data and predict opioid usage among veterans, emphasizing the importance of early intervention.[15]

This article is organized as follows: Section 2 briefly describes the Naive Bayes algorithm, Weight of Evidence (WoE) and Information Value (IV), and model evaluation, respectively. Section 3 presents the results of applying the Naive Bayes algorithm after the IV method and assessing the performance of the algorithm. All mentioned analyses were applied in R software by using the "caret," "Information," and "e1071" packages. Section 4 presents the conclusion.

METHODS

Naive Bayes classification algorithm

Machine learning methods are divided into two categories: predictive and descriptive models. Predictive models consist of classification and regression methods that predict what the outcome will be by using the obtained variables; descriptive models consist of clustering methods to reveal the character of the data.[16]

In descriptive models, the classification method is used when the target (dependent) variable is categorical. One of the most widely used classification techniques in recent years is the Naive Bayes method.

The Naive Bayes method is an algorithm that uses Bayesian theory to assign patterns in the data to previously defined classes. Bayes' theorem shows the relationship between the conditional probabilities and the marginal probabilities for a random variable within a probability distribution. Bayes' theorem describes a relationship accepted by all statisticians, and some statisticians also use the name Bayes' rule or Bayes' law for this concept. Bayes' theorem is a fundamental tool for updating and changing subjective beliefs about probability value in light of new evidence, not as an objective property but as a subjective value of the observer. The Naive Bayes classification algorithm

is a predictive and descriptive classification algorithm that analyzes the relationship between the target variable and the feature (independent) variables. Naive Bayes obtains the prior probability by calculating the number of times each outcome appears in the training set during model learning. Naive Bayes also calculates the conditional probability for each feature variable based on the target variable classes. These probabilities are combined with the prior probabilities to obtain the posterior probabilities and are used in classification prediction.[17] The main idea of the classification of the Naive Bayes classification algorithm is based on the principle of posterior probability.18 The class with the highest posterior probability is predicted.[19]

The Naive Bayes classification algorithm assumes that each feature variable has an equal impact on the target variable and is unrelated to the others.[20] This situation plays an important role in the fast operation of the Naive Bayes algorithm.[21] One benefit of using Naive Bayes is that it needs small training data to estimate the classification parameters.[22]

For a given problem to be classified, each of the K possible outcomes or classes C_k, represented by a vector $x = (x_1,...,x_n)$ encoding some n feature variables, the naive Bayes model assigns conditional probabilities $p(C_k \mid x_1,...,x_n)$. If the number of feature variables (n) is large or can take on a large number of values, then relying on probability tables is infeasible. As a result, the model must be redesigned to become more manageable. Using Bayes' theorem, conditional probability may be deconstructed as

$$p(C_k \mid x) = \frac{p(C_k) \times p(x \mid C_k)}{p(x)} \quad (1)$$

Using Bayesian probability terminology, the above equation may be expressed as

$$posterior = \frac{prior - likelihood}{evidence} \quad (2)$$

In practice, only the numerator of a fraction is important because the denominator is constant and does not depend on C, and the values of the feature variables x_i are known. The numerator is equivalent to the joint probability model

$$p(C_k, x_1, ..., x_n) \quad (3)$$

It can be rewritten as follows, using the chain rule to apply the concept of conditional probability repeatedly:

$$p(C_k, x_1, ..., x_n) = p(x_1, ..., x_n, C_k) = p(x_1 \mid x_2, ..., x_n, C_k)\, p(x_2, ..., x_n, C_k) = p(x_1 \mid x_2, ..., x_n, C_k) p(x_2 \mid x_3, ..., x_n, C_k)\, p(x_3, ..., x_n, C_k) = ...$$
$$= p(x_1 \mid x_2, ..., x_n, C_k) p(x_2 \mid x_3, ..., x_n, C_k) ... p(x_{n-1} \mid x_n, C_k)\, p(x_n \mid C_k)\, p(C_k) \quad (4)$$

The "naive" conditional independence assumption states that all feature variables in x are mutually independent based on the category C_k. Based on this assumption,

$$p(x_i \mid x_{i+1}, \ldots, x_n, C_k) = p(x_i \mid C_k) \quad (5)$$

Consequently, the joint model can be stated as

$$p(C_k \mid x_1, \ldots, x_n) \; \alpha \; p(C_k, x_1, \ldots, x_n)$$

$$= p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(x_3 \mid C_k) \ldots$$

$$= p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \quad (6)$$

The symbol α represents proportionality, as the denominator $p(x)$ is omitted. Under the independence assumptions, the conditional distribution over the class variable C is:

$$p(C_k \mid x_1, \ldots, x_n) = \frac{1}{z} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \quad (7)$$

where, $Z = p(x) = \sum_k p(C_k)\, p(x \mid C_k)$. If the feature variables values are known, the scaling factor is constant and only reliant on $x_1, \ldots, x_n$.

When a class label and a particular attribute value do not occur together, the likelihood estimate based on frequency is zero. In this case, the posterior probability will be zero, leading to a classification error. This is often described as a "zero-frequency problem". To solve this, the smoothing technique can be used. In a Bayesian framework, adding one to the count for each attribute value-class combination is one method of resolving this "zero-frequency problem". One of the simplest smoothing techniques is "Laplace smoothing".[23] The Laplace smoothing approach adds the least positive value to the classification process to assist in correcting the current data and avoiding classification errors. To ensure that the prior distributions of any subcategory remain unchanged, Laplace correction adds a "k" value between 0 and 1 to each subcategory. In general, a value of 1 is preferable.[24]

All of the data that the algorithm takes into account is split into "training" and "test" data. The data is often split into 20% or 30% test data and 80% or 70% training data. Here, the objective is to compare the algorithm learned using the training data with the test data in order to assess the classification performance.[25]

Weight of Evidence and Information Value

In recent years, Weight of Evidence (WoE) and Information Value (IV) have drawn more attention for applications including segmentation and variable reduction. They are based on information theory, first created for scorecard development in the late 1940s 26. This analysis technique is typically straightforward and takes less time overall 27. In order to produce the greatest distinction between the recoded variable values, WoE recodes them into distinct categories and gives each one a distinct WoE value. Here, it is crucial to assume that the target variable must be binary in order to show

whether an event occurred or not. When dealing with continuous variables, categorize them into intervals or bins. Every category is regarded as a distinct bin for categorical variables. Determine the Weight of Evidence (WoE) by applying the following formula:

$$WoE = ln\left(\frac{\text{Percentage of event}}{\text{Percentage of non-event}}\right) \quad (8)$$

Information Value (IV) evaluates the overall predictive ability of the variables that have been employed, whereas WoE examines a variable's predictive capacity with regard to its intended outcome. The predictive power of competing variables can be compared using IV. The IV calculation is as follows:

$$IV=\sum_i WoE_i \times (\text{Percentage of event}_i - \text{Percentage of non-event}_i) \quad (9)$$

Stronger predictive power of the variable is shown by higher IV values 28,29. The following table summarizes an interpretation of IV:

Table 1. Rules Related to Information Value (IV)

Information Value (IV)	Variable Predictiveness
<0.02	Unpredictive
0.02 to 0.1	Weak
0.1 to 0.3	Medium
0.3 to 0.5	Strong
> 0.5	Suspicious

Information Value (IV) is a widely used feature selection method in credit scoring and risk modeling, measuring the predictive strength of a variable in relation to a binary target.[30] Compared to the Chi-Square test, which assesses independence between categorical variables, IV provides a more interpretable ranking of predictor strength.[31] LASSO regression, on the other hand, applies L1 regularization to shrink irrelevant feature coefficients to zero, making it effective for high-dimensional numerical datasets but less suited for categorical variables.[32] Unlike IV, LASSO assumes linearity, limiting its application in non-linear relationships. While mutual information is another alternative that measures dependency between variables and the target, it lacks the intuitive interpretation of IV, particularly in financial modeling.[33]

Model Evaluation

Although classification models are supposed to correctly classify all data, it is undeniable that a model's performance can yield accurate findings. The confusion matrix can be computed to evaluate the model. Table 2 displays a cross-tabulation of the confusion matrix, which compares the response

feature data from the prediction class with the actual.[29]

Table 2. The Confusion Matrix

	Actual Positive (AP)	Actual Negative (AN)
Predicted Positive (PP)	True Positive (TP)	False Positive (FP)
Predicted Negative (PN)	False Negative (FN)	True Negative (TN)

The accuracy value obtained from Table 2 is as follows:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \tag{10}$$

Recall is another name for the true positive rate (TPR), which is the percentage of all real positives that were appropriately identified as positives, given as

$$\text{Recall (or TPR)} = \frac{TP}{TP + FN} \tag{11}$$

The percentage of all positive classifications in the model that are truly positive is known as precision. It has the following mathematical definition:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

The F1 score is a statistic that combines recall and precision. There is a trade-off between precision and recall, and F1 can be used to assess how well the models handle that trade-off.

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \tag{13}$$

The F1 score has a significant feature in that it returns zero if any of the components (precision or recall) go to zero. It penalizes extreme negative values in either component.

RESULTS

Exploring the propensity to take substance abuse is the goal of this study. The data used in this study were obtained from Giresun Province Prof. Dr. A. Ilhan Özdemir State Hospital Community Mental Health Center, from 205 patients in the four-year period between 2011 and 2014, with the permission number 42991614/770 of the relevant institution. Using yes/no responses, the target variable is regarded as a substance use status. Because of its normal distribution ($p>0.05$), age is the only continuous variable that is categorized according to its mean value (43) (Figure 1).

Table 3 summarizes the variables considered in the analysis.

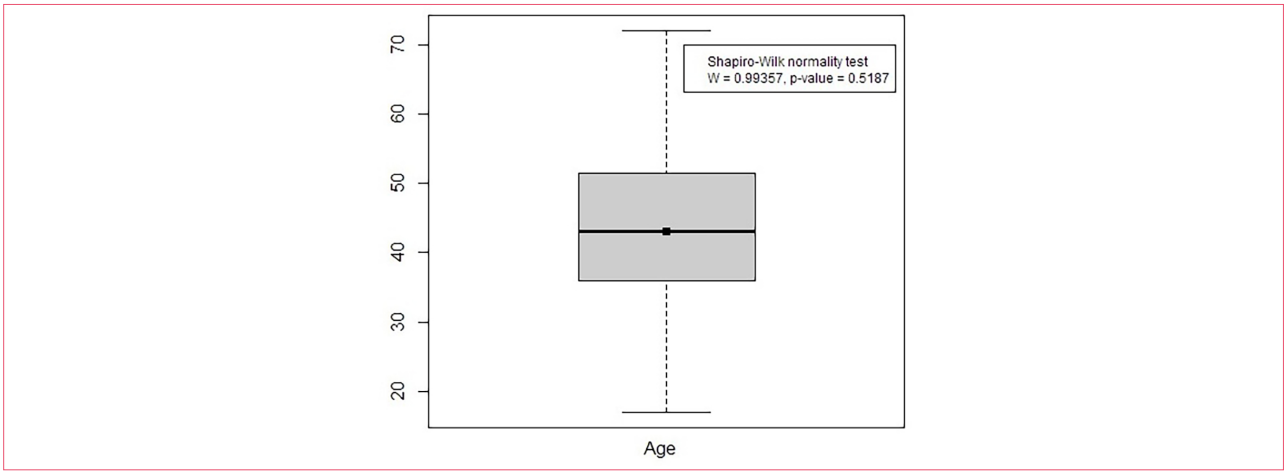

Figure 1. Box plot and normality test for age

Table 3. Summarizes the variables considered in the analysis

Variables	N(%)	Variables	N(%)
Substance Abuse		Urban Statu	
No	91 (44.4)	City	143 (69.8)
Yes	114 (55.6)	Bent	62 (30.2)
Gender		Live Alone	
Male	133 (64.9)	Yes	18 (8.8)
Female	72 (35.1)	No	187 (91.2)
Education Level		Family Disease	
Illiterate	16 (7.8)	No	110 (53.7)
Primary School	87 (42.4)	Yes	95 (46.3)
Secondary School	33 (16.1)	Relation with Family and Environmentv	
High School	45 (22)	Not good	60 (29.3)
College	7 (3.4)	Not good not bad	44 (21.5)
Undergraduate	16 (7.8)	Good	101 (49.3)
Graduate	1 (0.5)	Activity Statu	
Marital Status		Pasive	62 (30.2)
Single	112 (54.6)	Active	143 (69.8)
Married	68 (33.2)	<43	104 (50.7)
Divorced	10 (4.9)	Age	
Widow	15 (7.3)	≥43	101 (49.3)
Income Statu			
Unavailable	35 (17.1)		
Working	19 (9.3)		
Someone is looking	46 (22.4)		
State protection	70 (34.1)		
Retired	35 (17.1)		

Substance abusers made up about 56% of the patients. There were 35% women and 65% men in the sample. Primary school was the most educated level obtained (42%), while graduate school was the

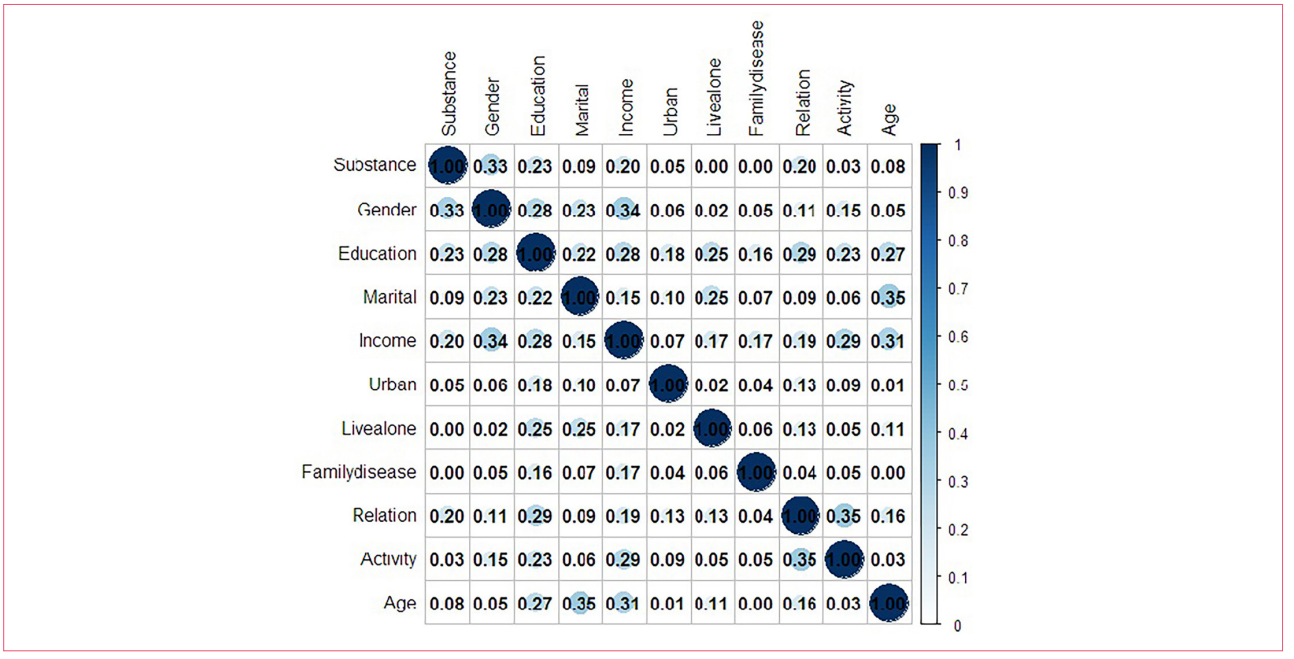

Figure 2. Correlation matrix based on Cramér's V

least educated (0.5%). Most of them were single (55%). The patients' income status was as follows: 34% were protected by the state, and 9% were employed. About 70% of the patients lived in the city. Roughly 92% of patients did not live alone. A history of family disease was present in 46% of the patients. Relationships with the environment and family were good for 49% of the patients. Seventy percent of patients had active lives. Finally, regarding the mean age, half of the patients were under 43.

Cramér's V is used to determine the association between the categorical variables, which measures the strength of the association between two categorical variables, ranging from 0 (no association) to 1 (perfect association). A small value (0.3 or less) suggests a weak or negligible association, moderate values (around 0.3-0.5) indicate a meaningful but not strong association, and high values (above 0.5) suggest a strong association 34. The association matrix of variables based on Cramér's V is displayed in Figure 2.

Substance abuse has a moderate association with gender but a weak association with education, income, and relationships (Figure 2). Generally, there were weak or negligible associations between features.

Detection and eliminating variables with low predictive information before creating a classification model using the Naive Bayes algorithm increases the unbiasedness of the predicted results. The predictive power of the features considered can be compared using IV. It is rational to select features with medium and strong predictive power. Hence, the features with IV values ranging between 0.1 and 0.5 are considered.

Figure 3 displays the features IV values ranging between 0.1 and 0.5. Respectively, "gender" had

the strongest predictive power (0.45), "education" (0.2), "income" (0.17), and "relationship" (0.17) have medium predictive power for substance abuse. The other six features' IV values were smaller than 0.03; hence, they were not displayed.

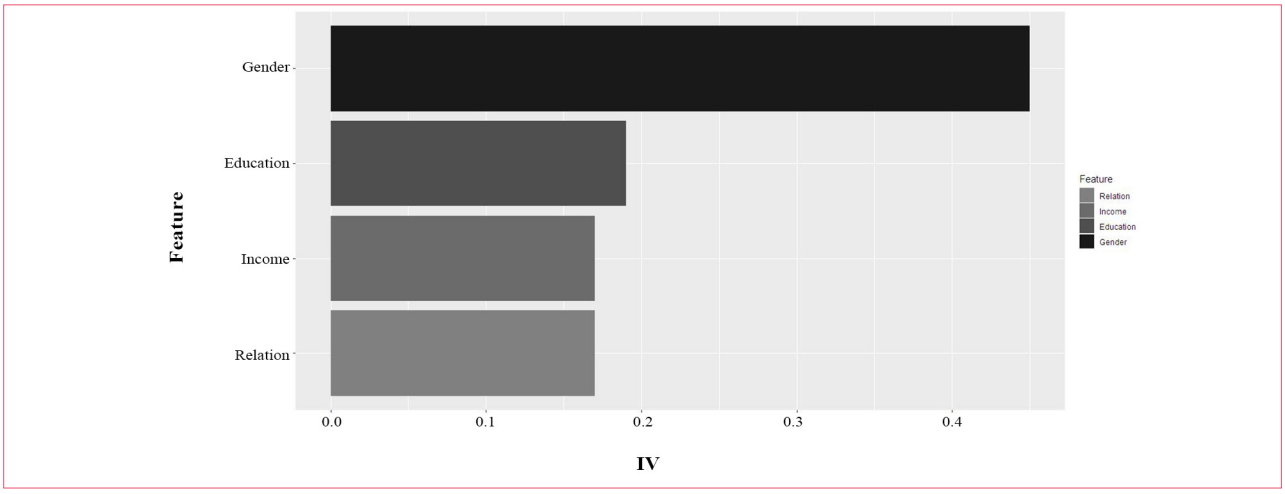

Figure 3. IV values of features ranging between 0.1 and 0.5

The Naive Bayes algorithm is used to generate a classification model once features with medium and strong prediction values for substance abuse have been chosen. In order to evaluate the prediction performance of the algorithm, the data was randomly split into 75% training and 25% test data. Thus, the algorithm trained with the training data will be tested with the test data. The algorithm was applied using the Laplace smoothing method to prevent classification errors in the case of zero-frequency problems.

Table 4 and Table 5, respectively, provide summaries of the prior and conditional probabilities that the Naive Bayes method obtained.

According to Table 4, 53% of the patients have substance abuse, while 47% do not.

Male substance abuse was nearly four times higher (78%) than female substance abuse (22%), according to Table 5. The majority of substance abuse propensity was found at the primary school level (38%), with secondary and high school education levels following (23% and 23%, respectively). Those with education levels above college and those who are illiterate (14%) have lower rates of substance abuse tendencies. Compared to other socioeconomic categories, patients under state protection (39%) had a higher tendency toward substance abuse. Substance abuse was not found in 58% of patients who had good relationships with their families and environment, but it was found in 30% and 27% of patients who had bad or neither good nor bad relationships, respectively.

The model's evaluation metrics indicate a balanced performance. Sensitivity (0.75) and recall (0.75) suggest that the model effectively identifies positive cases, while specificity (0.76) shows its ability to correctly classify negatives. Precision (0.65) indicates that some false positives occur. The F1-

score (0.70) balances precision and recall, reflecting overall effectiveness. Accuracy (0.76) confirms that 76% of predictions are correct. Finally, the AUC (0.76) suggests moderate discriminatory power between classes. Overall, the model performs fairly well (Table 6).

Table 4. A priori probabilities

	No	Yes
Substance Abuse	0.47	0.53

Table 5. Conditional probabilities

Features /Category	Substance Abuse	
	No	Yes
Gender		
Male	0.52	0.78
Female	0.48	0.22
Education Statu		
Illiterate	0.14	0.07
Primary School	0.32	0.38
Secondary School	0.13	0.23
High School	0.21	0.23
College	0.06	0.02
Undergraduate	0.12	0.06
Graduate	0.03	0.01
Income Statu		
Unavailable	0.18	0.15
Working	0.12	0.11
Someone is looking	0.26	0.24
State protection	0.26	0.39
Retired	0.17	0.12
Relation with Family and Environment		
Not good	0.28	0.30
Not good not bad	0.14	0.27
Good	0.58	0.43

Table 6. Evaluation metrics

Sensitivity	Specificity	Recall	Precision	F1	Accuracy	Area Under Curve (AUC)
0.75	0.76	0.75	0.65	0.70	0.76	0.76

The confusion matrix plot, which compares the predicted values from the Naive Bayes algorithm with the test data, is shown in Figure 4.

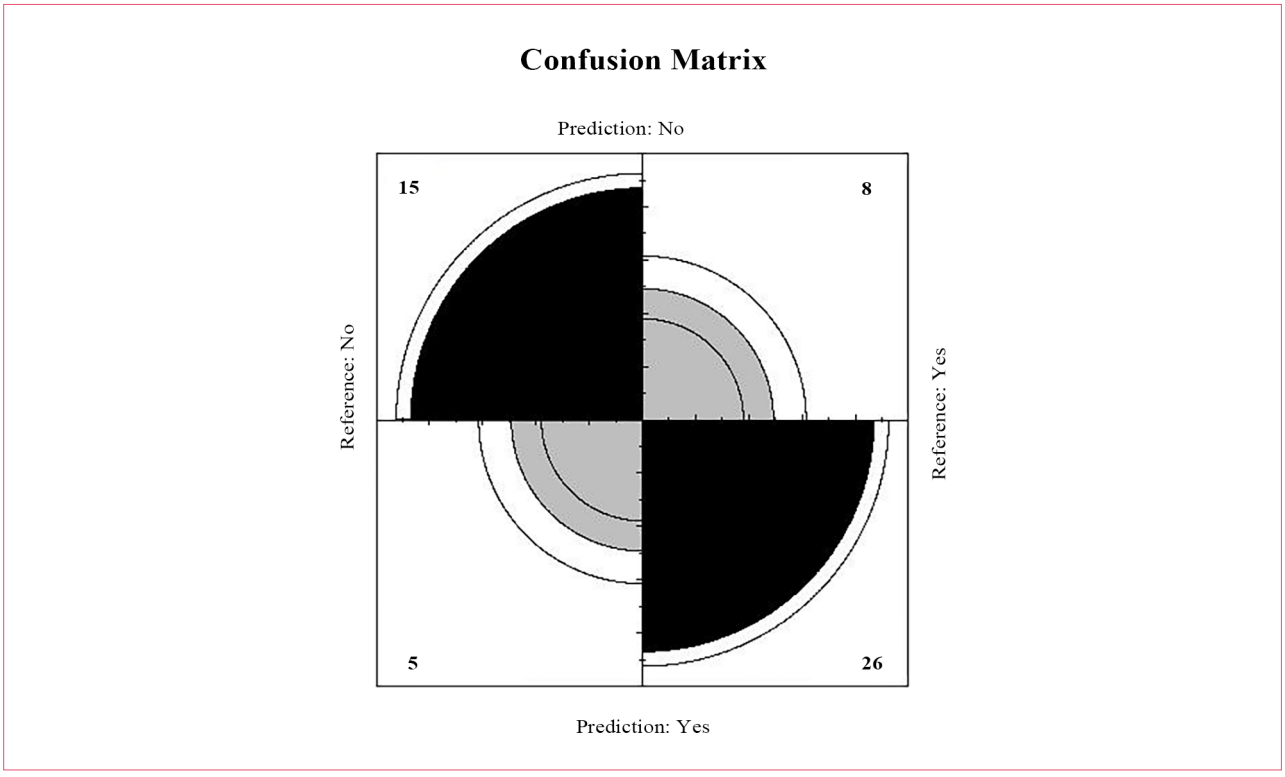

Figure 4. Confusion matrix plot

In the confusion matrix plot, the true class (target class) is represented by the columns, while the predicted class (output class) is represented by the rows. Observations that have been appropriately classified are represented by diagonal cells. The observations with off-diagonal cells are those that were misclassified. According to the plot, the Naive Bayes algorithm successfully identified 15 out of 20 patients in the test data who were not being substance abused as not being substance abused, while the other 5 were incorrectly classified as substance abused. Similarly, it can be seen that 26 out of 34 people in the test data who were substance abusers were correctly classified, while 8 were incorrectly classified as not being substance abusers.

DISCUSSION

Classifying substance abuse tendencies is critical for early intervention and effective treatment planning, particularly in mental health care settings. This study demonstrates that applying the Naive Bayes algorithm, coupled with feature selection via the Information Value (IV) method, yields valuable insights into the predictors of substance addiction. The findings align with existing literature, reinforcing the importance of demographic and psychosocial factors in understanding addiction risk.

Gender as a Predictor

The study revealed that males are approximately four times more likely to develop substance abuse

tendencies compared to females. This finding corroborates prior research, which has consistently reported higher substance abuse rates among males due to social, biological, and cultural factors.[35] These disparities highlight the need for gender-specific prevention strategies that address unique risk factors.

Educational Attainment

Education level emerged as a significant predictor, with individuals possessing only primary or high school education at greater risk than those with higher education levels. Lower educational attainment often correlates with limited awareness of health risks and reduced access to resources, which may contribute to higher substance use.[36] These results suggest that enhancing educational opportunities could indirectly reduce addiction prevalence.

Income and State Protection

Patients under state protection or with lower income statuses were more likely to exhibit substance abuse tendencies. Economic hardship and social vulnerability are well-documented risk factors, as financial instability can increase stress and exposure to environments conducive to substance use.[37] Policy-level interventions aimed at alleviating economic disparities may thus play a crucial role in mitigating addiction risks.

Relationship Dynamics

Poor or neutral relationships with family and environment were also significant predictors. This aligns with the theory that strong social support networks act as protective factors against addiction.[38] Conversely, strained familial or environmental relationships may foster feelings of isolation, making individuals more susceptible to substance use. Interventions focusing on strengthening interpersonal relationships could provide an additional layer of prevention.

Strengths and Limitations

This study highlights the utility of the Naive Bayes algorithm in identifying key predictors of substance abuse, offering a robust, probabilistic approach for classification. However, the findings are constrained by the dataset's representativeness, as well as potential biases inherent in self-reported data. Future studies should incorporate larger, more diverse populations and explore the integration of other machine learning techniques, such as support vector machines or neural networks, to enhance prediction accuracy.

Implications for Practice

The insights gained from this classification model can inform tailored interventions targeting high-

risk groups. For instance, community mental health centers could develop gender-specific educational programs, provide economic support for vulnerable populations, and implement family therapy to strengthen relational dynamics. Additionally, the predictive model offers clinicians a data-driven tool for risk assessment, enabling more precise and timely interventions.

CONCLUSION

In everyday life, the application of intelligence and intuition can serve scientific objectives. Another use of probability is the Bayesian approach, which is used in scientific investigations of such events. One of the classification algorithms using the Bayesian approach is the Naive Bayes algorithm.

In this study, substance abuse tendency was examined using the Information Value (IV) feature selection method that will increase the unbiasedness of the prediction results in the Naive Bayes algorithm. According to IV values, features ranging between 0.1 and 0.5 are selected as medium and strong power of predicted.

Gender, education level, income status, and relation with family and environment were the selected features for substance abuse data. The recall, precision, F1 score, and accuracy of the Naive Bayes algorithm based on these four features were obtained as 75%, 65%, 70%, and 76%, respectively. It is concluded that patients who were male, had a primary to high school education level, were state protected, and had bad or neither bad nor good relations with their family were more likely to be substance abusers.

Classifying substance abuse tendencies using machine learning techniques such as Naive Bayes represents a promising direction for mental health research and practice. By leveraging predictive analytics, healthcare providers can better understand and address the multifaceted nature of addiction, ultimately improving patient outcomes.

Availability of Data and Materials

Due to the nature of the research, due to [ethical/legal/commercial] supporting data is not available.

Disclosure statement

No potential conflict of interest was reported by the author.

REFERENCES

1. Raschka S. STAT 479: Machine Learning Lecture Notes [Internet]. 2018. Available from:

https://sebastianraschka.com/pdf/lecture-notes/stat479fs18/01_ml-overview_notes.pdf. Accessed 2024 Aug 10.

2. Hall P, Dean J, Kabul IK, Silva J. An overview of machine learning with SAS® Enterprise Miner™. Cary: SAS Institute Inc.; 2014.

3. Harrington P. Machine learning in action. 5th ed. Greenwich, CT: Manning; 2012.

4. Mitchell TM. Machine learning. Burr Ridge, IL: McGraw-Hill; 1997.

5. Daumé III H. A course in machine learning [Internet]. 2017. Available from: http://ciml.info/, chapter 5, p. 69. Accessed 2017 Sep.

6. Schapire RE. COS 511: Theoretical Machine Learning [Internet]. 2008. Available from: http://www.cs.princeton.edu/courses/archive/spr08/cos511/scribe_notes/0204.pdf. Accessed 2017 Mar 19.

7. Camastra F, Vinciarelli A. Machine learning for audio, image, and video analysis. In: Advanced Information and Knowledge Processing. 2008. p. 83–9.

8. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Cambridge: Cambridge University Press; 2008.

9. Rish I, et al. An empirical study of the Naive Bayes classifier. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence; 2001 Aug 4–6; Seattle. p. 41–6.

10. Zhang H. The optimality of Naive Bayes. Proc 17th Int Florida Artif Intell Res Soc Conf; 2004 May 12–14; Menlo Park. p. 562–7.

11. Ksir C, Ksir O. Drugs, society, and human behavior. 9th ed. Boston: McGraw-Hill; 2002. ISBN: 978-0072319637.

12. Mosby's Medical, Nursing, & Allied Health Dictionary. 6th ed. St. Louis: Mosby; 2002. ISBN: 978-0-323-01430-4.

13. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19(6):1236–46. https://doi.org/10.1093/bib/bbx044.

14. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychol Med. 2019;49(9):1426–48. https://doi.org/10.1017/S0033291719000151.

15. Han DH, Lee S, Seo DC. Using machine learning to predict opioid misuse among U.S. adolescents. Prev Med. 2020;130:105886. https://doi.org/10.1016/j.ypmed.2019.105886.

16. Alpaydin E. Introduction to machine learning. Cambridge: MIT Press; 2014. ISBN: 0262325756.

17. Ceci M. Naive Bayesian learning from structural data [dissertation]. Bari, Italy: Dipartimento di Informatica, University of Bari; 2005.

18. Panda M, Patra MR. Network intrusion detection using Naive Bayes. Int J Comput Sci Netw Secur. 2007;7(12):258–63.

19. Murty NM, Devi VS. Pattern recognition: an algorithmic approach. 2011. p. 86–102. ISBN: 978-0857294944.

20. Gupta P. Naive Bayes in machine learning [Internet]. Towards Data Science; 2024. Available from: https://towardsdatascience.com/naive-bayes-in-machine-learning-f49cc8f831b4. Accessed 2024 Aug 10.

21. Roman V. Machine learning introduction: a comprehensive guide [Internet]. Towards Data Science; 2024. Available from: https://towardsdatascience.com/machine-learning-introduction-a-comprehensive-guide-af6712cf68a3. Accessed 2024 Aug 10.

22. Understanding the mathematics behind Naive Bayes [Internet]. 2018. Available from: https://shuzhanfan.github.io/2018/06/understanding-mathematics-behind-naive-bayes/. Accessed 2024 Jan 24.

23. Randy H, Musdar AI. Aplikasi prediksi kerusakan smartphone menggunakan metode Naive Bayes dan Laplace Smoothing. J Tek Ind Syst Inf (JTRISTE). 2018;5(2):8–16.

24. Narayanan V, Arora I, Bhatia A. Fast and accurate sentiment classification using an enhanced Naïve Bayes model. In: IDEAL 2013. Berlin Heidelberg: Springer; 2013. p. 8206.

25. Witten IH, Frank E, Hall MA. Data mining: practical machine learning tools and techniques. 4th ed. Morgan Kaufmann; 2016.

26. Lin A. Variable reduction in SAS by using information value and weight of evidence. In: Proc SUGI Conf; 2015.

27. Alsabhan AH, Singh K, Sharma A, Alam S, Pandey DD, Rahman SAS, et al. Landslide susceptibility assessment in the Himalayan range based along Kasauli–Parwanoo road corridor using weight of evidence, information value, and frequency ratio. J King Saud Univ Sci. 2022;34(2).

28. Xia Y, Yan S. Feature selection based on weight of evidence and information value. Int J Inf Technol Decis Mak. 2015;14(4):769–94.

29. Kuhn M. Feature engineering and selection: a practical approach for predictive models. Springer; 2021.

30. Agresti A. Statistical methods for the social sciences. 5th ed. Pearson; 2018.

31. Cover TM, Thomas JA. Elements of information theory. 2nd ed. Wiley-Interscience; 2006.

32. Siddiqi N. Credit risk scorecards: developing and implementing intelligent credit scoring. Wiley; 2005.

33. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc B (Methodol). 1996;58(1):267–88.

34. Cramér H. Mathematical methods of statistics. Princeton University Press; 1946.

35. Addiction Center. The differences in addiction between men and women [Internet]. Available from: https://www.addictioncenter.com/addiction/differences-men-women/.

36. Lopez-Quintero C, de los Cobos JP, Hasin DS, Okuda M, Wang S, Grant BF, et al. Probability and predictors of remission from lifetime nicotine, alcohol, cannabis or cocaine dependence: results from the National Epidemiologic Survey on Alcohol and Related Conditions. Addiction. 2015;106(3):657–69.

37. Miller DP, Chang J. Parental substance use and child health outcomes: a look at health care utilization for Medicaid-insured children. Med Care Res Rev. 2019;76(2):267–86. https://doi.org/10.1177/1077558717722590.

38. Taylor OD. Adolescent depression as a contributing factor to the development of substance use disorders. J Hum Behav Soc Environ. 2017;27(7):715–22. https://doi.org/10.1080/10911359.2017.1339652.