# A Comparative Study of Machine Learning Algorithms for DrugAddiction Prediction

**Dr. S. Bharathidason[1] and C. Sujdha[2]**

[1]Assistant Professor and Head, Department of Computer Science, Loyola College (Autonomous), Chennai - 600 034.

[2]Assistant Professor, Department of Computer Science, Mar Gregorios College of Arts & Science, Chennai - 600 037.

Corresponding authour: bharathidasan@loyolacollege.edu

**Abstract:** Drug addiction is a pressing societal issue with far-reaching consequences. Accurate prediction of individuals at risk of drug addiction can greatly aid in prevention and intervention efforts. This study presents a comprehensive comparative analysis of three prominent machine learning algorithms: Classification and Regression Trees (CART), Support Vector Machines (SVM), and Random Forest, for the purpose of drug addiction prediction. The dataset used in this analysis contains a diverse set of attributes, including demographic information, mental and emotional health indicators, family dynamics, and prior experiences with drugs, making it a valuable resource for studying this complex issue. This study investigates the performance of these algorithms in predicting drug addiction based on the provided attributes, considering the factor accuracy. The results of this comparative analysis will contribute to the development of more accurate and efficient tools for identifying individuals at risk of drug addiction, ultimately assisting in the formulation of targeted prevention and intervention strategies.

**Keywords:** Drug addiction, Machine learning algorithms. Comparative analysis, Classification and Regression Trees (CART).Support Vector Machines (SVM), Random Forest.Prediction, Accuracy.

## INTRODUCTION

Substance abuse poses a significant public health concern worldwide, with far-reaching consequences for individuals, families, and society as a whole. Predicting and addressing substance abuse is a complex challenge that requires effective tools and methodologies. Machine learning (ML) techniques have shown promise in various healthcare applications, including the prediction and prevention of substance abuse.

Recent research articles on drug addiction highlight a range of profound problems faced by individuals grappling with substance abuse. Firstly, physical health deterioration is a significant concern, with drug addiction leading to a host of medical complications, including cardiovascular issues, respiratory diseases, and increased vulnerability to infectious diseases such as HIV/AIDS. Additionally, mental health challenges are prevalent among those struggling with addiction, including anxiety, depression, and heightened risk of suicidal ideation. Social consequences, such as strained relationships, isolation, and stigmatization, are also commonly reported, exacerbating the psychological toll of addiction. Economic hardships, including job loss and financial instability, further compound these problems. Moreover, the legal ramifications of drug abuse, including criminal charges and incarceration, pose substantial barriers to rehabilitation and reintegration into society. In sum, recent research underscores the multifaceted and deeply interconnected problems faced by individuals grappling with drug addiction, emphasizing the urgent need for comprehensive and compassionate support systems and interventions to address this complex issue.

To conduct this comparative analysis, we will employ a structured methodology that includes data preprocessing, feature selection, model training, and performance evaluation. Furthermore, we will explore the impact of various hyperparameters and settings on the algorithms' predictive capabilities. This research is essential as it contributes to the field of healthcare and addiction prevention by identifying the most suitable machine learning approach for predicting substance abuse.

## LITERATURE REVIEW

This section summarizes the related Machine Learning techniques employed for prediction. We have found many recent machine learning approaches, few recent research work in the domain of interest is summarized.

Previously in the year 2006 Mark G. Myers and John F. Kelly conducted a re- search in "Cigarette Smoking Among Adolescents with Alcohol and Other Drug Use Problems", they discussed about prevalence of cigarette smoking among adolescents with AOD use problems, smoking cessation efforts in this population, and special considerations for adolescent smoking cessation treatment, including peer influences, motivation, and nicotine dependence. They used a real-time data from the participants of myers and brown 1997 research paper, this paper was published in the Journal Alcohol Research & Health.

Ethan Sahker, Laura Acion and Stephan Arndt conducted a re-search named "Prediction of Adverse Drug Reactions Using Decision Tree Modelling" in the year of 2015they targeted the college student and non-student, clients without prior treatment admissions, aged 18–24, not in methadone maintenance therapy, and in non-intensive and ambulatory intensive outpatient treatment, for this analysis "Decision tree algorithm" was used to find the risk factors of students who are addicted to drug abuse. This paper was published in the Journal of American College Health.

In another paper based on Predicting the public and private life behaviours of a drug abuser in Bangladesh the authors emerged the statistical analysis of more than 8 algorithms. They had demonstrated the statistical differences between many algorithms based on ac-curacy measures. As we see most people are used Random Forest, Decision Tree, Logistic Regression, KNN, Naïve Bayes, Neural Network, etc.

### About the dataset

The research paper focuses on studies conducted in Bangladesh related to substance abuse and its pattern among the local people, which was conducted in the year 2019 and published in the online platform Kaggle for public availability. This dataset is modified with the help of Synthetic Minority Over-sampling Technique (SMOTE) to avoid the imbalance. The dataset contains 22 attributes related to individuals and their relationship with drugs or substance abuse, out of which one is categorical. The values of the categorical attributes are the outcome of people responding to the questions raised. Each row represents a different individual, and the columns represent different characteristics or factors related to their situation.

**Figure 1: Age distribution bar graph**

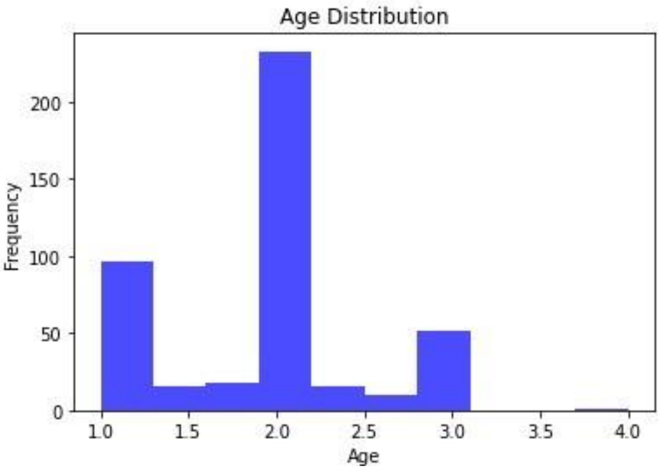
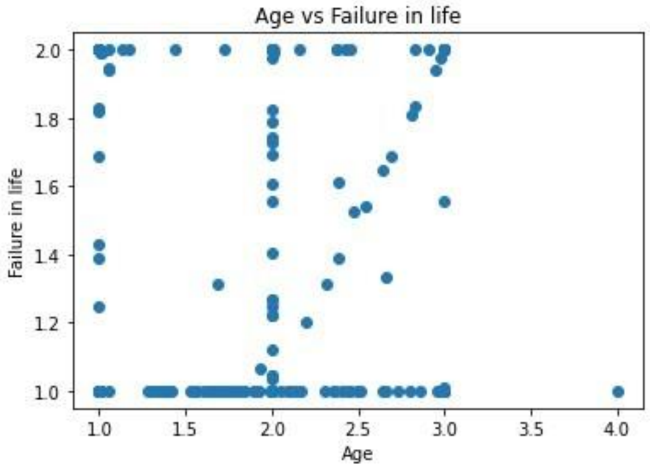


**Figure 2: Scatter plot for age vs failure in life**



***Total number of rows:** 441; **Total number of columns:** 22*

In summary, the dataset provided consists of various attributes related to individuals, capturingtheir demographics, education, family background, mental and emotional well-being, as wellas their experiences with drugs and related behaviours. The dataset includes information such as age, gender, education level, living situation, motives behind drug use, time spent on different activities, perceptions of failure in life, mental and emotional problems, suicidal thoughts, family relationships, family financial status, presence of addicted individuals in the family, withdrawal symptoms, workplace satisfaction, legal issues, living arrangements with drug users, smoking habits, history of drug use, influence of friends, willingness to try drugs if given the chance, and perceived ease of controlling drug use.

## MATERIALS AND METHOD

The data preprocessing phase includes learning about the dataset's summary, such as the total number of rows and columns, datatype, removing duplicate values and dealing with missing values, identifying relevant features using techniques such as correlation analysis, and scaling numerical features to bring them within a similar range. To overcome the class imbalance problem, the Synthetic Minority Over-sampling Technique (Smote) is employed to high sample the minority class with the objective to improve the minority class's predictive performance. SMOTE develops a more balanced class distribution through the creation of synthetic samples from the minority class. Exploratory data analysis (EDA) is also used to understand the characteristics of dataset before model building. Head of the first few rows are seen to understand the structure and format of the data examining each feature's datatype.

The dataset is split into two parts: one for training the model and one for testing its performance. This division helps in predicting how effectively the predictive model will perform on data that was previously unknown. Here the columns Age, Gender, Education, Live with, Motive about drug, Spend most time, Failure in life Mental/emotional problem, Suicidal thoughts, Family relationship, Financials of family, Addicted person in family, Withdrawal symptoms, Satisfied-with workplace, Case in court, Living with drug user, Smoking, Ever taken drug, Friends influence, If chance given to taste drugs, Easy to control use of drug are the independent variables and class is the dependent variable, here we have split the dataset 4 times because we have used 4 model for prediction, this has helped the model to perform good in the prediction of drug abuse.

To predict, six classification models are used:

- Classification and regression tree,
- K Nearest Neighbor
- Artificial Neural Network
- Random Forest Classifier,
- Support Vector Machine, and
- Naive Bayes.

### Evaluation Metrics

Various evaluation metrics are available for executing various machine learning models and determining the best one. Different evaluation techniques based on confusion metrics such as accuracy, precision, recall, and f-measure are introduced, and our model evaluation is based on the accuracy criteria.

### Models used for prediction

### K-nearest neighbors (KNN)

KNN is a type of algorithm used for both classifying and predicting values in a dataset. It works by finding the nearest data points to a given point and using them to make predictions. The "k" in KNN stands for the number of nearest neighbors to consider, which needs to be chosen before using the algorithm.

For classification tasks, KNN assigns a data point to the class that most of its nearest neighbors belong to. So, if most neighbors are in a certain class, that's the predicted class for the data point. For regression tasks, KNN predicts the value of a data point by averaging the values of its nearest neighbors. The distance between data points is crucial in KNN. It determines which points are considered neighbors. Common distance measures include Euclidean, Manhattan, and Minkowski distances.

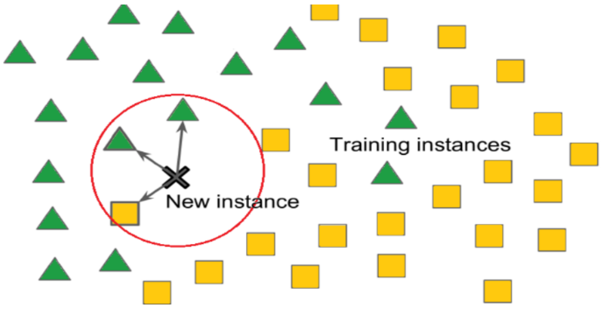


**Figure 3: K-nearest neighbors (KNN)**

**Classification and regression tree**: CART or Decision tree is a Supervised learning technique that can be utilized for solving regression as well as classification problems, but it is usually employed to solve classification problems. It is a tree-structured classifier in which internal nodes represent dataset features, branches represent decision rules, and every node in the leaf represents the result. It is a visual representation of all possible answers to a problem/decision given certain requirements.

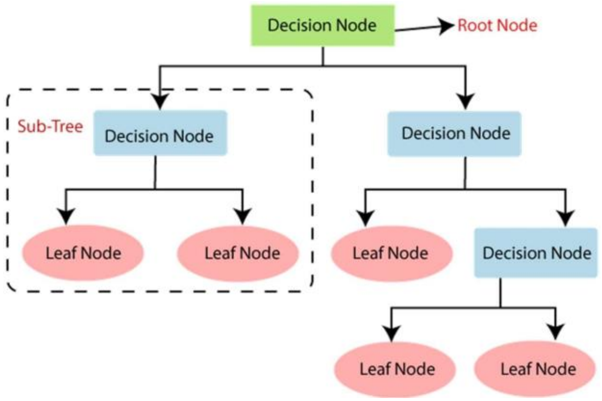


**Figure 4: Decision Tree**

Because of these features, CART is used as the first algorithm in this comparative analysis for prediction, with 75% of the data used for training and the remaining 25% used for testing. Import the classifier from the package tree and build the model with the parameters criterion as gini and splitter as best, which gives the model a higher accuracy rate for drug abuse prediction. Fit the model to the training dataset, predict with the testing data, tabulate the resultsas a data-frame, compare the real and predicted results, with the help of plot_tree function the tree structure is plotted and finally compute the Model's accuracy score.

**Random Forest Algorithm:** is a popular machine learning algorithm from the supervised learning technique. It can be utilized for both classification and regression problems in machine learning. It is based on the concept of ensemble learning, which is the process of combining multiple classifiers to solve a complex problem and improve the performance of the model. A random forest is a classifier that uses a number of decision trees on different subsetsof a given dataset and the mean them to improve the accuracy of prediction of that dataset. The greater the number of trees in the forest, the higher the accuracy and the lesser the risk of overfitting. These are the main reasons to use random forest as a second comparative algorithm.

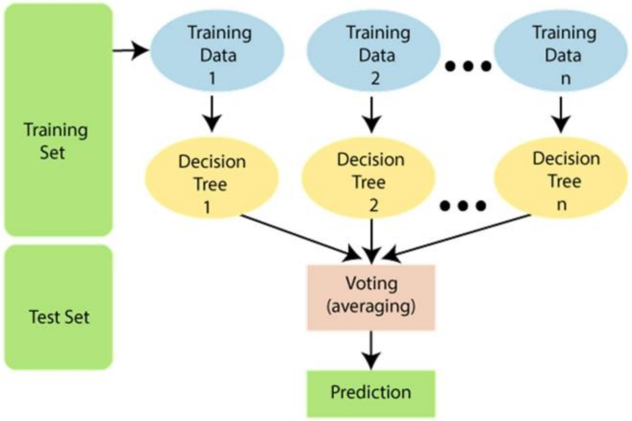


**Figure 5: Random Forest Algorithm**

Before developing the random forest model, the data is scaled using the technique standard scaler to standardize the features of a dataset. This standardization process is also known as z- score normalization, and it helped to enhance the model's accuracy. With 80% of the data for training and 20% for testing, import the classifier from the ensemble package and build the model with default parameters, train the model, test it, and compute the accuracy.

**Support vector machine**: offers very high accuracy compared to other classifiers such as logistic regression, and decision trees. It is known for its kernel trick to handle nonlinear input spaces. SVM is an exciting algorithm and the concepts are relatively simple. The classifier separates data points using a hyperplane with the largest amount of margin. That is why an SVM classifier is also known as a discriminative classifier. SVM finds an optimal hyperplane which helps in classifying new data points, It can easily handle multiple continuous and categorical variables. SVM constructs a hyperplane in multidimensional space to separate different classes. SVM generates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVM is to find a maximum marginal hyper-plane (MMH) that best divides the dataset into classes.

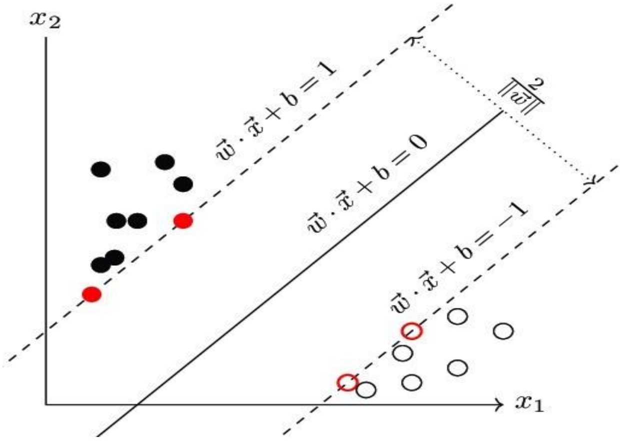

$x_2$
$\vec{w}\cdot\vec{x}+b=1$
$\frac{2}{\|\vec{w}\|}$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=-1$
$x_1$

**Figure 6: Support vector machine**

For support vector machine, the data used for training and testing are split as 80% for training and 20% for testing, the model is imported from support vector classifier, and the model is built with the parameter kernel='poly', resulting in the model a higher accuracy rate for drug prediction, fit the model in the training data, and predict with testing data, that givesthe model a higher accuracy rate and compute the accuracy percentage.

**Naive Bayes:** is a probabilistic machine learning algorithm based on Bayes' theorem. It is widely used for classification tasks, particularly in natural language processing, spam filtering, and sentiment analysis due to its simplicity and efficiency. Naive Bayes is a supervised learning algorithm that makes predictions based on the probability that a given input belongs to a certain class. It assumes that features (variables) are conditionally independent, meaning the presence of one feature does not affect the presence of another (hence the term "naive"). Despite this simplistic assumption, Naive Bayes often performs surprisingly well in practice, especially for text data.

At the core of the Naive Bayes algorithm is Bayes' theorem, which calculates the probability of a hypothesis (class) based on the observed evidence (features). The formula for Bayes' theorem is as follows

$$\mathrm{P}\ (\mathrm{Class} \mid \mathrm{features}) = \frac{\mathrm{p}\ (\ \mathrm{features} \mid \mathrm{class}) * \mathrm{p}(\mathrm{class})}{\mathrm{P}\ (\mathrm{features})}$$

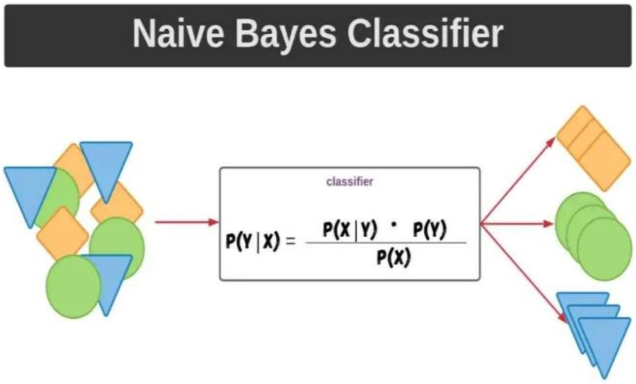


**Figure 7: Naive Bayes classifier**

For naïve bayes, the data used for training and testing are split as 75% for training and 25% for testing, the model is imported from Gaussian NB, and the model is built with the parameter the default parameter, to give the high accuracy rate for drug prediction, fit the model in the training data, and predict with testing data, and compute the accuracy percentage.

### Artificial Neural Network

Artificial Neural Networks (ANNs) are inspired by biological neural networks found in the human brain. Like the brain's neurons, ANNs consist of interconnected nodes organized into layers. These nodes, or neurons, communicate with each other across the network, mimicking the interconnected structure of the human brain.

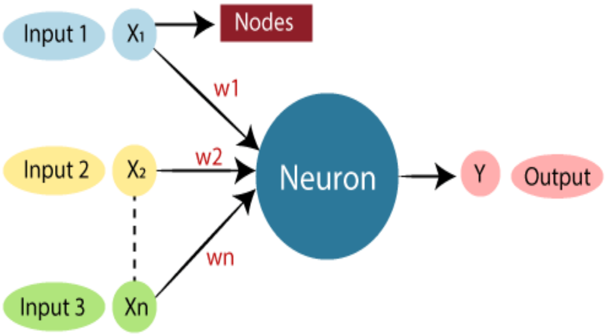


**Figure 8: Artificial Neural Network**

## RESULTS AND DISCUSSION

In this study, we applied four different machine learning algorithms—CART (Classification and Regression Trees), Random Forest, Support Vector Machine (SVM), and Naive Bayes— to predict drug usage patterns based on a comprehensive dataset. After rigorous evaluation, Random Forest emerged as the top-performing model in terms of accuracy. The results from the comparative analysis are as follows:

**CART:** CART, a decision tree-based algorithm, demonstrated an accuracy of **78.27 %** on the drug usage prediction task.

**KNN:** KNN, a distance based algorithm, demonstrated an accuracy of **78.18 %** on the drug usage prediction task.

**Naive Bayes:** The Naïve Bayes classifier, despite its simplicity, achieved an accuracy of 72**.72 %** in predicting drug usage patterns.

**SVM:** Support Vector Machine, known for its ability to handle complex decision boundaries, exhibited an accuracy of **85.45 %** on our dataset.

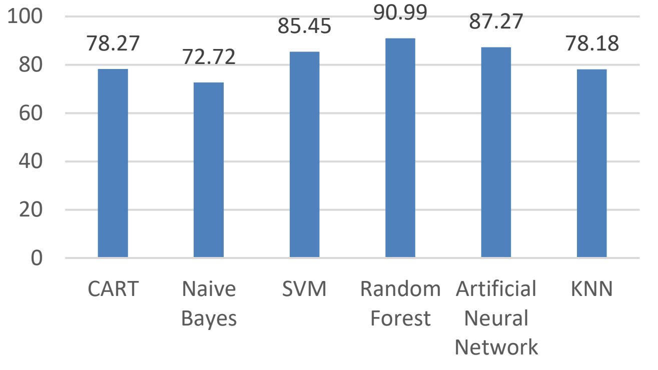
**ANN:** Artificial Neural Network, known for its ability to handle complex datasets, exhibited an accuracy of **87.27 %** on our dataset.

**Random Forest:** Random Forest, an ensemble learning method, outperformed the other algorithms with an impressive accuracy of **90.99 %**. This model's superior performance can be attributed to its ability to capture complex relationships within the data and mitigate overfitting through ensemble techniques.

Table 1: Comparative performance analysis of Machine Learning Algorithms

| **S. No** | **Model** | **Accuracy (%)** |
|---|---|---|
| 1 | CART | 78.27 |
| 2 | Naive Bayes | 72.72 |
| 3 | SVM | 85.45 |
| 4 | Random forest | **90.99** |
| 5 | Artificial Neural Network | 87.27 |
| 6 | KNN | 78.18 |

Figure 7 : Comparative performance analysis of Machine Learning Algorithms



The outstanding performance of Random Forest in predicting drug usage patterns highlights its suitability for this particular problem. Several factors contribute to RandomForest's success:

**Ensemble Learning:** Random Forest leverages the power of ensemble learning, combining multiple decision trees to form a robust and accurate model. This ensemble approach enables the model to generalize well to unseen data, capturing intricate patterns that might be missed by individual decision trees.

**Feature Importance:** Random Forest provides a feature importance score, indicating the relevance of each feature in predicting drug usage. By analyzing these scores, we gain valuable insights into which factors have the most significant influence on an individual's drug usage behavior. This information can be invaluable for targeted interventions and prevention strategies.

**Handling Missing Data:** Random Forest can effectively handle missing data, a common challenge in real-world datasets. Its inherent ability to impute missing values allows for a more comprehensive analysis, ensuring that valuable data points are not discarded.

**Hyperparameter Tuning:** Through careful tuning of hyperparameters such as the number of trees and the maximum depth of each tree, Random Forest can be optimized for specific datasets. This fine-tuning process enhances the model's performance and generalizability.

**Mitigating Overfitting:** Random Forest incorporates techniques like bagging and feature randomization, reducing the risk of overfitting. This ensures that the model does not memorize the training data but instead learns meaningful patterns that can be applied to unseen data effectively.

## CONCLUSION

In this study, we explored the application of various machine learning algorithms, includingKNN, ANN,CART (Classification and Regression Trees), Random Forest, SVM (Support Vector Machine), and Naive Bayes, to predict drug usage patterns among individuals. Leveraging a rich dataset encompassing diverse socio-demographic factors and psychological attributes, our objectivewas to identify the most effective predictive model. Upon rigorous evaluation and comparison, Random Forest emerged as the most robust and accurate predictive model for drug usage in our analysis. This result was particularly intriguing given the complexity and multidimensionality of the dataset. Random Forest, a versatile ensemble learning method, demonstrated superior performance in handling both numerical and categorical features, allowing it to capture intricate patterns within the data effectively.

The Random Forest algorithm's remarkable accuracy in predicting drug usage underscores its efficacy in real-world applications, especially in the context of social and behavioural studies. Its ability to mitigate overfitting, handle missing data, and interpret feature importance makes it a valuable tool for researchers and policymakers alike.

However, it is essential to acknowledge the limitations of our study. While Random Forest outperformed other models in terms of accuracy, further investigations are warranted to delve deeper into the interpretability of the model. Understanding the specific features that contribute significantly to its predictions can provide valuable insights into the underlying factors driving drug usage behaviours.

## REFERENCES

[1] Dataset - Drug Addiction in Bangladesh - Reasons: SMOTE

[2] F Hammann, H Gutmann, N Vogt, C Helma, J Drewe Prediction of Adverse Drug Reactions Using Decision Tree Modelling.

[3] Myers, Mark G., and John F. Kelly. ”Cigarette smoking among adolescents withalcohol and other drug use problems.” Alcohol Research &Health 29.3 (2006): 221

[4] Shahriar, Arif, et al. ”A Machine Learning Approach to Predict Vulnerability to Drug Addiction.” 2019 22nd International Conference on Computer and Information Technology (ICCIT). IEEE, 2019

[5] Maxwell, Aaron E., Timothy A. Warner, and Fang Fang. ”Implemen-tation of machine-learning classification in remote sensing: An appliedreview.” InternationalJournal of Remote Sensing 39.9 (2018): 2784-2817.

[6] Koizumi, Yuma, et al. ”SNIPER: Few-shot learning for anomalydetection to minimize false-negative rate with ensured true-positiverate.” ICASSP 2019-2019 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP).IEEE, 2019.

[7] Mitrpanont, Jarernsri, et al. ”A study on using Python vs Weka on dial-ysis data analysis.” 2017 2nd International Conference on InformationTechnology (INCIT). IEEE, 2017.

[8] Viloria, Amelec, et al. ”Comparative Analysis Between DifferentAutomatic Learning Environments for Sentiment Analysis.” Interna-tional Symposium on Distributed Computing and Artificial Intelligence.Springer, Cham, 2020