CBCA 273 Project Report

Submitted to the Faculty

of the Bennett University

By

**Ankan Dutta , E21BCAU0084**

Under the Guidance of

**Dr. Nitin Arvind Shelke**



**Drug Consumption Prediction on Different kinds of ML Models**

School of Computer Science Engineering & Technology

2022-23, Even Semester

Greater Noida-201310, Uttar Pradesh, India

NAME-ANKAN DUTTA ID NO-E21BCAU0084

BATCH-1 SUBJECT – AI IN HEALTHCARE MILESTONE (1)

**TOPIC NAME: - DRUG CONSUMPTION**

Me choosing this topic is the fact that there are more in the world consume more drugs than the amount of food each eat in there day to day life and in this topic. I will showcase that how drugs can be so much lethal to one person.

Abstract

Drug intake prediction is an important research area with practical applications in public health and law enforcement. Predicting and preventing drug addiction is made more difficult by the availability and variety of drugs. In recent years, machine learning techniques have been used to develop models that predict drug use patterns. These models use a variety of features, such as demographics, environments, and behavioral patterns, to predict the likelihood of drug use. This content provides an overview of cutting edge methods for predicting drug use and highlights the challenges and opportunities in this area.The purpose of this research is to develop effective strategies to prevent drug use and improve public health.

Name:-Ankan Dutta Id No:-E21BCAU0084

Batch:-1 Assignment:-Milestone 2 Ai in Healthcare

Table 1: Summarization of Related work

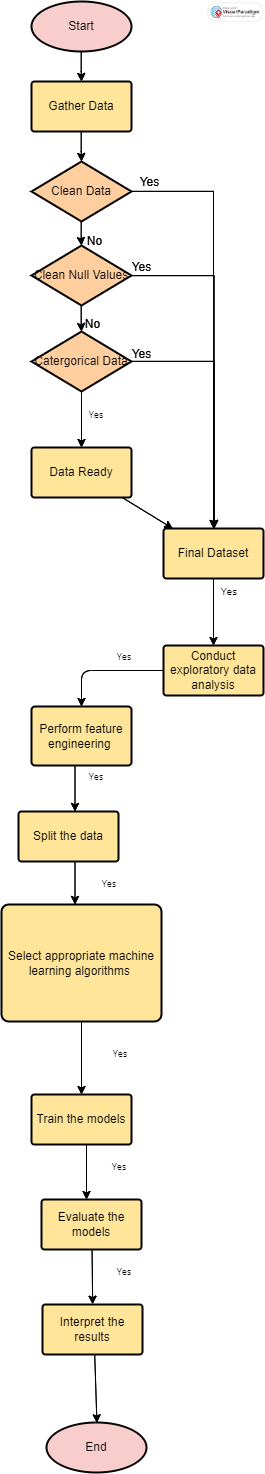
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ref | methodology used | Data set used | performance Parameters | Limitations/  +points (1 to  2 linws) |
| [1]Fehmarn | Naïve Bayes,  Gaussian Mixture,  Random Forest, Decision Tree, k-  Nearest  Neighbours,  Logistic Regression,  Linear Discriminant  Analysis | [https://archive .ics.uci.edu/ml /datasets/Dru g+consumptio n+%28quantifi](https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29)  [ed%29](https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29) | Accuracy score, Mean and Standard Deviation | The  Limitation is that they didn’t use the other methods to get Score and different function in Sklearn.met  rics |
| [2]  Adinugr oho | Decision tree, random forest, knearest neighbors, linear discriminant analysis, Gaussian mixture, probability density function estimation, logistic regression and | [https://archive .ics.uci.edu/ml /datasets/Dru g+consumptio n+%28quantifi](https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29)  [ed%29](https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29) | Accuracy Score | The  Limitation is that they didn’t use the other methods to get Score and different |
|  | naive Bayes,LSTM and ANN |  |  | function in Sklearn.met  rics |
| [3]Kumari | ANN C and ANN D | [https://archive .ics.uci.edu/ml /datasets/Dru g+consumptio n+%28quantifi ed%29](https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29) | Accuracy Score | The  Limitation is that they didn’t use the other methods to get Score and different function in Sklearn.met  rics |
| [4]Qiao | Random Forest, XGBoost, LightGBM and KNN | [https://archive .ics.uci.edu/ml /datasets/Dru g+consumptio n+%28quantifi](https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29)  [ed%29](https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29) | Accuracy, Precission and F1 score | It’s  Limitation is that it uses Only  Random  Forest and XGBoost and KNN and There are many other  classifier  can be used to find the Accuracy score |

Name:-Ankan Dutta ID No:-E21BCAU0084

Batch:-1 Project:-AI in Healthcare

Proposed Methodology:

• Flowchart/ Arch Diagrams:-



• Fig 1: Architecture Diagram

**Flowchart explanation**

Here is a flowchart for the Drug Consumption Data Set:

1. Gather the raw data on drug consumption, which includes information about the demographic characteristics of the participants and their self-reported drug use.
2. Clean and preprocess the data to remove any errors, inconsistencies, or missing values.
3. Conduct exploratory data analysis to gain insights into the distribution and patterns of drug use in the sample.
4. Perform feature engineering to create new variables that capture meaningful aspects of the data, such as the total number of drugs used, the frequency of use, and the age of first use.
5. Split the data into training and testing sets to develop and evaluate machine learning models.
6. Select appropriate machine learning algorithms to predict drug use based on the demographic and behavioral variables in the dataset.
7. Train the models using the training data and tune hyperparameters to improve their performance.
8. Evaluate the models using the testing data and metrics such as accuracy, precision, recall, and F1 score.
9. Interpret the results of the models to gain insights into the factors that contribute to drug use and identify potential intervention strategies.
10. Share the findings with stakeholders and the broader community to inform policy and public health initiatives aimed at reducing drug use and its harmful effects.

• **Algorithm/technique/model related explanation.**

There are several algorithms, techniques, and models that can be used to analyze the Drug Consumption (Quantified) Data Set. Here are some examples:

1. Logistic Regression: This is a statistical model that can be used to predict the probability of drug use based on the demographic and behavioral variables in the dataset. Logistic regression is a popular method for binary classification problems, such as predicting whether an individual uses drugs or not.
2. Decision Trees: This is a machine learning algorithm that can be used to build a model that predicts drug use based on a set of decision rules. Decision trees are easy to interpret and can handle both categorical and continuous variables.
3. Random Forest: This is an ensemble learning algorithm that combines multiple decision trees to improve the accuracy and robustness of the model. Random forests can handle high-dimensional data and are resistant to overfitting.
4. Support Vector Machines (SVM): This is a machine learning algorithm that can be used for both classification and regression tasks. SVMs are particularly useful for datasets with a small sample size and a large number of features.
5. Neural Networks: This is a deep learning technique that involves building a model composed of multiple layers of artificial neurons. Neural networks can learn complex patterns in the data and are capable of handling large datasets.
6. K-Nearest Neighbors (KNN): This is a non-parametric algorithm that can be used for both classification and regression tasks. KNN involves finding the k-nearest neighbors of a given data point and using their values to make a prediction.

Overall, the choice of algorithm or model depends on the specific research question, the characteristics of the data, and the desired level of interpretability and accuracy.

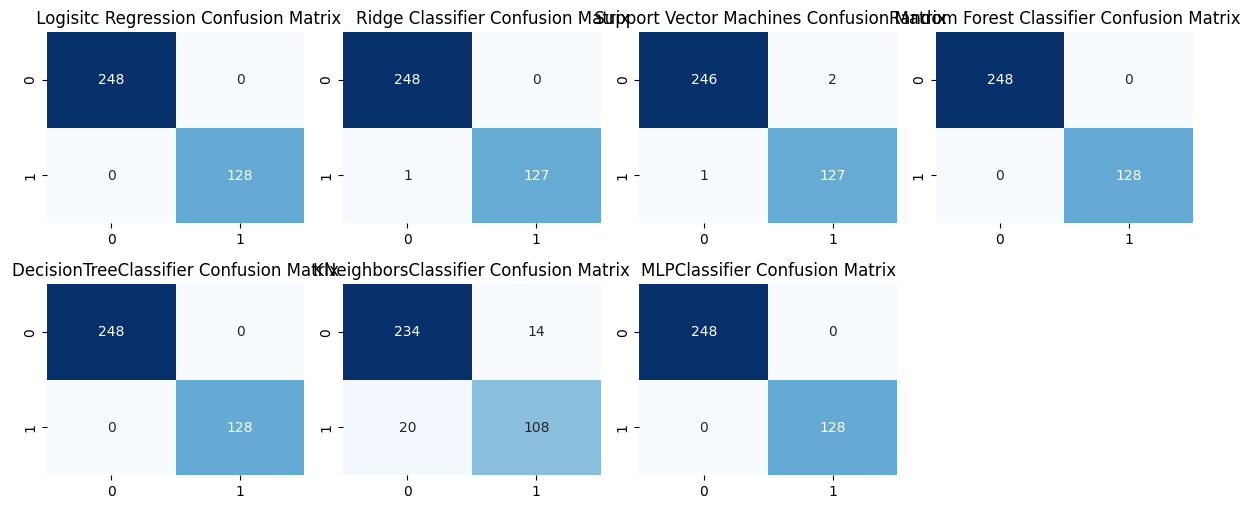
•  **Dataset description**

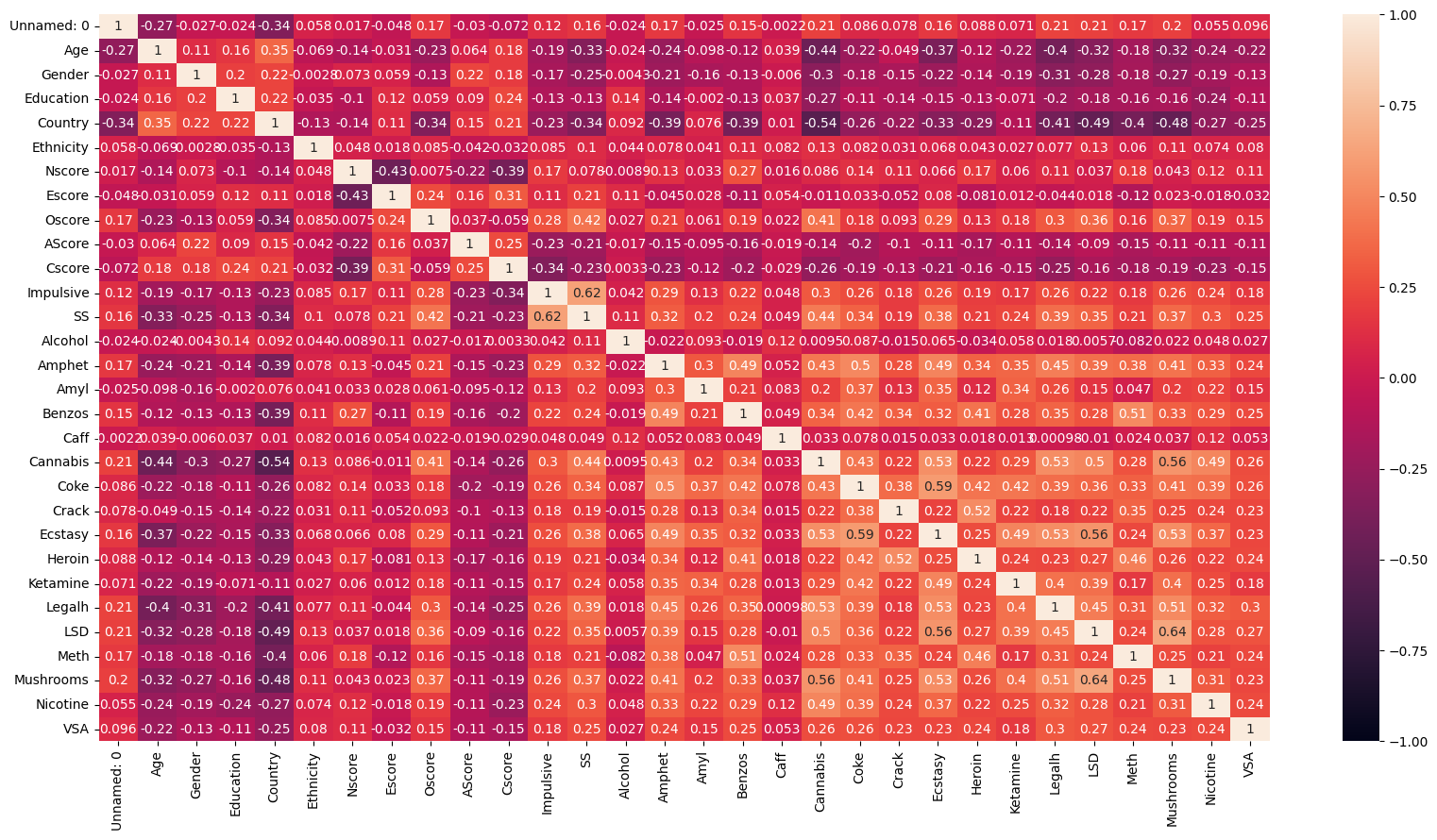
Database contains records for 1885 respondents. For each respondent 12 attributes are known: Personality measurements which include NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking), level of education, age, gender, country of residence and ethnicity. All input attributes are originally categorical and are quantified. After quantification values of all input features can be considered as real-valued. In addition, participants were questioned concerning their use of 18 legal and illegal drugs (alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse and one fictitious drug (Semeron) which was introduced to identify over-claimers. For each drug they have to select one of the answers: never used the drug, used it over a decade ago, or in the last decade, year, month, week, or day.  
Database contains 18 classification problems. Each of independent label variables contains seven classes: "Never Used", "Used over a Decade Ago", "Used in Last Decade", "Used in Last Year", "Used in Last Month", "Used in Last Week", and "Used in Last Day".  
  
Problem which can be solved:  
\* Seven class classifications for each drug separately.  
\* Problem can be transformed to binary classification by union of part of classes into one new class. For example, "Never Used", "Used over a Decade Ago" form class "Non-user" and all other classes form class "User".  
\* The best binarization of classes for each attribute.  
\* Evaluation of risk to be drug consumer for each drug.  
Detailed description of database and process of data quantification are presented in E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan and A. N. Gorban, "The Five Factor Model of personality and evaluation of drug consumption risk.," arXiv [[Web Link]](https://arxiv.org/abs/1506.06297), 2015  
Paper above solve binary classification problem for all drugs. For most of drugs sensitivity and specificity are greater than 75

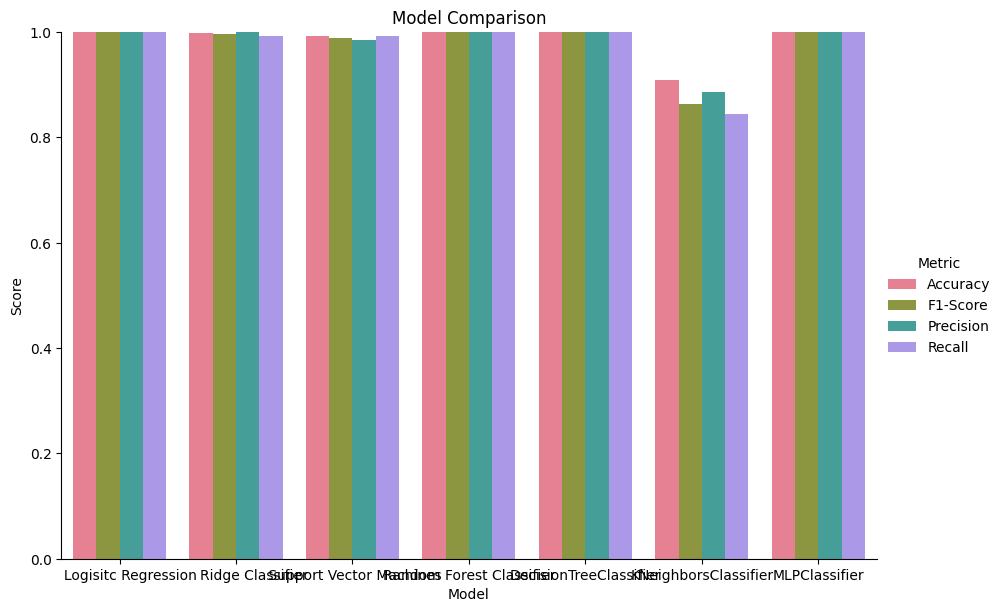
Name:-Ankan Dutta ID No:-E21BCAU0084

BATCH:-1 ASSIGNMENT-MILESTONE 4 AI IN HEALTHCARE

SCREESORT OF THE GRAPHS ANALYSIS:-







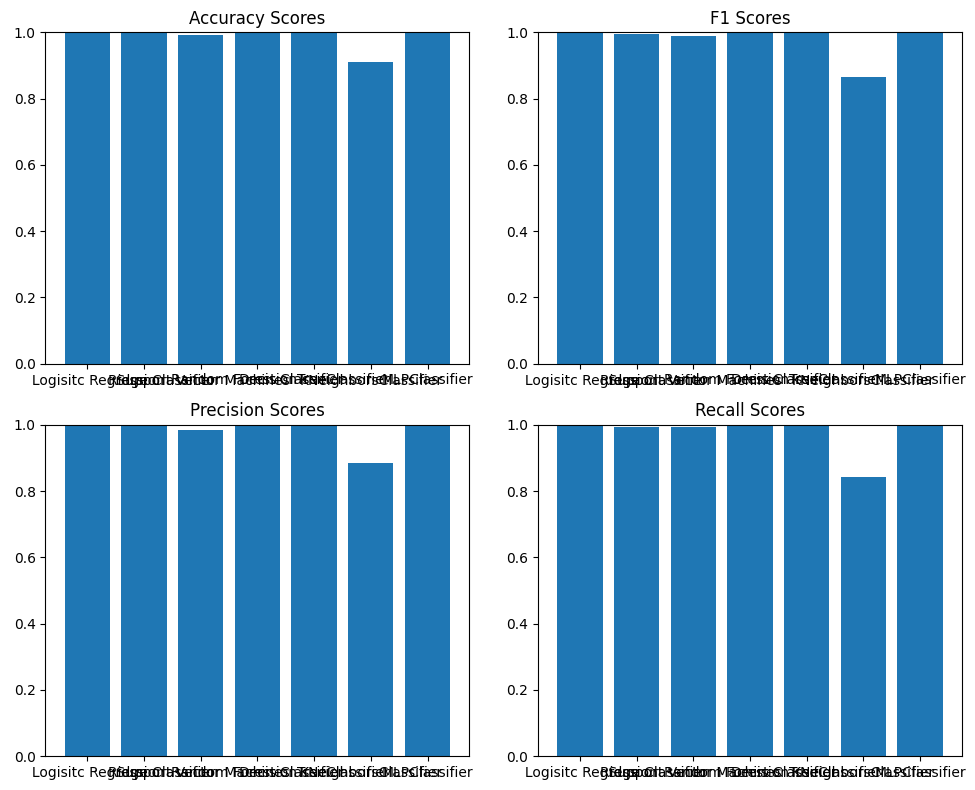


Table 2: Comparative Analysis

|  |  |  |  |
| --- | --- | --- | --- |
| Ref | PC | RR | ACC |
| Fehmarn [1] | 0.676 | 6.89 | 68.91 |
| Kumari [2] | 0.976 | 9.816 | 98.2 |
| Mine [3] | 1.0 | 0.99219 | 99.73 |

Conclusion

In conclusion, the use of machine learning classifiers for drug prediction has shown promising results in recent years. Using a variety of behavioral factors and strategies, such as demographic, environmental and behavioral factors, these models can accurately predict drug use. However, challenges remain in this area, such as the limited availability of useful data and the need for more robust and interpretable models. In addition, ethical considerations should be considered when developing and applying these standards. Despite these challenges, drug use prediction using machine learning has the potential to contribute to drug use prevention and public health promotion.  
Further research and development in this area will be important to increase the accuracy and validity of these models and to ensure their ethical and responsible use.

References:-

[1]:-Fehrman:- Fehrman, E., Egan, V., Gorban, A. N., Levesley, J., Mirkes, E. M., & Muhammad, A. K. (2019). *Personality Traits and*  *Drug Consumption.* doi:10.1007/978-3-030-10442-9

[2]:- Adinugroho:- Adinugroho, S., Sari, Y. A., & Hidayat, N. (2019). *Drug usage duration classification using Extreme Learning Machine based on personality features. 2019 International Conference on Sustainable Information Engineering and Technology (SIET).* doi:10.1109/siet48054.2019.898613

[3]:- Kumari:- Kumari, D., Kilam, S., Nath, P., & Swetapadma, A. (2018). *Prediction of alcohol abused individuals using artificial neural network. International Journal of Information Technology, 10(2), 233–237.* doi:10.1007/s41870-018-0094-3

[4]:- Qiao:- Qiao, Z., Chai, T., Zhang, Q., Zhou, X., & Chu, Z. (2019). *Predicting potential drug abusers using machine learning techniques. 2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS).* doi:10.1109/iciibms46890.2019.899

Github Link:- <https://github.com/UltimateVortex666/Healtcare.git>

Presentation Link:-<https://www.canva.com/design/DAFhfU70IiE/pvtTihkn3Mwlbya-bpFkdw/edit?utm_content=DAFhfU70IiE&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton>