Research Article

Prabhakar Gantela*, Merlin Linda George, Sudheer Reddy Bandi, Nagamani Samineni, and Sonkoju Nagarjuna Chary

A novel behavioral health care dataset creation from multiple drug review datasets and drugs prescription using EDA

https://doi.org/10.1515/comp-2025-0025
received October 11, 2024; accepted February 11, 2025

Abstract: In the current age, the mental condition of people varies rapidly due to various factors such as social relationships, eating disorders, and economic crisis. There are various factors that can be overcome, such as regular exercise, community engagement, meditation, reading books, and yoga. But there are situations where people cannot undergo the mentioned strategies. As advances in data science and big data continue, there is an increasing availability of drug review datasets. There are various manual and traditional approaches to identify the proper drug and condition with some flaws such as overtime, measurement error of the drug, and high computational complexity. Due to these barriers, cutting-edge technologies are involved in data exploration (data cleaning, data transformation, data integration, etc.) to identify the proper prescription of the condition with machine learning approaches. Furthermore, the proposed work has a threefold unique approach that includes the integration of datasets, the creation of a new dataset, and the focus on exploratory data analysis. In the final step, a novel dataset is created from multiple datasets on behavioral healthcare drug reviews that are compared with individual datasets. The main objective of the work is to satisfy the customer's health in all aspects. The work is verified by identifying the prescription for popular health conditions such as anxiety, depression, insomnia, panic disorder, and bipolar disorder.

Keywords: mental health, anxiety, bipolar disorder, data analysis, prescription, machine learning

*** Corresponding author: Prabhakar Gantela,** Department of Information Technology, Mizan Tepi University, Tepi Campus, Tepi, Ethiopia, e-mail: prabhakar@mtu.edu.et
Merlin Linda George: Department of Computer Science and Engineering, Vidya Jyothi Institute of Technology, Hyderabad, India, e-mail: merlingcse@vjit.ac.in
Sudheer Reddy Bandi: Department of Computer Science and Engineering, Swarna Bharathi Institute of Science and Technology, Khammam, India, e-mail: sudheer653@gmail.com
Nagamani Samineni: Department of Computer Science and Engineering, Swarna Bharathi Institute of Science and Technology, Khammam, India, e-mail: maniramesh2004@gmail.com
Sonkoju Nagarjuna Chary: Department of Electronics and Instrumentation Engineering, VNR Vignan Jyothi College of Engineering and Technology, Hyderabad, India, e-mail: nagarjunachary_s@vnrvjiet.in

1 Introduction

Every government is taking priority over mental health over physical health. Also, the study suggests the importance of mental health in all age groups. Various factors affect individual mental health, such as eating disorders, sociodemographic factors, environmental characteristics, and financial situation. The reports by the WHO state that even mental health affects the world economy by 2$ trillion each year. Further, due to these situations, the individual life is affected by social issues, unemployment, educational challenges, and community outreach. Even without overcoming these situations, the increase in the COVID 19 pandemic had a good impact on start-ups in mental health, healthcare companies, etc. [1]. Another study suggested that 10 people are affected by mental health issues for every 100 persons. Other statistics show that over 150 million people in the European region are suffering from mental health conditions, according to a report by WHO. The most important conditions that affect mental health are anxiety, depression, panic disorder, insomnia, stress, obsessive compulsive disorder, and mood variations [2].

Before entering artificial intelligence (AI), exploratory data analysis (EDA) is a good methodology to prepare the dataset for feature extraction and prediction/classification [3]. In addition, it is an important tool in data science that extracts meaningful insights into noise, outliers, visualization, and knowledge exploration with the combination of mathematics and statistics. In addition, the results are very useful for analyzing the results, decision making, and

knowledge extraction. The major works of a data scientist are as follows: (a) Defining the case study. (b) Data/knowledge extraction. (c) EDA. (d) Feature engineering. (e) Developing the model. (f) Working with the model and the case study. EDA is used in the early stages of data mining to understand the data in both a theoretical and a graphical manner [4]. Hence, EDA is a data-driven model to establish the relationship between exploratory variables and outcome variables in various application scenarios such as credit card fraud detection, disaster management analysis, infrastructure management, and retail business management for customer satisfaction. The process of EDA is shown in Figure 1.

In the rapidly evolving landscape of healthcare care, understanding and improving customer satisfaction is paramount. This article delves into a novel analysis of customer satisfaction in mental health care through the lens of machine learning techniques [5]. In the current scenario, AI is used in many applications such as satellite image analysis [6], medical image segmentation [7], biometric recognition [8], and target detection [9]. Using the power of advanced algorithms, our objective is to unravel intricate patterns and insights that can revolutionize the way healthcare providers perceive and respond to patient satisfaction. For more than five decades, researchers have diligently worked on defining patient satisfaction and devising measures to quantify it. In the United States, patient satisfaction surveys play a crucial role in assessing healthcare quality and performance, integral to the medicare formula for hospital reimbursement, all aimed at improving healthcare and patient outcomes. The hospital consumer assessment of health care providers and systems uses a 27-question self-reported feedback survey, emphasizing the patient's hospital recommendations to friends and family. This survey collects data from discharged patients, serving as a valuable tool to assess and improve the performance of healthcare providers.

This research is to bridge the gap between traditional approaches and cutting-edge technology, offering a fresh perspective on optimizing patient experiences in the realm of mental healthcare as given by by Lee et al. [10]. Further, this research aims to address the limitations of previous studies on user satisfaction with healthcare services. Using EDA and machine learning approaches, the focus is to predict and explore factors that influence user satisfaction, loyalty, and continuous user experience in health care services. The objective is to predict customer satisfaction with health care services, enhancing predictive accuracy through various EDA mechanisms and embedding techniques. Exploring patient satisfaction is the aim of this review, emphasizing its crucial role in health care planning and delivery around the world. In addition, the prescription of the drugs is also identified for various top-rated conditions. Identifying the viewpoints of unsatisfied patients is essential to pinpoint areas that need improvement, making it a cornerstone in evaluating and enhancing health care services.

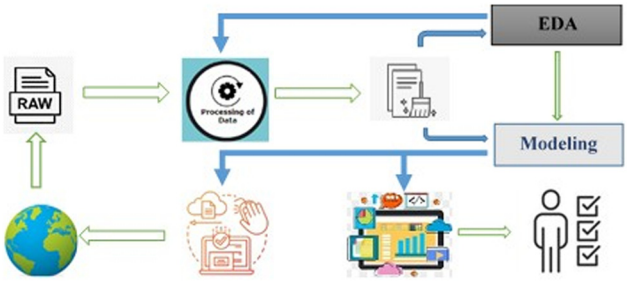

Figure 1: Step by Step procedure to implement the process of EDA.

The proposed work novelty is divided into four folded works as follows:

(1) Creation of a novel dataset: This focuses on creating a new dataset specifically for behavioral health services. This dataset is a useful resource for research and analysis because it compiles data from multiple sources.
(2) Combining different datasets for drug reviews: The collection of data from several drug review websites guarantees that the dataset is extensive and encompasses a wide variety of user experiences and viewpoints about pharmaceuticals.
(3) Prescription drug data emphasis: To enhance the study, the dataset links user assessments with legitimate medical procedures by including prescription facts in addition to reviews.
(4) Applying techniques for EDA: Using EDA as a primary methodological approach implies that patterns, trends, and correlations within the integrated dataset are explored to obtain insights. This stage lays the groundwork for a more in-depth analysis.

This article is divided into the following sections. The literature survey is presented in Section 2, whereas the complete EDA is given in Section 3. Section 4 gives the prescription results to satisfy the patient's needs and machine learning (ML) algorithms to enhance the study. Finally, Section 5 discusses the conclusion and the need for future enhancements.

2 Materials and methods

2.1 Overview

In this section, the overview of the data preprocessing is described in detail and how it is relevant to the proposed

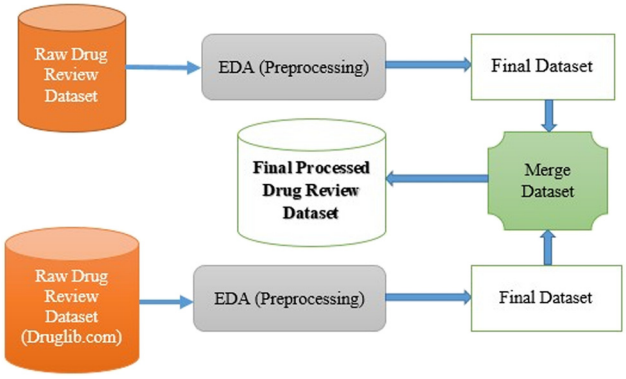

Figure 2: Process of combining two drug review datasets.

work. Next, the two clean datasets are obtained after the preprocessing work is completed. In the next section, the dataset merging process is explained in detail, and the process is depicted in Figure 2. In addition, the general framework of the proposed work is demonstrated in Figure 3.

2.2 Dataset preparation

The dataset is prepared by Kallumadi and Grer available at [11] with 215,063 samples. The various attributes of the data are drug name, date, useful count, condition, review, and rating. The dataset consists of text and multivariate data where the related tasks are classification, clustering, and regression. The main objectives of the dataset are (i) transfer of dataset model to another dataset based on the health condition, and (ii) sentiment analysis on side effects and effectiveness based on drug experience using natural language processing. The dataset is provided at the URL for research purpose https://archive.ics.uci.edu/dataset/462/drug+review+dataset+drugs+com. The second available dataset is prepared by the same people and is available in the study by Kallumadi and Grer [12]. The number of instances are 4,143 with the columns urlDrugName, condition, benefitsReview, sideEffectsReview, commentsReview, rating, and effectiveness. The objectives of the dataset are similar to the aforementioned dataset. The dataset is freely available at https://archive.ics.uci.edu/dataset/461/drug+review+dataset+druglib+com.

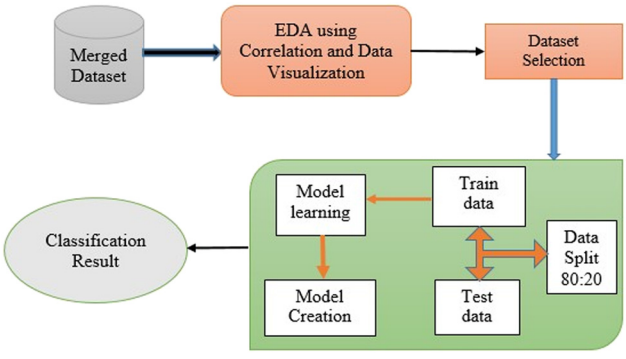

Figure 3: Overview of the proposed work.

2.3 Data cleaning

Data cleaning, which can also be referred to as data cleansing or data scrubbing, is an essential phase in the data mining process. This process includes detecting and rectifying errors, inconsistencies, and inaccuracies within the dataset to ensure that the data are of superior quality and appropriate for analysis [13]. Clean data are essential for accurate and reliable results in data mining. Here are key aspects of data cleaning in data mining.

2.3.1 Handle missing values

Identifying and addressing missing values in the dataset is a common data cleaning task. Strategies include removing instances with missing values, imputing missing values using statistical methods, or considering advanced imputation techniques. In the current working dataset, there are no missing values, and hence, this step is not applicable. In practical situations, absent data in pharmaceutical datasets (such as dosage, adverse effects, and contraindications) can result in insufficient or inaccurate information during real-time prescription evaluations. The proposed solution will substitute missing values for estimated figures (e.g., mean, median, or other more sophisticated techniques).

2.3.2 Dealing with duplicates

Detecting and removing duplicate records is important to avoid bias in analysis and prevent overfitting. Duplicates may arise due to data entry errors, system problems, or merging of datasets. Data redundancy occurs when the same information is represented in multiple ways. Identifying and eliminating redundancy help streamline the dataset and improve efficiency. Duplicated entries can distort data analysis, exaggerate statistics, and create misunderstandings during drug evaluations. The suggested approach offers a distinctive solution employing algorithms to detect, consolidate, or eliminate duplicate records utilizing unique identifiers or comparable attributes.

2.3.3 Outlier detection and treatment

Identifying and handling outliers are crucial to prevent them from disproportionately influencing the analysis. Techniques include visual inspection, statistical methods, and machine learning models to detect and handle outliers appropriately. Outliers, such as excessively high doses or uncommon adverse events, can skew the analysis and may result in erroneous clinical conclusions. A primary approach to address this issue is to employ statistical methods and techniques, such as box plots, scatter plots, or z scores, to pinpoint outliers. Furthermore, appropriate treatment should involve either eliminating outliers (if they are erroneous) or applying robust statistical techniques that are less affected by extreme values.

2.3.4 Handling noisy data

Noise in data refers to random variations or errors that can interfere with analysis. Filtering out noise involves the use of smoothing techniques, robust statistical measures, or outlier detection methods. Erroneous or random fluctuations in data can mask genuine patterns and impact the precision of drug evaluations in real time. This proposed project employs data smoothing methods such as moving averages or filtering to minimize noise.

2.3.5 Addressing incomplete data

Incomplete data refer to instances where certain attributes or variables are missing for some records. Decisions must be made on how to handle incomplete data, whether through imputation, removal of instances, or other strategies. The data in the first dataset are missing for the attribute (categorical) condition. As there are numerous types of conditions in health, the missing values cannot be filed, so the instances are removed from the dataset. Insufficient data can obstruct thorough evaluations of medications and may result in overlooked interactions or contraindications. The proposed approach incorporates data augmentation and collaboration with additional information from external sources or databases, as well as collaboration with data providers to improve data completeness.

2.3.6 Data standardization and normalization

This approach involves converting values to a consistent format or scale. This ensures that data from different sources or with varying units of measurement can be effectively compared and analyzed. These data are very much needed to maintain the standard deviation balance and helps maintain the uniformity and compatibility of the data. Varied data formats and scales can pose challenges in merging data from multiple sources and conducting significant comparisons. Different approaches, such as standardization (e.g., employing standard drug codes or measurement units) and normalization by adjusting data to a defined range (such as 0–1), help ensure that variables with higher values do not overshadow the analysis.

2.3.7 Data quality assessment

Data quality is a measure of the state of a dataset based on variables such as accuracy, completeness, consistency, reliability, and validity. These qualities guide the decision-making process during data cleaning as described in [14]. When such measurements are studied to diagnose the quality of data gathering, the phrase (data quality) evaluation is employed. Data quality sets corporate goals, reviews data across different dimensions, analyzes assessment results, designs and develops plans based on earlier analysis, implements solutions, and verifies at regular intervals to ensure consistent goals. Inaccurate drug reviews can result from poor data quality, which may endanger patient safety. The solutions proposed from this work include data profiling, data validation, and periodic audits.

2.4 Data integration

This mechanism includes merging and bringing data to a central point from several sources to build a strong and cohesive dataset for analysis [15]. The important aspects of data integration are error reduction, contribution to decision-making, data efficiency, and data collaboration. Data for data mining can come from databases, spreadsheets, text files, application programming interfaces, and more. These sources may include different data types, structures, and representations. The key steps in data integration (DI) are the collection of requirements, data profiling, the identification of the gaps between the requirements and the evaluation, the implementation of the DI process, and finally the verification, validation, and monitoring of the results of the DI process. The proposed datasets include two types of data: text and integers.

- Concatenation: Combining datasets by appending rows or columns.
- Merging/joining: Combining datasets based on common attributes.
- Aggregation: Combining data by summarizing or aggregating values.

2.5 Data transformation

Data transformation is a crucial step in the data mining process, where raw data are converted to a suitable format for analysis [16]. The main intention of data transformation is to strengthen the quality of the data, making it more appropriate for the specific requirements, and to perform statistical analysis of the data mining algorithms. Further, these are some key aspects of data transformation in the context of data mining. Data transformation often involves data cleaning to handle NaN values, empty cells, outliers, and inconsistencies. Techniques like imputation, removal of duplicates, and handling of outliers are applied during this phase. The strengths of these techniques are faster query response, enhanced data quality, data management, better data usage, and data organization.

- Normalization and standardization: Normalization involves scaling numerical attributes to a standard range, typically between 0 and 1, to ensure that different scales do not affect the performance of certain algorithms. Standardization involves transforming numerical attributes to have a mean of 0 and a standard deviation of 1.
- Encoding categorical data: Many data mining algorithms work with numerical data, so categorical variables need to be transformed into a numerical format through techniques such as one-hot encoding or label encoding.
- Aggregation and discretization: Aggregation involves combining multiple data values into a single value, often for summarization purposes. Discretization involves converting continuous numerical attributes into discrete intervals or categories.
- Variable transformation: Transforming variables can include mathematical operations (e.g., logarithmic transformations) to handle skewed distributions and make relationships between variables more linear. Creating new variables based on existing ones through feature engineering.

2.6 Data reduction

Reducing the dimensionality of the dataset using techniques such as principal component analysis (PCA) or feature selection. Dimensionality reduction can help simplify the dataset and speed up analysis. Data reduction in data mining refers to the process of reducing the volume but producing the same or similar analytical results as explored in the study by Huang et al. [17]. It involves selecting, transforming, or simplifying the data to make it more manageable but still representative of the underlying patterns and trends. Data reduction has potential applications when dealing with large datasets, as it helps to improve the efficiency of mining algorithms, reduces computational complexity, and mitigates the risk of overfitting. Here are some common techniques used for data reduction in data mining.

(1) Principal component analysis (PCA): PCA is a technique that transforms the original variables into a new set of uncorrelated variables, known as principal components. It helps reduce the dimensionality of the data while retaining as much variability as possible.
(2) Feature selection: Identify and select the most relevant features (attributes or variables) for the analysis, based on statistical measures, information gain, or other criteria. Removing redundant or irrelevant features can simplify the dataset. The features such as uniqueID, date, and useful count are removed using the drop() function in the Python library. Similarly, the columns such as sideEffectsReview and commentsReview are removed from the second dataset.
(3) Histograms and clustering: Representing data using histograms can help reduce the number of data points by grouping them into intervals, making it easier to analyze trends and patterns. Grouping similar instances into clusters and representing each cluster by its centroid or other summary statistics. This can significantly reduce the number of data points while maintaining the overall structure of the data.
(4) Binning: Grouping continuous data into bins or intervals. This helps reduce the impact of outliers and simplifies the data representation.
(5) Numerosity reduction: The alternative representation of data to reduce the data capacity. There are two popular methods: (i) parametric methods based on models such as log-linear and regression and (ii) nonparametric methods such as histograms, sampling, and clustering without any modeling approach.
(6) Wavelet transforms: These techniques are useful in signal processing and multiresolution analysis. The essence of wavelet transforms exists in the analysis of details at every level.
(7) Sampling: Sampling is a methodology used to select a subset of data from a larger dataset for analysis. The purpose of sampling is to make analysis more manageable, reduce computational complexity, and often provide faster results [18]. There are various sampling methods, and the choice of a particular method depends

on the nature of the data, the goals of the analysis, and the available resources. Here are some common sampling techniques in data mining: (a) random, (b) systematic, (c) stratified, (d) cluster, (e) strategic, (f) sequential, (g) reservoir, (h) under, and oversampling.

2.7 Discretization

Discretization, or feature (attribute) discretization, is a data preprocessing technique in data mining that converts continuous data into smaller intervals or bins, hence the name binning. The purpose of discretization is to display the data in a more manageable and understandable way [19]. This process is specifically useful for certain types of data mining algorithms that work more effectively with categorical or discrete data. Here are key aspects of discretization in data mining. The motivation for discretization lies in handling algorithm requirements, such as decision trees or association rule mining algorithms, which may perform better with categorical or discrete data rather than continuous data. Discretization may be applied to mitigate the impact of skewed distributions in continuous variables. Discretized data can be easier to interpret, especially when presenting results to nonexpert stakeholders.

Types of discretization:

(1) Equal width binning: Divides the range of continuous values into equal-width intervals. For example, if values range from 0 to 100, and you want five bins, each bin would cover a range of 20 (0–20, 21–40, etc.).
(2) Equal frequency binning: Divides the data into intervals containing approximately the same number of instances. This approach helps ensure that each bin has a similar distribution of data points. As the side effects and effectiveness are categorized into five object types, to handle the condition and rating effectively, the range of both these features is numbered from 0–4.
(3) Clustering-based discretization: Using clustering algorithms to group similar values together, creating intervals based on the clusters identified.
(4) Decision tree-based discretization: Involves the construction of a decision tree to find suitable split points that maximize information gain or other criteria. The decision tree identifies thresholds to create intervals.
(5) Unsupervised discretization: Does not use the target variable for determining bins. Equal width, equal frequency, and clustering-based discretization fall into this category.
(6) Supervised discretization: Takes the target variable into account during the discretization process. Decision tree-based methods, for example, involve finding split points that improve the predictive accuracy of the model.

2.8 Feature extraction

Feature extraction is a decisive technique in various fields such as data analysis, data mining, and machine learning, as analyzed in the study by Tran et al. [20]. The main motive is to transform the data from an actual source into a new representation that highlights the discriminative features needed for a specific task. It acts as a wrapper for the entire data mining process. The goal of feature extraction is dimensionality reduction, paying attention to key attributes, and improving the performance of data mining approaches. Feature extraction is a completely necessary step in the data preparation pipeline that contributes to a more compact and useful representation of the data. The technique used is determined by the dataset's features and the goals of the data mining or machine learning task.

Here are key aspects of feature extraction in data mining:

→ Dimensionality reduction: In this technique, the primary objective of feature extraction is to reduce the number of features in the dataset. Dimensionality reduction helps overcome the curse of dimensionality, making the data more manageable and improving the efficiency of algorithms.
→ Feature selection vs feature extraction: Feature selection involves choosing a subset of original features based on their relevance to the task. Selected features are retained, and others are discarded [21]. However, feature extraction involves transforming the original features into a new set of features, often using mathematical transformations. The new features are a combination of the original ones.
→ Benefits of feature extraction: To begin with, extracting features can enhance the performance of ML models by focusing on the most relevant elements of the data. Secondly, a reduced feature set leads to faster training and prediction times. Finally, extracted features are often more amenable to visualization, aiding in the interpretation of the data.
→ Considerations and challenges: Feature extraction involves summarizing or transforming data, which can lead to data loss. In addition, the selection of a particular attribute may affect the performance of algorithms in different ways. Furthermore, depending on the task, it is necessary to observe the data mobility from training to prediction, that is, the causes and effects of the decision.
→ Splitting a dataset into training and testing sets: As part of the development, construction, and evaluation of machine learning models, data are divided into three categories: training, validation, and testing sets. The reason for this division is to evaluate the model's ability

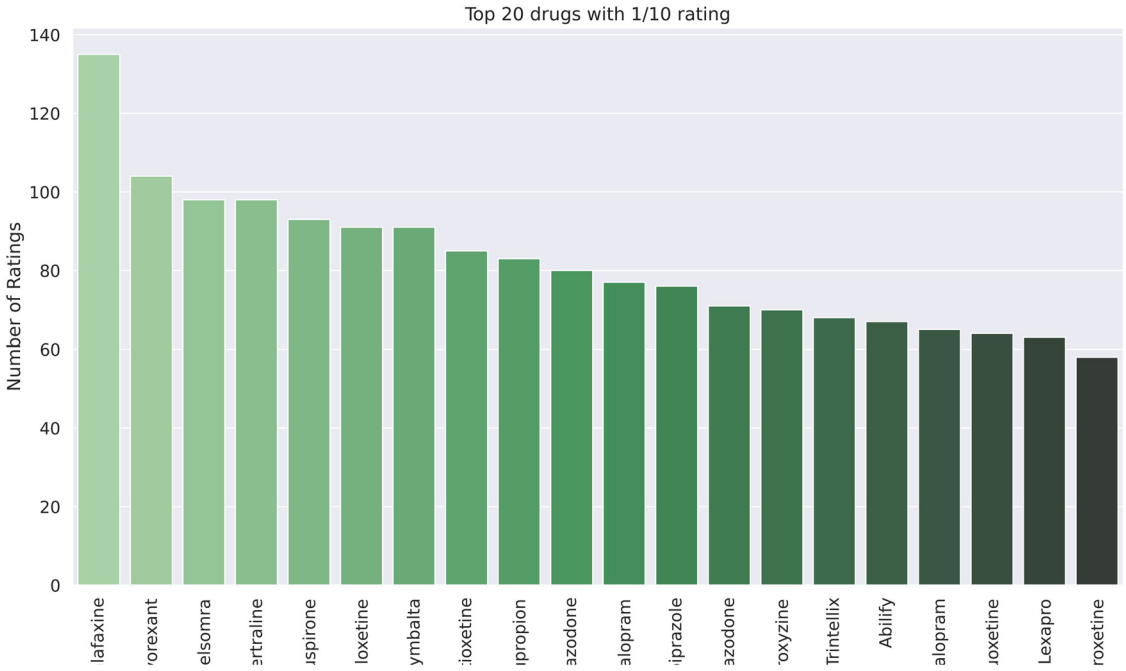

Figure 4: First drug review dataset with one rating for twenty drugs.

to generalize to unfamiliar data. The following are the essential components of dividing a dataset into training and testing sets:

(a) Training set: The training set is the subset of the dataset used to train the machine learning model. The model recognizes patterns, correlations, and features based on this selection of data. Typically, the training set accounts for a significant portion of the total dataset, often 70–80%.

(b) Testing set: The testing set is set aside to evaluate the model's performance on new, previously unknown data. The testing set is used to assess the model's

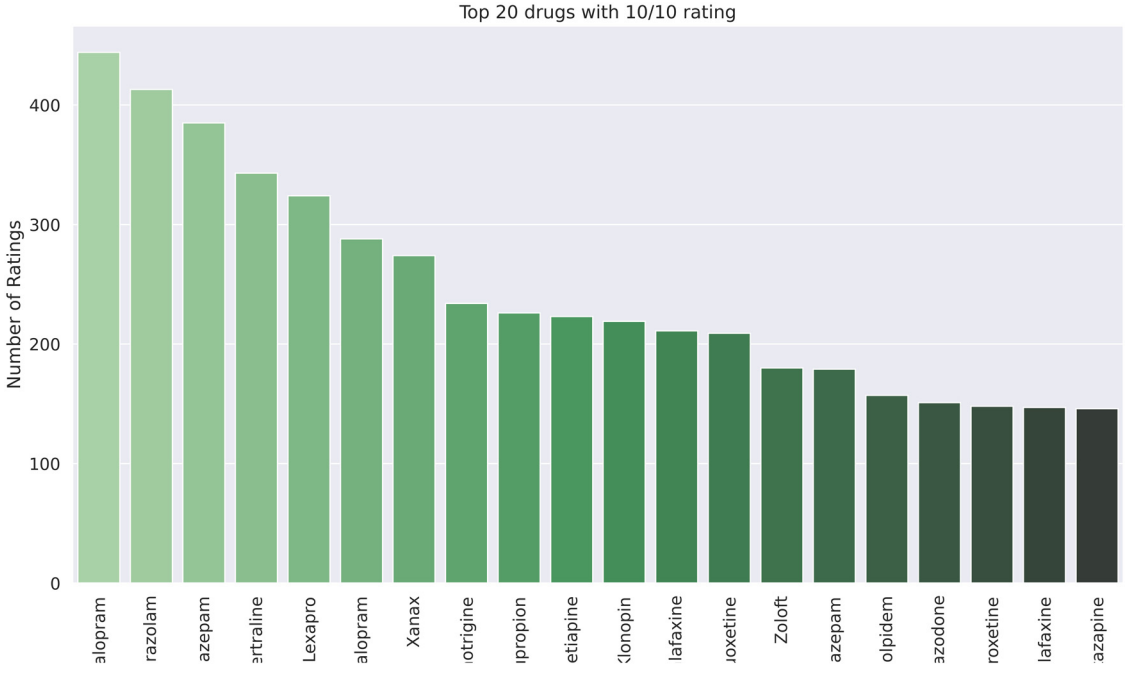

Figure 5: First drug review dataset with ten rating for 20 drugs.

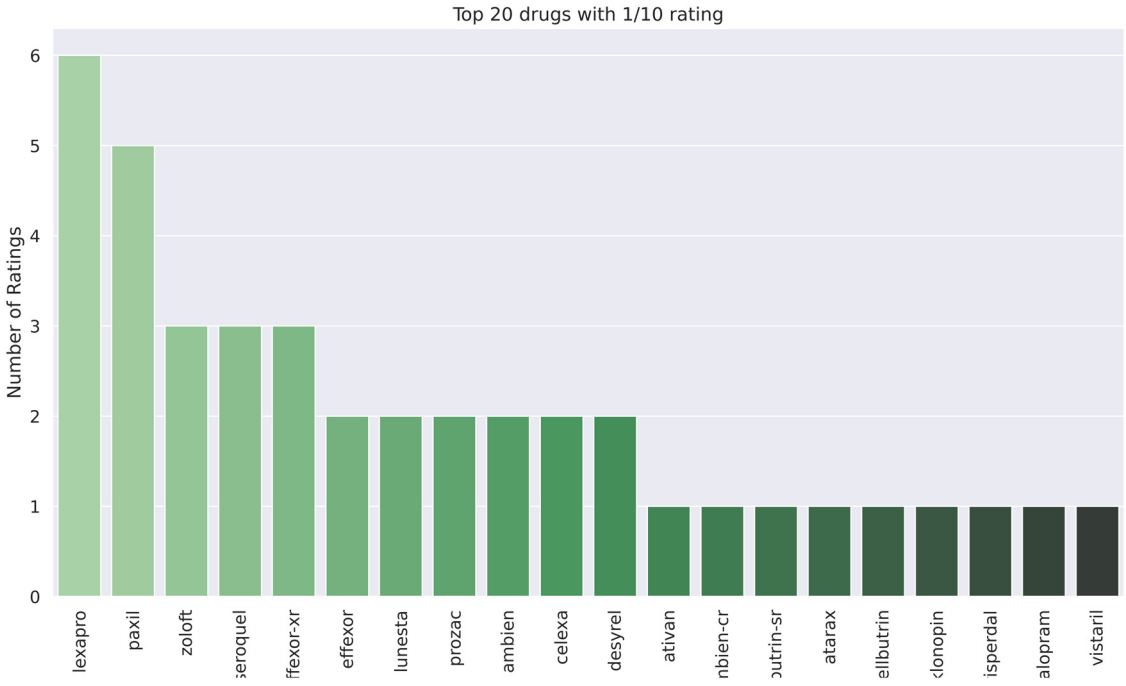

Figure 6: Second drug review dataset with one rating for 20 drugs.

ability to generalize and predict data that they were not exposed to during training. The testing set is critical for determining the model's performance and detecting any overfitting. The test set should be kept confidential during model development. Model adjustments based on test set performance can lead to overfitting the test set, compromising its ability to provide an unbiased evaluation.

→ Scaling features: Transforming numerical features through scaling ensures that all features have a

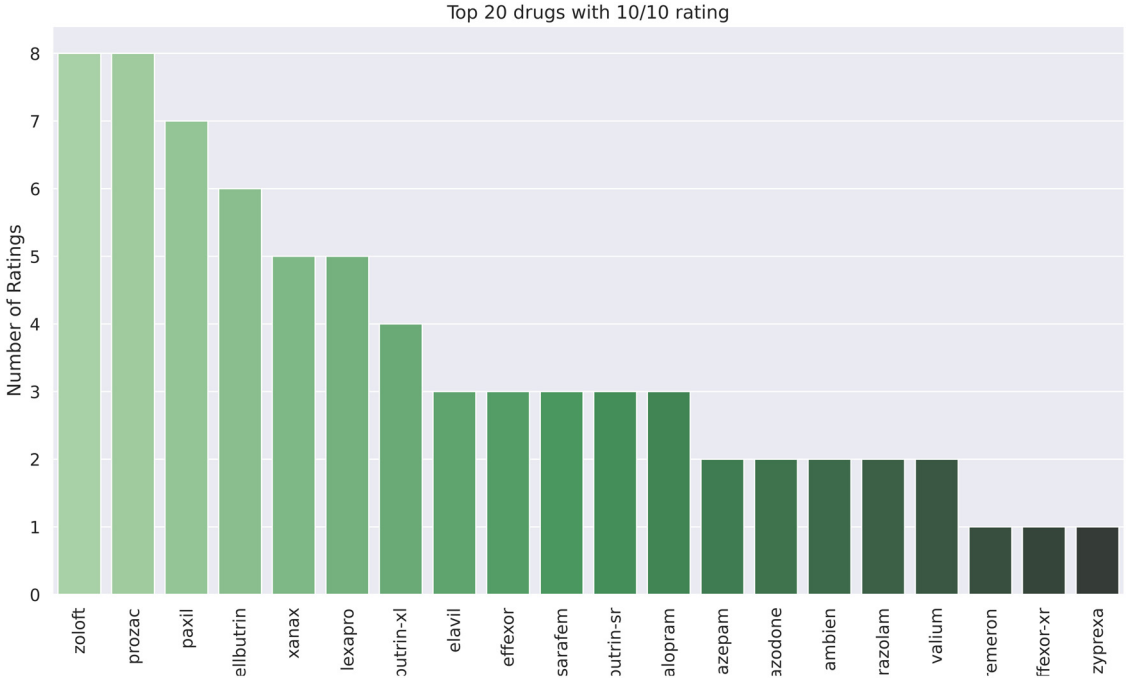

Figure 7: Second drug review dataset with ten rating for 20 drugs.

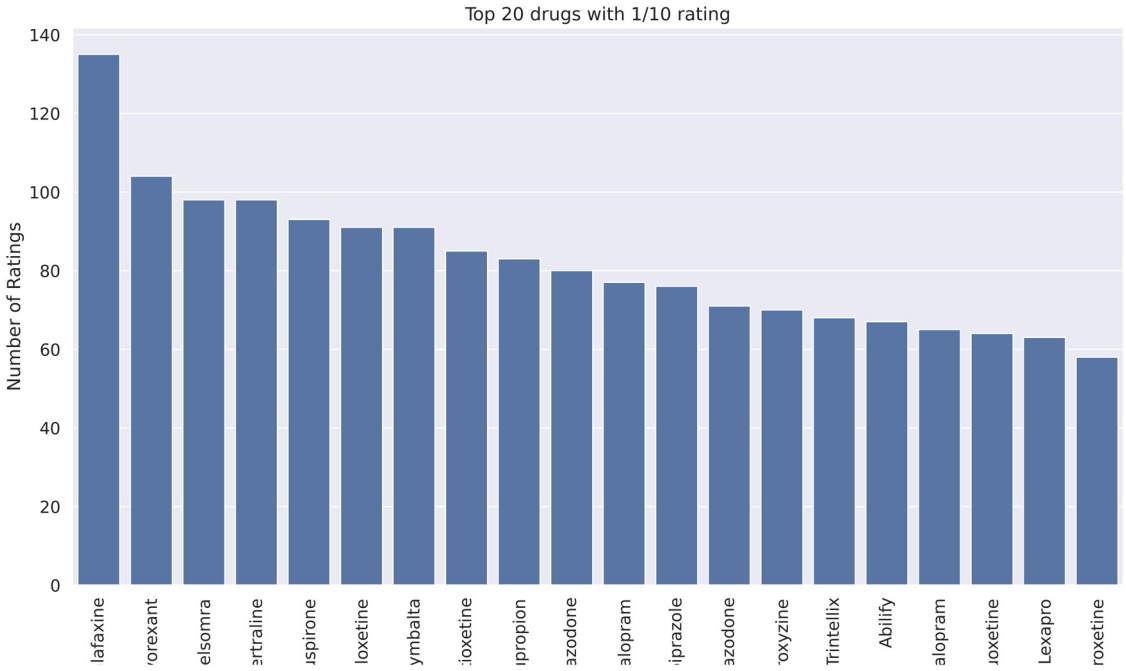

Figure 8: Integrated drug dataset with one rating for 20 drugs.

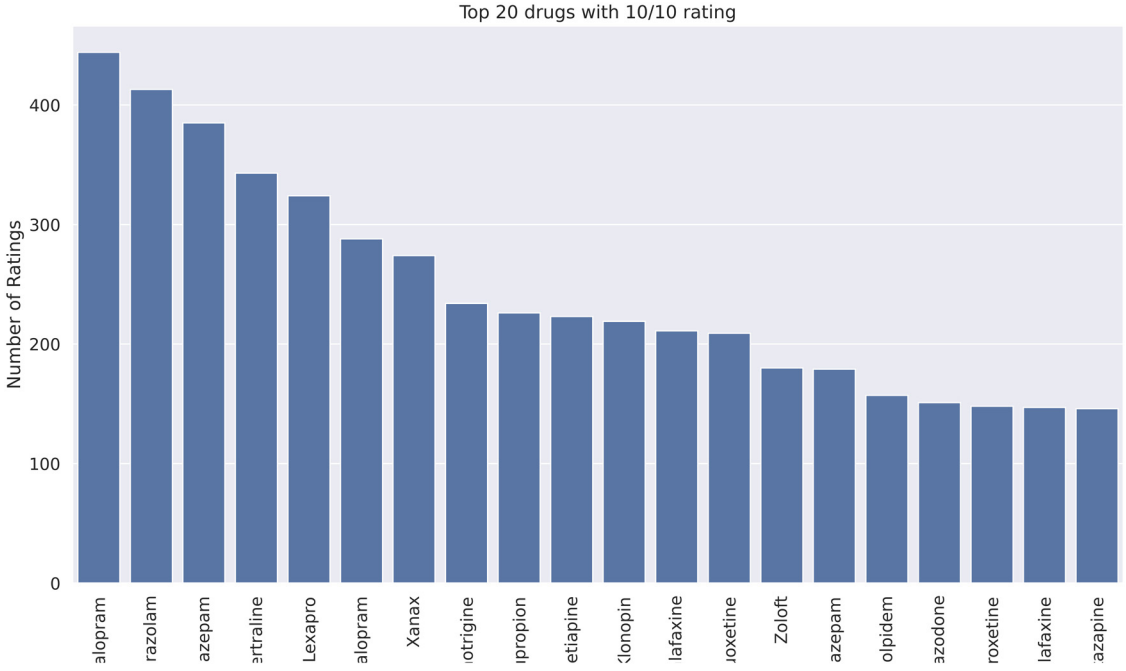

Figure 9: Integrated drug dataset with ten rating for 20 drugs.

similar scale or range, preventing any single feature from dominating the learning process or negatively impacting the performance of specific algorithms.

2.9 Preparation of the integrated dataset

As we have seen in the aforementioned sections, the working of EDA in data science, after completion of

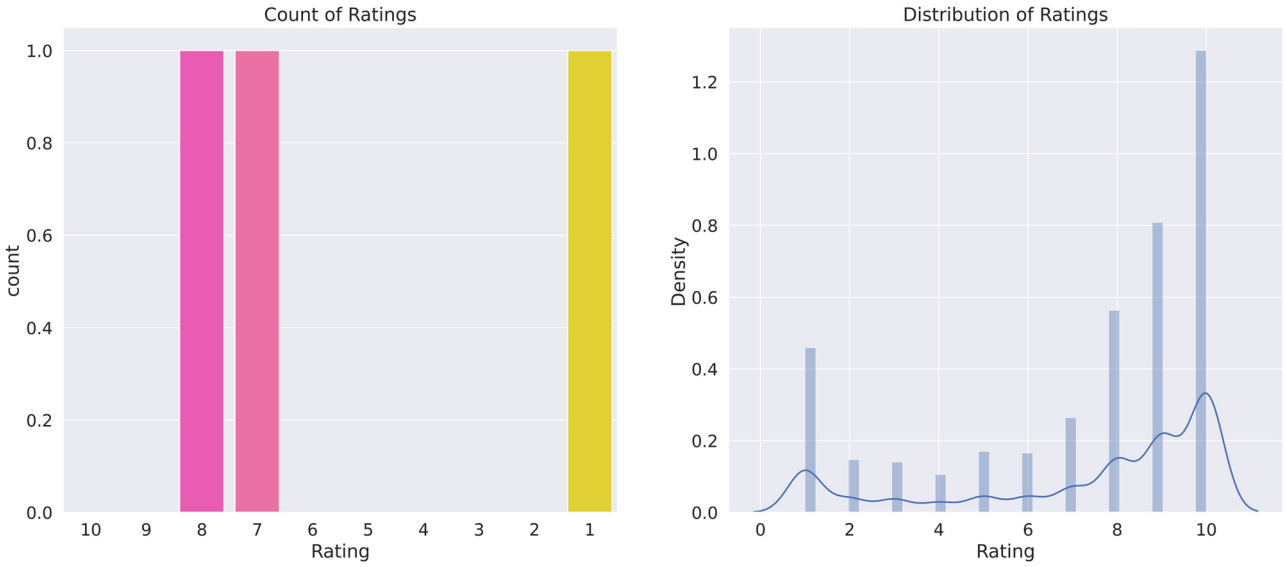

Figure 10: Distribution of rating for the first drug review dataset.

the EDA process, a clean and proper dataset is formed in the two csv files. The two files have different column names and different column positions. The column positions are rearranged, and the names are changed in common to the first dataset. Hence, the merged dataset consists of 34,589 instances with six attributes. The attributes are defined as drugName, condition, review, rating, date, and usefulCount.

3 Experimental results and analysis

3.1 EDA results

In this section, the EDA results are described from the integrated datasets. After integration, the dataset has six important columns named as drugName, rating, effectiveness, sideEffects, condition, and review. Further, the

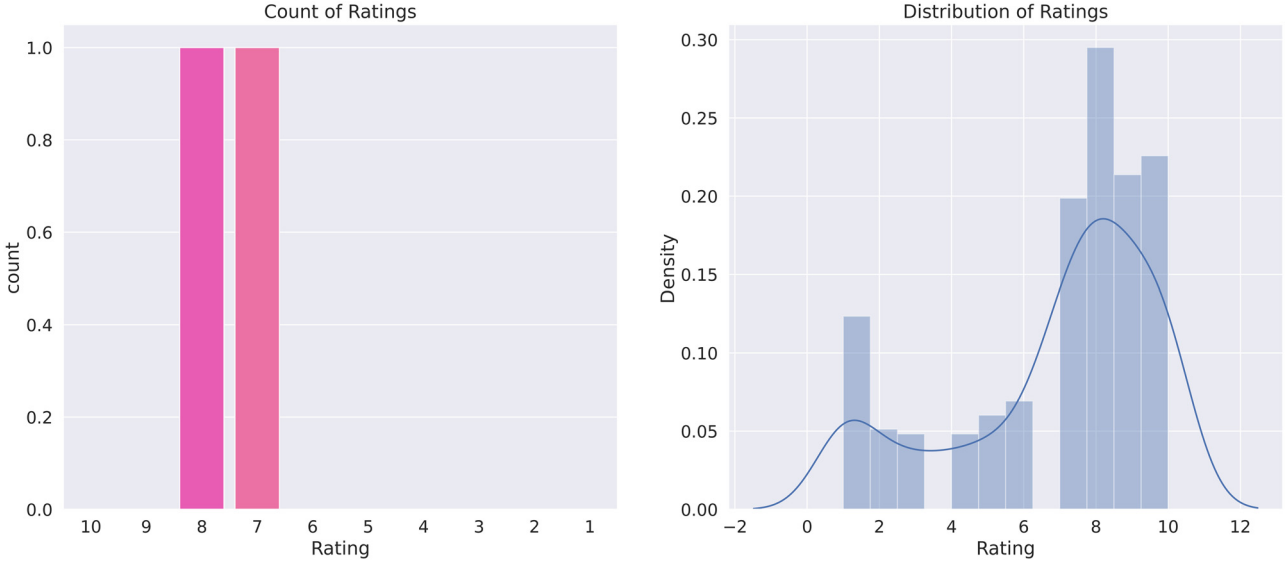

Figure 11: Distribution of rating for the second drug review dataset.

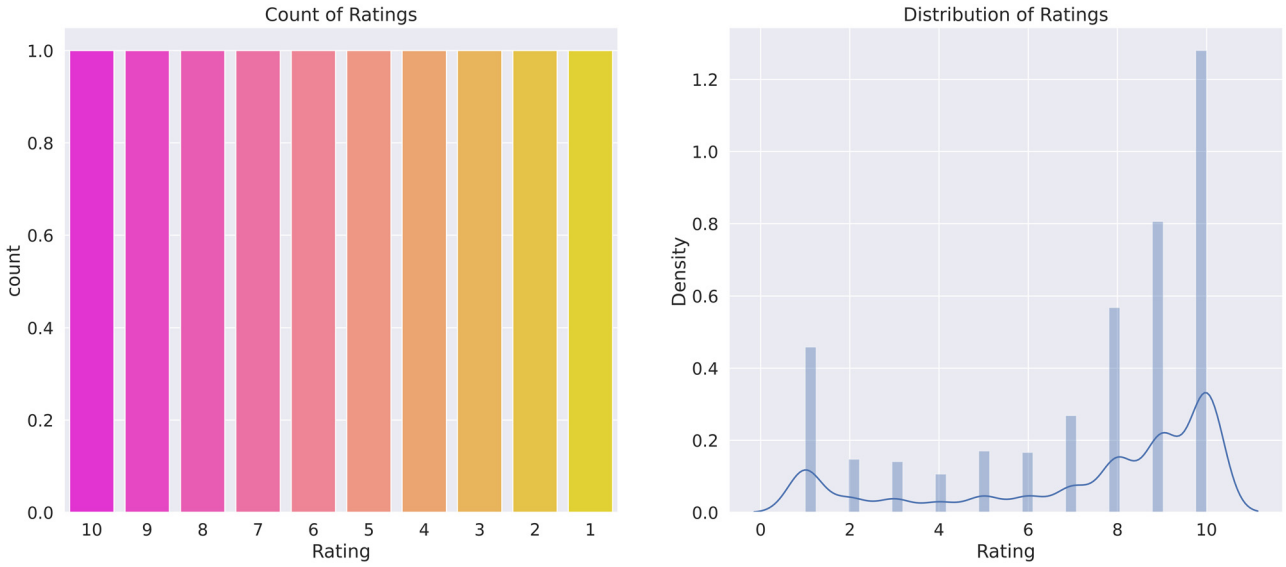

Figure 12: Distribution of rating for the integrated drug review dataset.

respective data types are object, integer, integer, integer, object, and object type. The drugs with the highest and lowest rating for the two datasets are shown in Figures 4–7 respectively. After merging the dataset, the results are given as Figures 8 and 9 for the lowest and highest rating. In addition, the distribution of drugs before and after merging is denoted in Figures 10–12. The diagrammatic representation indicates that the first dataset and the resultant dataset have a better distribution than the second. The uniformity increases to a certain extent after the merging process.

3.2 Classification algorithms

The various classifiers used for classification are multilayer perceptron (MLP), naive Bayes classifier (NBC), random forest classifier, and logistic regression classifier (LRC). First, MLP is based on neurons, layers, etc., whereas NBC depends on the Bayes theorem and probabilities. On the other hand, random forest classifier (RFC) is based on the number of nodes and trees with a huge impact on the large dataset. Finally, the LRC depends on the logistic function to finalize the value of the dependent variable. All algorithms are formulated in the Python language in Google colaboratory, and the metrics and accuracy values are referred to in Tables 1 and 2 respectively. The results in the table prove that MLP has greater dominance than traditional approaches. NBC and LRC are less effective for the large dataset and need more computational resources.

Table 1: Comparison of different classifier performance metrics for all the mental health care drug review datasets

Classifier	Dataset	Precision	Recall	AUC curve
MLP	First	95.5	96.0	96.2
	Second	94.4	94.9	95.0
	Integrated	98.7	98.5	98.7
NBC	First	92.8	93.3	93.5
	Second	91.9	92.4	92.5
	Integrated	94.3	95.0	94.8
RFC	First	94.0	94.5	94.7
	Second	92.9	93.4	93.6
	Integrated	95.2	95.7	96.0
LRC	First	93.3	93.8	94.0
	Second	92.0	92.5	92.7
	Integrated	96.2	96.8	97.0

Table 2: Comparison of different classifier performance for all the mental health care drug review datasets

Classifier	First dataset	Second dataset	Merged dataset
MLP	95.8	94.67	98.7
NBC	93.07	92.1	94.66
RFC	94.27	93.16	95.48
LRC	93.56	92.3	96.51

Table 3: Comparison with the existing approaches

Classifier	[22]	[23]	[24]	Proposed
MLP	96.0	95.8	96.3	98.7
NBC	92.3	92.8	93.0	94.66
RFC	94.0	93.8	94.7	95.48
LRC	94.0	94.2	94.4	96.51

Finally, the popular drugs used by most people are Venlafaxine, Lamotrigine, Clomipramine, Alprazolam, Quetiapine, Klonopin, Gabapentin, Pristiq, Parnate, and Xanax.

3.3 Analysis of results and algorithms

The values in Tables 1 and 2 indicate the good performance of the integrated dataset with the MLP technique due to the diversity and sample size of the dataset. The MLP shows the highest performance because of the layered architecture and robustness in handling the diverse data. The NBC gives poor performance because of the independence nature of attributes and simple assumptions with the dataset features. Finally, the proposed work is compared with three review techniques in Table 3 such as as by Askr et al., Qiu et al., and Vo et al. [22–24]. The aforementioned approaches were implemented with multiple drug dataset.

4 Conclusion and future recommendation

With the rapid growth of mental health issues, there are barriers to overcome the situations of communal life. Hence, this work proposed unique approaches in the data preprocessing to obtain the merged dataset from two different datasets. The solution of the merged dataset was to improve the awareness of patients about various drugs based on side effects, effectiveness, rating, and comments. This improves patient satisfaction with a particular condition/drug. In addition, from the merged dataset various results such as popular conditions, popular drugs, and useful reviews were obtained. In addition, classification algorithms were used to know the performance of the dataset obtained. The proposed work might face difficulties in ensuring the quality and representativeness of the dataset due to the potential biases or inconsistencies found in drug reviews and prescription data sources. Furthermore, relying solely on EDA could limit the ability to obtain deeper causal insights or establish predictive modeling abilities, which are crucial for effective solutions in behavioral health care. In the future, the proposed work will be used to implement various conditions such as birth control, diabetes, hypertension, etc. In addition, the dataset can be employed with various deep learning algorithms to compare performance with machine learning algorithms.

Acknowledgements: The authors extend their gratitude to the authors Surya Kallumadi and Felix Grer who provided the Mental Healthcare Drug Review dataset for research purposes.

Funding information: This research did not receive external funding.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: The authors declare that there is no conflict of interest with respect to the publication of this article.

Data availability statement: Data used to support the findings of this study are included in the article.

References

[1] R. Guthold, L. Carvajal-Velez, E. Adebayo, P. Azzopardi, V. Baltag, S. Dastgiri, et al., "The importance of mental health measurement to improve global adolescent health," *J. Adol. Health*., vol. 72, no. 1, pp. s3–s6, 2023.

[2] Z. Fu, H. Burger, R. Arjadi, and L. Bockting, "Effectiveness of digital psychological interventions for mental health problems in low-income and middle-income countries: a systematic review and meta-analysis," *The Lancet Psychiatry*, vol. 7, no. 10, pp. 851–864, 2020.

[3] X. Li, Y. Zhang, J. Leung, C. Sun, and J. Zhao, "EDAssistant: Supporting exploratory data analysis in computational notebooks with J. Zhao, EDAssistant: code search and recommendation," *ACM Trans. Inter. Int. Sys*. vol. 13, no. 1, pp. 1– 27, 2023.

[4] M. R. Sieber, C. Russ, and K. Kurz, "Organizational culture and business-IT alignment in COVID-19: A swiss higher education case study," *Int. J. Inn. Tech. Mgmt*., vol. 20, no. 3, p. 2242004, 2023.

[5] S. Graham, C. Depp, E. E. Lee, C. Nebeker, X. Tu, H. C. Kim, et al., "Artificial intelligence for mental health and mental illnesses: an overview," *Curr. Psy. Reports*, vol. 21, pp. 1–18, 2019.

[6] S. R. Bandi, M. Anbarasan, and D. Sheela, "Fusion of SAR and optical images using pixel-based CNN," *Neu. Net. World*, vol. 32, no. 4, pp. 197–213, 2022.

[7] S. Huang, M. Huang, Y. Zhang, J. Chen, and U. Bhatti, "Medical image segmentation using deep learning with feature enhancement," *IET Image Proc*., vol. 14, no. 14, pp. 3324–3332, 2020.

[8] M. L. George, T. Govindarajan, K. Angamuthu Rajasekaran, and S. R. Bandi, "A robust similarity based deep siamese convolutional neural network for gait recognition across views," *Co. Intel*., vol. 36, no. 3, pp. 1290–1319, 2020.

[9] T. Fan, "Research and realization of video target detection system based on deep learning," *Int. J. Wav. Multi. Inf. Proc*., vol. 18, no. 1, p. 1941010, 2020.

[10] Y. Lee, R. M. Ragguett, R. B. Mansur, J. J. Boutilier, J. D. Rosenblat, and A. Trevizol, et al., "Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta analysis and systematic review," *J. Aff. Dis.*, vol. 241, pp. 519–532, 2018.

[11] S. Kallumadi and F. Grer, *Drug Review Dataset (Drugs.com). UCI Machine Learning Repository*, 2018, doi: https://doi.org/10.24432/C5SK5S.

[12] S. Kallumadi and F. Grer, *Drug Review Dataset (Druglib.com). UCI Machine Learning Repository*, 2018, doi: https://doi.org/10.24432/C55G6J.

[13] O. Oyebode, F. Alqahtani, and R. Orji, "Using machine learning and thematic analysis methods to evaluate mental health apps based on user reviews," *IEEE Access*, vol. 8, pp. 111141–111158, 2020.

[14] R. Tornero-Costa, A. Martinez-Millana, N. Azzopardi-Muscat, L. Lazeri, V. Traver, and D. Novillo-Ortiz, "Methodological and quality flaws in the use of artificial intelligence in mental health research: systematic review," *JMIR Mental Health*, vol. 10, no. 1, p.e42045, 2023.

[15] K. L. Grazier, M. L. Smiley, and K. S. Bondalapati, "Overcoming barriers to integrating behavioral health and primary care services," *J. Pri. Care Community Health*, vol. 7, no. 4, pp. 242–248, 2016.

[16] J. W. Foreman, *Data smart: Using data science to transform information into insight*, 2nd edn., John Wiley and Sons, New York, 2013.

[17] X. Huang, L. Wu, and Y. Ye, "A review on dimensionality reduction techniques," *Int. J. Patt. Rec. Art. Int.*, vol. 33, no. 10, p. 1950017, 2019.

[18] R. Latpate, J. Kshirsagar, V. K. Gupta, and G. Chandra, *Advanced sampling methods*, 1st edn., Springer, Singapore, 2021.

[19] P. Barlas, I. Lanning, and C. Heavey, "A survey of open source data science tools," *Int. J. Intel. Comp. Cyber.*, vol. 8, no. 3, pp. 232–261, 2015.

[20] T. Tran, W. Luo, D. Phung, S. Gupta, S. Rana, and R. L. Kennedy, et al., "A framework for feature extraction from hospital medical data with applications in risk prediction," *BMC Bioinformatics*, vol. 15, no. 1, pp. 1–9, 2014.

[21] G. Tsang, S. M. Zhou, and X. Xie, "Modeling large sparse data for feature selection: hospital admission predictions of the dementia patients using primary care electronic health records," *IEEE J. Transl. Eng. Health Medi.*, vol. 9, pp. 1–13, 2020.

[22] H. Askr, E. Elgeldawi, H. Aboul Ella, Y. A. Elshaier, M. M. Gomaa, and A. E. Hassanien, "Deep learning in drug discovery: an integrative review and future challenges," *Artif. Intell. Rev.*, vol. 56, no. 7, pp. 5975–6037, 2013.

[23] Y. Qiu, Y. Zhang, Y. Deng, S. Liu, W. Zhang, "A comprehensive review of computational methods for drug-drug interaction detection," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 19. no. 4, pp. 1968–85, 2021.

[24] T. H. Vo, N. T. Nguyen, Q. H. Kha, N. Q. Lenq, "On the road to explainable AI in drug-drug interactions prediction: A systematic review," *Comput. and Struct. Biotechnol. J.*, vol. 20, pp. 2112–23, 2022.