Journal of Cardiovascular Disease Research

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 10, 2023

Multi-class Drug Classification using Machine Learning Models

C. Rashmi[1], A. Geethika[2], B. Lakshmi[2], G. Rakshitha[2], Ch. Swathi[2]

[1] Assistant Professor, Department of Information Technology, Mallareddy Engineering College for Women, (UGC-Autonomous), Hyderabad, India, rrashmi.cigiri@gmail.com.

[2] Student, Department of Information Technology, Mallareddy Engineering College for Women, (UGC-Autonomous), Hyderabad, India.

Abstract

In the world of medicine, drug classification holds immense importance as it helps determine the most suitable drugs for patients based on their unique characteristics and medical history. The dataset containing various features plays a vital role in assessing which drugs are best suited for individuals. This process is known as multi-class drug classification, where drugs are categorized into different classes based on their specific uses and therapeutic effects. Traditionally, drug classification has been carried out through manual or rule-based approaches, where physicians and medical experts rely on their knowledge and experience to prescribe drugs based on patient attributes. However, this method can be time-consuming and may not be efficient when dealing with a large number of drugs and patients. That's where machine learning comes in to revolutionize the process. Machine learning models are capable of analyzing vast amounts of data, learning complex patterns, and making predictions in a more automated and effective manner. The main objective of this project is to develop a machine learning model that accurately classifies drugs into multiple classes using specific features like Age, Sex, Blood Pressure (BP), Cholesterol level, and the Sodium to Potassium ratio. The target variable, in this case, is the "Drug," representing the name or class of the drug. Additionally, this project also involves exploratory data analytics, which focuses on data visualization and representation techniques. By exploring and visualizing the data, we can gain valuable insights and better understand the relationships between the features and drug classes, which will ultimately aid in building a robust and accurate machine learning model for drug classification.

1. Introduction

Drug classification plays a pivotal role in healthcare and pharmaceuticals. Accurate categorization of drugs based on patient information is vital for tailoring treatments to individual needs. In the past, this task heavily relied on expert human judgment, but today, machine learning models offer a promising way to automate and enhance the process. This project taps into the latest advancements in data science and healthcare technology to make drug classification more efficient and effective. Over the years, machine learning has gained prominence in the realm of drug classification. This shift has been driven by the ever-growing volumes of healthcare data and the pressing need for data-driven decision-making [1]. Traditional methods often required manual intervention, but machine learning brings automation and precision to the forefront, making it a game-changer in the healthcare industry. The motivation behind this project stems from the surging amount of healthcare data and the desire for more data-informed healthcare decisions [2]. Automating drug classification can yield significant benefits, including saving time and resources for healthcare professionals, reducing errors, and elevating patient care [3]. Furthermore, machine learning models can unearth intricate data patterns that might elude human analysis, making them invaluable in the quest for improved healthcare. Therefore, this project employed machine learning to address a crucial challenge: classifying drugs based on patient data. The project is not just about crunching numbers; it’s about using technology to make healthcare decisions smarter and more personalized. To do this, we employ a variety of data

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 10, 2023

visualization tools like Seaborn, Plotly, and Matplotlib. These tools help us explore the data and test the performance of three different machine learning models tailored for drug classification. To assess our models' effectiveness, we rigorously evaluate their performance. We use tools like confusion matrices, accuracy scores, and classification reports to gauge their accuracy and reliability in classifying drugs. Ultimately, our goal is to provide valuable insights that empower healthcare professionals to make more informed decisions about drug prescriptions, ultimately enhancing patient care and streamlining drug selection processes.

2. Literature Survey

Harpaz et al. [4], explored the field of text mining for adverse drug events (ADEs). It discusses the potential and challenges associated with utilizing text data, such as medical narratives, for detecting and understanding ADEs. The authors provide an overview of the current state of the art in text mining techniques for ADE detection, highlighting the promise and limitations of these approaches. The study emphasizes the importance of harnessing unstructured textual data to enhance drug safety surveillance. Sarker et al. [5], focused on the use of social media data for pharmacovigilance, particularly in the context of monitoring adverse drug reactions (ADRs). The authors examine the potential of social media platforms as a valuable source of information for ADR detection. They discuss various methodologies and challenges associated with mining social media data for pharmacovigilance purposes, providing insights into the opportunities and limitations of this emerging approach.

Benton et al. [6], introduced a novel approach to hypothesis generation for potential adverse effects of drugs by mining web-based information. The authors propose a methodology for utilizing web data, such as search engine queries and online forums, to identify potential adverse effects associated with specific medications. While the study demonstrates the potential of web mining for hypothesis generation, it acknowledges the need for validation and addresses the limitations of the approach. In [7], Liu and Chen presented Azdrugminer, an information extraction system designed to mine patient-reported adverse drug events from online patient forums. The authors highlight the significance of extracting valuable information from user-generated content and discuss the capabilities and limitations of Azdrugminer in identifying adverse drug events from unstructured forum discussions. Yates and Goharian, focused on Adrtrace, a system for detecting expected and unexpected ADRs from user reviews on social media platforms [8]. The authors emphasize the importance of identifying both expected and unexpected ADRs and present Adrtrace as a solution. They discuss the system's performance and limitations in distinguishing between these types of reactions.

In [9] Turdakov et al., discussed its capabilities in handling various text analysis tasks. While the paper provides an overview of Texterra's features and applications, it also acknowledges that any text analysis framework has inherent limitations in addressing the complexities of natural language understanding. In [10] Doshi-Velez and Kim focused on the interpretability of machine learning models. It discusses the importance of making machine learning models more interpretable and provides insights into the challenges and potential solutions. While the paper emphasizes the need for interpretable models, it also acknowledges that achieving full interpretability is a complex and ongoing research endeavor. Li et al., [11] introduced an evidential decision tree based on belief entropy as a decision-making framework. The paper discussed the theoretical foundation and applications of this approach. However, it also recognizes that decision-making under uncertainty is a challenging task, and the proposed approach has limitations in handling complex scenarios. Gala et al., [12] presented a study on drug classification using machine learning and interpretable models. It discusses the use of interpretable machine learning techniques for drug classification. While the study

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 10, 2023

emphasizes the importance of model interpretability, it also acknowledges that interpretability may come at the cost of model complexity and performance. In [13] Gururaj et al., focused on the classification of drugs based on their mechanism of action using machine learning techniques. It discusses the potential of machine learning in this context. However, it also recognizes that the accuracy of classification models may be influenced by data quality and the complexity of drug mechanisms.

3. Proposed System Model

The primary objective of this project is to develop a machine learning solution for multi-class drug classification. The project employs various data visualization libraries and machine learning tools to explore, preprocess, and classify drugs based on patient data, providing valuable insights for healthcare decision-making. This project contributes to the healthcare domain by offering a data-driven approach to drug classification. By leveraging machine learning and data visualization, it aims to improve the accuracy and efficiency of drug classification, potentially leading to more personalized and effective treatments for patients. The evaluation of multiple models helps identify the most suitable approach for this critical healthcare task.

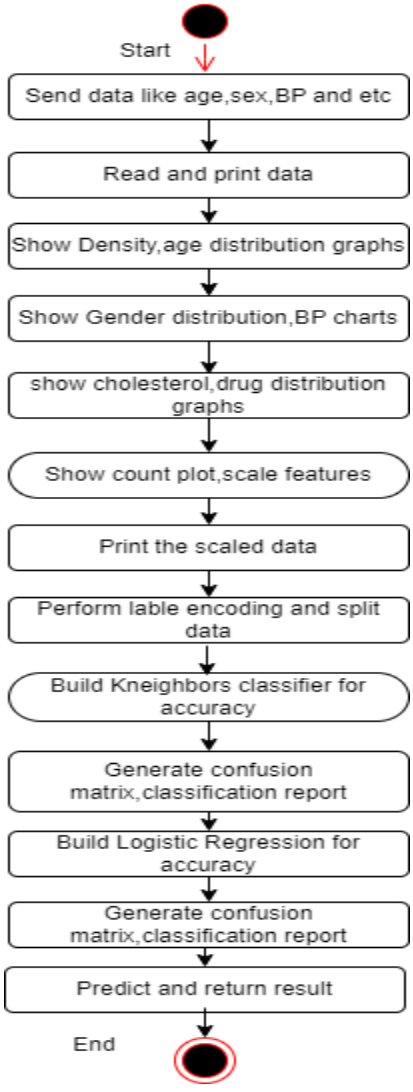


Figure 1: Detailed design of proposed methodology.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 10, 2023

Multilayer perceptron (MLP)

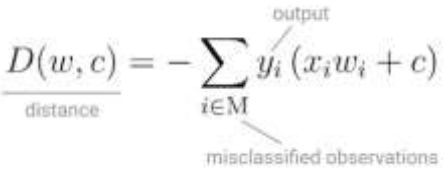
Although today the Perceptron is widely recognized as an algorithm, it was initially intended as an image recognition machine. It gets its name from performing the human-like function of perception, seeing, and recognizing images. In particular, interest has been centered on the idea of a machine which would be capable of conceptualizing inputs impinging directly from the physical environment of light, sound, temperature, etc. — the "phenomenal world" with which we are all familiar — rather than requiring the intervention of a human agent to digest and code the necessary information. Rosenblatt's perceptron machine relied on a basic unit of computation, the neuron. Just like in previous models, each neuron has a cell that receives a series of pairs of inputs and weights. The major difference in Rosenblatt's model is that inputs are combined in a weighted sum and, if the weighted sum exceeds a predefined threshold, the neuron fires and produces an output.

Figure 2. Perceptron neuron model (left) and threshold logic (right).

Threshold *T* represents the activation function. If the weighted sum of the inputs is greater than zero the neuron outputs the value 1, otherwise the output value is zero.

Perceptron for Binary Classification

With this discrete output, controlled by the activation function, the perceptron can be used as a binary classification model, defining a linear decision boundary. It finds the separating hyperplane that minimizes the distance between misclassified points and the decision boundary. The perceptron loss function is defined as below:



To minimize this distance, perceptron uses stochastic gradient descent (SGD) as the optimization function. If the data is linearly separable, it is guaranteed that SGD will converge in a finite number of steps. The last piece that Perceptron needs is the activation function, the function that determines if the neuron will fire or not. Initial Perceptron models used sigmoid function, and just by looking at its shape, it makes a lot of sense! The sigmoid function maps any real input to a value that is either 0 or 1 and encodes a non-linear function. The neuron can receive negative numbers as input, and it will still be able to produce an output that is either 0 or 1.

But, if you look at Deep Learning papers and algorithms from the last decade, you'll see the most of them use the Rectified Linear Unit (ReLU) as the neuron's activation function. The reason why ReLU became more adopted is that it allows better optimization using SGD, more efficient computation and is scale-invariant, meaning, its characteristics are not affected by the scale of the input.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 10, 2023

The neuron receives inputs and picks an initial set of weights random. These are combined in weighted sum and then ReLU, the activation function, determines the value of the output.

Figure 3. Perceptron neuron model (left) and activation function (right).

Perceptron uses SGD to find, or you might say learn, the set of weight that minimizes the distance between the misclassified points and the decision boundary. Once SGD converges, the dataset is separated into two regions by a linear hyperplane. Although it was said the Perceptron could represent any circuit and logic, the biggest criticism was that it couldn't represent the XOR gate, exclusive OR, where the gate only returns 1 if the inputs are different. This was proved almost a decade later and highlights the fact that Perceptron, with only one neuron, can't be applied to non-linear data.

4. Results description

Our project highlights the potential of machine learning, especially the MLPs, in tackling the intricate multi-class drug classification challenge. By amalgamating data exploration, visualization, preprocessing, and model evaluation, we have laid the foundation for more precise drug categorization, a development poised to significantly benefit healthcare practitioners and patients as shown in Table 1.

Table 1. Overall performance comparison of existing KNN, and LR, and proposed MLP classifier models.

Model	Accuracy	Precision	Recall	F1-score
KNN classifier	0.90	0.92	0.90	0.90
LR model	0.95	0.95	0.95	0.95
MLP classifier	0.975	0.98	0.97	0.97

Figure 4 consists of three confusion matrices, each obtained using a different classifier model (K-Nearest Neighbors, Logistic Regression, and Multi-Layer Perceptron). These matrices illustrate the performance of each model in drug classification. Rows in the matrix represent actual classes, columns represent predicted classes, and the values in the cells show the number of samples classified correctly (diagonal) and incorrectly (off-diagonal). These matrices are essential for evaluating the models' accuracy and performance in predicting drug categories.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 10, 2023

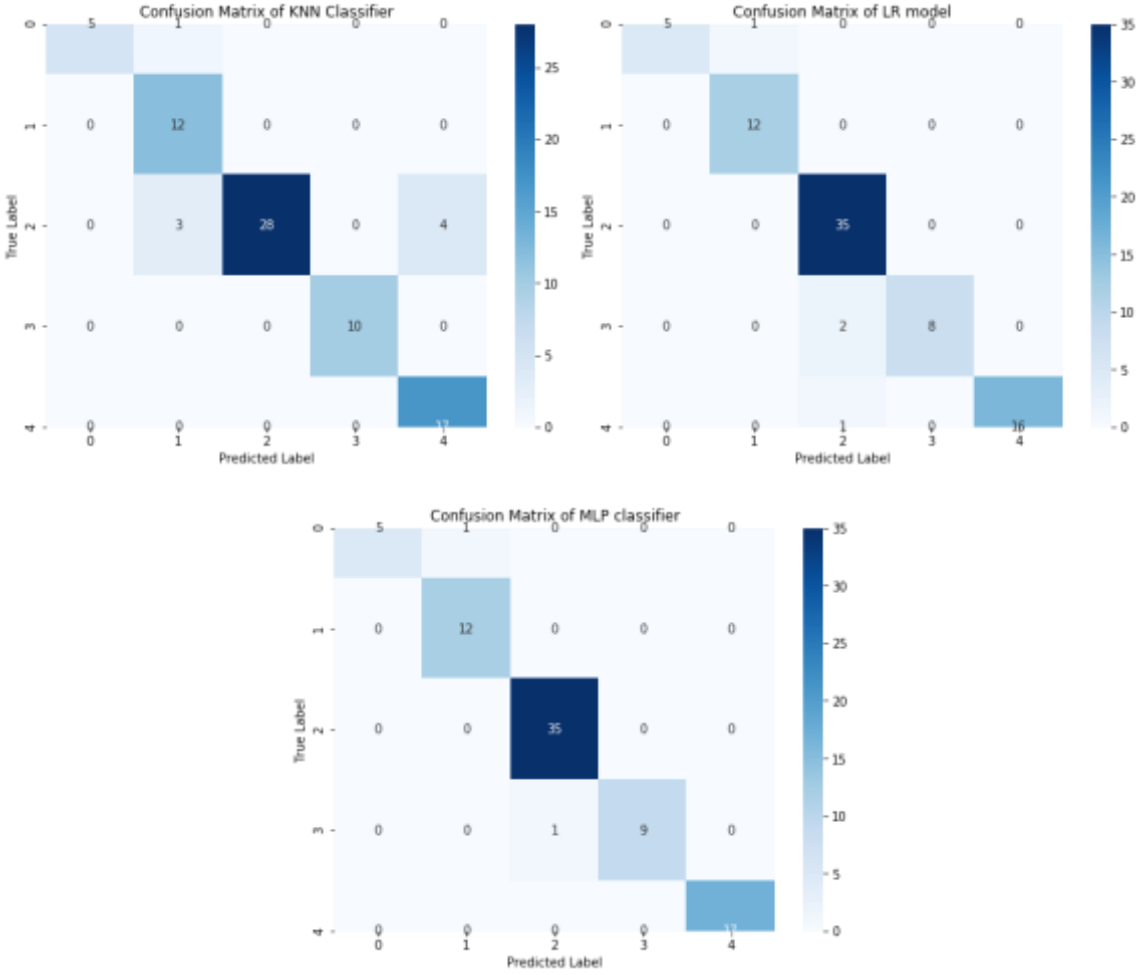


Figure 4: Confusion matrices obtained using KNN, LR, and MLP classifier models for drug classification system.

5. Conclusion

This project implemented a multi-class drug classification, harnessing the power of machine learning techniques and data visualization tools. Our primary goal was to enhance drug categorization accuracy, ultimately benefitting both healthcare professionals and patients. Initially, we delved into the dataset, conducting extensive data exploration to grasp its structure, characteristics, and inherent distributions. This preliminary analysis formed the basis for subsequent investigations. To provide intuitive insights, we harnessed data visualization libraries such as Seaborn, Plotly, and Matplotlib. These visualizations offered clear and accessible interpretations of the data, ranging from age distributions to drug category counts. To prepare our data for machine learning, we undertook essential data preprocessing steps. This involved standardizing numeric features and encoding categorical variables, ensuring that our models could effectively process and learn from the data. Then, this project employed three distinct machine learning models: the K-Nearest Neighbors (KNN) Classifier, the Logistic Regression (LR) Classifier, and the Multi-Layer Perceptron (MLP) Classifier. These models presented varying levels of complexity and were thoroughly evaluated to gauge their suitability for the multi-class drug classification task.

ISSN: 0975-3583, 0976-2833 VOL14, ISSUE 10, 2023

References

[1] L. Medina-Franco, M. A. Giulianotti, G. S. Welmaker and R. A. Houghten, "Shifting from the single to the multitarget paradigm in drug discovery", Drug discovery today, vol. 18, no. 9, pp. 495-501, 2016.

[2] H.-M. Lee and Y. Kim, "Drug repurposing is a new opportunity for developing drugs against neuropsychiatric disorders”", Schizophrenia research and treatment, vol. 20, 2016.

[3] R. Sloane, O. Osanlou, D. Lewis, D. Bollegala, S. Maskell and M. Pirmohamed, "Social media and pharmacovigilance: a review of the opportunities and challenges", British journal of clinical pharmacology, vol. 80, no. 4, pp. 910-920, 2015.

[4] R. Harpaz, A. Callahan, S. Tamang, Y. Low, D. Odgers, S. Finlayson, et al., "Text mining for adverse drug events: the promise challenges and state of the art", Drug safety, vol. 37, no. 10, pp. 777-790, 2018.

[5] A. Sarker, R. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, et al., "Utilizing social media data for macovigilance: A review", Journal of biomedical informatics, vol. 54, pp. 202-212, 2017.

[6] Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, et al., "Identifying potential adverse effects using the web: A new approach to medical hypothesis generation", Journal of biomedical informatics, vol. 44, no. 6, pp. 989-990.

[7] X. Liu and H. Chen, "Azdrugminer: an information extraction system for mining patient-reported adverse drug events in online patient forums", In International Conference on Smart Health, 2017.

[8] A. Yates and N. Goharian, "Adrtrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites", In European Conference on Information Retrieval, 2019.

[9] D. Y. Turdakov, N. A. Astrakhantsev and Y. R. Nedumov, "Texterra: A framework for text analysis", Programming and Computer Software, vol. 40, no. 5, pp. 288-295, 2018.

[10] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning", arXiv: Machine Learning, 2017.

[11] M. Li, H. Xu and Y. Deng, "Evidential Decision Tree Based on Belief Entropy", School of Computer Science and Engineering University of Electronic Science and Technology of China, 2019.

[12] D. V. Gala, V. B. Gandhi, V. A. Gandhi and V. Sawant, "Drug Classification using Machine Learning and Interpretability," 2021 Smart Technologies, Communication and Robotics (STCR), Sathyamangalam, India, 2021, pp. 1-8, doi: 10.1109/STCR51658.2021.9588972.

[13] Gururaj, H.L., Flammini, F., Kumari, H.A.C. et al. Classification of drugs based on mechanism of action using machine learning techniques. Discov Artif Intell 1, 13 (2021). https://doi.org/10.1007/s44163-021-00012-2.