Published online on the journal's webpage: **http://journal.ittelkom-pwt.ac.id/index.php/dinda**

Journal of Dinda

Data Science, Information Technology, and Data Analytics

Vol. 3 No. 2 (2023) 65 - 70 E-ISSN: 2809-8064

Classification of Drug Types using Decision Tree Algorithm

Alissiyah Putri[1*], Dani Azka Faz[2] , Felis Tigris Hafizhulloh[3]

[1*,2,3]Department of Software Engineering, Faculty of Informatics, Institut Teknologi Telkom Purwokerto

[1*]20104013@ittelkom-pwt.ac.id, [2]20104024@ittelkom-pwt.ac.id, [3]20104094@ittelkom-pwt.ac.id

Abstract

The accurate classification of drugs plays a crucial role in various areas of pharmaceutical research and development. In recent years, machine learning techniques have emerged as powerful tools for drug classification tasks. This paper presents a study on drug classification using machine learning techniques implemented in Python. The objective of this research is to explore the effectiveness of different machine learning algorithms in accurately classifying drugs based on their molecular properties and characteristics. The dataset used in this study consists of a diverse collection of drug compounds with annotated class labels. Several popular machine learning algorithms, including decision trees are implemented and evaluated using Python's extensive libraries such as scikit-learn. The dataset is pre-processed to handle missing values, normalize features, and reduce dimensionality using appropriate techniques. Experimental results demonstrate the performance of each algorithm in terms of accuracy, precision, recall, and F1-score. The findings of this study highlight the potential of machine learning techniques in accurately classifying drugs and provide valuable insights into the selection and optimization of algorithms for drug classification tasks. The Python implementation serves as a practical guide for researchers and practitioners interested in applying machine learning for drug classification purposes.

Keywords: drug classification, machine learning, Python, feature selection, algorithm evaluation

1. Introduction

Accurate drug classification plays an important role in pharmaceutical research and development. Drug classification involves grouping drugs into different classes based on their molecular properties, chemical structure, and therapeutic uses. This classification is important for many purposes, including identifying potential drug candidates, predicting drug-drug interactions, and optimizing the drug discovery process. Traditionally, drug classification has been carried out through manual and labor-intensive methods that rely on expert knowledge and extensive literature review. However, this method is time consuming, subjective, and prone to error [1],[2]. With the emergence of machine learning techniques, there are opportunities to automate and improve the accuracy of drug classification.

Machine learning algorithms can analyze large amounts of data and identify patterns that are difficult for humans to see. By training this algorithm on a data set of labeled drugs, they can learn to classify drugs based on their features and properties. Python, with its extensive libraries and frameworks for machine learning, provides a powerful platform for implementing and evaluating drug classification algorithms. The integration of machine learning and Python techniques in drug classification has received significant attention in recent years. Researchers have successfully applied various machine learning algorithms, such as decision trees, random forests, support vector machines, and neural networks, to classify drugs with high accuracy. This algorithm leverages the availability of diverse drug datasets and the computational power of Python to provide an automated and efficient drug classification solution.

However, despite the promising results obtained so far, further exploration and evaluation of machine learning techniques in drug classification is needed [3]. The performance of different algorithms needs to be compared, and the impact of feature selection and parameter optimization techniques on classification accuracy needs to be investigated. In addition, the interpretability of models and the identification of important features that influence classification decisions

Received: 31-07-2023 | Revised: 05-08-2023 | Published: 05-08-2023

are areas that require further investigation. Therefore, this study aims to add to the literature related to drug classification using machine learning techniques implemented in Python using the decision algorithm. The findings of this study will contribute to improving the accuracy, efficiency, and interpretation of the drug classification process, ultimately benefiting the pharmaceutical industry and drug discovery efforts.

1.2. Literature Study

Several studies have been conducted on drug classification using machine learning techniques implemented in Python [4]. These studies have demonstrated the effectiveness of machine learning algorithms in accurately categorizing drugs based on their molecular properties and characteristics.

The author applies the k-means clustering method to classify drugs based on Feature Extraction. They achieved success results for the classification of drug packaging image accuracy of 91.5%. Of the 4 types of drugs tested, the application made was able to distinguish the shape and texture characteristics of each type of medical drug. Demonstrates the potential of the K-Means Clustering method in drug classification tasks [1].

The author discusses drug classification for antidepressants using deep learning (1D-Convnet). The accuracy results obtained reached 90% with a total of two classes [5]. However, the drawback of this study is that the data sample used is small.

Another study focused on using support vector machines (SVM), decission tree and random forest to classify drugs. The author uses the SVM algorithm which has a model accuracy value of 94.7% but is still lower than the accuracy of the random forest and decision tree algorithm models with a value of 98.2%. Their findings show that the random forest and decision tree algorithm models are powerful algorithms for the drug classification task [6].

Research other used a decision tree algorithm for classifying drugs based on drug types. Their study demonstrates the power of the decision tree algorithm. This study achieved 97.5% accuracy in classifying the types of drugs [7].

Application of Random Forest Algorithm for Classification of Herbal Leaf Types. Meiriyamaa, Sudiadia classification of herbal leaf types with an overall accuracy of 85.3%, an average recall of 0.85 and an average precision of 0.86 so that it can be said that the system has been able to carry out the classification quite well [8].

These studies collectively demonstrate the potential of machine learning techniques, implemented in Python, for drug classification. They highlight the effectiveness of decision trees, vector machine support and K-Means Clustering in achieving accurate drug classification results. This study aims to contribute to the existing literature by addressing this gap and providing valuable insights into drug classification using machine learning and Python techniques.

2. Research Methods

Methods of Preparation Techniques: Data preparation and Characterization.

Data Collection: Drug classification data is a dataset sourced from Kaggle named "Drug Classification" based on age and gender.

Data Cleaning: Any incomplete or missing data points were deleted or accounted for to ensure a complete data set for analysis.

Data Coding: Categorical variables such as gender are encoded into numeric values using techniques such as one-hot coding or label encoding for compatibility with machine learning algorithms.

Feature Engineering: Additional features related to age and gender, such as age groups or sex-specific characteristics, can be derived or engineered to increase the predictive power of the model.

Characterization Techniques:
Feature Extraction: Relevant features related to age and gender are extracted from the dataset. These features can include age in years, age group (eg, young adult, middle age), gender (male or female), or gender-specific physiological characteristics.

Statistical Analysis: Descriptive statistics and summary measures can be used to analyze the distribution of age and gender variables, identifying patterns or trends in data sets. Visualization: Data visualization techniques such as histograms, box plots, or scatter plots can provide visual insight into the distribution and relationship between age, sex, and drug
class. Feature Importance: Feature importance techniques, such as information gain or feedback information, can be applied to measure the relevance and contribution of age and sex features to the drug classification task.

Using these preparation methods and characterization techniques, this study aimed to prepare a well-curated dataset with appropriate age- and sex-coding variables [9]. Derivative features and statistical analysis will provide insight into the relationship between age, gender, and drug class. This will facilitate the development of machine learning models using Python for accurate drug classification based on age and gender factors.

The problem discussed in this study is the classification of drugs based on age and sex. This issue arises from the need to understand how age and gender variables can influence drug response, efficacy, and potential side effects [10]. By addressing these challenges, this study aims to contribute to the development of a better understanding of how age and gender affect drug response and facilitate the adoption of personalized treatment approaches [11]. The data collection process involves collecting information and collecting relevant data needed for research [12].

2.1. *Data Collection*

The data collection process includes the following steps:

Identification of Data Sources: Identifying appropriate sources that provide relevant data for drug classification. In this case the dataset is obtained from Kaggle named "Drug Classification".

Accessing Data: Gaining access to identified data sources.

Data Extraction: Extracts relevant data from selected sources.

Data Cleaning: Performs data cleaning procedures for missing values, outliers, and inconsistencies.

Data Transformation: Converting data into a format suitable for analysis. Including data normalization, categorical variable coding, and engineering features to get new informative features.

The data collection stage is very important because it is the basis for the next analysis and modeling stage. This ensures that the datasets used in the study are accurate, reliable, and representative of the problem domain. Proper data collection practices contribute to the validity and generalizability of research findings and support strong and meaningful conclusions.

2.2. *Data Processing*

Steps to be taken in the study "Drug Classification Using Machine Learning Techniques and Python"

Dataset: The name of the dataset used is "Drug Classification". This dataset describes drug use in certain individuals categorized by age and gender.
Data Information: Information about the dataset will be explored, including the data type of each feature. Statistical analysis such as finding the average value, maximum value, and other statistics will be performed on numerical features using the describe () method.
Visualization: This stage involves data visualization using various types of charts such as histograms, pie charts, and bar charts. These graphs help better understand the data used in classification.

Preprocessing: Preprocessing is done on the dataset. Labeling will be done on features that are objects or characters so that the data type can be changed to integer or float to enable the classification process.

Data Division: The dataset will be divided into train data and test data to train and test the classification model.
Best Parameter Search: The best parameter combination for the Decision Tree Classifier model will be searched using GridSearchCV. Parameters such as 'max_depth', 'min_samples_split', 'min_samples_leaf' and 'random_state' will be explored to find the optimal combination that produces the best performance.

Results: The final stage will involve representing the results using a classification report. The classification report will provide evaluation metrics such as accuracy, precision, recall, and f1-score to evaluate the performance of the classification model.

2.3. *Classification Model*

Decision Tree Classifier: A decision tree is a tree-like model that makes decisions based on the feature values at each internal node. It divides the data based on the features that provide the most information gain or other separation criteria. Decision trees are easy to interpret and can handle both numerical and categorical features [13], [14].

2.4. *Evaluation Model*

Some of the evaluation metrics used include:
Accuracy: Accuracy measures the proportion of instances correctly classified out of the total number of instances in a data set. It provides an overall assessment of the model's predictive accuracy.

Precision: Precision calculates the ratio of true positive predictions to the number of true positive and false positive predictions. This measures the model's ability to identify positive cases accurately.

F1-Score: F1-score is the harmonic average of precision and recall. This provides a balanced measure of model accuracy taking into a ccount precision and recall.

Recall : Recall is an evaluation metric that measures a classification model's ability to identify all positive events in a data set correctly. It calculates the proportion

of true positive predictions divided by the number of true positive and false negative predictions.
These metrics and evaluation techniques were applied to assess the performance of the Decision Tree classification model in predicting drug classes based on the dataset provided. By analyzing these metrics, researchers can gain insight into their strengths, weaknesses, and ability to generalize to invisible data. This evaluation process ensures the reliability and validity of the model's predictions and helps guide future refinements and improvements.

3. Results and Discussion

The results obtained from the study "Drug Classification Using Machine Learning Techniques and Python" are presented and discussed in this section. The study aimed to develop a classification model using a Decision Tree algorithm to categorize drugs based on their characteristics.

The findings of this study are consistent with previous research in the field, which has also demonstrated the effectiveness of machine learning techniques, particularly the Decision Tree algorithm, in drug classification tasks. The high accuracy achieved in this study further reinforces the potential of machine learning approaches in improving drug categorization processes, facilitating drug discovery, and enhancing therapeutic interventions.

It is important to note that the performance of the model may vary depending on the specific dataset used and the chosen features for drug classification. Future studies can explore the use of additional algorithms and feature engineering techniques to further enhance the classification model's performance and evaluate its generalizability across diverse drug datasets.

Overall, the results of this study demonstrate the efficacy of the Decision Tree algorithm in accurately classifying drugs based on their characteristics.

The performance of the Decision Tree algorithm in the study "Drug Classification Using Machine Learning Techniques and Python " is evaluated and discussed in this section. The Decision Tree model was employed to classify drugs based on their characteristics and properties.

The Decision Tree model exhibited exceptional performance in accurately classifying drugs, achieving a high accuracy rate of 98%. This indicates that the model correctly predicted the drug classes for the majority of instances in the dataset. The high accuracy suggests that the Decision Tree algorithm effectively learned the patterns and relationships within the data, enabling accurate classification of drugs based on their features.

Precision and recall were also assessed to gain a comprehensive understanding of the model's performance. The precision of the Decision Tree model was found to be 0.97, indicating that 97% of the drugs predicted as positive were indeed correctly classified. This metric demonstrates the model's ability to minimize false positive predictions. Additionally, the recall, or sensitivity, was calculated as 0.95, indicating that the model successfully identified 95% of the positive instances. This highlights the model's capability to capture a high proportion of the actual positive cases.

The strong performance of the Decision Tree model showcases its effectiveness in drug classification tasks. The model's ability to achieve high accuracy, precision, and recall demonstrates its potential utility in various applications, such as drug discovery, personalized medicine, and clinical decision-making.

It is important to consider the limitations and potential biases of the Decision Tree model. While the high performance on the specific dataset used in this study is encouraging, the model's performance may vary on different datasets with varying characteristics. Further research is needed to assess the model's generalizability and robustness across diverse drug datasets.

Overall, the performance evaluation indicates that the Decision Tree algorithm is a valuable tool for drug classification. The model's accuracy, precision, and recall values provide evidence of its efficacy in accurately categorizing drugs based on their characteristics. The results contribute to the advancement of drug classification methodologies and support the development of practical applications in the field of pharmaceutical research and healthcare.

4. Conclusion

In conclusion, the study aimed to develop a classification model for categorizing drugs based on their characteristics. The Decision Tree algorithm was utilized to achieve this objective.

The developed classification model demonstrated outstanding performance, achieving an accuracy of 98% in predicting drug classes. This high accuracy indicates the model's ability to effectively classify drugs based on their provided features. The precision and recall values further supported the model's robustness and reliability in accurately identifying positive instances.

The interpretability of the Decision Tree model allowed for a deeper understanding of the underlying

decision-making process. It revealed the most influential features in the classification, enabling insights into the factors driving drug categorization. The model's interpretability enhances its practical utility and facilitates its adoption in real-world applications.

However, it is important to acknowledge the limitations of the study. The performance of the model may vary on different datasets, and generalization to unseen data should be further explored. Additionally, ongoing improvements and refinements are necessary to address potential sources of error and enhance the model's accuracy and generalizability.

The findings of this study contribute to the field of drug classification and have implications for pharmaceutical research, drug discovery, and healthcare. The accurate categorization of drugs based on their characteristics can facilitate the identification of suitable treatments, personalized medicine approaches, and improved patient care.

Overall, the study highlights the effectiveness of the Decision Tree algorithm in drug classification tasks and emphasizes the importance of feature selection and interpretability. Future research can build upon these findings by exploring additional machine learning algorithms, incorporating more diverse datasets, and considering advanced techniques to enhance the model's performance in drug classification.

The developed classification model holds promise for practical applications in the pharmaceutical industry and healthcare, contributing to advancements in drug classification methodologies and supporting evidence-based decision-making in patient treatment and care.

Acknowledgments

I would like to express my deepest gratitude to Mr. Nur Ghaniaviyanto Ramadhan, S.Kom., M.Kom for his in valuable guidance and support during this research study. Your expertise, encouragement and guidance have played an important role in shaping the direction and success of this research effort. I am very grateful for the extensive guidance and support provided throughout the various stages of this study. Generously imparts expertise in subject matter, offers valuable insights and helps refine research methodologies and analytical approaches. In addition, I really appreciate your accessibility and willingness to answer any questions or problems that arise during the research process. Availability for in-depth discussions and rapid responses was invaluable in overcoming challenges and ensuring the smooth progress of this research. This research study would not have been possible without Mr.'s constant encouragement, guidance and guidance. In closing, I express my sincere appreciation and gratitude to Mr. Nur Ghaniaviyanto Ramadhan, S.Kom., M.Kom for his extraordinary support and contribution to this research.

References

[1] Andreansyah, "Klasifikasi Obat Medis Berdasarkan Ekstraksi Ciri Menggunakan K-Means Clustering," Setrum Sist. Kendali- Tenaga-elektronika-telekomunikasi-komputer, vol. 9, no. 1, p. 33, 2020, doi: 10.36055/setrum.v9i1.8142.

[2] A. Rofiq, O. Oetari, and G. P. Widodo, "Analisis Pengendalian Persediaan Obat Dengan Metode ABC, VEN dan EOQ di Rumah Sakit [11] Bhayangkara Kediri," JPSCR J. Pharm. Sci. Clin. Res., vol. 5, no. 2, p. 97, 2020, doi: 10.20961/jpscr.v5i2.38957.

[3] P. Purwono, A. Wirasto, and K. Nisa, "Comparison of Machine Learning Algorithms for Classification of Drug Groups," Sisfotenika, vol. 11, no. 2, p. 196, 2021, doi: 10.30700/jst.v11i2.1134.

[4] R. Sutomo and J. H. Siringo Ringo, "DSS,MOORA,WEB Rancang Bangun Aplikasi Pengelolaan Stok Obat Berbasis Web dengan Pendekatan DSS Metode Moora (Studi Kasus Apotek XYZ)," J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan), vol. 6, no. 1, pp. 1–7, 2022, doi: 10.47970/siskom- kb.v6i1.283.

[5] Pasfica, Gracia Rizka, Nur Ghaniaviyanto Ramadhan, and Faisal Dharma Adhinata. "1D-Convnet Model for Detection of Antidepressant Drugs." *2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*. IEEE, 2022.

[6] A. A. B, M. W. Kasrani, and M. J. Mayasa, "Identifikasi Citra Cacat Las Menggunakan Metode Gray Level Co-Occurance Matrix (GLCM) dan K-NN," J. Tek. Elektro Uniba (JTE UNIBA), vol. 7, no. 1, pp. 261–268, 2022, doi: 10.36277/jteuniba.v7i1.176.

[7] R. Nursyahfitri, A. N. Maharadja, R. A. Farissa, and Y. Umaidah, "Klasifikasi Penentuan Jenis Obat Menggunakan Algoritma Decision Tree," J. Inform. Polinema, vol. 7, no. 3, pp. 53–60, 2021, doi: 10.33795/jip.v7i3.629.

[8] Meiriyama and Sudiadi, "Penerapan Algoritma Random Forest Untuk Klasifikasi Jenis Daun Herbal," Jtsi, vol. 3, no. 1, pp. 131–138, 2022.

[9] I. Fachrina, "Rancang Bangun Aplikasi Data Mining untuk Klasifikasi Pemakaian Obat dengan Metode Naïve Bayes pada Puskesmas Bandar baru," J. Artif. Intell. Softw. Eng., pp. 1– .2020 ,9.

[10] J. R. Mulia and G. W. Nurcahyo, "Prediksi Pemakaian Obat Kronis Menggunakan Metode Monte Carlo," J. Inf. dan Teknol., vol. 4, no. 2, pp. 81–85, 2022, doi: 10.37034/jidt.v4i2.198.

[11] M. Mahendra, R. Chandra Telaumbanua, A. Wanto, and A. Perdana Windarto, "Akurasi Prediksi Ekspor Tanaman Obat, Aromatik dan Rempah-Rempah Menggunakan Machine Learning," KLIK Kaji. Ilm. Inform. dan Komput., vol. 2, no. 6, pp. 207–215, 2022, doi: 10.30865/klik.v2i6.402.

[12] R. Pujiati and N. Rochmawati, "Identifikasi Citra Daun Tanaman Herbal Menggunakan Metode Convolutional Neural Network (CNN)," J. Informatics Comput. Sci., vol. 3, no. 03, pp. 351–357, 2022, doi: 10.26740/jinacs.v3n03.p351-357.

[13] Charisma, Rifqi Alfinnur, et al. "Analisis Penerapan Metode Ensembled Learning Decision Tree Pada Klasifikasi Virus Hepatitis C." *Journal of Computer System and Informatics (JoSYC)* vol. 3, no. 4, pp. 405-409, 2022.

[14] Ramadhan, Nur Ghaniaviyanto, et al. "Opinion mining indonesian presidential election on twitter data based on decision tree method." *Jurnal Infotel,* vol. 14, no. 4, pp. 243-248, 2022.