Mobile-Based Intelligent Drug Search Engine Using Machine Learning

Bata, Lydia[1], Dr. Amadu Asabe[2] , Bala P. Bulus[3]

[1]Modibbo Adama University, Yola, Nigeria

(Email: lydiaishaya@gmailcom)

[2]Modibbo Adama University, Yola, Nigeria, asabe@mau.edu.ng)

[3]Federal College of Education, Yola, Nigeria

Corresponding Author: bpbala@fceyola.edu.ng)

Abstract

Hospitals and clinics serve as primary locations for medical services, where doctors often recommend specific medications for patients. However, patients sometimes stru,ggle to obtain high-quality medication due to a lack of knowledge about where to find it. This study aimed to develop a mobile-based intelligent drug search engine using a decision tree technique to predict the availability of drugs and pharmacies. The dataset comprised 600 records from six pharmacies in the Yola North Local Government, Adamawa State. After preprocessing and cleaning the data, the data were split into training (80%) and testing (20%) sets. A decision tree model was built using the Python Jupyter Notebook to predict the availability of malaria drugs and pharmacy locations. The model achieved high prediction accuracies of 94% and 98% on the training and test datasets, respectively, with precision and recall scores of 98%. Additionally, 5-fold cross-validation yielded a mean accuracy of approximately 95.57%. This study demonstrated the feasibility of using supervised learning to predict the availability of malaria and typhoid drugs, providing valuable insights for both pharmacies and clients. It is recommended that pharmacies leverage this ML model to ensure the optimal availability of medicines, thereby addressing the problem of drug accessibility. Individuals can utilize the mobile application to quickly identify pharmacies with the required drugs, potentially saving lives in critical situations.

Keywords: Mobile devices, data preprocessing Machine learning technique, and Mobile applications

Moteur de recherche de médicaments intelligents sur mobile utilisant l'apprentissage automatique

Resume

Les hôpitaux et les cliniques servent de lieux principaux pour les services médicaux, où les médecins recommandent souvent des médicaments spécifiques pour les patients. Cependant, les patients ont parfois du mal à obtenir des médicaments de haute qualité en raison d'un manque de connaissances sur l'endroit où le trouver. Cette étude visait à développer un moteur de recherche de médicaments intelligent sur mobile à l'aide d'une technique d'arbre de décision pour prédire la disponibilité des médicaments et des pharmacies. L'ensemble de données comprenait 600 dossiers de six pharmacies du gouvernement local de Yola North, dans l'État *d'Adamawa. Après le prétraitement et le nettoyage des données, les données ont été divisées en entraînement (80%) et en tests (20%). Un modèle d'arbre de décision a été construit à l'aide du cahier Python Jupyter pour prédire la disponibilité de médicaments du paludisme et de lieux de pharmacie. Le modèle a atteint des précisions de prédiction élevées de 94% et 98% sur les ensembles de données de formation et de test, respectivement, avec des scores de précision et de rappel de 98%. De plus, la validation croisée 5 fois a donné une précision moyenne d'environ 95,57%. Cette étude a démontré la faisabilité de l'utilisation d'apprentissage supervisé pour prédire la disponibilité du paludisme et des médicaments typhoïdes, fournissant des informations précieuses pour les pharmacies et les clients. Il est recommandé de tirer parti de ce modèle ML pour assurer la disponibilité optimale des médicaments, abordant ainsi le problème de l'accessibilité des médicaments. Les individus peuvent utiliser l'application mobile pour identifier rapidement les pharmacies avec les médicaments requis, potentiellement sauver des vies dans des situations critiques.*

Mots-clés: Appareils mobiles, technique d'apprentissage automatique de prétraitement des données, et applications mobiles

المستشفيات والعيادات بمثابة مواقع رئيسية للخدمات الطبية حيث يوصي الأطباء في كثير من الأحيان بأدوية محددة للمرضى.ومع ذلك، يكافح المرضى
أحياناً للحصول على أدوية عالية الجودة بسبب عدم المعرفة حول مكان العثور عليه . تهدف هذه الدراسة إلى تطوير محرك بحث ذكي عن الأدوية قائم على
الهاتف المحمول باستخدام تقنية شجرة القرار للتنبؤ بتوافر الأدوية والصيدليات. تضمنت مجموعة البيانات 600 سجل من ست صيدليات في حكومة يولا
الشمالية المحلية، ولاية أداماوا بعد المعالجة المسبقة وتنظيف البيانات، تم تقسيم البيانات إلى مجموعات تدريب (٪80) واختبار (٪20). تم بناء نموذج شجرة
للتنبؤ بتوافر عقاقير الملاريا ومواقع الصيدلة. حقق النموذج دقة تنبؤ عالية بنسبة ٪94 و ٪98 على Python Jupyter القرار باستخدام دفتر
مجموعات بيانات التدريب والاختبار بدقة واستدعاء درجات ٪98. بالإضافة إلى ذلك، أسفر التحقق المتقاطع 5 أضعاف عن متوسط دقة يبلغ حوالي
٪95.57 أظهرت هذه الدراسة جدوى استخدام التعلم الخاضع للإشراف للتنبؤ بتوافر عقاقير الملاريا والتيفوئيد، مما يوفر رؤى قيمة لكل من الصيدليات
هذا لضمان توافر الأدوية، وبالتالي معالجة مشكلة إمكانية الحصول على الأدوية يمكن للأفراد ML والعملاء. يوصى بأن تستفيد الصيدليات من نموذج
استخدام تطبيق الهاتف المحمول لتحديد الصيدليات بسرعة بالعقاقير المطلوبة، مما قد ينقذ الأرواح في المواقف الحرجة

Introduction

In Nigeria, one in seven persons aged 15-64 years had used a drug (other than tobacco or alcohol) in the past year. The past year's prevalence of any drug use is estimated at 14.4% (range 14% – 15%), corresponding to 14.3 million people aged 15-64 years (United Nations Office On Drugs and Crime ,2019).

Hospitals and clinics are the most official places where medical services are provided. Patients received medical counselling and treatment via ailments from doctors in these places. Diagnosis is carried out on a patient with an ailment, and prescriptions and recommendations for appropriate medications are ultimately made. Sometimes doctors recommend the use of certain company products for patients. Patients try all means to see that they get drugs recommended by experts, but sometimes they end up buying low-quality medication; according to Devarajan & Das (2012), lack of access to quality medicines is one of the constraints to poor people's health in Africa.

Patients had different visit units/times inside the hospital. However, many outpatients visit the pharmacy inside the hospital for their prescriptions. These patients leave the doctors' facilities and any of the different units on different occasions, subsequently establishing an arbitrary entry rate at the pharmacy where the dispensing activity occurs (Almeziny, 2017).

Pharmacy and dispensaries within hospitals and clinics dispense medications as prescribed by doctors. Medications are less expensive in pharmacies in government hospitals than in private hospitals (Ohinmaa et al., 2016). However, certain drugs are sometimes not found in such medical stores, which require them to go outside. Information on where to obtain the right medicine outside hospitals or clinics has become a concern. Poor-quality medicine is one of the obstacles to improving health in developing countries (Idowu, 2017) . As a result, patients purchase substandard medications from nonpharmacy outlets. The nonpharmacy trade concerns the sale of medicinal products outside of pharmacies, such as limited-service pharmacies, supermarkets, petrol stations, and shops open to the public (Oleszkiewicz, et al .,2021).

According to the World Health Organization (WHO), one in 10 medicines may not meet acceptable standards. Access to medications through nonpharmacy outlets is observed. Increasing the availability of medicines by allowing patients to obtain them outside of pharmacies contributes to the spread of self-medication.

Medications at public hospitals and clinics are less expensive than those at private clinics and hospitals; however, good and high-quality medications are generally not always available in either public or private clinics. The lack of availability of medication leads to patients going out to receive prescribed medication without knowledge of the right pharmaceutical store to visit, leading to the purchase of substandard medication that threatens life. Some people lost their loved ones as a result of not being able to receive the right medications at the right time.

Only approximately 35% of children in Nigeria receive treatment at government health facilities (National Population Commission Nigeria, 2013). The rest are treated by patent medicine vendors (PMVs), pharmacy shops, private health facilities, or traditional medicine practitioners (Oguonu, & Edelu, 2016). Some studies have linked poor adherence to an increase in malaria-related mortality and increased drug costs as underlying factors (Obiebi , 2019). The drawback of this practice, for example, is the progression to severe malaria and its consequences, which include prolonged hospitalization, increased treatment costs, avoidable blood transfusions, and even death. These problems can be minimized significantly by providing online mobile-based pharmaceutical services.

This study aimed to develop a mobile-based intelligent drug prediction model for accessing medication from pharmacies. This is imperative because malaria is a common infectious disease caused by parasitic protozoa of Plasmodium that infect, feed on, and destroy humans. Recently, the parasite has developed resistance to drugs, resulting in the production of various drug combinations by pharmaceutical companies and hence the proliferation of many malaria drugs and increased prices. Thus, there is a need for intelligent drug prediction models.

The specific objectives of the study are to pre-process and transform the data into a suitable format for building a model. A model was built using the training dataset, and the model performance was evaluated using the test dataset in the Python programming language. The decision tree algorithm was applied for the classification and prediction of the best drugs. The model performance was evaluated using K-fold cross-validation.

Overview of Mobile Devices

A mobile device is a portable handheld computer device designed to be carried in hand. Aungst (2013) asserted that the use of mobile devices by healthcare professionals (HCPs) has transformed many aspects of clinical practice. This device has developed from being a handset that was initially used to make calls and send short text messages (SMS) to smartphones with functionality and speed almost equivalent to that of personal computers and has become ubiquitous among the public.

Smartphones are information and communication technology (ICT) tools that are widely adopted by healthcare professionals. Today, most medical sciences universities have provided smartphones as educational aid tools and acquisition licenses for medical app resources in the training of their students (Jebraeily et al., 2017). The use of smartphones has increased in recent years, and they have been widely adopted by health specialists, especially among medical students. A smartphone is a tool that has an operating system and can be used in various applications. Mobile devices are used in every aspect of human endeavour. Additionally, according to Wallace, S. et al.,(2012), mobile devices have become commonplace in healthcare settings, leading to rapid growth in the development of medical software applications (apps) for these platforms. Numerous apps are now available to assist HCPs with many important tasks, such as information and time management, health record maintenance and access, communications and consulting, reference and information gathering, patient management and monitoring, clinical decision-making, and medical education and training(Divali, et al., *2013).*

Applications (ap)

Medical knowledge is rapidly expanding and updating. Timely access to information and the latest scientific evidence without any time or place limitations is very important. Popularly called an app is a type of application software designed to run on a mobile device, such as a smartphone or tablet computer (Onyekachukwu, 2022). These software programs have been developed to run on a computer or mobile device to accomplish a specific purpose (Application Software for Personal, Enterprise & Workgroup Objectives, 2013, October 18).
Many mobile apps are not intended to replace desktop applications but are meant to complement them to provide a resource that has the potential to improve outcomes at the point of care. From the internet to email, they offer on-the-go access to information never before possible (Payne, et al.,2012).

The rapid integration of mobile devices into clinical practice has, in part, been driven by the increasing availability and quality of medical software applications, or "apps. Faster processors, improved memory, smaller batteries, and highly efficient open-source operating systems that perform complex functions have paved the way for the development of a flood of medical mobile device apps for both professional and personal use (Ventola, 2014).

The portability and ease of download of medical apps on mobile devices have made mobile clinical resources available to medical students for practice. Medical apps used for many various purposes are available. According to (Gallagher Healthcare 2020). whether you are a nurse, doctor, physician's assistant, or surgeon, there is an app—in fact, there are several—that can help you become a better medical professional. Apps in the medical field, including

a. Medscape

The_Medscape is a versatile medical reference app that allows physicians to stay up-to-date in the medical field.

b. Kareo

Kareo is a cloud-based electronic health record (EHR) app designed for iOS devices, such as the iPad and iPhone. With Kareo, you can: Add and edit appointments, manage patients, and view patient history.

c. *Epocrates*

As one of the earliest medical apps with millions of downloads, 50% of doctors in the U.S. currently use EOs. The app is known for being a reliable, time-saving tool that keeps you focused on making the best possible clinical decisions.

d. Lexicomp

Lexicomp is a clinical reference app that provides access to a wide range of knowledge. The app references a library of information regarding infectious diseases, oral diseases, and toxicology. There are even mobile apps that simulate surgical procedures or that can conduct simple medical exams, such as hearing or vision tests. The use of medical apps has become frequent and widespread; 70% of medical school HCPs and students reported using at least one medical app regularly, with 50% using their favourite app daily.

Benefits of mobile apps in the medical field

To ensure the quality of medical apps, it seems very important that academic and healthcare organizations develop and update the apps and provide guidelines for the accreditation of apps. It is recommended that for the promotion of knowledge and skills, students provide essential education (Jebraeily, et al., 2017).

Mobile apps provide many benefits for HCPs, perhaps most significantly increasing access to point-of-care tools, which have been shown to support better clinical decision-making and improved patient outcomes. However, some HCPs remain reluctant to adopt their use. Despite the benefits they offer, better standards and validation practices regarding mobile medical apps need to be established to ensure the proper use and integration of these increasingly sophisticated tools into medical practice. These measures will increase barriers to entry into the medical app market, increasing the quality and safety of the apps currently available for use by HCPs.

Machine Learning Techniques

Machine learning is an evolving branch of computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment. These horses are considered working horses in the new era of so-called big data (Jordan, & Mitchell, 2015). Techniques based on machine learning have been applied successfully in diverse fields ranging from pattern recognition, computer vision, spacecraft engineering, finance, entertainment, and computational biology to biomedical and medical applications. Machine learning addresses the question of how to build computers that improve automatically through experience (Jordan, & Mitchell, 2015), and machine learning research has made great progress in many directions. It is one of today's most rapidly growing technical fields, as it lies at the intersection of computer science and statistics and at the core of artificial intelligence and data science. Recent progress in machine learning has been driven both by the development of new learning algorithms and theory and by the on-going explosion in the availability of online data and low-cost computation. Jenni, & Chris (2019) asserted that machine learning techniques are attracting substantial interest from medical researchers and clinicians. We address the need for capacity development in this area by providing a conceptual introduction to machine learning alongside a practical guide to developing and evaluating predictive algorithms using freely available open-source software and public domain data. The adoption of data-intensive machine learning methods can be found throughout science, technology and commerce, leading to more evidence-based decision making across many walks of life, including health care, manufacturing, education, financial modelling, policing, and marketing.

Machine learning techniques are being applied to new kinds of problems, including knowledge discovery in databases, language processing, robot control, and combinatorial optimization, as well as to more traditional problems such as speech recognition, face recognition, handwriting recognition, medical data analysis, and game playing.

Machine learning has increasingly been employed in combination with natural language processing (NLP) to make sense of unstructured text data. By combining ML with NLP techniques, researchers have been able to derive new insights from comments from clinical incident reports, social media activity (Greaves et al, 2013), doctor performance feedback (Gibbons et al, 2017), and patient reports after successful cancer treatments. Automatically generated information from unstructured data could be exceptionally useful not only for gaining insight into quality, safety, and performance but also for early diagnosis. Recently, automated analysis of free speech collected during in-person interviews resulted in the ability to predict the transition to psychosis with perfect accuracy in a group of high-risk youths (Bedi et al, 2015).

Related Work on Machine Learning Algorithms in Medicine

In their work, Mswahil et al (2021) developed five machine learning models to predict the antimalarial bioactivities of a drug against Plasmodium falciparum from the features (i.e., molecular descriptor values) obtained from PaDEL software from the SMILES of compounds and compared the machine learning models via experiments with the collected data of 4794 instances. Therefore, they found that three of the five models, namely, the artificial neural network (ANN), extreme gradient boost (XGB), and random forest (RF) models, outperform the others in terms of accuracy while observing that, using roughly a quarter of the promising descriptors picked by the feature selection algorithm, the five models achieved equivalent and comparable performance.

The objective of this study was to explore the associations between tetronic acid (T1304) and LAP (Laponite) at concentrations of 1–20% (w/w) and 0–3% (w/w), respectively. Response surface methodology (RMS) and two types of machine learning methods (multilayer perceptron (MLP) and support vector machine (SVM)) were used to evaluate the physical behaviour of the systems and the solubility of β-lapachone (β-Lap) in the systems. β-Lap (a model drug with low solubility in water) has antiviral, antiparasitic, antitumour, and anti-inflammatory properties. The results show an adequate machine learning approach to predict the physical behaviour of nanocarrier systems with and without the presence of LAP. Additionally, the analysis performed with SVM showed better results ($R2 > 0.97$) in terms of data adjustment in the evaluation of β-Lap solubility. Furthermore, this work presents a new methodology for classifying phase behaviour using ML. The new methodology allows the creation of a phase behaviour surface for different concentrations of T1304 and LAP at different pH values and temperatures. The machine learning strategies used were excellent for assisting in the optimized development of new nanohybrid platforms.

A universal model built using a neural network and quantitative structure-activity relationships to predict the binding energy between guest and host molecules using input features was proposed by (Kerner & Recum, 2020). The trained model returned correlation values, R2, of 0.9806 and 0.9958 between the predicted and experimental binding affinities for the training and validation sets, respectively. This correlates to mean absolute errors of 0.951 kJ/mol and 0.771 kJ/mol for the training and validation sets, respectively. They concluded that while limited to the current polymers used to train the model, the dataset can be expanded, and models can be retrained for further applications.

In a study conducted by Lo-Ciganic et al (2019) identified individuals at high risk for opioid overdose, targeting many patients who are not truly at high risk. This study aimed to develop and validate a machine learning algorithm to predict opioid overdose risk among Medicare beneficiaries with at least 1 opioid prescription. A prognostic study was conducted between September 2017 and 2018. Participants included fee-for-service Medicare beneficiaries without cancer who completed 1 or more opioid prescriptions from January 1, 2011, to December 31, 2015. The beneficiaries were randomly and equally divided into training, testing, and validation samples. EXPOSURES Potential predictors (n = 268), including socio-demographic, health status, patterns of opioid use, and practitioner-level and regional-level factors, were measured in 3-month windows, starting 3 months before initiating opioids until loss to follow-up or the end of observation. Multivariate logistic regression (MLR), least absolute shrinkage and selection operator–type regression (LASSO), random forest (RF), gradient boosting machine (GBM), and deep neural network (DNN) methods were applied to predict overdose risk in the subsequent 3 months after the initiation of treatment with prescription opioids. Prediction performance was assessed using the C statistic and other metrics (e.g., sensitivity, specificity, and number needed to evaluate [NNE] to identify one overdose).

The Youden index was used to identify the optimized threshold of the predicted score that balanced sensitivity and specificity. The results show that Beneficiaries in the training (n = 186 686), testing (n = 186 685), and validation (n = 186 686) samples had similar characteristics (mean [SD] age of 68.0 [14.5] years; approximately 63% were female, 82% were white, 35% had disabilities, 41% were dual eligible, and 0.60% had at least 1 overdose episode). In the validation sample, the DNN (C statistic = 0.91; 95% CI, 0.88-0.93) and GBM (C statistic = 0.90; 95% CI, 0.87-0.94) algorithms outperformed the LASSO (C statistic = 0.84; 95% CI, 0.80-0.89), RF (C statistic = 0.80; 95% CI, 0.75-0.84), and MLR (C statistic = 0.75; 95% CI, 0.69-0.80) methods for predicting opioid overdose. At the optimized sensitivity and specificity, the DNN had a sensitivity of 92.3%, specificity of 75.7%, NNE of 542, positive predictive value of 0.18%, and negative predictive value of 99.9%.

The DNN classified patients into low-risk (76.2% [142 180] of the cohort), medium-risk (18.6% [34 579] of the cohort), and high-risk (5.2% [9747] of the cohort) subgroups, with only 1 in 10 000 patients in the low-risk subgroup having an overdose episode. More than 90% of overdose episodes occurred in the high-risk and medium-risk subgroups, although positive predictive values were low, given the rare overdose outcome. They concluded that machine learning algorithms appear to perform well for risk prediction and stratification of opioid overdose, especially in identifying low-risk subgroups that have minimal risk of overdose.

Machine learning technique was demonstrated in developing three predictive models for cancer diagnosis using descriptions of nuclei sampled from breast masses(Jenni. & Chris, 2019). These algorithms include regularized general linear model regression (GLM), support vector machines (SVMs) with a radial basis function kernel, and single-layer artificial neural networks. The algorithms were trained on data from the evaluation sample before they were used to predict the diagnostic outcome in the validation dataset. They compared the predictions made on the validation datasets with real-world diagnostic decisions to calculate the accuracy, sensitivity, and specificity of the three models. The results showed that the trained algorithms were able to classify cell nuclei with high accuracy (.94 - .96), sensitivity (.97 - .99), and specificity (.85 - .94). The maximum accuracy (.96) and area under the curve (.97) were achieved using the SVM algorithm. The prediction performance increased marginally (accuracy=.97, sensitivity = .99, specificity = .95) when the algorithms were arranged into a voting ensemble. They used a straightforward example to demonstrate the theory and practice of machine learning for clinicians and medical researchers and concluded that the principles demonstrated can be readily applied to other complex tasks, including natural language processing and image recognition.

Additionally, Li et al (2020) drug combinations are currently a hot research topic in the pharmaceutical industry, but experiment-based methodologies are extremely costly in terms of time and money. Many computational methods have been proposed to address these problems by starting from existing drug combinations. However, in most cases, only molecular structure information is included, which covers too limited a set of drug characteristics to efficiently screen drug combinations. Here, they integrated similarity-based multifeatured drug data to improve the prediction accuracy by using the neighbour recommender method combined with ensemble learning algorithms. By conducting feature assessment analysis, we selected the most useful drug features and achieved an AUC of 0.964 for the ensemble models. The comparison results showed that the ensemble models outperform traditional machine learning algorithms such as support vector machine (SVM), naïve Bayes (NB), and logistic regression (GLM). Furthermore, we predicted 7 candidate drug combinations for a specific drug, paclitaxel, and successfully verified that two of the predicted combinations have promising effects.

The inability to access prescribed medications at the right time indeed poses significant challenges and has serious consequences for patients' health and well-being. Patients often rely on healthcare facilities, such as public hospitals and clinics, to provide them with the necessary medications, and delays or unavailability of medications can lead to undesirable outcomes.

The potential risks associated with patients seeking medications from unverified or substandard sources or not knowing where to receive prescribed medications highlight the importance of having a well-functioning healthcare system that provides access to high-quality medications. It also underscores the significance of patient education and awareness about the potential risks of obtaining medications from unreliable sources and knowing where to receive the right medication. However, in an attempt to address these issues, studies carried out through the application of machine learning and data mining were reviewed.

Most of the studies reviewed focused on the prediction of malaria drug production, overdose drug use, medical adherence, drug combination, etc., using different machine-learning techniques. Although the techniques applied in all these studies yielded positive results, none were used in the prediction of drug prices, and a decision tree, which is a supervised learning technique, was not used. Thus, this study will adopt a decision tree to predict good malaria and typhoid drugs and their prices from pharmacies.

Material and Methods

Overview of the Proposed Model Life Cycle

The life cycle of a model refers to the stages involved in the creation, training, evaluation, and deployment of a machine learning-based model. This chapter presents the proposed system life cycle, analysis of the system, data collection, method used in collecting data, and design and implementation of the proposed system. An overview of the proposed model life cycle is shown in Figure 1.

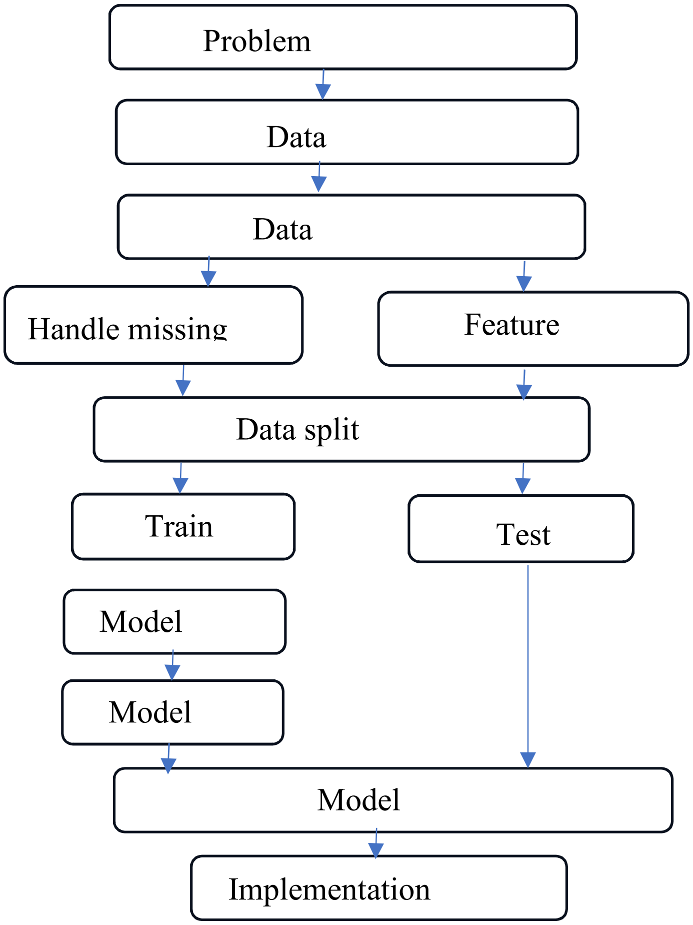

Figure 1. Proposed model life cycle.

Data Collection

Primary data on antimalarial drugs were used for this study. The data were collected from 6 pharmacies in the Yola North Metropolis. The data used consisted of 600 anti-malaria and typhoid drugs of different products and brands, which made up the population of the study.

The purpose of the sample technique was to select the pharmacies used. Purposive sampling is a method of sampling in which the selection of individuals or cases is based on specific characteristics that are relevant to the study's objectives (Andrade, 2021). The historical records of the pharmacies in the provision of anti-malaria and typhoid drugs were the yardstick used by the researcher to select the six pharmacies.

The data were collected through careful observation and recording of all malaria and typhoid drugs available in the selected pharmacies. Table 3.1 shows a sample of the collected data.

Pre-processing

Pre-processing in machine learning refers to the series of steps and techniques applied to raw data before they are fed into a machine learning algorithm for training or prediction. The goal of pre-processing is to transform the data into a format that is suitable for the chosen algorithm, improve model performance, and enhance the quality of predictions.

Data Cleaning

Handling missing values is crucial because they can affect the quality of analysis and machine learning models. Missing values in a dataset refer to the absence of data for one or more variables (features) in certain observations (rows) that occur due to data entry errors, incomplete data collection, or data processing issues. In this stage, the entire rows containing incomplete data were completely removed from the dataset.

Feature selection and Data Encoding

To build the model, features relevant for the model prediction of drug availability were selected. These features include drug name, drug brand (publishers), user, drug type, and pharmacy. These dataset features are all categorical and are of different lengths. Hence, there is a need to encode categorical variables to ensure the compatibility of the data with the algorithm.

Data encoding in machine learning involves converting categorical features into a numerical format that can be used by machine learning algorithms. To achieve this, binary encoding was carried out to create a binary column for each category in the original variable using the binary encode function in Python. This function is given as

encoder = ce. Binary Encoder(cols=['categorical_column']) encoded_df = encoder.fit_transform (dataframe)

Each column represents whether a particular category is present or not for a given observation. Zero (0) encoding represents non availability of drug name, drug brand (publishers), user, drug type, or pharmacy. One (1) encoding represents the availability of a drug name, drug brand (publisher), user, drug type, or pharmacy. The target class (available) is encoded as 1 to represent non availability of the drug and pharmacy, and 2 represents the availability of the drug in a particular pharmacy sample of features; their encoding is shown in Table1.

Table 1: Features encoded

Drug_Name	Brand	User	pharmacy	Available
1	1	0	1	1
1	0	0	1	1
1	0	0	1	1
1	1	0	1	1
1	1	1	1	2
1	1	1	1	2
1	1	1	1	2
1	1	1	1	2
1	1	1	1	2
1	1	1	1	2
1	1	1	1	2
1	1	1	1	2
0	0	0	1	1
0	0	1	1	1
0	0	1	1	1
0	0	1	1	1
1	1	1	1	2
1	1	1	1	2

Data Splitting

This stage involves dividing a dataset into multiple subsets for the purpose of training, validating, and testing machine learning models with the aim of evaluating the model's performance on unseen data and preventing over fitting, where the model performs well on the training data but poorly on new, unseen data.

The dataset was divided into 560 training sets constituting 80% of the dataset. This subset is used to train the machine learning model. The model learns patterns, relationships, and features from these data. A total of 140 test sets constituting 20% of the dataset were used to evaluate the model's final performance. It provides an unbiased estimate of the model's ability to make accurate predictions on new, unseen data.

Splitting criteria

The decision of how to split the data at each node is determined by a splitting criterion, such as the Gini impurity (for classification). The criterion aims to create subsets with homogeneous classes or similar target values.

The decision tree algorithm builds the tree by recursively choosing the best features and thresholds to split the data, resulting in a hierarchy of nodes and branches.

Model Selection

A decision tree is a versatile and widely used machine learning algorithm that can be used for both classification and regression tasks. It is a graphical representation of a decision-making process that recursively splits data into subsets based on the values of input features. Each split or decision is made based on a particular feature and a threshold, leading to a tree-like structure of nodes and branches.

.***Build and train model***

Tools for Building Model

Python was used to construct the decision tree model. Python is a high-level, versatile, and widely used tool for programming, and data analysis is known for its simplicity, readability, and ease of use. It was created by Guido van Rossum and first released in 1991.

Python is one of the most popular languages for data analysis, manipulation, visualization, and machine learning. NumPy, pandas, Matplotlib, Seaborn, and scikit-learn were the libraries used.

The model is built and trained to predict the availability of the right drug in a particular pharmacy based on certain conditions. decisionTreeClassifier () and fit () functions.

Model Training and Testing

Training is a process that helps ML algorithms extract useful information to train models capable of producing accurate predictions and hence desired outcomes. The technique used in this study is a decision tree. In building the model, features were specified. The features used to predict the target (available) are drug name, brand, type, user, and pharmacy. The decision tree algorithm was fit to the training set CART, which is a decision tree algorithm that was used to develop and train the model on 80% of the dataset. The best attribute is picked using the information gain, which is the measurement of changes in entropy after the segmentation of a dataset using the formula below:

i. Entropy= Entropy(S) [(Weighted Avg) *E ntropy(attribute) (1)

After training the model, 20% of the data were used to test the model to make predictions on the test dataset. The model was tested on the testing dataset to predict the availability of drugs in a pharmacy and was checked to determine how well the model performed on the testing dataset using a tree. score () method as well.

The model decision rule is based on entropy. Mathematically, the entropy of a node is calculated as:

ii. Entropy= -Σ p_i * log2(p_i) = -Σ p_i * log2(p_i) (2)

where:

p_i represents the proportion of samples of class i in the node.

Model performance evaluation

Model performance was evaluated using the following methods and metrics to determine how well the model made predictions on both the training and test datasets.

i. Accuracy: Accuracy is a basic metric that calculates the proportion of correctly predicted instances out of the total number of instances in the dataset. It is suitable for balanced datasets.
ii. Confusion matrix: A confusion matrix provides a detailed view of the model's performance. It breaks down predictions into true positives, true negatives, false positives, and false negatives. From the confusion matrix, researchers calculated the precision, recall, and F1-score.
iii. Accuracy= TP+TN/TP+TN+FP+FN (3)
iv. Precision: Precision is the ratio of true positive predictions to the total number of predicted positives. It quantifies the model's ability to correctly identify positive cases and is particularly useful when the cost of false positives is high.

Precision = TP/(TP+FP) (4)

v. Recall (sensitivity or true positive rate): Recall is the ratio of true positive predictions to the total number of true positives. It measures the model's ability to identify all positive cases and is important when the cost of false negatives is high.

Recall = T P/(T P + F N) (5)

vi. F1-Score: The F1-score is the harmonic mean of the precision and recall. This approach provides a balanced view of a model's performance by considering both false positives and false negatives.

Results and Discussion

The accuracy, precision, recall, and confusion matrix are the techniques and metrics used to evaluate the model to assess its performance and generalizability. Table 2 shows the model accuracy, precision, and recall.

Table 2: Accuracy, precision, and recall results

Metric	**0**	**1**	**Accuracy %**	
		%	**Training set**	**Testing set**
Precision	**100**	**100**		
Recall	**100**	**100**	**98**	**100**

The accuracy of the model is the proportion of correctly predicted instances to the total number of instances.

Precision measures the ratio of the number of correctly predicted positive observations to the total number of predicted positives.

Recall is the ratio of correctly predicted positive observations to the number of actual positives in the dataset. The confusion matrix creates visualized true positive, true negative, false positive, and false negative predictions that reflect the model's performance across different classes. Figure 2 shows the confusion matrix results of the model.

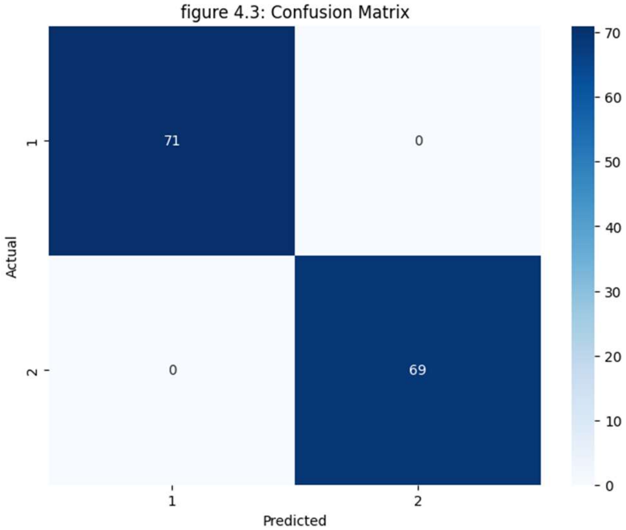

Figure 2:

Model Cross-Validation Results

Cross-validation is a technique used to assess the general performance of a machine learning model by splitting the dataset into multiple folds for training and testing. Fivefold cross-validation was carried out to assess the model's general performance. Table 3 shows the 5-fold cross-validation results, means and standard deviations of the models.

Table 3: 5 Cross-Validation Scores

Fold1%	Fold2%	Fold3%	Fold4%	Fold5%	Mean%	SD
1.0	**79.29**	**1.0**	**98.6**	**1.0**	**95.57**	**0.00816**

These scores represent the accuracy of the decision tree model on different folds of the cross-validation process. Each score corresponds to a separate fold. The scores indicate that the model achieved perfect accuracy (100%) on three of the five folds and very high accuracy (79.29%) and accuracy (98.57%) on the remaining two folds.

The mean accuracy, which is the average accuracy across all folds of the cross-validation, was 0.9957142. The mean accuracy is approximately 99.57%. This suggests that, on average, the model correctly predicted approximately 99.57% of the instances in the test sets.

Similarly, the standard deviation measures the variability of the accuracy scores across the different folds and was found to be `0.0286163248976326`. This standard deviation is approximately 0.02861, which indicates the degree of variation in accuracy between the folds. A lower standard deviation generally implies that the model's performance is consistent across folds. The low standard deviation suggests that the model's accuracy is consistent across different folds, indicating stable performance.

Overall, the cross-validation scores indicate that this decision tree model performed very well on the dataset. It achieved perfect accuracy on three out of the five folds and very high accuracy on the remaining two folds. The average accuracy across all folds is high, reflecting the model's ability to generalize well to unseen data. The potential reasons for the differences in accuracy between folds can be attributed to variability in the data, specific instances in the test sets, or the small size of the dataset.

This result agreed with the results of a study by (Pall, et al, 2023), in which the authors were able to correctly predict drug shortages with 69% accuracy and a kappa value of 0.44. Additionally, 59% of the shortages were predicted to be the most impactful. Additionally, Mbonyinshuti et al (2022) focused on the application of machine learning (ML) to forecast future trends in the demand for essential drugs in Rwanda. The random forest was able to predict 10 selected medicines with an accuracy of 88% with the training set and 76% with the test set, and it can thus be used to forecast future demand based on past consumption data by inputting a month, year, district, and medicine name.

Conclusion

Given that accessing the right medicine in the right pharmacy at the right time poses some challenge for demand and supply to drug seekers, the need to find ways to overcome this challenge to maximize availability and increase access to the right drugs is imperative through the application of predictive modelling.

This study focused on the application of a decision tree, an ML technique, for predicting the availability

of malaria drugs in Yola, Adamawa State, Nigeria. In building the model, 80% of the data were used as the training set, and 20% of the data were used as the test set.
The model prediction accuracy on the training set was 95%, and that on the test set was 98%. Thus, it can be used to make predictions by inputting the drug name, brand, and type. Finally, this study was implemented by developing a mobile android-based app to assist drug seekers in finding and accessing drugs easily.

The challenge encountered during this study was that some pharmacies were unwilling to provide the necessary data. This challenge informed the decision to use a limited number of datasets.

Contribution to knowledge

In this study, we identified and processed relevant data and transformed it into a suitable format for building a predictive model. We used the training dataset to create the model and evaluated its performance with the test dataset, all within the Python programming language. Specifically, we applied a decision tree, a supervised machine learning algorithm, to classify and predict the best drugs. Subsequently, we implemented the model's frontend using a full-stack (FE, M, BE) approach.

Our work presents a supervised learning approach that predicts the availability of malaria and typhoid drugs using the decision tree machine learning technique. We also designed and implemented the model for use by both pharmacies and clients seeking drugs, with the hope that it will address accessibility issues in obtaining medications from pharmacies.

Recommendation

Given the results of the study, the researcher recommends the following:

i. Pharmacies should take advantage of this ML model to make drugs available for easy access to individuals in need. The problem of not knowing where to obtain drugs will be mitigated greatly if pharmacies use the model to ensure the optimal availability of medicines.

ii. Individuals should take advantage of this platform for quick identification of the right pharmacies to access drugs at the right time where needed by simply downloading and installing the app on an android device. This will help save individuals in critical condition who require urgent attention.

References

Almeziny, M. A. (2017). Factors Influencing Patient Waiting Time in the PSMMC Outpatient Pharmacy. Factors Influencing Access to Medicines in Nigeria: Views and Experiences of Residents of the Federal Capital Territory

Application Software for Personal, Enterprise & Workgroup Objectives. (2013, October 18). Retrieved from https://study.com/academy/lesson/application-software-for-personal-enterprise-workgroup-objectives.html.

Aungst, T. D. (2013). *Medical applications for pharmacists using mobile devices*. Ann Pharmacother;47(7–8):1088–1095.

Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., Ribeiro, S., Javitt, D. C., Copelli, M., & Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset inhigh-riskyouths. *NPJschizophrenia*, *1*,15030. https://doi.org/10.1038/npjschz.2015.30

De Melo B., R. & Lima, C., Oliveira, F., Câmara, G., Viseras, C., Moura, T., Souto, E. B., Severino, P., Raffin, F. & Fernandes, M. (2022). New Machine Learning Approach for the Optimization of Nano-Hybrid Formulations. Nanomanufacturing. 2. 82-97. 10.3390/nanomanufacturing2030007.

Devarajan, S., & Das, J. (2012). *Improving access to drugs: Fitting the solution to the problem available. at* https://blogs.worldbank.org/africacan/improving-access-to-drugs-fitting-the-solution-to-the-problem

Divali, P., Camosso-Stefinovic, J. & Baker, R.. *(2013).* Use of personal digital assistants in clinical decision making by health care professionals: a systematic review. *Health Informatics J*;19(1):16–28. 19(3),

Gallagher Healthcare (2020). *The 23 Best Medical Apps for Doctors Retrieved* on 27/06/2022 from https://www.gallaghermalpractice.com/blog/post/the-23-best-medical-apps-for-doctors

Gibbons, C, Richards, S., Valderas, J. M.,& Campbell, J. (2017). *Supervised Machine Learning Algorithms Can Classify Open-Text Feedback of Doctor Performance With Human-Level Accuracy*, J Med internet Res. 2017;19(3): 65. https://doi.org/10.2196/jmir.6533

Greaves, F, Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). *Use of sentiment analysis for capturing patient experience from free-text comments posted online*,. J Med internet Res. 15(11):239. https://doi.org/10.2196/jmir.2721.6

Idowu,S. O.(2017) *What I've learnt about drug testing in Nigeria over the last 20 years.* https://theconversation.com/what-ive-learnt-about-drug-testing-in-nigeria-over-the-last-20-years-129561

Jebraeily, M., Fazlollahi, Z. Z., & Rahimi, B. (2017). *The Most Common Smartphone Applications Used By Medical Students and Barriers of Using Them. Acta Inform Med.*;25(4):232-235. doi: 10.5455/aim.2017.25.232-235. PMID: 29284911; PMCID: PMC5723200.

Jenni, A. M. S. & Chris, J. S. (2019). *Machine learning in medicine: a practical introduction*. BMC Medical Research Methodology 19:64 https://doi.org/10.1186/s12874-019-0681-4

Jordan, M. .I.,& Mitchell, T.M. (2015). Machine learning: Trends, perspectives, and prospects. Science. 349(6245):255-60. doi: 10.1126/science.aaa8415. PMID: 26185243.

Kerner, J. & Recum, H. (2020). Predicting Drug Interactions to Unassociated Biomedical Implants Using Machine Learning Techniques and Model Polymers. 10.1101/2020.11.10.374900.

Li, J., Tong, X., Zhu, L & Zhang, H.(2020). *A Machine Learning Method for Drug Combination Prediction.* https://www.frontiersin.org/articles/10.3389/fgene.2020.01000/full

Lo-Ciganic W, Huang, J. L., Zhang, H. H. (2019) Evaluation of Machine-Learning Algorithms for Predicting Opioid Overdose Risk Among Medicare Beneficiaries With Opioid Prescriptions. *JAMA Netw Open*. 2(3):e190968. doi:10.1001/jamanetworkopen.2019.0968

Mbonyinshuti F, Nkurunziza J, Niyobuhungiro J, & Kayitare E. (2022). The Prediction of Essential Medicines Demand: A Machine Learning Approach Using Consumption Data in Rwanda. *Processes*. 2022; 10(1):26. https://doi.org/10.3390/pr10010026

Mswahili, M. E., Martin, G. L., Woo, J., Choi, G. J., & Jeong, Y. S. (2021). Antimalarial Drug Predictions Using Molecular Descriptors and Machine Learning against Plasmodium Falciparum. *Biomolecules*, *11*(12), 175https://doi.org/10.3390/biom11121750

National Population Commission Nigeria, ICF International. Nigeria Demographic and Health Survey (2013). Abuja, Nigeria, and Rockville, Maryland, USA: NPC and ICF International. 2014; p. 217.

Obiebi I. P. (2019). Adherence to antimalarial drug policy among doctors in Delta State, Nigeria: implications for malaria control. *Ghana medical journal*, *53*(2), 109–116. https://doi.org/10.4314/gmj.v53i2.5

Oguonu, T., & Edelu, B. O. (2016). Challenges of Managing Childhood Malaria in a Developing Country: The Case of Nigeria. InTech. doi: 10.5772/65488

Oleszkiewicz, P., Krysinski, J., Religioni, U.& Merks, P. (2021). *Access to Medicines via Non-Pharmacy Outlets in European Countries—A Review of Regulations and the Influence on the Self-Medication Phenomenon. Healthcare*, 9, 123. https://doi.org/10.3390/healthcare9020123

Onyekachukwu, S. D. (2022). Reason why your business needs a mobile app .available at https://www.linkedin.com/pulse/reason-why-your-business-

Pall, R., Gauthier, Y., Auer, S., & Mowaswes, W. (2023). Predicting drug shortages using pharmacy data and machine learning. *Health care management science*, *26*(3), 395– 411. https://doi.org/10.1007/s10729-022-09627-y

Payne, K.F. B., Wharrad, H.& Watts, K. (2012). *Smartphone and medical related App use among medical students and junior doctors in the United Kingdom (UK): a regional survey.BMC Medical Informatics and Decision Making* 2012,12:121 Page 2 of 11 http://www.biomedcentral.com/1472-6947/12/121

United Nations Office on drugs and Crime (2019).https://www.unodc.org/documents/data-and analysis/statistics/Drugs/Drug_Use_Survey_Nigeria_2019_book.pdf

Ventola, C. L. (2014). *Mobile devices and apps for health care professionals: uses and benefits.* P T. 2014 May;39(5):356-64. PMID: 24883008; PMCID:PMC4029126.

Wallace,S., Clark, M., &White, J.(2012). 'It's on my iPhone': *attitudes to the use of mobile computing devices in medical education, a mixed-methods study*. BMJ Open.