CSCI 6443

Data Mining

Term Paper

Title: **Improve diagnostic accuracy**

Submitted by: **Harshini Naidu Ganapathy (GWID – G39946084)**

Abstract

There is a critical global health issue: misdiagnosis. Each year, incorrect diagnoses affect millions of individuals, leading not only to adverse health outcomes, including potential fatalities, but also imposing significant financial strains on patients, their families, insurance providers, and governmental health bodies. Moreover, the professional lives of physicians are at risk due to unintended errors in medication prescription or disease identification, potentially damaging their careers and reputation in the medical community.

Recognizing the urgency and importance of improving diagnostic accuracy, our research focuses on leveraging data mining techniques to analyze a dataset of medical prescriptions. Our objective is to unearth patterns and knowledge that can enhance the diagnostic process, ultimately aiming to support patient health, reduce unnecessary financial expenditure, and assist physicians in making accurate diagnoses.

To achieve this, we employed four distinct single classification algorithms: decision trees, random forest, Naive Bayes, and K-nearest neighbors. These methods were chosen for their ability to handle complex datasets and their varied approaches to classification, which include decision rules, ensemble methods, probabilistic predictions, and similarity measurements, respectively. Our initial predictions targeted outcome: identifying the disease based on prescriptions they are using.

However, to refine our results and push the boundaries of diagnostic accuracy, we integrated these algorithms into an Ensemble Learning framework. Ensemble Learning combines the strengths of individual models to improve overall performance, aiming for higher accuracy and reliability in predictions. This approach is particularly effective in complex problem areas like medical diagnosis, where the intricacy of data and the critical nature of outcomes demand precision.

Upon conducting extensive experiments to compare the efficacy of different data mining techniques, our proposed Ensemble Learning model demonstrated superior performance. It achieved an accuracy and kappa score of 98.73% and 0.9218 for disease prediction. These results not only show a promising improvement over other studies in the field but also highlight the potential of data mining and Ensemble Learning in enhancing diagnostic processes.

The implications of our study are far-reaching. By improving diagnostic accuracy, we can directly contribute to better patient outcomes, preventing health complications and potentially saving lives. Additionally, by reducing misdiagnoses, we can alleviate the financial burden on patients, insurance companies, and governments, ensuring resources are used efficiently. For physicians, our findings offer a valuable tool to minimize diagnostic errors, supporting their decision-making process and safeguarding their professional integrity.In conclusion, our research underscores the transformative potential of data mining in healthcare, specifically in improving diagnostic accuracy. As we move forward, these insights not only benefit the immediate stakeholders but also set a foundational benchmark for future investigations aimed at optimizing health outcomes through technological innovation.

Table of contents

1. Introduction

Research indicates that annually, around 12 million people globally are impacted by incorrect medical diagnoses, translating to about 5% (or one in 20) of patients receiving the wrong diagnosis. Among these, 10% to 20% are in severe condition. It's estimated that misdiagnosis leads to the deaths of 40,000 to 80,000 individuals each year. Disparities exist as women and minorities are disproportionately affected, experiencing higher rates of misdiagnosis by 20% to 30%. Specifically, misdiagnoses are associated with 44% of cancer cases, with prostate, breast, and thyroid cancers having the highest misdiagnosis rates. Additionally, 51% of individuals have reported receiving a different diagnosis after seeking a second opinion following a breast x-ray.[1]

One third of medical mistakes leading to death or disability stem from incorrect or delayed diagnoses. The repercussions of such diagnostic errors are significant, encompassing unnecessary medical treatments, increased financial burdens for both patients and governments, emotional and physical distress, and, in some cases, death. Notably, diagnostic mistakes are a primary cause of malpractice claims, accounting for 28.6% of such claims and 35.2% of the total compensation paid. According to studies, the average payout for a claim related to diagnostic errors, adjusted for 2011 inflation, was $386,849, with a median of $213,250. Over a decade, the compensation for diagnostic errors amounted to $1.8 billion.[1]

Disease diagnosis involves identifying the illness that most accurately corresponds to an individual's symptoms. The primary challenge in this process is that certain symptoms and signs can be ambiguous, making accurate identification of the disease crucial for effective treatment of any condition.[2]

Machine learning is a discipline capable of forecasting disease diagnoses by utilizing previously gathered training data. Numerous researchers have developed a variety of machine learning algorithms aimed at accurately recognizing a broad spectrum of illnesses. These algorithms can construct models that predict diseases and suggest appropriate treatments.[3]

The abundance of data accessible has made disease prediction a critical area of study. Researchers utilize these extensive datasets to develop models for disease prediction within decision-making systems, facilitating enhanced early-stage disease detection and treatment. Prompt diagnosis and immediate intervention stand as the most efficient strategies to reduce mortality rates associated with diseases.[4]

In pursuit of these goals, data mining within the healthcare sector offers a suite of tools and methodologies that can analyze data to unveil underlying patterns. Broadly, these data mining techniques are categorized into descriptive and predictive types. Descriptive techniques encompass methods like clustering and association rule mining, while predictive techniques involve classification and forecasting approaches.[5,6]

This study aims to leverage data mining techniques to extract insights from a dataset of medical prescriptions obtained from the Data.CMS.gov website. By examining the prescribed medications for various diseases, our methodology seeks to predict both the category and specific type of disease affecting the patient. We employed various classification methods to forecast diseases based on the prescription data, and our experiments indicate that the prediction outcomes are satisfactory. The structure of this paper is as follows: Section 2 provides the background information. The methodology we propose is detailed in Section 3. Section 4 discusses the experimental results and their implications. The conclusion of the study is found in Section 5, and Section 6 includes the declarations.

1.1.Abbreviations and Important Terms

Table 1 below, presents a comprehensive list of the important terms and abbreviations that are referenced throughout the remaining text.

Table 1: Abbreviations and Terms

Term / Abbreviation	Definition
RF	Random Forest
LM	Linear Models
ANN	Artificial Neural Network
HRFLM	combining random forest (RF) features with a linear method (LM)

HRFLC	combining random forest, AdaBoost, and the Pearson coefficient
IGFS	Information Gain-based Feature Selection
SVM	Support Vector Machine
CKD	Chronic kidney disease
CNN	Convolutional Neural Network
KNN	K-Nearest Neighbors
PCA	Principal Component Analysis
CDSS	Clinical Decision Support Systems
ML	Machine Learning

2. Literature Survey: Past Related Work

Recent years have seen a surge in research focused on forecasting various diseases, their treatments, and the discovery of new drugs globally. Various data mining techniques have been applied to disease detection, yielding diverse outcomes. This includes significant work on diseases like heart disease, diabetes, cancer, and more. Heart disease, being one of the most prevalent conditions affecting humans today, necessitates early prediction and diagnosis to decrease mortality rates. Among the notable studies in this area, Kondababu and colleagues (2021) achieved promising results in predicting heart disease using machine learning algorithms. Their study highlighted several existing approaches, with the HRFLM technique—combining random forest (RF) features with a linear method (LM)—standing out for its high accuracy rate of 88.7%.[7]

In 2021, Jeyaranjani and colleagues devised a decision support system utilizing a supervised learning model to determine the status of coronary heart disease via angiography. Their findings showcased the Artificial Neural Network (ANN) model's effectiveness, boasting a 97% accuracy rate in predicting the stages of the disease. This system plays a crucial role in facilitating the early detection of coronary heart conditions.[8]

In 2021, Maini and team introduced a machine learning system designed to predict heart disease specifically within the Indian demographic. This online-accessible system excels in the early detection of cardiovascular diseases. It showcased impressive performance metrics, with the Random Forest (RF) algorithms achieving an accuracy of 93.8%, sensitivity of 92.8%, and specificity of 94.6%.[9]

In 2021, Kumari and her team introduced a soft voting classifier model, incorporating three algorithms: random forest, logistic regression, and Naive Bayes, aimed at predicting diabetes in patients. They tested their model on two datasets: the Pima Indians Diabetes Database and the Breast Cancer Database. The accuracy achieved by their model was 79.08% on the diabetes dataset and 97.02% on the breast cancer dataset.[10]

In 2021, Pavithra and Jayalakshmi introduced an innovative HRFLC feature selection approach, combining random forest, AdaBoost, and the Pearson coefficient. This method significantly enhances disease prediction efficiency and accuracy.[11]

Despite the availability of various data mining classification algorithms for heart disease prediction, there is a lack of data specifically for forecasting heart disease in individuals with diabetes. To address this, Arumugam and colleagues (2021) refined the decision tree model to enhance its prediction accuracy for heart disease risk in diabetic patients, finding it to consistently surpass the performance of simple vector and Naive Bayes models.[12]

Diabetes represents a significant health issue globally, contributing to numerous fatalities annually, prompting extensive research focused on its prediction. Among these studies, Jain and co-authors (2021) utilized artificial intelligence algorithms to forecast diabetes within the Pima Indians Diabetes dataset. Their findings highlighted that the neural network algorithm, demonstrating an 87.88% accuracy rate, performed optimally. This level of precision is advantageous for medical professionals aiming to treat the disease in its initial stages.[13]

In 2021, Jothi and associates introduced a model designed to forecast heart disease risk using the decision tree algorithm. Their research demonstrated that applying the Decision Tree algorithm to the dataset enabled the prediction of a patient's heart disease risk with an accuracy of 81%.[14]

In the current era, cancer remains a leading cause of mortality, with breast cancer particularly significant among causes of death in women globally. In light of this, extensive research efforts have been dedicated to this area. Recognizing the critical importance of early detection and intervention for lymphedema in enhancing the lives of breast cancer survivors, Wei and colleagues (2021) aimed to create a model that could provide early warnings for breast cancer-related lymphedema. They introduced a logistic regression model that excelled in performance, evidenced by an AUC of 0.889 (range 0.840–0.938), sensitivity of 0.771, specificity of 0.883, accuracy of 0.825, Brier score of 0.141, and satisfactory calibration.[15]

In 2021, Ramesh and team introduced an algorithm for selecting features designed to improve the efficacy of machine learning techniques, termed the Information Gain-based Feature Selection (IGFS). Their research demonstrated that, when applied, the Support Vector Machine (SVM) and Random Forest (RF) algorithms exhibited the best results, achieving an accuracy rate of 88%.[16]

In 2021, Khaleel and Al-Bakry developed a model capable of determining the presence of diabetes in an individual. Their findings indicated that the logistic regression model they proposed was more successful in predicting diabetes than other algorithms, demonstrating a high accuracy of 94%.[17]

In 2015, Kumar and Sahoo introduced an innovative algorithm that merges the Naive Bayes and genetic algorithms to enhance the classification accuracy for heart disease. This algorithm is designed to distinguish between heart disease data sets as either indicative of illness or health. The experimental outcomes from six different data sets in their study affirm that this method is effective for classification purposes. Their predictive model provides a useful tool for doctors, enabling them to diagnose heart disease more effectively using fewer data points.[18]

Chronic kidney disease (CKD) is marked by a progressive decline in kidney performance, often without symptoms in its initial phases, making early identification crucial for mitigating later-stage risks. In 2020, Pinto and colleagues applied the CRISP-DM methodology to develop a predictive system for CKD conditions. Their findings revealed that the J48 algorithm, which they proposed, delivered the most effective outcomes, achieving an accuracy of 97.66%, a sensitivity of 96.13%, a specificity of 98.78%, and a precision of 98.31%.[19]

Recognizing the significance of data on the spread of infectious and chronic diseases as a key aspect of epidemiological information for assessing community health, Teimouri and team (2016) estimated the prevalence of diseases treated in outpatient settings by analyzing outpatient prescription data. Of the various classification methods evaluated in their research, the support vector machine stood out with an impressive accuracy of 95.32%. Subsequently, to enhance the performance of singular data mining techniques, combination strategies were employed. Among these, the Weighted Voting algorithm emerged as the most effective, achieving an accuracy of 97.16%.[20]

In 2020, Dhanya and colleagues leveraged both established ensemble methods and a mix of supervised machine learning algorithms to create a novel model aimed at predicting breast cancer. Recognizing that not all data features are essential for accurate breast cancer prediction, and that selecting relevant features can enhance model efficiency, they applied feature selection techniques. The results indicated that their suggested stacking ensemble approach, combined with f-test feature selection, proved to be a dependable and effective strategy for breast cancer prediction.[21]

In the current era, the importance of early diagnosis cannot be overstated. In 2021, Malladi and team employed machine learning techniques to predict diseases based on symptomatic data. Their findings highlighted that the Convolutional Neural Network (CNN) algorithm outperformed the K-Nearest Neighbors (KNN) algorithm in predicting general diseases, with a reliability rate of 84.5%.[22]

In 2015, Onan devised a technique for constructing a cancer diagnosis system that integrates fuzzy-rough nearest neighbor classification, consistency-based subset evaluation, and fuzzy-rough instance selection. This approach employs feature selection to enhance the system's understandability, reduce training duration, and improve model generalization. The assessment of this method revealed a remarkable accuracy of 99.71%, demonstrating its reliability as an automated tool for diagnosing breast cancer.[23]

In 2019, Dehkordi and colleagues conducted a study to forecast the kind of physician—whether public or private—a patient was referred to, as well as the specific disease they were diagnosed with. The dataset utilized in this research comprised 70 distinct disease types, 386 different drugs, and encompassed a total of 600 patient records. To enhance the predictive capability of their model, a stacking approach was employed. The findings indicated that the model achieved an accuracy of 73.17% in predicting the type of physician and 57% in determining the disease type.[24]Nowadays, obesity poses a significant health risk globally, acting as a precursor to complex conditions like stroke, heart disease, and liver cancer. In 2021, Ferdowsy and the team utilized machine learning algorithms to forecast the likelihood of obesity. Their findings demonstrated that the logistic regression algorithm they proposed performed effectively, achieving an accuracy of 97.09%.[25]

3.Ensemble Learning models

The development of Ensemble Learning models, including this particular model, is primarily aimed at lowering the rate of errors. The foundational principle behind Ensemble Learning is that the likelihood of incorrectly classifying or positioning a new sample is significantly reduced when multiple models are combined, as opposed to relying on a singular model. Stacking, which is a form of Ensemble Learning, operates on a similar premise to Boosting and Bagging (Bootstrap Aggregation). Boosting is a collective algorithm in machine learning that aims to minimise both variance and bias by transforming a group of weak learners into strong ones. Specifically, Boosting targets bias reduction and typically involves models that are low in variance but high in bias, with the Adaboost algorithm being a notable technique that adjusts the weights of each training sample. Conversely, Bagging seeks to enhance the stability and accuracy of machine learning algorithms applied in statistical classification and regression, by generating a composite model more reliable than individual base models. This method not only reduces variance but also aids in mitigating overfitting.[26]

3.1Stacking

To reduce the error rate, the approach of combining several machine learning techniques into a single predictive model is adopted, aiming to enhance predictive accuracy. This methodology capitalizes on the premise that the probability of misclassifying the category or position of a new sample is significantly diminished when utilizing multiple models, rather than a solitary one. In this system, predictions from each base model serve as inputs for a meta-level classifier, which then produces the final prediction. All the base models are trained using the available data. Subsequently, a hybrid model that incorporates the strengths of these individual models is trained to make the final prediction, effectively leveraging the collective insight for more accurate outcomes.[27]

3.2The Proposed Method

In this section, in the first part, the data collection method is explained and then in the second part, a suitable model for predicting the disease and the disease category is presented.

3.2.1Data collection

In this project I used dataset from Data.CMS.gov site.I used the dataset Medicare Part D Prescribers by Geography and Drug.The dataset for Medicare Part D Prescribers by Geography and Drug provides details on prescription medications issued by physicians and other healthcare professionals, covered under the Medicare Part D Prescription Drug Program. For every medication, it records the total count of dispensed prescriptions, encompassing both initial prescriptions and refills. Additionally, the dataset accounts for the total cost of the drug, incorporating the cost of the drug's ingredients, dispensing fees, sales tax, and any relevant administration fees. This cost is calculated based on payments made by the Part D plan, the Medicare beneficiary, government subsidies, and contributions from any other third-party payers.It consists of 114.9k rows and 22 columns.

3.2.2Model

This project presents a comprehensive exploration into the development and evaluation of an ensemble machine learning (ML) model, aiming to predict drug classifications based on prescription data. The endeavor is set against the backdrop of a dataset that contains detailed prescription information, including the total number of claims, total drug costs, and the generic names of drugs. The primary challenge is to classify drugs into one of four categories: Opioid, Long-Acting Opioid (Opioid_LA), Antibiotic, and Antipsychotic, leveraging the given features.

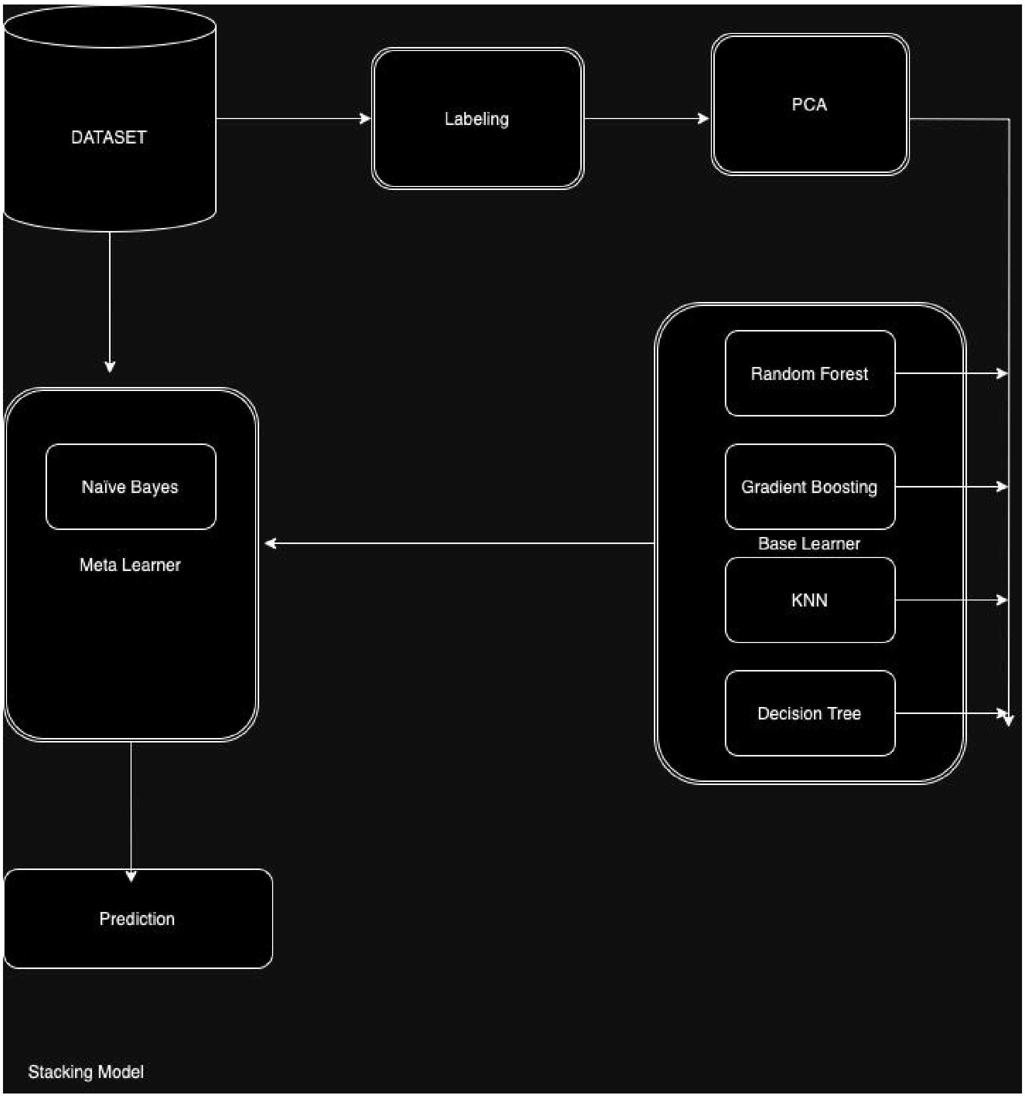


Figure 1- Ensemble Model Architecture

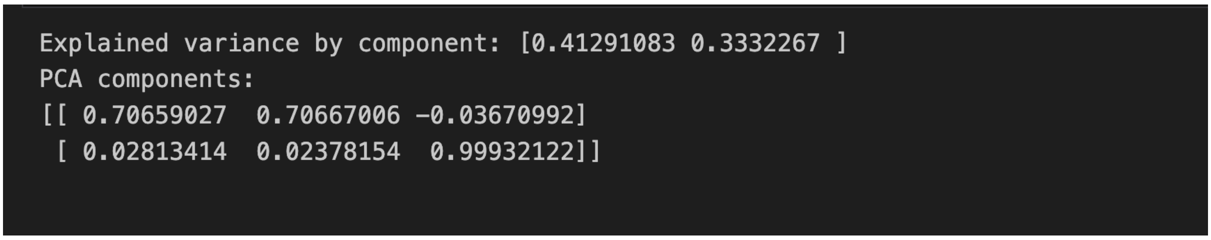
3.2.3Data Preprocessing and Transformation

The initial stage of the project focuses on preparing the dataset for ML modeling. A crucial part of this process involves encoding the generic names of drugs using Label Encoding, a technique that converts categorical text data into a numerical format that can be understood by ML algorithms. This transformation is vital because most ML models are designed to work with numerical data.

Moreover, the dataset features several binary flags that indicate whether a drug falls into specific classifications, such as whether it is an opioid or an antibiotic. These flags are combined into a single target variable, `Drug_Classification`, which simplifies the task into a multiclass classification problem. This target variable is subsequently encoded, providing a numerical representation of the drug categories to be predicted.

Following the encoding process, three features are selected for the modeling phase: the total number of claims, total drug costs, and the encoded generic names. These features are believed to hold significant information that could influence the drug's classification. To ensure that all features contribute equally to the model's decision-making process, they are standardized using `StandardScaler`. This standardization ensures that the features have a mean of zero and a standard deviation of one, mitigating the risk that larger scale features will dominate the learning process.

To further refine the feature set, Principal Component Analysis (PCA) is applied for dimensionality reduction. PCA is a statistical technique that transforms the original correlated features into a set of values of linearly uncorrelated variables called principal components. This transformation is performed while attempting to retain as much of the variance in the original dataset as possible. For this project, the feature space is reduced to two principal components, aiming to simplify the model while preserving essential information.Approximately 41.29% of the variability in your dataset can be represented by the first principal component alone and second principal component, which is about 33.32%.The total variance explained by the first two components together, which in this case is approximately 74.61% The goal is to reduce the dimensionality of the dataset by selecting the first few principal components, thereby simplifying the dataset while retaining as much information as possible.



```
Explained variance by component: [0.41291083 0.3332267 ]
PCA components:
[[ 0.70659027  0.70667006 -0.03670992]
 [ 0.02813414  0.02378154  0.99932122]]
```

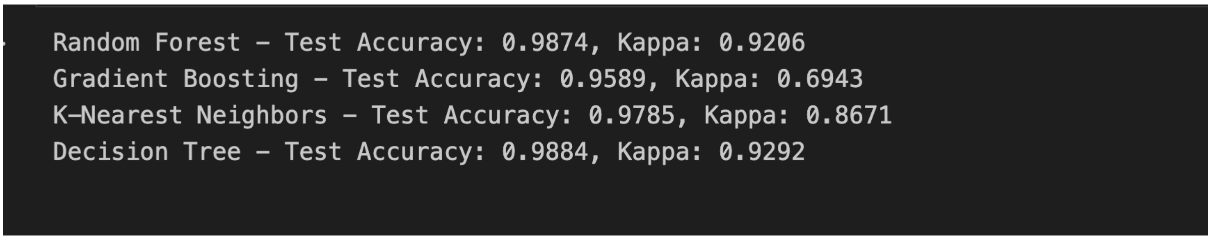
Figure 2 - Variance after PCA

3.2.4Model Training and Evaluation

The core of the project involves training and evaluating four distinct base ML models: Random Forest, Gradient Boosting, K-Nearest Neighbors (KNN), and Decision Tree. Each model offers unique strengths and approaches to learning from data, making them suitable candidates for ensemble learning. After training on the processed dataset, the performance of each model is assessed using two key metrics: accuracy and Cohen's Kappa score. Accuracy measures the proportion of correctly predicted instances out of all predictions, providing a straightforward assessment of performance. In contrast, Cohen's Kappa score offers a more nuanced evaluation by accounting for the agreement between the predicted and actual classifications, adjusted for chance agreement. This metric is particularly useful in multiclass classification tasks, where random chance could otherwise inflate the perceived accuracy.

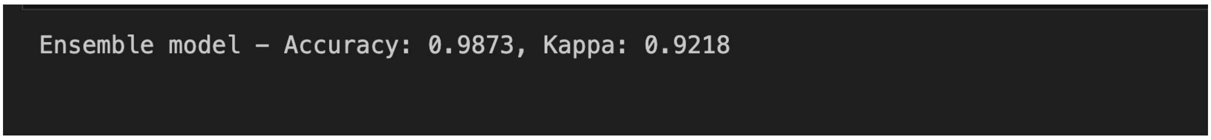
The predictions made by these base models on the training set serve as input features for a meta model, a Naïve Bayes classifier, in a technique known as stacking. This second-level model aims to learn how to best combine the predictions of the base models to improve the overall prediction accuracy. By training on the predictions rather than the original features, the meta model effectively learns the strengths and weaknesses of each base model, optimizing the final predictions.

Finally, the ensemble model's performance is evaluated on a separate test set, using the same accuracy and Cohen's Kappa score metrics. This evaluation offers insights into the ensemble model's ability to generalize to new, unseen data, highlighting the effectiveness of the stacking approach.Evaluating each base model individually gives you a benchmark of their performance. This is useful for understanding if the ensemble model is actually improving upon the base models' predictions or if it's merely averaging their performance.



```
Random Forest - Test Accuracy: 0.9874, Kappa: 0.9206
Gradient Boosting - Test Accuracy: 0.9589, Kappa: 0.6943
K-Nearest Neighbors - Test Accuracy: 0.9785, Kappa: 0.8671
Decision Tree - Test Accuracy: 0.9884, Kappa: 0.9292
```

Figure 3- Accuracy and Kappa Score of Base Models



```
Ensemble model - Accuracy: 0.9873, Kappa: 0.9218
```

Figure 4- Ensemble Model Accuracy and Kappa Score

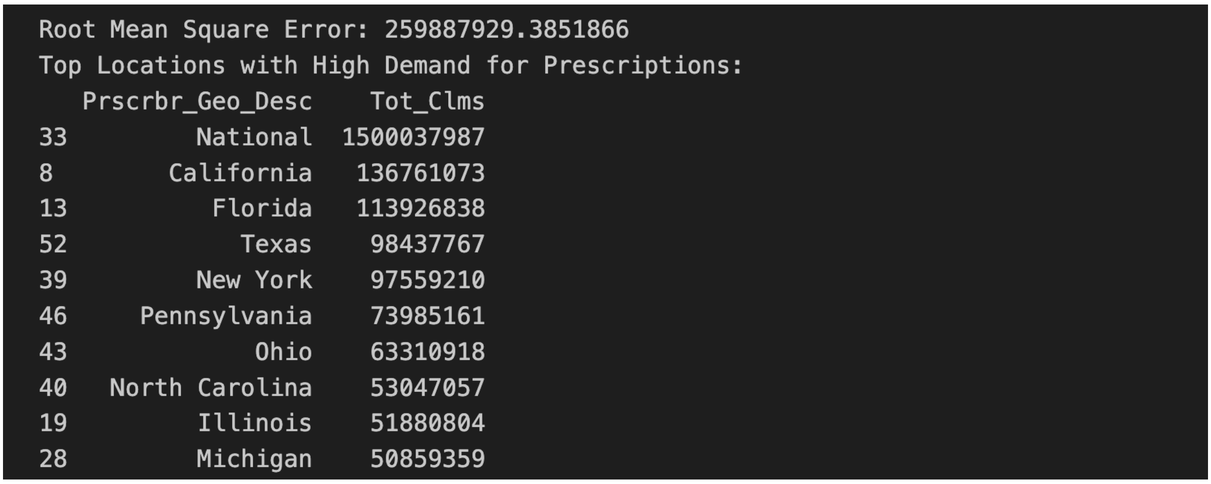
In conclusion this project exemplifies the power of ensemble learning, specifically through a stacking methodology, in tackling complex classification tasks. By combining the predictions of multiple base models through a meta model, the ensemble approach seeks to capitalize on the diverse perspectives and strengths of different algorithms. The preprocessing and transformation steps, including encoding, standardization, and PCA, prepare the dataset for effective learning. The evaluation of both the base models and the ensemble model provides a comprehensive understanding of how each component contributes to the overall performance. Through this project, we demonstrate that ensemble models can achieve superior predictive performance compared to individual models, offering a robust solution to the challenging task of drug classification based on prescription data.

4.Application

The proposed model outlines a comprehensive approach to building a predictive model for classifying drugs into different categories based on their characteristics and usage indicators. The application of machine learning techniques, including preprocessing, dimensionality reduction, multiple classification algorithms, and ensemble modeling, makes this model serve several valuable purposes in healthcare and pharmaceutical sectors. Here are some potential applications:

1. Drug Utilization Research: By classifying drugs into categories for example, in this model as opioids, antibiotics, and antipsychotics, one can analyze patterns in drug utilization. This help in understanding the prescribing trends, identifying potential overuse or misuse of certain drug classes, and developing strategies for more effective and safer drug use.

2.Pharmacy Inventory Management: Pharmacies can use this model to optimize their inventory levels based on the classification of drugs. Understanding drug categories that are more in demand in specific locations can help in better stock management, ensuring the availability of essential medications while reducing the wastage of less frequently used drugs.



```
Root Mean Square Error: 259887929.3851866
Top Locations with High Demand for Prescriptions:
   Prscrbr_Geo_Desc    Tot_Clms
33         National  1500037987
8        California   136761073
13          Florida   113926838
52            Texas    98437767
39         New York    97559210
46     Pennsylvania    73985161
43             Ohio    63310918
40   North Carolina    53047057
19         Illinois    51880804
28         Michigan    50859359
```

Figure 5 - Top Locations with High Demand for Prescriptions

3. Clinical Decision Support Systems (CDSS): Integrating this model into CDSS can assist healthcare providers in selecting the most appropriate medication for their patients. By accurately classifying drugs, the system can offer recommendations that are aligned with clinical guidelines and patient-specific factors, enhancing the quality of care.
4. Pharmaceutical Marketing and Sales Strategy: Pharmaceutical companies can utilize insights from this model to tailor their marketing and sales strategies. By understanding the classifications of drugs that are most prescribed or have the highest demand, companies can focus their efforts on developing and promoting medications that meet the current needs of the healthcare sector.
5. Public Health Policy Making: Policymakers can use the data generated by this model to inform public health policies related to drug use. For example, identifying trends in opioid prescription can guide the development of policies aimed at preventing opioid misuse and addiction.
6. Educational Tools for Healthcare Professionals: This project can serve as an educational resource for healthcare professionals, helping them to better understand the drug classification system. It could be incorporated into continuing education programs to update professionals on drug classifications and their implications for patient care.
7. Patient Education and Engagement: By simplifying complex drug information into understandable classifications, this model can be used to develop educational materials for patients. This can empower patients to be more informed about their medications, leading to better adherence to treatment plans and improved health outcomes.

8. Regulatory Compliance and Monitoring: Regulatory bodies can use the classification model to monitor drug prescriptions and distributions more effectively, ensuring compliance with regulations related to drug safety and use. This can help in identifying irregularities and preventing the illicit distribution of controlled substances.

These applications demonstrate the broad potential impact of your project, from improving clinical outcomes and patient safety to informing public health policies and enhancing pharmaceutical management practices.

5.Conclusion

This comprehensive approach to drug classification through advanced machine learning techniques, including preprocessing, dimensionality reduction, ensemble modeling, and the evaluation of multiple classifiers, showcases its potential to significantly impact the healthcare and pharmaceutical industries. By effectively categorizing drugs into specific classes such as opioids, antibiotics, antipsychotics, and others, the project not only aids in understanding prescribing patterns and drug utilization but also serves as a foundational tool for a multitude of applications ranging from pharmacy inventory management to clinical decision support, and public health policy formulation.

The successful implementation and validation of the model, demonstrated by its accuracy and Cohen's kappa scores, underline the viability of using machine learning for complex classification tasks in the medical field. Furthermore, the ensemble model's superior performance, achieved by combining predictions from several base models, highlights the value of leveraging diverse algorithms to enhance predictive capabilities.

This project also illuminates the importance of data preparation and feature engineering in building effective machine learning models. By encoding categorical variables and applying dimensionality reduction through PCA, it demonstrates how to manage high-dimensional data and extract meaningful patterns, which is crucial in the context of large and complex healthcare datasets.

Moreover, the project underscores the dynamic capabilities of machine learning in adapting to the intricacies of drug classification, providing a scalable solution that can accommodate new data and evolving classification needs. This adaptability is essential in the rapidly changing landscape of healthcare and pharmaceuticals, where new drugs are continuously introduced, and prescribing guidelines are frequently updated.

In conclusion, this project represents a significant step forward in the application of machine learning in healthcare. It offers a robust framework for drug classification that can support a wide range of stakeholders, from healthcare providers and pharmacists to policymakers and pharmaceutical companies. By enhancing understanding, improving decision-making, and facilitating more efficient and effective drug management, this project has the potential to contribute to better health outcomes and more rational drug use across the healthcare system. The methodologies and insights derived from this work can serve as a foundation for further research and development in the field, encouraging the integration of machine learning tools into various aspects of healthcare and pharmaceutical management.

6.References

[1] Balogh EP, Miller BT, Ball JR. Improving diagnosis in health care. Washington, DC: National Academies Press (US); 2015.

[2]Kumari, S.; Kumar, D.; Mittal, M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *Int. J. Cogn. Comput. Eng.* **2021**, *2*, 40–46.

[3]Latha, C.B.C.; Jeeva, S.C. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Inform. Med. Unlocked* **2019**, *16*, 100203.
4.Igodan, E.C.; Thompson, A.F.-B.; Obe, O.; Owolafe, O. Erythemato Squamous Disease Prediction using Ensemble Multi-Feature Selection Approach. *Int. J. Comput. Sci. Inf. Secur. IJCSIS* **2022**, *20*, 95–106.

[5]Subanya B, Rajalaxmi R. Feature selection using Artificial Bee Colony for cardiovascular disease classification. 2014 International Conference on Electronics and Communication Systems (ICECS). Coimbatore: IEEE; 2014. p. 1–6.

[6]GHazanfari M, Alizadeh S, Teimourpour B. Data mining knowledge discovery. Tehran: Iran University of Science and Technology; 2014.

[7]Kondababu A, Siddhartha V, Kumar BB, Penumutchi B. A comparative study on machine learning based heart disease prediction. In: Materials Today: Proceedings; 2021.

[8]Jeyaranjani J, Rajkumar TD, Kumar TA. Coronary heart disease diagnosis using the efficient ANN model. In: Materials Today: Proceedings; 2021.

[9]Maini E, Venkateswarlu B, Maini B, Marwaha D. Machine learning–based heart disease prediction system for Indian population: an exploratory study done in South India. Med J Armed Forces India. 2021;77(3):302–11.

[10]Kumari S, Kumar D, Mittal M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. Int J Cogn Comput Eng. 2021;2:40–6.

[11]Pavithra V, Jayalakshmi V. Hybrid feature selection technique for prediction of cardiovascular diseases. In: Materials Today: Proceedings; 2021.

[12]Arumugam K, Naved M, Shinde PP, Leiva-Chauca O, Huaman-Osorio A, Gonzales-Yanac T. Multiple disease prediction using machine learning algorithms. In: Materials Today: Proceedings; 2021.

[13]Jain B, Ranawat N, Chittora P, Chakrabarti P, Poddar S. A machine learning perspective: to analyze diabetes. In: Materials Today: Proceedings; 2021.

[14]Jothi KA, Subburam S, Umadevi V, Hemavathy K. Heart disease prediction system using machine learning. In: Materials Today: Proceedings; 2021.

[15]Wei X, Lu Q, Jin S, Li F, Zhao Q, Cui Y, et al. Developing and validating a prediction model for lymphedema detection in breast cancer survivors. Eur J Oncol Nurs. 2021;54:102023.

[16] Ramesh G, Madhavi K, Reddy PDK, Somasekar J, Tan J. Improving the accuracy of heart attack risk prediction based on information gain feature selection technique. In: Materials Today: Proceedings; 2021.

[17]Khaleel FA, Al-Bakry AM. Diagnosis of diabetes using machine learning algorithms. In: Materials Today: Proceedings; 2021.

[18]Kumar S, Sahoo G. Classification of heart disease using naive bayes and genetic algorithm. In: Computational intelligence in data mining-volume 2: Springer; 2015. p. 269–82.

[19]Pinto A, Ferreira D, Neto C, Abelha A, Machado J. Data mining to predict early stage chronic kidney disease. Procedia Comput Sci. 2020;177:562–7.

[20]Teimouri M, Farzadfar F, Alamdari MS, Hashemi-Meshkini A, Alamdari PA, Rezaei-Darzi E, et al. Detecting diseases in medical prescriptions using data mining tools and combining techniques. Iran J Pharm Res. 2016;15(Suppl):113.
[21]Dhanya R, Paul IR, Akula SS, Sivakumar M, Nair JJ. F-test feature selection in stacking ensemble model for breast cancer prediction. Procedia Comput Sci. 2020;171:1561–70.

[22]Malladi R, Vempaty P, Pogaku V. Advanced machine learning based approach for prediction of skin cancer. In: Materials Today: Proceedings; 2021.

[23]Onan A. A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. Expert Syst Appl. 2015;42(20):6844–52.

[24]Dehkordi SK, Sajedi H. Prediction of disease based on prescription using data mining methods. Heal Technol. 2019;9(1):37–44.

[25]Ferdowsy F, Rahi KSA, Jabiullah MI, Habib MT. A machine learning approach for obesity risk prediction. Curr Res Behav Sci. 2021;2:100053.

[26]Han J, Pei J, Kamber M. Data mining: concepts and techniques. 3rd ed: The Morgan Kaufmann Series in Data Management Systems; 2011.

[27]Sulzmann JN, F¨urnkranz J. Rule stacking: an approach for compressing an ensemble of rule sets into a single classifier. In: International conference on discovery science. Heidelberg: Springer; 2011. p. 323–34.