Optimizing Medicine Recommendation Systems: A Comparative Analysis of SVM, XGBoost and Multinomial NB Models

AVANTIKA SINGH[1], KOMAL SAXENA[2]
AIIT, Amity University, Noida U. P, INDIA

Abstract: - This paper explores the establishment and testing of a healthcare recommendation system for personalized medication suggestions. It employs advanced machine learning methods on various patient profiles, drug prescriptions, and medical conditions using UCI ML repository dataset. A rigorous data collection process is conducted, cleaning and exploratory analysis are performed to elicit insights for model building. Three main models – SVM, XGBoost, Multinomial NB were compared in terms of their performance in medication recommendation tasks where XGBoost performed better than all other models. The research highlights the need for large-scale datasets and more complex algorithms in improving patient care optimization. Additionally, it points out some directions for future work such as feature integration and model refinement to increase adaptability across different clinical settings.

Keywords: - Support Vector Machine (SVM), XGBoost, Multinomial Naive Bayes (NB), Exploratory data analysis (EDA)

Received: March 15, 2024. Revised: August 16, 2024. Accepted: September 19, 2024. Published: October 31, 2024.

1. Introduction

The study "XGBRS Framework Integrated with Word2Vec Sentiment Analysis for Augmented Drug Recommendation" presents a system of recommending drugs which lacks inclusion of deep learning, scalability consideration imbalanced data handling and interpretability of XGBoost making it only applicable within pharmaceuticals. [1]
The paper, "A Catalogue of Machine Learning Algorithms for Healthcare Risk Predictions," presents a healthcare risk prediction framework with seven algorithms; however, it does not include XGBoost nor scalability. The absence of cloud deployment and federated learning restricts real-world applicability which means that there is need for future research. [2]In the article "Comparative Analysis of Naive Bayesian Techniques in Health-Related for Classification Tas" The paper evaluates healthcare ML techniques but misses advanced methods like XGBoost and scenario evaluation. It advocates for comprehensive methodologies like CRISP-DM, contrasting with other studies. [3]The research paper named "Application of Support Vector Machine for Prediction of Medication Adherence in Heart Failure Patients" predicts heart failure medication adherence using SVM (77.63% accuracy), yet faces limitations in sample size and study design, urging improvements for practical use. [4]The document entitled "Support Vector Machines (SVMS) Based Advanced Healthcare System Using Machine Learning Techniques" emphasizes SVMs and machine learning in healthcare but overlooks scalability, interpretability, and ethics, urging improvements for responsible integration and better patient outcomes. [5]. The article named "Comparative Performance Evaluation of Classification Algorithms for Clinical Decision Support Systems" by the author compares ML algorithms in clinical settings for disease prediction without considering diverse contexts. In order to make the algorithms more applicable, more studies should be carried out. [6]. The paper titled "Performance Evaluation of Different Machine Learning Classification Algorithms for Disease Diagnosis" examines ML algorithms for diagnosing diseases and shows their efficacy but identifies such difficulties as having unfair datasets. It calls for standardized approaches to improve consistency and accuracy of the algorithm thereby enhancing patient care. [7]
The document "COVID-19 Risk Prediction For Diabetic Patients Using Fuzzy Inference System And Machine Learning Approaches" presents a model that predicts risk levels among diabetics suffering from coronavirus; however it does not take into account demographic factors while favouring CatBoost. To ensure strongness during pandemics this must plug gaps found during research stages which also helps direct clinical decisions. [8]

2. Literature Review

Digitalizing has entirely transformed the availability of information in relation to health. These large volumes of data on health can be hard to understand simply because of medical terminologies or know which sources are credible especially with them being accessible all the time. Therefore, this article proposes that we come up with a Health Recommendation System (HRS) that will help individuals find trustworthy information about wellness. "In this context, the recommender systems may provide the patients with extra laymen-friendly details helping to better comprehend their health condition as represented by their past records. However, such systems must be adapted to cope with the certain requirements in the health sector to provide highly relevant information for patients. These are known as health recommender systems (HRS)." [9] This study employs machine learning algorithms to analyze data from UCI Machine Learning repository which is a collection of databases and domain theories. [10] The data that is tested and trained is comes from a research paper titled "Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning," presented at the 2018 International Conference on Digital Health. [11]Machine learning models, such as Support Vector Machine (SVM), Multinomial Naive Bayes (NB), and XGBoost, are employed to analyse the data and make predictions about the dataset. [12] The model training on the given dataset helps to know the health-related patterns, that helps the HRS to provide personalized recommendations to the users. By examining the reviews of the users as well as integrating the machine learning, the HRS aims to revive the process of health information retrieval. This paper analyses various methodologies, challenges and insights to advance the process of personalized healthcare information retrieval, enabling the individuals to make informed decisions among the vast online health landscape.

3. Problem Statement

The global nature of healthcare seeking behaviour has changed significantly over time due to evolving landscape of health information on internet. Henceforth, a Health Recommendation System (HRS) using SVMs, Multinomial_NB, XGBoost models integrated with web-based rating tools having mining feature of reviews. The main aim is recommending suitable doctors and drugs both for individual users also beneficial to e-commerce platforms like Amazon. The aim is adapt to the unique demands of health sector along with the effective addressal of users challenges with medical vocabulary and distinct information sources.

4. Methodology

4.1 Data Collection

Data collection lays the groundwork for analysis by providing a basis for research. In this section we describe the methodology used to gather secondary data for our dissertation project; this included extracting from "Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning" (Felix Gräßer, Surya Kallumadi, Hagen Malberg, Sebastian Zaunseder – "International Conference on Digital Health 2018". [11] UC Irvine Machine Learning Repository was used as a trustworthy source because it provides reliable material. [10] When electronic databases are involved secondary sources were selected based on relevance and authenticity.

4.2 Pre-processing and Data Cleaning

To form a health recommendation system, the data needs to be processed carefully so that it can be used for modelling properly. [13].The initial step is dividing the dataset into four parts: 'review', 'condition', 'rating' and 'useful Count' as independent variables (X), and drug Name as dependent variable (y). After that, we should encode categorical variables and split the data into training set and testing set by Label Encoder and train_test_split respectively. Then, we further refine the data through transformations such as scaling ('rating' and 'useful Count'), TF-IDF vectorization ('review') and Count Vectorizer ('condition'). These steps make sure that all the information are ready for analysis consistently. Besides this, evaluate_model function has been created which helps us to thoroughly assess our classification models allowing comprehensive comparison within research framework.

4.3 Exploratory Data Analysis

In this research paper, an exploratory data analysis (EDA) [14]was conducted to understand patterns, trends, and relationships in the dataset. It involved descriptive statistics, histograms, scatter plots, and box plots to uncover insights.

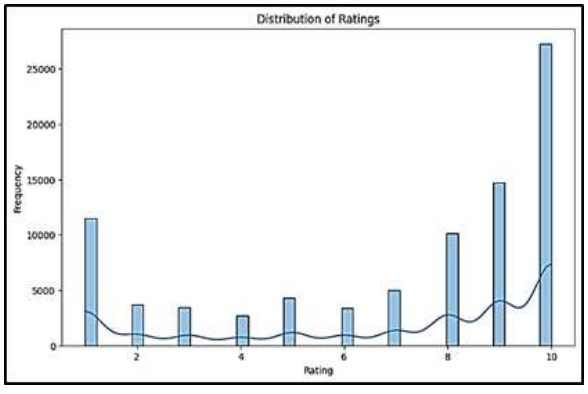

Figure 1: Distribution of Ratings

The graph illustrates the spread of rating values. On the x-axis, you see the rating value, ranging from 0 to 10, while the y-axis represents the frequency, or how many ratings each value has received. It appears that the distribution is skewed positively, with a peak at the higher end (ratings of 8 or 10), indicating that there are more positive reviews than negative ones.

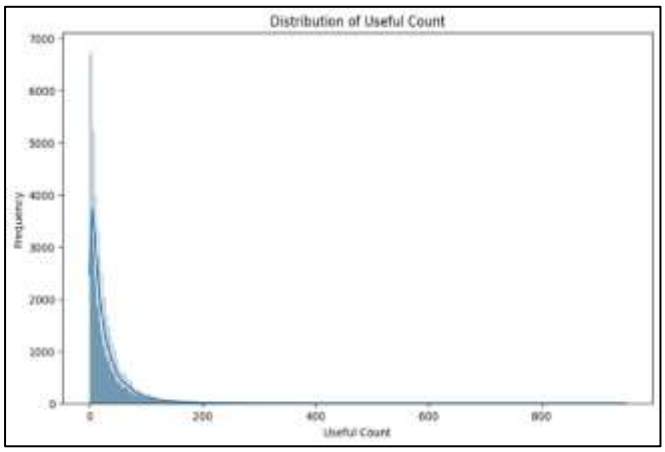

Figure 2: Distribution of Useful Count

This graph shows the distribution of values for "useful count." The y-axis represents frequency, or how many times each value appears on the x-axis. Notice that the right tail of the distribution is longer than the left - there are many more high "useful count" values (around 800-1200) than lower ones (around 0-200).

Figure 4: Word Cloud Conditions

This word cloud represents user queries related to birth control by showing how frequently certain words are mentioned in relation to it. [15].For example anxiety depression migraines were frequently mentioned alongside birth control indicating that people want to know more about hormonal contraception's impact on mental health states. Among other things it also indicates diverse user needs (weight control).

Figure 5: Word Cloud for Reviews

The analysis highlights upon "side effects" and mental health impacts like 'anxiety' and 'depression'. Also, the terms like 'emergency

contraception' and 'IUD", as shown in word cloud, is an indication to contraceptive options. However, many of the concerns expressed were about the side effects of different contraceptives on mental health, physical wellness, as well as various methods themselves; this does not give any sentiment analysis or contextual information about word relationships in the word cloud image shown here.

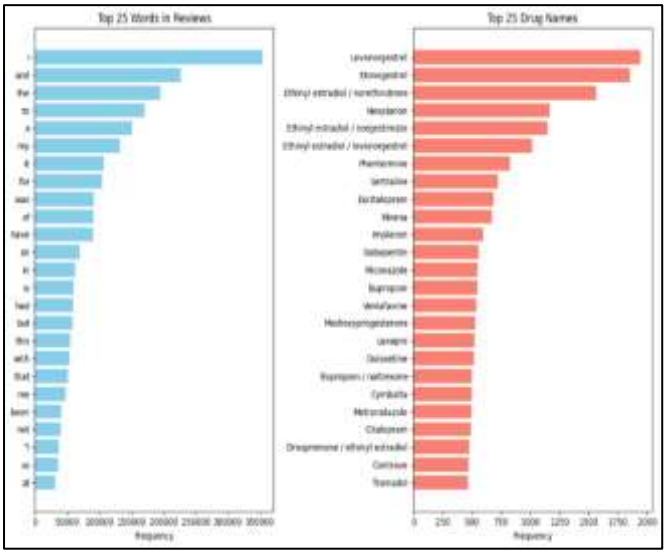

Figure 6: Top 25 in Reviews and Top 25 Drug Names

The graph shows frequency of words in review data where "estrogen" and "progesterone" are highlighted as major terms indicating focus on hormonal birth control; additionally, brand names such as Mirena or Implanon representing hormonal IUDs stand out but it lacks sentiment analysis and specific context about drugs.

5. Model Training

5.1 Multinomial Naive Bayes

We developed Multinomial Naive Bayes (Multinomial_NB) classifier as part of our machine learning toolkit for building an effective healthcare recommendation system. [16] We implemented scikit-learn's Multinomial_NB classifier for text classification within a pipeline designed to smoothen data processing. Model training involved tuning parameters via data splitting in order to achieve the best predictive performance possible based on evaluation results that also showed how well this model can generalize over new datasets — an important aspect for real world applications.

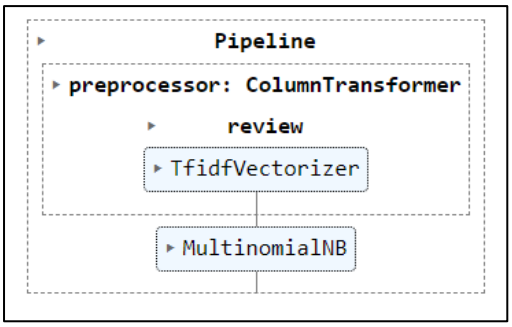

Figure 7: Multinomial Naive Bayes pipeline

To get text classification done, we followed a two-step machine learning pipeline. To begin with, scikit-learn's Column Transformer and TfidfVectorizer were used in preprocessing the training data (X_train) into numerical TF-IDF representations. Thereafter, we classified by using multinomial Naive Bayes classifier (MultinomialNB). pipe1.fit(X train, y train) was used to fit the pipeline which enabled us to make predictions on new data (X_new) using pipe1.predict(X_new).

For this study's model training stage, scikit-learn in Python was used to create an SVM (Support Vector Machine) algorithm. [17] . We designed a Linear Support Vector Classification (LinearSVC) with specific parameters including squared hinge loss function and fixed random state for reproducibility. Scikit-learn's Pipeline class combines pre-processing steps seamlessly with SVM models for improved efficiency during training as well as workflow simplification purposes.

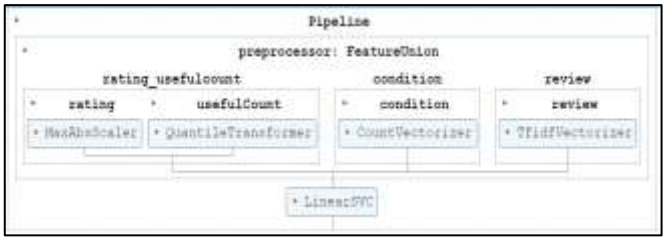

Figure 8: Schematic pipeline process showing text processing steps including Feature Union, MaxAbsScaler, QuantileTransformer, Count Vectorizer, TfidfVectorizer, and LinearSVC

The picture represents a process of sentiment analysis conducted through pipeline fitting via pipe2.fit (X_train, y_train). It pre-processes text data using methods like Count Vectorizer and may scale features with MaxAbsScaler. The given review classification task was performed by LinearSVC algorithm implemented within this pipeline which had been trained on provided data for further analysis.

The training data showed promising results with the support vector machine (SVM) pipeline (pipe2) after the model was trained and evaluated. An accuracy of 0.848 was achieved while precision, recall and F1 scores were around 0.85 Similarly, on testing data, accuracy fell sharply to 0.292 which means it might be overfitting — but also lower precision, recall and f1 score all indicate poor performance on new data too; Moreover, there is an Undefined Metric Warning that suggests imbalanced or problematic modelling may need investigation. Such findings can help guide future iterations of this model better so they don't fail again like they did her.

5.2 XGBOOST

In order to get ready, the dataset for use with the XGBoost Algorithm, a number of steps were taken. [18]. First, it had to be determined how many unique drug names were included in our dataset (n). After that information was pre-processed so that it can be converted into DMatrix for XGBoost which is an effective learning format. For training models on all CPU cores quickly, various XGBoost parameters were set such as 'multi: SoftMax' which is used when working on multi-class classification problems and making predictions using all available central processing units simultaneously. A round began by training a model for some number of rounds (num rounds) where predictions are iteratively refined until there is no improvement anymore. But even after training much and tuning parameters; performance evaluation gave low accuracy scores like precision, recall or F1 score showing overfitting i.e., when a model performs extremely well on training data but fails to generalize on unseen records. This means that perhaps there are intricate relationships within our data sets thus we should look further into it so as to enhance its representational power for other similar cases outside those already seen during training.

6. Model Compression

	Accuracy_Train	Accuracy_Test
Multinomial_NB	0.108231	0.094090
SVM	0.848056	0.292121
XGB	0.025738	0.026188

Figure 9: Machine Learning Model Comparison: Training vs. Test Accuracy (Multinomial_NB, SVM, XGB)

To assess the impact of model compression [19] ,efficient performance tracking is crucial. In this paper,we have demonstrated it through Pandas DataFrames. The code constructs a dictionary (Accuracy_d) which is used to store training and test accuracies, then converts it into a DataFrame (Accuracy). By doing so, it makes an easy comparison of pre and post compression metrics, which is important for evaluating how effective the compression techniques are, in terms of maintaining the model feneralizability with minimum performance degradation

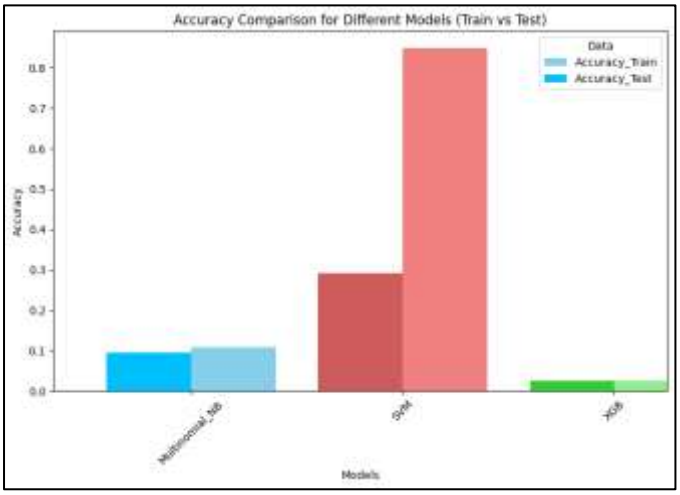

Figure 10: Accuracy Comparison for Different Models (Train vs Test)

The above image illustrates the pipeline analysis for the text data, which begins with the features such as "rating_usefulcount" and "review" which are extracted from product reviews. To clean and normalize the data, we have employed a pre-processing sub-pipeline. Techniques like “CountVectorizer" or "TfidfVectorizer." Converts the text feature into numerical features. Ultimately, a machine learning model such as "LinearSVC" conducts sentiment classification.

.

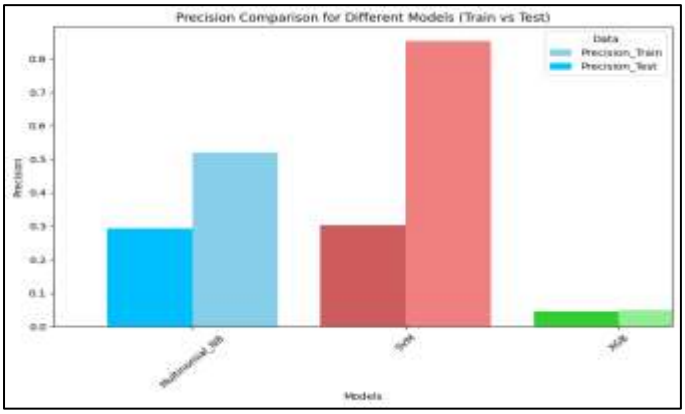

Figure 10: Precision Comparison for Different Models (Train vs Test)

The above image demonstrates a analysis pipeline which has various pre-processing stages, which includes normalization and feature extraction techniques such as "Count Vectorizer" and "TfidfVectorizer". Thereafter, a machine learning model named as “LinearSVC” is trained on the processed data. The accompanying code snippet explains the training pipeline on the training dataset, that enables the sentiment analysis on new reviews.

7. Conclusion

In conclusion, this paper employs various data analysis and demonstrates the models and their methodologies. The paper includes data collection, exploratory analysis which demonstrates the data with help of various graphical figures, then the data is pre-processed and thereafter trained with Multinomial Naive Bayes, Support Vector Machine (SVM), and XGBoost. It is seen that among the three algorithms used, Multinomial Naïve Bayes is most precise, but overfitting suggests a need of regularization techniques. Then comes SVM which demonstrates moderate precision and stability after Multinomial Naïve Bayes, making it a more practical choice. After that, the XGBoost, which shows the lowest precision in the medicine recommendation system. Thus, through this analysis, the paper provides insights and solutions, by evaluating the effectiveness of the algorithms, also it discusses the importance of model compression for the practical implementation in a limited resource environment.

.

References

[1] S. Paliwal, A. K. Mishra, R. K. Mishra, N. Nawaz and M. Senthilkumar, “XGBRS Framework Integrated with Word2Vec Sentiment Analysis for,” Tech Science Press, p. 18, 2022.
[2] A. K.,. S. K.,. K. M. Argyro Mavrogiorgou, “A Catalogue of Machine Learning Algorithms for Healthcare,” sensors, p. 45, 2022.
[3] N. H. S. B. Marzuki Ismail, “Comparative Analysis of Naive Bayesian Techniques in,” JOURNAL OF SOFT COMPUTING AND DATA MINING, Vols. VOL.1 NO. 2 (2020) 1-10, 2020.
[4] H.-G. K.-H. K. S. C. S.-K. L. Youn-Jung Son, “Application of Support Vector Machine for Prediction of Medication Adherence in Heart Failure Patients,” Healthcare Informatics Research.
[5] T. K. S. A. A. Manju Payal, “Support Vector Machines (SVMS) Based Advanced Healthcare System Using Machine Learning Techniques,” Researchgate, 2022.
[6] S. L. Bayu Adhi Tama, “A Comparative Performance Evaluation of Classification Algorithms for Clinical Decision Support Systems,” MDPI, 2020.
[7] A. M. A. Q. Munder Abdulatef Al-Hashem, “Performance Evaluation of Different,” International Journal of E-Health and Medical Communications, vol. 12, no. 6, 2021.
[8] M. C. M. S. B. K. S. K. G. S. H. A. A.-D. Alok Aggarwal, “COVID-19 Risk Prediction for Diabetic Patients Using Fuzzy Inference System and Machine Learning Approaches,” Hindawi Journal of Healthcare Engineering, vol. 2022, p. 10, 2022.
[9] D. P. Martin Wiesner, “Health Recommender Systems: Concepts, Requirements, Technical Basics and Challenges,” International Journal of Environmental Research and Public Health, p. 28.
[10] “UC Irvine Machine Learning Repository,” [Online]. Available: https://archive.ics.uci.edu/.
[11] S. K. H. M. a. S. Z. Felix Gräßer, “Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning,” in International Conference on Digital Health, New York, 2018.
[12]N. Pavanam, “www.analyticsvidhya.com,” [Online]. Available: https://ww.analyticsvidhya.com/blog/2020/11/understanding-naive-bayes-svm-and-its-implementation-on-spam-sms/.
[13] S. V. Dorpe, “Preprocessing with sklearn: a complete and comprehensive guide,” [Online]. Available: https://towardsdatascience.com/preprocessing-with-sklearn-a-complete-and-comprehensive-guide-670cb98fcfb9.
[14] “Kaggle,” [Online]. Available: https://www.kaggle.com/code/imoore/intro-to-exploratory-data-analysis-eda-in-python.
[15] “Word Clouds and Qualitative Data Analysis,” [Online]. Available: https://sambodhi.co.in/word-clouds-and-qualitative-data-analysis/.
[16]C. d. u. t. M. N. B. algorithm. [Online]. Available: https://developer.ibm.com/tutorials/awb-classifying-data-multinomial-naive-bayes-algorithm/.
[17] M. H. P. Himani Bhavsar, “A Review on Support Vector Machine for Data Classification,” International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 1, no. 10.

[18] "Practical Applications of XGBoost in Data Science," [Online]. Available: https://medium.com/@harshitaaswani2002/practical-applications-of-xgboost-in-data-science-72e34992326.
[19] "Efficient Deep Learning: Unleashing the Power of Model Compression," [Online]. Available: https://towardsdatascience.com/efficient-deep-learning-unleashing-the-power-of-model-compression-7b5ea37d4d06.
[20] C. d. u. t. M. N. B. algorithm. [Online]. Available: https://developer.ibm.com/tutorials/awb-classifying-data-multinomial-naive-bayes-algorithm/