Drug Recommendation System Using Machine Learning Technique

Prakash S. Chougule [1], Tejaswi S.Kurane[2] ,R.D.Shinde[3] ,G.A.Gadhari[4] ,Mrs.V..C.Shinde[5] ,Prashant S.Chavan[6]

[1] Associate Professor, Rajarshi Chhatrapati Shahu College, Kolhapur(MS), India

[2,4]Assistant Professor, Rajarshi Chhatrapati Shahu College, Kolhapur (MS), India

[3]Assistant Professor, Jaysingpur College, Jaysingpur (MS), India

[4] Associate Professor, Vivekanand College, Kolhapur(MS), India

[5] Research Student, Rajarshi Chhatrapati Shahu College, Kolhapur (MS), India

Abstract: In today's digital era healthcare is one among the major core areas of the medical domain. People trying to find suitable health-related information that they are concerned with. The Internet could be a great resource for this kind of data, however you need to take care to avoid getting harmful information. In medical emergencies, rapid and precise drug recommendations are crucial for patient survival and effective treatment. A drug Recommendation System in machine learning (ML) is a software application designed to assist healthcare professionals and patients in selecting the most appropriate medication. The system processes comprehensive patient data, including medical histories and real-time health indicators, to provide accurate drug recommendations In this article our main aims to introduce a machine learning-based drug recommendation system based on four different methods like Random Forest, SVM, Passive-Aggressive and Logistic Regression (Over) classifiers. Which are analyses the reviews and predict conditions based on these conditions recommended relevant top five drugs. Our study shows that random forest method is superior as compared to logistic regression, passive aggressive, support vector machine. Passive aggressive method is give excellent result as compared to support vector and logistic regression. Support vector machine is better as compared to logistic regression. In overall comparison the performance of logistic regression is worst.

Keywords: Drug recommendations system, machine learning, text analysis, TF-IDF, Random Forest, Logistic Regression, SVM, Passive-Aggressive.

Introduction: The rise in coronavirus cases is rapid, leading to a shortage of doctors. There are fewer doctors in rural areas compared to cities.[1] Since medical school takes between 6 and 12 years to complete, this issue has worsened, making it hard to bring in more doctors. A telemedicine framework should be fully utilised during this challenging period.[1]The frequency of clinical mistakes has increased in modern healthcare environments. Medication-related errors affect more than 200,000 people in China and 100,000 in the US annually. Over 40% of doctors, especially experts, make mistakes while writing prescriptions. This occurs because their decisions are based on a limited scope of knowledge, which influences the solutions they provide [3][4]. For patients who need professionals with an in-depth understanding of microscopic organisms, antibacterial agents, and patient care, choosing the best prescription is essential [2]. Due to new research, clinical personnel can access more medications and diagnostics daily. As a result, doctors find it more and more challenging to choose the best course of action or drugs for patients based on their symptoms and medical background.

Computer applications have significantly improved with the introduction of artificial intelligence (AI). Emulation of human cognitive processes in digital systems is at the heart of artificial intelligence. The development of AI is mainly dependent on machine learning methods, which include data collection, rules for information extraction, the production of both approximations and exact conclusions, and the confirmation of these results. The effectiveness of artificial intelligence is largely contingent upon the precision of the algorithms employed in machine learning. The precision of machine learning algorithms is predominantly contingent upon substantial training datasets. In contemporary times, a vast amount of data is available for training systems. The integration of artificial intelligence into the drug development process has increased. Currently, AI is pivotal in the analysis and advancement of drug discovery. Pharmaceutical companies, research and development institutions focused on AI, and medical professionals can collaborate to investigate new medicinal solutions tailored to specific needs. Clinical trials and medical techniques are typically used to assess the safety of drugs. However, based on their personal experiences, patients—significant stakeholders—can provide insightful

information on how their drugs impact [15]. To express their ideas and opinions regarding their [14] health, physicians, and drugs, many people use social media and online health forums such as RateMDs, WebMD, Ask a Patient, DrugLib.com, Drugs.com, MedHelp, and Daily Strength. The online reviews from patients give direct insights into the effectiveness of drugs and their side effects. These reviews can be negative or positive, depending on how the medication impacts their health. Analysing these patient reviews is essential for healthcare professionals and drug developers, as it offers critical information that can significantly affect healthcare practices.

A recommender framework is a standard system that provides consumers with recommendations based on their requirements and preferences. These systems use user feedback to understand customer mood and provide tailored suggestions. Based on patient assessments, a medicine recommender system recommends medications for specific conditions using sentiment analysis and feature engineering. Sentiment analysis employs tools and techniques to find and extract emotional information, including opinions and attitudes, from texts [5]. However, by creating new features from old ones, feature engineering enhances model performance.There are five sections in this research paper. The significance of this work is briefly summarised in the introduction. A summary of earlier studies in this area may be found in the related works section. The techniques employed in this investigation are described in the methodology section. Several metrics are used in the results section to evaluate the implemented model's results. The framework's limits are described in the following section of the conclusion.

Literature Review: One of the most important sectors is healthcare, where mistakes could have fatal results. Medical errors, which occur when doctors prescribe the incorrect drug based on their limited experience, are one of the major problems facing the healthcare industry. An estimated 250,000 people die in the US each year as a result of medical errors, which include a broad spectrum of healthcare failures from surgical blunders to incorrect pharmaceutical prescriptions (National Academies of Sciences, Engineering, and Medicine). Notably, Anderson and Abrahamson (2017) discussed how medical errors significantly impact patient safety and stressed the significance of implementing policies and procedures to lower these risks. The most frequent of these errors is prescribing the incorrect drug, which highlights the urgency of developing tools such as our drug recommendation system to assist healthcare providers in making better choices [6].Their research aimed to propose supervised machine learning models capable of accurately classifying pharmaceutical reviews into two distinct categories, positive and negative, utilising text reviews as input. Using the Term Frequency-Inverse Document Frequency (TF-IDF), they developed machine-learning models specifically tailored to address five common factors: acne, high blood pressure, pain, depression, and birth control use.Bow, TF-IDF, Word2Vec, and Manual Feature Analysis were used for vectorisation [7]. With a sharp increase in AI advancement, there has been an increase in the application of machine learning strategies to recommender frameworks. These days, recommended frameworks are common in the travel industry, e-commerce, restaurants, etc.

Unfortunately, few studies in the drug proposal framework utilising sentiment analysis are available. The medication reviews are more intricate to analyse as they incorporate clinical wordings like infection names, reactions, and synthetic names used in drug production [8]. Leilei Sun [9] examined large-scale treatment records for the best prescription. The idea was to use an efficient semantic clustering algorithm to estimate the similarities between treatment records. Similarly, it established a framework to evaluate the adequacy of the proposed treatment. This structure can recommend the best treatment regimens for new patients based on their demographic locations and medical complications. An Electronic Medical Record (EMR) of patients gathered from numerous clinics for testing. The result shows that this framework improves the cure rate. Xiaohong Jiang et al. [10] examined three distinct algorithms: a decision tree algorithm, a support vector machine (SVM), and a neural network on treatment data. SVM was picked for the medication proposal module as it performed well in the three unique boundaries - model exactness, proficiency, and versatility. The mistake check system was also proposed to ensure analysis, precision and administration quality. In this article we comparative study of drug recommendation system based on four methods of classifications like Logistic Regression, Passive

Aggressive, SVM and Random Forest.We use Jupiter code in python to identify the best model and based on this model we create drug recommendation search engine.

Methodology: The dataset used in this research is the Drug Review Dataset (Drugs.com), taken from the UCI ML repository [11]. This dataset contains six attributes: name of drug used (text), review (text) of a patient, condition (text) of a patient, helpful count (numerical), which suggests the number of individuals who found the review useful, date (date) of review entry, and a 10-star patient rating (numerical) determining overall patient contentment. It contains a total of 215063 instances.

Fig. 1 shows the proposed model used to build a medicine recommender system. It contains four stages: Data preparation, classification, evaluation, and Recommendation.

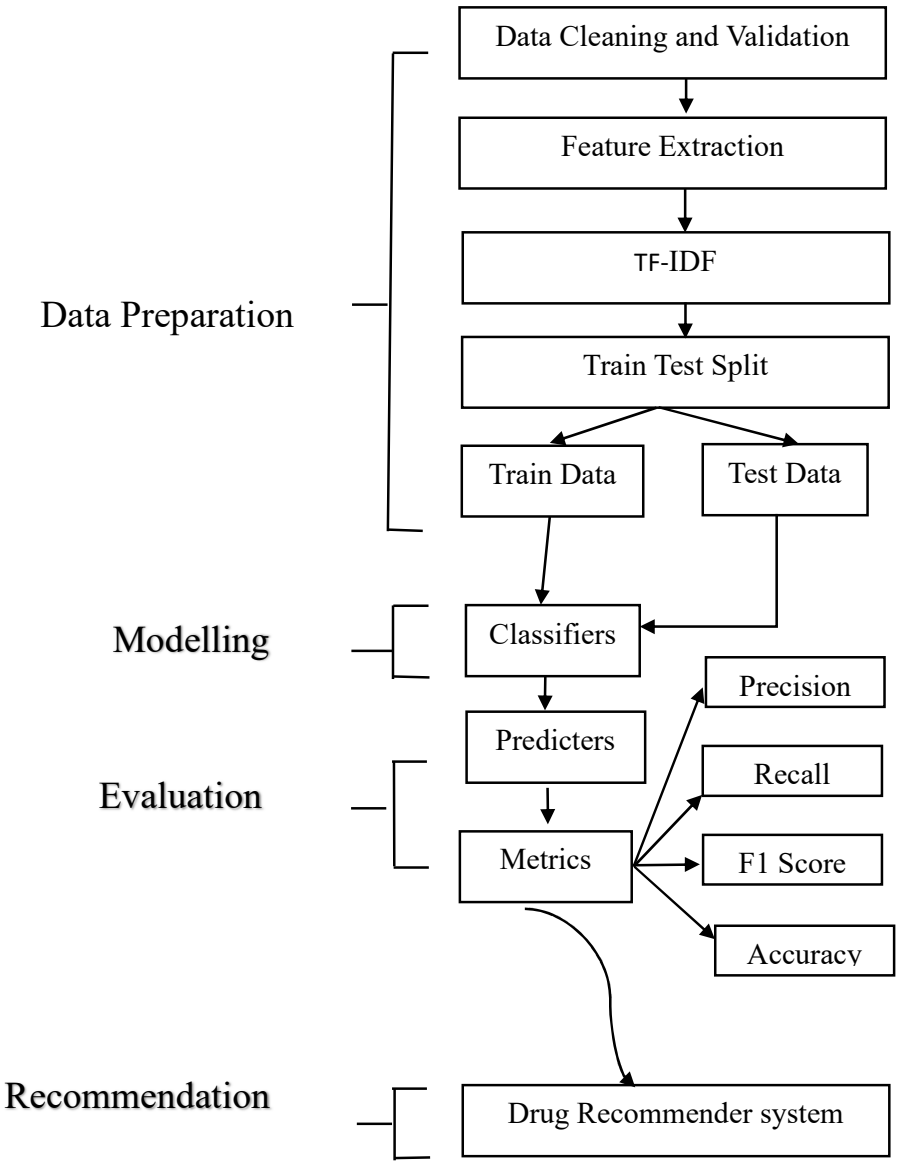

Fig. 1. Flowchart of the proposed model

A.Data Cleaning: Applied standard data preparation techniques like checking null values, duplicate rows, and removing unnecessary values and text from rows were used in this research. Subsequently, all 1194 null values

rows were removed in the conditions column, as shown in Fig. 2. We make sure that a unique ID is unique to remove duplicacy.

B.Feature Extraction:After text preprocessing, a proper set-up of the data is required to build classifiers for sentiment analysis. Machine learning algorithms can't work with text straightforwardly; they should be changed to numerical format. In particular, vectors of numbers. The TF-IDF is a well-known and straightforward strategy for feature extraction with text information used in this research [12].

TF-IDF: TF-IDF [12] is a popular weighting strategy in which words are offered with weight, not count. The principle was to give low importance to the terms often appearing in the dataset, which implies that TF-IDF estimates relevance, not a recurrence. Term frequency (TF) can be called the likelihood of locating a word in a document.

$$idf(t,d) = \log(1 + freq(t,d)) \quad (1)$$

Inverse document frequency (IDF) is the opposite of the number of times a specific term appeared in the whole corpus. It catches how a particular term is document-specific.

$$idf(t,d) = \log\left(\frac{N}{count(d \in D : t \in d)}\right) \quad (2)$$

TF-IDF is the multiplication of TF with IDF, suggesting how vital and relevant a word is in the document.

$$tfidf(t,d,D) = tf(t,d).idf(t,D) \quad (3)$$

C.Train Test Split: We created four datasets using TF-IDF. These datasets were split into 75% of training and 25% of testing. While breaking the data, we set an equal random state to ensure the same set of random numbers was generated for the train test split of the generated dataset.

D.Classifiers: Distinct machine-learning classification algorithms were utilised to develop a classifier capable of predicting the overall review. Logistic Regression, Passive-Aggressive, SVM, and Random Forest classifiers were experimented with using the TF-IDF model, given its sparse matrix, which makes applying tree-based classifiers time-consuming. A significant problem with this dataset is around 210K reviews, which takes substantial computational power. We selected only machine learning classification algorithms that reduce the training time and give faster predictions.

E.Metrics

The predicted Reviews were measured using five metrics, namely precision (Prec), recall (Rec), f1score (F1), and accuracy (Acc.). [14]. Let the letter be: T_p = True positive or occurrences where the model predicted the positive reviews truly, T_n = True negative or occurrences where the model predicted the negative class truly, F_p = False positive or occurrence where the model mispredicted the positive class, F_n = False negative or occurrences where the model mispredicted the negative class, Precision, recall, accuracy and f1score shown in equations given below,

$$Precision = \frac{Tp}{Tp+Fp} \quad (4)$$

$$Recall = \frac{Tp}{Tp+Fn} \quad (5)$$

$$Accuracy = \frac{Tp+Tn}{Tp+Fp+Tn+Fn} \quad (6)$$

$$F1\ score = 2.\frac{Precision.Recall}{Precision+Recall} \quad (7)$$

The area under the curve (Auc) helps distinguish a classifier's capacity to compare classes and is utilised to review the region operating curve (roc). Roc curve visualises the relationship between actual positive rate (Tpr) and false positive rate (Fpr) across various thresholds.

F.Drug recommendation system : After assessing the metrics, one predicted result was picked and joined to produce the prediction. The higher the useful count, the better the drug. In our recommendation system, you can enter a review of your health. The engine predicts the condition and recommends the top five drugs based on useful count and rating.

Graphical Analysis :

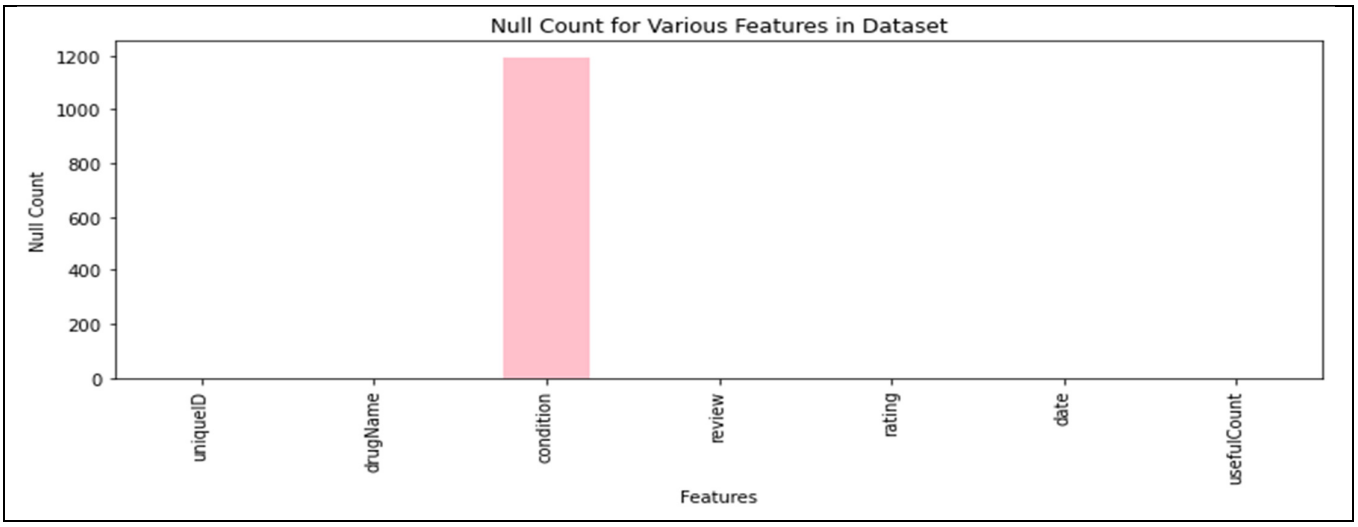

Fig. 2. Bar plot of the number of null values versus attributes

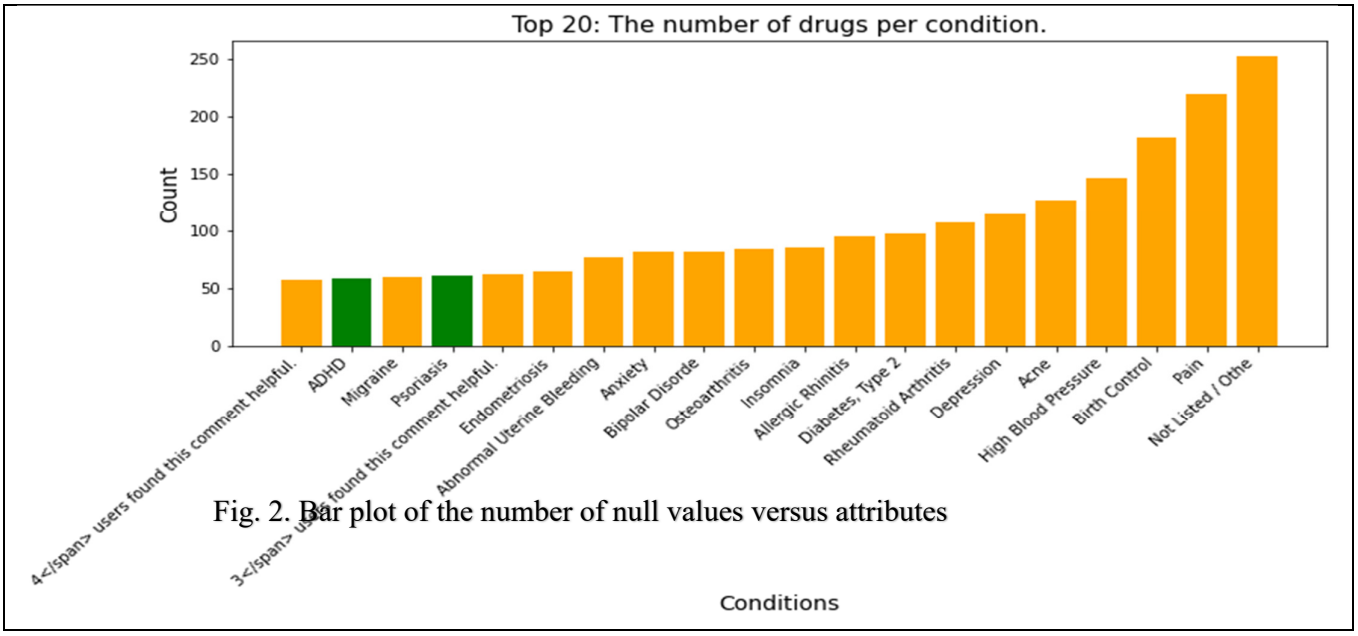

Fig. 3. Bar plot of Top 20 conditions that have a maximum number of drugs available

Fig. 3 shows the top 20 conditions with the maximum number of drugs available. One thing to notice in this figure is that two green-coloured columns show the conditions with no meaning. Removing all these conditions from the final dataset makes the total row count equal to 212141.

Fig. 4 shows the visualisation of value counts of the 10-star rating system. The rating beneath or equivalent to five featured a cyan tone; otherwise, it was blue. The vast majority pick four qualities; 10, 9, 1, 8, and 7 are more than twice the same number. It shows that the positive level is higher than the negative, and people's responses are polar. The condition and drug column were joined with the review text because the condition and medication words also have predictive power. Before proceeding to the feature extraction part, cleaning up the review text before vectorisation is critical. This process is also known as text preprocessing. We first cleaned the reviews after removing HTML tags, punctuations, quotes, URLs, etc. The cleaned reviews were lowercase to avoid duplication, and tokenisation was performed to convert the texts into small pieces called tokens—additionally, stopwords.

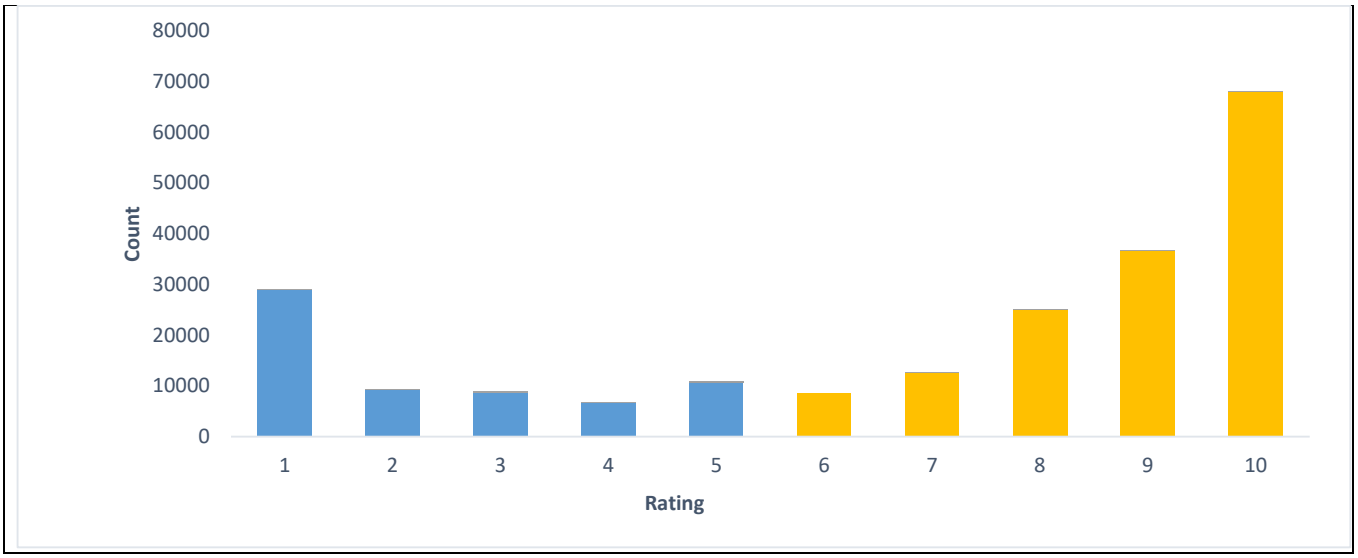

Fig. 4. Bar plot of count of rating values versus 10 rating number

Statistical Analysis: This work classified each review depending on the user's star rating. Ratings above five are classified as positive, while negative ratings are from one to five-star ratings. Initially, the number of positive and negative ratings in training data were 111583 and 47522, respectively. TF-IDF was used as the text representation method, and various machine learning algorithms were applied for classification.

Table I

Model	Prec	Rec	F1	Acc.
Logistic Regression	0.922	0.918	0.916	0.918
Passive Aggressive	0.94	0.94	0.939	0.94
SVM	0.93	0.928	0.926	0.928

Random Forest	0.954	0.951	0.95	0.951

After evaluating all the models, the prediction results of Random Forest (TF-IDF). The main intention is to ensure that all four models correctly classify the recommended top drugs. Table II shows the top four drugs recommended by our model for the top five conditions: Acne, Birth Control, High Blood Pressure, Pain and Depression.

Table II

Condition	Recommended Drug Name	Condition	Recommended Drug Name
Acne	Adapalene	Depression	Sertraline
	Accutane		Cymbalta
	Isotretinoin		Citalopram
	Differin		Celexa
	Spironolactone		Vilazodone
Pain	Oxycodone	Birth Control	Etonogestrel
	Elavil		Levonorgestrel
	Amitriptyline		Implanon
	hydrocodone		Etonogestrel
	OxyContin		NuvaRing
High Blood Pressure	Telmisartan	High Blood Pressure	Cardizem
	Amlodipine		Metoprolol
	Micardis		

This table represents the recommendation of the top five drugs on the top five conditions.

Conclusion:

Table 2 shows that, according to review Patient we predict the condition and based on these conditions recommended top five drugs using passive aggressive model. Here we use passive aggressive method because it take less time than random forest for prediction In this article we make the comparative study of four different methods of classification. We observed that random forest method is superior as compared to logistic regression, passive aggressive, support vector machine. Passive aggressive method is give excellent result as compared to support vector and logistic regression. Support vector machine is better as compared to logistic regression. In overall comparison the performance of logistic regression is worst.

References:

[1] *Telemedicine*. (n.d.). https://www.mohfw.gov.in/pdf/Telemedicine.pdf.

[2] Bartlett, J. G., Dowell, S. F., Mandell, L. A., File, T. M., Musher, D. M., & Fine, M. J. (2000). Practice guidelines for managing community-acquired pneumonia in adults. *Clinical Infectious Diseases*, *31*(2), 347–382. https://doi.org/10.1086/313954

[3] Wittich, C. M., Burkle, C. M., & Lanier, W. L. (2014). Medication errors: An overview for clinicians. *Mayo Clinic Proceedings*, *89*(8), 1116–1125. https://doi.org/10.1016/j.mayocp.2014.05.007

[4] Huai-fu, W. (2013). The reason and prevention of hospital medication errors. *Practical Journal of Clinical Medicine*. https://www.semanticscholar.org/paper/The-reason-and-prevention-of-hospital-medication-Huai-fu/9a3df70d921576a4340324c4a28f77276228c741

[5] Duggan, S. F. and M. (2013, January 15). Health Online 2013. *Pew Research Center*. https://www.pewresearch.org/internet/2013/01/15/health-online-2013/

[6] Anderson, J. G., & Abrahamson, K. (2017). Your health care may kill you: Medical errors. *Studies in Health Technology and Informatics*, *234*, 13–17.

[7] Garg, S. (2021). Drug recommendation system based on sentiment analysis of drug reviews using machine learning. *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 175–181. https://doi.org/10.1109/Confluence51648.2021.9377188

[8] Tekade, T. N., & Emmanuel, M. (2016). Probabilistic aspect mining approach for interpretation and evaluation of drug reviews. *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, 1471–1476. https://doi.org/10.1109/SCOPES.2016.7955684

[9] Doulaverakis, C., Nikolaidis, G., Kleontas, A., & Kompatsiaris, I. (2012). GalenOWL: Ontology-based drug recommendations discovery. *Journal of Biomedical Semantics*, *3*(1), 14. https://doi.org/10.1186/2041-1480-3-14

[10] Bao, Y., & Jiang, X. (2016). An intelligent medicine recommender system framework. *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*, 1383–1388. https://doi.org/10.1109/ICIEA.2016.7603801

[11] *Search | kaggle*. (n.d.). Retrieved February 2025, from https://www.kaggle.com/search?q=drug+recommendation+system+based+on+sentiment+analysis+in%3Adatasets

[12] Ramos, J. E. (2003). *Using tf-idf to determine word relevance in document queries*. https://www.semanticscholar.org/paper/Using-TF-IDF-to-Determine-Word-Relevance-in-Queries-Ramos/b3bf6373ff41a115197cb5b30e57830c16130c2c

[13] Powers, D. M. W. (2020). *Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation* (No. arXiv:2010.16061). arXiv. https://doi.org/10.48550/arXiv.2010.16061

[15] Insilico Medicine Unveils First-Ever Drug Discovered and Designed by Artificial Intelligence. https://digitalmarketreports.com/news/11731/insilico-medicine-first-ai-drug/

[14] Nimaga Nimaga, M. (2022). Mejora de la destreza escrita en inglés por medio de la literatura: La escritura de una historia breve con Inklewriter. https://core.ac.uk/download/544378814.pdf

[16] Customer Service Reviews: What They Are & How to Manage?. https://feedeo.spreading.ai/blog/customer-service-review