International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 01 | January - 2023 Impact Factor: 7.185 ISSN: 2582-3930

Valetudinarian Situation Categorize Using Drug Reviews

Ayush Kumar Chaudhary[1], Kushager Kumar[1]

[1] Final year student , Dept. of Information Technology, Maharaja Agrasen Institute of technology, Delhi, India

---

Abstract - Human healthcare is one of society's most important issues. To ensure that patients receive the care they require as quickly as possible, it searches for the best diagnosis and thorough disease. For the health-related component of searching, other disciplines are required, such as statistics and computer science, because this recognition is frequently complex[3]. These disciplines must go beyond the traditional ones in order to follow innovative methodologies. Because there are so many new strategies, it is possible to give a general summary without focusing on certain details. In order to do this, we propose a thorough examination of machine learning-related disorders in humans. In order to find interesting trends, make unimportant forecasts, and aid in decision-making[1], this research focuses on existing approaches connected to machine learning growth applied to the diagnosis of human disorders in the medical industry. In order to produce suitable decision support, this study examines distinctive machine learning methods employed in healthcare applications[4], [5]. With the goal of developing a practical decision support system for medical applications, the research gap is to be filled up by this work.

Key Words: **Human disease, Machine learning, Neural Networks.**

1. INTRODUCTION

One of the most significant aspects of the economy and of human life is medicine and healthcare. The world we live in today and the world from a few weeks ago have undergone a great deal of change[7]. Everything has changed to be horrifying and bizarre. The doctors and nurses are working as hard as they can to save people's lives in this situation where everything has gone virtual, even at the risk of putting their own lives in peril. In some isolated villages, there are no medical facilities.

Board-certified medical professionals known as virtual doctors prefer to conduct telephone and video appointments over in-person visits, while this is not an option in an emergency[6]. Machines are always regarded as superior to mankind because they can complete jobs more rapidly and consistently with a high level of accuracy without the possibility of human error. A disease predictor, often referred to as a virtual doctor[3], [8], is able to accurately predict any patient's symptoms without the involvement of a human. A disease predictor can also be helpful for diseases like COVID-19 and Pandemic because it can detect a person's illness without coming close touch with them. Although there are some virtual doctor models out there, they lack the necessary accuracy because not all the necessary criteria are taken into account. To determine which model makes the most accurate predictions[5], [9], it was important to create a variety of models. Although the size and complexity of ML projects vary, they all follow a similar general format. To recall the creation and application of the predictive model, several rule-based techniques were taken from machine learning. Various machine learning (ML) methods were used to start some models by collecting raw data and dividing it into groups based on gender, age, and symptoms. The data set was subsequently processed using a variety of machine learning (ML) models, including fine, medium, and coarse decision trees[4], [7], gaussian naive Bayes, and passive-aggressive classifiers. Every model was given the same set of input parameters to process the data, and as a result, each model produced the disease with a different level of accuracy. The model with the best degree of precision has been chosen.

2. LITERATURE SURVEY

As shown by several publications on this subject, there have been numerous studies on the harvesting of adverse drug reactions using reviews in social networks. These studies employ a variety of techniques to find negative customer opinions. The strategy based on definitions is the most popular.[1] , [2] Adverse drug reactions are taken from drug instructions, clinical testing records, and user evaluations from forums devoted to health in dictionaries. The majority of papers describe methods that employ machine learning techniques. Support Vector Machine (SVM) was employed by the authors in Convolutional Fields as well as the random forest approach [6], [7]. N-grams, components of speech tags, semantic categories from

Larger working, the number of negated contexts, the belonging lexicon-based features for ADRs, drug names, and word embeddings are the most prominent features of machine learning. Existing research demonstrates that features based on subjectivity analysis, topic modeling[8], sentiment analysis, and polarity classification can be applied to enhance ADR detection outcomes.

Convolutional neural networks have recently been used in several studies for ADR classification, recurrent neural networks for extracting ADRs, and encoder-decoder networks for ICD coding[10], [8]. Authors have also investigated convolutional neural networks for predictions of socio-demographic variables based specifically on medical reviews[5].

The preponderance of works on sentiment analysis for drug reviews falls into one of two categories: those that use supervised classification to learn sentiments or those that apply lexicons with sentiment scores.

One of the first papers on sentiment analysis of drug reviews To use numerous polarity classifiers, Xia et al. created a topic classifier from patient data[3], [4], [5]. A clause-level sentiment analysis system that takes into account several review features, such as overall satisfaction, effectiveness, side effects, and condition, is demonstrated by Na et al. Here, a rule-based technique is used to compute the sentiment polarity of individual phrases based on a lexicon while taking into consideration grammatical relations and lexical annotation[7] , [8]. The sentiment classification of patient feedback on oncological medications is aspect-based. Using a lexical resource, opinion words are recognized here, and overall sentiments are derived.

Utilizing supervised learning sentiment analysis, examine patient drug satisfaction. Three levels of polarity were determined in this study when comparing SVM and neural network-based techniques. Numerous studies have tried to enhance domain adaptation or cross-domain sentimental analysis [1], [2],[9], but not at the level of medication review aspects, but rather among

different entities like products, movies, or dining establishments. A thorough survey of the cross-domain sentiment analysis literature is presented.

	Unnamed: 0	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combination of Bystolic 5 Mg and Fish Oil"	9.0	May 20, 2012	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of Intuniv. We became concerned when he began this last week, when he started taking the highest dose he will be on. For two days, he could hardly get out of bed, was very cranky, and slept for nearly 8 hours on a drive home from school vacation (very unusual for him.) I called his doctor on Monday morning and she said to stick it out a few days. See how he did at school, and with getting up in the morning. The last two days have been problem free. He is MUCH more agreeable than ever. He is less emotional (a good thing), less cranky. He is remembering all the things he should. Overall his behavior is better. \r\nWe have tried many different medications and so far this is the most effective."	8.0	April 27, 2010	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, which had 21 pill cycle, and was very happy- very light periods, max 5 days, no other side effects. But it contained hormone gestodene, which is not available in US, so I switched to Lybrel, because the ingredients are similar. When my other pills ended, I started Lybrel immediately, on my first day of period, as the instructions said. And the period lasted for two weeks. When taking the second pack- same two weeks. And now, with third pack things got even worse- my third period lasted for two weeks and now it's the end of the third week- I still have daily brown discharge.\r\nThe positive side is that I didn't have any other side effects. The idea of being period free was so tempting. Alas."	5.0	December 14, 2009	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth control. I'm glad I went with the patch, I have been on it for 8 months. At first it decreased my libido but that subsided. The only downside is that it made my periods longer (5-6 days to be exact) I used to only have periods for 3-4 days max also made my cramps intense for the first two days of my period. I never had cramps before using birth control. Other than that im happy with the patch"	8.0	November 3, 2015	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around. I feel healthier, I'm excelling at my job and I always have money in my pocket and my savings account. I had none of those before Suboxone and spent years abusing oxycontin. My paycheck was already spent by the time I got it and I started resorting to scheming and stealing to fund my addiction. All that is history. If you're ready to stop, there's a good chance that suboxone will put you on the path of great life again. I have found the side-effects to be minimal compared to oxycontin. I'm actually sleeping better. Slight constipation is about it for me. It truly is amazing. The cost pales in comparison to what I spent on oxycontin."	9.0	November 27, 2016	37

Figure 1: drugs reviews

3. METHODLOGY

We generated an excel sheet from an open-source dataset that included every symptom associated with each condition. Age and gender were then established as components of the dataset based on the disorders. Over 230 ailments were listed, each with over a thousand distinct symptoms and a thousand drug reviews. Various machine learning algorithms were fed data on a person's symptoms, age, and gender.

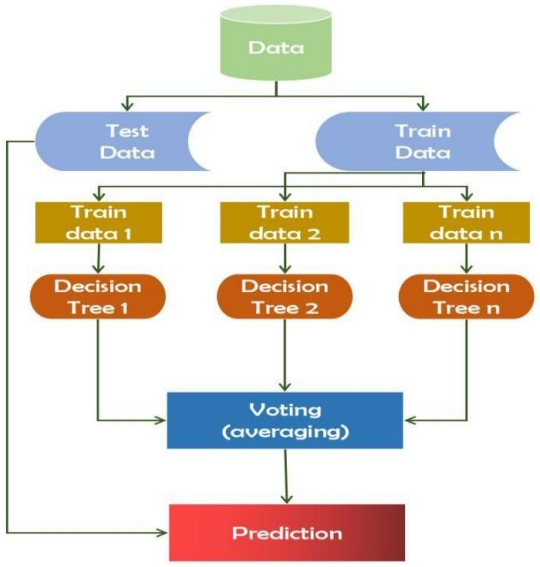


Figure 2: Working mechanism of random forest technique.

3.1 NAÏVE BAYES

It is a machine learning algorithm that uses Bayes' probability theorem to solve classification issues. This is mostly used for text categorization, which requires large training data sets[8].

$$P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)}$$

Figure 3: naïve bayes theorm

Where P(h|d) is the likelihood that hypothesis h will hold given the data d. The posterior probability refers to this. P(d|h) represents the likelihood that data d would exist if hypothesis h were accurate. P(h) is the probability that hypotheses h are correct (regardless of the data). The probability previous to h is what we refer to as. The probability of the data is P(d) (regardless of the hypothesis).

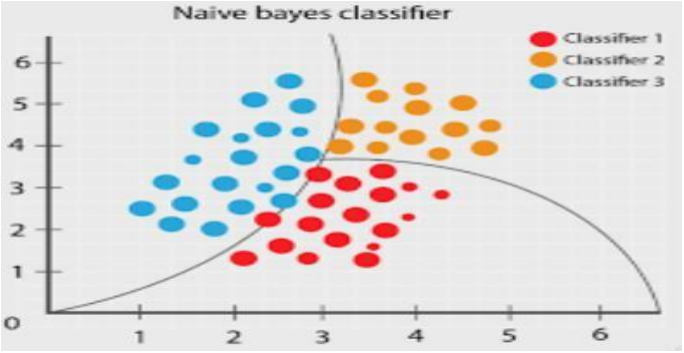


Figure 4 : graph representation of naïve bayes

3.2 DECISION TREES

The supervised learning algorithm family includes the decision tree algorithm. Regression and classification are done using it. The top of the decision tree employs the tree diagram approach for prediction. It starts with a root node, splits in the dominating input feature, and then splits once more. These procedures continue until every input has been deposited [9], [11], and at the extreme last node, which holds the weights, the input is classified based on these weights. The maximum number of splits from each node in a coarse tree is four. In contrast, the maximum number of splits from each node in a medium tree is 20. The maximum number of splits from each node in a fine tree is 100.

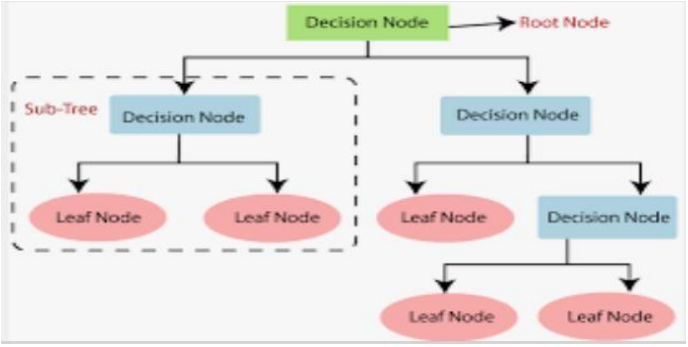


Figure 5: representation of Descision trees

3.3 PASSIVE AGGRESSIVE CLASSIFIER

Large-scale learning typically uses passive-aggressive algorithms. One of the few "online-learning algorithms" is this one. As opposed to batch learning, when the full training dataset is used at once, online machine learning techniques employ sequential input data and update the machine learning model one step at a time. This is especially helpful when there is a large amount of data and training the complete dataset would be computationally impossible due to the magnitude of the data[11], [12]. Simply put, an algorithm for online learning will acquire a training example, update the classifier, and then discard the sample.

Passive: Maintain the model and make no changes if the prediction is accurate. In other words, the example's data are insufficient to alter the model in any way.

Aggressive: Modify the model if the prediction turns out to be inaccurate. In other words, a model modification could make it right.

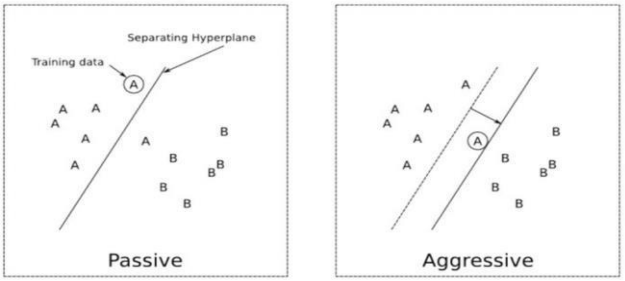


Figure 5 : passive Aggressive

4. RELATED WORK

The first idea that entered my head was to categorize patients' conditions using reviews. Online systems may find it useful to recognize human situations. The second step is to anticipate the reviewer's mood[3]. The third step is to categorize medications for the same ailment (in my instance, "Birth Control").

Data was divided into two.csv files; the larger dataset was used for training, while the smaller dataset was used for testing[7]. This was done simply by concatenating the two datasets, which made cleaning and transforming the data easier to accomplish all at once.

sample template format ,Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

Text data will inevitably contain some artifacts. I initially think I'm looking at HTML apostrophe code. I used the OSEMN methodology for data science, which involves cleaning (Scrubbing) data as soon as you obtain it (assuming there is still more to clean) [9]. Because I used three different words to explain the same subject, I hope there is no misunderstanding (cleaning, scrubbing, preprocessing).

Preprocessing

Making text appropriately clean and converted is the initial step in natural language processing. If you don't, your model either won't work or performs extremely poorly.

Taking Out Stop Words

Words like "I," "was," "there," "me," etc. don't provide a neural network with much meaning; instead, they create noise and slow down the learning process.

Lemmatization

To minimise dimensionality and increase classification accuracy, verbs could be changed into their root form since

words like learning, learned, and learn all imply the same thing.

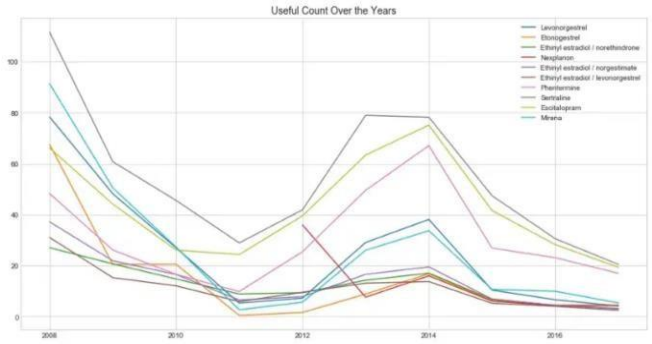
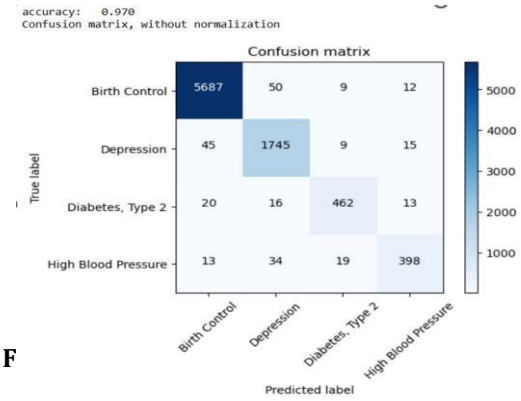


Fig 6 : top most reviewed drugs have similar pattern

5. RESULT

Since the Passive Aggressive approach is more sophisticated than Multinomial Naive Bayes and counting the test, we use it. We evaluate using accuracy as the metric, and for better presentation, we plot the confusion matrix.



F

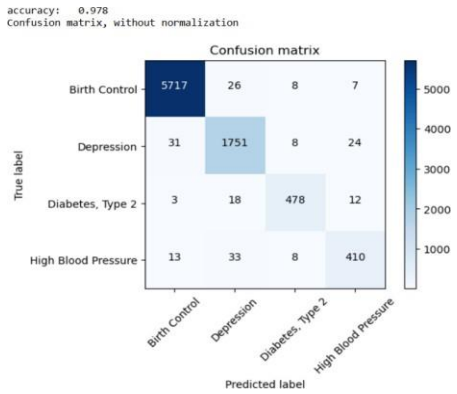


Fig 8 : Confusion Matrix for Passive Aggressive

Using the Bag of Words model with the Passive Aggressive algorithm, we were able to achieve 97.8% accuracy. The majority of classes are for birth control, hence this category contains the most forecasts.

Table I shows the results using evaluation metrics on a bag of words vectorization technique. We can easily see that perceptron outperforms all other classification algorithms. Multinomial Naïve Bayes have accuracy of 97% and Passive Aggressive accomplished a 97.8% AUC score.

Model	Accuracy
Multinomial Naïve Bayes	0.970
Passive Aggressive	0.978

6. CONCLUSIONS

Given that the user has provided specific drug review information, this research offered a way of identifying and predicting the presence of a disease in an individual utilizing machine learning algorithms such Naive Bayes, Random Forest Classifier, K-Nearest Neighbors, and Support Vector Machines. As a result, the suggested techniques also result in a comparison of various machine learning methods for multiclass classification. The accuracy of the disease prediction does depend on the review's authenticity, but it is strongly thought that the proposed method can diagnose diseases and identify them while also lowering the cost of medical diagnosis, treatment, and doctor consultation.

Using the aforementioned parameters, the passive-aggressive categorization had a disease prediction accuracy of 97.8%, which was the highest. Nearly every ML model produced good accuracy values.

REFERENCES

[1] Diana Cavalcanti and Ricardo Prudêncio. 2017. Aspect- Based Opinion Mining in Drug Reviews. In Progress in Artificial Intelligence, Eugénio Oliveira, João Gama, Zita Vale, and Henrique Lopes Cardoso (Eds.). Springer International Publishing, Cham, 815–827M.

[2] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain Adaptation for Large-scale Sentiment Classification: A Deep Learning Approach. In Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML'11). Omnipress, USA, 513–520

[3] [12] A. Mishra, A. Malviya, and S. Aggarwal. 2015.Towards Automatic Pharmacovigilance: Analysing PatientReviews and Sentiment on Oncological Drugs. In 2015IEEE International Conference on Data Mining Workshop(ICDMW)1402–1409. https://doi.org/10.1109/ICDMW.2015.230.

[4] Vinodhini Gopalakrishnan and Chandrasekaran Ramaswamy. 2017. Patient opinion mining to analyze drugs

satisfaction using supervised learning. Journal of Applied Research and Technology 15, 4 (2017), 311 – 319. https://doi.org/10. 1016/j.jart.2017.02.005

[5] K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Computational and structural biotechnology journal 13, 8 (2015)

[6] A.U. Haq, J.P. Li, M.H. Memon, S. Nazir, R. Sun, A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms, Mobile Information Systems 2018 (2018)

[7] M. Maniruzzaman, M.J. Rahman, B. Ahammed, M.M. Abedin, Classification and prediction of diabetes disease using machine learning paradigm, Health Information Science and Systems 8(1), 7 (2020)

[8] D.R. Langbehn, R.R. Brinkman, D. Falush, J.S. Paulsen, M. Hayden, an International Huntington's Disease Collaborative Group, A new model for prediction of the age of onset and penetrance for huntington's disease based on cag length, Clinical genetics 65(4), 267 (2004)

[9] T.V. Sriram, M.V. Rao, G.S. Narayana, D. Kaladhar, T.P.R. Vital, Intelligent parkinson disease prediction using machine learning algorithms, International Journal of Engineering and Innovative Technology (IJEIT) 3(3), 1568 (2013)

[10] S. Vijayarani, S. Dhayanand, Liver disease prediction using svm and na¨ıve bayes algorithms, International Journal of Science, Engineering and Technology Research (IJSETR) 4(4), 816 (2015)

[11] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz ,Yoram Singer, Online Passive-Aggressive Algorithms ,School of Computer Science and Engineering The Hebrew University Jerusalem, 91904, Israel

[12] S. Karthika , N. Sairam , A Naïve Bayesian Classifier for Educational Qualification, School of Computing, SASTRA University