iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VIII **Month of publication:** August 2021

DOI: https://doi.org/10.22214/ijraset.2021.37537

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

Drug Disease Association Prediction Using NLP & Machine Learning

Suchith Reddy Vemula
Systems Engineer, TCS, Hyderabad.

Abstract: ***The relevance of analysing user-generated data on the internet has lately increased owing to the vast amount of information that can be obtained via proper study of such data. The majority of this information can be found on social media sites like Facebook, Twitter, and LinkedIn. Opinions and reviews on goods, movies, prescriptions, hotels, and other items are among the data accessible on such platforms. Companies are increasingly relying on data mining and analysis to get a better understanding of public opinion on a certain topic. There has been sufficient study on the use of sentiment analysis in many areas such as product reviews, movies, hotels, and so on. However, in the field of medicine, such techniques must be given more weight, since the US Food and Medication Administration has done multiple research on the consequences of adverse drug responses on patients. Pharmaceutical firms must study the impact of regularly used pharma products on patients in order to understand the good and bad impacts of medications on patients. The goal of this study is to use ML models to analyse the review of patient evaluations in order to evaluate if the opinions represented in the reviews are positive (or) negative.***

Keywords: Data Pre-Processing, Exploratory Data Analysis, Feature Extraction, Sentimental Classification, ML Algorithms, Prediction, Visualisation .

I. INTRODUCTION

Drug development is a difficult, time taking, and costly procedure. A medicine typically takes ten to fifteen years of research and one to fifteen billion dollars to develop from an abstract notion to a commercially accessible product. 90 percent of medications fail to get FDA approval in the United States each year, prohibiting them from being used in clinical trials. As a result, Drug Repositioning (DR) is a computational procedure. This strategy avoids various pre-approval testing that are important for newly created medicinal molecules, and it may cut the development cycle for repositioning a medicine in half, from three to ten years. Governments, business entities, and university scholars have been more interested in DR in recent years. They must analyse user-generated data, which has lately gained prominence owing to the wealth of information that can be gleaned through quantitative analysis of such data. There has been considerable study on the use of sentiment analysis in many areas such as product reviews, movies, and so on. However, there is a greater need for such approaches in the area of medicine, since the US FDA has done multiple research on the impacts of medication responses on patients. Pharma firms must study the impact of regularly used pharmaceuticals on patients in order to get a clear vision on both the good and bad impacts of pharmaceuticals on patients.

II. RELATED WORK

There are few studies which focus on applying sentiment analysis techniques using machine learning models. Data in the form of reviews for different subjects like camera, laptops, summer camps, lawyers, drugs, radio, restaurant and television were chosen for the study. Related Work Of ducted sentiment analysis on reviews posted on hearing loss forums using Naive Bayes Algorithm with Pipeline Technique was done. Further analysis was carried out using traditional bag of words approach. A bag of words approach is the process of extracting features from the text, it involves extraction of words and the infrequency of occurrence in text. The words are assigned a score based on their polarity.

Proposed machine learning model with the help of sentimental Analysis in NLP where a model trained in one domain and can be used as a classifier in another domain. Support Vector Machine (SVM) was used as a base classifier. To test the performance of classifier trained in one domain and its classifying power in another domain, a new SVM classification model was trained for each model and was used to test the model on the rest of the dataset with a K-fold of 25. The classification accuracy was calculated as the average of K-fold tests.

By building two different ensembles where one used simple majority vote of its component models for each new classification and another used weighted majority votes, the authors demonstrated that it is possible to improve the accuracy by selecting the cross-domain models with lexicons similar to target domain lexicon. This work demonstrated that it is possible to deploy a model trained in one domain to a different domain and achieve an acceptable accuracy in classifying the sentiment.

The two methods were compared; in each of the 3 test cases, machine learning algorithms outperformed the bag of words approach. Algorithm gave the best performance with the over-all agreement of 0.64 which is a good agreement with the model. This study showcased that machine learning with NLP methods can perform better when compared to existing traditional sentiment analysis methods. And thus this paper demonstrates the use of machine learning model along with the sentimental analysis in-order to predict the results.

III. METHODOLOGY

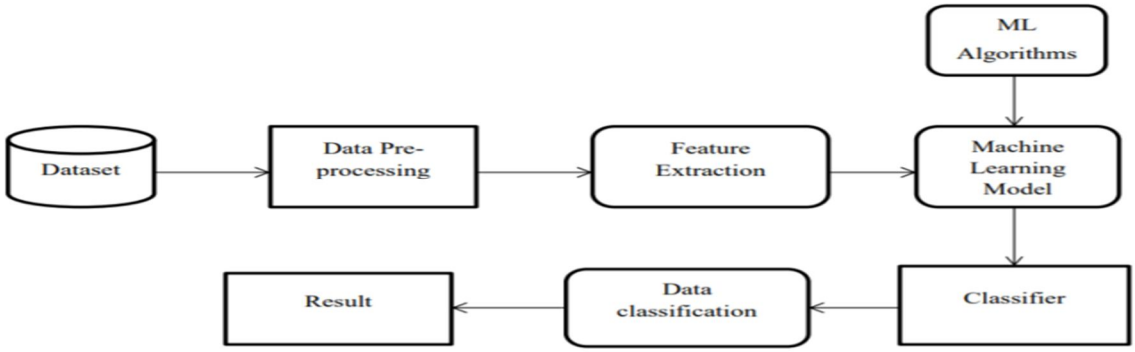


Fig 1. Architecture

A. Dataset Collection

Collection of data regarding the drug and disease and effect of drug on the patients disease either from web or by a survey . The collected data can be either observational, experimental, simulative, derived or computed. This process of accumulating information about our target variables helps us in achieving our outcomes. The collected data might not be in a refined way and thus we need to perform further processes to make use of that data well.

	drugName	condition	review	rating	date	usefulCount
0	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	20 May 2012	27
1	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	27 April 2010	192
2	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	14 December 2009	17
3	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	03 November 2015	10
4	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	27 November 2016	37

Fig 2. Sample DataSet

B. Exploratory Data Analysis (EDA)

In ML, it is way of visualizing, summarizing and interpreting the information that is hidden in rows and columns. It is a approach to analyzing data sets to summarize their main characteristics, often with visual methods. Primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task in the form of histogram. Exploratory Data Analysis is done before pre-processing to understand the characterisation of data .

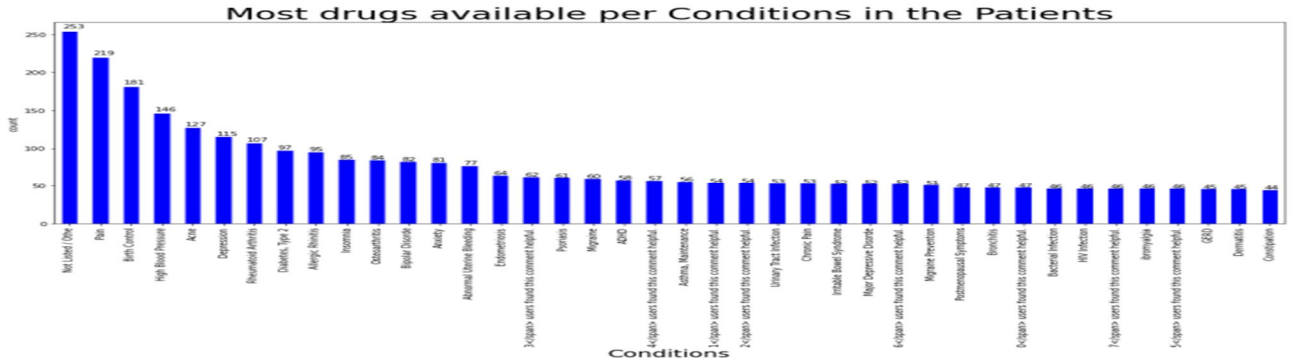


Fig 3. No of Drugs available per Conditions (disease)

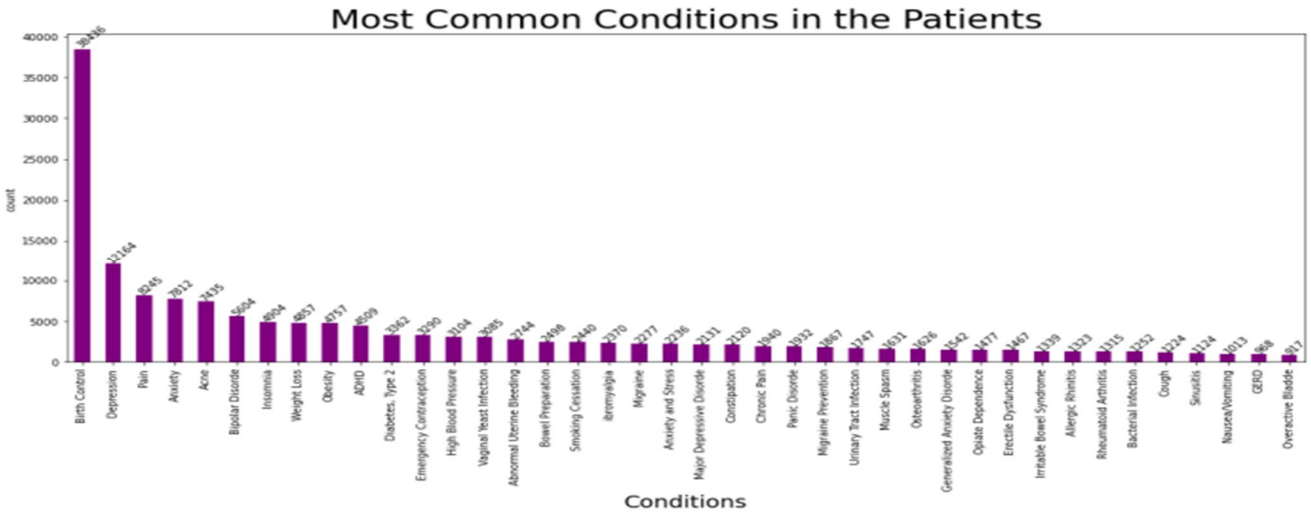


Fig 4. No of Disease conditions in patients

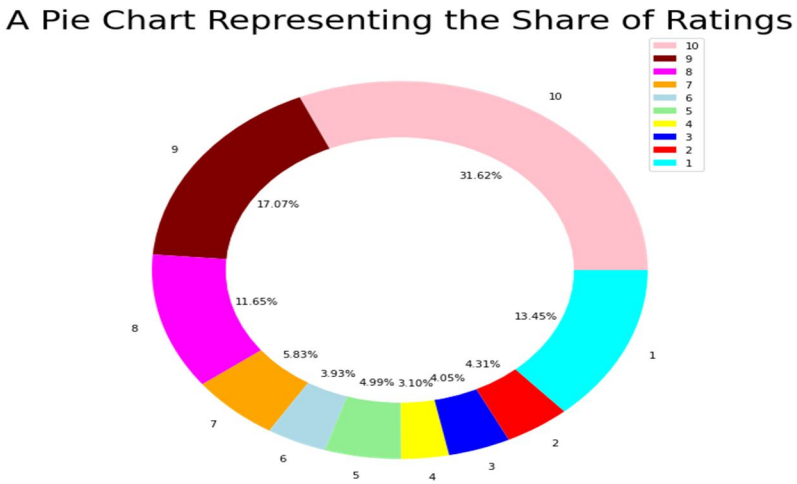


Fig 5. Percentage of No of ratings on scale of (1-10)

C. Data Pre-Processing

Data pre-processing is used to refine our raw data so that it will be worth to use, it becomes a value that can be used for further analysis. There might be missing values, null values, unwanted noisy data which may affect the final outcome and thus to avoid that we use some of the data pre-processing techniques like Formatting, Cleaning, Sampling etc.

```
dframe=dframe.drop(['date'],axis=1)
dframe.head()
```

	drugName	condition	review	rating	usefulCount	Sentiment_review
0	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	27	1.0
1	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	192	1.0
2	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	17	1.0
3	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	10	1.0
4	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	37	1.0

```
dframe['condition'].isnull().sum()
```

```
1194
```

```
dframe = dframe.dropna(axis = 0)
dframe = dframe.drop(['usefulCount'],axis=1)
dframe.shape
```

```
(213869, 5)
```

Fig 6. Screenshot showing Data Pre-Processing Techniques

Then assign manually the review sentiment based on the rating to training data. If rating is > 5 , classify it as 1 and else 0. (works as an output label)

```
dframe.loc[(dframe['rating'] >= 5), 'Sentiment_review'] = 1
dframe.loc[(dframe['rating'] < 5), 'Sentiment_review'] = 0
dframe['Sentiment_review'].value_counts()
```

```
1.0    161491
0.0     53572
Name: Sentiment_review, dtype: int64
```

Fig 7. Screenshot showing output label assignment of training set

D. Feature Extraction

Here the main NLP techniques comes into picture where NLP is a research and application field, which is concerned with the manipulation and understanding of natural languages. In NLP, this task does have tremendous impact on the success of text analysis. This is mostly caused by the unstructured and arbitrary nature of text data. Furthermore, machines need structure and numerical data. In order to get so we use following techniques :

1) *Tokenization:* For processing written natural language it is inevitable to split texts into smaller units, which are called tokens. Computers need to distinguish single entities of a text and tokenization is used to create them. Usually tokens represent simple words, which are the smallest independent units of natural language. Tokenization breaks running texts into short text entities and is the very first task in any text pre-processing cycle.
2) *Stemming:* Stemming is a method which removes stop words**.** Stop words are those which are least required by the model for predicting output Besides stop word elimination, stemming is a useful technique to map words to their word stems and further reduce the input dimension. This helps to extract the real meaning of a text and makes the unstructured data better accessible for a machine.

```
stemmer = SnowballStemmer('english')

def review_to_words(raw_review):
    # 1. Delete HTML
    review_text = BeautifulSoup(raw_review, 'html.parser').get_text()
    # 2. Make a space
    letters_only = re.sub('[^a-zA-Z]', ' ', review_text)
    # 3. lower letters
    words = letters_only.lower().split()
    # 5. Stopwords
    meaningful_words = [w for w in words if not w in mystopwords]
    # 6. Stemming
    stemming_words = [stemmer.stem(w) for w in meaningful_words]
    # 7. space join words
    return( ' '.join(stemming_words))
```

```
dframe['review_clean'] = dframe['review'].apply(review_to_words)
```

Fig 8. Screenshot showing stemming

3) *Vectorization:* Besides pre-processing the words themselves, their representations have to be changed into a machine readable format. Vectorization is an approach that transforms a text into one vector.

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.pipeline import Pipeline
from sklearn.metrics import confusion_matrix,classification_report
```

```
cvz = CountVectorizer(max_features = 20000, ngram_range = (5, 5))
pipeline = Pipeline([('vect',cvz)])
```

```
dframe_train_features = pipeline.fit_transform(x['review_clean'],x['drugName'])
#print(cvz.vocabulary_)
```

Fig 9. Screenshot using Vectorization

E. Machine Learning Model

The Vectorized data is now sent to the machine learning model first to train, As we are using supervised Learning here, the model is given the input and output for training. The algorithms we are using is Support Vector Machine (SVM) which helps model get trained regarding the output classification and then we give the test data to model to classify the input and predict output.

1) *Logistic Regression:* It is a linear algorithm with a non-linear output. Logistic regression is used when the variables are categorical. It uses logistic function to model the conditional probability. It uses sigmoid function an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

1 / (1 + e^-value)

```
logreg = LogisticRegression()
logreg.fit(X_train, y_train)
```

```
LogisticRegression()
```

```
predictlog=logreg.predict(X_test)
logreg.coef_
```

```
array([[ 0.13705465, -0.7762381 , -0.7762381 , ..., -0.01791968,
         0.08832454,  0.04758498]])
```

```
from sklearn import metrics
```

```
model1=metrics.accuracy_score(y_test,predictlog)
```

```
print(model1)
```

```
0.7567991631799164
```

Fig 10. Screenshot showing data fitting in model of logistic regression algorithm

2) *Support Vector Machine:* Support vector machine is another simple algorithm in machine learning which is highly preferred by many as it produces significant accuracy with less computational power. It is used for both regression and classification objectives. The objective of the support vector machine algorithm is to find a hyper plane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. As a supervised classifier, a standard support vector machine (SVM) aims to find a hyper plane separating 2 classes which maximizes the distance to the closest points from each class. The closest points are called support vectors.

```
svm=SVC()
```

```
svm.fit(X_train, y_train)
```

```
SVC()
```

```
predictsvm=svm.predict(X_test)
```

```
model2=metrics.accuracy_score(y_test,predictsvm)
```

```
print(model2)
```

```
0.7723430962343096
```

Fig 11. Screenshot showing data fitting in model of SVM algorithm

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

3) *kNN Classifier:* The k-nearest neighbour (kNN) algorithm classifies unlabeled instances based on a voting of the labels of k closest training examples in the feature space. kNN is a lazy learning algorithm since it defers data processing until a classification request arises. Because kNN uses local information, it can achieve highly adaptive performance. On the other hand, kNN involves large storage requirement and intensive computation, and the value of k also needs to be determined properly.

```
from sklearn.neighbors import KNeighborsClassifier
k_neighbour = KNeighborsClassifier(n_neighbors=2)
```

```
k_neighbour.fit(X_train, y_train)
```

```
KNeighborsClassifier(n_neighbors=2)
```

```
predictknn=k_neighbour.predict(X_test)
```

```
model3=metrics.accuracy_score(y_test,predictknn)
print(model3)
```

```
0.7659414225941422
```

Fig 12. Screenshot showing data fitting in model of kNN algorithm

F. Data Classification & Results

Test Data is thus classified by model which is trained by algorithms and predict whether the drug-disease association has positive effect or negative effect and the results are visualised. Here we are also doing a comparative study of which algorithm has more accuracy and for that we are also using Logistic regression and KNN Classifier algorithm.

G. Proposed Algorithm

The Algorithm or the total process is according to the following steps :

1) *Step-1:* Exploratory Data Analysis is done before pre-processing to understand the characterisation of data .
2) *Step-2:* . Data Pre-Processing is done by removing null values, and reducing noisy data and use cross-validation to split the data into training and test data.
3) *Step-3:* Then assign manually the review sentiment based on the rating to training data. If rating is > 5 , classify it as 1 and else 0. (works as an output label)
4) *Step-4:* Use the NLP techniques like stemming, tokenisation, vectorization to extract features from the text given as review and then give the vectorized result to model. (works as input labels)
5) *Step-5:* Vectorized training set is then used to train the model with input and output labels .
6) *Step-6:* After training the model, the model is used to predict the review sentiment for testing data .
7) *Step-7:* The predicted result is analysed with the actual result and the predictive analysis is then visualised.
8) *Step-8:* Lastly the accuracy of the models which we are using is comparatively studied.

IV. RESULT ANALYSIS

Our Dataset Consists of approximately 2 lakh records in which we have reviews stated by the patients regarding the effect of the drug on their disease .We have split the total data into 70% training and 30% testing and then we train the data on model and we test the model on testing set. Here we develop model using three different algorithms which means 3 models are developed and their predicted results are visualised using confusion matrix. Here confusion matrix demonstrates the True Positives, True Negatives, False Positives and False Negatives so that we get an idea of how many positive and negative sentiments are classified correctly and vice versa.

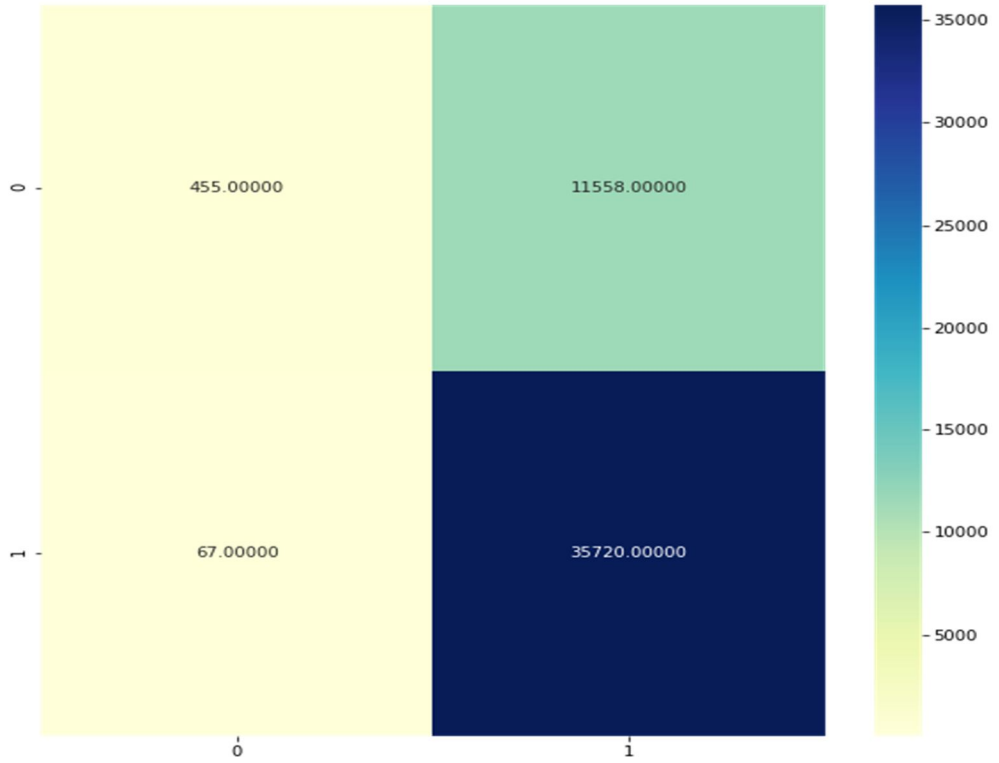


Fig 13. Visualised Prediction Result of Logistic regression model through Confusion Matrix

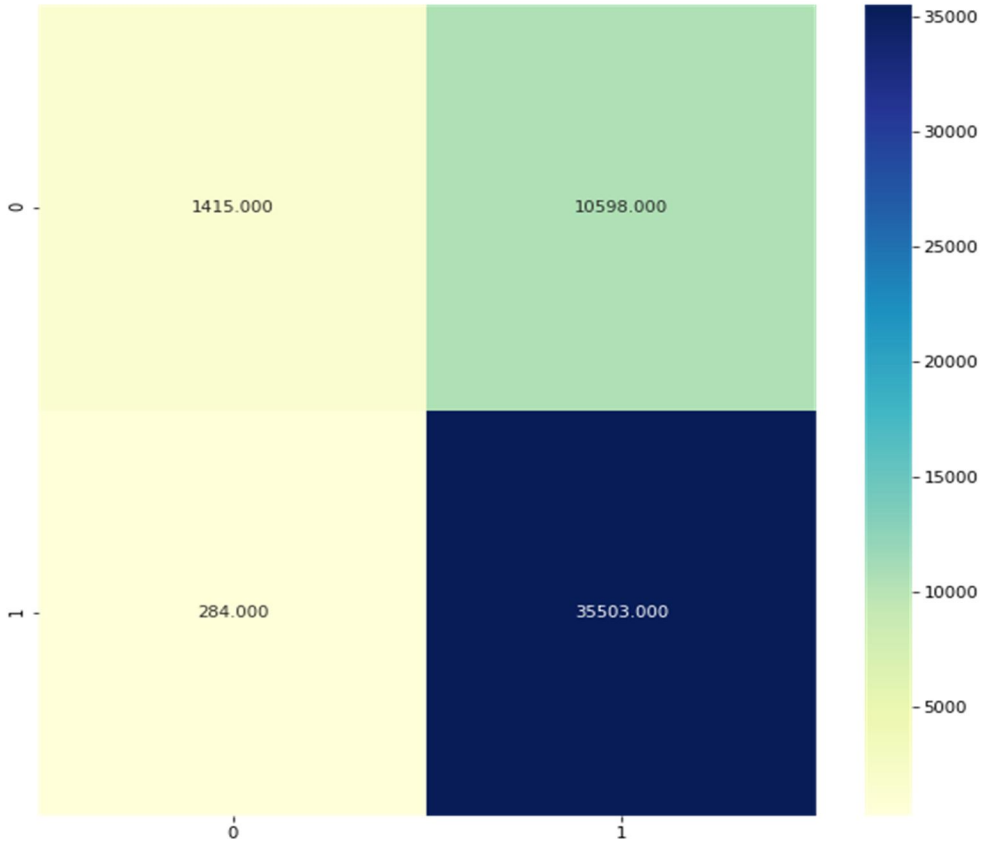


Fig 14. Visualised Prediction Result of Support vector Machine through Confusion Matrix

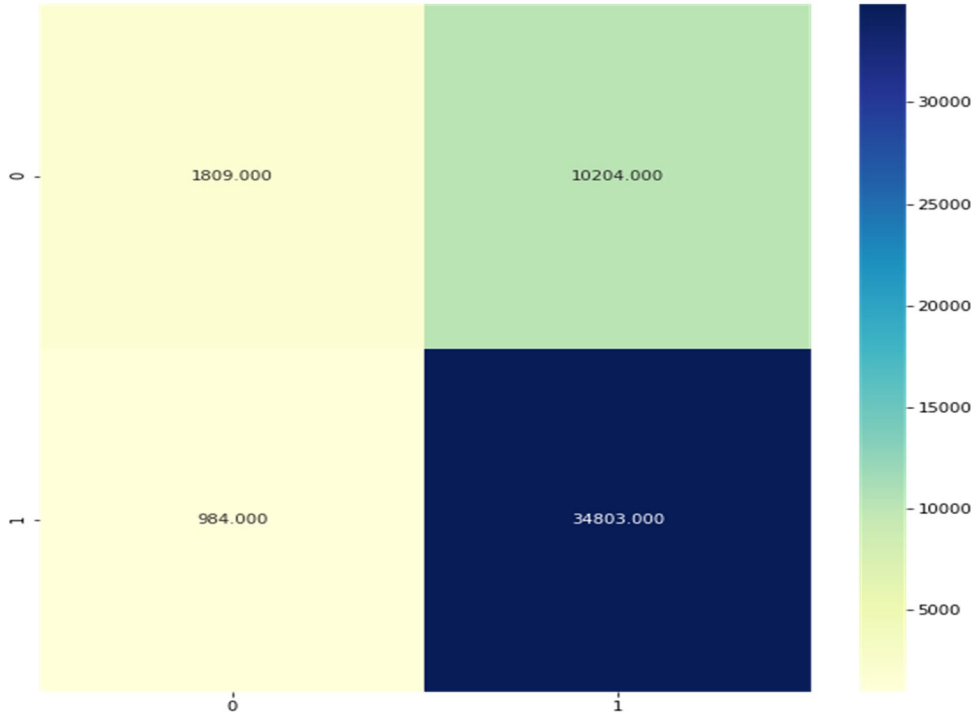


Fig 15. Visualised Prediction Result of k Nearest Neighbour Classifier through Confusion Matrix

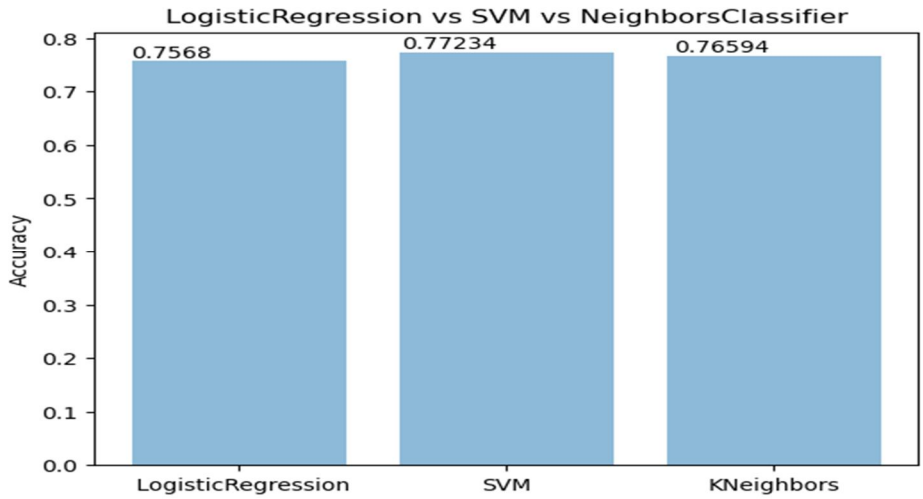


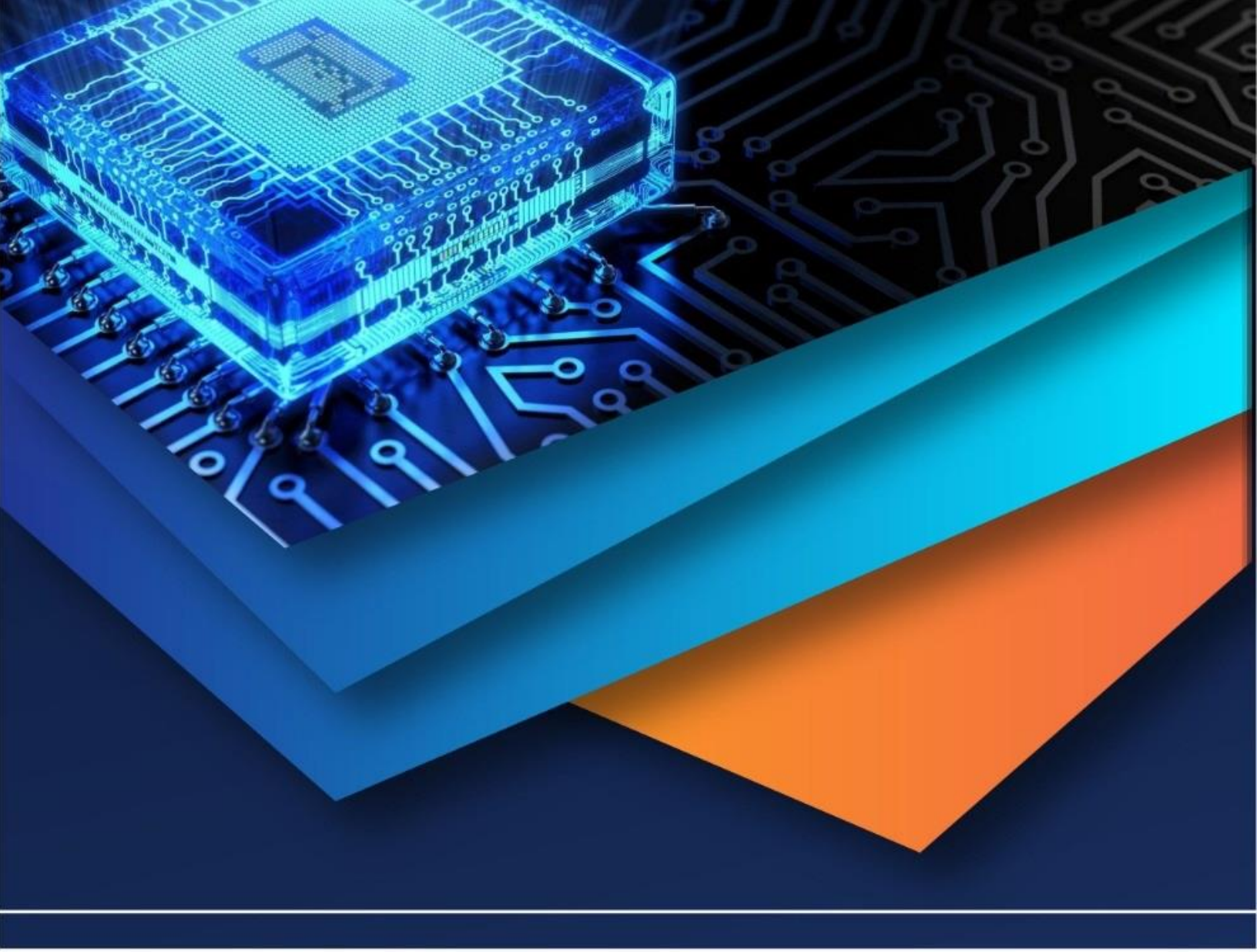
Fig 16. Comparative Chart of Accuracies of Algorithms

V. CONCLUSION AND FUTURE SCOPE

The key idea behind approaching drug-illness prediction as either a challenge will be to develop a perfect drug-illness relationship by integrating prior knowledge about the treatment and the disease. A number of solutions have been suggested and developed for the task. As a consequence, our technique may be improved in the future. Because these computational procedures are often utilized with little quantities of data, the time needed to process them increases fast as the number of data increases. The dataset would be upgraded over time, and even the algorithm will get easier to predict as time passes. This method of prediction of drug-disease association using Machine Learning and NLP can be considered as next best method for implementation of techniques which use only sentimental classification as it gives better accuracy and predictive results. Based on patient input, the project's purpose is to predict which treatment (drug) has been the most beneficial one. This system has a wide variety of applications: Drug Prediction Automation and to improve accuracy in order to help them in the future, and also manage disease-related information.

REFERENCES

[1] K. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabási, ‘‘The human disease network,’’ Proc.Nat.Acad.Sci.USA,vol.104,no.21, pp. 8685–8690, 2007.
[2] L. Weng, L. Zhang, Y. Peng, and R. S. Huang, ‘‘Pharmacogenetics and pharmacogenomics: A bridge to individualized cancer therapy,’’ Pharmacogenomics, vol. 14, no. 3, pp. 315–324, 2013.
[3] Y. H. Li et al., ‘‘Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs,’’ Briefings Bioinf., 2019.
[4] Y. H. Li et al., ‘‘Therapeutic target database update 2018: Enriched resourceforfacilitatingbench-to-clinicresearchoftargetedtherapeutics,’’ Nucleic Acids Res., vol. 46, no. D1, p. D1121, 2018.
[5] H. Yang et al., ‘‘Therapeutic target database update 2016: Enriched resource for bench to clinical drug target and targeted pathway information,’’ Nucleic Acids Res., vol. 44, no. D1, pp. 1069–1074, 2016.
[6] S. J. Cockell et al., ‘‘An integrated dataset for in silico drug discovery,’’ J. Integr. Bioinf., vol. 7, no. 3, pp. 15–27, 2010.
[7] S. Naylor and J. Schonfeld, ‘‘Therapeutic drug repurposing, repositioning and rescue PartI: Overview,’’ Drug Discovery World, vol.16,pp.49–62, Dec. 2014.
[8] F. Zhu, X. X. Li, S. Y. Yang, and Y. Z. Chen, ‘‘Clinical success of drug targets prospectively predicted by insilicostudy,’ ’Trends Pharmacol.Sci., vol. 39, no. 3, pp. 229–231, 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY