MEDICINE RECOMMENDATION SYSTEM FOR E-COMMERCE

MINI-PROJECT REPORT

Submitted by

Dilli B (20G022)
Sankara Narayanan S (20G088)
Rina E Veronica (20H039)
Thanush AA (20H055)

BACHELOR OF ENGINEERING/TECHNOLOGY

in

Mechanical and Computer Science and Business Systems

THIAGARAJAR COLLEGE OF ENGINEERING, MADURAI – 625 015
(A Government Aided Autonomous Institution Affiliated to Anna University)

August 02 to September 22, 2021

THIAGARAJAR COLLEGE OF ENGINEERING, MADURAI- 625 015
(A Government Aided Autonomous Institution Affiliated to Anna University)

TEQIP Phase III: Student Training on Future Skill technologies
(online mode)

BONAFIDE CERTIFICATE

Certified that this Mini-project report Medicine recommendation system is the bonafide work of **Dilli B (20G022), Sankara Narayanan S (20G088), Rina E Veronica (20H039), and Thanush AA (20H055)** of 3rd sem. Department of Mechanical and CSBS, who attended the NPIU MHRD supported Future Skill Training on **'Artificial Intelligence and Machine Learning'** in online mode and carr out the mini-project work under my supervision between Aug. 02 – Sep. 22, 2021

Submitted for Evaluation held at Thiagarajar College of Engineering on September 22,2021

TEQIP Coordinator
TCE, Madurai

Faculty trainer
(Dr. M.Suguna, Assistant Professor – CSE Dept.)

Abstract

We learned about the basic concepts in Artificial Intelligence (AI) and Machine Learning (ML) in this course organized by NPIU MHRD. We learned theoretical topics like Depth First Search (DFS), Breadth-First Search (BFS), Uniform Cost Search (UCS), Informed Search Algorithms, and so on. These search algorithms gave us exposure to the basics of Artificial Intelligence (AI). Also, these topics gave us a small exposure on the topic of Data Structures and Algorithms (DSA). Later on, we learned topics related to Machine Learning (ML). This includes steps involved in Machine Learning Application: Data Collection, Data Preparation, Choosing a model, Training a model, Evaluation, Hyperparameter Tuning, and Prediction. We also learned in detail about each and every component of Machine Learning like Data, feature-parameters, and algorithms. We came to know about sites where we can get datasets like Kaggle and UCI repository, and also resources where we can explore research papers like SciHub. We learned about the types of Machine Learning Algorithms like Supervised Learning Algorithm, Unsupervised Learning Algorithm, and reinforcement Learning algorithms. We learned about Neural Networks and how it works. Later on, we had sessions that were more practically oriented, we implemented algorithms like, linear regression, Logistic Regression, K-nearest Neighbors, Decision Trees .etc. in python. Through these practical sessions, we learned how these models work. We were taught about various python packages like pandas, matplotlib, sklearn, seaborn etc. We learned how to visualise data using matplotlib. We learned about various methods to pre-process data and metrics that are used for measuring accuracy like Confusion Matrix, Mean Absolute error, and many more. We learned about their advantages and disadvantages and under which situations do we need to implement these metrics. We formed a team of four members for the mini-project. We learned a lot of concepts other than those taught in the course like text analysis, transfer learning, Chi2 distributions, and Using TdFidVectorizer module to calculate the Inverse Document Frequency (IDF) of descriptions to simplify the analysis process of the model, Naïve Bayes Algorithm, Linear SVC model, and Random Forests. Later on, we learned about dimensionality reduction and Principal Component Analysis (PCA) in detail. We also learned about the application of PCA using python.

Table of Contents

List of Figures

List of Abbreviations

AI	Artificial Intelligence
AI	Artificial Intelligence
ML	Machine Learning
IDF	Inverse Document Frequency
AD	Alzheimer's Disease
MRI	Magnetic Resonance Imaging
SVM	Support Vector Machines
SVC	Support Vector Classifier
EHR	Electronic Health Records

CHAPTER 1
INTRODUCTION

Machine learning (ML) was considered as an integral part of Artificial Intelligence (AI), also a data analysis technique that computerizes the explanatory model structure. In most scenarios, based on the learning method, two types of ML algorithms (supervised & unsupervised) were used .At present, these algorithms are engaging in all the major industries like healthcare, banking, transport, social media, etc. .Above all, the medical industry is advancing quickly with high volumes of information and increasing difficulties in inventory and patient outcomes. Economically developed nations such as the USA, Japan, European countries are even facing problems with the enormous collection of medical data . However, by using conventional techniques, it is not possible to analyze this significant volume of information because of time consumption and efforts. Therefore, ML techniques are coming up with various algorithms and programs to avoid these issues. Besides that, the selection of a proper algorithm is not an easy task since it depends on multiple factors such as data volume, information type, and outcomes related to industry requirements . Nowadays, ML algorithms are progressively utilized in neuroimaging studies like a prediction of Alzheimer's disease (AD) from auxiliary MRI. Also, many studies attempted different ML strategies in predicting AD and their causes . In the study of AD prediction and retrieval, a multistage classifier utilizing ML, including Naive Bayes classifier, support vector machine (SVM), and K-nearest neighbor (KNN) was used to group Alzheimer's illness in the more acceptable and effective way.

Similarly, a study concluded that the utilization of locally linear embedding (LLE) kind of unsupervised learning was utilized to categorize AD based on fundamental MRI data. Besides, some preliminary studies with ML techniques concluded that these methods are valid and accomplished with high precision (up to 98%) in diagnosing clinical events with analysis of patient medical records .

Similarly, we built a machine learning model that uses supervised machine learning methods to suggest a best alternative medicine that cures the same disease with no or less side-effects. In other words, There are many medicines available in the market manufactured by different

pharmaceutical companies where people do not know the same formula works with the composition of the medicine. Whenever a required medicine isn't available , effective, safe alternative medicines are to be suggested to the user. During times of high demand for medicines and when stocks run out soon, this system would be very helpful to patients. With E-pharmacy receiving a great motivation in the pandemic, patients received their medication despite lockdown. Nowadays people are aware of E-pharmacy , so our model will be very useful for all the normal people.

CHAPTER 2
BACKGROUND

Time and tide waits for none. Medicines are important and must be given and taken timely. Unavailability of medicines is a commonly faced problem and affects people of all ages from toddlers to old people. When suggested medicines are not available, alternative medicines are sure available but some people might not really know them or they might be confused as to which alternative medicine they must go for, and it is really important to know what the medicines they are going to go for and consume does when they face the situation, people must be made aware as to whatever medicine they are being suggested. A medicine recommendation system that suggests the best alternative medicine with proper description of its purpose when the first preferred one is not in stock quickly and effectively would solve the dilemma. Most medicines have the same properties and would be manufactured by different names and companies. And even though most is known to people , some might not have been aware of it and a recommendation system cannot lag in educating people with the medicines available in the market with same or similar properties. To some people, this might be a very precious piece of information since the chances of having the alternative medicine available in their nearest pharmacy more as compared to the medicine their doctor suggested is not zero. When the medicine the doctor suggested is not available at the time, then there is no other choice but to search for another medicine that would cure the ailment. The Medicine recommendation that is going to be talked about a lot here is designed to be fast, efficient and accurate. The best preferred medicine would be analyzed and be suggested to the user.

CHAPTER 3
PROBLEM FORMULATION

Hospitals have access to a vast amount of data about patients and their health parameters. Thus, there is a need for a convenient way for medical professionals to utilize this information effectively. An example would be the access to aggregated information from the existing database on a specific problem at the point of care when it is necessary. Moreover, there are more drugs, tests, and treatment recommendations (e.g. evidence-based medicine or clinical pathways) available for medical staff every day. Thus, it becomes increasingly difficult for them to decide which treatment to provide to a patient based on her symptoms, test results, or previous medical history. On the other hand, all this data can be used to strive for personalized healthcare which is currently on the rise and predicted to get a major disruptive trend in healthcare in the upcoming years. And moreover, in times of demand for medicines and when pharmacies and sites run out-of-stock of the most suggested one, people are left with no other choice but to go for other medicines that help cure their ailments. There are many medicines available in the market manufactured by different pharmaceutical companies where the user does not know the same formula works with the composition of the medicine. Whenever a required medicine isn't available, effective, safe alternative medicines are to be suggested to the user. During times of high demand for medicines and when stocks run out soon, this system would be very helpful to patients. With E-pharmacy receiving a great motivation in the pandemic, patients received their medication despite lockdown where this model can be implemented.

CHAPTER 4
LITERATURE REVIEW

We conducted our literature review in several steps. We followed the guidelines defined in . First, we defined search terms based on population, intervention, outcome of relevance and experimental design. However, we concluded that for our approach the population contains all healthcare facilities. Since this population is so comprehensive and non-specific, we excluded keywords about the population. This resulted in the following major keywords: Intervention: medication recommendation system Outcome of relevance: system for medication recommendation Experimental Design: empirical studies, systematic literature reviews, solution descriptions The intervention and outcome of relevance category are the same. Therefore, they were only included one time. Once this has been agreed on, the search algorithm was constructed. The logical operators AND as well as OR were used to combine the search terms defined in the previous step. The following synonyms were considered: Medication: Medication, drug, Drug Recommendation: Recommendation, Recommender, recommender System: System, framework, Framework, algorithm, Algorithm, engine, Engine This resulted in the following search algorithm: {{medication OR Medication OR drug OR Drug} AND {recommendation OR Recommendation OR recommender OR Recommender} AND {system OR System OR Engine OR engine OR framework OR Framework OR algorithm OR Algorithm}}. To verify the algorithm and the terms used, we (IJACSA) International Journal of Advanced Computer Science and Applications, conducted a test for some papers we already knew. The test was successful as we could find relevant papers. Afterwards, we chose the databases to search in based on available access which led to five databases: ACM Digital Library IEEExplore ScienceDirect Elsevier John Wiley Inc. We also agreed on using Google Scholar as a search engine as a sixth source because it provides results from a high variety of databases which we might not have included and thus, can lead to a higher quantity of relevant papers. Then, we agreed on inclusion and exclusion criteria which are defined as follows: Inclusion criteria:

1) Conference Proceedings and Journals published after 1999

2) Studies focusing on medicine recommendation systems in general and/or specified for any disease

3) Studies focusing on medicine recommendation systems based on graph databases

Exclusion criteria:

1) Papers published before 2000

2) Manuscripts written in another language than English

3) Technical reports and white papers as well as Graduation projects, Master thesis and PhD dissertation

4) Textbooks (print and electronic)

5) Studies in other domains of knowledge Finally, some quality criteria for the papers which met the inclusion criteria were defined to guarantee a selection of high quality papers only. A scored system was used. For each of the following criteria met, a paper is assigned one point: Logical and reasonable in results and findings regarding the domain of knowledge Clearly stated objectives, results and findings regarding the domain of knowledge Well-presented and justified arguments Reasonably tested and/or applied system Well referenced with a minimum of ten sources Only papers which met all criteria, thus had 5 points, were observed.

CHAPTER 5
PURPOSE OF THE WORK

The management and quality of hospital services depend to a large extent on individual medical decisions. For example, based on their experience, each physician may select treatments coded into specific keys for disease management and the cost of treatment; each of these treatments is then combined with a cost refund respectively which is encoded in specified keys in the Electronic Health Records (EHR) and on medical bills. These keys would have an ‘economic effect’ on the health system. The reimbursement system's behavior, in some cases, enables physicians to offer more health services to help a patient, regardless of whether the additional care is economically optimal. At the same time, physicians can also ignore their colleagues' experiences by using alternative treatments that might be more suitable for treating a disease. Other deficiencies may stem from the fact that physicians are either overloaded by a large number of patients and/or by several therapeutic options to consider, which is part of a portfolio that has more than 100 different alternative treatments per patient. Because of this, physicians do not usually have the time or knowledge to evaluate all of the different alternative options accurately and individually.

By using recommender systems, physicians can reduce their workload, yet still have each decision made under their control. They will also get an insight into how other physicians recommend a treatment in the same given situation.

Nowadays, recommender systems have proven to be invaluable for online users to cope with information overload in many different fields (e.g. e-commerce, decision support systems, etc.) . However, the architecture of recommender systems and their evaluation in real-world problems is still an active area of research. We have investigated and implemented a recommender system intending to transfer this technology in the field of medicine to support physicians seeking to predict the rating or preference of a treatment key for new patients.

Recommendation systems in medicine are not new. There are about 911 search items in published medical journals, with articles reporting recommender systems for patients using personal health systems. Until the last few years, most of these techniques were used for analyzing rich EHR data

based on traditional machine learning and statistical techniques such as logistic regression, support vector machines (SVM) , and random forests. However, recently, deep learning techniques have achieved great success in many domains through deep hierarchical feature construction and capturing long-range dependencies in data effectively, particularly for the analysis of EHRs in medicine. For a recent review article on recommender systems in healthcare see e.g. Tran et al.

Here, we are introducing a new type of recommender system which is a combination of methods to generate synthetic populations while keeping personal data protected when needed for Artificial Intelligence (AI) applications and the use of novel methods based on continuous-valued logic and multi-criteria decision operators aimed for robust, safer, and more understandable use of deep learning. By combining neural networks with continuous logic and multi-criteria decision-making tools, thus reducing in this way the black-box nature of neural network models .. By doing so, we are exploring the (often overlooked) possibility of combining neural networks with continuous logical systems. This strategy provides a clear advantage in the medical field since it is a system that, due to its nature, can be easily understood by physicians and/or medical practitioners, who often make their decisions relying on continuous logical rules. We aim to reach more transparency of AI applications in medicine while preserving efficient deep learning methods. The synthetic populations are mainly used for training the Deep Learning machinery. They are completely anonymized yet keep their original structure of the original data sample used for ML use.

Our customers consist of physicians handling individual patient diagnoses. In this way, by providing these recommendations, we are seeking to reach both an economic and clinical efficacy. Persuading patients to make use of best-suited treatment keys aims to reduce the costs of patient management. Improve the management of a diagnosis of the disease by helping the physicians to discover additional keys that could improve the treatment of patients.

CHAPTER 6
OBJECTIVES

- To design a ML model that would recommend alternative medicines from the best manufacturer during the unavailability of preferred medicines.

- To suggest the best alternative medicine that cures the same disease effectively.

- To suggest cheaper and effective medicine to cure a particular disease.

- To show the superior medicine for a disease.

- Fast and accurate suggestions based on medical regulations.

- To develop a medical system where no compromises are made with the safety and effectiveness of the system.

- Best medical care for the patient with optimum economic impact (i.e. cost efficiency) for those who have to pay the bill finally. So, if there are two equal treatments available, the system should recommend the cheaper one with the same if not better treatment quality result than the more expensive option.

- Depending on the health system of a particular country and how billing/charging is carried out, there may be another conflict that the physician has to contend with: optimizing the cost-income ratio of the organization that is treating patients.

CHAPTER 7
SYSTEM DESIGN

We worked on a sample EHR dataset created only for testing purposes and it is obtained from Kaggle, the dataset contained Drug Name, Disease, Description, Manufacturer, NSE symbol, and its Rating. First, the dataset is subjected to pandas tool for cleaning, which is then subjected to train various machine learning models for prediction, with the help of confusion matrix, the accuracy of the classifiers were calculated, and the classifier with high accuracy were applied to the test dataset for predicting and recommending drugs to the patients. The proposed framework is shown in Fig. 6.1.

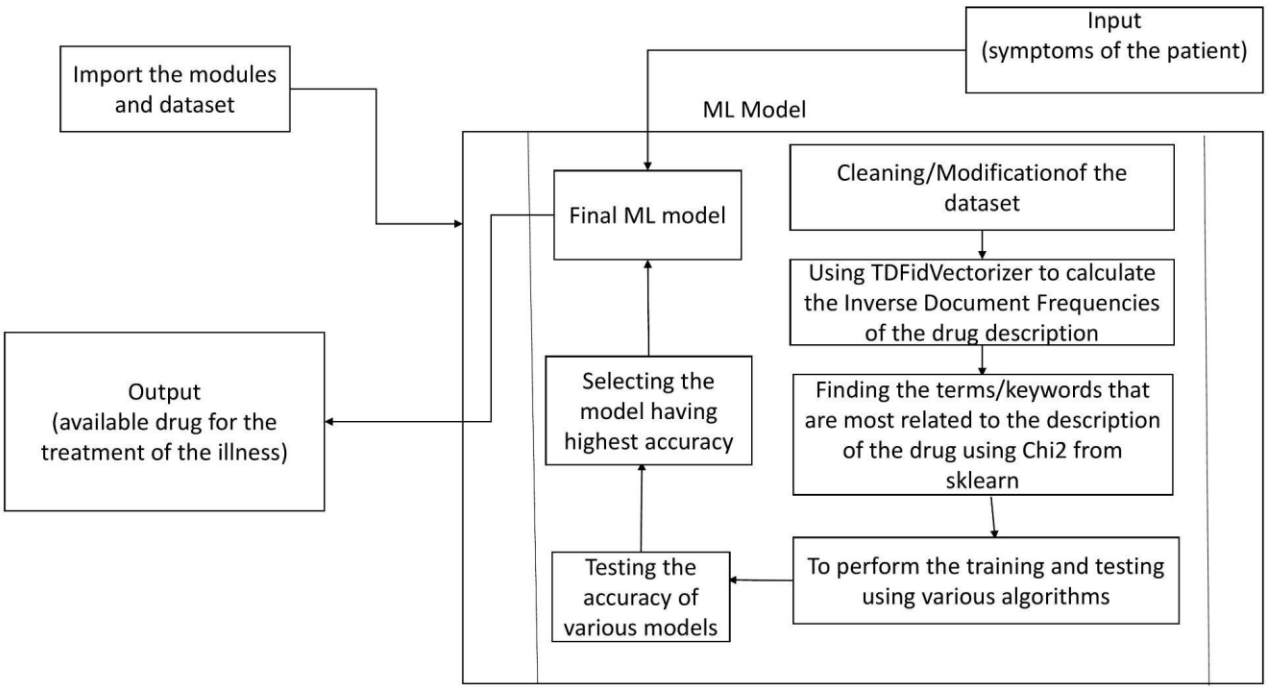


Fig 6.1.: Flowchart representation of the system design of the ML model

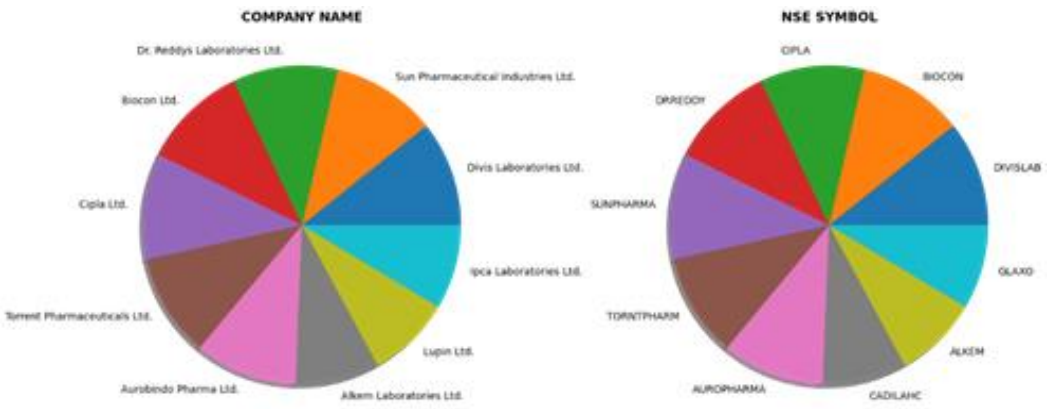
The main steps in providing the Medicine recommendation for the disease are as follows:

Step 1: Importing and cleaning Dataset

In this step, the dataset was imported and cleaned using pandas, a python based data analysis tool.

Step 2: Visualizing and Reduction of data.

In this step, the dataset is visualized using mathematical plotting functions like matplotlib to check the feasibility of data values. A total of 22481 testing records were visualized. Data visualizations help the data scientist to check the feasibility of the dataset attributes and their values. First figure shows the different manufacturers available. NSE symbol, ratings, and industry type is been visualized below respectively



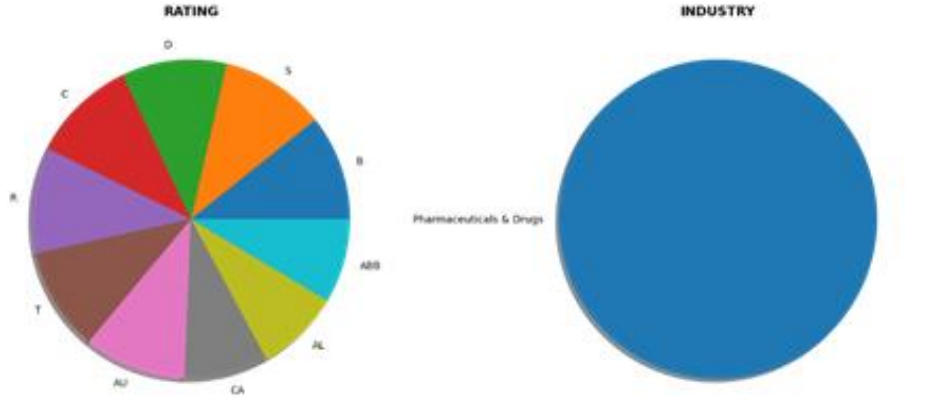


Fig 7.1. : Pie charts showing the distribution of Company name, NSE symbol, Rating, and industry type in the dataset

Step 3: Converting the text representation to vectors

In this step, the cleansed data is converted into vectors using TFID module. Because the computer can understand any data only in the form of numerical value. So, for this reason, we vectorize all of the text so that the computer can understand the text better.

Step 4: Applying supervised machine learning classifiers

In this step, the vectors are subjected to different machine learning classifiers such as Naive Bayes, logistic Regression, Support Vector Machine (SVM) and K-nearest neighbor.

SVM:

Support Vector Machines are a well-known ML technique for classification and other learning activities. SVM is a discriminative classifier and formally characterized by an optimal hyperplane. It produces an outcome of the optimal hyperplane, which classifies new examples and datasets that support hyperplanes are called support vectors [13]. In a two-dimensional (2D) region, this hyperplane is a line isolating into two segments wherein each segment lay on either side. For instance, multiple line data classification is done with two distinct datasets (i.e., squares and dots) and ready to propose an affirmative interpretation. However, the selection of an optimal hyperplane is not an easy job as it should not be noise sensitive, and generalization of data sets should be accurate [14]. Pertinently, SVM is trying to find an optimized hyperplane that provides considerable minimum distance to the trained data set . In mathematical notation, for 2D space, a line can distinguish the linearly separable data. The equation of the line is $y = ax + b$. By renaming x with x1 and y with x2, the equation will change to

$ax1 - x2 + b = 0$. If we specify $X = (x1, x2)$ and $w = (a, -1)$, we get $w \cdot x + b = 0$, which is called the equation of the hyperplane.

Step 5: Accuracy measurement

In this step, The accuracy of each of them shall be predicted using a metric and the best out of them would be chosen for implementation.

Step 6: Termination

Terminate the process when the machine learning classifier model is built for a recommendation.

CHAPTER 8
MACHINE LEARNING BASED MODULE

This section contains the experimental results and analysis of a recommendation system for multi-disease. A dual-core i3 system with 4 GB of RAM, pandas, NumPy, csv, SkLearn, TfidfVectorizer, and Matplotlib libraries was used in the experimental analysis for generating recommendations. Machine learning classifier algorithms were evaluated for their accuracy in recommending appropriate drug recommendations to overcome certain medical issues. The experimental analysis takes place in two steps. In the first step, the dataset is imported, cleansed and dimension reduced using the pandas tool, the second step involves training machine learning models for the recommendation system where the test data is subjected to various classifiers such as SVM, Native Bayes, Logistic Regression, and K-NN and accuracy is measured, then the classifier with high accuracy is used in predicting disease.
The description of the algorithm is as follows:

Naive Bayes

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Logistic regression

Logistic regression, also known as logit regression or logit model, is a mathematical model used in statistics to estimate (guess) the probability of an event occurring having been given some previous data. Logistic regression works with binary data, where either the event happens (1) or the event does not happen (0).

K-nearest neighbors

K-nearest neighbors (kNN) is a supervised machine learning algorithm that can be used to solve both classification and regression tasks. It is seen as an algorithm that comes from real life. People tend to be affected by the people around them: Our behavior is guided by the friends we grew up with.

SVM

Support Vector Machines are a well-known ML technique for classification and other learning activities. SVM is a discriminative classifier and formally characterized by an optimal hyperplane. It produces an outcome of the optimal hyperplane, which classifies new examples and datasets that support hyperplanes are called support vectors. In a two-dimensional (2D) region, this hyperplane is a line isolating into two segments wherein each segment lay on either side. For instance, multiple line data classification is done with two distinct datasets (i.e., squares and dots) and ready to propose an affirmative interpretation. However, the selection of an optimal hyperplane is not an easy job as it should not be noise sensitive, and generalization of data sets should be accurate. Pertinently, SVM is trying to find an optimized hyperplane that provides considerable minimum distance to the trained data set . In mathematical notation, for 2D space, a line can distinguish the linearly separable data. The equation of the line is $y = ax + b$. By rename x with x1 and y with x2, the equation will change to $ax1 - x2 + b = 0$. If we specify $X = (x1, x2)$ and $w = (a, -1)$, we get $w \cdot x + b = 0$, which is called the equation of the hyperplane.

CHAPTER 9
Results and Discussion

We have used an accuracy score metric to evaluate the accuracy of the models. The accuracy of the models which we have tested are given below:

1. Multinomial Naive Bayes - 0.963302752293578
2. Logistic Regression - 0.963302752293578
3. K - Nearest Neighbors - 0.1559633027522936
4. Linear SVC - 0.963302752293578

It was surprising to note that the accuracy scores of three of the models were exactly the same. K nearest neighbors model had the lowest accuracy score of about 15.596%.

We evaluated the accuracy of models again using the f1 score with weighted average as a parameter to make sure that the results given above are not due to some errors. The f1 scores of the models are given below:

1. Multinomial Naive Bayes - 0.966360856269113
2. Logistic Regression - 0.963302752293578
3. K - Nearest Neighbors - 0.20305810397553517
4. Linear SVC -0.963302752293578

This time, the Multinomial Naive Bayes model had the highest accuracy of 96.636%, followed by Logistic Regression and Linear SVC with the same accuracy of 96.33%. K - Nearest Neighbors had the lowest accuracy of 20.305%.

From the results which we have obtained, we can conclude that Multinomial Naive Bayes Classifier had the highest accuracy among the models which we have trained so far.

CHAPTER 10
Conclusion

In this work, a drug recommendation system for multi-disease using machine learning for healthcare was developed using a sample dataset which was created only for testing purposes. The proposed method using Multinomial Naive Bayes model showed that it can be an effective drug recommendation tool in healthcare. The machine learning classifiers used in the analysis include K-nearest neighbors,naïve bayes, linear svc, logistic regression to achieve accuracy and provide drug recommendations for the patients suffering from short-term disease. The experimental results showed that the proposed method using the Multinomial Naive Bayes yielded a higher predictive performance compared with the other classifiers under analysis. Based on the experimental results obtained, the proposed method is found to be effective in improving the quality of drug recommendation, thereby improving the healthcare industry.

CHAPTER 11
Future Enhancements

This project is meant to help people and serve as an extra helping hand at the time of crisis. If we successfully implement it, our new-found goals would be to bring this project to life over on different platforms. This Medicine Recommendation System would be made into an user-friendly application probably with updates to make the application useful for the public. The Medicine Recommendation system shall be made for the public accordingly so that it does not encourage self-medication , a website that can be accessed anywhere and updates to improvise the application from time to time. With more advancements and accuracy in medicine suggestions , this project can be implemented and brought into the online pharmaceutical industry.

This system would not only serve as a medicine recommendation system but also as a system that would tell the user about the medicines they want to know about, with just them having to enter the name of the medicine or a picture of the medicine's packaging. This recognition of a medicine through images would be an advancement that would be designed later in the future. This knowledge of medicines could be available to people in the form of an application. A part of this whole medicine recommendation system is the system that lets the user know the description of the medicine they are going to go for. This part of giving the user the knowledge of what the medicine they want to know about, can be separated and implemented as a whole project.

Since regular updation of medicines available in the market is implemented in the system. This could even be beneficial to doctors, letting them know about the new medicines in the market with same or similar compositions keeping the doctors open to suggesting their patients with alternative medicines themself when a patient asks them for suggestions. The system's main purpose is to be of any use to the public, and reaching out to people in any way is welcomed.

REFERENCES

1. https://www.springerprofessional.de/en/a-drug-recommendation-system-for-multi-disease-in-health-care-us/18279852
2. https://towardsdatascience.com/multi-class-text-classification-with-scikit-learn-12f1e60e0a9f
3. https://www.kaggle.com/saratchendra/medicine-recommendation
4. https://www.kdnuggets.com/2018/11/multi-class-text-classification-model-comparison-selection.html
5. https://www.researchgate.net/publication/350625836