A COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS

by Kinjal Sarkar

Submission date: 05-Apr-2022 09:01AM (UTC+0530)
Submission ID: 1802068480
File name: Major.docx (1.36M)
Word count: 9514
Character count: 51099

A **MAJOR PROJECT** report on :

A COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS

submitted in partial fulfillment of the requirements for the degree of
B. Tech
In
Electronics and Electrical Engineering
By

Souvik Karmakar	1807228
Sudeshna Dutta	1807232
Indrashis Mitra	1807274
Kinjal Sarkar	1807277
Pratyay Basu	1807291

under the guidance of
Prof. K.B. Ray

School of Electronics Engineering

April 2022

CERTIFICATE

This is to certify that the project report entitled **"A COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS"** submitted by

Souvik Karmakar	1807228
Sudeshna Dutta	1807232
Indrashis Mitra	1807274
Kinjal Sarkar	1807277
Pratyay Basu	1807291

in partial fulfilment of the requirements for the award of the **Degree of Bachelor of Technology** in **Electronics and Electrical Engineering** is a bonafide record of the work carried out under my guidance and supervision at School of Electronics Engineering, KIIT (Deemed to be University).

Signature of Supervisor
Prof.K.B.Ray
School of Electronics Engineering
KIIT (Deemed to be University)

The Project was evaluated by us on _____

EXAMINER 1 EXAMINER 2

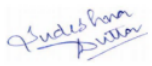
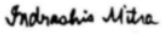
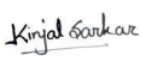
EXAMINER 3 EXAMINER 4

ACKNOWLEDGEMENT

We feel immense pleasure and feel privileged in expressing our deepest and most sincere gratitude to our supervisor **Prof. K.B. Ray**, for her excellent guidance throughout our project work. Her thoughtfulness, devotion, hard work, and attention to detail have been a source of great inspiration for us. Our sincere gratitude to you, ma'am, for your unending support and patience. We would especially want to thank her for her assistance in painstakingly and meticulously revising all of our papers.

We are really grateful to **Prof. Suman Roy, Prof. Pravat Biswal, Prof. M. Ramana** FIC project (EEE), Associate Dean **Dr. Amlan Datta and Prof. Suprava Pattanaik**, Dean (School Of Electronics) for their support and suggestions during our course of the project work in the final year of our undergraduate course.

STUDENT SIGNATURE:-

Roll Number	Name	Signature
1807228	**Souvik Karmakar**	
1807232	**Sudeshna Dutta**	
1807274	**Indrashis Mitra**	
1807277	**Kinjal Sarkar**	
1807291	**Pratyay Basu**	

Date:-

ABSTRACT

Indian pharmaceutical companies are the world's leading provider of generic medicines. Over 50% of the world's vaccine need is supplied by the Indian pharmaceutical sector, which accounts for 40% of pharmaceuticals needed in the United States and 25% of all medicines in the United Kingdom. Pharmaceutical production in India ranks third globally in terms of volume and fifteenth globally in terms of value. The domestic pharmaceutical industry consists of over 3,000 medical firms and 10,500 production facilities.

India is a major player in the global pharmaceutical sector. There is a large hub of budding engineers and scientists in the nation that can take the business to the next level. Antiretroviral drugs manufactured in India currently account for more than 80% of the world's AIDS medicine supply.

However,people in India are taking desperate steps to keep loved ones alive as a disastrous spike of new coronavirus infections overwhelms the country's health-care infrastructure. They are turning to dubious medical therapies in some circumstances, and to the underground market for life-saving pharmaceuticals in others.

Hence our project aims to solve this by developing a medicine recommendation system so that pharmacists can know in advance what medicines are the medicines being bought together the most,and keep stocks accordingly. We do not recommend which medicine is to be taken,rather the focus is on finding the drug combinations bought frequently together so that an idea can be had of the most-selling medicines;thus keeping their stock replenished would eliminate the black market,help to earn profits and help the customers.

PROPOSED SYSTEM: Our primary goal is to use Machine learning to aid in medicine supply due to good accuracy[15]. Using the support metrics of the Apriori algorithm, we plan to make a recommendation system of the medicine a particular customer is most likely to buy so that there is a win-win situation for both the customer and the shop owner - the customer gets the most appropriate medicine they want, at all times and do not have to face the hassles of out of stock medicines; while the pharmacist also learns the particular combination of medicines, which is made available easily, will yield the maximum benefit in the upcoming future. Also,we intend to make a comparison of Apriori and Eclat algorithms to understand the differences in both of the recommendation methods. Furthermore,we compare classification algorithms based on different parameters so that we have an understanding of which algorithm behaves better.

TABLE OF CONTENTS

Abstract

Table of Contents

List of Figures

List of Tables

List of symbols/ abbreviations

CHAPTER I: INTRODUCTION 1-3

1.1 Background 1

1.2 Literature Survey 1-3

1.3 Organization 3

CHAPTER 2: BASIC CONCEPTS

2.1 Introduction 4

2.2 How does Machine learning work? 5-7

2.3 How do machines learn? 7

2.4 Types of Machine Learning algorithms 8

2.5 Problems in supervised learning 8

2.6 What is classification? 8-9

2.7 What is recommendation? 9-10

2.8 Dataset 10

CHAPTER 3: PROJECT ANALYSIS

3.1 Project conceptual analysis 9-12

3.2 Project implementation 12-14

3.3 Working of the Apriori model 14-15

3.4 Working of ECLAT model 15-18

3.5 Dataset features 18-19

3.6 Proposed model 19

3.7 Classification models 20-23

3.8 Evaluating Classification models 23-26

CHAPTER 4 : RESULTS ANALYSIS

4.1 Apriori model 27-30

4.2 ECLAT model 31-34

4.3 Classification models 34-37

CHAPTER 5 : CONCLUSION

5.1: Summary 38

5.2: Future scope 39-41

5.3: Constraints 42

5.4: Social Impact 42

CHAPTER 6 : PLANNING & REFERENCES 40-44

Project Summary **44-45**

Publications **46**

LIST OF FIGURES

Fig no.	Description	Page No.
2.1	Cycle of Machine Learning	4
2.2	Components of learning	5
3.1	If-else concept	12
3.4	Working of Apriori model	19

3.5	Decision tree strucure	21
3.6	Working of Random forest	23
3.7	Bias	24
3.8	Example of Variance	25
4.1	Visualisation of ECLAT model results	31
4.2	Comparision of various classifiers	36

LIST OF TABLES

Table ID	Table Title	Page
7.1	Showing details about project planning and management	49

LIST OF SYMBOLS / ABBREVIATIONS

Symbol / Abbreviations	Description
ML	Machine Learning
RNN	Recurrent Neural Network
ODE	Ordinary Differential Equation
kNN	k Nearest Neighbour
ECLAT	Equivalence Class Clustering and bottom-up Lattice Traversal
DNA	Deoxyribonucleic acid
UI	User Interface
UX	User Experience

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

Healthcare is now one of the fastest-growing businesses in the world, and it is experiencing a complete transformation and revolution on a global scale. Global healthcare costs, which are now projected at $6 trillion to $7 trillion, are predicted to hit $12 trillion in only seven years, according to Russell Reynolds and Associates [1].

Machine learning is a sub-discipline of computer science which originated from the study of data patterns and an artificial intelligence based computational learning theory. It rose from an environment that was the integration of the interaction between available data, computing power, and statistical methodologies. The perfect blend in these three widely differing and rapidly developing areas gave birth to what is now known as machine learning .It's a first-class ticket to today's most exciting data analytics jobs.[2].

Deep learning, support vector machines, inductive logic programming, reinforcement learning, genetic algorithms, similarity and metric learning , sparse dictionary learning, and decision making , and so on are all sub-problems of machine learning[4]. K-nearest neighbour, Bayesian network,neural network,decision trees, decision rule, logistic regression, naive Bayes, and support vector machine are some of the machine learning techniques utilised in the medical field. In a perfect world, a model based on an example set would properly categorise an unknown example, which would need the model's capacity to generalise from the training set in a satisfactory way.

1.2 LITERATURE SURVEY

All across the world, healthcare plays an important part in improving people's health status and well-being. In healthcare data classification, ambiguity and high-dimensionality are two factors that add to the difficulty[3]. Patients, physicians, and medical treatments are all recorded in the healthcare big data set, which grows in volume so quickly that typical data analytics tools are not able to keep up with it and evaluate it effectively. Human experts have typically built and tested deep neural network models using a trial-and-error approach[10] .

Some machine learning technologies are used in conjunction with the big data analytics framework as a means of addressing these issues. Data mining has emerged as a critical study issue in the advancement of computing applications in health care and biology[5].Patients with difficult-to-treat diseases have a better chance at recovery because of advances in medication discovery and development made possible by technological advances. Several large tech companies, including IBM and Google, have developed machine learning tools that can help doctors uncover novel treatment options for patients[17]. Precision medicine is a significant concept in this discussion since it involves developing new ways to treat complex disorders and uncovering the underlying causes. However, even though several semi-supervised approaches have been presented to give additional training data, automatically produced labels are frequently too noisy to adequately retrain models[2]. The impact of COVID-19 pandemic on healthcare was catastrophic,mainly due to lack of preparedness. Hence in this project we have tried to make things easier in whatever way we can.Yoo et al. investigated the benefits and drawbacks of using data mining techniques in the biomedicine field[6]. Yadav et al. employed ordinary differential equations (ODEs) to generate time series to facilitate enhancement of RNNs[7].

Mainly we propose a model to help patients . COVID-19 has a greater risk of serious problems in some susceptible groups, such as the elderly, fragile, or those with several chronic illnesses. Using such a classification we can implement a variety of measures for their betterment,such as a vaccine scheduler[6]. Or ,as all of us know,shortage of medicines was a huge factor behind the large number of deaths we have witnessed. Hence our project also proposes a method to resolve this,by applying machine learning techniques to stock up medicines,which have been observed to be of significant demand,so that there is no dearth and we can give them to those in need.Big Data and the Cloud are two examples of new technologies that are helping to solve healthcare issues. Healthcare data is expanding at an exponential rate these days, necessitating an efficient, effective, and timely solution to cut mortality rates[8].

The importance of data collection, processing, integration, and reporting of underlying knowledge has been emphasised in the development of the notion of business intelligence and analysis, as well as how this understanding can assist in making more appropriate business decisions and gaining a vivid knowledge of market behaviours and trends. Efforts to automate Pap smear and colposcopy screening have already been

attempted, and a review of Pap smear was published in 2018[22]. A humongous growth of the data has facilitated us to reveal the hidden truth from data. Big Data analysis can be used for efficient decision making in the medical domain by some modifications to the existing machine learning algorithms [3]. The ability to detect these problems automatically and with high accuracy might considerably improve real-world diagnosis processes[23]. According to our findings, many academics are motivated to study machine learning algorithms in the health-care industry. However, selecting the appropriate algorithm to predict disease based on the data set generated by the researcher is always tough[10]. Chronic diseases can be diagnosed faster through modern technologies[16].

1.3 ORGANIZATION OF THE REPORT

This report has been divided into 6 chapters: -

Chapter 1 – Introduction

Chapter 2 - Ideation

Chapter 3 - Project Implementation

Chapter 4 - Results and Discussion

Chapter 5 – Conclusion & Future Scope

Chapter 6 – Planning & References

CHAPTER 2

IDEATION

Objective - To leverage machine learning to predict customer behavior patterns concerning buying medicines

Introduction

Machine learning is a branch of computer science that arose from the study of data patterns and a computational learning theory in artificial intelligence[11]. It rose from an environment that was the integration of the interaction between available data, computing power, and statistical methodologies. The perfect blend in these three widely differing and rapidly developing areas gave birth to what is now known as machine learning. Growth in available data compelled a spurt in computing power, which in turn stimulated the development of statistical methods to analyze large datasets, thus facilitating the collection and analysis of even larger and more complex, interesting data.

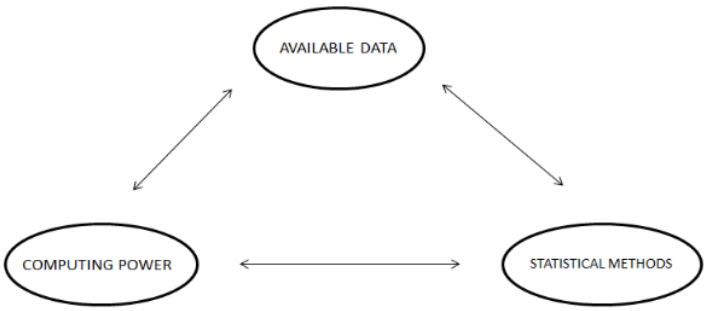


Fig 2.1 – Cycle of Machine Learning

WORKING OF MACHINE LEARNING :

Machine learning is made up of 3 parts:-

1. The computational algorithm that works at the crux of making determinations
2. Features and variables that influence the decision
3. Base knowledge for which the answer is known, which helps(trains) the system to learn

The model is given parameter data for which the solution is known at the start. The algorithm is then run, and changes are made until the output (learning) of the algorithm matches with the known solution. At this step, the system is given increasing volumes of data as input to help it learn and process higher computational judgments.

How do machines learn?

The fundamentals of learning are similar. It may be broken down into three sections:

• Data input: It establishes a factual foundation for later thinking through observation, memory storage, and recall.

•Abstraction: It entails the transformation of data into more generalised representations.

•Generalization: It is built on the foundation of abstracted data.

Fig 2.2 - Components of learning

Types of Machine Learning Algorithms :

Based on the type of input available during the training process or the desired outcome, there are 4 main types of machine learning algorithms:-

1. Supervised learning - used in those situations where the output is known, for a particular input; i.e. trained on labeled examples
2. Unsupervised learning - used in those situations where the output is not known, for a particular input;i.e. trained on unlabelled examples

3. Semi-supervised learning - works in those situations in which the combination of supervised and unsupervised learning is required to generate appropriate function or classifier
4. Reinforcement learning - is like a reward/punishment kind of a scenario. Desired manners are rewarded, while undesired ones are punished. Thus the agent behaves in such a way that a sequence of actions that lead to desirable outcomes are produced more times.

PROBLEMS AND ISSUES IN SUPERVISED LEARNING

Before getting started,we should judiciously pick an algorithm for use. While picking one,we need to be wary of the following factors -

1. Heterogeneity of Data: Many algorithms, such as support vector machines and neural networks, need homogenous numeral and normalised feature vectors[21]. Because algorithms that employ metrics of distance are extremely vulnerable to this, these approaches should be used only as a last resort if the data is diverse. Decision Trees provide a lot of flexibility when it comes to handling diverse data.
2. Data Redundancy: If the data contains redundant information, such as strongly correlated values, distance-based approaches are worthless due to numerical instability. In this circumstance, data can be subjected to some form of regularisation to avoid this problem.
3. Dependent Features: When feature vectors are dependent on one another, algorithms that monitor complicated relationships, such as Neural Networks and Decision Trees, perform better than other algorithms.
4. Bias - Variance Tradeoff : An algorithm for learning is biased for a particular input x if it is mostly wrong when anticipating the proper output for "A" when trained on each of these data sets, whereas a learning algorithm has high variance for a specific input "A" if it predicts different output values when trained on different training sets. The total of bias and variance of the learning process may be connected to the prediction error of a learnt classifier, and neither can be large since the prediction error would be high. Machine learning algorithms have the ability to automatically control the balance between bias

and variance, or manually tweak the balance using bias parameters, and adopting such methods will remedy this dilemma.

5. The Dimensionality Curse: The machine learning algorithm may be confused by the huge number of dimensions if the problem has a big number of dimensions and the problem only depends on a subspace of the input space with modest dimensions., resulting in a high variance method. In reality, the accuracy of the trained function is likely to increase if the data scientist can manually eliminate unnecessary characteristics from the input data. In addition, several feature selection techniques, such as Principal Component Analysis for unsupervised learning, aim to discover the important characteristics while excluding the unnecessary ones[30]. This decreases the number of dimensions.
6. Overfitting: As a result of human or sensor failures, the programmer should be aware that the output values might contain inherent noise. The algorithm should avoid attempting to infer a function that exactly matches all of the data in this case. When data is fitted too precisely, overfitting occurs, resulting in the model answering perfectly for all training examples but with a very high error for unknown samples. Two practical strategies to avoid this are to stop the learning process early and add filters to the data in the pre-learning phase to limit noises.

What is recommendation?

Recommendation engines are a type of machine learning that deals with ranking or evaluating items or people. A recommender system is a system that anticipates the ratings that a user will give to a certain item. After then, the predictions will be rated and returned to the user.

They are often utilized by huge corporations like Google, Instagram, Spotify, Amazon, Reddit, and Netflix to boost interaction with users and the platform. Spotify, for example, would propose tracks similar to those consistently listened to or enjoyed so that user may continue listening music on their site. Amazon utilises recommendations to recommend goods to different users depending on the data that Amazon has acquired for that user.

Recommender systems are frequently seen as a "black box," with the models developed by these major corporations being difficult to comprehend. The produced results are

frequent recommendations for the user for things that they need / desire but are unaware that they need / want until it is recommended to them.

There are several techniques to develop recommender systems. Some employ algorithmic and formulaic approaches, such as Page Rank, while others use more modeling-centric approaches, such as collaborative filtering, content-based, link prediction, and so on. The complexity of each of these techniques varies, but complexity does not imply "excellent" performance. Simple solutions and implementations frequently produce the best outcomes.

What is classification?

The act of identifying, analysing, and categorising objects into specified groupings, often known as "sub-populations," is known as classifications. Machine learning systems translate future datasets into appropriate and relevant categories using pre-classified training datasets.

In machine learning, input training data is utilised by classification algorithms to predict the likelihood or probability that subsequent data will fall into one of the predefined categories. Today's main email service providers utilise categorization to divide emails into "spam" and "non-spam" categories, which is one of the most prevalent uses of classification. This is a first step toward classification based on similarity, which has significant clinical implications for computer-assisted diagnosis[27].

Categorization is a sort of "pattern recognition," to put it another way. In this situation, the same pattern (similar number sequences, phrases or attitudes, and so on) is detected in future data sets by classification algorithms that were applied to the training data.

CHAPTER 3

Project Implementation

DATA PREPROCESSING :

The preprocessing of the dataset is very much required to support the regulations and syntax which the particular ML model asks for.

The different phases of the preprocessing include :

- Importing of libraries
- Importing the dataset
- Handling the missing data
- Encoding the categorical data
- Encoding the dependent variable
- Splitting the dataset (Training and test set)
- Feature scaling

IMPORTING THE LIBRARIES :

For general use cases the following libraries are imported to support the model structure:

1. **NUMPY**: It will facilitate working on arrays.
2. **MATPLOTLIB**: It will allow plotting of very attractive graphs for visual representation.
3. **PANDAS**: It will allow us to not only import the datasets but also create the matrix of features and dependent variable vectors.

IMPORTING THE DATASET :

A new variable is to be created which will contain the exact copy of the dataset we are aiming to deploy. The next target will be creating a data frame. We need to call a function from the panda's library that is read_csv.

This data frame now created will be the same as the dataset variable.

This is not enough as the need is to create 2 more entities that are:

13
Matrix of features and the dependent variable vector.

In most of the ML models, the dependent variable is at the end of the dataset and the beginning comprises the matrix of features.

HANDLING THE MISSING DATA :

13
We try to replace the missing values with the average of all the values in that particular column.

We take help from a reputed data science library that is SCIKIT LEARN.

Inside that, we take the help of a module named IMPUTE. We now create a tool/method in the object imputer to connect the object to the matrix of features.

Now the imputer transform replaces the missing values with the mean value and is stored or returned to the dataset portion which had the missing values.

ENCODING THE CATEGORICAL DATA

We try to encode the strings to certain numbers to let the ML model understand them and establish a correlation between them.

The encoding procedure used here is *ONE HOT ENCODING* (which allows the representation of categorical data to be more expressive)

ENCODING THE DEPENDENT VARIABLE

Label Encoder is a class in the pre-processing module of the scikit learn library that has no arguments if the dependent variable had only 2 categories. So we can just encode them into 1 and 0.

SPLITTING THE DATASET

The dataset is generally split into the Training set and Test set.

Training set: ML model has to be trained with the sets of data to identify attributes and features and patterns of the data.

Test set: To test the ML model with the new feature data to do performance monitoring.

We need to specify the test size to clarify how much % of the dataset we desire in the test set and the remaining in the training set.

We set the random_state as 1 to choose the data from the dataset randomly so that we just do not feel lucky just for this particular dataset.

FEATURE SCALING :

In ML models, some of the features dominate over some of the features. To tackle this problem feature scaling is important.

Feature scaling is not required for all ML models; it is a model-specific procedure.

The method of feature scaling is Standardisation and Normalisation.

Normalization is usually preferred and recommended where we have a normal distribution in most of the features.

We don't need feature scaling when the data is encoded or binary data.

ASSOCIATION RULE LEARNING :

Associative rule learning is a category of unsupervised learning that examines the dependency relationship between two data items and maps them appropriately so that benefit may be gained. It does so by looking for any relevant relationships or linkages among the dataset's variables. It also uses a variety of criteria to look for intriguing connections between the database's variables.

Working:-

Association learning works on the if-else concept.

Fig 3.1 If – Else concept

The If element of association is called the Antecedent.

The Then statement is called the Consequent.

This type of relationship is called Single Cardinality.

The key parameters to find the association is:

1. **Support,**
2. **Confidence and**
3. **Lift**

Support

The frequency of X, or the frequency with which a certain item shows up in the dataset, is known as support. The amount of the transaction T that is made up of the itemset X is what we use to calculate this value. The following may be expressed for transactions T if X datasets exist:

$$Supp(X) = \frac{Freq(X)}{T}$$

Confidence

The frequency with which a rule holds true is indicated by its confidence. In other words, number of times X and Y appear in a dataset where X is known. In other words, it's the set of items that include X divided by the number of transactions that include X.

$$Confidence = \frac{Freq(X,Y)}{Freq(X)}$$

Lift

Lift is the strength of a rule. It may be determined using the following formula:

$$Lift = \frac{Supp(X,Y)}{Supp(X) \times Supp(Y)}$$

Lift is the ratio between the actual support measure and the expected support measure assuming X and Y are independent of each other. It can be one of three things:

- **Lift = 1:** Antecedent and subsequent occurrence probabilities are independent of one another**.**
- **Lift > 1:** Determines the degree to which the two itemets are interdependent.
- **Lift 1:** It indicates that one object is a replacement for another, implying that one item causes harm to another.

Apriori Model specifications:

In this project, we have used the Apriori model for recommending the medicine combination that the customer is most likely to buy.

The Apriori algorithm, given by R.Agrawal and Srikant in the year 1994, uses recurrent itemsets to develop association rules. It is mainly designed to work on databases that comprise of transactions. Using these rules, it is possible to determine how weakly or how strongly these objects are interlinked.

To calculate the associations effectively, the Apriori algorithm uses a Hash tree and a breadth-first search to find frequent items from a large dataset, iteratively, also called level-wise search, wherein *k*-frequent itemsets are used to find *k+1* itemsets.

An important attribute referred to as the Apriori property is utilized to improve the level-wise generation efficiency of frequent itemsets by minimizing the search space.

Apriori Property is a term that refers to a property that exists before it.
All non-empty subsets of the repeated itemset must be frequent. The Apriori method relies heavily on the anti-monotonicity of the support measure.
In Apriori, it is assumed that -

- A frequent itemset's subgroups must all be frequent (Apriori property).
- If an itemset is rare, all of its supersets are rare as well.

What is a Frequent Itemset?

Frequent itemsets are ones that receive more support than the threshold limit or otherwise known as user-specified minimum support. In other words, if X and Y are the most common sets collectively, then X and Y should be the most common set separately.

Eg - Assume there are 2 itemsets X{2,4,7,8} and Y{3,4,5,8}. In these two transactions, {4,8} are the frequent itemsets.

Eclat model :-

ECLAT stands for Equivalence Class Clustering and bottom-up Lattice Traversal.
It is one of the most widely used mining strategies for Association Rules. It's a more efficient and scalable version of the Apriori approach. The Apriori approach works on

a horizontal level, replicating the behaviour of a graph's Breadth-First Search, whereas the ECLAT algorithm works in a vertical sense, simulating a graph's Depth-First Search. The ECLAT algorithm is faster than the Apriori algorithm because of its vertical approach.

The ECLAT model only comprises the support parameter , here we will be talking only about sets in a specific manner . There will be a set of two or more tasks which need to be analysed and fit into the model for smooth functioning . Here we need to set a minimum support and then take in account all the subsets in transactions having support less than minimum support and then we need to sort these according to decreasing order of support .

Working:-

The primary idea is to compute a candidate's support value using Transaction Id Sets(tidsets) intersections rather than generating subsets that are not found in the prefix tree. All single items, as well as their tidsets, are used in the function's first call. The function is then executed recursively, with each recursive call verifying and combining item-tidset pairings with additional item-tidset pairs. This method is repeated until there are no more candidate item-tidset pairings to combine.

Here we need to set a minimum support and then take in account all the subsets in transactions having support less than minimum support and then we need to sort these according to decreasing order of support .

We just need to organise the list into a Pandas Dataframe and we don't need to create tuples for the confidence and lift section .

Let's look at an example of the above-mentioned working:

Transact id	Bread	Jam	Milk	Butter	Chips
T1	1	0	1	1	1
T2	0	1	0	1	1
T3	0	1	1	0	1

T4	1	1	0	1	1
T5	0	1	1	0	1
T6	1	0	1	0	1
T7	1	1	1	0	1
T8	1	1	1	0	0

Boolean matrix illustration

The above data is a boolean matrix, with the data showing whether or not the j'th item is included in the i'th transaction . 1 denotes truth, while 0 denotes falsity.

In a tabular format, we call the function and arrange each item with its tidset:

-k = 1, minimum support = 2

Itemset	**Transaction ID Set**
Bread	**{T1,T4,T6,T7,T8}**
Jam	**{T2,T3,T4,T5,T7,T8}**
Milk	**{T1,T3,T5,T6,T7,T8}**
Butter	**{T1,T2,T4}**
Chips	**{T1,T2,T3,T4,T5,T6,T7}**

Tab – Iteration 1

We now recursively call the function till no more item-tidset pairs can be combined**:-**

k = 2

Itemset	**Transaction ID Set**

Bread,Jam	**{T4,T7,T8}**
Bread,Milk	**{T1,T6,T7,T8}**
Bread,Butter	**{T1,T4}**
Bread,Chips	**{T1,T4,T6,T7}**
Jam,Milk	**{T12,T14,T16,T17}**
Jam,Butter	**{T3,T5,T7,T8}**
Jam,Chips	**{T2,T3,T4,T5.T7}**
Milk,Chips	**{T1,T3,T5,T6,T7}**

Tab – Iteration 2

k = 3

Itemset	Transaction ID sets
{Bread,Jam,Milk}	**{T7,T8}**
{Bread,Jam,Chips}	**{T4,T7}**
{Bread,Butter,Chips}	**{T1,T4}**
{Jam,Butter,Chips}	**{T2,T4}**

Tab – Iteration 3

k = 4

Itemset	Transaction ID sets
{Bread,Jam,Milk,Chips}	**{T7}**

Tab – Last iteration - stop

Because there are no more item-tidset pairs to merge, stop at k = 4.

Because the supplied dataset has a minimum support of 2, the following rules can be derived:-

Bought Goods	**Products Recommended**
Bread	**Jam**
Bread	**Milk**
Bread	**Chips**
Jam	**Milk**
Jam	**Butter**
Jam	**Chips**
Bread and jam	**Milk**
Bread and jam	**Chips**

Tab - Results

WORKING OF APRIORI MODEL :

The following are the steps of the Apriori algorithm:-

1. Identify things present in the transactional database that have support and then pick the ones with the least amount of trust and support.
2. Identify all of the data points that have a support value greater than the currently specified or minimum support value.
3. Be sure to take note of any subset rules whose confidence value exceeds the threshold limit.

4. In descending order of lift, arrange the rules in the database.
5. As the elevator moves down the list, we'll have a clearer picture of the connections between the drugs.

Dataset:

In the first part, we created a dataset that simulates the transaction data of a pharmacy's customers. Rows indicate each transaction and contain the list of drugs purchased by the individual in question, totaling 7500.

This dataset comprises E. Coli promoter gene sequences (DNA) with incomplete domain theory from the UCI Molecular Biology (Promoter Gene Sequences) dataset.

The description of the attribute is as follows -

1. One of {+/-}, indicating the class ("+" = promoter).

2. The name of the instance which is essentially a 1000+ sequence of nucleotides prepared by T.Record.

3. 3-59. The remaining 57 fields are the sequence, (p-50) to (p7). Each of these fields is filled by one of {a, c,g,t}.

PROPOSED MODEL

Fig 3.4 - Workflow of the Apriori model

The goal of this research is to use machine learning to help with drug supply. Using the Apriori algorithm's support metrics, the goal is to create a recommendation system for the medicine that a specific customer is most likely to buy, resulting in a win-win situation for both the customer and the shop owner: the customer gets the most appropriate medicine they want at all times and does not have to deal with the problems of out-of-stock medicines; and the pharmacist learns the specific combination of

medicines that is made available quickly. A lack of drug supply implies the involvement of medical black market is drastically reduced. The complete workflow of the proposed model is given in Fig.4.4 .

Classification Models

1. **Naive Bayes -**

   Naive Bayes is a classification technique based on the assumption that predictors in a dataset are unrelated. This implies that the traits are unconnected to one another. The Bayes theorem underpins the naïve Bayes algorithm,that is as follows:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \qquad \text{.....(i)}$$

Where :

P(A | B) = Chance of A occurring if B occurs

P(A) = likelihood that A will occur

P(B) = likelihood that B will occur

P(B | A) = Chance of B occurring if A occurs

2. Decision Trees:

A visual depiction of decision-making is a decision tree. Creating an option tree begins with providing a yes/no question and then dividing the answer into two pieces. Nodes and leaves are both involved in the decision-making process. Internal nodes indicate properties of the dataset; offshoots indicate rules of decision; while each leaf node represents a categorization process result..

Leaf node and Decision node together make up a decision tree.Outcomes of decisions are represented via leaf nodes, while Decision nodes are the means by which decisions are made. Decision nodes contain several branches, whilst Leaf nodes have none.

The evaluations or tests are made depending on the data given. It's a visual representation of all the alternative workable solutions to complex problems or choices, based on particular parameters being satisfied or not met.

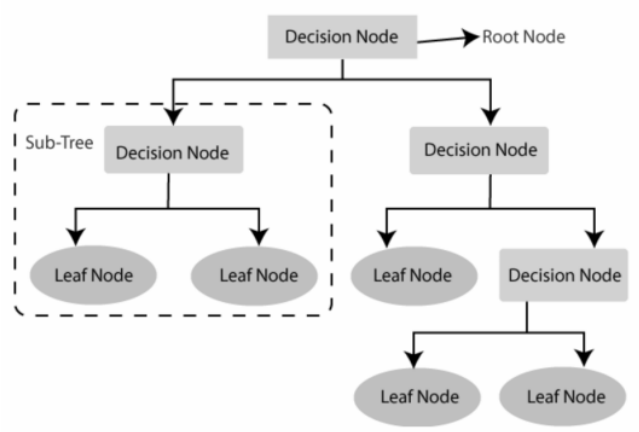


Fig 3.5 - Decision tree structure

3. Gaussian Classifier: (Radial basis function)

Gaussian Classifier is non-parametric machine learning method that might be utilized to develop newer non-parametric regression-classification algorithms.

For classification predictive modelling, Gaussian methods can be used as a machine learning method. Gaussian processes, such as SVMs, are a variation of kernel - based approach, but unlike SVMs, provide highly accurate predictions.

In Gaussian processes, a kernel specifies how samples relate to one another; it determines the covariance function of the data. This is also known as the latent function or "nuisance" function. The Gaussian Processes Classifier is a non-parametric technique for binary classification applications.

4. K-Nearest Neighbour:

K-Nearest Neighbor is a rudimentary Machine Learning method that utilizes the Supervised Learning approach. The K-NN approach assumes that the new case/data and old cases are comparable, and it places the new case in the category that is closest to the current categories. The K-NN algorithm keeps all of the current data and categorises incoming data points based on their similarities. This means that as new data is created, the K-NN approach can swiftly categorise it into a suitable category.

The K-NN approach may be used for both regression and classification, however it is more commonly utilised for classification tasks.K-NN is a non-parametric method, which means it makes no assumptions about the underlying data.

5. Random Forest:

One of the most well-known machine learning algorithms belonging to the supervised learning method is Random Forest. In machine learning, it can be used for classification and regression problems. In order to deal with a complex problem and improve the precision, ensemble training, which combines many classifiers, is used.

Random Forest is a classifier that comprises a number of decision trees on various subsets of the provided dataset and averages the results to enhance the predicted accuracy of that dataset Instead of depending on a single decision tree, the random forest collects the forecasts from each tree and predicts the final output based on the majority vote of predictions.

The bigger the number of trees in the forest, the higher the accuracy and the lower the risk of overfitting.

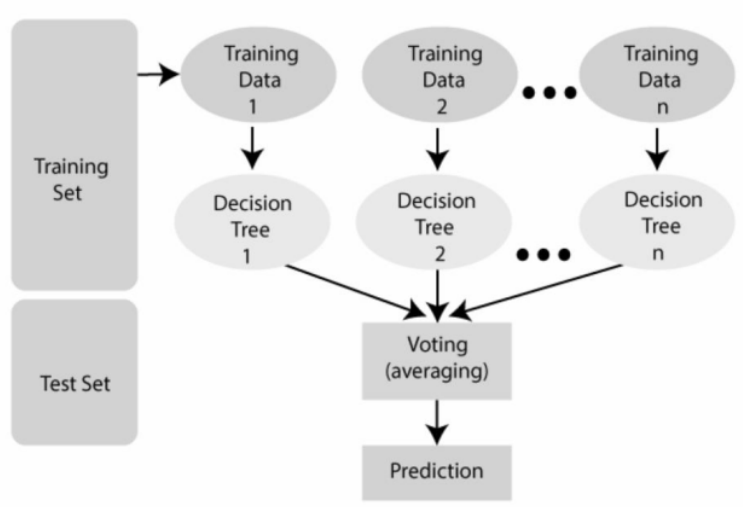


Fig 3.6 - Working of random forest

Evaluating Classification Algorithms

Some accuracy measurements are needed to assess the accuracy of any classifier model. To test how effectively the classifiers predict, the following methods are utilised:

a. Method of Withholding/Holdout - It is one of the most often used methods for measuring the accuracy of classifiers. This approach divides the data into two sets: a Training set and a Testing set. The training set is supplied to the model, and it learns from the data in it. The data in the testing set is hidden from the model, and it is used to assess the correctness of the model after it has been trained.

Both the features and label will be included in the training set, but the model will only need to predict the features in the testing set.

The model's accuracy is assessed by calculating how many labels it accurately predicted when the predicted labels are compared to the actual labels.

b. Using Variance-Bias - Bias is the difference between actual and expected numbers. The underlying assumptions about the data that the model makes in order to anticipate new data are known as the model's bias. It's a dead ringer for the patterns seen in the data. The model's assumptions are too simplistic when the Bias is big, and the model is unable to represent the major aspects of our data; this is referred to as underfitting.

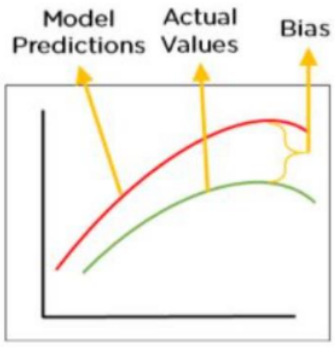


Fig 3.7 - Bias

The model's susceptibility to data disparities can be quantified as variance. It's highly likely that the model will be able to learn from the noise. As a result, it will place a high value on minor traits. When the variance is high, the model will collect all of the features of the data it is given, adjust itself to the data, and forecast successfully on it; but, new data may not have the same features, and the model will be unable to forecast accurately on it. This is referred to as the variance.

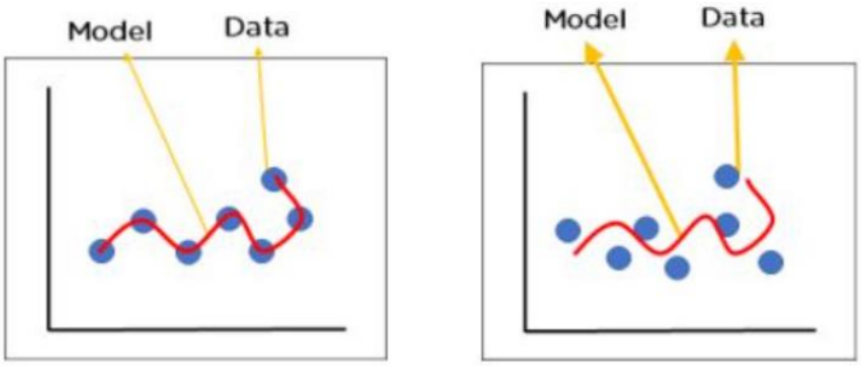


Fig 3.8 - Example of Variance

c. Precision - Recall - Precision is used to assess a model's ability to categorise values correctly. It's determined by dividing the total number of correctly identified data points by the number of correctly classified data points for each class label.

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}$$

When and where:

A true positive(TP) is one in which the model accurately predicts the positive outcome. A true negative(TN), is a result in which the model accurately predicts the negative one.

A false positive(FP) is a situation in which the model incorrectly predicts the positive class.. A false negative(FN) is when the model forecasts the negative class incorrectly.

The capacity of the mode to anticipate positive values is measured by recall. It is a measure of how often the model forecasts true positive values. The ratio of genuine positives to the total number of real positive values is used to compute this value.

d. F1 - Score - The weighted harmonic mean of accuracy and recall is the F1. The closer the F1 score number is to 1.0, the higher the model's projected performance.

$$\text{F1 Score} = \frac{2*(Recall*Precision)}{Recall+Precision}$$

A dataset's support consists of the number of instances of a class within it. It does not differ between models; it just diagnoses the process of performance evaluation.

Imbalanced support in the training data may suggest fundamental problems in the classifier's reported scores, indicating the necessity for stratified sampling or rebalancing. The support does not differ between models, but rather diagnoses the evaluation process.

A classification report is used to display the trained classification model's recall, accuracy, support and F1 score.True and false positives, as well as true and false negatives, are used to generate the metrics. In this scenario, positive and negative are general names for the projected classes.

The software and technology used for the project is Google Colab and python.

Python is an interpretable general-purpose programming language with a high level of abstraction. Its approach to design is based on the usage of important identifiers to focus on code readability. It is an object-oriented strategy that assists programmers in writing clean and logical code for any project.

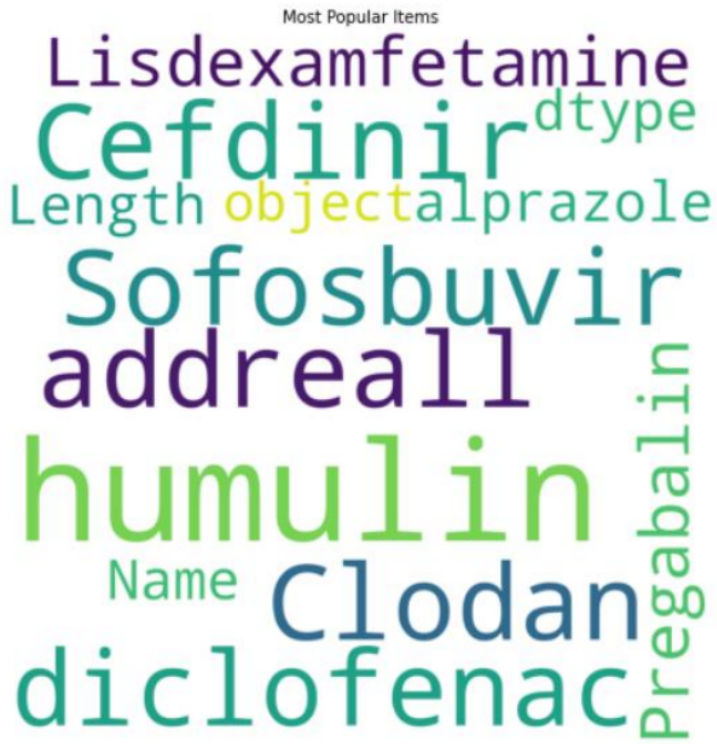
CHAPTER 4

Results and Discussion

Part 1 (A)

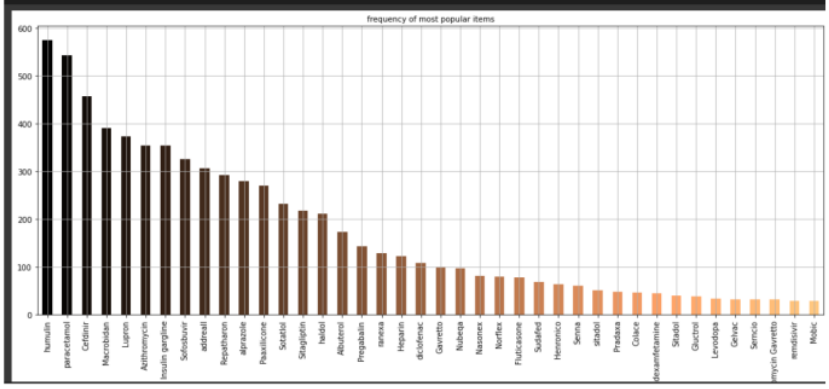
Apriori Recommendation system :

Word Cloud showing the most popular items



Frequency of the most popular items :

```
plt.rcParams['figure.figsize'] = (18, 7)
color = plt.cm.copper(np.linspace(0, 1, 40))
dataset[0].value_counts().head(40).plot.bar(color = color)
plt.title('frequency of most popular items', fontsize = 10)
plt.xticks(rotation = 90 )
plt.grid()
plt.show()
```



Displaying the results

Displaying the first results coming directly from the output of the apriori function

```
[ ] results = list(rules)
```

```
[ ] results
```

```
RelationRecord(items=frozenset({'Lisdexamfetamine', 'Azithromycin'}), support=0.01973070257299027, ordered_statistics=[OrderedStatistic(items_base=frozenset({'Lisc
RelationRecord(items=frozenset({'Lumbradorsiva', 'Azithromycin'}), support=0.0050659912011731 77, ordered_statistics=[OrderedStatistic(items_base=frozenset({'Lumbra
RelationRecord(items=frozenset({'Azithromycin', 'Lupron'}), support=0.0333288894814024 8, ordered_statistics=[OrderedStatistic(items_base=frozenset({'Lupron'}), ite
RelationRecord(items=frozenset({'Mobic', 'Azithromycin'}), support=0.0147980269297427 02, ordered_statistics=[OrderedStatistic(items_base=frozenset({'Mobic'}), item
RelationRecord(items=frozenset({'Azithromycin', 'Neomycin'}), support=0.0158645513931475 8, ordered_statistics=[OrderedStatistic(items_base=frozenset({'Neomycin'}),
RelationRecord(items=frozenset({'Pradaxa', 'Azithromycin'}), support=0.0178642847620317 3, ordered_statistics=[OrderedStatistic(items_base=frozenset({'Pradaxa'}), i
RelationRecord(items=frozenset({'Senna', 'Azithromycin'}), support=0.0083988801493134 25, ordered_statistics=[OrderedStatistic(items_base=frozenset({'Senna'}), item
RelationRecord(items=frozenset({'Serncio', 'Azithromycin'}), support=0.0206639114784695 4, ordered_statistics=[OrderedStatistic(items_base=frozenset({'Serncio'}), i
RelationRecord(items=frozenset({'Shringix', 'Azithromycin'}), support=0.0073323556859085 46, ordered_statistics=[OrderedStatistic(items_base=frozenset({'Shringix'}
RelationRecord(items=frozenset({'Sitagliptin', 'Azithromycin'}), support=0.0317291027862951 6, ordered_statistics=[OrderedStatistic(items_base=frozenset({'Sitaglipt
RelationRecord(items=frozenset({'Sofosbuvir', 'Azithromycin'}), support=0.0231969070790561 25, ordered_statistics=[OrderedStatistic(items_base=frozenset({'Sofosbuvi
RelationRecord(items=frozenset({'diclofenac', 'Azithromycin'}), support=0.0279962671643780 82, ordered_statistics=[OrderedStatistic(items_base=frozenset({'diclofena
RelationRecord(items=frozenset({'Azithromycin', 'gemazar'}), support=0.0034662045060658577, ordered_statistics=[OrderedStatistic(items_base=frozenset({'gemazar'}),
RelationRecord(items=frozenset({'glucagon', 'Azithromycin'}), support=0.0057325689908012 26, ordered_statistics=[OrderedStatistic(items_base=frozenset({'glucagon'}
RelationRecord(items=frozenset({'humulin', 'Azithromycin'}), support=0.0275963204906012 55, ordered_statistics=[OrderedStatistic(items_base=frozenset({'humulin'}),
RelationRecord(items=frozenset({'lydicasone', 'Azithromycin'}), support=0.0059992001066524 46, ordered_statistics=[OrderedStatistic(items_base=frozenset({'lydicasor
RelationRecord(items=frozenset({'ranexa', 'Azithromycin'}), support=0.0077323023596853 75, ordered_statistics=[OrderedStatistic(items_base=frozenset({'ranexa'}), it
RelationRecord(items=frozenset({'Azithromycin', 'remdisivir'}), support=0.0142647646980402 62, ordered_statistics=[OrderedStatistic(items_base=frozenset({'remdisivi
```

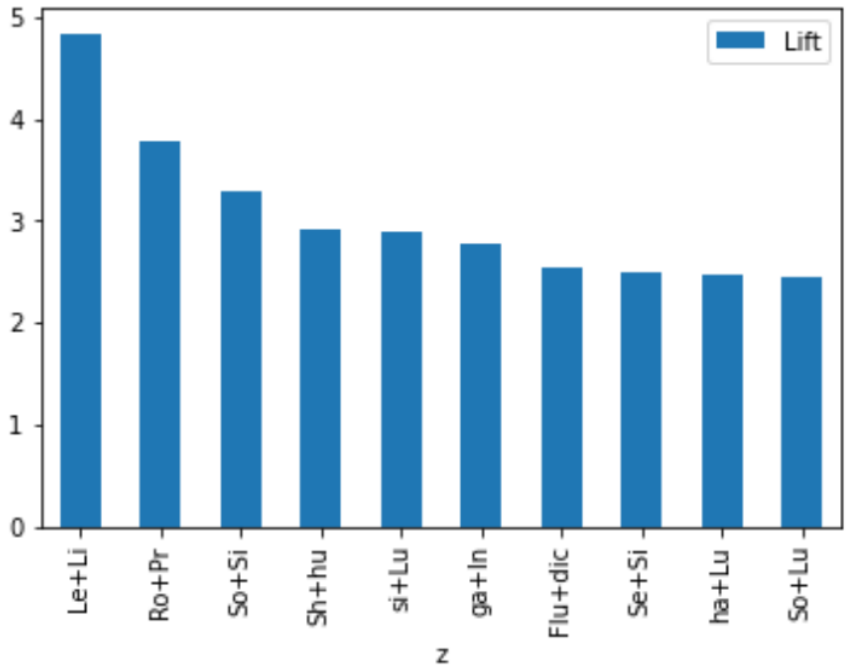
Displaying non - sorted results

```
resultsinDataFrame
```

	Left Hand Side	Right Hand Side	Support	Confidence	Lift
0	Albuterol	Azithromycin	0.021197	0.302857	1.522608
1	Alprimine	Insulin gargline	0.003200	0.369231	2.120674
2	Fluticasone	Azithromycin	0.019864	0.393140	1.976503
3	Hofine	Azithromycin	0.009865	0.383420	1.927635
4	Azithromycin	Insulin gargline	0.054793	0.275469	1.582155
...	...	...	...	...	...
116	glucagon	paracetamol	0.004933	0.345794	1.553176
117	ranexa	paracetamol	0.007732	0.347305	1.559963
118	remdisivir	paracetamol	0.014798	0.352381	1.582760
119	remdisivir Macrobidan	paracetamol	0.005333	0.449438	2.018704
120	remeron	paracetamol	0.007199	0.382979	1.720194

	Left Hand Side	Right Hand Side	Support	Confidence	Lift
68	Levothyroxine	Lisdexamfetamine	0.004533	0.290598	4.843951
99	Rosuvastatin	Pregabalin	0.005733	0.300699	3.790833
105	Sotatlol	Sitagliptin	0.015998	0.323450	3.291994
103	Shringix	humulin	0.005199	0.254902	2.923577
77	sitadol	Lupron	0.005466	0.275168	2.886760
57	gabapentin	Insulin gargline	0.003733	0.482759	2.772720
29	Fluticasone	diclofenac	0.011332	0.224274	2.545056
101	Senna	Sitagliptin	0.006532	0.246231	2.506079
75	haldol	Lupron	0.016131	0.235867	2.474464
74	Sofosbuvir	Lupron	0.016664	0.233209	2.446574

	Left Hand Side	Right Hand Side	Support	Confidence	Lift	var
68	Levothyroxine	Lisdexamfetamine	0.004533	0.290598	4.843951	Le+Li
99	Rosuvastatin	Pregabalin	0.005733	0.300699	3.790833	Ro+Pr
105	Sotatlol	Sitagliptin	0.015998	0.323450	3.291994	So+Si
103	Shringix	humulin	0.005199	0.254902	2.923577	Sh+hu
77	sitadol	Lupron	0.005466	0.275168	2.886760	si+Lu
57	gabapentin	Insulin gargline	0.003733	0.482759	2.772720	ga+In
29	Fluticasone	diclofenac	0.011332	0.224274	2.545056	Flu+dic
101	Senna	Sitagliptin	0.006532	0.246231	2.506079	Se+Si
75	haldol	Lupron	0.016131	0.235867	2.474464	ha+Lu
74	Sofosbuvir	Lupron	0.016664	0.233209	2.446574	So+Lu



1(B)

ECLAT Recommendation system :

ECLAT model result

```
resultsinDataFrame.nlargest(n = 10,
                            columns = 'Support')
```

	Left Hand Side	Right Hand Side	Support
4	Azithromycin	Insulin gargline	0.054793
53	Insulin gargline	Sitagliptin	0.039195
106	Sitagliptin	paracetamol	0.038128
9	Lupron	Azithromycin	0.033329
76	Lupron	paracetamol	0.033196
92	Norflex	paracetamol	0.032662
16	Sitagliptin	Azithromycin	0.031729
111	humulin	alprazole	0.028796
56	diclofenac	Insulin gargline	0.028396
18	diclofenac	Azithromycin	0.027996

Results sorted by descending support

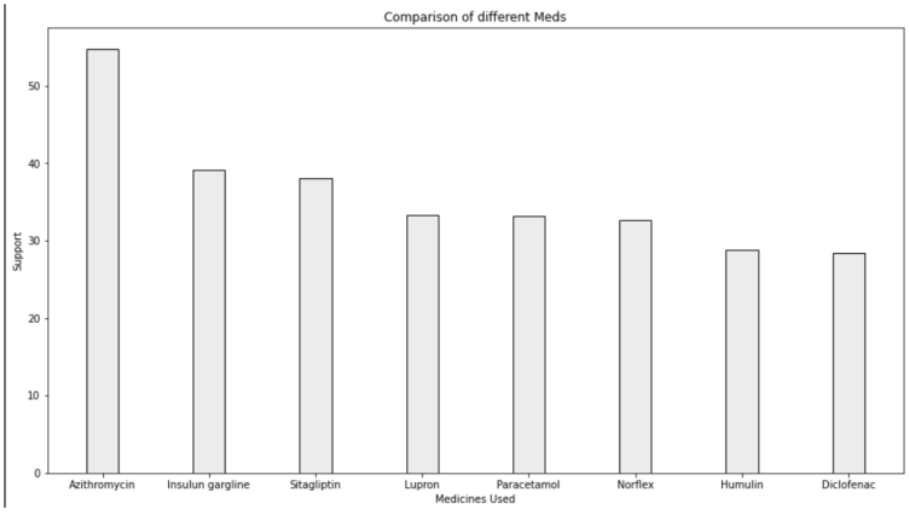


Fig 4.1 - ECLAT model result visuals

There are two types of association rule mining algorithms now available: those with horizontal format and those with vertical format. For eg., there is a matrix displaying transactions with objects. This matrix can be displayed horizontally or vertically.

The horizontal data arrangement is the most widely utilized. Each transaction, in other words, contains a transaction identifier (TID) and an item - list that occurs in that transaction, i.e., TID:itemset. The vertical data structure, in which the database is made up of a series of objects, each of which is followed by a series of transaction IDs with the item, i.e., {item:TID set} is another often used layout. The horizontal configuration is shown in table 1; the vertical layout is shown in table 2:

TRANSACTION ID	ITEMSETS
T1	Milk , Biscuit, Sauce
T2	Milk Sauce
T3	Slice

Horizontal Format Illustration

ITEMSET	TRANSACTION
Milk	T1 , T2
Biscuit	T1
Sauce	T1,T2
Slice	T3

Vertical Format Illustration

Apriori algorithm works with horizontal format whereas Eclat can be applied for those with vertical data sets. The Eclat algorithm is a data mining algorithm for locating frequently occurring things[31]. In association rule mining, certain algorithms generate frequent itemsets in a horizontal format, while others use vertical formats.

The ECLAT algorithm is quicker than the Apriori algorithm because of its vertical approach . The database is to be searched repeatedly in the Apriori technique to locate common itemsets; this limitation is alleviated in Eclat by employing a vertical dataset. Only one scan of the database is required by Eclat.

While the Apriori algorithm imitates a graph's Breadth-First Search in a horizontal sense, the ECLAT method imitates a graph's Depth-First Search in a vertical sense, which is generally faster than Breadth-First search.

Eclat represents transactions entirely vertically. There is no requirement of subset tests or subset generation for finding out support.

All single items of data, as well as their respective tidsets, are used in the first invocation of the function. The function is then called recursive, and each item in the tidsets pair is validated and combined with the other items in the tidsets pair in each recursive call. This operation is performed until there are no more candidate items in tidsets pairs to merge.

The main advantage of the vertical format is that it allows for quick frequency counting using intersection operations on transaction ids (tids) as well as automated data reduction. The basic problem with these techniques is that the algorithm's scalability decreases as the intermediate results of vertical tid lists get too large for memory.

When the data set size is small or medium, the Eclat method is inherently quicker than the Apriori approach. When the data set size is big, Apriori may outperform Eclat because intermediate Tidsets formed by the Eclat method take more memory than Apriori. When we have a huge dataset, the intermediate results of vertical tid lists grow too enormous for memory, limiting the scalability of the method. As a consequence, the Eclat technique outperforms the Apriori approach for small and medium datasets, whereas the Apriori approach outperforms the Eclat method for big datasets.

2

Classification algorithms :

The classification report of each model is included below,which prints the mean and standard deviation of the values.

The accuracy of the classification algorithm can be viewed just below,as a percentage.

```
Nearest Neighbors: 0.828571 (0.130785)
Nearest Neighbors
78.125
              precision    recall  f1-score   support

           0       1.00      0.65      0.79        20
           1       0.63      1.00      0.77        12

    accuracy                           0.78        32
   macro avg       0.82      0.82      0.78        32
weighted avg       0.86      0.78      0.78        32
```

```
Gaussian Process: 0.762500 (0.167143)
Gaussian Process
87.5
              precision    recall  f1-score   support

           0       1.00      0.80      0.89        20
           1       0.75      1.00      0.86        12

    accuracy                           0.88        32
   macro avg       0.88      0.90      0.87        32
weighted avg       0.91      0.88      0.88        32


Decision Tree: 0.678571 (0.178750)
Decision Tree
81.25
              precision    recall  f1-score   support

           0       0.94      0.75      0.83        20
           1       0.69      0.92      0.79        12

    accuracy                           0.81        32
   macro avg       0.81      0.83      0.81        32
weighted avg       0.84      0.81      0.82        32


Random Forest: 0.680357 (0.162382)
Random Forest
71.875
              precision    recall  f1-score   support

           0       1.00      0.55      0.71        20
           1       0.57      1.00      0.73        12

    accuracy                           0.72        32
   macro avg       0.79      0.78      0.72        32
weighted avg       0.84      0.72      0.72        32
```

```
Naive Bayes: 0.830357 (0.142690)
Naive Bayes
90.625
              precision    recall  f1-score   support

           0       1.00      0.85      0.92        20
           1       0.80      1.00      0.89        12

    accuracy                           0.91        32
   macro avg       0.90      0.93      0.90        32
weighted avg       0.93      0.91      0.91        32
```

Classification reports of different algorithms

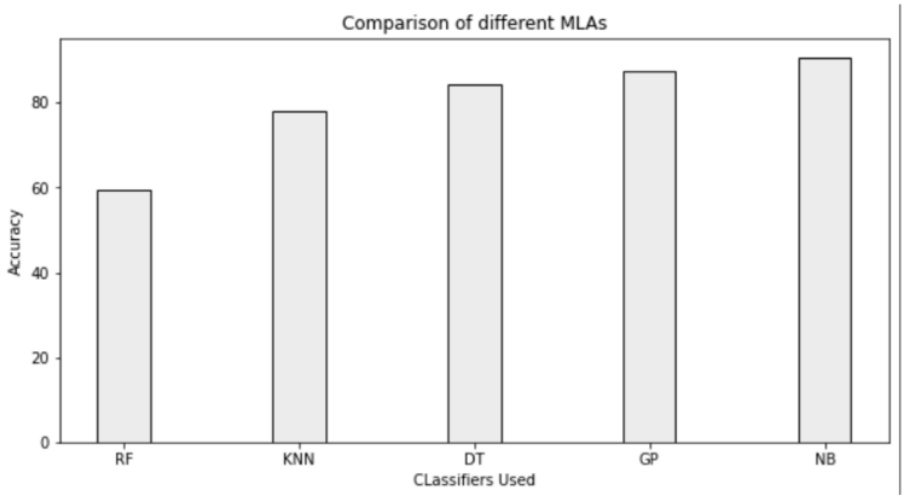


Fig 4.2 – Comparative plot of various classifiers

The best model for this particular classification is the Naive-Bayesian algorithm with an accuracy of 90.625%.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 Summary:-

We have mainly focused on two objectives. The first one was a medicine recommendation system that will be helpful for the healthcare sector. People won't have to face the problem of unavailable medicines, since the stores will be stocked well in advance since they can know which medicines are most likely to be bought. Moreover, the economy will be helped since the medical black market will be eliminated as medicines are readily available so there is no shortage, thus no scope of dishonest people to dupe others by profiteering from selling medicines at exorbitant rates to needy people. Secondly,our focus is on the study and comparison of various classifiers[14]. We have taken 5 classifiers and processed the same dataset containing DNA sequences through each of them,so that we can understand which classifier works best in our case.
However, there are a few shortcomings to the recommendation too. Basic knowledge of the operation is to be learned by the caregivers. Furthermore, the drugs that have been anticipated may not always be available, and individuals may require additional medications. People might be allergic to the medicines that are being recommended together,thus requiring other types of medicines.

The classifier system is fairly accurate,but not foolproof[30]. It might not work well with variations in data or there might be mislabeled data,leading to difficulties in maintaining compliance. In worst cases,there might be natural disasters or other difficulties that disrupt the entire structure,causing extensive damage and going back to the start all over again.

5.2 Future Scope:-

The main motive behind this model is to ensure that the common people get the best possible variant of the medicines available in the market at all points of time.

This model will just recommend the best associated combination of medicines that go along with each other in a certain manner based on the previous sales of those medicines. Thus, it allows the optimal transaction to happen between the patient and the chemist shop.

On the flip side, it also allows the chemist to update his/her stocks to its full potential at any given point in time so that the patient can get the best possible variants of the medicines, whenever he/she has a necessity of it .

In future, the proposed Apriori/ECLAT based machine learning recommendation model can be enhanced to allow low infrastructural casualties in a healthcare center as it will always ensure that the best possible medicine or other health equipment are available at all times of the year . This will boost the lack of technical and managerial policies that are lacking today in different healthcare centres across India .

Because of the effects of globalisation and high mobilisation of the global population, new viruses are expected to originate and spread as quickly as the present COVID-19. Identifying infections sooner will aid in the prevention of outbreaks such as COVID-19 and will aid in medication development. As a result, DNA sequence categorization is critical in computational biology. Our classification model tries to explore different classifiers and ascertain which classifier works best in our case. Thereafter in future the tests can be performed on a larger dataset and studied in detail.

This model can be further enhanced by focussing on the UI/UX aspect which will allow a patient and his/her family to get a clear visual understanding of the current status of the different healthcare facilities that are available at a healthcare center in some developed areas without even travelling long distances in search of a preferable diagnostic centre for the patient .

This method will save many lives and, as a result, lead to a better policymaking mindset for the general public.

5.3 Constraints

However, there are a few shortcomings to the project .

Basic knowledge of the operation is to be learned by the caregivers. Furthermore, the drugs that have been anticipated may not always be available, and individuals may require additional medications. People might be allergic to the medicines that are being recommended together,thus requiring other types of medicines.

The classifier system is fairly accurate, but not fool proof. It might not work well with variations in data or there might be mislabelled data, leading to difficulties in maintaining compliance. In worst cases, there might be natural disasters or other difficulties that disrupt the entire structure, causing extensive damage and going back to the start all over again.

5.4 Social Impact

Medicine as a discipline cannot function in isolation from society. The sort of medication that is in use is determined by society. Members of society, either directly or via their representatives, determine what resources are required for healthcare professional training and delivery across all medical specialties. Furthermore, society will establish and prescribe deviance and how deviance will be dealt with, particularly in psychiatry. The social compact that existed between monarchs and their subjects is now being replicated between physicians and society as a whole, but through their representatives, who will also decide how the professions are controlled. Physicians must answer to regulatory agencies about clinical practise and healthcare delivery standards.

Medicine recommendation systems help to mitigate problems of shortage,drug trafficking and various other parasitic problems related to pharmaceutical industry which might threaten to rot away the foundation of healthcare.

CHAPTER 6

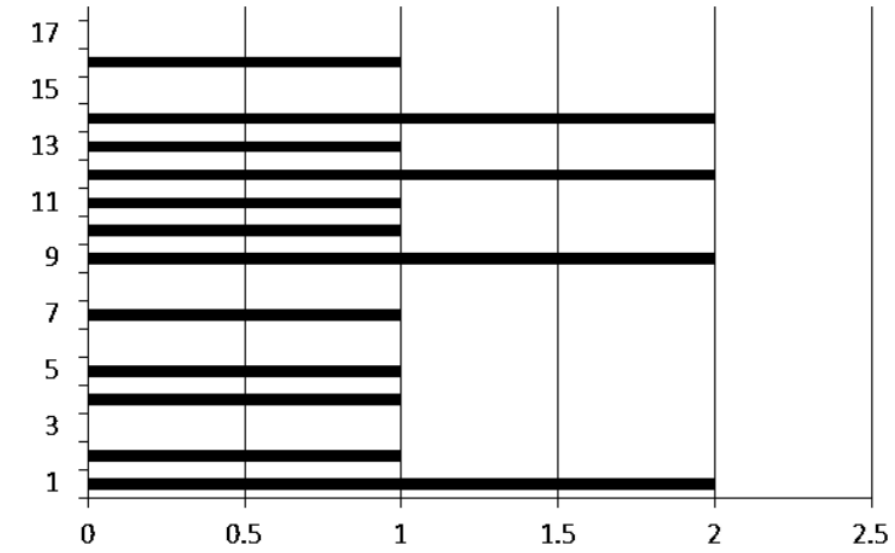
PLANNING & REFERENCES

6.1 Planning and project management

S.No.	Activity	Starting Week	Number of Weeks
1.	Literature Review	1st-2nd week of Dec	2
2.	Required software setup, coding,calibration	3rd week of Dec	1
3.	Code Integration & Debugging	1st week of Jan	1
4.	Inclusion of ECLAT model	2nd week of Jan	1
5.	Medicine overview along with a basic understanding of python and data manipulation and preprocessing	3rd week of Jan	1
6.	Preparing the model(train)	4th week of Jan	2
7.	Fitting model	2nd week of Feb	1
8.	Checking model	3rd week of Feb	1

9.	Analysis after training the model	4th week of Feb	2
10.	Metrics evaluation	1st week of Mar	1
11.	Preparation of project report	2nd week of Mar	2
12.	Preparation of Project presentation	4th week of Mar	1

The Gantt Chart is shown below:-



6.2 REFERENCES –

[1]. J. Bouwens, “Embracing Change: The healthcare industry focuses on new growth drivers and leadership requirements.”

[2]. "Intro to Machine Learning | Udacity." Intro to Machine Learning | Udacity. Accessed April 27, 2016.

[3]. Thanh Nguyen, Abbas Khosravi, Douglas Creighton, Saeid Nahavandi, Classification of healthcare data using genetic fuzzy logic system and wavelets, Expert Systems with Applications, Volume 42, Issue 4, 2015, Pages 2184-2197

[4]. Schölkopf, Bernhard, Christopher J. C. Burges, and Alexander J. Smola,Advances in Kernel Methods: Support Vector Learning. Cambridge, MA: MIT Press, 1999.

[5]. Witten, I. H., and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques. Amsterdam: Morgan Kaufman, 2005

[6]. Zhiyong Ma, Juncheng Yang, Taixia Zhang, Fan Liu. (2016). An Improved Eclat Algorithm for Mining Association Rules Based on Increased Search Strategy. International Journal of Database Theory and Application, 9(5), 251-266

[7]. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, et al. Data mining in healthcare and biomedicine: a survey of the literature. J Med Syst 2012; 36:2431–48.

[8]. Yadav M, Malhotra P, Vig L, Sriram K, Shroff G. Ode-augmented training improves anomaly detection in sensor data from machines. arXiv preprint arXiv:1605.01534,2016:1–5.

[9]. Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction: With 200 Full-color Illustrations. New York: Springer, 2001

[10]. O.Stephen et.al.,”An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare”,Journal of Healthcare Engineering ,Volume 2019

[11]. P. S. Mung and S. Phyu. Effective analytics on healthcare big data using ensemble learning. In 2020 IEEE Conference on Computer Applications(ICCA), pages 1–4, 2020.

[12]. Samer Ellahham**,** Artificial intelligence in the diagnosis and management of COVID-19: a narrative review, *Journal of Medical Artificial Intelligence*,2021

[13]. Dave DeCaprio, Joseph Gartner, Carol J. McCall, Thadeus Burgess, Kristian Garcia, Sarthak Kothari, Shaayaan Sayed, Building a COVID-19 vulnerability index, *Journal of Medical Artificial Intelligence*,2020

[14]. Pahulpreet Singh Kohli and Shriya Arora. Application of machine learning in disease prediction. In 2018 4th International Conference on Computing Communication and Automation (ICCCA), pages 1–4. IEEE, 2018.

[15]. Munira Ferdous, Jui Debnath and Narayan Ranjan Chakraborty, Machine Learning Algorithms in Healthcare: A Literature Survey, In, that on 2020 11th International Conference on Computing, Communication, and Networking Technologies (ICCCNT)

[16]. Shweta Ganiger and KMM Rajashekharaiah. Chronic diseases diagnosis using machine learning. In 2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET), pages 1–6. IEEE, 2018.

[17]. Dharavath Ramesh, Pranshu Suraj, and Lokendra Saini. Big data analytics in healthcare: A survey approach. In 2016 International Conference on Microelectronics, Computing and Communications (MicroCom), pages 1–6. IEEE, 2016

[18]. Kumar SP, Samson VRR, Sai UB, Rao PLSDM, Eswar KK. Smart health monitoring system of patient through IoT. In: 2017 International conference on I-SMAC (IoT in social, mobile, analytics and cloud) (I-SMAC). IEEE; 2017. p. 551–6.

[19]. Trivedi S, Cheeran AN. Android based health parameter monitoring. In: *2017 International conference on intelligent computing and control systems (ICICCS)*. IEEE; 2017. p. 1145–9

[20]. Acharya AD, Patil SN. IoT based health care monitoring kit. In: 2020 Fourth international conference on computing methodologies and communication (ICCMC). IEEE; 2020. p. 363–8.

[21]. Geron, “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow”, O’Reilly Media, Inc., Canada, 2019

[22]. William W, Basaza-Ejiri AH, Obungoloch J, Ware A. A review of applications of image analysis and machine learning techniques in automated diagnosis and classification of cervical cancer from pap-smear images.

[23].T. I. Mohammad, A. A. Md, T. M. Ahmed, and A. Khalid, “Abnormality detection and localization in chest x-rays using deep convolutional neural networks,” 2017, http://arxiv.org/abs/1705.09850.

[24].P. Huang, S. Park, R. Yan et al., “Added value of computer-aided CT image features for early lung cancer diagnosis with small pulmonary nodules: a matched case-control study,” *Radiology*, vol. 286, no. 1, pp. 286–295, 2017.

[25].G. Varun, P. Lily, C. Marc et al., “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2017

[26].J. Melendez, G. B. Van, P. Maduskar et al., “A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest x-ray,” *IEEE Transactions on Medical Imaging*, vol. 34, no. 1, pp. 179–192, 2015.

[27].U. Avni, H. Greenspan, E. Konen, M. Sharon, and J. Gold-berger, “X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words,” *Med Imaging, IEEE Transactions*, vol. 30, no. 3, 2011.

[28].H. Boussaid and I. Kokkinos, “Fast and exact: ADMM-based discriminative shape segmentation with loopy part models,” in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA, June 2014

[29]. X. Z. Yang, C. P. En and Z. Y. Fang, “Improvement of Eclat algorithm for association rules based on hash Boolean matrix”, Application Research of Computers, vol. 27, no. 4, (2010), pp. 1323-1325.

[30]. M. J. Zaki and K. Gouda, “Fast vertical mining using diffsets”, Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, USA, (2003), pp. 326- 335.

[31]. F. P. En, L. Yu, Q. Q. Ying and L. L. Xing, “Strategies of efficiency improvement for Eclat algorithm”, Journal of Zhejiang University (Engineering Science), vol. 47, no. 2, (2013), pp. 223-230.

Project Summary

Project Title	A Comparative Analysis of Machine Learning Algorithms
Team Members (Names)	Souvik Karmakar ,Sudeshna Dutta ,Indrashis Mitra , Kinjal Sarkar & Pratyay Basu
Faculty Guide	**Prof. K.B. Ray**
Semester / Year	VIII / IV year
Project Abstract	Indian pharmaceutical companies are the world's leading provider of generic medicines. Over 50% of the world's vaccine need is supplied by the Indian pharmaceutical sector, which accounts for 40% of pharmaceuticals needed in the United States and 25% of all medicines in the United Kingdom. Pharmaceutical production in India ranks third globally in terms of volume and fifteenth globally in terms of value. The domestic pharmaceutical industry consists of over 3,000 medical firms and 10,500 production facilities. India is a major player in the global pharmaceutical sector. There is a large hub of budding engineers and scientists in the nation that can take the business to the next level. Antiretroviral drugs manufactured in India currently account for more than 80% of the world's AIDS medicine supply.However,people in India are taking desperate steps to keep loved ones alive as a disastrous spike of new coronavirus infections overwhelms the country's health-care infrastructure. They are turning to dubious medical therapies in some circumstances, and to the underground market for life-saving pharmaceuticals in others.Hence our project aims to solve this by developing a medicine recommendation system so that pharmacists can know in advance what medicines are the medicines being bought together the most,and keep stocks accordingly. We do not recommend which medicine is to be taken,rather the focus is on finding the drug combinations bought frequently together so that an idea can be had of the most-selling medicines;thus keeping their stock replenished would eliminate the black market,help to earn profits and help the customers.
List **codes** and **standards** that significantly affect your project.	• IEEE 7000.7001 are two standards for evaluating the algorithms • IEEE P7002 standard was followed while analysing dataset of medicine combinations
Briefly explain two **significant trade-offs** considered in your design, including options considered and the solution chosen	• This recommendation is a random one,not based on actual medical conditions • It is not tested for real-time DNA datasets from labs

PUBLICATIONS

Paper presented in International Conference on Robotics, Control and Computer Vision (ICRCCV 2022)

To be published by Springer as proceedings in Lecture Notes in Electrical Engineering series.

SELF DECLARATION FOR PLAGIARISM CHECK

We, Souvik Karmakar(1807228), Sudeshna Dutta(1807232),Indrashis Mitra(1807274), Kinjal Sarkar(1807277) and Pratyay Basu(1807291) are declaring that our Project report on "A COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS" has plagiarism well within the limits prescribed to us. We take full responsibility for it.

A COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS

ORIGINALITY REPORT

17%	10%	7%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	2%
2	Submitted to University of North Texas Student Paper	2%
3	www.simplilearn.com Internet Source	1%
4	Submitted to University of Bradford Student Paper	1%
5	Submitted to Higher Education Commission Pakistan Student Paper	1%
6	Submitted to National College of Ireland Student Paper	1%
7	content.iospress.com:443 Internet Source	1%
8	Submitted to Middlesex University Student Paper	<1%
	Submitted to Sheffield Hallam University	

9	Student Paper	<1%
10	Submitted to University of Hertfordshire Student Paper	<1%
11	www.geeksforgeeks.org Internet Source	<1%
12	Submitted to University of Essex Student Paper	<1%
13	levelup.gitconnected.com Internet Source	<1%
14	Submitted to The University of Memphis Student Paper	<1%
15	dl.acm.org Internet Source	<1%
16	gams.com Internet Source	<1%
17	Submitted to October University for Modern Sciences and Arts (MSA) Student Paper	<1%
18	Mahdi Abdollahi, Xiaoying Gao, Yi Mei, Shameek Ghosh, Jinyan Li, Michael Narag. "Substituting clinical features using synthetic medical phrases: Medical text data augmentation techniques", Artificial Intelligence in Medicine, 2021 Publication	<1%

19	Pei, Zhi. "Simplification of fuzzy multiple attribute decision making in production line evaluation", Knowledge-Based Systems, 2013. Publication	<1 %
20	Submitted to University of East London Student Paper	<1 %
21	Submitted to University of Dundee Student Paper	<1 %
22	Munira Ferdous, Jui Debnath, Narayan Ranjan Chakraborty. "Machine Learning Algorithms in Healthcare: A Literature Survey", 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020 Publication	<1 %
23	Submitted to Swinburne University of Technology Student Paper	<1 %
24	Submitted to Bangladesh University of Professionals Student Paper	<1 %
25	www.ukessays.com Internet Source	<1 %
26	Submitted to Polytechnics Mauritius Student Paper	<1 %
	Submitted to International College	

27	Student Paper	<1%
28	www.ijraset.com Internet Source	<1%
29	Munetoshi Akazawa, Kazunori Hashimoto. "Artificial intelligence in gynecologic cancers: Current status and future challenges – A systematic review", Artificial Intelligence in Medicine, 2021 Publication	<1%
30	Submitted to Universiti Teknologi MARA Student Paper	<1%
31	smartech.gatech.edu Internet Source	<1%
32	Avni, Uri, Hayit Greenspan, Eli Konen, Michal Sharon, Jacob Goldberger, and Bram van Ginneken. "", Medical Imaging 2011 Computer-Aided Diagnosis, 2011. Publication	<1%
33	Zhiyi Liu, Rui Chang. "Study on efficient algorithm of frequent item-set mining", Proceedings of 2011 International Conference on Electronics and Optoelectronics, 2011 Publication	<1%
34	hppcb.nic.in Internet Source	<1%

35	Submitted to Chester College of Higher Education Student Paper	<1%
36	Mahdi Abdollahi, Xiaoying Gao, Yi Mei, Shameek Ghosh, Jinyan Li. "A Dictionary-based Oversampling Approach to Clinical Document Classification on Small and Imbalanced Dataset", 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2020 Publication	<1%
37	pdfcoffee.com Internet Source	<1%
38	Submitted to University of Wollongong Student Paper	<1%
39	docs.neu.edu.tr Internet Source	<1%
40	downloads.hindawi.com Internet Source	<1%
41	es.scribd.com Internet Source	<1%
42	phdprojects.org Internet Source	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		