DEVELOPMENT OF A HYBRID CLUSTERING ALGORITHM FOR EFFICIENT MEDICAL RESOURCES ALLOCATION

BY

AKHIGBEMIDU, OZEMOYA REX (B.SC.)

Reg. No: 20124832278

A THESIS SUBMITTED TO

THE SCHOOL OF POSTGRADUATE STUDIES
FEDERAL UNIVERSITY OF TECHNOLOGY, OWERRI, IMO STATE

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE (M.Sc.) DEGREE IN COMPUTER SCIENCE

AUGUST, 2024

CERTIFICATION

I certify that this work "**Development of A Hybrid Clustering Algorithm for Efficient Medical Resources Allocation**" was carried out by **AKHIGBEMIDU OZEMOYA REX** (Reg. No. **20124832278**) in partial fulfilment of the requirements for the award of Master of Science (M.Sc) degree in Computer Science in the Department of Computer Science, Federal University of Technology Owerri.

30/09/2024
....
Dr. I. I. Ayogu
(Supervisor)

30/09/2024
Date

....
Dr. (Mrs.) C. G. Onukwugha
(Co Supervisor)

30/09/2024
Date

....
Dr. (Mrs.) J. N. Odii
(Head of Department)

30/09/2024
Date

....
Prof. (Mrs.) U. F. Eze
(Dean, SICT)

30/09/24
Date

....
Prof. (Mrs.) J. N. Nwosu
(Dean, School of Postgraduate Studies)

....
Date

....
Prof. Moses Okechukwu Onyesolu
(External Supervisor)

....
Date

DEDICATION

This work is dedicated to Almighty God for His love lavished on me.

ACKNOWLEDGEMENTS

I acknowledge my Supervisors, Dr. I. I. Ayogu and Dr. (Mrs.) C. G. Onukwugha for their contributions, criticisms and guidance that led to the successful completion of this work. May God richly bless them.

I appreciate Dr. (Mrs.) J. N. Odii who is the Head of Department, Computer Science, as well as other lecturers of the Department of Computer Science, Federal University of Technology Owerri for their willingness to impart knowledge without reservation. On this same rope, I sincerely appreciate the contributions and supports of the Dean, SICT, Prof. Mrs. U. F. Eze. Her leadership in SICT has evidently improved academic activities in SICT. I am equally very grateful to Dr. Adetokunbo John-Otunu for his supports and contributions.

Special thanks to my lovely wife, Mrs. Ezinne B. Akhigbemidu and my children: Emmanuel and Daniel for their prayers and support throughout the period of the programme. I pray for favour of God upon them.

My thanks also go to my parents and in laws: Mr. and Mrs. Akhigbemidu and Dr. and Mrs. Ewulonu, respectively. I pray God to bless them. I also appreciate my brothers and my sisters for their prayers and support during the period of this programme. I pray for God's reward upon them.

I acknowledge the contributions, encouragement and co-operation from my professional colleagues and friends especially Chinedu Temple, Ihuoma and among others. May God bless them. My appreciations to many who offered support in one way or another, whose names are not mentioned here. May the blessing of God be with all of them.

Above all, my profound gratitude goes to almighty God for provisions, journey mercies, protection and favours granted me during the course of this programme. To Him alone be all the glory.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

LIST OF APPENDICES

ABSTRACT

Efficient medical resource allocation is a critical challenge in healthcare systems, particularly with increasing demand and limited resources for managing in-patient and out-patient treatment datasets. This project is motivated by the need to address this challenge, as clustering algorithms offer a promising approach for grouping healthcare data, enabling more effective distribution of medical resources. This project aimed to develop a hybrid clustering algorithm that combines the strengths of density-based and partitioning methods to optimize medical resource allocation. The project used a combination of K-representative and K-means clustering algorithms. Adopting the Object-Oriented Analysis and Design (OOAD) methodology, the proposed algorithm analyzes medical datasets to produce more effective clusters, revealing insights that enhance resource distribution. The hybrid algorithm, implemented using the JAVA object-oriented programming tool, generated better-defined clusters of in-patients and out-patients, providing actionable knowledge and intelligence for optimizing medical resource allocation. The results demonstrate the algorithm's potential to improve decision-making in healthcare systems by enhancing the efficiency of resource allocation. The findings further suggest that this hybrid algorithm can serve as a robust tool for healthcare providers, contributing to more efficient resource management and better patient outcomes.

Keywords: Hybrid, Clustering, Algorithm, Medical Resources, Allocation.

CHAPTER ONE

INTRODUCTION

1.1 Background Information

Efficient medical resource allocation is fundamental to the effective functioning of healthcare systems, particularly in times of crisis and resource scarcity (Gou *et al.,* 2024). It involves the strategic distribution of limited resources, including medical personnel, equipment, and medications, to ensure optimal patient care and maximize health outcomes (Gong & Kang, 2023). The complexity of this task is heightened by the need to balance multiple factors such as urgency of care, patient demographics, and geographical constraints. Addressing these challenges requires innovative approaches to improve allocation processes, reduce disparities, and enhance overall healthcare delivery (Barbosa *et al.,* 2020).

Clustering algorithms, as a category of unsupervised learning techniques, are employed to partition objects within a dataset into clusters based on shared characteristics (Sambantham, 2023). The primary objective is to increase the similarity among objects within the same cluster while decreasing the similarity between objects in different clusters. This is achieved by maximizing intra-cluster similarity and minimizing inter-cluster similarity. The fundamental concept underlying clustering algorithms is to uncover inherent structures and patterns within the dataset without relying on pre-defined category labels. These algorithms determine whether objects should be grouped into the same cluster by evaluating their similarity or distance from each other (Yin *et al.,* 2024).

Clustering algorithms find extensive applications across diverse domains. These include data mining, as demonstrated by density-based algorithms for discovering clusters in noisy large spatial databases, image processing as illustrated in efficient clustering methods for colour segmentation of spatial images, and bioinformatics for tasks such as protein clustering for structural prediction (Sambantham, 2023). They are also used in social network analysis, encompassing the identification of critical nodes through graph entropy and market analysis, which includes customer segmentation using clustering techniques and social network analysis.

Clustering evaluation involves the manner of figuring out datasets which might be very different to apprehend the records variations in addition, the similarities within the data. The aim of clustering and cluster evaluation is to organize and distinguish similar units and to split them from differing units (Kristina & Miin-shen, 2020). K-means clustering algorithm is a well-known partitioning clustering algorithm that is used to analyze very large numerical datasets (Haesik, 2020). It was extended to K-mode to enable it to analyze categorical datasets. Also, the weakness of the k-mode algorithm was enhanced to develop a k-representative clustering algorithm. Moreover, Xiaoliang, Yulin, Yi, Honglian, Muhammad, and Joshua (2020) combined the K-means and K-mode techniques to create the K-prototype clustering algorithm. This innovative approach addresses the complexities of working with mixed datasets in diverse fields such as Agriculture diseases (Pinisetty, Valaboju & Rao, 2015), Stock exchange, laboratory analysis, customers sales analysis, and social media activities., etc (Oleji, Nwokorie, and Chukwudebe, 2020; Haesik, 2020).

These algorithms have its limitation concerning accuracy for a decision support system; this cluster give up, assessment includes a huge type of statistical techniques. In cluster assessment, one tries to organize large numbers of individuals, process or objects into smaller numbers of collectively wonderful instruction in wherein within the individuals have similar characteristics (Haesik, 2020).

This thesis focuses on the development of a hybrid clustering algorithm designed to optimize the allocation of medical resources, thereby improving efficiency and ensuring equitable access to healthcare services, particularly the tertiary health care institution.

1.2 Problem Statement

Resource management and allocation in tertiary health institutions have been a challenging issue due to the large volume of treatment records, drug inventory, and health cases of both inpatients and outpatients acquired over the years in hospitals. Medical directors and health personnel often allocate resources based on assumptions that lack reliable evidence; such assumed predictions can lead to incorrect resource allocation and wastage of limited resources. Clustering algorithm performance is limited due to the nature of the datasets such as mixed, numerical, and categorical datasets when creating clusters (Xiaoliang *et al.,* 2020; Prasad & Raju, 2020).

This research seeks to address the problem of predicting inaccurate mining methods with wrong estimation of results for future planning. Mostly, it is very harmful in the hospital management system because patient's life will be at risk. Better clustering results for the analysis of medical resources allocation were achieved in this work by developing a hybrid clustering algorithm with a mechanism that outperformance the existing clustering algorithms.

1.3 Objectives

The primary objective of this work is to develop a hybrid clustering algorithm for efficient medical resources allocation. The specific objectives are:

1. To develop a hybrid clustering algorithm using K-representative and K-means clustering algorithms.

2. To preprocess the historical datasets of inpatient and outpatient in tertiary health institution to train the hybrid model.
3. To cluster the preprocessed healthcare dataset for the medical resource's allocation guide.
4. To test and evaluate the performance of the developed model.

1.4 Justification of the Study

The justifications of the study are as follows:

1. Considering the large complex datasets of inpatient and outpatients at tertiary health institutions this project has the potential to implement a hybrid clustering algorithm that can optimize the historical healthcare records at tertiary health institutions. Its outcome may guide medical personnels on the proper management of the limited healthcare resources.
2. Through efficient preprocessing, this project is justified by its potential to implement a framework that sufficiently manages challenging factors such as handling outliers in datasets, as well as size of data during clustering.
3. The implementation of this work has the potential to reduce wastage of the scarce medical resources in tertiary health institutions.

1.5 Scope of the Study

The proposed hybrid clustering algorithm consists of K representative and K means clustering algorithms to cluster patient records. This work will use clustering procedures to aid an understanding of how clustering methods will be used to develop patient resource utilization classification schemes and decisions are required to use such method in a typical tertiary health institution.

CHAPTER TWO

LITERATURE REVIEW

2.1 Conceptual Framework

The conceptual framework in this report discussed the concept of clustering, clustering methods and data mining preprocessing techniques.

2.1.1 Concept of Clustering

Cluster creation is a form of unsupervised learning technique. An unsupervised learning approach involves techniques in which references are drawn from datasets, including input data without classified responses. Generally, it is widely used as a method to discover significant structural pattern**s**, explanatory underlying tactics, generative traits and groupings inherent in constant of examples (Srya, 2021). Clustering is a system studying method that involves the grouping of information points. Given a hard and fast of information points, clustering algorithm is used to categorize records factor into particular groups. In theory, facts factors which are within side the identical institution need to have comparable residence and functions at the same time as facts factors in specific corporations need to have tremendously diverse residences and functions (Mohanavalli, & Jaisakthi, 2015). Clustering is a technique of unsupervised learning and is a not unusual place method for statistical data analysis used in many fields (George, 2018). Clustering are challenges of dividing the populace or data points into several groups such that fact factors with inside the equal corporation are extra just like different records factor with inside the identical instruction and diverse to the records factors in different corporations. It is basically a hard and fast of items on the right of some similarity and dissimilarity amongst them (Srya, 2021).

Furthermore, according to Aravind (2021), clustering (cluster evaluation) involves the process of categorizing gadgets based on their similarities. Clustering perhaps utilized in a number of areas machine learning, computer graphics, sample recognition, snap shots evaluation and statistics retrieval, bioinformatics, and facts compression. Clusters are an elaborate concept this is why there are lots of specific clustering algorithms. Distinctive cluster models are employed and for each of those cluster models unique algorithms may be given (Ramakrishna & Jayaraj, 2014). Clusters determined via way of means of one clustering algorithm will in reality be unique from clusters determined with the aid of using unique algorithm (Aravind, 2021). In machine learning systems, step one in the direction of information the dataset is to organization the examples. Grouping an unlabeled instance is known as clustering. As the samples are unlabeled, clustering prediction on unsupervised system gaining knowledge of (Opara, Eze & Oleji, 2020). If examples are classified, they become classified in the cluster (Aravind, 2021).

2.1.2 Why Clustering?

Clustering is essential in various fields, particularly in healthcare, where efficient resource allocation is critical. In the context of medical resource allocation, clustering helps group similar data points, such as patients with similar needs, medical facilities, or resource demands. This enables healthcare providers to identify patterns, optimize resource distribution, and ensure that medical services reach the right population at the right time. Clustering also allows for the segmentation of patients based on factors like disease type, severity, or geographic location, which helps tailor resource allocation strategies and reduce inefficiencies. Additionally, as the volume of healthcare data grows, traditional methods struggle to manage complexity. Clustering provides a data-driven approach to handle this, improving decision-making and enhancing operational efficiency. With advancements in hybrid clustering algorithms, which combine different clustering techniques, there is potential to achieve more accurate and scalable solutions, ultimately leading to better resource management in healthcare systems. (Srya, 2021).

2.1.3 Clustering Methods

Clustering techniques are as follows**:**

i. **Density-Based Methods:** This method groups data points based on regions of high density, separating clusters by areas of low point density. It is effective for identifying arbitrarily shaped clusters and can efficiently handle noise and outliers in large datasets, making it suitable for applications like spatial data analysis (Srya, 2021).

ii. **Hierarchical Based Methods:** The clusters fashioned in these strategies form a tree kind formed based completely at the hierarchy. New clusters are formed using miles previously formed one. It's divided into two categories.

    a. **Agglomerative** (bottom-up method)

    b. **Divisive** (top-down method)

iii. **Partitioning Methods:** Those strategies partition objects into k clusters and each partition form one cluster (Wei, Chow & Chan, 2015). Partitioning methods, like K-Means, divide the dataset into a predefined number of clusters by iteratively assigning data points to the nearest cluster centroid. It is computationally efficient for large datasets but requires prior knowledge of the number of clusters and may struggle with complex, non-linear data distributions (Srya, 2021).

iv. **Grid-based Methods:** This method divides the data space into a finite number of grid cells and performs clustering based on the number of points within each cell. It is highly efficient for large datasets and works independently of data distribution, as seen in algorithms like STING, which focus on multi-resolution clustering.

2.1.4 Data Preprocessing in Data Mining

Data preprocessing is a truth mining technique that is used to transform raw statistics to useful format (Deepak & Abhishek, 2021). The diagram for data preprocessing is shown in Figure 2.1.

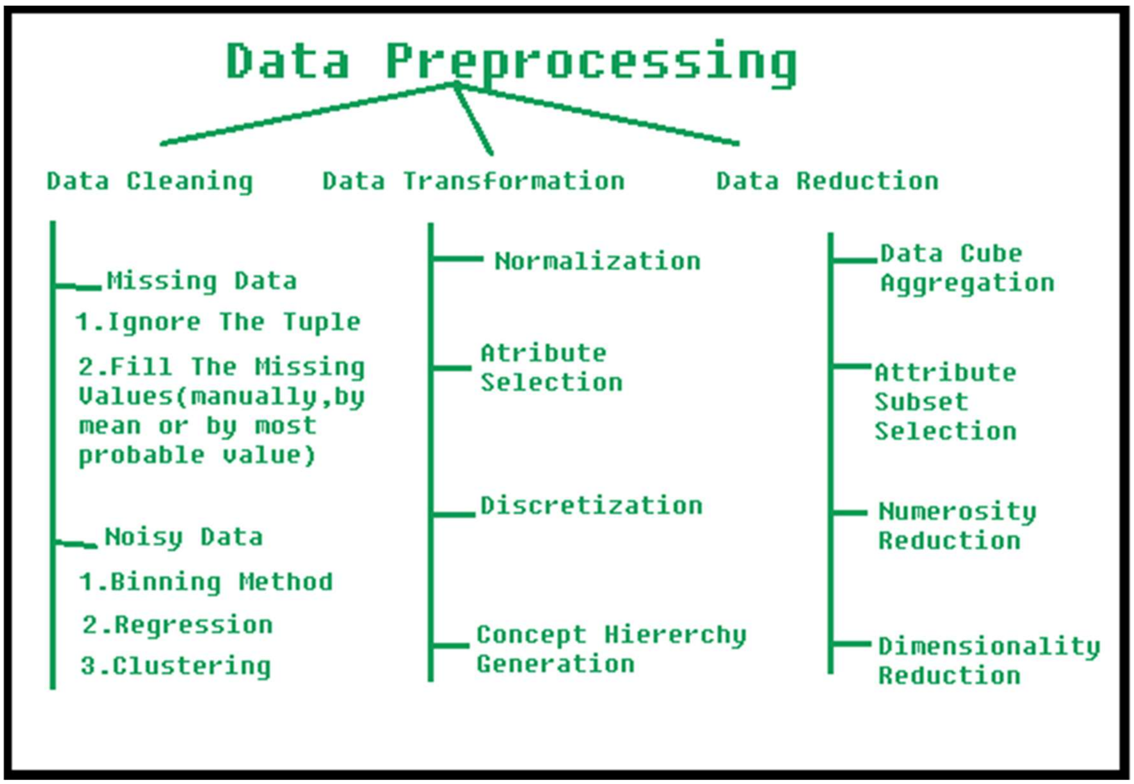


Figure 2.1: Data Preprocessing (Source: Deepak & Abhishek, 2021).

2.1.4.1 Steps Involved in Data Preprocessing:

1. **Data Cleaning:** The record should have many beside the factor and missing component. To cope with this component, recording cleaning is done. It consists of managing missing statistics data, noisy records etc.

i. Missing Data: This statistic arises at the same time as some statistics is missing with inner data. It can be handled innumerous ways.

ii. Ignore the tuples: This method is suitable most effective at the same time as the dataset we've is quite big and a couple of values are missing inner tuple.

iii. Fill the Missing values: There are numerous methods to do that task. You can pick out to fill the lacking values manually with aid of using characteristics suggest or the maximum possible value.

iv. Noisy Data: Noisy information is incomprehensible facts that cannot be interpreted via way of means of machines. It could be generated because of detective records collection, record access errors and so forth. It could be treated a lot of ways:

a. Binning Method: This method works on sorted that facts will let you easy it. The complete records are out up into segments of same duration and then several strategies are carried out to finish the task. Each segmented is sorted separately. One nonetheless replaces all records in a segment through manner of approach of its recommended or boundary values to complete the task.

b. Regression: Right here records can be made clean smooth through turning into it to a regression function. The regression used may be linear (having one impartial variable) or a pair of (having a couples of variables).

c. Clustering: This technique organizes the same information in a cluster. The outliers can be undetected or it's going to fall out of the clusters.

2. **Data Transformation:** This step is taken so that it will redesign the information in appropriate forms suitable for mining process. This consists of the subsequent ways:

i. Normalization: It is far some distance an awesome manner to scale the data values in a specific range (-1.0 to 1.0 or 0.0 to 1.0)

ii. Attribute Selection: On this strategy, new attributes are crafted from a given set of attributes to help the mining process.

Discretization: This is done to replace the uncooked values of numeric attribute with resource C language levels or conceptual levels.

iii. Concept Hierarchy Generation: Here attributes are converted from lower to higher in hierarchy. For instance, the attribute "city" may be transformed to "country".

3. **Data Reduction:** The goal here is, to optimize the overall performance and reduce data storage and assessment costs since the essence of mining is to seek knowledge from massive amount of data (Deepak & Abhishek, 2021).

The several steps to data bargain are:

i. **Data Cube Aggregation:** Aggregation operation is applied to data for the improvement of the information cube.

ii. **Attribute Subset Selection:** The quite relevant attributes need to be used rest all can be discarded. For performing traits selection, you will use diploma of significance and p-value of the attribute. The attribute having p-value greater than significance diploma can be discarded.

iii. **Numerosity Reduction:** This lets in preparing the version of data instead of whole statistics as an instance Regression Models.

iv. **Dimensionality Reduction:** This reduces the dimensions of data thru manner of method of encoding mechanisms. It could be lessee or lossless. If after reconstruction from compressed data, particular data can be retrieved, such reduction is referred to as lossless reduction else it’s far some distance too as loss reduction. The two effective techniques of dimensionality reduction are: Wavelet transforms and PCA (Principal Component Analysis) (Deepak & Abhishek, 2021).

2.2 Theoretical Review

This thesis observed the best of partitioning clustering algorithm for the assessment and development of the proposed hybrid clustering algorithms.

2.2.1 Partitioning clustering algorithm

In centroid/partitioning clustering, clusters are represented through an essential vector, which might not continuously some of the dataset. Even on this specific clustering type the value of K wants to be chosen. That is an optimization troubleshooting locating the quantity of cancroids or the value of K and assigning the items to nearly cluster centers. The one step wishes to be completed on this taken care of manner that the square distance from clusters is maximized. One of the maximum significantly used centroid primarily based absolutely clustering algorithms is K-Means and actually taken considered one in all drawbacks is which you want to select a K value in advance (Aravind, 2021).

2.2.2 K-means Clustering Algorithms

K-means is a partitioning-based clustering algorithm that aims to divide a dataset into **K distinct clusters**, where each cluster has a centroid representing its center. The algorithm starts by selecting **K initial centroids**, then assigns each data point to the nearest centroid, forming clusters. It iteratively recalculates the centroids by taking the average of all points within each cluster and reassigns points to the closest new centroids. This process continues until the centroids no longer change significantly, aiming to minimize the within-cluster variance and produce compact, well-separated clusters (Surya, 2021).

K-means clustering is a very widely recognized and powerful unsupervised machine learning algorithm (Aditya, 2020).

A K-means clustering algorithm attempts to compare items in the form of clusters. The range of groups is represented through K. Let's take an illustration in Figure 2.1 and 2.2.

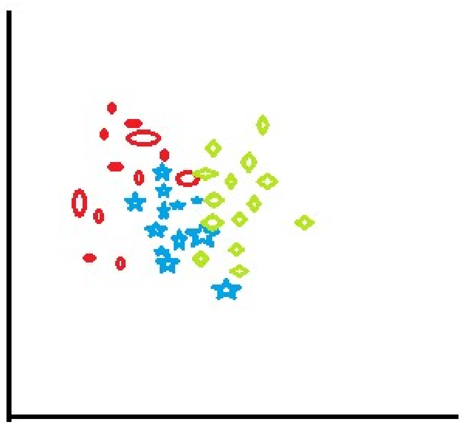


Figure 2.2: Before applying k-means clustering

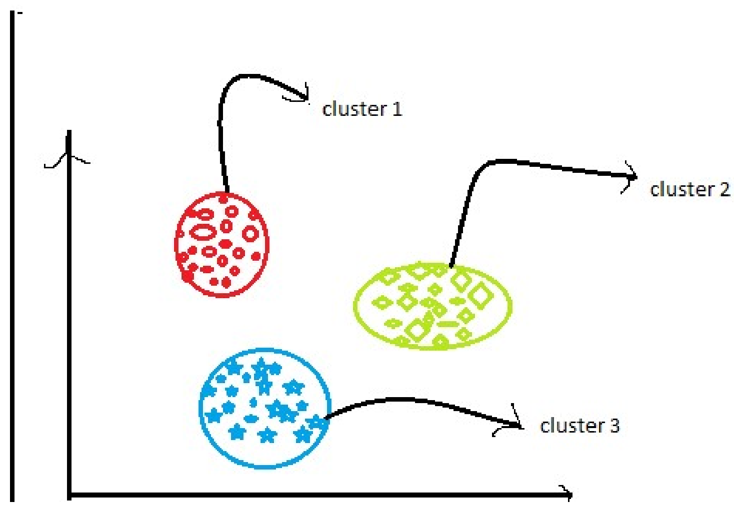


Figure 2.3: After applying K-means clustering

(Source: Aditya, 2020).

Figure 2.2 shows the data in advance before using the k-means clustering algorithm. It can be observed clearly that all of the unique classes are messed up. Figure 2.3 shows the data after using the K-means clustering algorithm. Notice that three unique objects are found and classified into three unique classes which are probably called clusters (Aditya, 2020). K-Means has gain that it is quite rapid as all it calls for is computing the distances between points and group centers facilities only few computations. It for that reason has a linear complexity O (n) (Morissette & Chartier, 2013). At the opportunities hand, K-Means has a couple of disadvantages. Groups and classes are first picked and counted. This is not always trivial and flawless with a clustering algorithm we did want it to parent the only out for us because of reality the issue of the factor of it is miles to benefit some belief from the data. K-means moreover starts off evolved advance with a random choose of cluster centers and consequently it could yield special clustering effect on unique runs of the algorithm (Nanopoulos & Yannis, 2001; Overlade, Oladipupo & Obagbuwa, 2010). Consequently, the results may now no longer also not be repeatable and lack consistency (George, 2018).

2.2.3 Steps in K-Means

Step 1: Firstly, choose a number of classes/groups to apply and randomly initialize their respective middle points.

Step 2: Every data point is assessed through computing the distance between that point and each group center, and organization of middle after which classifying the point to be in the group whose middle is closest to it.

Step 3: Primarily based totally on these classified points, we recomputed the group centers through taking the suggest of all the vectors in the group.

Step 4: Repeat those steps for a fixed variety of iterations or till the group centers don't change much between iterations (George, 2018; Aditya, 2020).

Every time clusters are made centroids are up to date; the up to date centroid is the middle of all points which within side the cluster. This method keeps until continues till the centroid now not alternative i.e. answer converges.

2.2.4 Advantages of K-means

Aditya (2020) stated the advantages K-means such as:

i. It is simple to implement.
ii. It is scalable to a large data set and quicker to big datasets.
iii. It adapts to brand-new examples very frequently.
iv. Generalization of clusters for one of a kind of shapes and sizes (Aditya, 2020).

2.2.5 Disadvantages of K-means

The disadvantages of K-means are as follows:

i. It is far touchy to the outliers (Zhang& Fang, 2013).
ii. Choosing the k values manually is a difficult job.
iii. Because the quantity of dimensions will increase its scalability decreases (Aditya, 2020).

2.3 Empirical Literature

Jae and Sang (2021) presented k-representatives algorithm: A clustering algorithm with getting to know distance degree for categorical values. Their method changed distinction metric that measures the gap among all values of every characteristic statistically and clusters objects in dataset primarily based totally in this metric. The algorithm changed into proved the use of actual international dataset of Mushroom and Nursery from the University of California Ivory repository. “The consequences display that example of various classes are separated via way of means of clusters a very success. Moreover, the range of clusters is adjusted thru getting rid needless clusters for the duration running of the algorithm. Their future work stated that the algorithm may be integrated with different distance degree in numerical area such as Euclidean distance with the right scaling techniques among our distance degree and numerical distance degrees.

Also, Toan and Van-Nam (2021) proposed a Locality-Sensitive Hashing (LSH) based representatives clustering technique for large categorical data. The mechanism of the device carries the LSH approach into the k-means-like clustering for you to make it able to predict the higher preliminary clusters for reinforcing clustering effectiveness (Toan & Van-Nam, 2021). To this case, the mechanism first applied a data pushed dissimilarity degree for categorical data to assemble a circle of relatives of binary hash capabilities that used to generate the preliminary clusters. They used a nearest neighbor seek at every iteration for cluster reassignment of data items to enhance the clustering complexity. These answers are included into the k-representatives algorithm ensuring with inside the so referred LSH-k-representatives algorithm. The consequence display that the newly advanced algorithm yields similar or higher clustering effects in assessment to the prevailing intently associated works.

Mohanavalli and Jaisakthi (2015) proposed a chi-square primarily based totally statistical methods to decide the burden of the attributes. This weight vector issued to derive the space matrix of the mixed dataset. The gap matrix is used to cluster the data factors use of the conventional clustering algorithms. Experiments were done usage of the UCI benchmark datasets, heart, credit, and vote. Other than those data sets we've get additionally examined our proposed approach use of an actual time financial institution data set. The accuracy of their clustering effects acquired became higher than the one of the present works. They could not increase their proposed machine to enhance the computation of characteristic applicable for use because the corresponding weights in distance computation in blended data clustering. It nonetheless surfer for the preliminary choice of k values. This may cause untimely convergence.

Jain and Verma (2014) defined an approximate algorithm primarily based totally on k-means. It's a far unique approach for huge data evaluation which could be very fast, scalable and has excessive accuracy. It overcomes the disadvantage of k-means of an unsure quantity of iterations through solving the quantity of iterations, without dropping the precision. The efficacy and precision of algorithm is proven on numerous actual and artificial datasets. The algorithm supplied here can't deal with categorical data properly till its miles transformed into equal numerical data. Exploring clustering huge data in phrases of categorical data will be any other viable extension. Finding out primary and secondary attributes is taken into consideration in ideal to be supplied as an enter through the user (which shows the view factor of study). Their proposed device inherited the hassle of preliminary choice of k values. This can cause untimely convergence. Machine learning concepts may be used to determine the concern of attributes instead of asking from the user.

Manjinder*, et al.,* (2014) proposed adaptive K-means clustering approach for data clustering to adjust k-means clustering. It adapts in line with the image primarily based totally on color primarily based totally clustering. The no. of clusters the use of the color capabilities are computed primarily based totally on histogram evaluation in gray format. The height of the histogram is the primarily supply of computation of range of colors with inside the image and primarily based on totally same image data are clustered.

The research of Jae and Sang, (2021) modified the value of distinction metric that measures the space among all values of every characteristic of the categorical datasets statistically and clusters objects in dataset to enhance K-representative clustering algorithm. The effects display that the instances of different classes are separated with the aid of using clusters very successfully. Jae and Sang (2021) counseled that the progressed K-mode algorithm (K-representatives algorithm) must be included with different distance measure in numerical domain name including Euclidean distance with the deal scaling technique among our distance measure and numerical distance measures.

Moreover, the study of Toan and Van-Nam (2021) utilized a data-driven dissimilarity measure for categorical data to assemble a family of binary hash features that turned into used to generate the preliminary clusters. The mechanism of usage of the closest neighbor seeks at every iteration for cluster reassignment of data objects advance the clustering complexity for developing clusters. They referred to as their proposed algorithm LSH-k-representatives algorithm. It yields similar or higher clustering outcomes in evaluation to the present intently associated works. Consequently, the observations from those present algorithms guided this work to breaded a hybrid clustering algorithm use of K-means and K-representatives clustering algorithms for clustering mixed dataset.

Kanjanawattana *et al.* (2020) proposed a new algorithm and applied same to mixed types of dataset, which consisted of numeric and categorical types. He opined that most data in the actual world is always mixed of two types, the numeric and categorical one. The experimental results show work better than K-means even though their input is a mixed dataset. And their experimental results are better in terms of speed and performance than K-means.

The proposed combination of supervised and unsupervised learning of clustering data without any preliminary assumption on the cluster shape. The experimental results show that the clusters are created correctly when the attributes are petal length versus petal width good accuracy (Kolhe *et al.,* 2020).

Zhang *et al.* (2019) proposed idea of the K-means clustering algorithm analysis. They offered two methods of improving the K-means clustering algorithm based on improving the initial focal point and secondly, determining the value of K. There simulation experiments proved that the improved clustering algorithm was not only stable in the clustering process but reduced or even avoided the impact of the noise data in the data set object. Their experimental results showed an improvement of K-means clustering algorithms are still not solved completely; rather, it requires further attempt and exploration.

Dewangan *et al.* (2020) proposed a Fuzzy C-means clustering algorithm, which allows one piece of data to belong to two or more clusters. Their new fuzzy k-modes algorithm is effective and better than the other existing k-modes algorithms, their algorithm could not combine the two implementations of fuzzy c-means and fuzzy k-modes for mixture of data items set. Moreover, they could not find out the cluster area and center of cluster. The paper suggested that the Fuzzy C-mean (FCM) algorithm for numeric data should be implemented and the Fuzzy K-modes algorithm for categorical data be also implemented separately and then combine the both implementations for mixture of data items that is numeric as well as categorical data items to produce the final results.

Yiu-ming *et al.* (2018) proposed a unified metric for categorical and numerical attributes in data clustering, in which the attributes are in numerical, categorical or mixed. Their experimental results show the efficacy of the proposed approach, there is no user-assigned parameter in the proposed algorithm.

Ahirwar *et al.,* (2020) proposed efficient algorithm that uses techniques of divide and conquer to cluster large datasets. He applied the Squared Euclidean Distance measure in the measuring the similarity between data points. The approach is very efficient in identifying the points and assigning the data points to the best cluster.

CHAPTER THREE

METHODOLOGY

3.1 Methodology

This project adopted Object Oriented Analysis and Design methodology to develop and design hybrid clustering algorithm for analysis of medical resources allocation datasets. The steps of the methodology are as follows: System Analysis, System Design, Object Design and Implementation. The Steps for Object Oriented Analysis and Design methodology are shown in Figure 3.1.

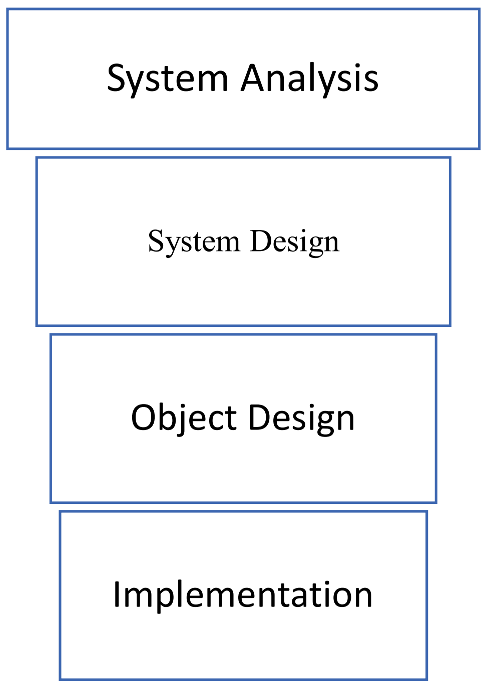


Figure 3.1: Steps for Object Oriented Analysis and Design Methodology (Mukherjee, 2016).

3.2 System Analysis

Clustering mixed dataset is a challenging task. K-means are used to cluster numerical dataset while k-mode is used to cluster categorical dataset. The integration of K-means and K-mode produced K-prototype that is used to analyze mixed dataset. K-means is one of the most effective unsupervised learning algorithms, addressing a well-known clustering problem. The method is straightforward, classifying a given data set into a fixed number of clusters (let's assume k clusters). The primary idea is to define k centroids, one for each cluster. These centroids must be carefully placed, as their initial positions can significantly impact the results. Ideally, they should be positioned as far apart as possible. However, K-means faces challenges when dealing with categorical data and is sensitive to outliers in the dataset. This is where the hybrid K-prototype clustering algorithm comes into play. Numerous extensions of k-modes algorithms were evolved to decorate the clustering overall performance for categorical data. However, the k-modes algorithm faces a critical hassle choosing modes while two or extra categorical values at a positive function have the equal maximum frequency, which probably makes it volatile as distinct mode decide on might produce unique result. It is far really well worth noting right here that via way of mean definition, the mode of a cluster is not always precise in trendy and the clustering end result strongly relies upon on the choice of modes at some point of the clustering process. Curiously, this problem of the k-modes algorithm may be triumph over through k-representatives algorithm. With inside (Toan and Van-Ham, 2021), the k-representatives algorithm changed into offered to address the non-precise mode disadvantage of the k-modes algorithm the aid of using introducing the idea of representatives rather than of modes for clusters.

3.3 Architecture of the Existing System

The architecture of the existing system is shown in Figure 3.2.

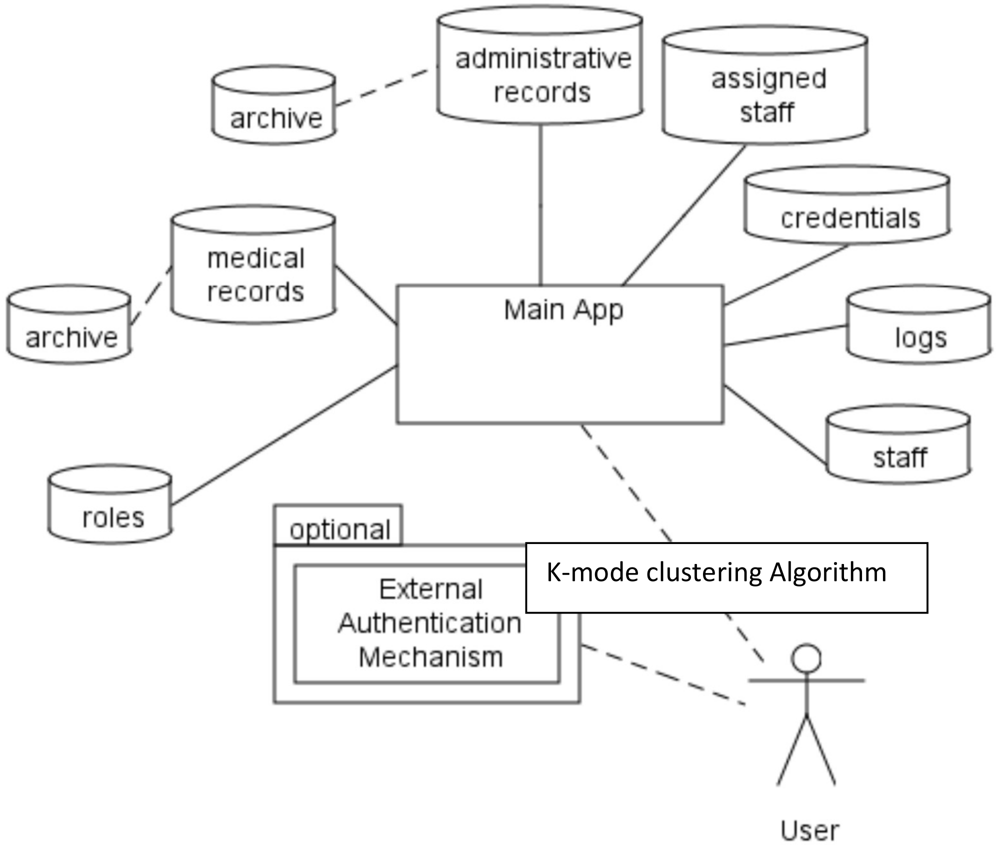


Figure 3.2: Architecture of the Existing System (Slawomir *et al.*, 2011).

The box in the middle (MainApp) symbolizes the self-containing system which will process the patient record data. It is surrounded by a number of cylinders representing the databases, these are:

1. **Administrative record** - This includes all the administrative information about the patients, like their name, next of kin, social security ID etc.

1 **Assigned staff** - Every patient is assigned a doctor and a nurse taking care of him. These staff members have got special privileges for accessing the record of the patient. This database contains information about assignment of the medical staff to the patients.

2 **Credentials** - The users have to log in to the system before they can perform any action on it. In this table the < user□name; has hedpassword > pairs are stored. The authentication of the users is done against this database.

3 **Logs** - the system is expected to log the security-related events. The logs go into this database.

4 **Medical records** - All the health-related information about the patients are stored, like what treatments they are undergoing, what medicine they are taking, what allergies they have got, etc.

5 **Roles** - The staff members of a hospital using the Hospital Information System assume various roles, like doctor, nurse or an administration clerk. This database embraces the information what roles the staff members assume.

6 **Staff** - The information about the hospital staff is stored.

Also, below the MainApp box there is another box symbolizing an optional authentication mechanism - this external mechanism could be used for authentication of the users if the password scheme is not sufficient. However, authorization of the action performed by users is expected to done by the MainApp. Summarizing the conceptual design, there is one self-containing application co-operating with a number of databases and possibly an external authentication mechanism. Now we will go to the functional design of the system, presented in the form of use cases (Slawomir *et al.,* 2011).

7. **K-Modes:** The K-Modes algorithm by extending the K-Means algorithm with a simple mismatching dissimilarity measure for categorical data, and a frequency-based method to update modes during clustering.

3.3.1 Advantages of the Existing System

The following are the advantages of the k-modes algorithm. They include:

i. It initializes the modes of k clusters.
ii. It computes the dissimilarity between the objects and the modes of the cluster and places the object in the cluster which results in minimum dissimilarity before updating the mode of the cluster.
iii. The modifications in the traditional k-modes algorithm have removed the numeric-only limitation of the k-means.
iv. Notwithstanding the changes, it still maintains the efficiency of clustering huge amount of categorical data sets

The frequency-based method is used to find (or update) the modes of clusters

3.3.2 Disadvantages of the Existing System

There are basically two problems that can arise when using either the original k-modes or the extended k-modes for cluster analysis. They include:

i. The method can only guarantee a locally optimal solution.
ii. There had been no reliable indices (or statistics) that can be used with both k-modes and extended k-modes to ascertain the true K (number of clusters) in the datasets
iii. The clusters can have convex shapes.
iv. It is not efficient in clustering mixed dataset

3.4 Analysis of the Proposed System

A hybrid clustering algorithm to cluster medical mixed dataset for decision making on resource management and allocation in the hospital is proposed. The proposed hybrid clustering algorithm consists of k-representative clustering algorithm and k-means clustering algorithm. K-representative clustering algorithm is one of the simplest unsupervised learning algorithms used for clustering categorical datasets, while k-means clustering algorithm clusters numerical data. The developed system will facilitate an understanding of how clustering methods will be used to develop patient resource utilization classification schemes, for the results of the proposed hybrid clustering algorithm for clustering mixed data sets.

This thesis proposed the development of a hybrid clustering algorithm using K-representative and K-means clustering algorithms for inpatients and outpatient's treatment dataset to optimize medical resources allocation. The outcome of the hybrid clustering algorithm overcomes the drawbacks inherited by K-prototype clustering algorithm from K-means and K-mode algorithms. The proposed hybrid algorithm produced better clusters of in-patients and out-patients treatment datasets which reveals knowledge and intelligence for medical resource allocation.

3.4.1 System Design

System Design was used to design the functions of the proposed hybrid clustering algorithm.

3.4.2 Architecture of the proposed system

Figure 3.3 illustrates the architecture of the proposed system

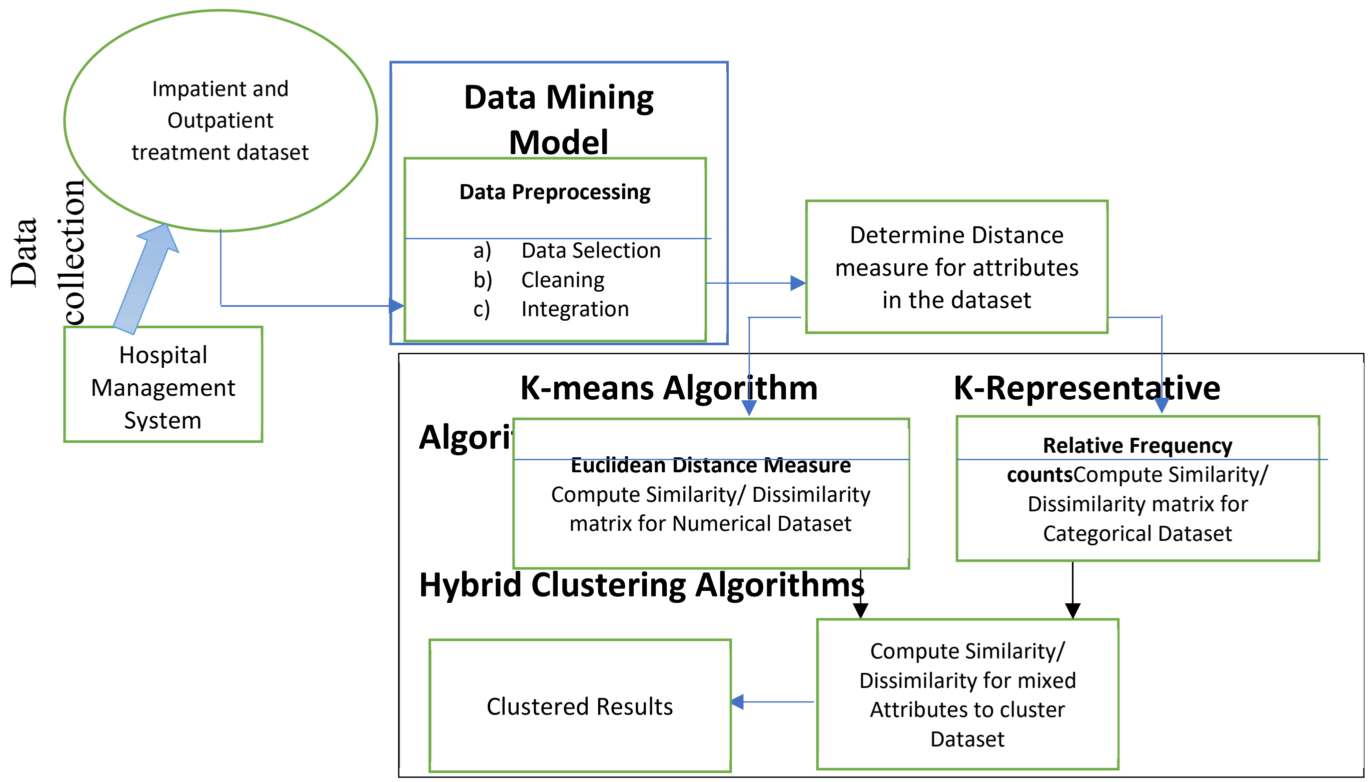


Figure 3.3: Architecture of the proposed system

From the architecture of the proposed system data of inpatients and outpatients are collected from the hospital management system. The data are preprocessed to form a metric for machine learning analysis by data: selection, cleaning and integration. The developed hybrid algorithm is used to cluster the dataset in various groups of clusters. The K-means algorithm handles the numerical datasets while the K-representatives handle the categorical dataset. Together, these algorithms processed the combined dataset comprising treatment records for both inpatients and outpatients.

3.4.3 Advantages of the Proposed System

The advantages of the proposed system are as follow:

1. It is efficient to cluster hospital large mixed datasets for decision making.
2. It is not time consuming
3. It is an efficient decision support system for resource management system in hospitals.
4. There are no duplications of result.

3.4.4 Disadvantages of the Proposed System

The disadvantages of the proposed system are as follow:

1. Outliers has effect on its' clustering centre thereby convergence to local minimum.
2. Wrong selection of initial value of k cause the algorithm to have a premature output convergence.

The diagram for High level model of the proposed system is shown in Figure 3.4

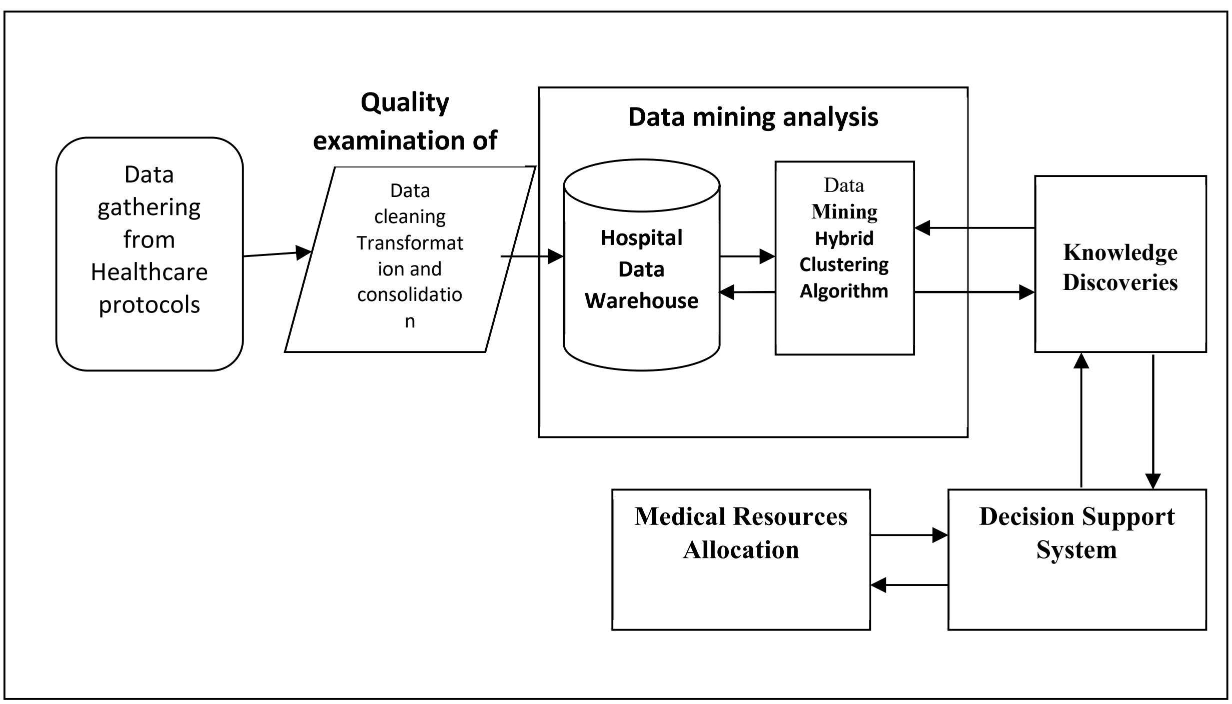


Figure 3.4: High Level Model of the proposed system

The components of the architecture of the proposed system are discussed as follow:

1. **Data gathering from healthcare protocols:** That is first segment of Medical Data Mining Life Cycle which begins off evolved with the gathering of big quantity of data generated with the aid of using healthcare transactions that is too complicated and voluminous to be processed and analyzed. Data from right here beneficent vicinity have been accumulated hospital, laboratories, radiology and administration.
2. **Data cleaning Transformation and consolidation**: Data mining is one of the obligations with inside the procedure of known how discovery from the database. The stairs with inside in information discovery Data mining (KDD) process contains:

   i. Data cleaning: On this segment nostril data and inappropriate data are eliminated from the collection.

   ii. Data integration: That is which more than one data sources often heterogeneous are blended in a common source.

   iii. Data selection: The data applicable to the evaluation is determined on and retrieved from the data collection.

   iv. Data transformation: It is also regarded as data consolidation; in this segment the chosen data is converted shape suitable for the mining procedure.

   v. Data mining: It is farther crucial step wherein smart strategies are implemented to extract doubtlessly beneficial patterns.

   vi. Pattern evaluation: On this step, exciting patterns representing know how are diagnosed primarily based totally on given measures.

vii. **Knowledge Discoveries (**Knowledge representation): It is miled the last segment wherein the discovered knowledge is visually offered to the user. This crucial step makes use of visualization strategies to assist users understand and interpret the data mining results.

3. **Data Warehouse:** Data warehouse (DW or DWH) additionally called an corporation data warehouse (EDW) is a devices used for reporting and data analysis. DWs are imperative repositories of incorporation data from one or more disparate sources. They store modern day and data ancient are used for growing analytical reports for knowledge workers throughout the enterprise. Examples of reports should variety from annual and quarterly assessment and tendencies to special daily sales analyses. The data saved with inside the warehouse is uploaded from the operational systems (such as marketing, sales, etc., shown in the figure to the right). The data might also additionally by skip via an operational data store for extra operations earlier than it is far the DW for reporting.
4. **Medical resources allocation:** The healthcare area faces robust strain to lessen value even as growing amounts of services are delivered**.** One approach that may be used to curb these with problems is the usage of healthcare information systems for support and expertise management. Healthcare facilities have at their disposal substantial quantity of data. Thorough evaluation of available data on a given problem can lead to more efficiency decision-making. The venture is to extract applicable knowledge from this data and act upon it in a well-timed manner. The era of information and knowledge calls for data prepared right into a beneficial form.
5. **Decision Support System:** understanding discovery project has first of all clean desires in thoughts for choice guide system.

3.5 Object Design

Object design was produced using the Unified Modeling Language (UMLs). It includes Algorithm. Pseudo code, use case diagram, and activity diagram.

3.5.1 Algorithm of the proposed Hybrid System

Apply the hybrid algorithm on every of the partition of the dataset to get the primarily clusters of every partition through preliminary initial value from the end result of swarm to the location of its fine particle.

Step 1: *Calculate the clustering centre of every clusters in all walls separately.*

Step 2: *Calculate the goal feature of the combined dataset.*

Step 3: *Repeat step 2 for all partitions.*

Step 4: *Ultimately, primarily based on the minimum distance criterion, assign every data factors to the cluster to which it has minimal distance, in order to limit the inside groups sum of squared errors.*

The Mathematical model for the proposed hybrid system which consist of numerical and categorical algorithm is shown in equation (3.1).

$$d_2(X,Y) = \left(\sum_{j=1}^{p} \left(x_j - y_j\right)^2 + \gamma \sum_{j=p+1}^{m} rf\delta\left(x_{j},,,, \right) \right) \tag{3.1}$$

Where:

i. The first term is numeric attributes and the second term is the easy matching dissimilarity degree on the categorical attributes.

ii. The weight γ is used right here to keep away from favoring both the categorical or numerical attributes. x_j *and* y_j are the datapoints of the dataset for j=1 …p, m.

iii. rf is the relative frequency.

3.5.2 Pseudo Code of the proposed Hybrid System

The pseudo codes are as follow:

FOR i = 1 TO QuantityOfObjects

Mindistance= **Distance**(X[i],O_prototypes[1])+ λ * **Sigma**(X[i],C_prototypes[1])

FOR j = 1 TO NumberOfClusters

// compute the distance measure for the dataset stored as a matrix andadd the weight,

distance= **Distance**(X[i],O_prototypes[j])+ λ * **Sigma**(X[i],C_prototypes[j])

IF (distance <Mindistance)

Mindistance = distance

Cluster = j

ENDIF

ENDFOR

Clustership[i]=cluster

ClusterCount[cluster] + 1

FOR j=1 TO NumberOfNumericAttributes

SumInCluster[cluster,j] + X[i,j]

```
        O_prototypes[cluster,j]=SumInCluster[cluster,j]/ClusterCount[cluster]

ENDFOR

FOR j=1 TO NumberOfCategoricAttributes

        FrequencyInCluster[cluster,j,X[i,j]] + 1

        C_prototypes[cluster,j]=HighestRelativeFreq(Cumulative Frequencyincluster,cluster,j)

ENDFOR

ENDFOR

//Initial allocation process.

moves=0

FOR i = 1 TO NumberOfObjects

        //(To find the cluster whose prototype is the nearest to object i.)…

        IF (Clustership[i]<>cluster)

        moves+1

        oldcluster=Clustership[i]

        ClusterCount[cluster] + 1

        ClusterCount[oldcluster] - 1

        FOR j=1 TO NumberOfNumericAttributes

                SumInCluster[cluster,j] + X[i,j]
```

```
            SumInCluster[oldcluster,j] - X[i,j]

            O_prototypes[cluster,j]=SumInCluster[cluster,j]/ClusterCount[cluster]

            O_prototypes[oldcluster,j]= SumInCluster[oldcluster,j]/ClusterCount[oldcluster]

            ENDFOR

FOR j=1 TO NumberOfCategoricAttributes

      RelativeFrequencyInCluster[cluster,j,X[i,j]] + 1

      FrequencyInCluster[oldcluster,j,X[i,j]] - 1

      C_prototypes[cluster,j]=HighestRelativeFreq (cluster,j)

      C_prototypes[oldcluster,j]=HighestRelativeFreq (oldcluster,j)

ENDFOR

ENDIF

ENDFOR
```

3.5.3 Activity Diagram of the Proposed Hybrid System

The Activity Diagram of the Proposed System is shown in Figure 3.5.

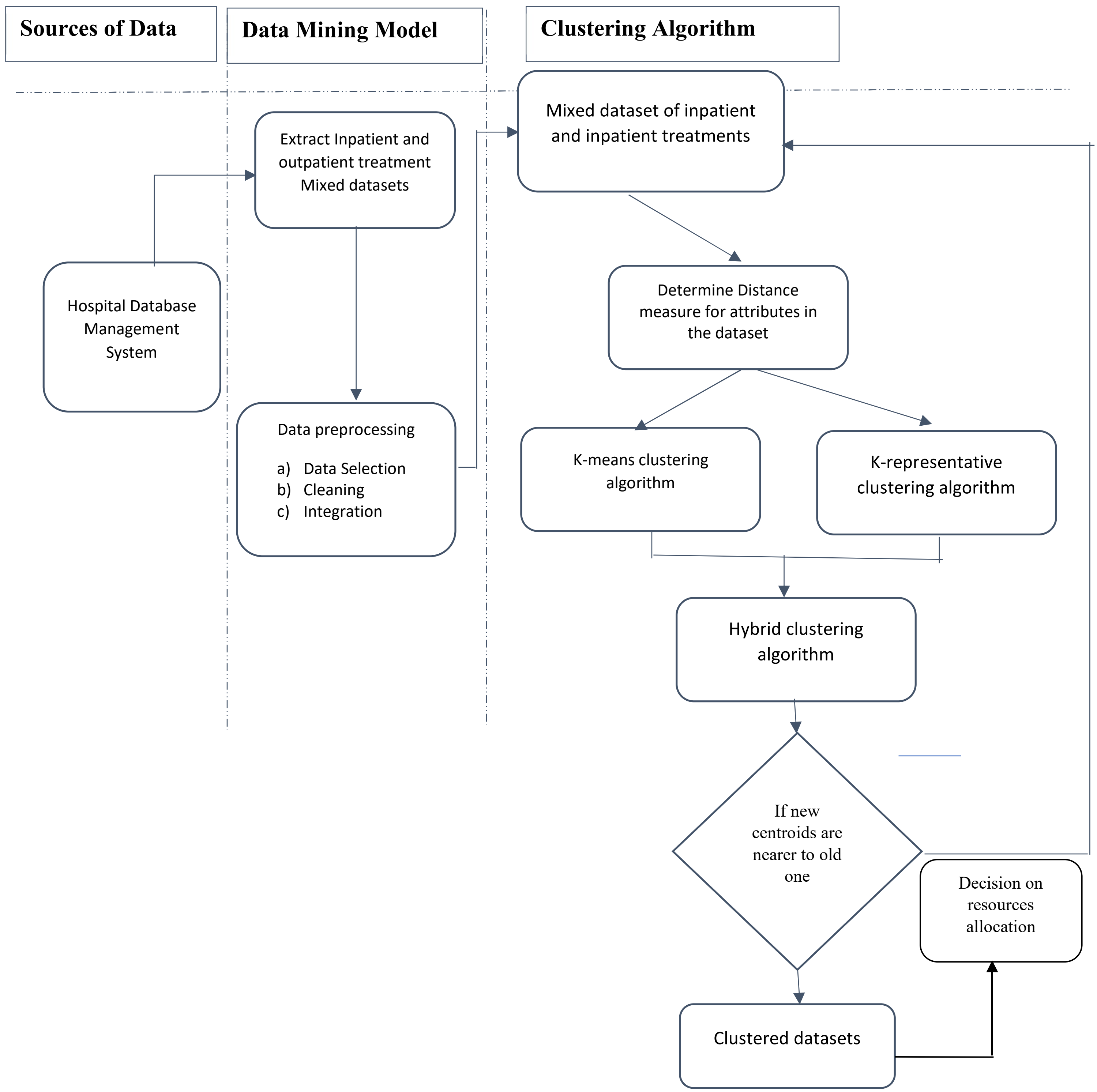


Figure 3.5 Activity Diagram of the Proposed System

Figure 3.5 Activity diagram described the connectivity of the proposed system objects or functions. The inpatients and outpatient's treatment mixed datasets are extracted from the Hospital database management system. The extracted datasets were preprocessed to obtain a matrix of computational mixed dataset patient's treatments. The proposed mechanism applied the proposed hybrid clustering algorithm to create clusters of patient's treatments. The clustering analysis is used for decision support guide for resources allocation in the Hospital.

The Use case diagram is shown in Figure 3.6.

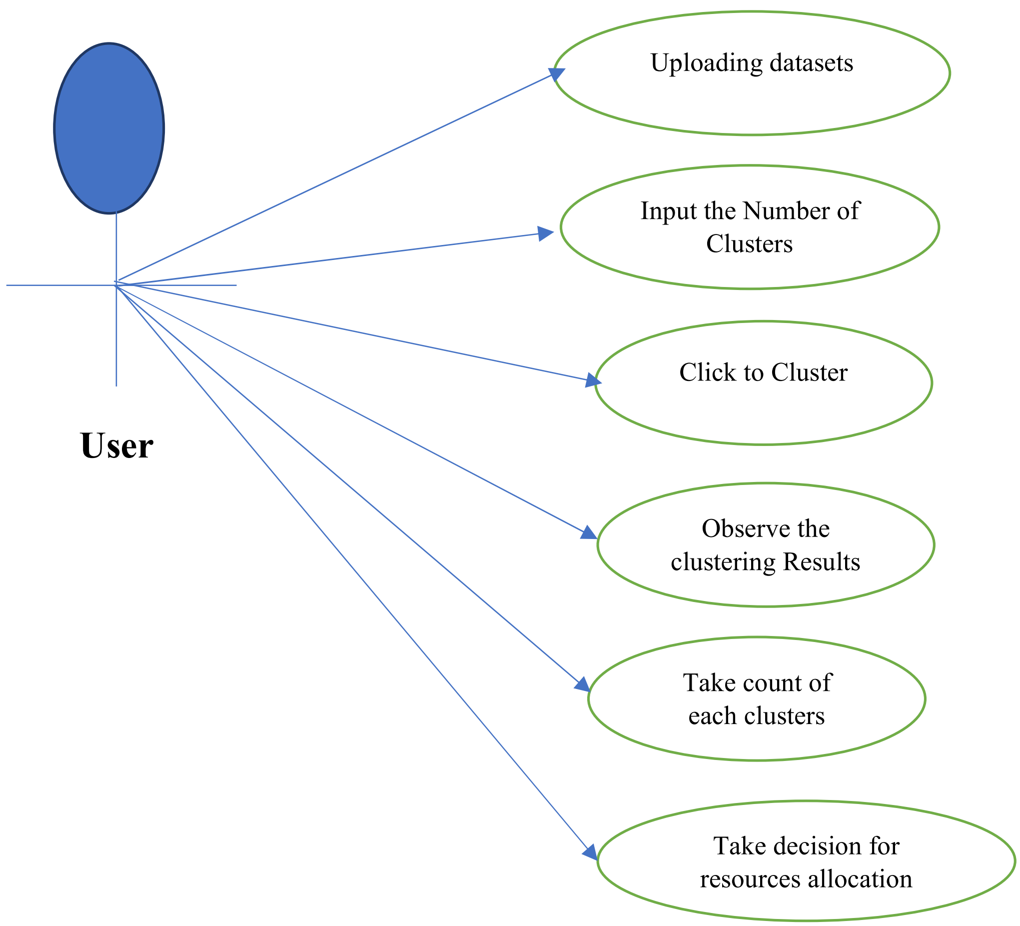


Figure 3.6: Use Case Diagram

Figure 3.6 describes the use case diagram shows the activities of the user on developed system. The user clicks to upload the dataset in the program. After that the use inputs the number of clusters to be created and clicks on cluster to analyze the patient's treatment datasets. The outputs of the program are observed and the user takes records of the count of treatment in each cluster. The results obtain are used for decision support guide in the hospital. The input/output design is shown in Figure 3.7.

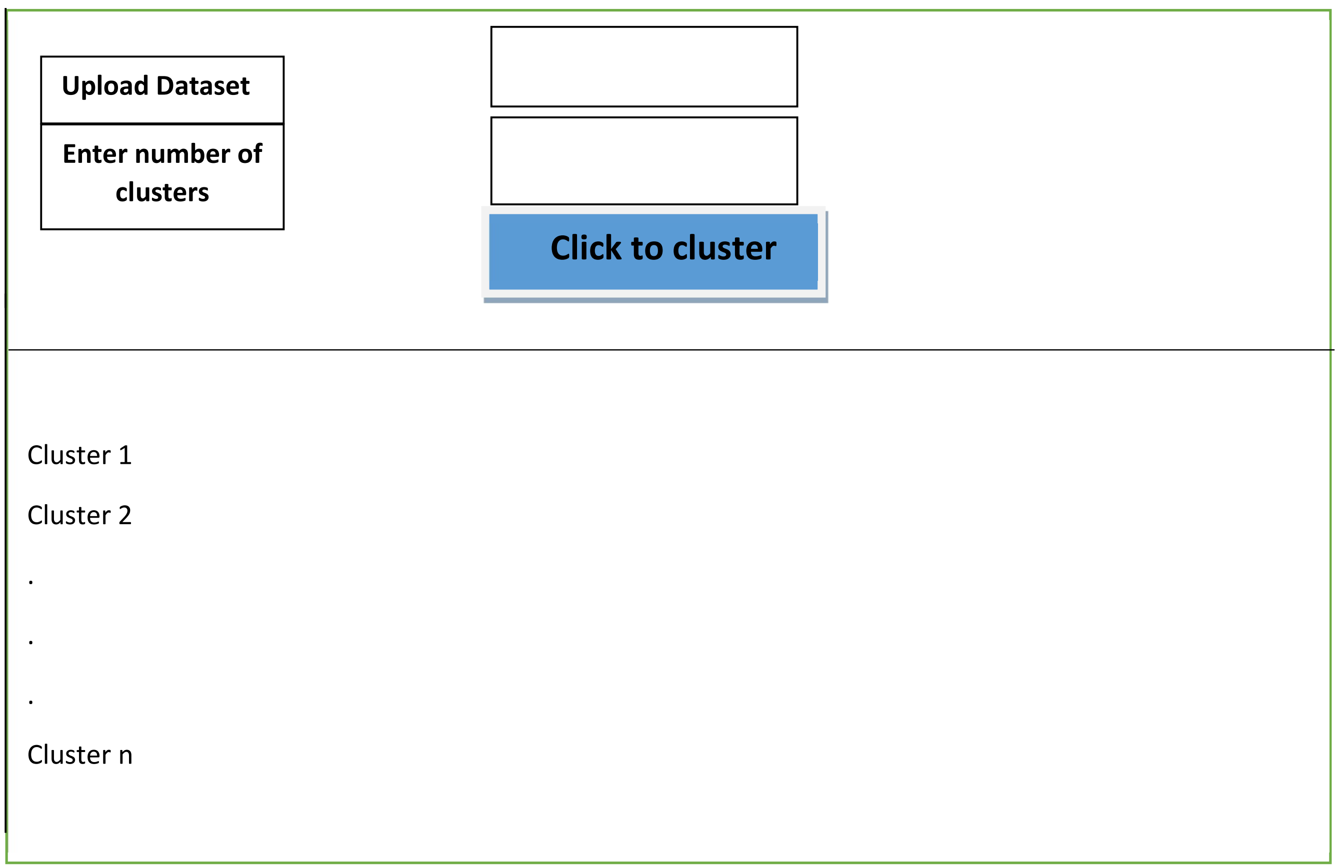


Figure 3.7: Input/output Design is shown

3.6 Implementation

Implementation of the machine includes writing a computer application, checking and ultimately extrude over to the new system.

3.6.1 Systems Requirement

The gadget requirement to useful resource the green and powerful usage of the newly designed software program is assessed into:

i. Hardware and
ii. Software program

3.6.2 Hardware Requirement

The hardware requirements are:

i. 4GB RAM minimum
ii. 2 GB Hard disk capacity.
iii. Printer: An EPSON, Hewlett Packard etc. with cultured printing attribute.
iv. UPS: This is an acronym for uninterrupted power supply. It maintains the desktop computer briefly on for a while after strength failure.
v. A voltage stabilizer regulates the voltage supplied to PC devices, ensuring it remains within safe operating limits to prevent electrical damage to the equipment

3.6.3 Software Requirement

The software program necessities for this device are namely:

i. An Operating System (Windows 7, 8, 10 and 11)
ii. Java run time
iii. Net Beans Environment

3.6.4 Choice of the Programming Language

The research uses k-representatives to determine the number of clusters in JAVA. With its emphasis on Object Oriented Programming (OOP), JAVA enforces the good deal wanted appropriate software program practices along with the make use of interfaces and right code organization. Also, JAVA comes with a package deal of optimized equipped to apply data structures that proper programming quicker and much less mistakes prone.

JAVA has been very frequently criticized for the gradual of execution, incomparable with the one of a natively-compiled code. At the same as that is nonetheless objectively authentic, with the latest optimizations together with JAVA Hotspot, JAVA code can obtain the walking instances similar with the one C++. That is particularly the case whilst the JAVA integrated data types are used. For this reason, we've determined to layout the implementation of k-representatives with inside the manner that the maximum time-eating operations are done on such types.

3.6.5 System Testing

The set-up component is to restore the specific software and hardware for the operation of the system. While initializing the software, one ought to make sure that the distinctive hardware of the computer system is functioning also with the operating system.

On the checking out stage, each declaration should be examined to make sure that there are not any mistakes in program. If this isn't done, a mistake within the program may linger undetected for months and the intended goal of the system may be defeated. The improvement of software system entails collection of production activities wherein possibilities for injection of human facilities are enormous. Blunder may also start to arise on the inception of the procedure wherein the cam be imperfectly unique advert properly as later design and improvement stages.

Due to human lack of ability to carry out and talk with perfection, software program improvement is a done via way of means of a first-class warranty activity. Software program checking out is an important detail of software program improvement organization, to anticipate forty percentage of the entire challenge attempt on checking out. Checking out software is the technique of executing software with an aim of locating error. Checking out software is the manner of executing software with an aid of locating error. Testing uncovers different classes of error such as:

i. Syntax Error
ii. Logical Error
iii. Runtime Error and
iv. Compilation Error
v. The errors that were found were debugged.

3.6.6 Sources of Data

The data used to analyze the performance of the developed hybrid clustering algorithm were collected from tertiary health institution. The primary health provider takes care of the students, staff and the communities' health treatments. The treatment prescriptions are recorded manually in a register. It was selected, clean, and transformed to analysis the developed clustering algorithm. The mixed dataset collected for the analysis in the work consist of treatment prescriptions for inpatients and outpatients.

The data was preprocessed by selection, cleaning, and integration.

i. Selection: The relevant information about the inpatients and outpatients was selected for the study. Information such as the name, matriculation number of the patients was discarded.
ii. cleaning: missing attributed and incomplete records was also discarded

iii. Transformation: Some of the categorical data were converted to numerical data for easy manipulation of the proposed hybrid clustering algorithm. The transformed data is shown in Table 4.1.

The dataset consists of three attributes such as sickness, patient's type and treatments. The 370 instances of the mixed dataset were collected. It includes: malaria, food poison, headache, restlessness, asthma, stroke and typhoid etc. The abridged form of the dataset is represented in Table 3.1.

Table 3.1: Summary of the treatment records of inpatients and Outpatients

S/N	Sickness	Patient_type	Treatments
1	Hearing Loss	1	3
2	Headache	1	3
3	Heart Disease	1	3
4	Food poison	2	3
5	Malaria	1	3
6	Asthma	1	3
7	Heart Attacks	1	3
8	Infertility	1	4
9	Restlessness	1	3
10	HIV/AIDS	2	3
11	Restlessness	1	4
12	Typhoid	2	3

13	HIV/AIDS	1	4
14	Typhoid	2	3
15	Restlessness	1	3
16	Typhoid	1	3
17	Anxiety	1	3
18	Anxiety	1	3
19	Anxiety	1	3
20	Malaria	1	3
21	Malaria	2	4
22	Malaria	1	3
23	Typhoid	1	3
24	Food poison	1	3
25	Malaria	2	4
26	Liver	2	4
27	Food poison	1	3
28	Food poison	1	3
29	Malaria	2	3
30	Bladder Cancer	2	3
31	Malaria	2	3
32	Malaria	1	3
33	Malaria	1	3
34	Malaria	1	3
35	Malaria	1	3

36	Malaria	1	4
37	Stroke	1	4
38	Malaria	1	4
39	Bipolar Disorder (BD)	1	4
40	Malaria	2	3

Where1= outpatients, 2= inpatients, 3 = tablets, 4 = injections

Source of data: Medical center of Federal University of Technology, Owerri. (2015).

3.6.7 Validation of the Hybrid Clustering Algorithm

The accuracy of the proposed system was determined using clustering accuracy. The formula is shown in equation (3.2)

Purity Measure=

$$\frac{1}{n}\sum_{i=1}^{j} d_i \tag{3.2}$$

`

A cluster is known as a natural cluster if all the terms belong to a single class. The clustering accuracy is defined as 'r', wherein di is the quantity of data objects that arises in each cluster C_l and its corresponding categorized elegance, which has the maximal value j and *n* is the number of objects in the data set. The error of creating clusters *e* is defined as e = 1 – r. (3.3).
If a partition has a clustering accuracy of a hundred percentage, it manner that it has most efficient natural clusters (Huang, 1998).

3.6.8 Output/Input Specifications

The output specification is the clustered results produced from the analysis after testing the scalability of the proposed hybrid clustering algorithm. Whereas the input specification consists of the number of clusters, the attributes of the dataset such as diagnoses, patient_type and treatments (tablets, injections). The output/input specifications are represented in Table 3.2.

Table 3.2: **Output/Input Specification**

Inputs Specification	**Output Specifications**
Number of clusters, Diagnoses, Type of patients and Treatments	Clusters of diagnoses, type of patients, and various treatment

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Results

The system interface for the developed system is shown in Figure 4.1.

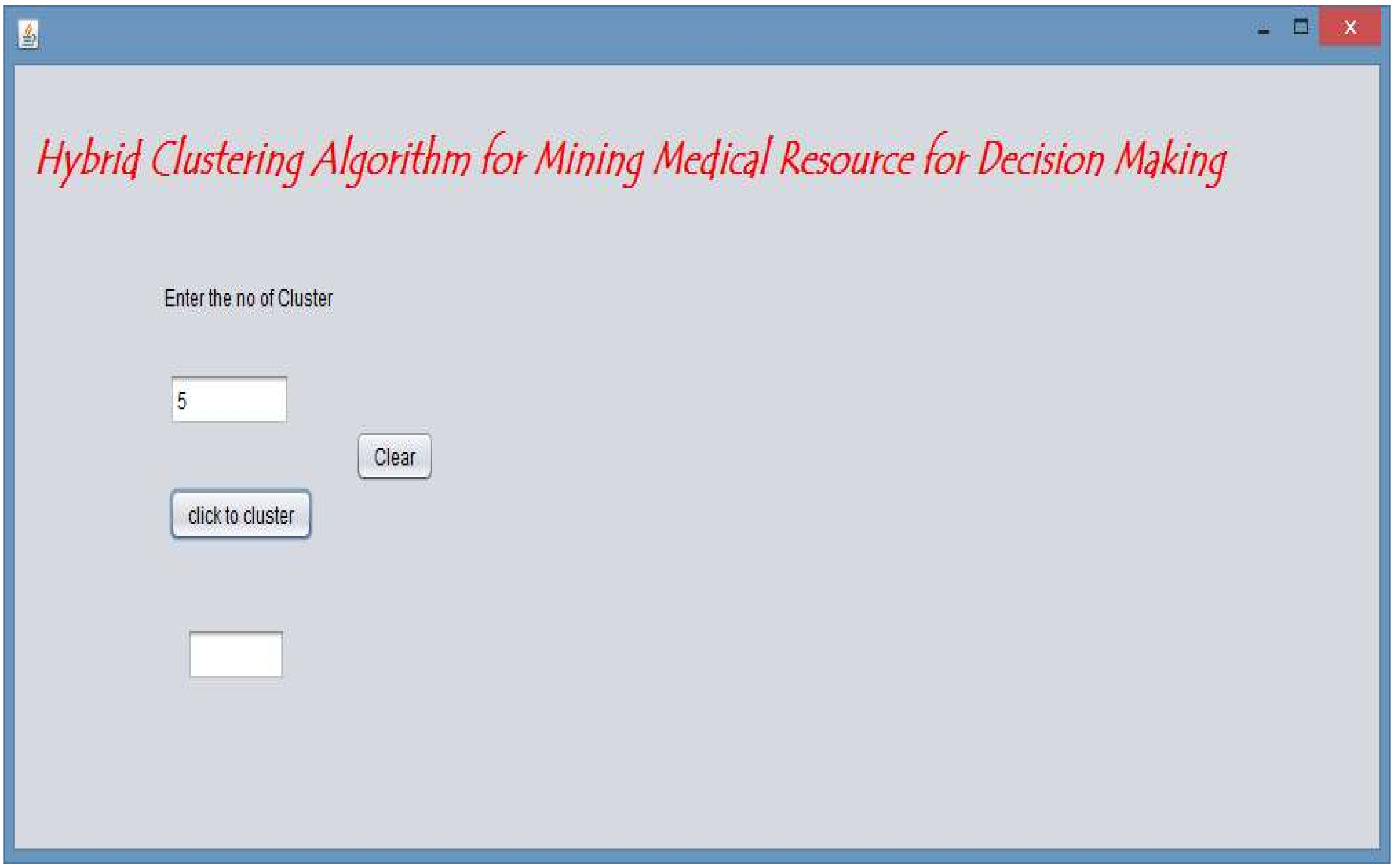


Figure 4.1: User Interface of the system

The clustered output is shown in various Figures 4.2.

In fig 4.2, we have a hybrid clustering algorithm for mining medical resources for decision making containing number of clusters and two buttons stating the clear and click to cluster.

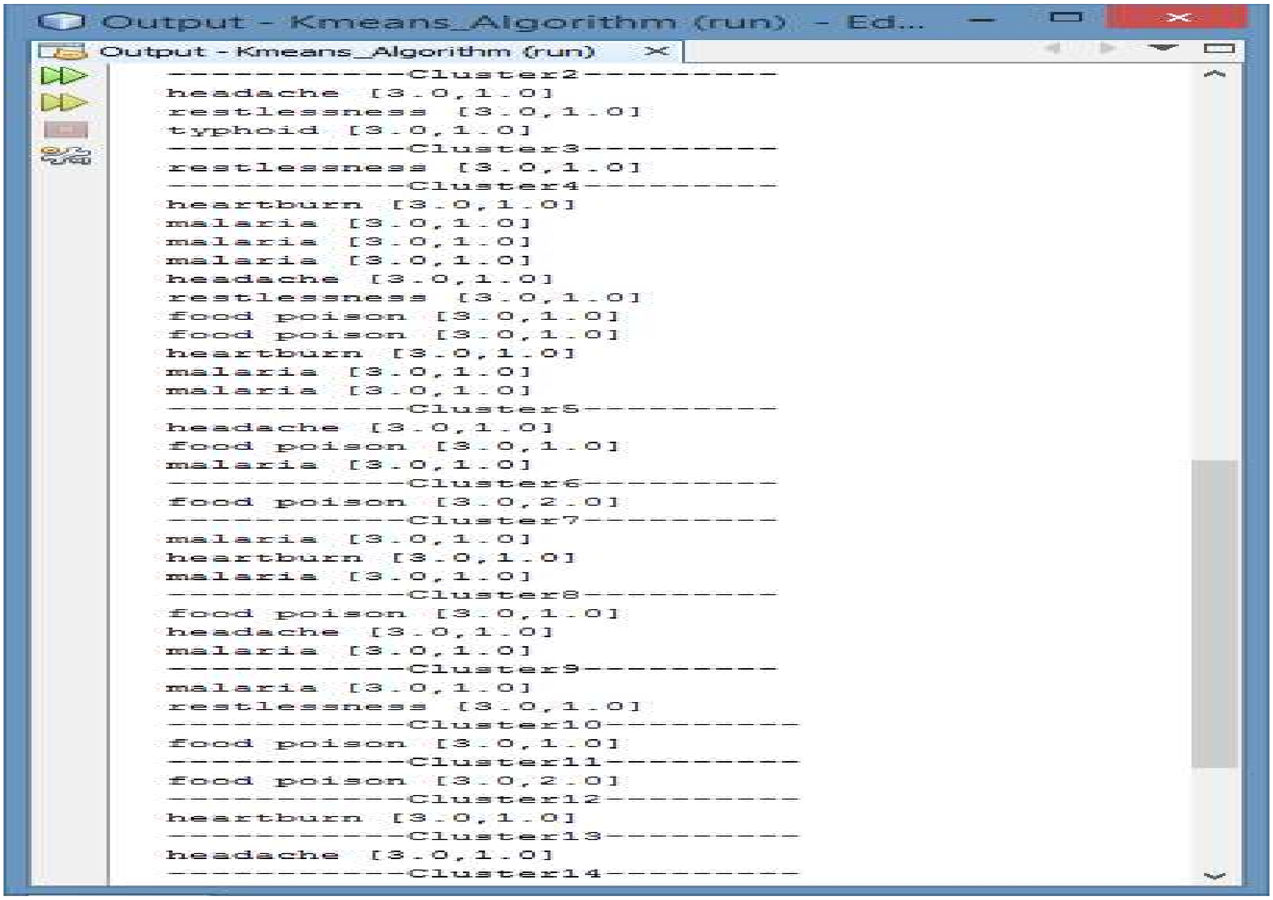


Figure 4.2a: Clustering results

In figure 4.2a shows the result for the hybrid clustering algorithm for mining medical resources for decision making containing the heartburn, malaria, headache, restlessness and food poison.

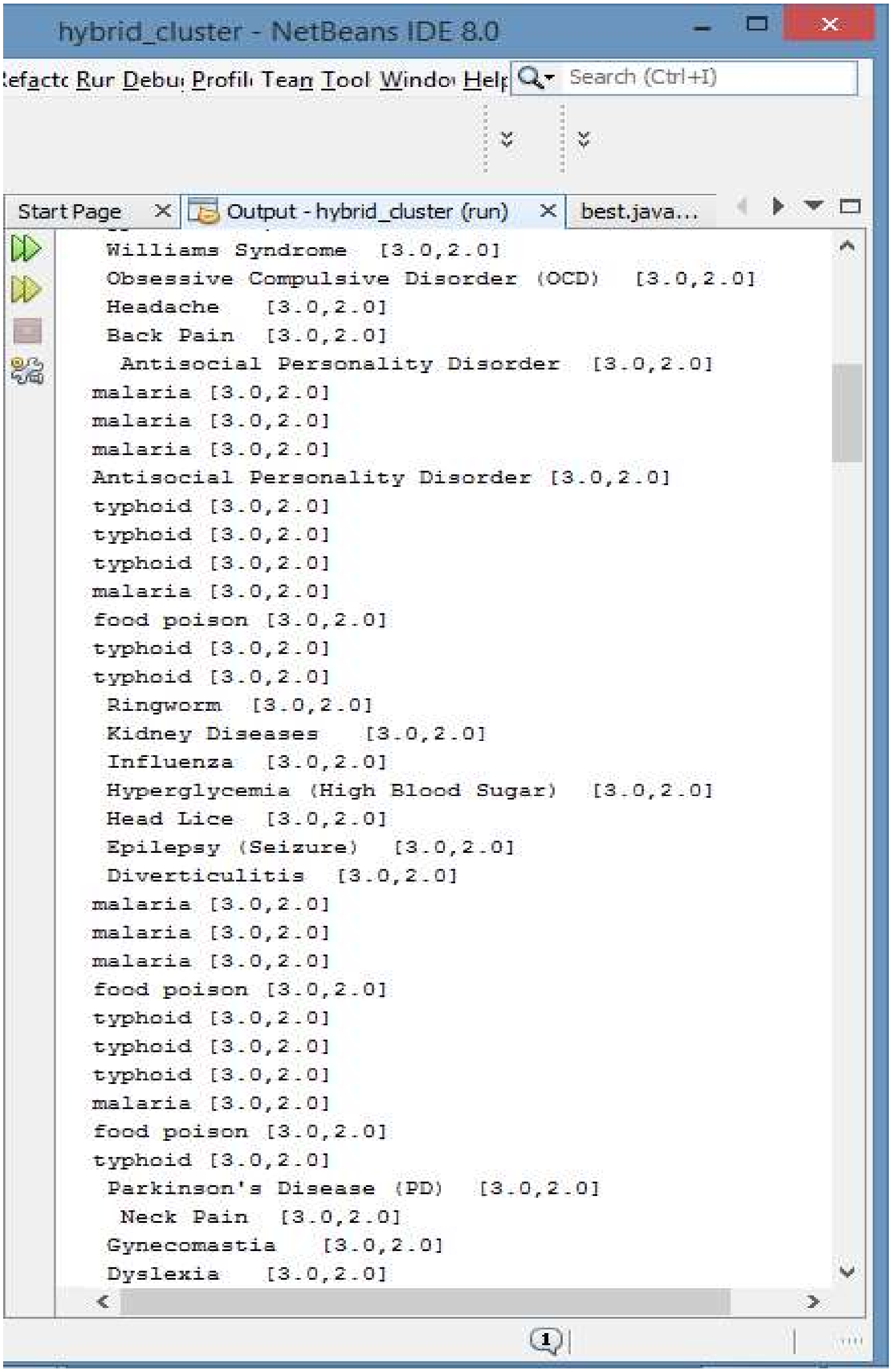


Figure 4.2b: Clustering results

The diagram in Figure 4.2b shows the result on hybrid clustering algorithm for mining medical resources for decision making for Williams Syndrome, Obsessive Disorder, Headache and back pain and others.

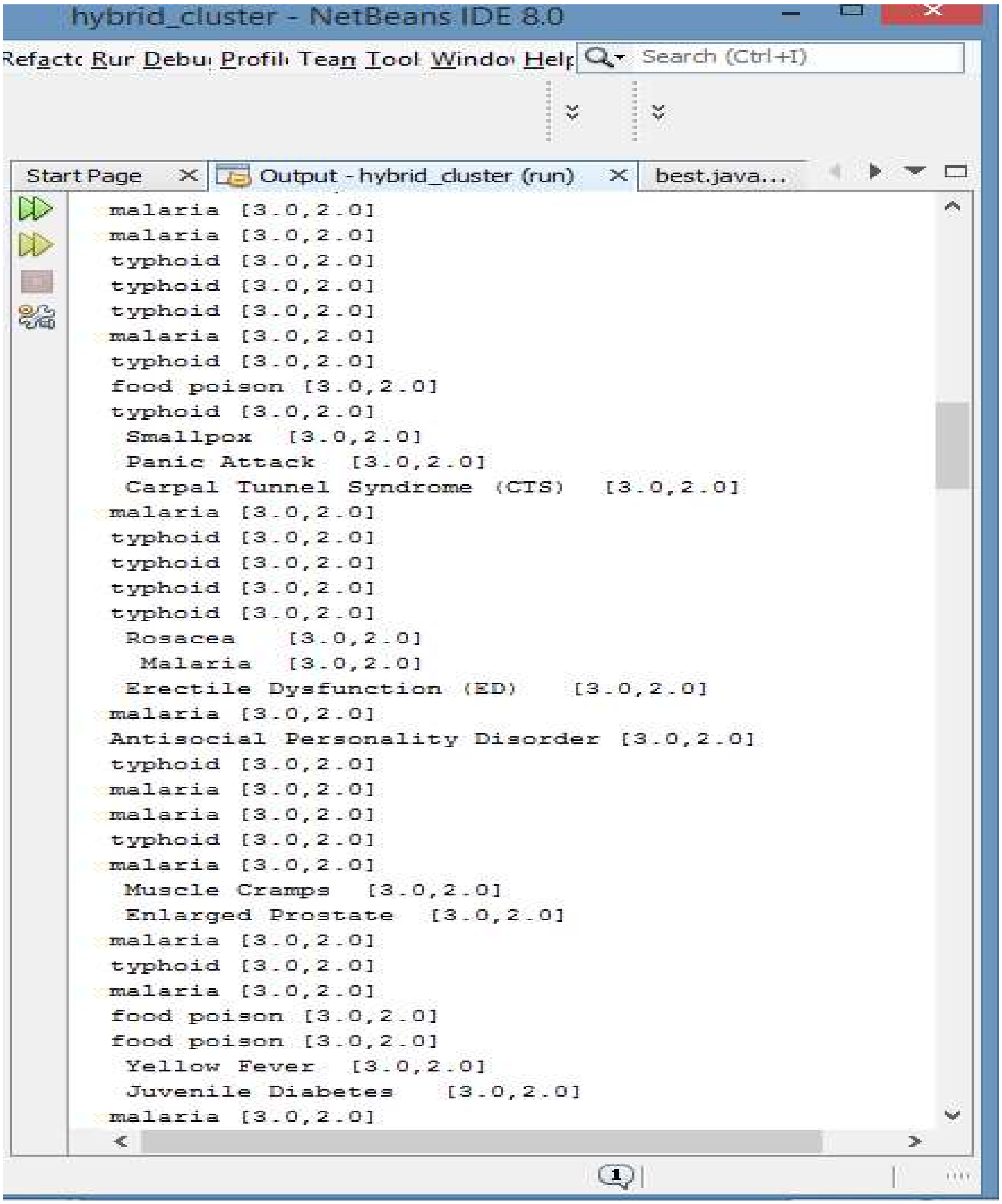


Figure 4.2c: Clustering results

The diagram In figure 4.2c shows the hybrid clustering algorithm for mining medical resources for decision making on malaria, typhoid, food poison, small pox, panic attack, carpal tunnel syndrome, rosacea, erectile dysfunction, muscle cramp and antisocial personality disorder.

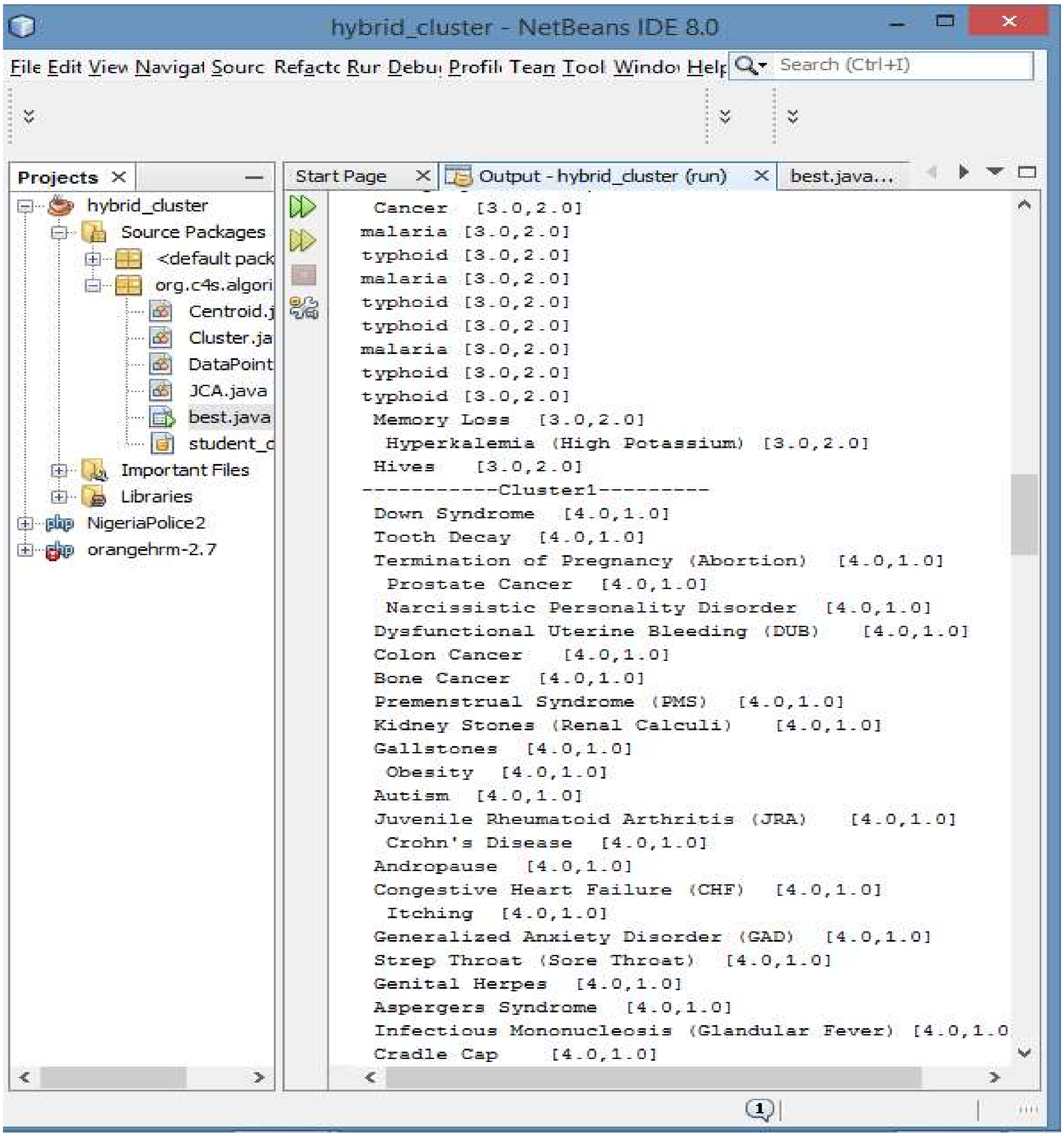


Figure 4.2d: Clustering results

The above diagram shows the clustering result for hybrid clustering algorithm for mining medical resources for decision making on Malaria, typhoid, Cancer, Memory Loss, Hyperkelemia (High potassium), Hives, Down syndrome, Tooth Decay, Termination Of Pregnancy, and Prostate Cancer.

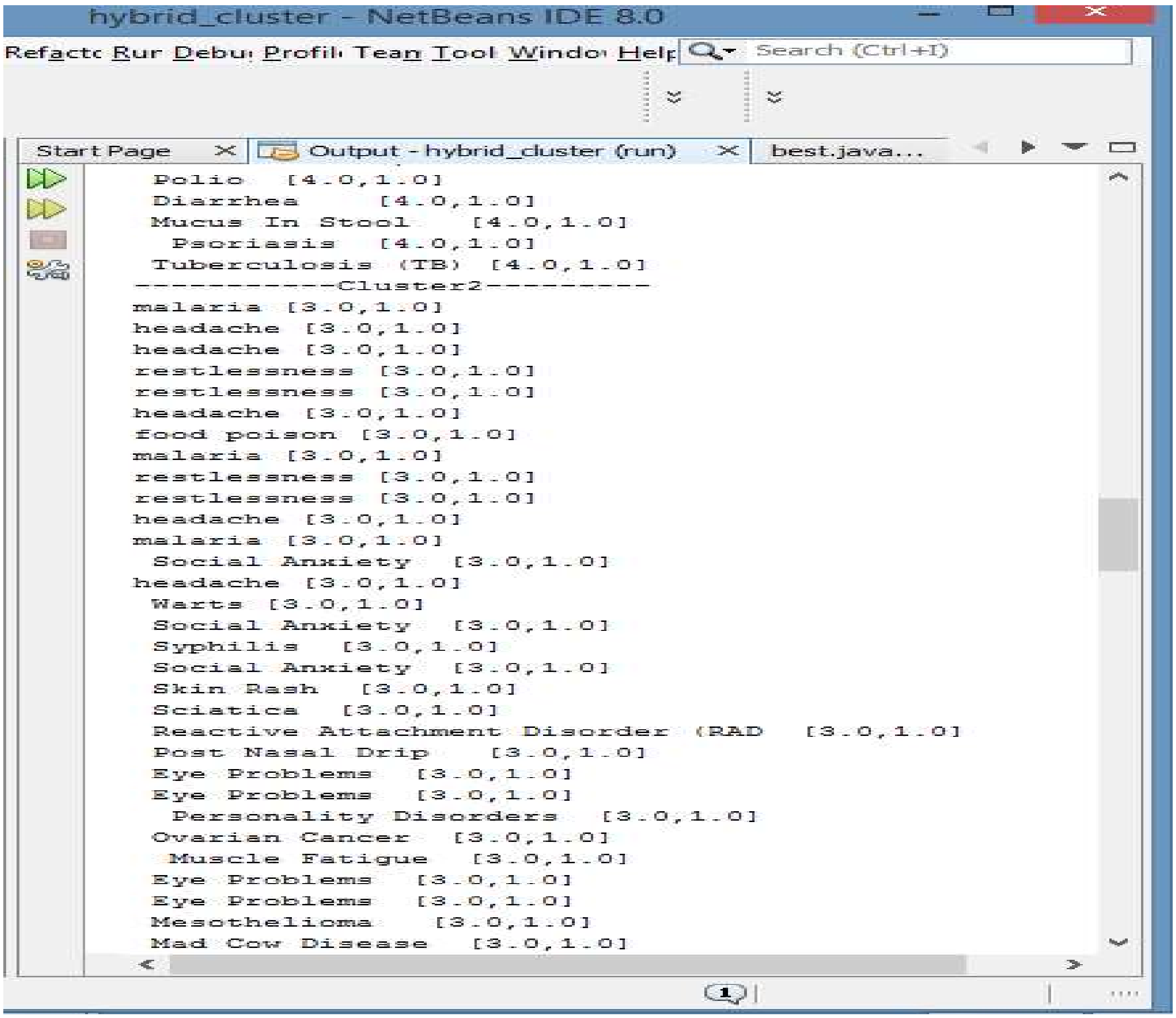


Figure 4.2e: Clustering results

In figure 4.2e is a clustering result on hybrid clustering algorithm for mining medical resources for decision making contains the result of Polio, Diarrhea, Mucus in Stool, Psoriaosis and Tuberculosis

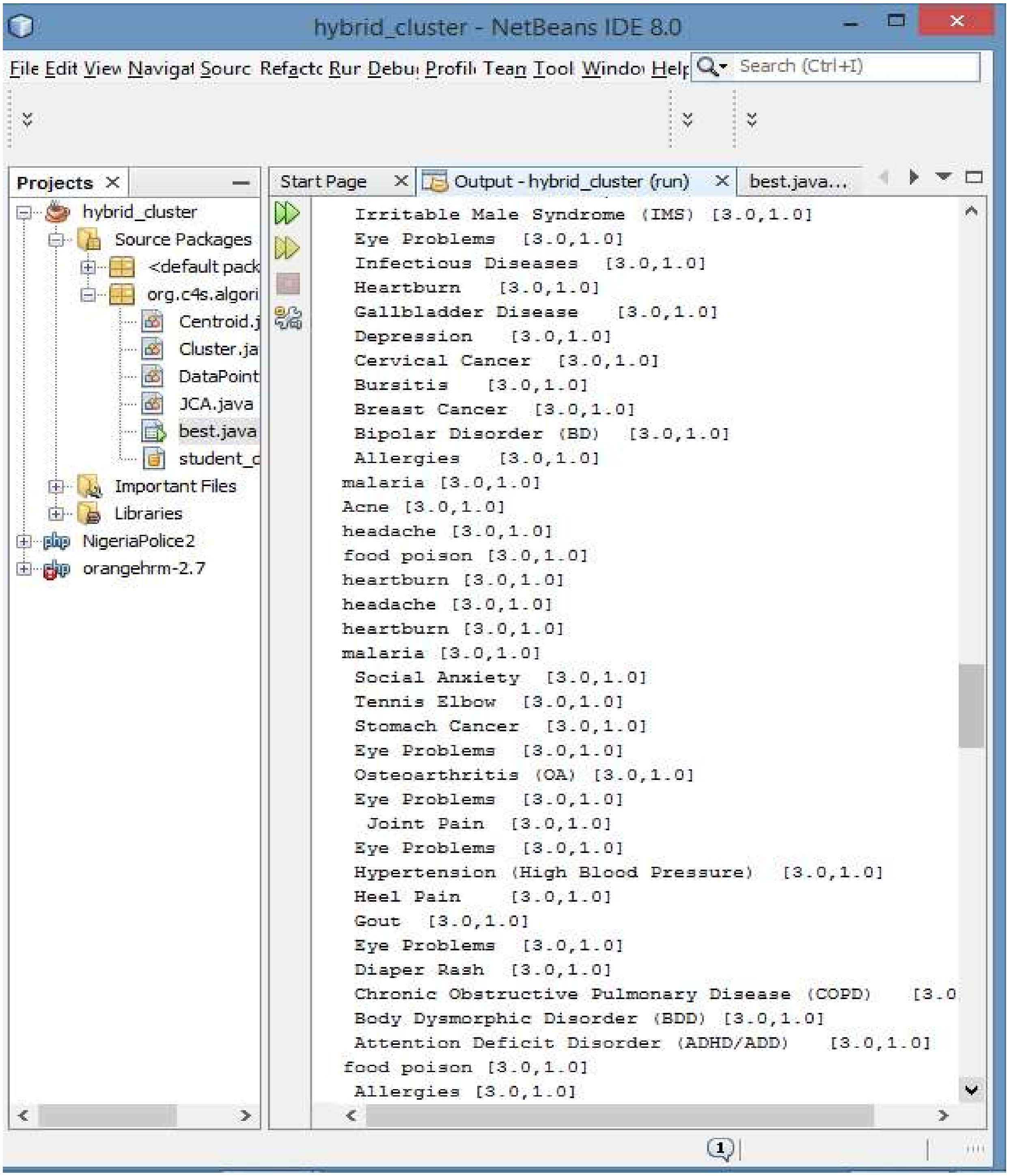


Figure 4.2f: Clustering results

The figure above is a clustering result of a hybrid clustering algorithm for mining medical resources for decision making on Irritable Male Syndrome, Eye Problem, Infectious Disease, Heartburn, Gallbladder Disease, Depression, Cervical Cancer, Bursitis, Breast Cancer, Bipolar Disorder and Allergies.

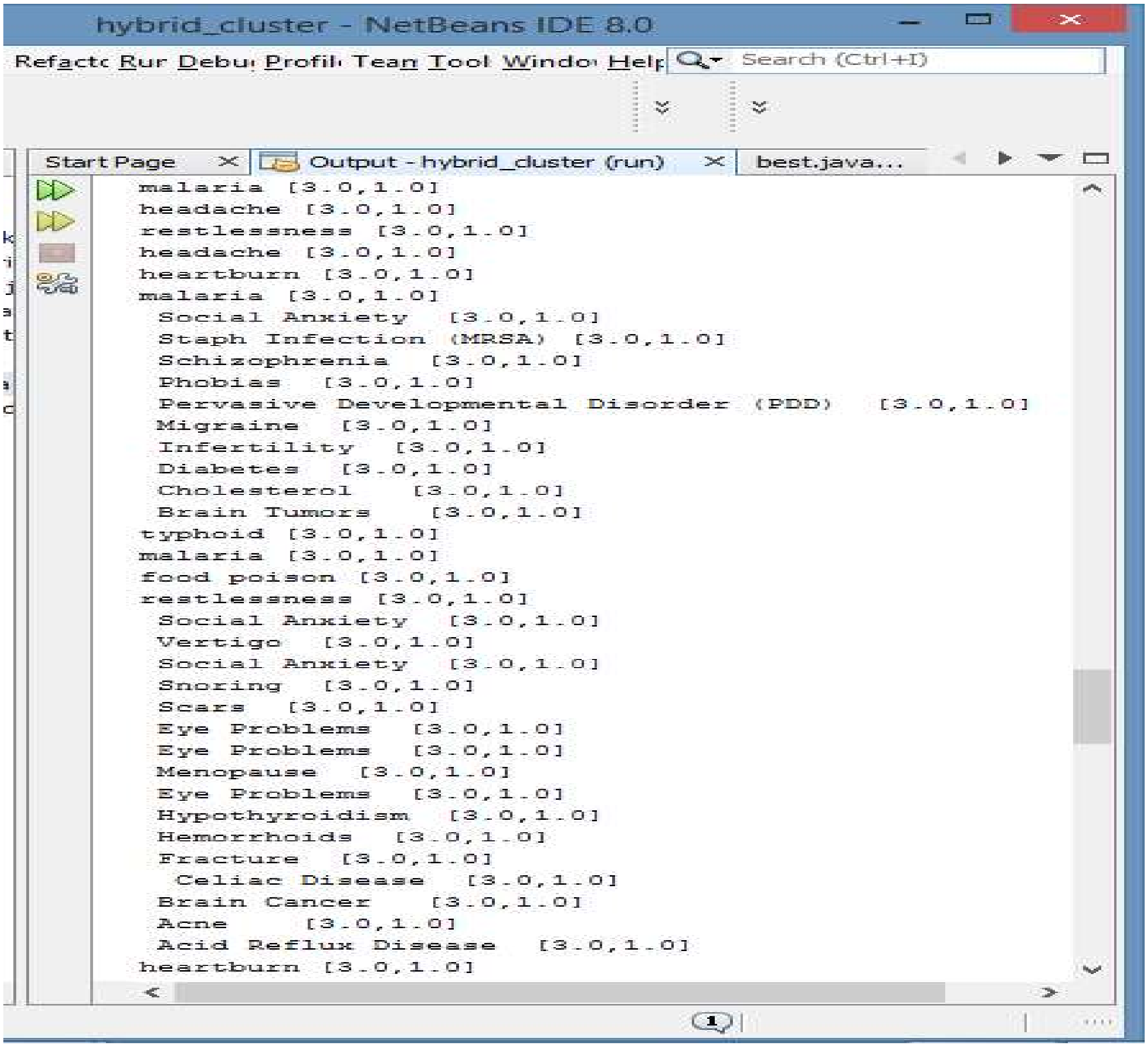


Figure 4.2g: Clustering results

In figure 4.2g is a result for hybrid clustering algorithm for mining medical resources for decision making on Malaria, Headache, Restlessness, and Heartburn, eye problem, food poison, social anxiety.

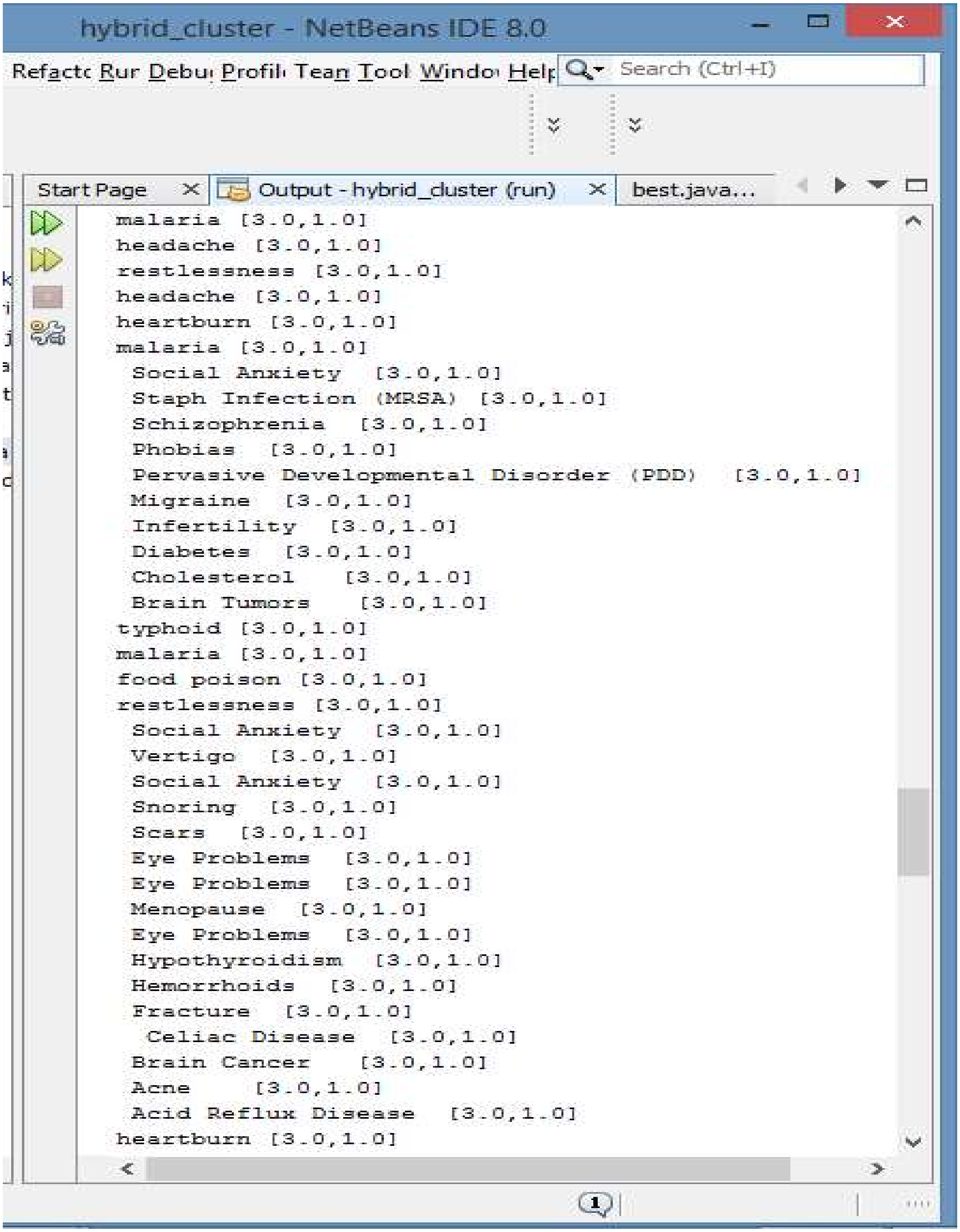


Figure 4.2h: Clustering results

Figure 4.2h is a clustering result for hybrid clustering algorithm for mining medical resources for decision making on Malaria, Headache and Heartburn.

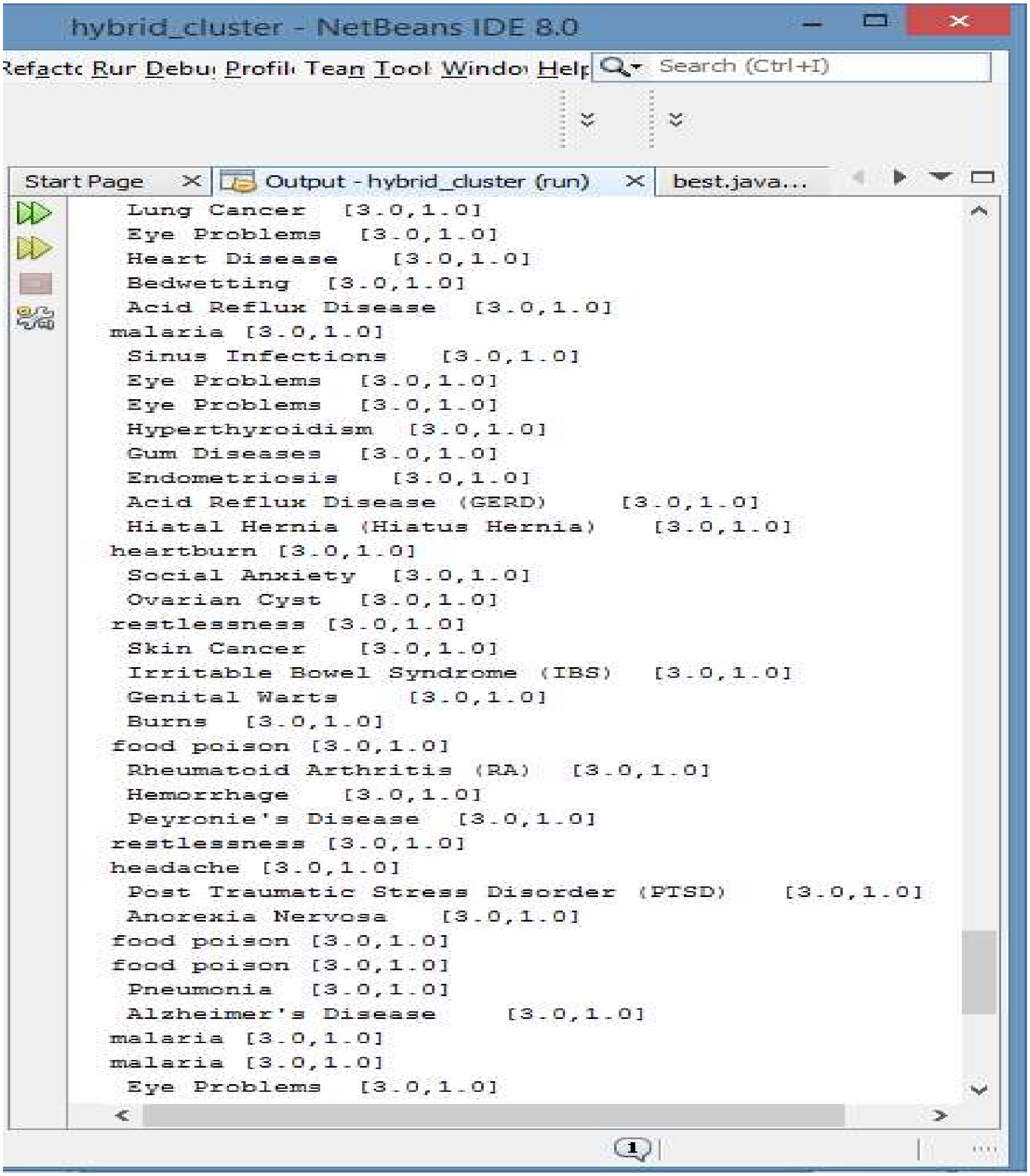


Figure 4.2i: Clustering results

Figure 4.2i shows the clustering result for hybrid clustering algorithm for mining medical resources for decision making on Lung Cancer, Eye Problem, Heart Disease, Bedwetting and Acid Reflux Disease and other health related issues.

The graphical representation of the clustered inpatient and outpatient treatments is shown in Figure 4.3.

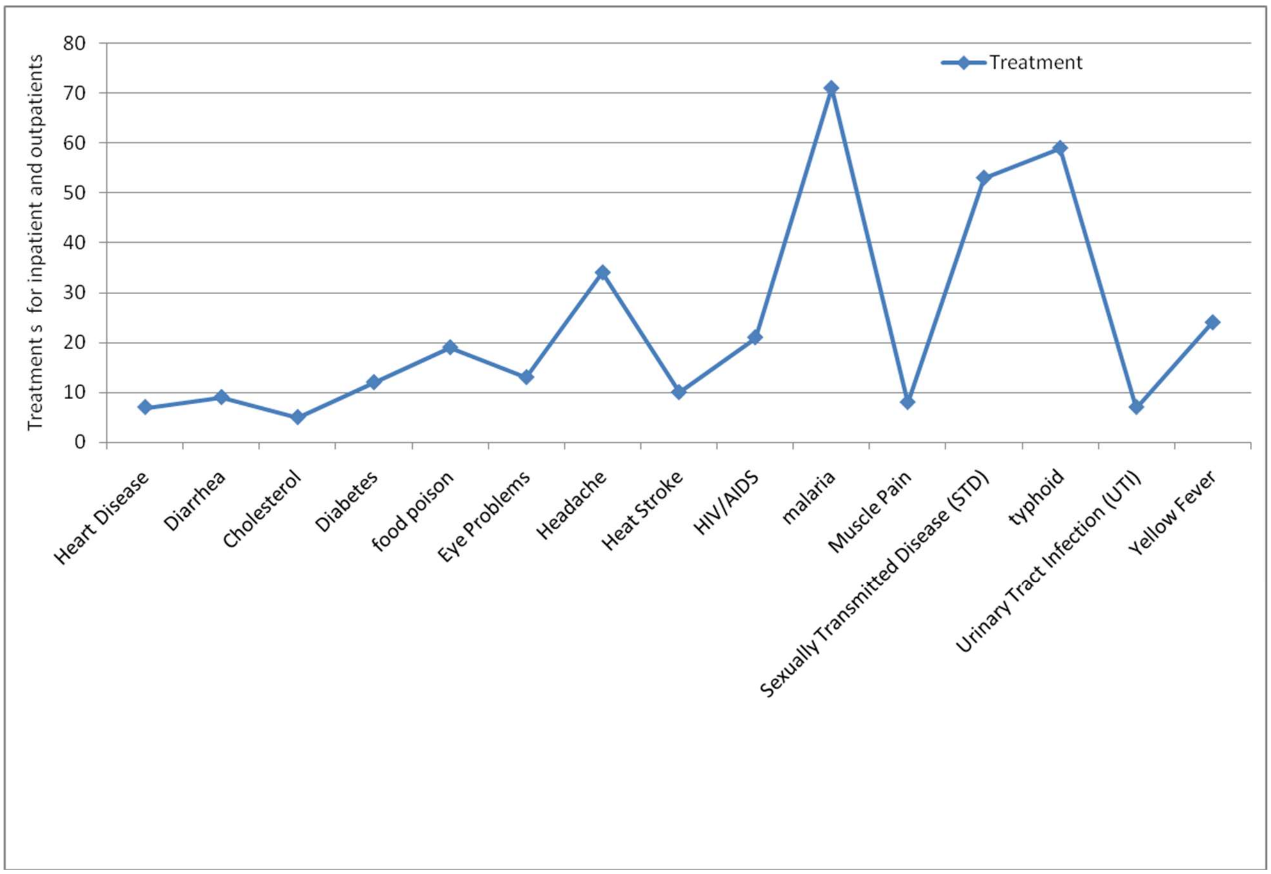


Figure 4.3: Representation of Cluster results of treatments of inpatient and outpatients

In figure 4.3, it shows the graphical representation of clustered result of treatments of patients and outpatients on Heart Disease, Diarrhea, Cholesterol, Diabetes, Food Poison, Eye problem, Headache, Head Stroke, HIV/AIDS, Malaria, Muscle Pain, Sexually Transmitted Diseases (STD), Typhoid, Urinary Tract Infection (UTI) and Yellow Fever.

For instance, the misclassification matrices for the test dataset are shown in Tables 4.1

Table 4.1: Misclassification matrices for the test data

Attributes	Cluster 1 (injection for outpatient treatment)	Cluster 2 (Injection treatment for inpatients)	Cluster 3 (Tablets for outpatient treatments)	Cluster 4 (Merged)	Total
Outpatient	126	0	195	1	322
Inpatient	0	31	0	17	48
					370

The attributes of inpatients and outpatients are in the first column of the matrices and the rest four columns are the four clusters. In Table 4.1 there is a one-to-one correspondence between clusters and inpatient and outpatient classes, which mean the instances in the same disease and treatments classes, were clustered into the same clusters.

The 27 clusters generated were classified into four (4) clusters based on treatment and injections. Cluster 1 represents injection for outpatient treatments while cluster 3 represents tablets for outpatient treatments. Also, cluster 2 represents injection treatments for inpatients while clusters.

The information in Table 4.1 was used to compute the misclassification matrix of each attribute of the dataset.

The misclassification matrix of each clustering result was used to define compute the *clustering accuracy* on the hybrid algorithms proposed by (Huang, 2010). The measure of clustering results called the *clustering accuracy,* as defined in equation (3.2) gave

$$r = \frac{(126+195+31+1+17)}{370} = 0.9972 \quad (4.1)$$

and e = 0.002703

4.2 Discussion of Results

This research work considered the inherited theories and underlining principles of the proposed hybrid clustering algorithm for the mining of hospital datasets. Previously, the existing problems in K-means was discussed and K-means is still left with some inherited problems like solving the problem of clustering mixed data to produce better clusters, which is the main interest of this work. Though some algorithms like representative K-means (which is a modification of K-modes algorithm (Xiaoliang *et al.,* 2020; Toan & Van-Nam, 2021) has proven to work well on categorical data, but not on mixed data. The forecourt is to look at possible ways of clustering mixed data with better results by considering a known similarity measure, mutual information (MI) with a defined model that will help produce the number of clusters as accurate as possible.

In this work, a hybrid clustering algorithm was developed using K-representatives and K-means Algorithms to mine medical records for medical resources allocation. The dataset of inpatients and outpatient treatment records was obtained from tertiary health institution primary healthcare medicals to test and validate the developed hybrid clustering algorithm.

The clustered results show that most of the cases of inpatient and outpatient's illness are Diarrhea, Cholesterol, Diabetes, food poison, Eye Problems, Headache, Heat Stroke, HIV/AIDS, malaria, Muscle Pain, Sexually Transmitted Disease (STD), typhoid, Urinary Tract Infection (UTI) and Yellow Fever. The results of the analysis revealed that Malaria has the highest clusters, followed by Sexually Transmitted Diseases (STD), Typhoid, Headache, and Food poison among the inpatients and out patients.

The trustworthiness of the proposed hybrid model was evaluated using misclassification matrices or confusion metric model. The accuracy of the proposed system produced 0.9972. The mechanisms for hybridizing K-representative and K-means clustering algorithms was effective to produce 99.72% reliability for creating clusters of mixed datasets.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

This work developed a hybrid model for mining large dataset of inpatient and outpatient treatments in tertiary health institution medicals for medical resources decision making using k-means and k-representative clustering algorithm. The application of data mining methods in the health sector is an interesting phenomenon. It sets to uncover the previously hidden data to meaningful information that could be used for both strategic decisions for resources optimization. The proposed hybrid model provided efficient algorithm for clustering datasets of patient's treatment for knowledge discovery. And it improved k-means clustering algorithm for optimal solution and efficient clustering of mixed dataset while the existing system the method of can only guarantee a locally optimal solution and as well as their result has no reliable indices or statistics that can be used with both k-modes and extended k-modes to ascertain the true k (number of clusters) in the datasets which includes the cluster can have convex shape and it is not efficient in clustering mixed datasets.

The produced hybrid clustering algorithm efficiently mine hospital resources for decision making. The accuracy (99.72%) of the hybrid algorithm shows that the instances of different classes were successfully separated by clusters. Also, the result of the misclassification matrix shows that the proposed hybrid clustering algorithm efficiently clustered the experimental datasets.

5.2 Contribution to Knowledge

This research has the following contribution to knowledge:

1. This work developed a hybrid clustering algorithm that optimizes the historical healthcare records at tertiary health institution medicals for efficient management of the limited healthcare resources.
2. The outcome of this work provided decision support system for healthcare management for tertiary health institution of students, university staff, and the community in general.

5.3 Recommendations

Applications of data mining techniques for the treatment of inpatient and outpatient’s resources allocation in tertiary health institution medicals were analyzed in this study to develop performance monitoring and evaluation tools.

It is recommended that hospital management system should implement the results of this work to guide the medical personnel provide the needed treatment for her patients at any point in time. Future work should apply other machine learning algorithms like supervised learning and optimization algorithms to model the limited hospital resource allocation.

REFERENCES

Aditya, K. P. (2020). A Simple Explanation of K-Means Clustering. Retrieved on 16th November, 2021 from https://www.analyticsvidhya.com/blog/2020/10/a-simple-explanation-of-k-means-clustering/

Ahirwar, D. K., Saxena, S. K., and Sisodia, M. S. (2020). Anomaly detection by naive Bayes & RBF network. *International Journal of Advanced Research in Computer Science and Electronics Engineering*, 1: 14-18.

Aravind, C. R. (2021). Exploring Clustering Algorithms: Explanation and Use Cases. Retrieved on 16 November, 2020 from https://neptune.ai/blog/clustering-algorithms.

Barbosa, E., Gonçalves, A., Guerra, M., & Cruz, C. (2020). A systematic review on resource allocation in public health. *European Journal of Public Health*, *30*(Supplement_5). https://doi.org/10.1093/eurpub/ckaa166.1309.

Dewangan, D., Kumar, M., and Qureshi, M. (2020). Power system transient stability analysis based on interval type-2 fuzzy logic controller and genetic algorithms, *International Journal of Innovative Science Engineering and Technology*, 1(4): 103-120.

Deepak, J, & Abhishek, K. (2021). Data Preprocessing in Data Mining, Retrieved on 17th November, 2021 from https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/?ref=lbp.

George, S. (2018). The 5 Clustering Algorithms Data Scientists Need to Know. Retrieved on 16th November, 2020 from https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68.

Gong, C., & Kang, H. (2023). Resource Allocation Efficiency of Urban Medical and Health Financial Expenditure Under the Background of Employees' Health. *Risk Management and Healthcare Policy*, *16*(June), 1059–1074. https://doi.org/10.2147/RMHP.S412514.

Gou, X., Xu, X., Deng, F., Zhou, W., & Herrera-Viedma, E. (2024). Medical health resources allocation evaluation in public health emergencies by an improved ORESTE method with linguistic preference orderings. *Fuzzy Optimization and Decision Making*, *23*(1), 1–27. https://doi.org/10.1007/s10700-023-09409-3.

Haesik, K. (2020). Performance analysis of k-means clustering algorithms for mMTC systems, *IEEE Institute of Electrical and Electronic Engineers*.30-35.

Jae, H. & Sang, Chan. (2021). k-representatives algorithm: a clustering algorithm with learning distance measure for categorical values, Korea Advanced Institute of Science and Technology. 1-6.

Kanjanawattana, M. & Anderberg, R. (2020). *Cluster Analysis for Application*. Academic Press, 2020.

Kolhe, G. V., Prasad, N., Hanumantha, R., Depa, P. & Alekhya, B. N. (2020). Unsupervised Learning Algorithms to Identify the Dense Cluster in Large Datasets, *International Journal of Computer Science and Telecommunications*, 2(4) , 345-353

Kristina, P. S. & Miin-shen, Y. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*. 8: 80716 – 80727.

Mohanavalli, S. & Jaisakthi, S. M. (2015). Precise distance metric for mixed data clustering using chi-square statistics. Research Journal of Applied Sciences, 10(12): 1441-1444. https://doi.org/10.19026/rjaset.10.1846

Morissette, L. & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematical. *Tutorials in Quantitative Methods for Psychology*, 9(1): 15-24.

Mehmood, R.; Zhang, G.; Bie, R.; Dawood, H.; Ahmad, H.(2021) Clustering by fast search and find of density peaks via heat diffusion. *Neurocomputing* , *208*, 210–217

Mukherjee, M. (2016). Object-Oriented Analysis and Design. *International journal of Advanced Engineering and Management,* 1: 18-24.

Nanopoulos, A., & Yannis, M. (2001). C2P: Clustering based on Closest Pairs. Retrieved on 29th October, 2019. From http://citeseerx.ist.psu.edu/showciting;jsessionid=2A8586B1F54B5BC067B3C8994A558275?cid=450412.

Opara. C. C., Eze, U. F. & Oleji, C. P. (2020). Hybrid Data Mining Model for Knowledge Discovery on Students Academic Performance. *International Journal of Research Publications,* 47(1): 1-12.

Oyelade, O. J., Oladipupo, O. O. & Obagbuwa, I. C. (2010). Applications of K-means clustering algorithm for prediction of Students Academic Performance. *International Journal of Computer Science and Information Security*, 7(1): 292-295.

Pinisetty, V. N. P., Valaboju, R. & Rao, N. R. (2015). Hybrid Algorithm for Clustering Mixed Data Sets. *IOSR Journal of Computer Engineering*, 6(2): 09-13.

Prasad, M. V. S. & Raju, N. O. (2020). Extended K-mean clustering technique using map reduced segmentation clustering for big data. *International Journal of Research and Analytical Reviews (IJRAR),* 7(2): 946-952.

Ramakrishnan, M. & Jayaraj, T. D. (2014). Modified k-means algorithm for effective clustering of categorical data sets. *International Journal of Computer Applications*, 89(7): 0975 – 8887. https://doi.org/10.5120/15518-4102.

Ramesh, V., Ramar, K., & Babu, S. (2013). Parallel k-means algorithm on agricultural database. IJCS *International Journal of Computer Science*, 10(1): 1694-0784.

Saini, A.; Saraswat, S.; Faujdar, N. (2022). Clustering Based Stock Market Analysis. *Int. J. Control. Theory Appl.*, *10*

Sambantham, V. (2023). Some Clustering Methods , Algorithms and their Applications. *International Journal of Recent and Innovation Trends in Computing and Communication*, *11*(6s), 401–410. https://doi.org/10.17762/ijritcc.v11i6s.6946

Srya, P. (2021). Clustering in Machine Learning. Retrieved on 17th November, 2020 from https://www.geeksforgeeks.org/clustering-in-machine-learning/?ref=lbp.

Toan, N. M. & Van-Nam, H. (2021). An LSH-based -representatives clustering method for large categorical data, Elsevier. 6, 29-40. https://doi.org/10.1016/j.neucom.2021.08.050

Wei, M., Chow, T. W. S. &. Chan, R. H. M. (2015). Clustering heterogeneous data with k-means by mutual information-based unsupervised feature transformation. Entropy, 17, 1535-1548.

Xiaoliang, Z., Yulin, He., Yi, J. Honglian, Q., Muhammad, A., & Joshua, Z. H. (2020). A Robust k-means clustering algorithm based on observation point mechanism.1-5. https://doi.org/10.1155/2020/3650926.

Yin, H., Aryani, A., & Petrie, S. (2024). A Rapid Review of Clustering Algorithms. *ArXiv*, 1–25.

Yiu, M. C. & Hong, J. (2018). Categorical and numerical data clustering based on a unified similarity metric without knowing cluster number in Pattern Recognition, *Elsevier*, 46(2013): 2228–2238.

Zhang, C., & Fang, Z. (2019). An Improved K-means Clustering Algorithm. *Journal of Information & Computational Science*, 10(1): 193-199.

APPENDIX A

Source code

```
/*---------------Centroid.java-----------------*/
package org.c4s.algorithm.cluster;
/**s
 * This class represents the Centroid for a Cluster. The initial centroid is calculated
 * using a equation which divides the sample space for each dimension into equal parts
 *depending upon the value of k.
  */
class Centroid {
private double mCx, mCy;
private Cluster mCluster;
public Centroid(double cx, double cy) {
this.mCx = cx;
this.mCy = cy;
   }
public void calcCentroid() { //only called by CAInstance
intnumDP = mCluster.getNumDataPoints();
doubletempX = 0, tempY = 0;
int i;
        //caluclating the new Centroid
for (i = 0; i<numDP; i++) {
tempX = tempX + mCluster.getDataPoint(i).getX();
            //total for x
tempY = tempY + mCluster.getDataPoint(i).getY();
            //total for y
        }
this.mCx = tempX / numDP;
this.mCy = tempY / numDP;
        //calculating the new Euclidean Distance for each Data Point
tempX = 0;
```

```
tempY = 0;
for (i = 0; i<numDP; i++) {
mCluster.getDataPoint(i).calcEuclideanDistance();
        }
        //calculate the new Sum of Squares for the Cluster
mCluster.calcSumOfSquares();
    }

public void setCluster(Cluster c) {
this.mCluster = c;
    }
public double getCx() {
returnmCx;
    }
public double getCy() {
returnmCy;
    }
public Cluster getCluster() {
returnmCluster;
    }
}
package org.c4s.algorithm.cluster;
importjava.util.Vector;
/**
This class is the entry point for constructing Cluster Analysis objects.
Each instance of JCA object is associated with one or more clusters,
and a Vector of DataPoint objects. The JCA and DataPoint classes are
the only classes available from other packages.
@see DataPoint
**/
public class JCA {
```

```
private Cluster[] clusters;
privateint miter;
private Vector mDataPoints = new Vector();
private double mSWCSS;
public JCA(int k, intiter, Vector dataPoints) {
clusters = new Cluster[k];
for (int i = 0; i< k; i++) {
clusters[i] = new Cluster("Cluster" + i);
        }
this.miter = iter;
this.mDataPoints = dataPoints;
   }
private void calcSWCSS() {
double temp = 0;
for (int i = 0; i<clusters.length; i++) {
temp = temp + clusters[i].getSumSqr();
        }
mSWCSS = temp;
   }
public void startAnalysis() {
        //set Starting centroid positions - Start of Step 1
setInitialCentroids();
int n = 0;
        //assign DataPoint to clusters
        loop1: while (true) {
for (int l = 0; l <clusters.length; l++)
            {
clusters[l].addDataPoint((DataPoint)mDataPoints.elementAt(n));
n++;
if (n >= mDataPoints.size())
break loop1;
```

```
        }
      }
      //calculate E for all the clusters
calcSWCSS();

      //recalculate Cluster centroids - Start of Step 2
for (int i = 0; i<clusters.length; i++) {
clusters[i].getCentroid().calcCentroid();
      }
      //recalculate E for all the clusters
calcSWCSS();
for (int i = 0; i< miter; i++) {
        //enter the loop for cluster 1
for (int j = 0; j <clusters.length; j++) {
for (int k = 0; k < clusters[j].getNumDataPoints(); k++) {
            //pick the first element of the first cluster
            //get the current Euclidean distance
doubletempEuDt = clusters[j].getDataPoint(k).getCurrentEuDt();
            Cluster tempCluster = null;
booleanmatchFoundFlag = false;
            //call testEuclidean distance for all clusters
for (int l = 0; l <clusters.length; l++) {
            //if testEuclidean<currentEuclidean then
if (tempEuDt> clusters[j].getDataPoint(k).testEuclideanDistance(clusters[l].getCentroid())) {
tempEuDt = clusters[j].getDataPoint(k).testEuclideanDistance(clusters[l].getCentroid());
tempCluster = clusters[l];
matchFoundFlag = true;
              }
              //if statement - Check whether the Last EuDt is > Present EuDt
              }
//for variable 'l' - Looping between different Clusters for matching a Data Point.
```

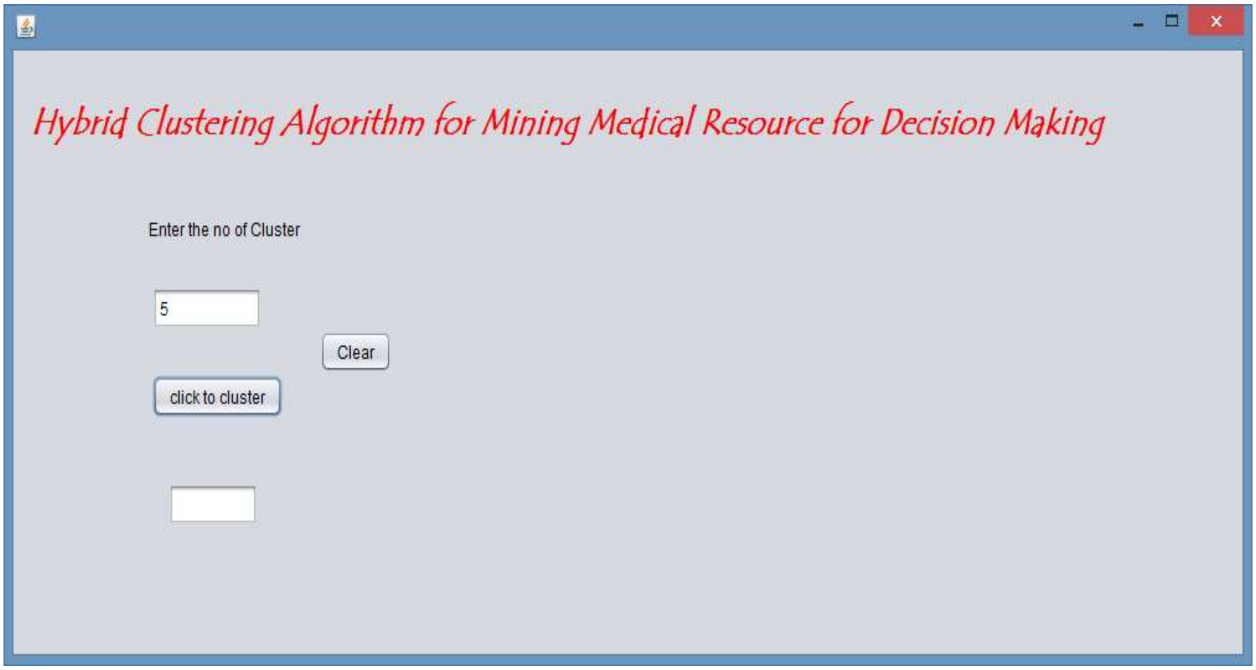
```
//add DataPoint to the cluster and calcSWCSS
if (matchFoundFlag) {
                tempCluster.addDataPoint(clusters[j].getDataPoint(k));
                clusters[j].removeDataPoint(clusters[j].getDataPoint(k));
for (int m = 0; m <clusters.length; m++) {
clusters[m].getCentroid().calcCentroid();
                }
//for variable 'm' - Recalculating centroids for all Clusters
calcSWCSS();
              }
//if statement - A Data Point is eligible for transfer between Clusters.
           }
//for variable 'k' - Looping through all Data Points of the current Cluster.
}//for variable 'j' - Looping through all the Clusters.
}//for variable 'i' - Number of iterations.
   }
public Vector[] getClusterOutput() {
      Vector v[] = new Vector[clusters.length];
for (int i = 0; i<clusters.length; i++) {
v[i] = clusters[i].getDataPoints();
      }
return v;
}    private void setInitialCentroids() {
      //kn = (round((max-min)/k)*n)+min where n is from 0 to (k-1).
double cx = 0, cy = 0;
for (int n = 1; n <= clusters.length; n++) {
cx = (((getMaxXValue() - getMinXValue()) / (clusters.length + 1)) * n) + getMinXValue();
cy = (((getMaxYValue() - getMinYValue()) / (clusters.length + 1)) * n) + getMinYValue();
        Centroid c1 = new Centroid(cx, cy);
clusters[n - 1].setCentroid(c1);
c1.setCluster(clusters[n - 1]);
```

```
        }
    }

private double getMaxXValue() {
double temp;
temp = ((DataPoint) mDataPoints.elementAt(0)).getX();
for (int i = 0; i<mDataPoints.size(); i++) {
DataPointdp = (DataPoint) mDataPoints.elementAt(i);
temp = (dp.getX() > temp) ? dp.getX() : temp;
        }
return temp;
    }
private double getMinXValue() {
double temp = 0;
temp = ((DataPoint) mDataPoints.elementAt(0)).getX();
for (int i = 0; i<mDataPoints.size(); i++) {
DataPointdp = (DataPoint) mDataPoints.elementAt(i);
temp = (dp.getX() < temp) ? dp.getX() : temp;
        }
return temp;
    }

private double getMaxYValue() {
double temp = 0;
temp = ((DataPoint) mDataPoints.elementAt(0)).getY();
for (int i = 0; i<mDataPoints.size(); i++) {
DataPointdp = (DataPoint) mDataPoints.elementAt(i);
temp = (dp.getY() > temp) ? dp.getY() : temp;
        }
return temp;
    }
```

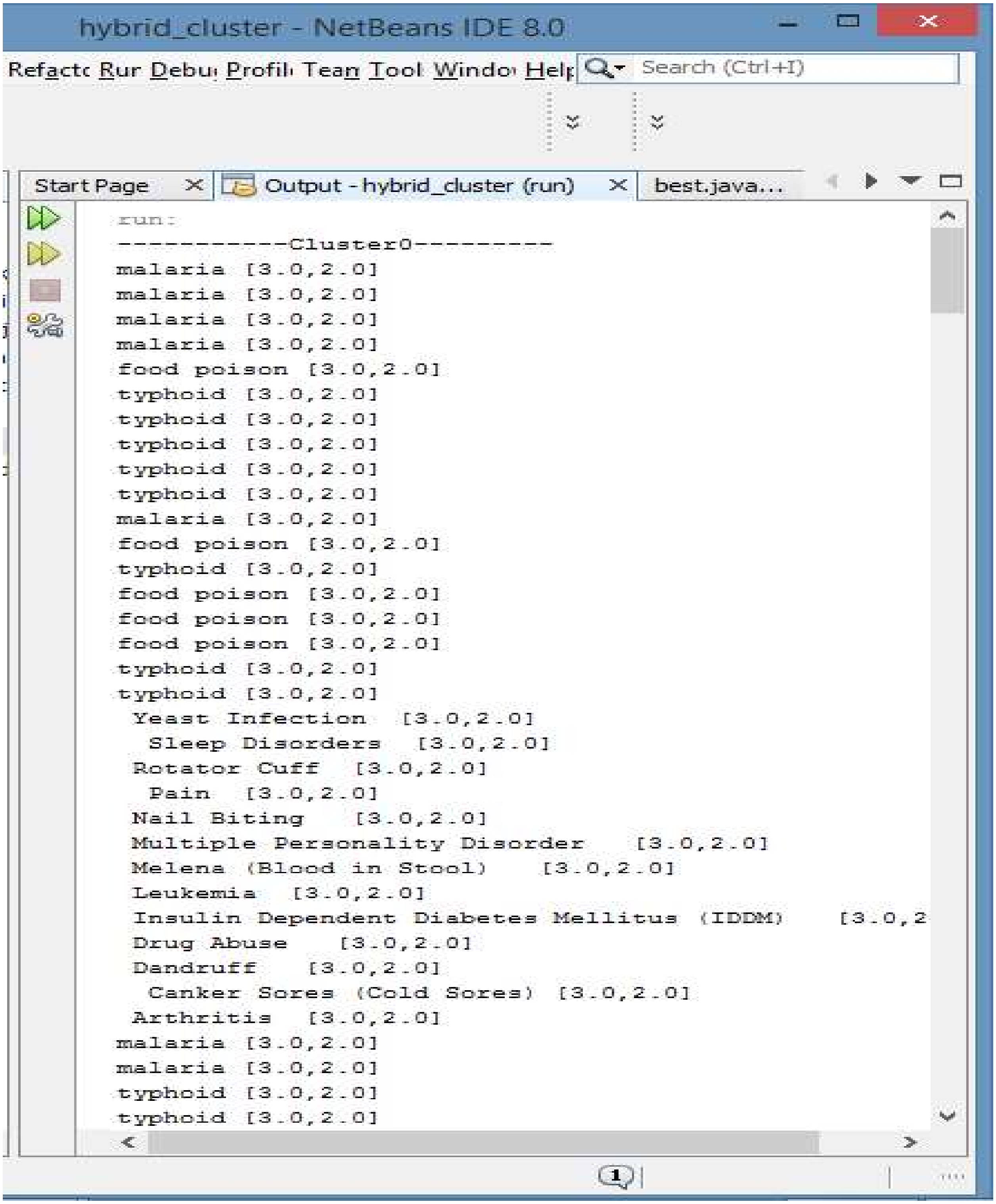
```
private double getMinYValue() {
double temp = 0;
temp = ((DataPoint) mDataPoints.elementAt(0)).getY();
for (int i = 0; i<mDataPoints.size(); i++) {
DataPointdp = (DataPoint) mDataPoints.elementAt(i);
temp = (dp.getY() < temp) ? dp.getY() : temp;
        }
return temp;
    }

publicintgetKValue() {
returnclusters.length;
    }

publicintgetIterations() {
return miter;
    }

publicintgetTotalDataPoints() {
returnmDataPoints.size();
    }

public double getSWCSS() {
returnmSWCSS;
    }
public Cluster getCluster(intpos) {
return clusters[pos];
    }
}
```

APPENDIX B:

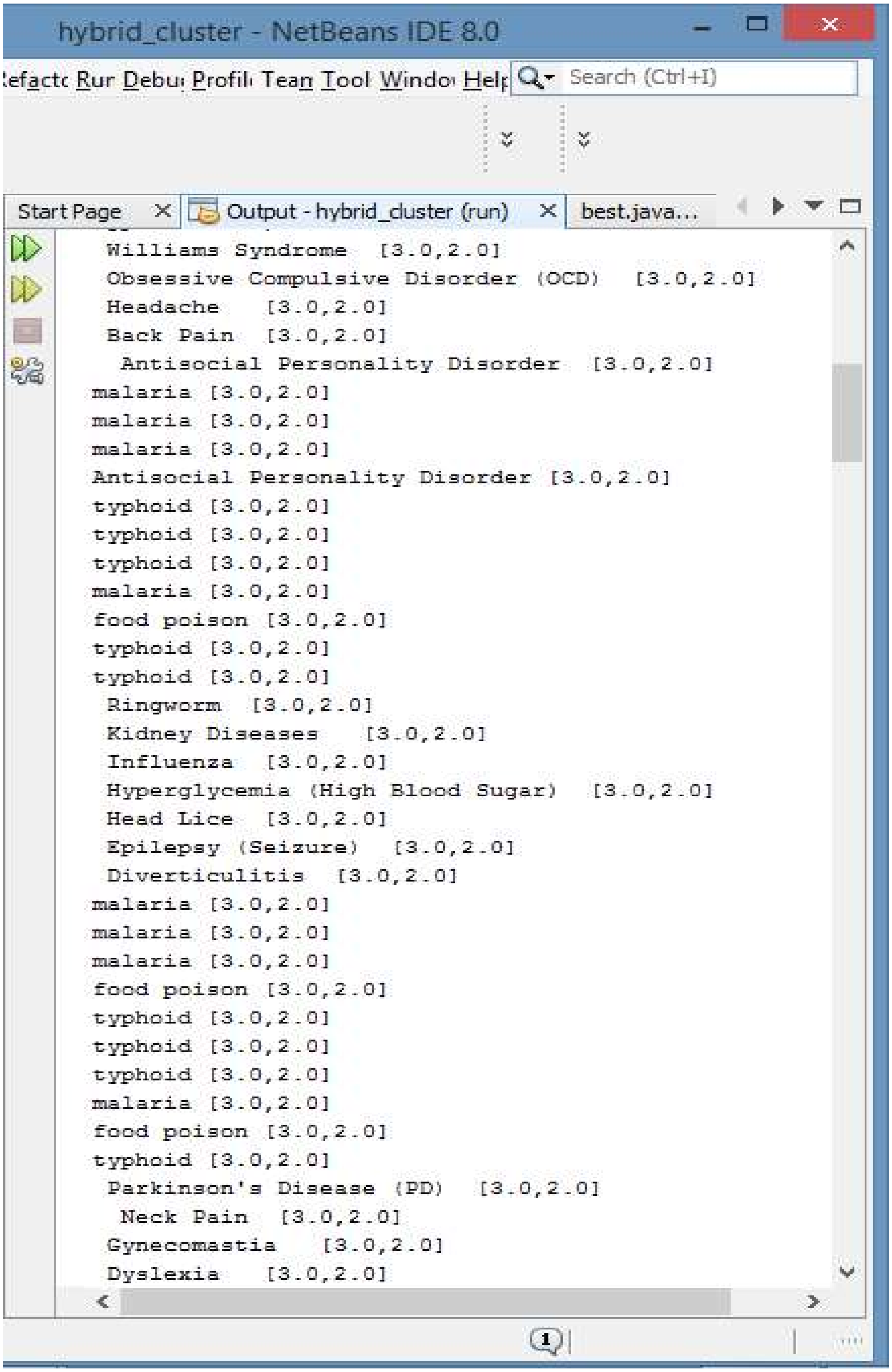
SAMPLE OUTPUTS



User interface for Mining medical record.



Clustering results



Clustering results

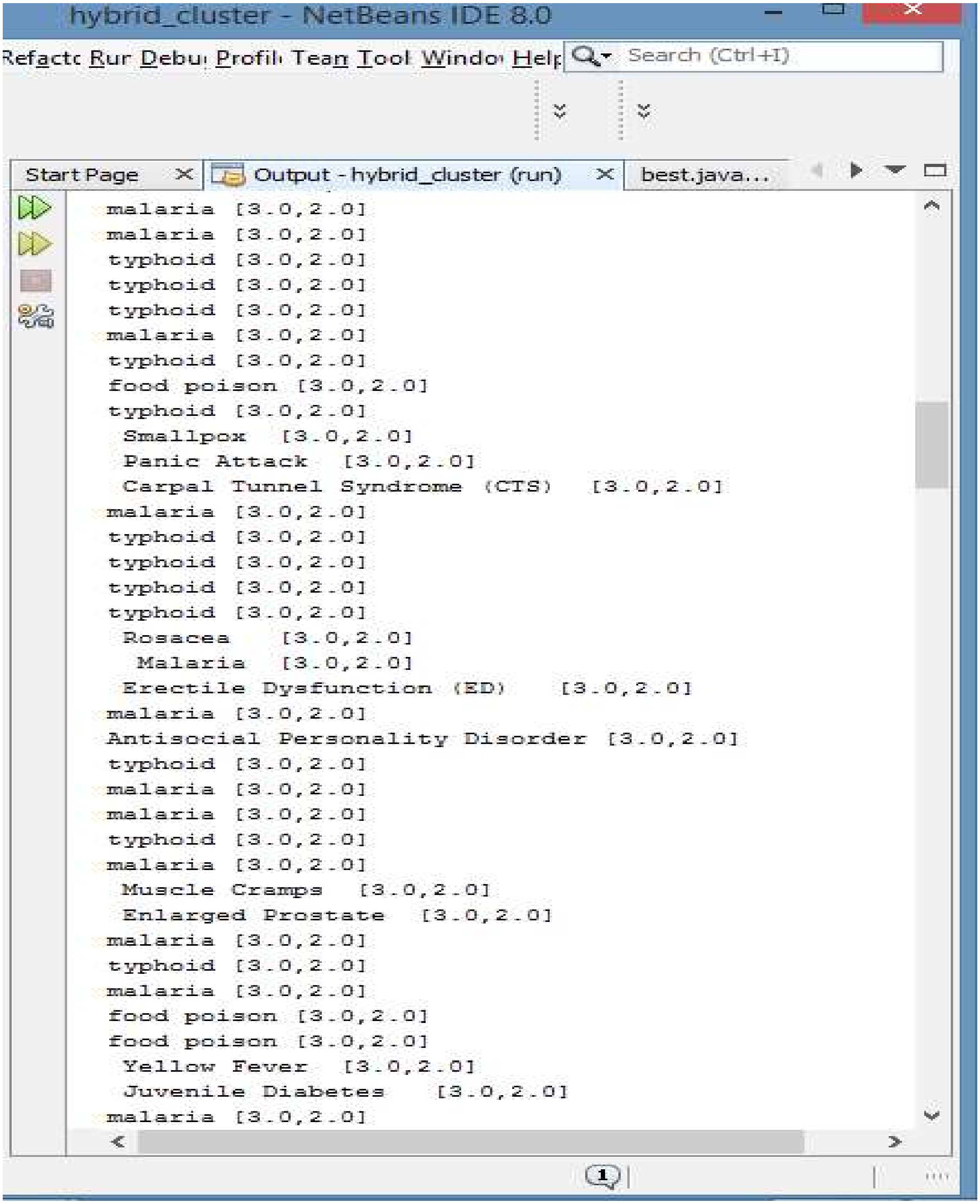
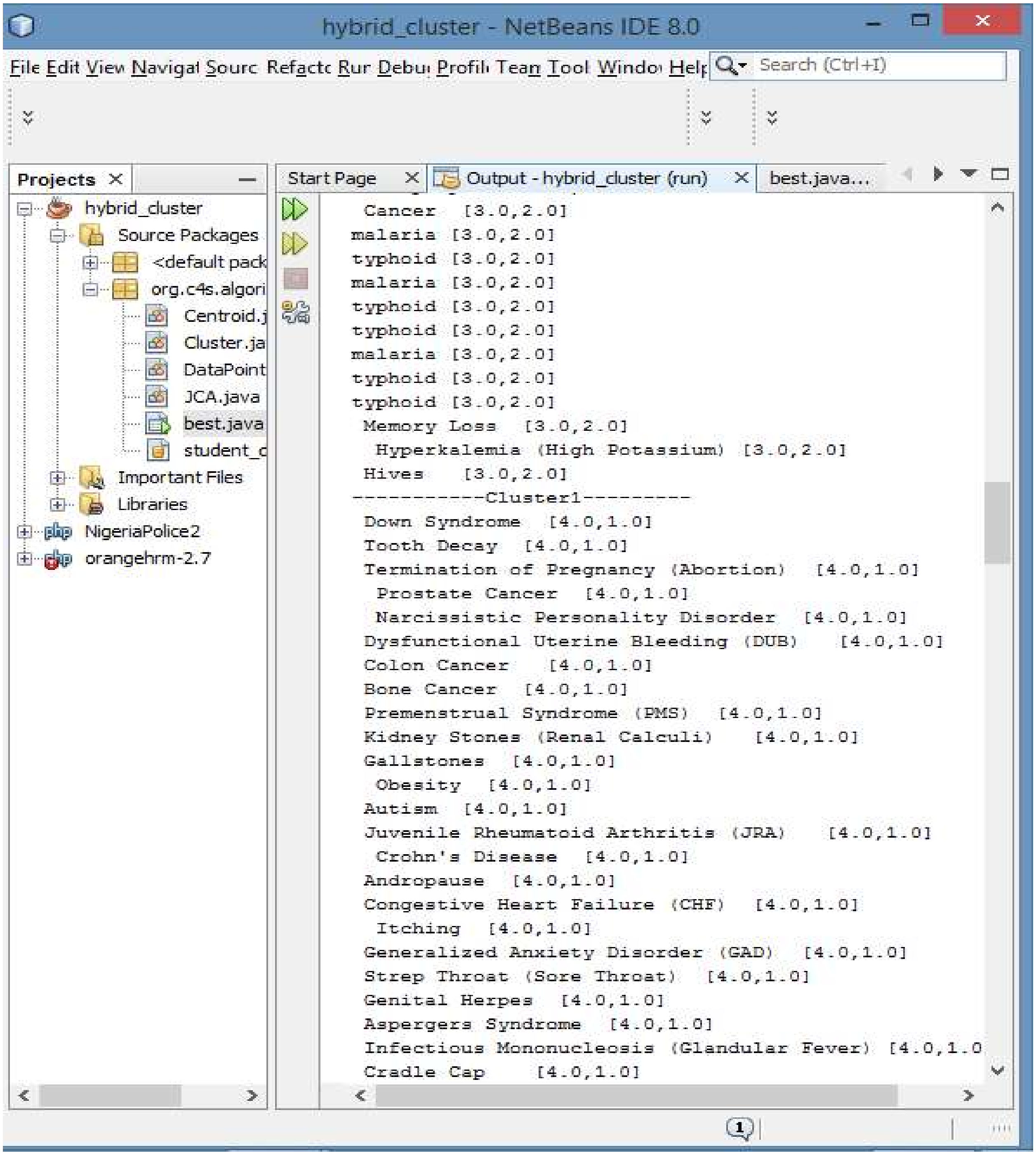
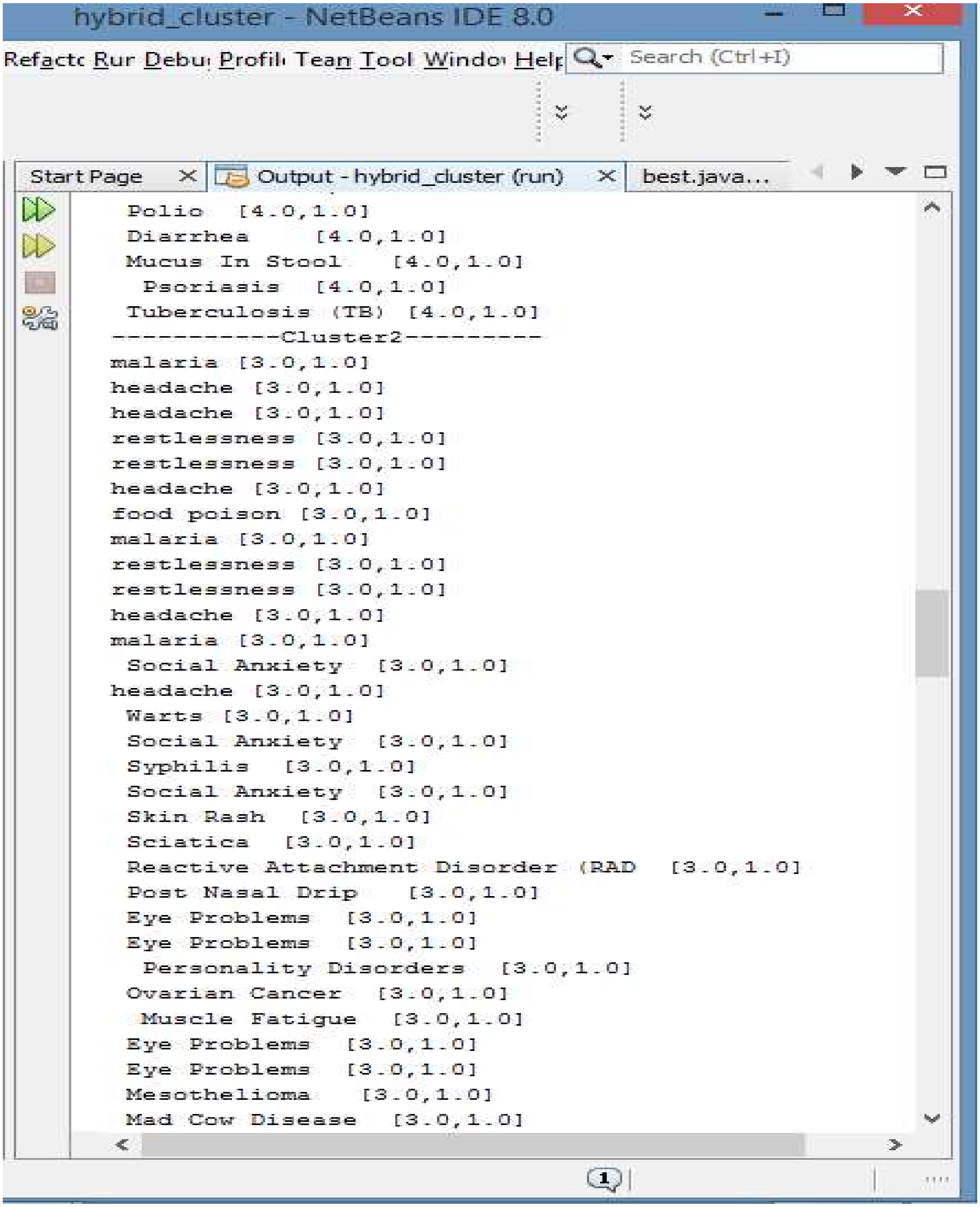


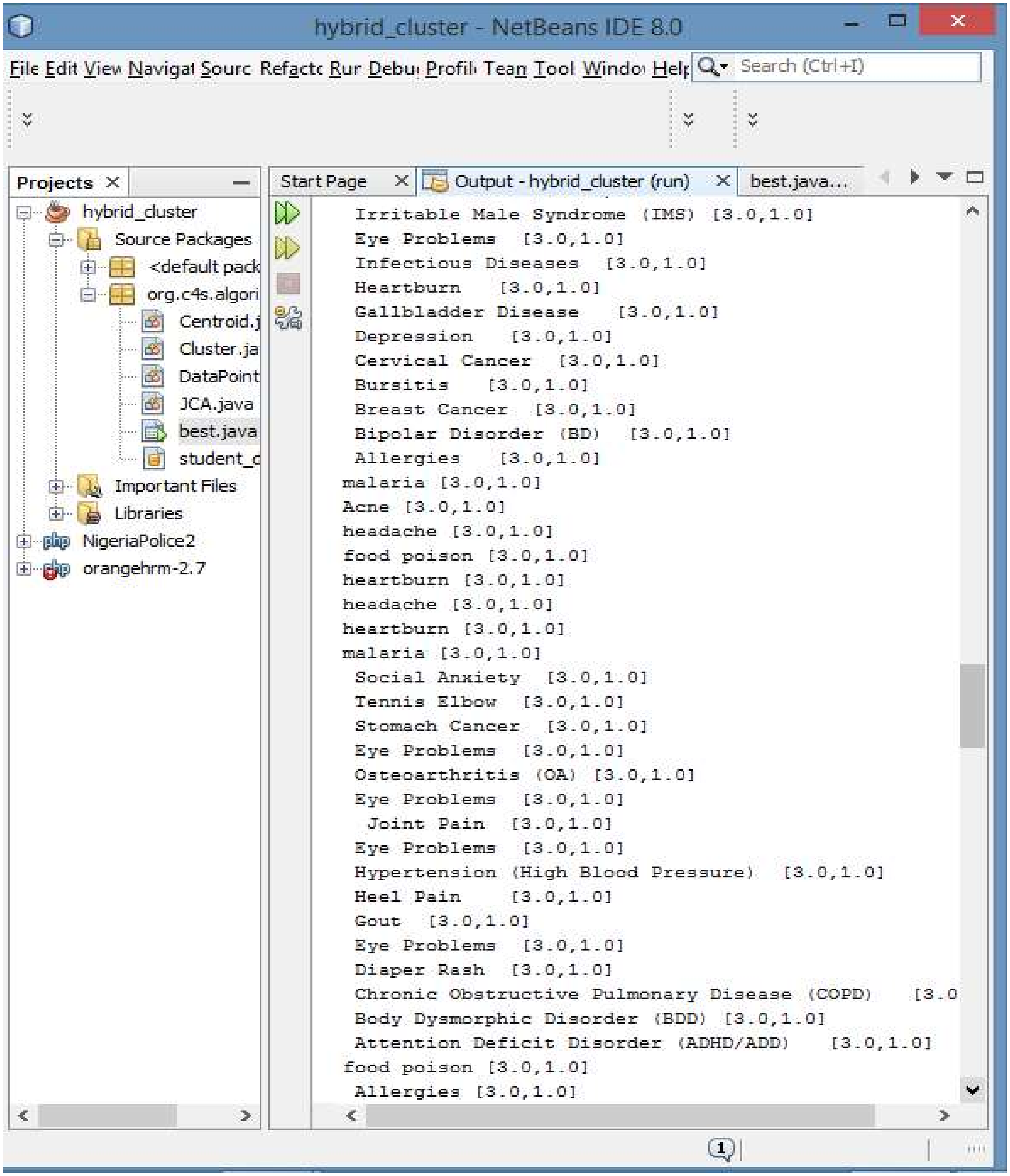
Figure 5.2c: clustering results



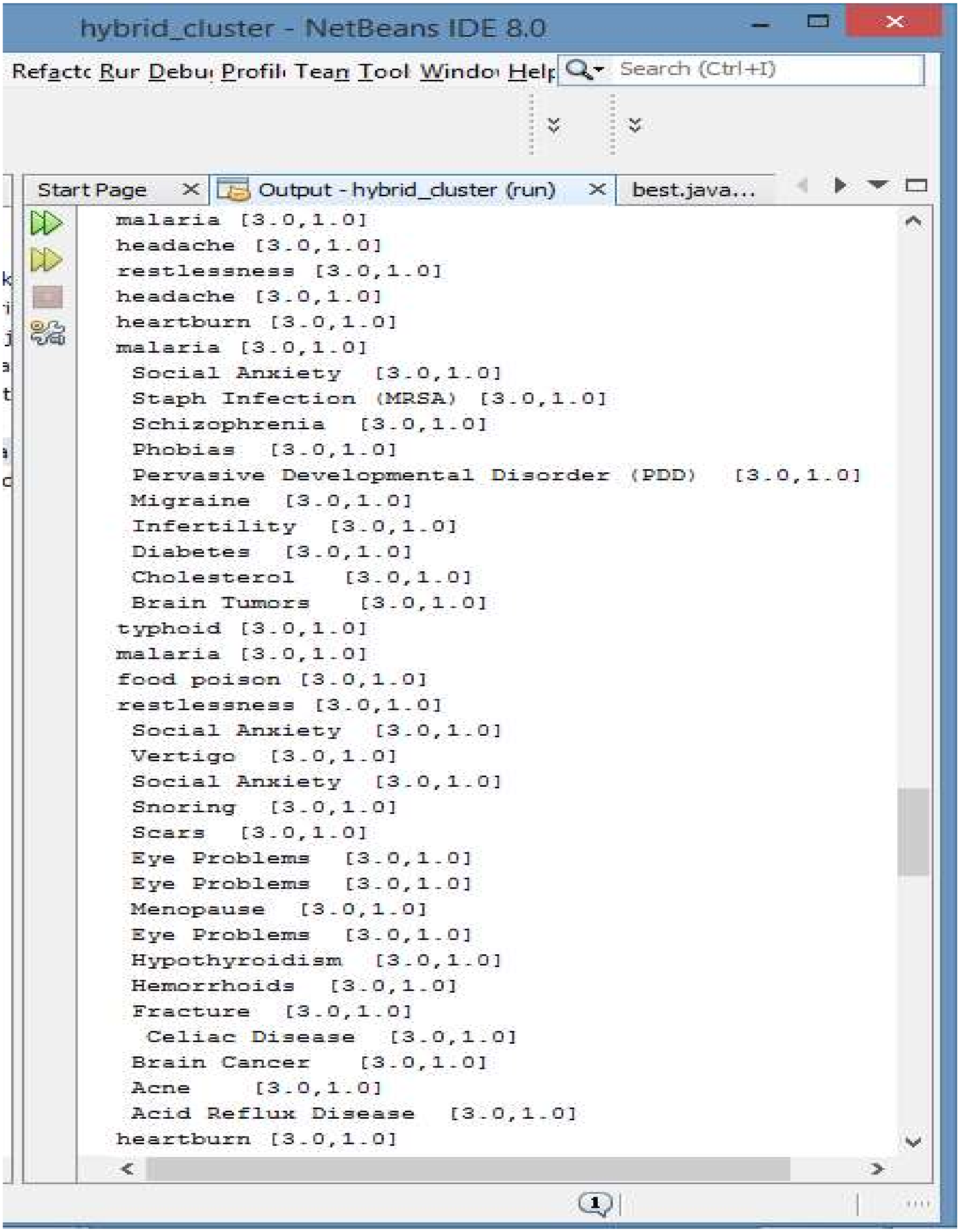
Clustering results



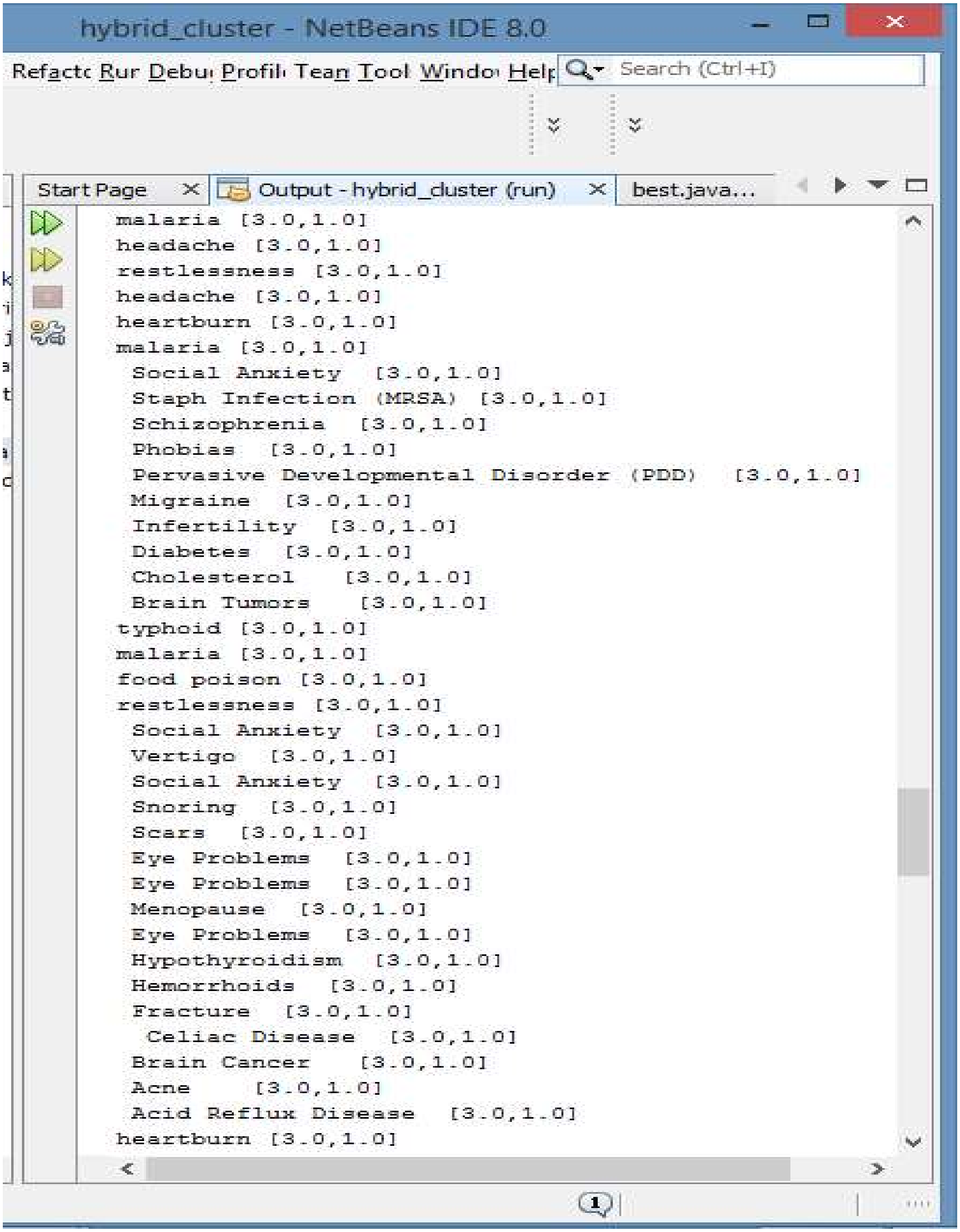
Clustering results



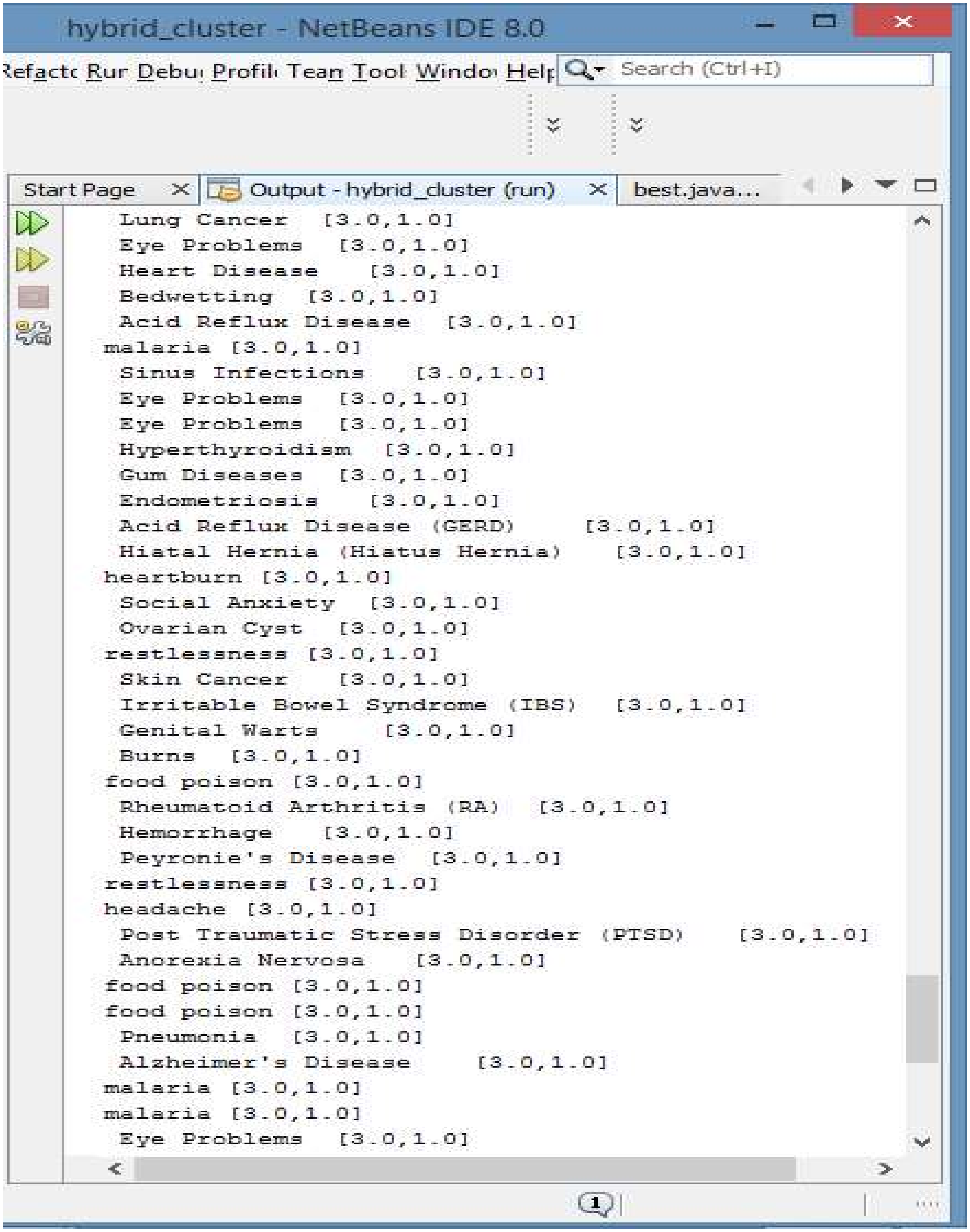
Clustering results



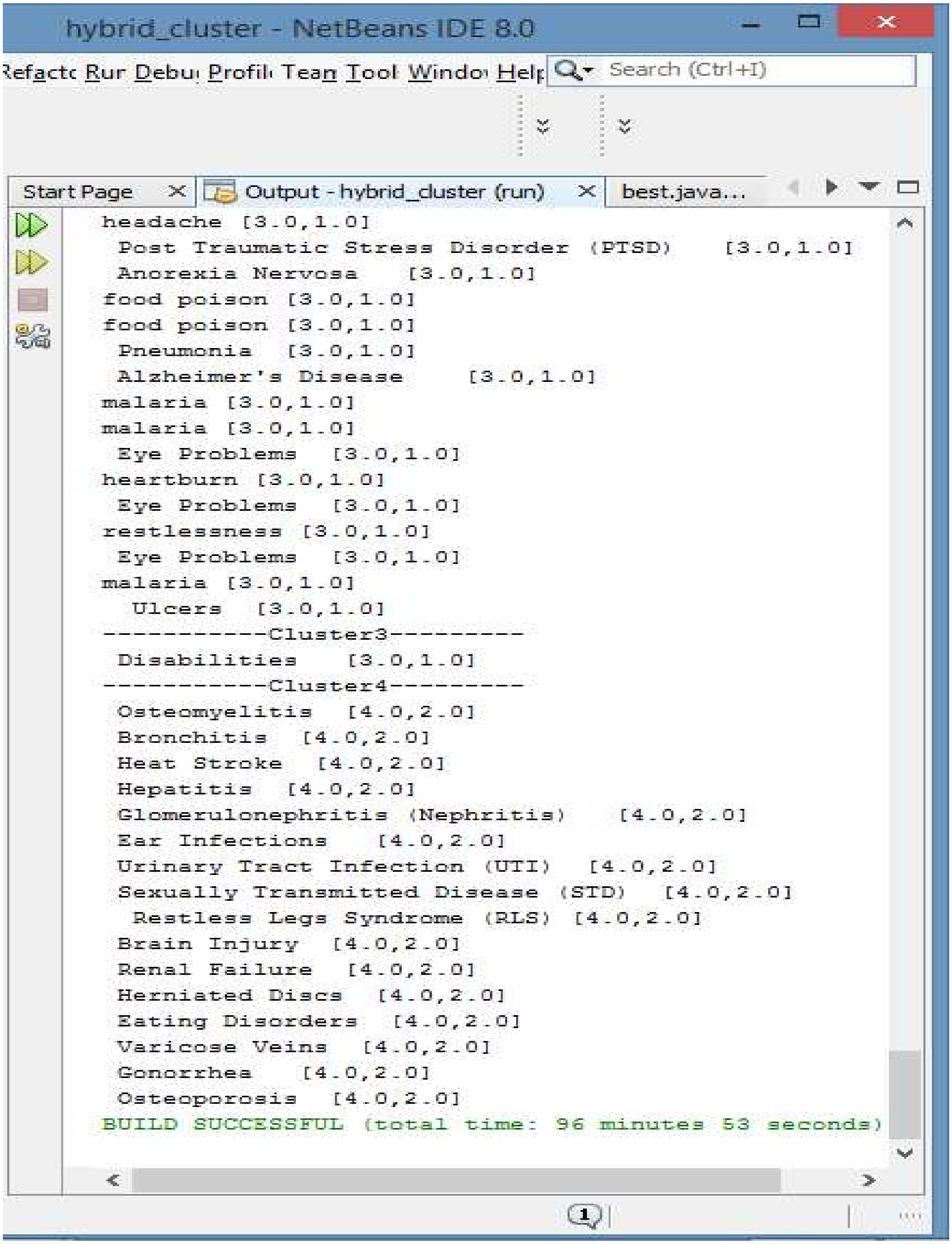
Clustering results



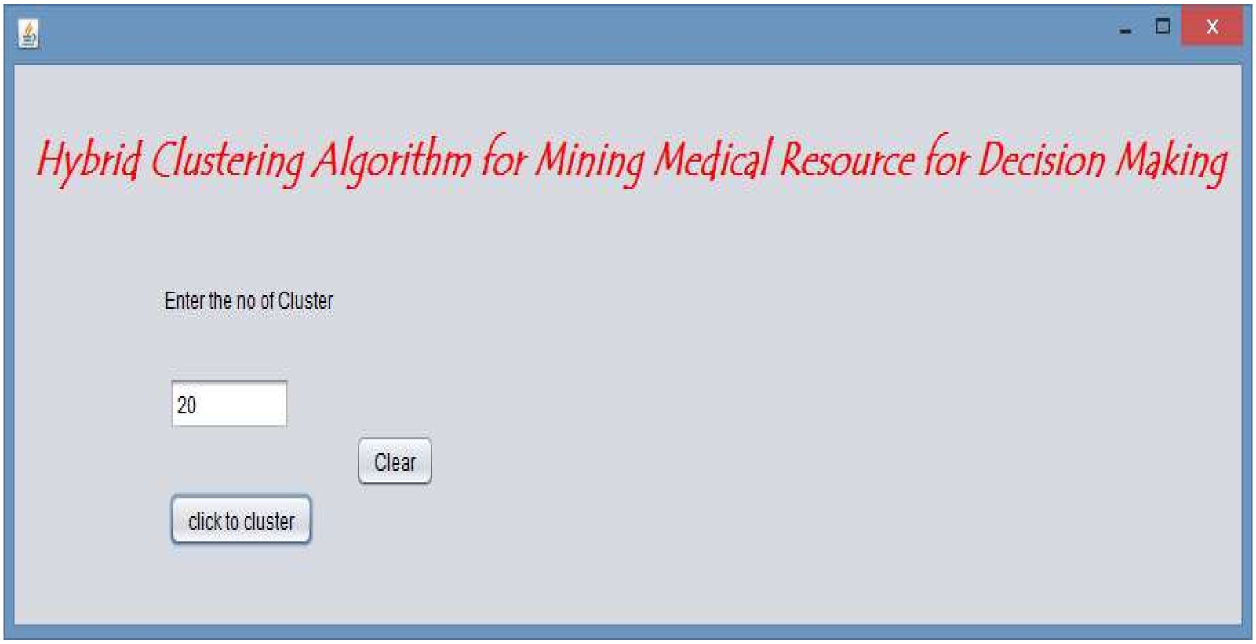
Clustering results

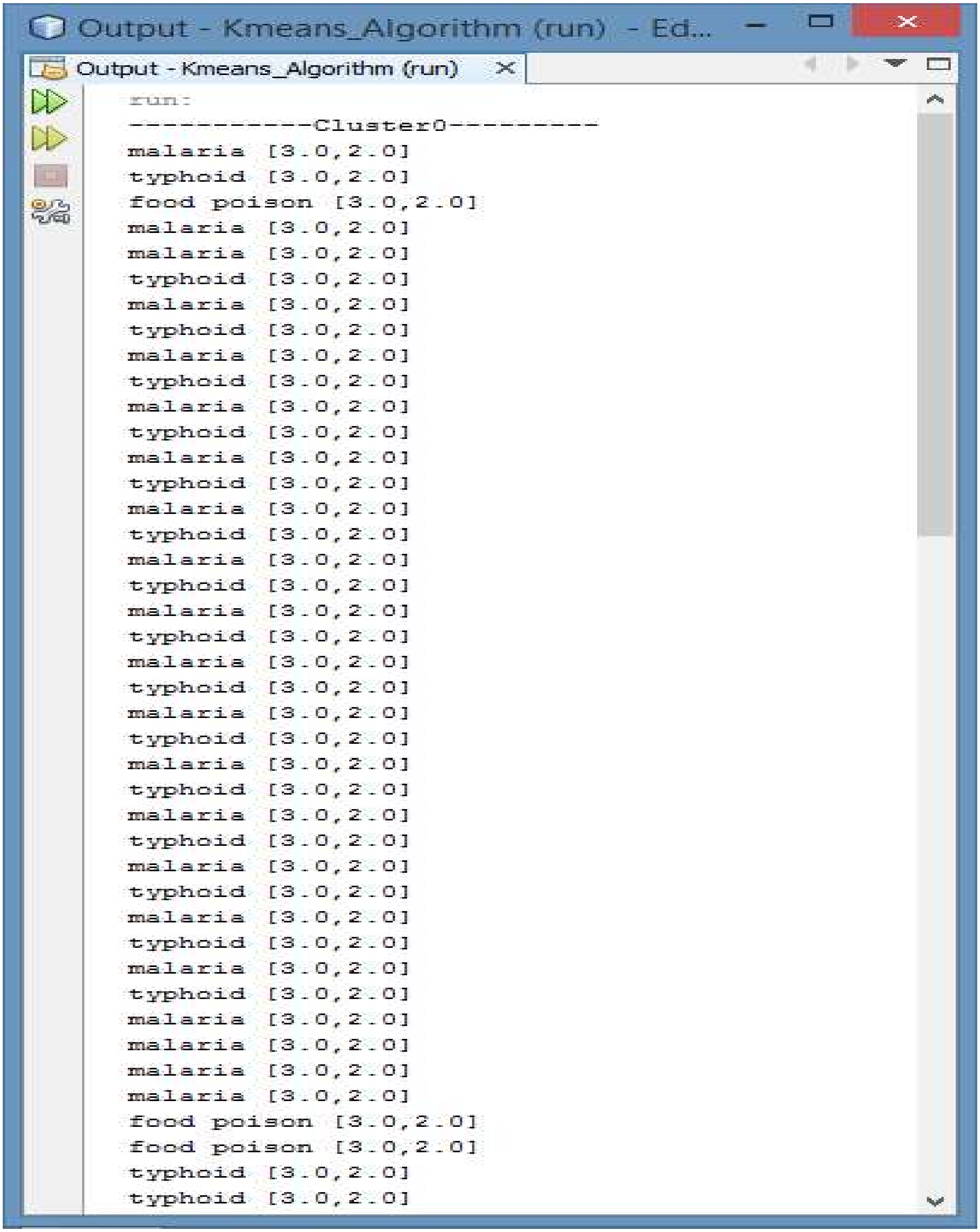


Clustering results

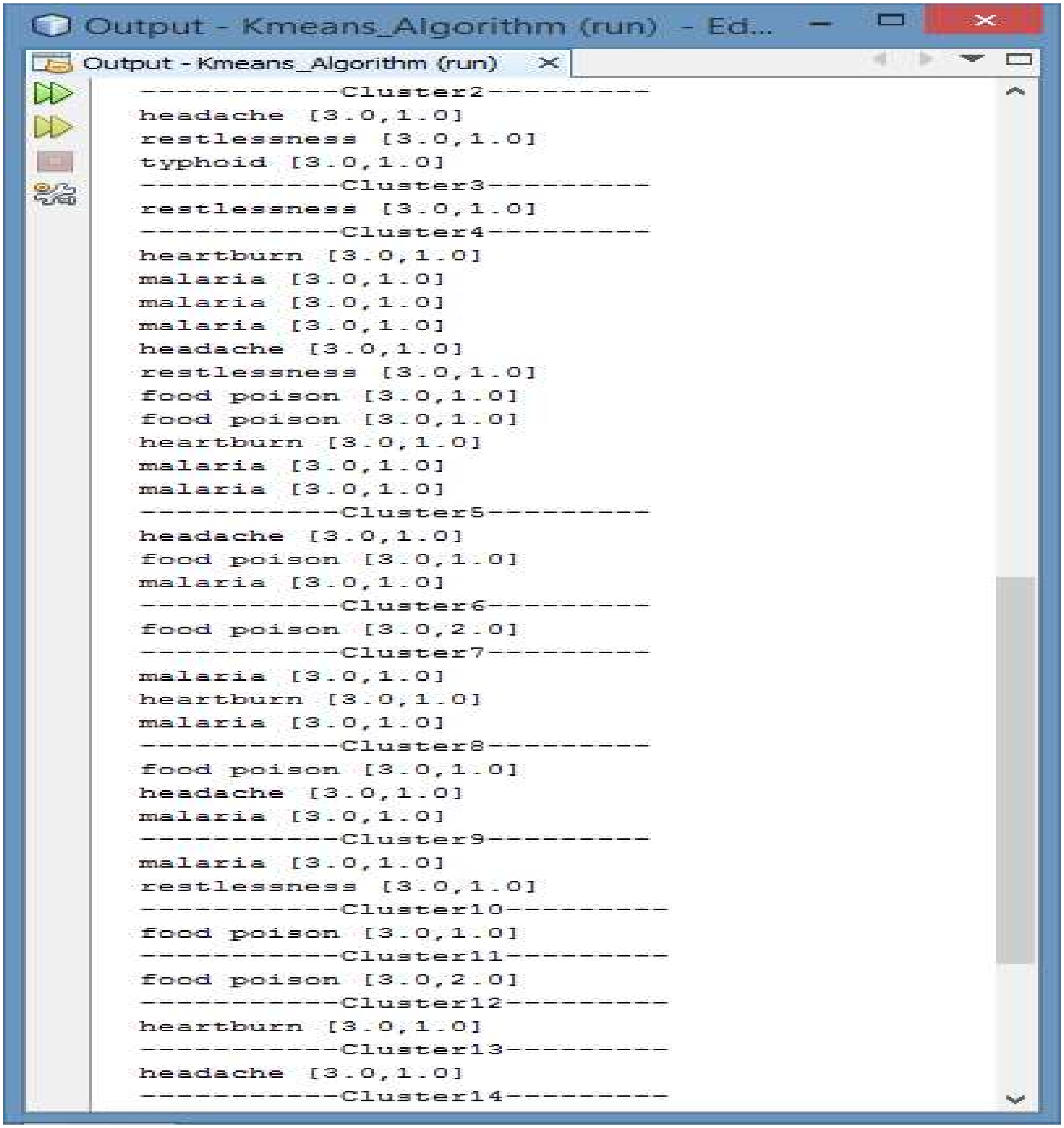


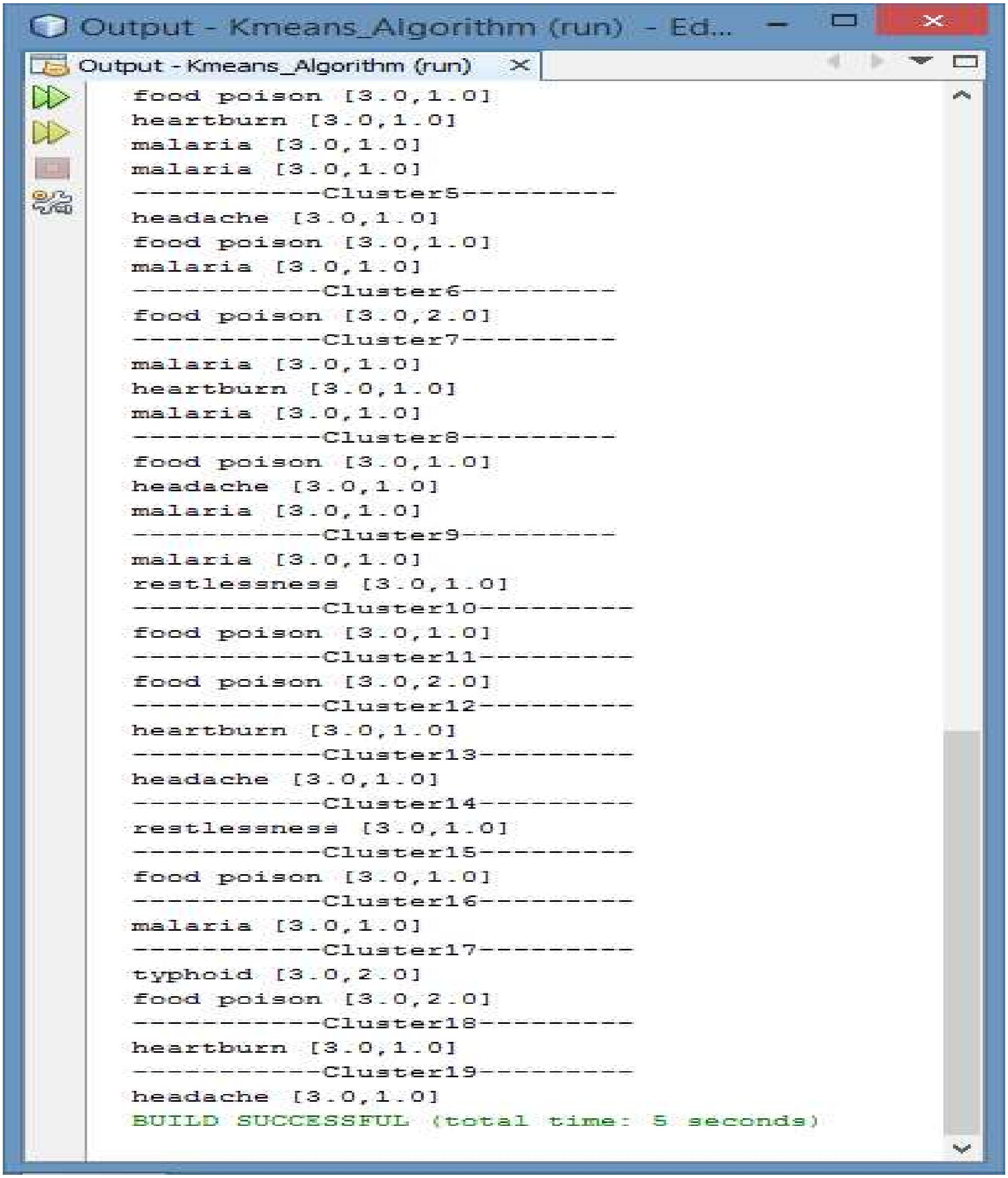
Clustering results


Hybrid Clustering Algorithm for Mining Medical Resource for Decision Making
Enter the no of Cluster
20
Clear
click to cluster

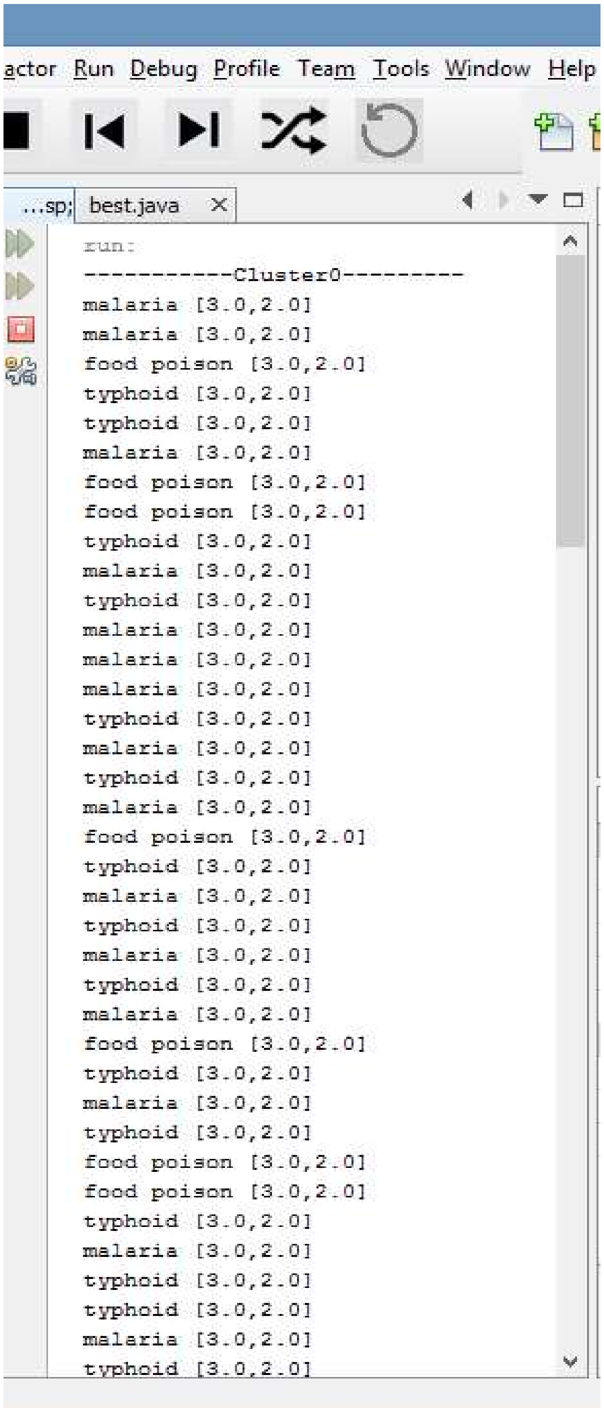


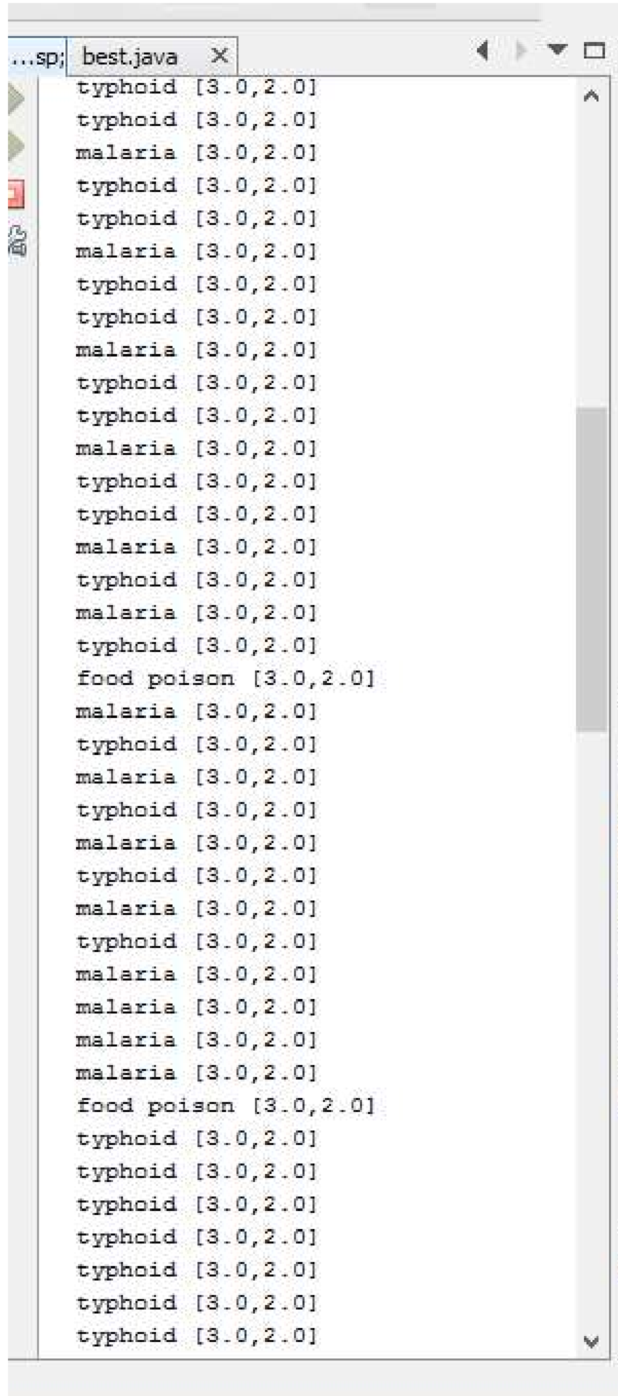
Output - Kmeans_Algorithm (run) - Ed...
Output - Kmeans_Algorithm (run)
run:
-----------Cluster0---------
malaria [3.0,2.0]
typhoid [3.0,2.0]
food poison [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
food poison [3.0,2.0]
food poison [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]

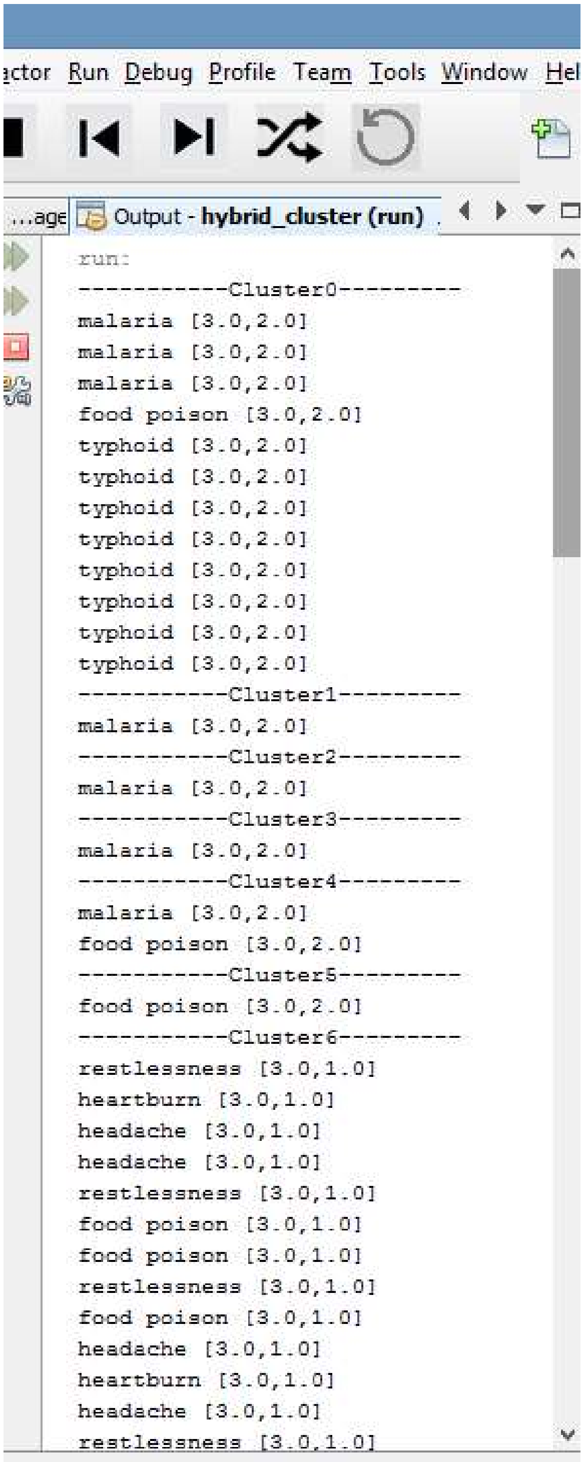

Output - Kmeans_Algorithm (run) - Ed...
Output - Kmeans_Algorithm (run)
-----------Cluster2---------
headache [3.0,1.0]
restlessness [3.0,1.0]
typhoid [3.0,1.0]
-----------Cluster3---------
restlessness [3.0,1.0]
-----------Cluster4---------
heartburn [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
headache [3.0,1.0]
restlessness [3.0,1.0]
food poison [3.0,1.0]
food poison [3.0,1.0]
heartburn [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
-----------Cluster5---------
headache [3.0,1.0]
food poison [3.0,1.0]
malaria [3.0,1.0]
-----------Cluster6---------
food poison [3.0,2.0]
-----------Cluster7---------
malaria [3.0,1.0]
heartburn [3.0,1.0]
malaria [3.0,1.0]
-----------Cluster8---------
food poison [3.0,1.0]
headache [3.0,1.0]
malaria [3.0,1.0]
-----------Cluster9---------
malaria [3.0,1.0]
restlessness [3.0,1.0]
-----------Cluster10---------
food poison [3.0,1.0]
-----------Cluster11---------
food poison [3.0,2.0]
-----------Cluster12---------
heartburn [3.0,1.0]
-----------Cluster13---------
headache [3.0,1.0]
-----------Cluster14---------

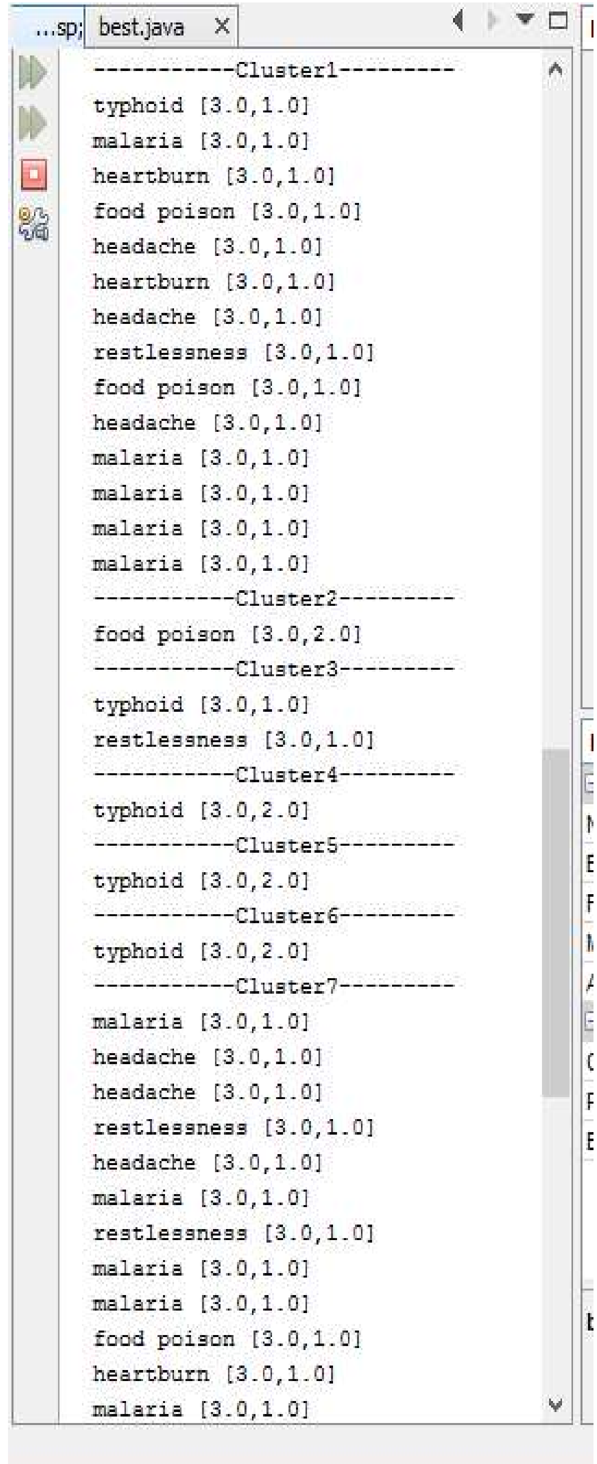

Output - Kmeans_Algorithm (run) - Ed...
Output - Kmeans_Algorithm (run)
food poison [3.0,1.0]
heartburn [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
-----------Cluster5---------
headache [3.0,1.0]
food poison [3.0,1.0]
malaria [3.0,1.0]
-----------Cluster6---------
food poison [3.0,2.0]
-----------Cluster7---------
malaria [3.0,1.0]
heartburn [3.0,1.0]
malaria [3.0,1.0]
-----------Cluster8---------
food poison [3.0,1.0]
headache [3.0,1.0]
malaria [3.0,1.0]
-----------Cluster9---------
malaria [3.0,1.0]
restlessness [3.0,1.0]
-----------Cluster10---------
food poison [3.0,1.0]
-----------Cluster11---------
food poison [3.0,2.0]
-----------Cluster12---------
heartburn [3.0,1.0]
-----------Cluster13---------
headache [3.0,1.0]
-----------Cluster14---------
restlessness [3.0,1.0]
-----------Cluster15---------
food poison [3.0,1.0]
-----------Cluster16---------
malaria [3.0,1.0]
-----------Cluster17---------
typhoid [3.0,2.0]
food poison [3.0,2.0]
-----------Cluster18---------
heartburn [3.0,1.0]
-----------Cluster19---------
headache [3.0,1.0]
BUILD SUCCESSFUL (total time: 5 seconds)

Results of 10 clusters





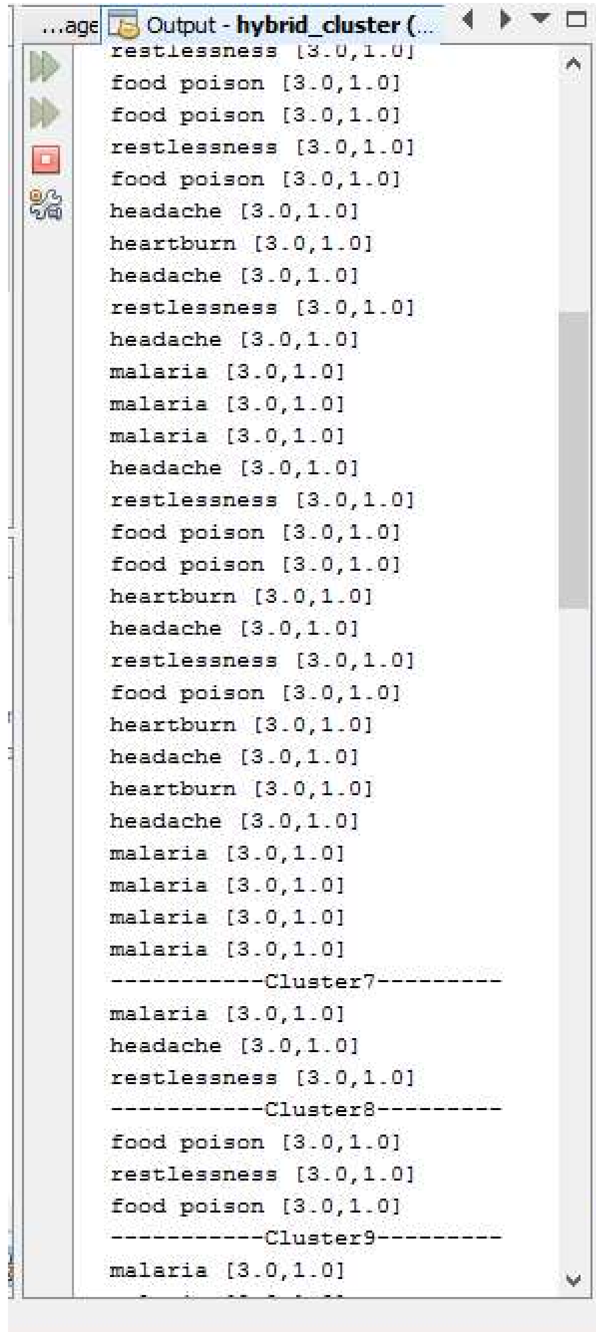

ctor Run Debug Profile Team Tools Window Hel
...age Output - hybrid_cluster (run)
run:
-----------Cluster0---------
malaria [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
food poison [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
-----------Cluster1---------
malaria [3.0,2.0]
-----------Cluster2---------
malaria [3.0,2.0]
-----------Cluster3---------
malaria [3.0,2.0]
-----------Cluster4---------
malaria [3.0,2.0]
food poison [3.0,2.0]
-----------Cluster5---------
food poison [3.0,2.0]
-----------Cluster6---------
restlessness [3.0,1.0]
heartburn [3.0,1.0]
headache [3.0,1.0]
headache [3.0,1.0]
restlessness [3.0,1.0]
food poison [3.0,1.0]
food poison [3.0,1.0]
restlessness [3.0,1.0]
food poison [3.0,1.0]
headache [3.0,1.0]
heartburn [3.0,1.0]
headache [3.0,1.0]
restlessness [3.0,1.0]


...sp; best.java
-----------Cluster1---------
typhoid [3.0,1.0]
malaria [3.0,1.0]
heartburn [3.0,1.0]
food poison [3.0,1.0]
headache [3.0,1.0]
heartburn [3.0,1.0]
headache [3.0,1.0]
restlessness [3.0,1.0]
food poison [3.0,1.0]
headache [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
-----------Cluster2---------
food poison [3.0,2.0]
-----------Cluster3---------
typhoid [3.0,1.0]
restlessness [3.0,1.0]
-----------Cluster4---------
typhoid [3.0,2.0]
-----------Cluster5---------
typhoid [3.0,2.0]
-----------Cluster6---------
typhoid [3.0,2.0]
-----------Cluster7---------
malaria [3.0,1.0]
headache [3.0,1.0]
headache [3.0,1.0]
restlessness [3.0,1.0]
headache [3.0,1.0]
malaria [3.0,1.0]
restlessness [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
food poison [3.0,1.0]
heartburn [3.0,1.0]
malaria [3.0,1.0]

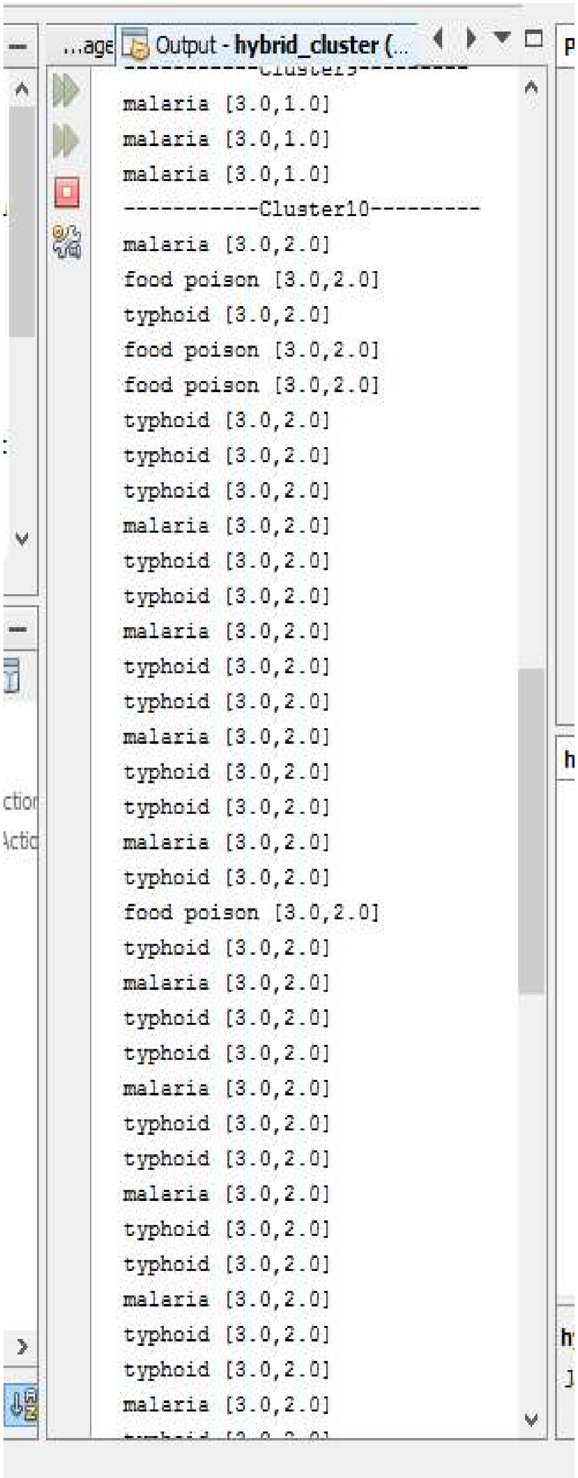
```
-----------Cluster1---------
typhoid [3.0,1.0]
malaria [3.0,1.0]
heartburn [3.0,1.0]
food poison [3.0,1.0]
headache [3.0,1.0]
heartburn [3.0,1.0]
headache [3.0,1.0]
restlessness [3.0,1.0]
food poison [3.0,1.0]
headache [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
-----------Cluster2---------
food poison [3.0,2.0]
-----------Cluster3---------
typhoid [3.0,1.0]
restlessness [3.0,1.0]
-----------Cluster4---------
typhoid [3.0,2.0]
-----------Cluster5---------
typhoid [3.0,2.0]
-----------Cluster6---------
typhoid [3.0,2.0]
-----------Cluster7---------
malaria [3.0,1.0]
headache [3.0,1.0]
headache [3.0,1.0]
restlessness [3.0,1.0]
headache [3.0,1.0]
malaria [3.0,1.0]
restlessness [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
food poison [3.0,1.0]
heartburn [3.0,1.0]
malaria [3.0,1.0]
```

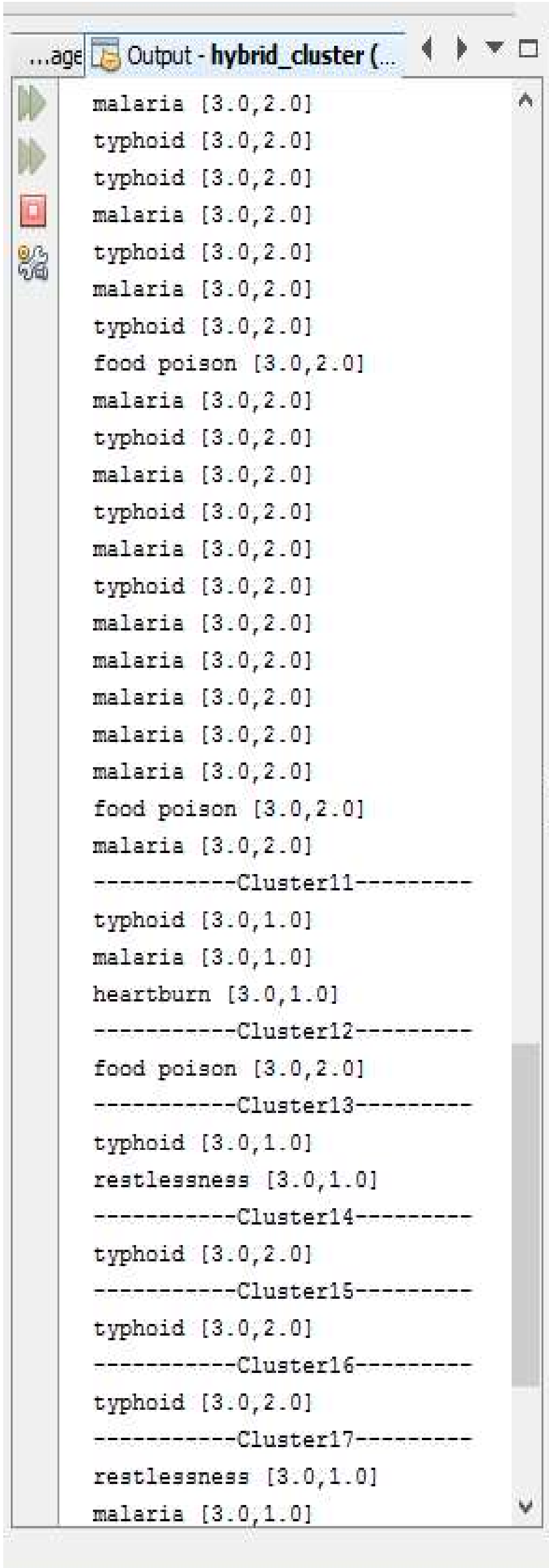
```
malaria [3.0,1.0]
restlessness [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
food poison [3.0,1.0]
heartburn [3.0,1.0]
malaria [3.0,1.0]
food poison [3.0,1.0]
headache [3.0,1.0]
headache [3.0,1.0]
malaria [3.0,1.0]
restlessness [3.0,1.0]
restlessness [3.0,1.0]
malaria [3.0,1.0]
food poison [3.0,1.0]
heartburn [3.0,1.0]
food poison [3.0,1.0]
headache [3.0,1.0]
heartburn [3.0,1.0]
restlessness [3.0,1.0]
heartburn [3.0,1.0]
heartburn [3.0,1.0]
headache [3.0,1.0]
food poison [3.0,1.0]
restlessness [3.0,1.0]
restlessness [3.0,1.0]
food poison [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
food poison [3.0,1.0]
malaria [3.0,1.0]
heartburn [3.0,1.0]
restlessness [3.0,1.0]
-----------Cluster8---------
food poison [3.0,1.0]
malaria [3.0,1.0]
-----------Cluster9---------
food poison [3.0,2.0]
```

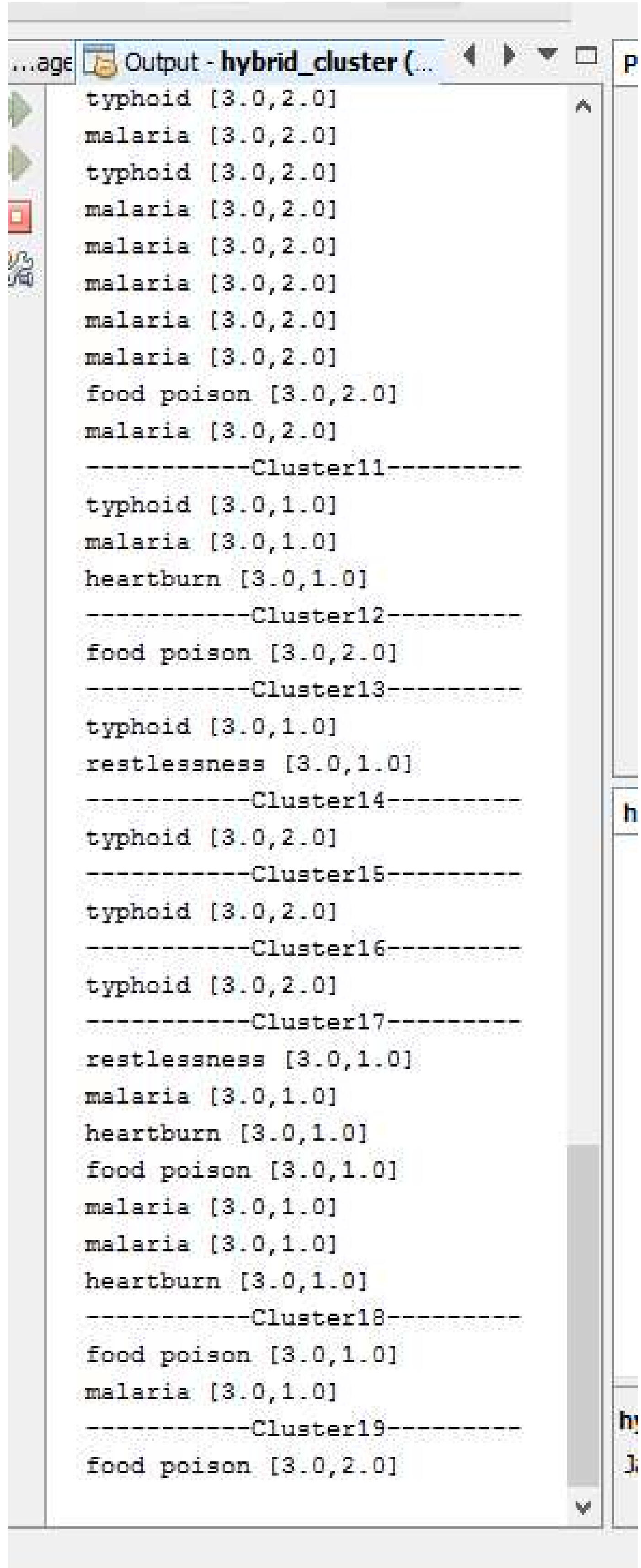
RESULTS OF CLUSTER 20



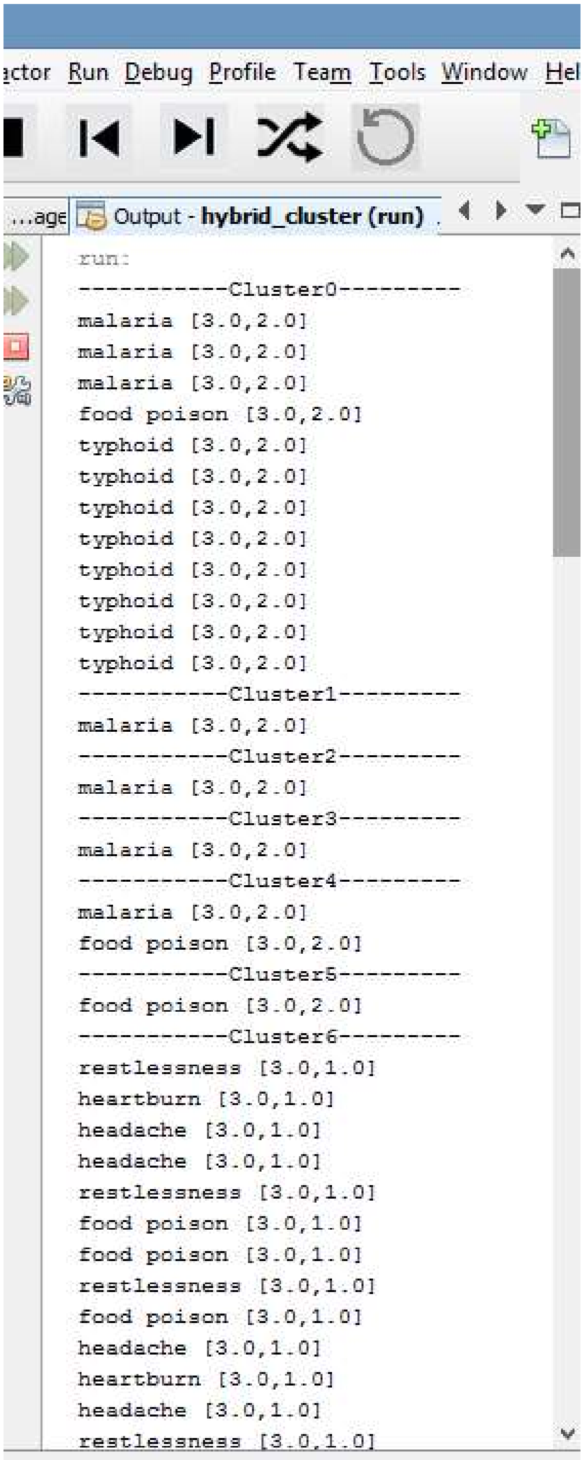

```
restlessness [3.0,1.0]
food poison [3.0,1.0]
food poison [3.0,1.0]
restlessness [3.0,1.0]
food poison [3.0,1.0]
headache [3.0,1.0]
heartburn [3.0,1.0]
headache [3.0,1.0]
restlessness [3.0,1.0]
headache [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
headache [3.0,1.0]
restlessness [3.0,1.0]
food poison [3.0,1.0]
food poison [3.0,1.0]
heartburn [3.0,1.0]
headache [3.0,1.0]
restlessness [3.0,1.0]
food poison [3.0,1.0]
heartburn [3.0,1.0]
headache [3.0,1.0]
heartburn [3.0,1.0]
headache [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
-----------Cluster7---------
malaria [3.0,1.0]
headache [3.0,1.0]
restlessness [3.0,1.0]
-----------Cluster8---------
food poison [3.0,1.0]
restlessness [3.0,1.0]
food poison [3.0,1.0]
-----------Cluster9---------
malaria [3.0,1.0]
```

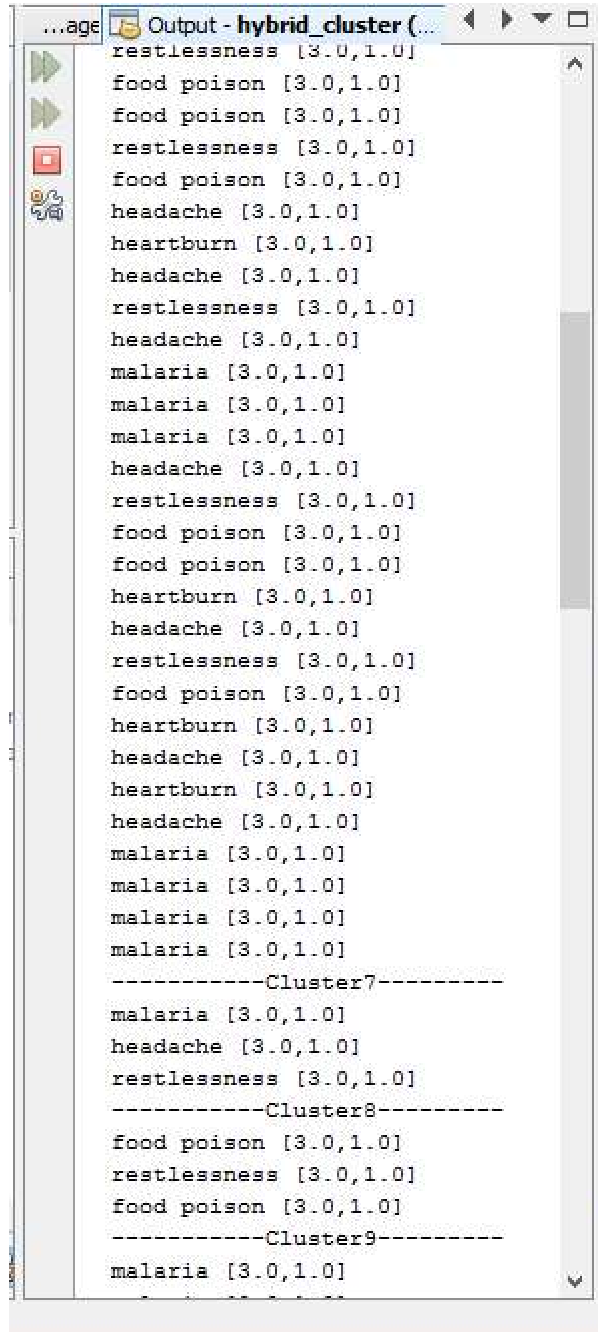

...age
Output - hybrid_cluster (...
malaria [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
-----------Cluster10---------
malaria [3.0,2.0]
food poison [3.0,2.0]
typhoid [3.0,2.0]
food poison [3.0,2.0]
food poison [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
food poison [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]


...age
Output - hybrid_cluster (...
malaria [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
food poison [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
food poison [3.0,2.0]
malaria [3.0,2.0]
-----------Cluster11---------
typhoid [3.0,1.0]
malaria [3.0,1.0]
heartburn [3.0,1.0]
-----------Cluster12---------
food poison [3.0,2.0]
-----------Cluster13---------
typhoid [3.0,1.0]
restlessness [3.0,1.0]
-----------Cluster14---------
typhoid [3.0,2.0]
-----------Cluster15---------
typhoid [3.0,2.0]
-----------Cluster16---------
typhoid [3.0,2.0]
-----------Cluster17---------
restlessness [3.0,1.0]
malaria [3.0,1.0]

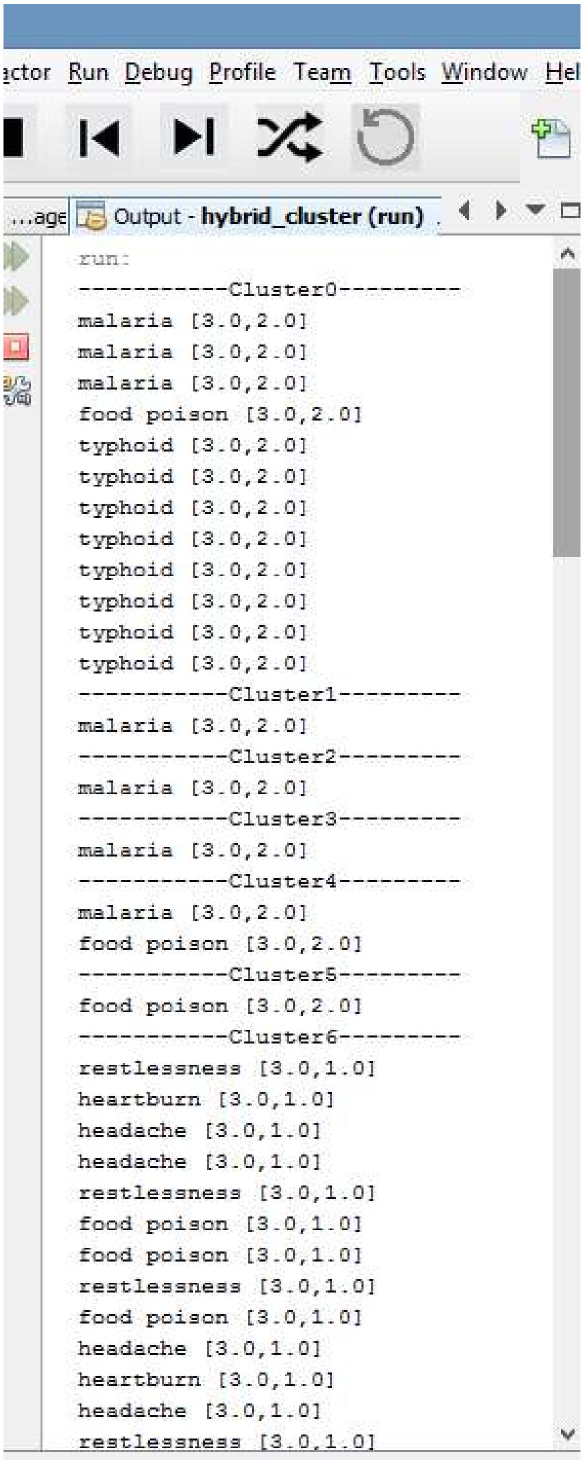

...age
Output - hybrid_cluster (...
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
food poison [3.0,2.0]
malaria [3.0,2.0]
-----------Cluster11---------
typhoid [3.0,1.0]
malaria [3.0,1.0]
heartburn [3.0,1.0]
-----------Cluster12---------
food poison [3.0,2.0]
-----------Cluster13---------
typhoid [3.0,1.0]
restlessness [3.0,1.0]
-----------Cluster14---------
typhoid [3.0,2.0]
-----------Cluster15---------
typhoid [3.0,2.0]
-----------Cluster16---------
typhoid [3.0,2.0]
-----------Cluster17---------
restlessness [3.0,1.0]
malaria [3.0,1.0]
heartburn [3.0,1.0]
food poison [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
heartburn [3.0,1.0]
-----------Cluster18---------
food poison [3.0,1.0]
malaria [3.0,1.0]
-----------Cluster19---------
food poison [3.0,2.0]

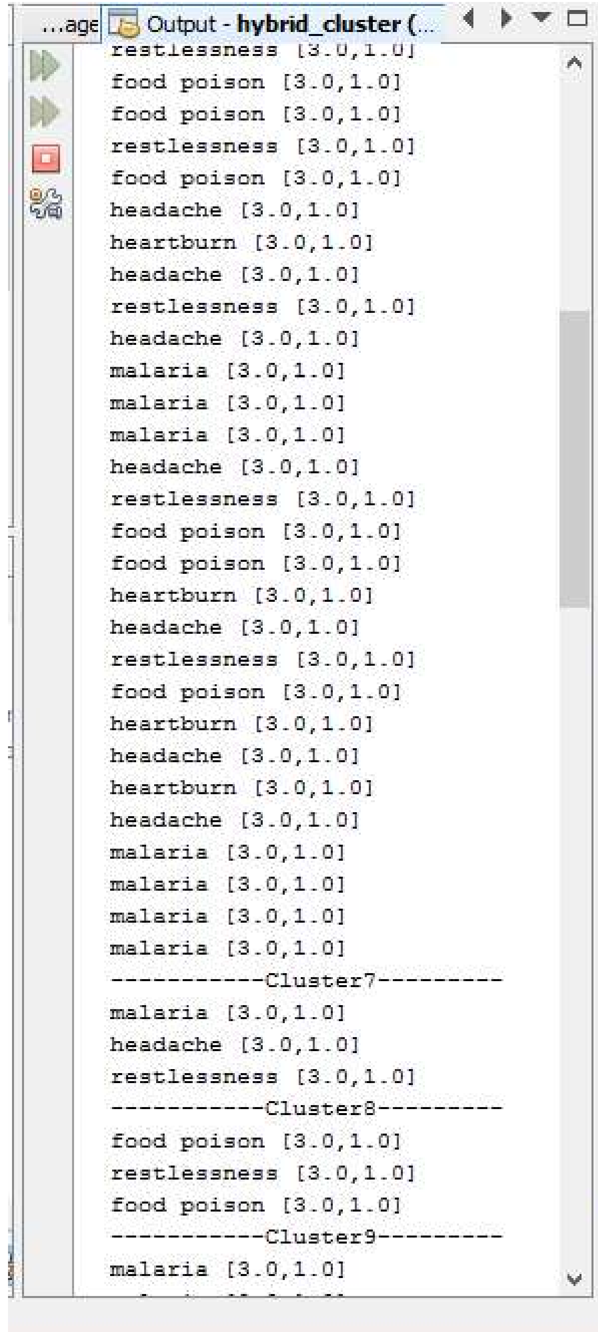
RESULTS OF CLUSTER 20

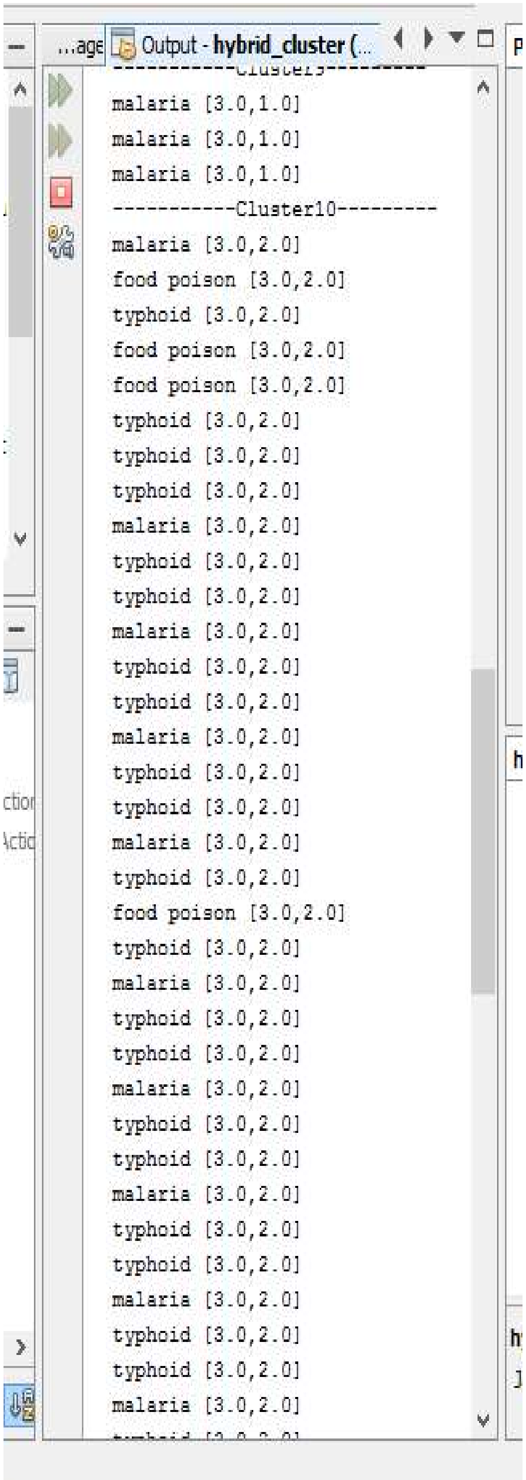


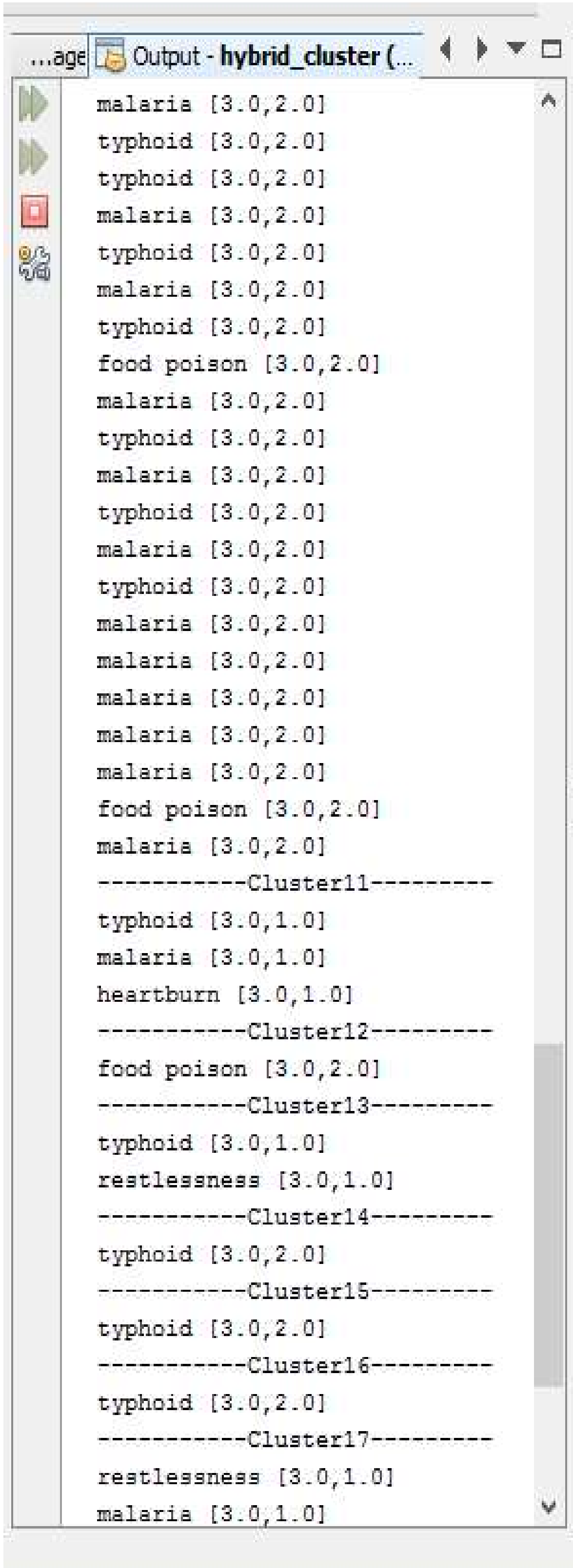


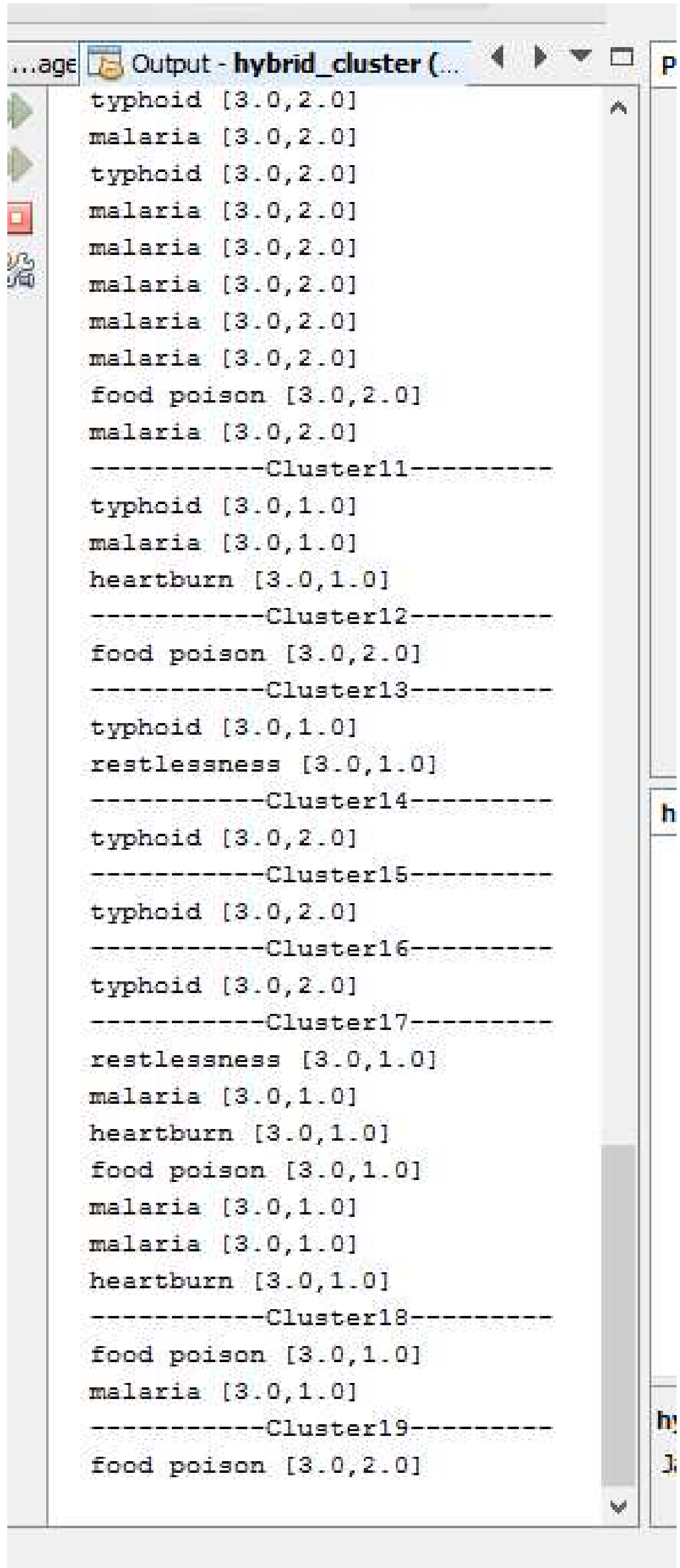
RESULTS OF CLUSTER 20

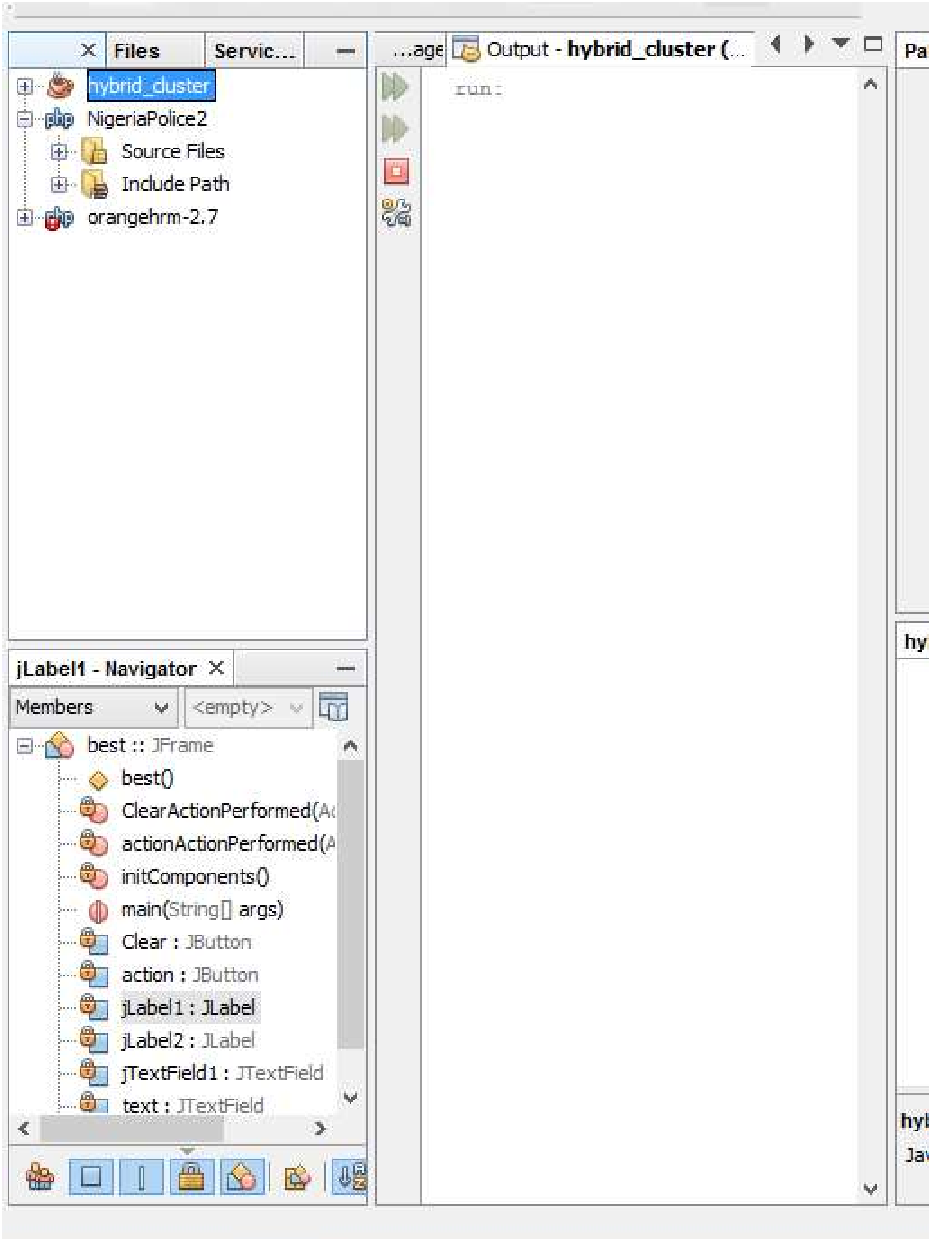


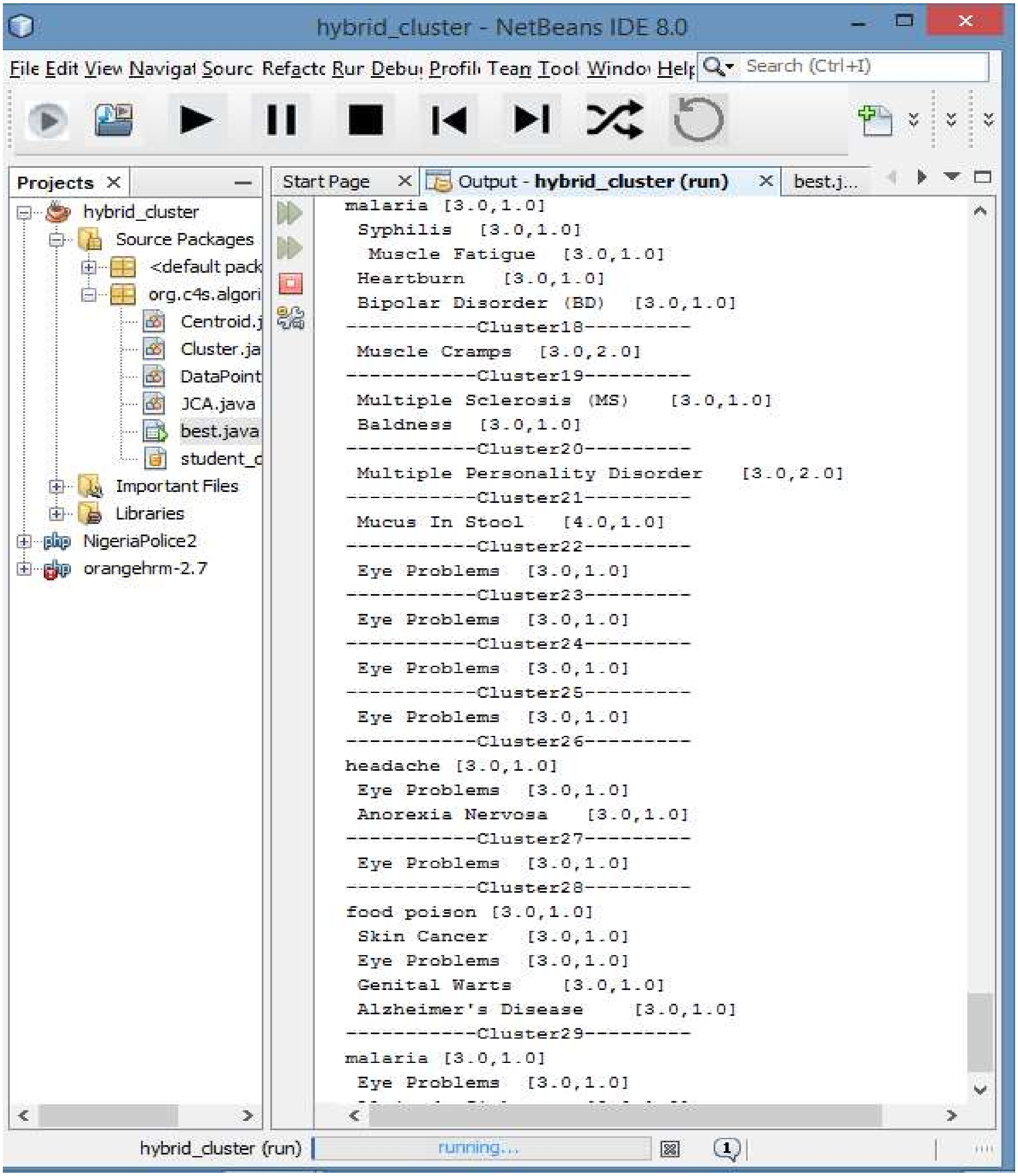



...age
Output - hybrid_cluster (...
malaria [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
-----------Cluster10---------
malaria [3.0,2.0]
food poison [3.0,2.0]
typhoid [3.0,2.0]
food poison [3.0,2.0]
food poison [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
food poison [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]


...age
Output - hybrid_cluster (...
malaria [3.0,2.0]
typhoid [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
food poison [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
food poison [3.0,2.0]
malaria [3.0,2.0]
-----------Cluster11---------
typhoid [3.0,1.0]
malaria [3.0,1.0]
heartburn [3.0,1.0]
-----------Cluster12---------
food poison [3.0,2.0]
-----------Cluster13---------
typhoid [3.0,1.0]
restlessness [3.0,1.0]
-----------Cluster14---------
typhoid [3.0,2.0]
-----------Cluster15---------
typhoid [3.0,2.0]
-----------Cluster16---------
typhoid [3.0,2.0]
-----------Cluster17---------
restlessness [3.0,1.0]
malaria [3.0,1.0]


...age
Output - hybrid_cluster (...
typhoid [3.0,2.0]
malaria [3.0,2.0]
typhoid [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
malaria [3.0,2.0]
food poison [3.0,2.0]
malaria [3.0,2.0]
-----------Cluster11---------
typhoid [3.0,1.0]
malaria [3.0,1.0]
heartburn [3.0,1.0]
-----------Cluster12---------
food poison [3.0,2.0]
-----------Cluster13---------
typhoid [3.0,1.0]
restlessness [3.0,1.0]
-----------Cluster14---------
typhoid [3.0,2.0]
-----------Cluster15---------
typhoid [3.0,2.0]
-----------Cluster16---------
typhoid [3.0,2.0]
-----------Cluster17---------
restlessness [3.0,1.0]
malaria [3.0,1.0]
heartburn [3.0,1.0]
food poison [3.0,1.0]
malaria [3.0,1.0]
malaria [3.0,1.0]
heartburn [3.0,1.0]
-----------Cluster18---------
food poison [3.0,1.0]
malaria [3.0,1.0]
-----------Cluster19---------
food poison [3.0,2.0]


Files
Servic...
hybrid_cluster
NigeriaPolice2
Source Files
Include Path
orangehrm-2.7
jLabel1 - Navigator
Members
<empty>
best :: JFrame
best()
ClearActionPerformed(A
actionActionPerformed(A
initComponents()
main(String[] args)
Clear : JButton
action : JButton
jLabel1 : JLabel
jLabel2 : JLabel
jTextField1 : JTextField
text : JTextField
...age
Output - hybrid_cluster (...
run:


hybrid_cluster - NetBeans IDE 8.0
File Edit View Navigat Sourc Refactc Rur Debu Profil Tean Tool Windo Helr
Search (Ctrl+I)
Projects
hybrid_cluster
Source Packages
<default pack
org.c4s.algori
Centroid.j
Cluster.ja
DataPoint
JCA.java
best.java
student_c
Important Files
Libraries
NigeriaPolice2
orangehrm-2.7
Start Page
Output - hybrid_cluster (run)
best.j...
malaria [3.0,1.0]
Syphilis [3.0,1.0]
Muscle Fatigue [3.0,1.0]
Heartburn [3.0,1.0]
Bipolar Disorder (BD) [3.0,1.0]
-----------Cluster18---------
Muscle Cramps [3.0,2.0]
-----------Cluster19---------
Multiple Sclerosis (MS) [3.0,1.0]
Baldness [3.0,1.0]
-----------Cluster20---------
Multiple Personality Disorder [3.0,2.0]
-----------Cluster21---------
Mucus In Stool [4.0,1.0]
-----------Cluster22---------
Eye Problems [3.0,1.0]
-----------Cluster23---------
Eye Problems [3.0,1.0]
-----------Cluster24---------
Eye Problems [3.0,1.0]
-----------Cluster25---------
Eye Problems [3.0,1.0]
-----------Cluster26---------
headache [3.0,1.0]
Eye Problems [3.0,1.0]
Anorexia Nervosa [3.0,1.0]
-----------Cluster27---------
Eye Problems [3.0,1.0]
-----------Cluster28---------
food poison [3.0,1.0]
Skin Cancer [3.0,1.0]
Eye Problems [3.0,1.0]
Genital Warts [3.0,1.0]
Alzheimer's Disease [3.0,1.0]
-----------Cluster29---------
malaria [3.0,1.0]
Eye Problems [3.0,1.0]
hybrid_cluster (run)
running...