

**SECTION ALLOCATION OPTIMIZATION FOR
COMPUTER SKILLS-2 STUDENTS USING DATA
MINING TECHNIQUES**

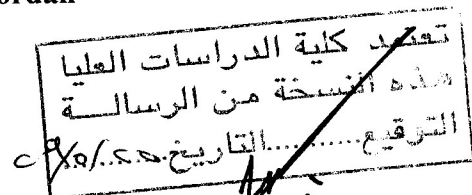
**By
Esra Fawaz Al-Zaghoul**

**Supervisor
Dr. Ammar Huneiti**

**Co-Supervisor
Dr. Imad Salah**

**This Thesis was Submitted in Partial Fulfillment of the Requirements for the
Master's Degree of Science in Information Systems
Faculty of Graduate Studies
The University of Jordan**

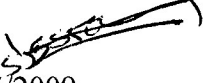
May, 2009



د. امّار هنيّتي

**The University of Jordan
Authorization Form**

I, Esra Fawaz Ahmad Al-Zaghoul, authorize the University of Jordan to supply copies of my Thesis/Dissertation to libraries or establishments or individuals on request, according to the University of Jordan regulations.

Signature: 
Date: 19/5/2009

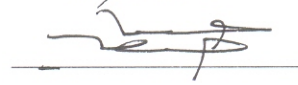
COMMITTEE DECISION

This Thesis/Dissertation (Section Allocation Optimization for Computer Skills-2 Students Using Data Mining Techniques) was Successfully Defended and Approved on 14/5/2009

Examination Committee

Dr. Ammar M. Huneiti, (Supervisor)
Assist. Prof. of Hypermedia-Based Performance
Support Systems for the Web

Signature



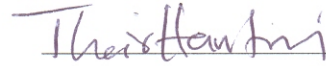
Dr. Imad K. Salah, (Co-Supervisor)
Assoc. Prof. of Complex Systems & Networks



Dr. "Mohammad Belal" Al-Zoubi, (Member)
Assoc. Prof. of Data Mining / GIS and
Computer Graphics



Dr. Thair Mahmoud Hamtini, (Member)
Assist. Prof. of E-learning



Dr. Ghassan Kanaan, (Member)
Assoc. Prof. of Natural Language Processing
/Information Retrieval
(The Arab Academy for Banking and Financial Sciences)



تعتمد كلية الدراسات العليا
هذه النسخة من الرسالة
التوقيع..... التاريخ 14/5/2009



Dedication

This thesis is dedicated to my family, especially Mum and Dad, who taught me that even the largest task can be accomplished if it is done one step at a time, and who always offered me unconditional love and support especially throughout the course of this thesis.

Acknowledgement

After many thanks to Allah, I would like first to thank my parents, who taught me how to express my ideas, for being supportive and loving, and for their hard work in bringing me up. They were always there to listen and to give advice. Dear Mum and Dad. What can I say? You've been with me, All along the way. Sweet family means everything in the world to me. I move ahead because of you. For all your love and your support a million words would be too short. I don't always show it but you know that I do appreciate how much the both of you have helped me with my life and given me all of the things that have gotten me here. Thank you for all those times you stood by me, for all the truth that you made me see, for all the joy you brought to my life, for all the wrong that you made right, and for every dream you made come true. I'll be forever thankful. You're the one who held me up, never let me fall. You were my strength when I was weak, my voice when I couldn't speak, and my eyes when I couldn't see. You saw the best there was in me. So thanks again to Mum and Dad, brother and sister - you're a dear- you were always there for me.

A special thanks to my supervisor Dr.Ammar Huneiti and my co-supervisor Dr.Imad Salah. A special thanks to Dr.Fawaz Zaghoul and Dr.Mousa Al-Akhras for helping me a lot. And a special thanks to my Web team at work and my colleagues.

Finally, I would like to thank everyone who helped me to complete my thesis.

Contents

COMMITTEE DECISION	ii
Dedication	iii
Acknowledgement	iv
Contents	v
List of Tables	ix
List of Figures	xi
Abstract	xiv
1 Introduction	1
1.1 Preface	1
1.2 Computer Skills	1
1.2.1 Computer Skills-1	1
1.2.2 Computer Skills-2	2
1.3 Problem Statement	2
1.4 Problem Signification	3
1.5 Research Objectives	3
1.6 Main Contribution	4
1.7 Chapters' Organization	4

2 Literature Review	5
2.1 Data Mining Background	5
2.2 Machine Learning Background	5
2.2.1 Supervised Learning	6
2.2.2 Unsupervised Learning	6
2.2.3 Data Clustering	6
2.3 Related Works	7
2.3.1 Purposes and Principle of Clustering	7
2.3.2 Students Levels Evaluation	8
2.3.3 Clustering Techniques	8
2.3.4 Modeling of Intelligent Learning	9
2.3.5 Intelligent Systems Approaches using Clustering Techniques	10
2.3.6 Data Normalization	11
2.3.7 K-Means Algorithm	13
2.3.8 K-means Clustering Algorithm Time Complexity	14
2.3.9 Clusters Evaluations	14
3 Materials and Methods	16
3.1 Data Collection	17
3.2 Data Preprocessing	18
3.3 Data Clustering	20
3.4 K-means Clustering Algorithm	20
3.5 Data Analysis	23
3.5.1 Significant Attributes Selection	23
3.5.2 Number of Clusters (Groups) Selection	23
3.5.3 Suitable Distance Function Selection	24
3.5.4 Suitable Cluster Identification Selection	27

3.6	Clusters Representation	27
3.7	Cluster Evaluation	27
3.7.1	User Inspection	28
3.7.2	Purity	28
3.7.3	Precision	29
3.7.4	Recall	29
4	Experimental and Theoretical Results	31
4.1	Data Preprocessing	34
4.2	Datasets Testing	37
4.3	Data Interpretation	38
4.3.1	Significant Attributes	38
4.3.2	Suitable Number of Clusters	39
4.3.3	Suitable Distance Function	42
	Experiment No.1	42
	Experiment No.2	42
4.4	Clusters Representation	44
4.5	Clusters Evaluation	44
4.5.1	Evaluation of Experiment No.2 with k=4	44
4.5.2	Evaluation of Experiment No.2 with k=7	47
4.6	Final Results	49
4.6.1	Suitable Grouping (Cluster Identification) for CS2 Students	49
4.6.2	Evaluation of Experiment No.2 with k=4 after categorization.	51
4.6.3	CS2 Students' Distribution	53
5	Conclusions and Recommendations	63
5.1	Conclusions	63
5.2	Recommendations	64

6 Future Work	65
6.1 Outliers Analysis	65
6.2 Dealing with Outliers	65
A Experiment No.1 Clusters	67
B Experiment No.2 with K=6	74
C Experiment No.2 Figures	84
References	101
Abstract (In Arabic)	108

List of Tables

4.1	Number of Students per Cluster in Experiment No.2 with k=4	45
4.2	Confusion matrix with purity values for (CS2) in Experiment No.2 with k=4 before categorization.	46
4.3	Precision values for (CS2) in Experiment No.2 with k=4 before categorization.	46
4.4	Recall values for (CS2) in Experiment No.2 with k=4 before categorization.	46
4.5	Number of Students per Cluster in Experiment No.2 with k=7.	47
4.6	Confusion matrix with purity values for (CS2) in Experiment No.2 with k=7.	49
4.7	Precision values for (CS2) in Experiment No.2 with k=7.	49
4.8	Recall values for (CS2) in Experiment No.2 with k=7.	49
4.9	Categories of CS2 students with their ranges.	50
4.10	Confusion Matrix with purity values for (CS2) in Experiment No.2 with k=4 after categorization.	51
4.11	Precision values for (CS2) in Experiment No.2 with k=4 after categorization.	52
4.12	Recall values for (CS2) in Experiment No.2 with k=4 after categorization.	52
B.1	Number of Students per Cluster in Experiment No.2 with k=6	74
B.2	Confusion matrix with purity values for (CS2) in Experiment No.2 with k=6.	77
B.3	Precision values for (CS2) in Experiment No.2 with k=6.	77

B.4 Recall values for (CS2) in Experiment No.2 with k=6. 78

List of Figures

3.1	Materials and Methods (Methodology) steps Flow Chart	17
3.2	A snapshot of the Students' Data records.	18
3.3	K-means Clustering Algorithm Flow Chart	21
3.4	K-means Clustering Algorithm	22
3.5	Silhouette Plot Example1.	25
3.6	Silhouette Plot Example2.	26
4.1	Experimental and Theoretical steps Flow Chart	33
4.2	Different datasets before and after preprocessing and discretization	35
4.3	Pie Graph for the whole (CS1) Records.	36
4.4	A snapshot of Students Data records after preprocessing task is completed.	37
4.5	Square Euclidean Distance Function Silhouettes for Experiment No.2 with different number of clusters (K).	40
4.6	Square Euclidean Distance Function Silhouettes for Experiment No.2 with different number of clusters (K).	41
4.7	Manhattan Distance Function Silhouettes for Experiment No.1.	43
4.8	Euclidean Distance Function Silhouettes for Experiment No.1.	54
4.9	Square Euclidean Distance Function Silhouettes for Experiment No.1.	55
4.10	Euclidean Distance Function Silhouettes for Experiment No.2.	56
4.11	Square Euclidean Distance Function Silhouettes for Experiment No.2.	57
4.12	3-D Diagram of CS2 clusters with k=4.	58

4.13	3-D Diagram of CS2 clusters with k=7.	59
4.14	3-D Diagram of CS2 categories for the results (Experiment No.2 with k=4).	60
4.15	CS2 attribute's values distribution all over the clusters (clusters 1 and 2).	61
4.16	CS2 attribute's values distribution all over the clusters (clusters 3 and 4).	62
6.1	Clustering with and without the effect of outliers (Liu, 2007).	65
A.1	Pie Graph for the whole (CS1) Records.	68
A.2	Data points distribution in Cluster1.	69
A.3	Data points distribution in Cluster2.	70
A.4	Data points distribution in Cluster3.	71
A.5	Data points distribution in Cluster4.	72
A.6	Data points distribution in Cluster5.	73
B.1	3-D Diagram of CS2 clusters with k=6.	75
B.2	3-D Diagram of CS2 clusters with k=6 (Backside view).	76
B.3	Data points distribution's Histogram in Cluster1 (k=6).	78
B.4	Data points distribution's Histogram in Cluster2 (k=6).	79
B.5	Data points distribution's Histogram in Cluster3 (k=6).	80
B.6	Data points distribution's Histogram in Cluster4 (k=6).	81
B.7	Data points distribution's Histogram in Cluster5 (k=6).	82
B.8	Data points distribution's Histogram in Cluster6 (k=6).	83
C.1	Data points distribution's Histogram in Cluster1 (k=4).	84
C.2	Data points distribution's Histogram in Cluster2(k=4).	85
C.3	Data points distribution's Histogram in Cluster3(k=4).	85
C.4	Data points distribution's Histogram in Cluster4(k=4).	86
C.5	HSGE Attribute's Histogram in Cluster1 (k=4).	86
C.6	HSGE Attribute's Histogram in Cluster2 (k=4).	87

C.7	HSGE Attribute's Histogram in Cluster3 (k=4).	87
C.8	HSGE Attribute's Histogram in Cluster4 (k=4).	88
C.9	CGPA Attribute's Histogram in Cluster1 (k=4).	88
C.10	CGPA Attribute's Histogram in Cluster2 (k=4).	89
C.11	CGPA Attribute's Histogram in Cluster3 (k=4).	89
C.12	CGPA Attribute's Histogram in Cluster4 (k=4).	90
C.13	CS1 Attribute's Histogram in Cluster1 (k=4).	90
C.14	CS1 Attribute's Histogram in Cluster2 (k=4).	91
C.15	CS1 Attribute's Histogram in Cluster3 (k=4).	91
C.16	CS1 Attribute's Histogram in Cluster4 (k=4).	92
C.17	English1 Attribute's Histogram in Cluster1 (k=4).	92
C.18	English1 Attribute's Histogram in Cluster2 (k=4).	93
C.19	English1 Attribute's Histogram in Cluster3 (k=4).	93
C.20	English1 Attribute's Histogram in Cluster4 (k=4).	94
C.21	Data points distribution's Histogram in Cluster1 (k=7).	94
C.22	Data points distribution's Histogram in Cluster2 (k=7).	95
C.23	Data points distribution's Histogram in Cluster3 (k=7).	95
C.24	Data points distribution's Histogram in Cluster4 (k=7).	96
C.25	Data points distribution's Histogram in Cluster5 (k=7).	96
C.26	Data points distribution's Histogram in Cluster6 (k=7).	97
C.27	Data points distribution's Histogram in Cluster7 (k=7).	97
C.28	Square Euclidean Distance Function Silhouettes for Experiment No.2 with k=2.	98
C.29	Square Euclidean Distance Function Silhouettes for Experiment No.2 with k=3.	99
C.30	Euclidean Distance Function Silhouettes for Experiment No.2 with k=2.	99
C.31	Euclidean Distance Function Silhouettes for Experiment No.2 with k=3.	100

SECTION ALLOCATION OPTIMIZATION FOR COMPUTER SKILLS-2 STUDENTS USING DATA MINING TECHNIQUES

By

Esra Fawaz Zaghoul

Supervisor

Dr. Ammar M. Huneiti

Co-Supervisor

Dr. Imad K. Salah

Abstract

Data clustering is one of the most important tools for analyzing the structure of datasets. It has been applied to various fields such as machine learning, data mining, pattern recognition, image analysis, bioinformatics, information retrieval. The most difficult problems in cluster analysis are the identification of the number of clusters in a dataset, determining to which cluster each member belongs, and finding the suitable distance function. This research investigated the problem of Computer Skills-2 (CS2) course's grade scale based on K-means clustering technique. It applied this technique to the University of Jordan (UJ) students who took the Computer Skills-2 course. In particular, the K-means clustering algorithm is employed to distinguish student groups who might share similar misconceptions.

The main objective of this research is to find a solution to the CS2 classes' classification problem. CS2 course problem resides in its grade scale at the end of the course semester. CS2 course is given to both medical and humanity colleges' students within the same sections. The CS2 grade scale usually consists of two categories of grades. These categories are two levels. The first level is the high scores, usually consist of the medical colleges' students, and the other is the low scores, which is usually from humanity colleges' students. If similar groups of students can be found depending on their previous backgrounds, knowledge, type of education, discipline, abilities and skills, then the coordinator of the CS2 course will be able to allocate them to same sections. This allocation will enable instructors to provide students with specific topics, help, mentor, materials, exams and exercises, according to a specific knowledge level that is suitable for each section. This may enable students to gain better knowledge, well understanding and get higher marks.

The main objectives of this research are:

1. Finding the relationship between students' attributes and their academic achievements in CS2.
2. Finding groups of students of similar interests i.e. backgrounds, knowledge, type of education, discipline, abilities and skills.
3. Adapting the CS2 course to these groups of similar interests.
4. Improving the students' achievements and close academic gaps among students of different backgrounds.

To solve the problem of CS2 classes' classification, this research endeavors to find the relationship between students' attributes and their academic achievements in CS2. It also introduces an approach that provides support in finding groups of students of similar interests. This approach will also help in adapting the CS2 course to these groups of similar interests. Many experimental tests have been done on many datasets, using MATLAB ®2008b (The MathWorks, 2008). The research endeavor will investigate the problem of CS2 course's grade scale based on K-means clustering technique. This technique will be applied to the UJ students who took the CS2 course and find out the results.

Clustering has been shown to be one of the most commonly used data analysis techniques. It also has a long history, and has been used in almost every field, e.g., medicine, psychology, botany, sociology, biology, archeology, marketing, insurance and library science.

1. Introduction

1.1. Preface

Now a days, huge amount of data is generated which contains important information that accumulates daily in databases and is not easy to extract. Data mining was developed as a means of extracting information and knowledge from databases to discover patterns or concepts that are not obvious. Machine learning provides the technical basis of data mining by extracting information from the raw data in the databases (Amershi and Conati, 2007). There are two types of machine learning (*i*) supervised learning and (*ii*) unsupervised learning. Data Clustering is one primary type of unsupervised learning (Bach and Jordan, 2006; Jain et al., 1999).

1.2. Computer Skills

There are two course subcategories of the Computer skills course; (*i*) Computer Skills-1 (CS1) and (*ii*) Computer Skills-2 (CS2).

1.2.1. Computer Skills-1

CS1 course is one of the prerequisite courses of the CS2. The CS1 course introduces students to the fundamentals of computer. It consists of an introduction to computer hardware and software. It also introduces students to a detailed functional description of the main parts of a modern computer and Microsoft office applications such as (Word, Excel, PowerPoint...etc.). Each student at the UJ has to take the CS1 course unless he/she had passed its qualification exam.

1.2.2. Computer Skills-2

One of the obligatory courses at the UJ is the CS2 course. CS2 is one of the fundamental courses at the UJ. It is given by Computer Information Systems (CIS) Department at King Abdullah II School for Information Technology (KASIT). The CS2 course introduces students to problem solving using computer variables, algorithms and its representation, data types and definitions. It also introduces them to use advanced applications using software packages such as MS Word templates, comparing documents, table of contents, indexing, inserting data, mailing merge, MS Excel charts, functions, sorting and filtering, MS Access tables, relations, forms, queries, reports, import and export files and data, and introduction to the web applications.

1.3. Problem Statement

CS2 course is given to both medical and humanity colleges' students within the same sections i.e. medical and humanity students are taking the same course sections. This affects the academic achievement of both colleges' students and makes an achievement gaps which persist between medical and humanity colleges' students. CS2 course problem resides in its grade scale at the end of the course semester. Because of the academic achievement gaps among students, the CS2 grade scale usually consists of two categories of grades. These categories are two levels. The first level is the high scores, usually consist of the medical colleges' students, and the other is the low scores, which is usually from humanity colleges' students.

The main reason of the problem is due to the differences between students and their backgrounds, knowledge, type of education, discipline, abilities, skills, understanding, possible misconceptions and many other attributes that will be discussed. In order to solve such problem, its causes have to be found. So a sample of student data was requested. This sample was the whole CS2 sections for the last Spring Semester (2007/2008).

1.4. Problem Signification

In a classroom, a teacher attempts to convey his/her knowledge to the students, thus it is important for the teacher to obtain formative feedback about how well students understand his/her material. By gaining insight into the students' understanding and possible misconceptions, the teacher will be able to adjust the teaching and to supply more useful learning materials as necessary. Therefore, the diagnosis of formative student evaluations is critical for teachers and learners, as is the diagnosis of patterns in the overall learning by a class in order to inform a teacher about the efficacy of his/her teaching.

The purpose of this research is to find solutions to the Computer Skills-2 (CS2) classes' classification problem that resides in its grade scale at the end of the course semester. If similar groups of students can be found depending on their previous backgrounds, knowledge, type of education, discipline, abilities and skills, then the coordinator of the CS2 course will be able to allocate them to same sections. This allocation will enable instructors to provide students with specific topics, help, mentor, materials, exams and exercises, according to a specific knowledge level that is suitable for each section. This may enable students to gain better knowledge, well understanding and get higher marks.

1.5. Research Objectives

The main objectives of this research are:

1. Finding the relationship between students' attributes and their academic achievements in CS2.
2. Finding groups of students of similar interests i.e. backgrounds, knowledge, type of education, discipline, abilities and skills.
3. Adapting the CS2 course to these groups of similar interests.

4. Improving the students' achievements and close academic gaps among students of different backgrounds.

1.6. Main Contribution

To solve the problem of CS2 classes' classification, this research endeavors to find the relationship between students' attributes and their academic achievements in CS2. It also introduces an approach that provides support in finding groups of students of similar interests. This approach will also help in adapting the CS2 course to these groups of similar interests. It will also improve the students' achievements and close academic gaps among students of different backgrounds. Many experimental tests have been done on many datasets, using MATLAB ®2008b (The MathWorks, 2008). The research endeavor will investigate the problem of CS2 course's grade scale based on K-means clustering technique. This technique will be applied to the UJ students who took the CS2 course and find out the results.

Clustering has been shown to be one of the most commonly used data analysis techniques. It also has a long history, and has been used in almost every field, e.g., medicine, psychology, botany, sociology, biology, archeology, marketing, insurance and library science.

1.7. Chapters' Organization

Chapter 2 gives an overview on the previous work that is related to the problem. Chapter 3 explains the methodology of this research. Chapter 4 introduces the result of the experimental tests. Next, Chapter 5 conclude with a summary on the result of the experiments and give recommendations. Finally, Chapter 6 view a suggestions for future work.

2. Literature Review

2.1. Data Mining Background

Humans have been "manually" extracting information from data for centuries, but the increasing volume of data in modern times has called for more automatic approaches. As datasets and the information extracted from them has grown in size and complexity, direct hands-on data analysis has increasingly been supplemented and augmented with indirect, automatic data processing using more complex and sophisticated tools, methods and models. Data mining is the process of using computing power to apply methodologies, including new techniques for knowledge discovery, to data (Maimon and Rokach, 2005).

Data mining identifies trends within data that go beyond simple data analysis. Through the use of sophisticated algorithms, non-statistician users have the opportunity to identify key attributes of processes and target opportunities (Christen, 2007; Pan et al., 2007).

Data mining is the process of extracting hidden patterns from large amounts of data. As more data is gathered, with the amount of data doubling every year, data mining is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery (Yin et al., 2005; Ciriani et al., 2008).

2.2. Machine Learning Background

Machine learning is the subfield of artificial intelligence that is concerned with the design and development of algorithms that allow computers to improve their performance over time based on data, such as from sensor data or databases (Mitchell, 1997). A major focus of machine learning research is to automatically produce and induce models, such as rules and patterns, from data. Hence, machine learning is closely related to fields such as

data mining, statistics, inductive reasoning, pattern recognition, and theoretical computer science (Weiss, 1997). Machine learning provides the technical basis of data mining by extracting information from the raw data in the databases (Jain et al., 1999; Bach and Jordan, 2006). There are two types of machine learning (1) supervised learning and (2) unsupervised learning. Data Clustering is one primary type of unsupervised learning (Amershi and Conati, 2007).

2.2.1. Supervised Learning

In this type, knowledge is acquired from training data. The training data consist of input objects and desired outputs. The output can predict a class label of the input object (classification). The task of the supervised learner is to predict the value of any valid input object after having seen a number of training examples, i.e. supervised learning discovers patterns in the data that relate data attributes to a class attribute (Kogan et al., 2006; Kim and Lee, 2000).

2.2.2. Unsupervised Learning

Unsupervised learning is a technique that is used to determine how the data are organized. This type of learning is distinguished from supervised learning in that the learner is given only unlabeled examples, i.e. the user wants to explore the data to find some common structures or patterns. From a theoretical point of view, supervised and unsupervised learning differ in the availability of the desired output. Clustering is one technology for finding such structures or patterns (Amershi and Conati, 2007; Liu, 2007).

2.2.3. Data Clustering

Data clustering is an unsupervised learning, because in order to get the number of clusters, we have to investigate, explore and fetch the data points we have. We need data clustering to find the common attributes in the data and then find their clusters (Jain et al., 1999). Data clustering is one of the most important tools for analyzing the structure

of data sets (Savvion Incorporated, 2006). It has been applied to various fields such as machine learning, data mining, pattern recognition, image analysis, bioinformatics, information retrieval...etc (Kogan et al., 2006). One of the most difficult problems in cluster analysis is the identification of the number of clusters in a data set, determining to which cluster each member belongs, finding the suitable distance function...etc. The k-means algorithm is the best known partitioning clustering algorithm. K-means computes cluster centroids differently for each distance measure, to minimize the sum with respect to the measure that has been specified (Bunn and Carminati, 1988; Pun and Ali, 2007).

2.3. Related Works

Here are some previous works which are related to the CS2 classes' classification problem:

2.3.1. Purposes and Principle of Clustering

A new approach, called CrossClus, which carried out cross-relational clustering with user's guidance, was proposed by Yin et al. (2005). CrossClus extracted the set of highly relevant features in multiple relations connected via linkages defined in the database schema, evaluates their effectiveness based on user's guidance, and identifies interesting clusters that fit user's needs. This method took care of both quality in feature extraction and efficiency in clustering (Yin et al., 2005).

A technique that can be used for various clustering purposes was proposed by Hoppner and Klawonn (2008). They also extended the approach to avoid extremely small or large clusters in their cluster analysis. They considered the problem of subdividing a set X of N objects into C homogeneous groups with size constraint (Klawonn and Hoppner, 2006; Hoppner and Klawonn, 2008).

An approach for assigning students into proper groups which was based on the fuzzy clustering techniques was proposed by (Al-Zoubi et al., 2008).

2.3.2. Students Levels Evaluation

Nakkrasae et al. (2004) employed a computational intelligence approach to classify software component repository into similarity component cluster groups based on Fuzzy Subtractive Clustering algorithm. His approach not only is suitable for multidimensional data, but also automatically decides the correct model classification (Nakkrasae, 2004).

A new formulation of the conceptual clustering problem was proposed by Mishra et al. (2004). Their goal was to explicitly output a collection of simple and meaningful conjunctions of attributes that define the clusters. The formulation differs from previous approaches since the clusters discovered may overlap and also may not cover all the points (Mishra et al., 2004).

Song and Rajasekaran (2005) studied the k-means clustering problem and proposed three constant approximation algorithms for the k-means clustering (Song and Rajasekaran, 2005).

Christen (2007) discussed the student population and presented a course structure and its assessments. His aim was to cover more than just the core data mining techniques and algorithms (like classification, prediction, clustering and association rule mining), but also to expose students to other important issues relevant to the knowledge discovery in databases (KDD) process, ranging from data quality, pre-processing and integration to privacy and social impacts of data mining. Additionally, he intended to give students insight into current data mining research (Christen, 2007).

2.3.3. Clustering Techniques

Jain et al. (1999) presented an overview of pattern clustering methods from a statistical pattern recognition perspective. They provided useful references to fundamental concepts accessible to the broad community of clustering practitioners. They presented taxonomy of clustering techniques, and identified cross-cutting themes and recent advances. They also described some important applications of clustering algorithms (Jain et al., 1999).

Kim and Lee (2000) discussed a new type of semi-supervised document clustering and attempted to isolate more semantically coherent clusters by employing the domain-specific knowledge provided by a document analyst. They used an external human knowledge to guide the clustering mechanism with some flexibility when creating the clusters. Their approach was based on a variant of complete-linkage agglomerative hierarchical clustering. They also developed the concepts of requested clusters by exploiting user relevance feedback (Kim and Lee, 2000).

Borodion et al. (2004) proved that computing an approximate solution can be done much more efficiently using fairly natural clustering rules. More specifically, for agglomerative clustering (used, for example, in the Alta VistaTM search engine), for the clustering defined by sparse partitions, and for a clustering based on minimum spanning trees we derive randomized approximation (Borodin et al., 2004).

Bach and Jordan (2006) derived a new cost function for spectral clustering based on a measure of error between a given partition and a solution of the spectral relaxation of a minimum normalized cut problem. They also developed a tractable approximation of their cost function that was based on the power method of computing eigenvectors (Bach and Jordan, 2006).

The novel correlation clustering algorithm COPAC (Correlation PARTition Clustering) that aimed to improve robustness, completeness, usability, and efficiency, was proposed by Achtert et al. (2007). Their experimental evaluation empirically showed that COPAC is superior over existing state-of-the-art correlation clustering methods in terms of runtime, accuracy, and completeness of the results (Achtert et al., 2007).

2.3.4. Modeling of Intelligent Learning

A user modeling framework was outlined that uses both unsupervised and supervised machine learning in order to reduce development costs of building user models, and facilitate transferability. It was applied to model student learning during interaction with the

Adaptive Coach for Exploration (ACE) learning environment (using both interface and eye-tracking data). In addition to demonstrating framework effectiveness, authors also compared their results from previous research on applying the framework to a different learning environment and data type. Their results also confirmed previous research on the value of using eye-tracking data to assess student learning (Amershi and Conati, 2007).

Web mining aims to discover useful information and knowledge from the Web hyper-link structure, page contents, and usage data. Although Web mining uses many conventional data mining techniques, it is not purely an application of traditional data mining due to the semistructured and unstructured nature of the Web data and its heterogeneity. It has also developed many of its own algorithms and techniques. Liu has written a comprehensive text on Web data mining. Key topics of structure mining, content mining, and usage mining are covered both in breadth and in depth. His book brings together all the essential concepts and algorithms from related areas such as data mining, machine learning, and text processing to form an authoritative and coherent text. The book offers a rich blend of theory and practice, addressing seminal research ideas, as well as examining the technology from a practical point of view. It is suitable for students, researchers and practitioners interested in Web mining both as a learning text and a reference book. Lecturers can readily use it for classes on data mining, Web mining, and Web search. Additional teaching materials such as lecture slides, datasets, and implemented algorithms are available online (Liu, 2007).

2.3.5. Intelligent Systems Approaches using Clustering Techniques

A class learning diagnosis problem was investigated by embedding important concepts in a test and analyzing the results with a hierarchical coding scheme. Based on previous research, the part-of and type-of relationships among concepts were used to construct a concept hierarchy that may then be coded hierarchically. These approaches were implemented as an integrated module in a previously developed system and applied to

two real classroom datasets, the results of which show the practicability of this proposed method(Cheng et al., 2005).

Preschool children attending Head Start programs between 3 and 5 years of age, over 95% African-American were observed to determine physical proximity to peers as well as rates of visual attention given and received. Sociometric data were used to derive peer acceptance scores, peer friendships, and sociometric status classifications. Three subgroup types (high mutual proximity (HMP), lower mutual proximity (LMP), and ungrouped children) were identified through complete linkage hierarchical clustering and chi-square procedures from the proximity data. HMP subgroups tended to be larger, to have higher sociometric acceptance scores, and children in these subgroups had more reciprocated friendships than was true for the other subgroup types. Significant within-group preferences and out-group biases were observed for both HMP and LMP subgroups using measures of visual attention and sociometric choice data, but these were more marked for HMP subgroups. Results are consistent with previous ethological studies of affiliative structures in preschool classrooms and also show that methods of data collection and analysis from social ethology and child psychology research traditions are mutually informing (Heather, 2006; Santos et al., 2008).

2.3.6. Data Normalization

The curse of dimensionality is a damning factor for numerous potentially powerful machine learning techniques. Widely approved and otherwise elegant methodologies used for a number of different tasks ranging from classification to function approximation exhibit relatively high computational complexity with respect to dimensionality. This limits severely the applicability of such techniques to real world problems. Rough set theory is a formal methodology that can be employed to reduce the dimensionality of datasets as a preprocessing step to training a learning system on the data. A utility of the Rough Set Attribute Reduction (RSAR) technique to both supervised and unsupervised learning

was investigated in an effort to probe RSAR's generality. FuREAP, a Fuzzy-Rough Estimator of Algae Populations, which is an existing integration of RSAR and a fuzzy Rule Induction Algorithm (RIA), is used as an example of a supervised learning system with dimensionality reduction capabilities. A similar framework integrating the Multivariate Adaptive Regression Splines (MARS) approach and RSAR is taken to represent unsupervised learning systems (Shen and Chouchoulas, 2001). The goal of data normalization and standardization is to eliminate data redundancy with the aim of ensuring the integrity of data. In data fusion, score normalization is a step to make scores, which are obtained from different component systems for all documents, comparable to each other. It is an indispensable step for effective data fusion algorithms such as CombSum and CombMNZ to combine them. Four linear score normalization methods were evaluated, namely the fitting method, Zero-one, Sum, and ZMUV, through extensive experiments. Experimental results showed that the fitting method and Zero-one appeared to be the two leading methods (Zighed et al., 1997; Ho and Scott, 1997).

Zalmai developed a fairly large number of sets of global parametric sufficient optimality conditions under various generalized assumptions for a discrete min-max fractional programming problem involving arbitrary norms (Wu et al., 2006). Min-max control is a robust control, which guarantees stability in the presence of matched uncertainties. The basic min-max control is a static state feedback law. Recently, the applicability conditions of discrete static min-max control through the output have been derived (Zalmai, 2007). Results for output static min-max control are further extended to a class of output dynamic min-max controllers, and a general parametrization of all such controllers is derived. The dynamic output min-max control is shown to exist in many circumstances under which the output static min-max control does not exist, and usually allows for broader bounds on uncertainties. Another family of robust output min-max controllers, constructed from an asymptotic observer which is insensitive to uncertainties and a state min-max control, is derived. The latter is shown to be a particular case of the dynamic min-max control

when the nominal system has no zeros at the origin. In the case where the insensitive observer exists, it is shown that the observer-controller has the same stability properties as those of the full state feedback min-max control. Information system discretization widens the use of rough set theory. Rough set attribute discretization should maintain the consistency of knowledge base classification with less cut-points (Chmielewski and Jerzy, 1996; Sharav-Schapiro et al., 1999; Yang et al., 2007).

2.3.7. K-Means Algorithm

Adaptive k-means clustering algorithms have been used in several artificial neural network architectures, such as radial basis function networks or feature-map classifiers, for a competitive partitioning of the input domain. An enhancement of the traditional k-means algorithm was presented. It approximates an optimal clustering solution with an efficient adaptive learning rate, which renders it usable even in situations where the statistics of the problem task varies slowly with time. That modification was based on the optimality criterion for the k-means partition stating that: all the regions in an optimal k-means partition have the same variations if the number of regions in the partition is large and the underlying distribution for generating input patterns is smooth (Chinrungrueng and Sequin, 1995; Wagstaff et al., 2001; Bhatia, 2004; ?).

Subspace clustering is a challenging task in the field of data mining. Traditional distance measures fail to differentiate the furthest point from the nearest point in very high dimensional data space. To tackle the problem, minimal subspace distance was designed which measures the similarity between two points in the subspace where they are nearest to each other. It can discover subspace clusters implicitly when measuring the similarities between points. We use the new similarity measure to improve traditional k-means algorithm for discovering clusters in subspaces. By clustering with low-dimensional minimal subspace distance first, the clusters in low-dimensional subspaces are detected. Then by gradually increasing the dimension of minimal subspace distance, the clusters get refined

in higher dimensional subspaces (Ayaquica-Martinez et al., 2005; Zhao et al., 2006).

A research that aimed to assign weights to m clustering variables, so that k groups were uncovered to reveal more meaningful within-group coherence. A new criterion to be minimized was proposed, which is the sum of the weighted within-cluster sums of squares and the penalty for the heterogeneity in variable weights (Huh and Lim, 2009).

2.3.8. K-means Clustering Algorithm Time Complexity

Clustering techniques have become very popular in a number of areas, such as engineering, medicine, biology and data mining. A good survey on clustering algorithms can be found in (Lingras and Yao, 2002). The k-means method is an old but popular clustering algorithm known for its speed and simplicity. Until recently, however, no meaningful theoretical bounds were known on its running time. Arthur and Vassilvitskii demonstrate that the worst-case running time of k-means is superpolynomial by improving the best known lower bound (Vrahatis et al., 2002; Arthur and Vassilvitskii, 2005; Tian et al., 2005). One approach is based on Genetic Algorithms, and the other is an adaptation of K-means algorithm. Both the approaches have been successful in generating intervals of clusters. The efficiency of the clustering algorithm is an important issue when dealing with a large datasets. Lingras and Yao provides comparison of the time complexity of the two rough clustering algorithms (Abascal et al., 2006).

2.3.9. Clusters Evaluations

Some of the established approaches to evaluating text clustering algorithms for information retrieval show theoretical flaws. These flaws were analyzed and new evaluation measure to overcome them was introduced. Based on a simple yet rigorous mathematical analysis of the effect of certain parameters in cluster based retrieval, it was shown that certain conclusions drawn in the recent literature must be taken with a grain of salt (Huijsmans and Sebe, 2001).

Performance evaluations in Probabilistic Information Retrieval are often presented as Precision-Recall or Precision-Scope graphs avoiding the otherwise dominating effect of the embedding irrelevant fraction. However, precision and recall values as such offer an incomplete overview of the information retrieval system under study: information about system parameters like generality (the embedding of the relevant fraction), random performance, and the effect of varying the scope is missed. Two cluster performance graphs were presented. In those cases where complete ground truth is available (both cluster size and database size) the Cluster Precision-Recall (Cluster PR) graph and the Generality-Precision=Recall graph are proposed (Mehlitz et al., 2007).

3. Materials and Methods

This Chapter will address how goals and objectives will be met. Also where the various activities took place are explained. The adopted methodology steps are show in Figure 3.1.

As seen in Figure 3.1, The adopted methodology steps are:

1. Data collection: this task was achieved at first. In this step, the needed information was acquired.
2. Data Preprocessing: this task was achieved after data was acquired. Data preprocessing step consists of three parts which are:
 - Dimension Reduction.
 - Reclassification.
 - Data Standardization.
3. Data Clustering: this task is the main step in this research. Data clustering step includes the K-means Clustering Algorithm.
4. Data Analysis: this was the final step. Data Analysis consists of four parts which are:
 - Significant Attributes Selection.
 - Number of Clusters Selection.
 - Suitable Distance Function Selection.
 - Suitable Cluster Identification Selection.

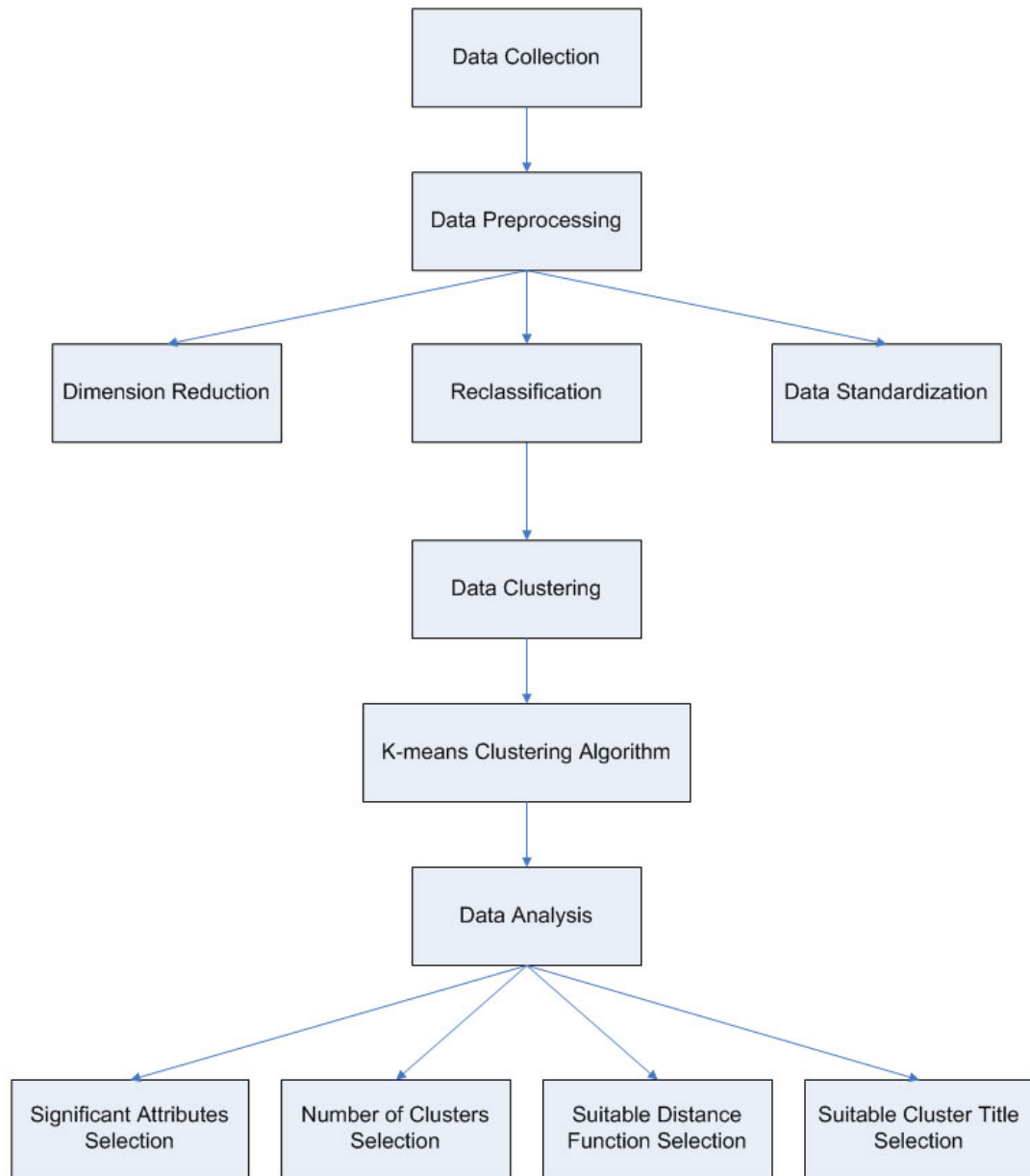


Figure 3.1: Materials and Methods (Methodology) steps Flow Chart

3.1. Data Collection

This is the task of gathering general information and resources. A sample of student data was acquired. This sample was the entire CS2 sections for the last Spring Semester (2007/2008). As seen in Figure 3.2, this sample consists of 2088 student records and 26 attributes as student's number, student's GPA, Faculty, the Computer Skills-1 mark and

others.

P	O	N	K	J	I	H	G	F	E	D	C	B	A	
English 1	Sem Crd hrs.	All Credit hrs.	compQ	Nationality	Sex	Specialization	Faculty	CS-1	GPA	Sem-GPA	Cert. Nat	HSGE	Student No.	1
0.0	3	087	بلا	1	1	090	23	9.90	2.42	1.00	01	718	2	
4.0	15	024	دلج	1	2	040	12	8.80	3.77	3.83	01	944	3	
8.8	13	027	دلج	1	2	110	05	8.80	3.80	3.89	01	970	4	
8.8	6	009	دلج	1	2	071	16	8.80	2.83	2.75	01	830	5	
3.0	12	021	دلج	1	2	010	22	8.80	3.00	3.37	01	772	6	
8.8	12	060	دلج	1	2	010	16	8.80	2.97	2.87	01	879	7	
8.8	6	023	دلج	1	2	110	05	8.80	3.74	4.00	01	918	8	
3.0	3	033	دلج	1	1	040	16	8.80	1.86	2.00	01	745	9	
8.8	3	009	دلج	1	1	040	09	8.80	3.50	3.00	01	947	10	
9.9	12	012	دلج	1	1	011	20	8.80	2.88	2.87	01	833	11	
1.0	11	102	راسب	1	1	010	07	1.00	1.97	1.00	01	666	12	
2.5	18	123	راسب	1	2	060	04	3.00	2.43	2.58	01	664	13	
1.5	15	123	راسب	1	2	010	17	1.00	2.20	1.60	01	772	14	
9.9	12	144	دلج	1	1	010	09	8.80	2.87	3.08	01	907	15	
8.8	17	141	دلج	5	1	070	09	8.80	2.82	2.44	01	938	16	
4.0	15	096	دلج	1	2	051	22	8.80	2.92	2.30	01	821	17	
4.0	16	091	دلج	4	2	040	12	8.80	2.27	2.06	01	839	18	
3.5	18	105	راسب	1	2	010	23	3.50	3.47	3.00	01	879	19	
2.0	18	111	راسب	1	2	010	23	2.00	3.16	3.08	01	877	20	
3.0	15	099	راسب	1	2	010	23	2.50	3.30	3.10	01	882	21	
3.5	15	096	راسب	1	2	051	22	2.50	2.75	2.80	01	919	22	
4.0	19	092	راسب	1	2	010	04	4.00	3.79	3.84	01	897	23	
2.5	18	099	راسب	1	2	010	04	2.50	2.32	2.33	01	808	24	
3.5	18	096	دلج	1	2	041	08	8.80	3.23	3.50	01	895	25	
1.0	18	090	راسب	1	2	061	08	2.00	2.10	1.25	01	651	26	
4.0	13	100	دلج	1	2	020	09	8.80	3.64	3.23	01	914	27	
3.0	9	051	راسب	1	1	010	16	1.50	1.85	0.66	01	655	28	

Figure 3.2: A snapshot of the Students' Data records.

3.2. Data Preprocessing

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Data preprocessing transforms the data into a format that will be more easily and effectively processed for our clustering purpose. Some of the important tasks in data preprocessing include:

1. Data Interpolation.
2. Smoothing of noisy data.
3. Identification and removal of noisy data records.
4. Data Standardization and Normalization.

Also, in the experiments, there were two kinds of noisy data:

1. Ones were because of errors in data recording; those were obvious, so they were deleted.
2. Others were very far away from all other data points in the same clusters; those data points were noticed over many iterations and then the decision was either to remove them or reassigning them to other closest clusters.

Each attribute in the sample data, that is shown in Figure 3.2, has different range of magnitude, as an example, the High Schools General Exam (HSGE) attribute's range is (0 - 1000) while the Cumulative Grade Point Average (CGPA) attribute's range is (0 - 4). These two attributes, as an example, have a huge difference in their ranges. In order to be able to test such attributes, within the same datasets, using K-means Clustering Algorithm, we need to make normalization and standardization on each attribute of our sample.

The goal of data normalization and standardization is to eliminate data redundancy with the aim of ensuring the integrity of data. Data preprocessing is a very important and time consuming task in data mining field. Preprocessing is needed in our algorithm in order to be sure that our data is cleaned. Our aim in preprocessing task is to get a suitable target dataset.

The preprocessing tasks were as follows:

1. Dimension Reduction.
 - Remove unwanted records.
 - Remove unwanted attributes.
2. Reclassification
 - Converting text fields to digits.
 - Categorize attributes in proper datasets.

3. Data Standardization

- Range standardization.
- Values standardization (discrete and continuous).

3.3. Data Clustering

In some applications, the data have no class attributes. The user wants to explore the data to find some essential and intrinsic structures in them. Clustering is one of the methods that are using for finding such structures. It organizes data points into similar groups, called clusters such that the data points in the same cluster are similar to each other and data points in different clusters are very different from each other.

Clustering is the assignment of objects into groups (called clusters) so that objects from the same cluster are more similar to each other than objects from different clusters. Often similarity is assessed according to a distance measure, distance functions will be explained in details in subsection 3.5.3. Clustering is a common technique for statistical data analysis, Data Analysis will be explained in details in section 3.5.

3.4. K-means Clustering Algorithm

K-means Clustering Algorithm, simply speaking, is an algorithm that classifies and groups objects based on some attributes/features into K number of group. K is positive integer number. The grouping is done by minimizing the distances between data and the corresponding cluster centroid. Thus, the purpose of K-mean clustering is to classify the data. The k-means algorithm is the best known partitioning clustering algorithm. It is also the most widely used among all clustering algorithms due to its simplicity and efficiency. In practice, several k values are tried and the one that gives the most desirable result is selected, K-means Clustering Algorithm Flow Chart is shown in Figure3.3.

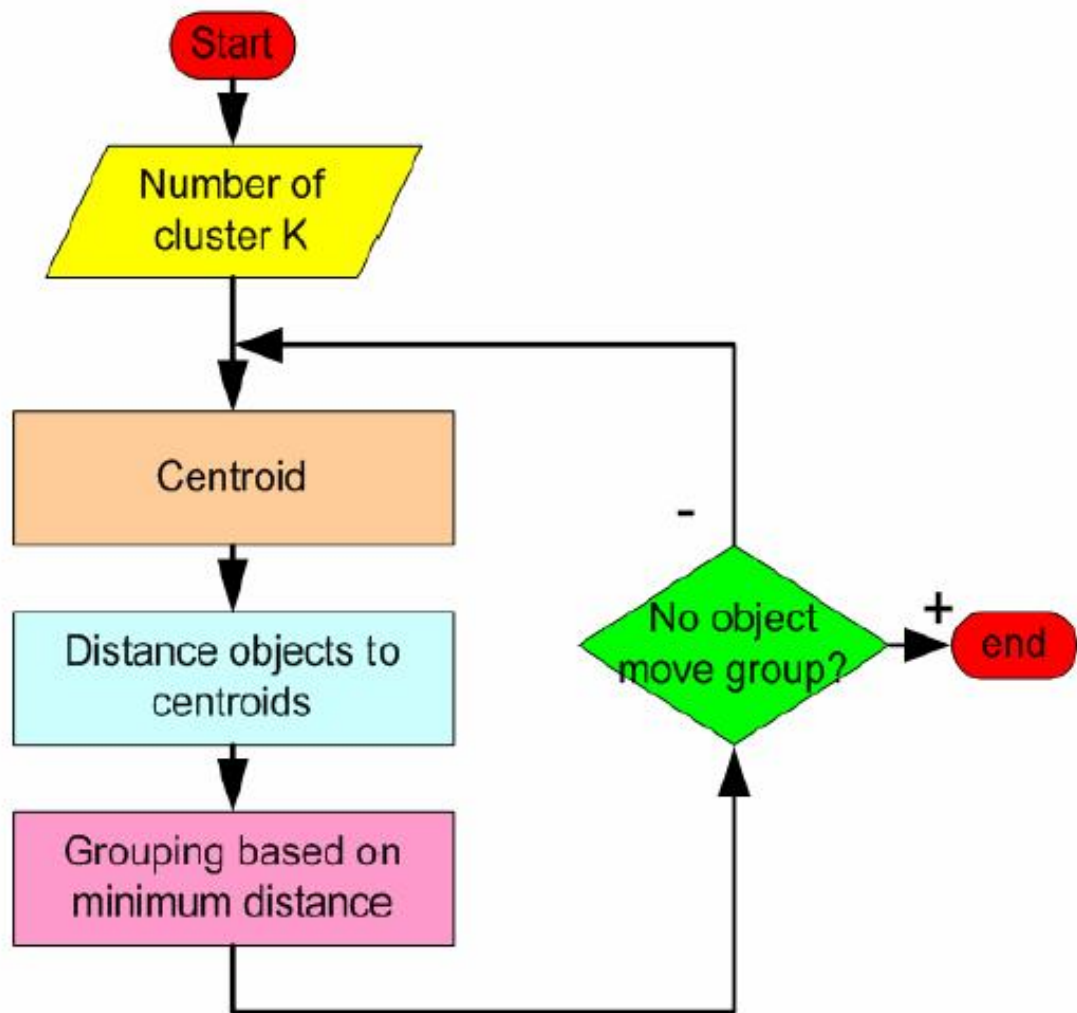


Figure 3.3: K-means Clustering Algorithm Flow Chart

Given a set of data points and the needed number of clusters (K), this algorithm iteratively partitions the data into k clusters based on a distance function. At the beginning, the algorithm randomly selects k data points as the centroids. It then computes the distance between each centroid and every data point. Each data point is assigned to the closest centroid. A centroid and its data points represent a cluster. Once all the data points in the data are assigned, the centroid for each cluster is recomputed using the data points in the current cluster. This process repeats until a stopping criterion is met. K-means Clustering Algorithm is shown in Figure 3.4.

The stopping criterion may be one of the following:

1. no (or minimum) re-assignments of data points to different clusters.
2. no (or minimum) change of centroids.

Algorithm k -means(k, D)

```

1  choose  $k$  data points as the initial centroids (cluster centers)
2  repeat
3      for each data point  $\mathbf{x} \in D$  do
4          compute the distance from  $\mathbf{x}$  to each centroid;
5          assign  $\mathbf{x}$  to the closest centroid      // a centroid represents a cluster
6      endfor
7      re-compute the centroid using the current cluster memberships
8  until the stopping criterion is met

```

Figure 3.4: K-means Clustering Algorithm

We were testing datasets using MATLAB ®2008b (The MathWorks, 2008). MatLab provides different toolboxes; one of those toolboxes is the Statistics Toolbox. Statistics Toolbox provides many functions that make the clustering task much easier for the user; one of those functions is the `kmeans` function. The function `kmeans` partitions data into k mutually exclusive clusters, and returns the index of the cluster to which it has assigned each observation, MatLab `kmeans` function usage is shown in formula (3.1).

$$IDX = kmeans(X, k), \quad (3.1)$$

where X is a matrix. `kmeans` function partitions the points in the data matrix X into K clusters. `kmeans` function returns an n -by-1 vector IDX containing the cluster indices of each point. By default, `kmeans` uses squared Euclidean distance function. Distance functions will be explained in details in subsection 3.5.3.

3.5. Data Analysis

Data analysis is an important stage of the research process. Data analysis is a process of gathering, modeling, and transforming data. Our goal is highlighting useful information, suggesting conclusions, and supporting decision making. Data mining is a particular data analysis technique that focuses on modeling and knowledge discovery for predictive rather than purely descriptive purposes.

3.5.1. Significant Attributes Selection

After Data preprocessing, some attributes were excluded while others were selected according to their influence on data clustering; this will be explained in details subsequently in chapter 4.

At the time of testing datasets whose most significant attributes were selected, we found that our selected attributes were divided into two types; Continuous and Discrete. It was also found that the continuous attributes (HSGE and CGPA) overpowered the discrete ones (CS1 and English1) in data clusters. So the decision was to categorize continuous attributes according to data normalization method that is called: Min-Max. Min-Max is a linear transformation of the original range into a newly specified data range, as seen in equation (3.2).

$$\frac{Max - Min}{n} \quad (3.2)$$

Where Max: is the highest value in the range, Min: is the lowest value in the range and n: is the number of classes in the new range.

3.5.2. Number of Clusters (Groups) Selection

To identify the most suitable number of clusters in the datasets, Silhouette Coefficients was used. Silhouette is used to plot clustered data. It shows which objects lie well within the cluster and which ones are merely somewhere in between clusters. The entire Silhouette plot shows the silhouettes of all clusters next to each other, so that the quality

of clusters can be compared. In order to determine the right number of clusters, the Silhouette Coefficients are used. For a given point i in a cluster A , the silhouette of I , $s(i)$, is shown in Equation (3.3) as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (3.3)$$

where $a(i)$ is the average dissimilarity of i -data point to all other data points in the same cluster (the cluster to which i belongs). And $b(i)$ is the minimum of average dissimilarity of i - data point to all data points in other cluster (in the closest cluster to which i belongs).

It is followed from Formula (3.3) that $-1 \leq s(i) \leq 1$ i.e. $s(i)$ lies somewhere between -1 and 1 . The value of $s(i)$ has three cases:

1. If silhouette value is close to 1, it means that the sample data is "well-clustered" and it was assigned to a very appropriate cluster.
2. If silhouette value is about zero, it means that the sample data could be assigned to another closest cluster as well, and the sample data lies equally far away from both clusters.
3. If silhouette value is close to -1 , it means that the sample data is "misclassified" and is merely somewhere in between the clusters.

The overall average silhouette width for the entire plot is simply the average of the $s(i)$ for all objects in the whole dataset. The largest overall average silhouette indicates the best clustering (number of cluster). Therefore, the number of cluster with maximum overall average silhouette width is taken as the optimal number of the clusters, Silhouette plot examples are shown in Figures 3.5 and 3.6.

3.5.3. Suitable Distance Function Selection

There are many different distance measures for K-means Clustering Algorithm, depending on the kind of data that is being clustered. Each cluster in the partition is defined

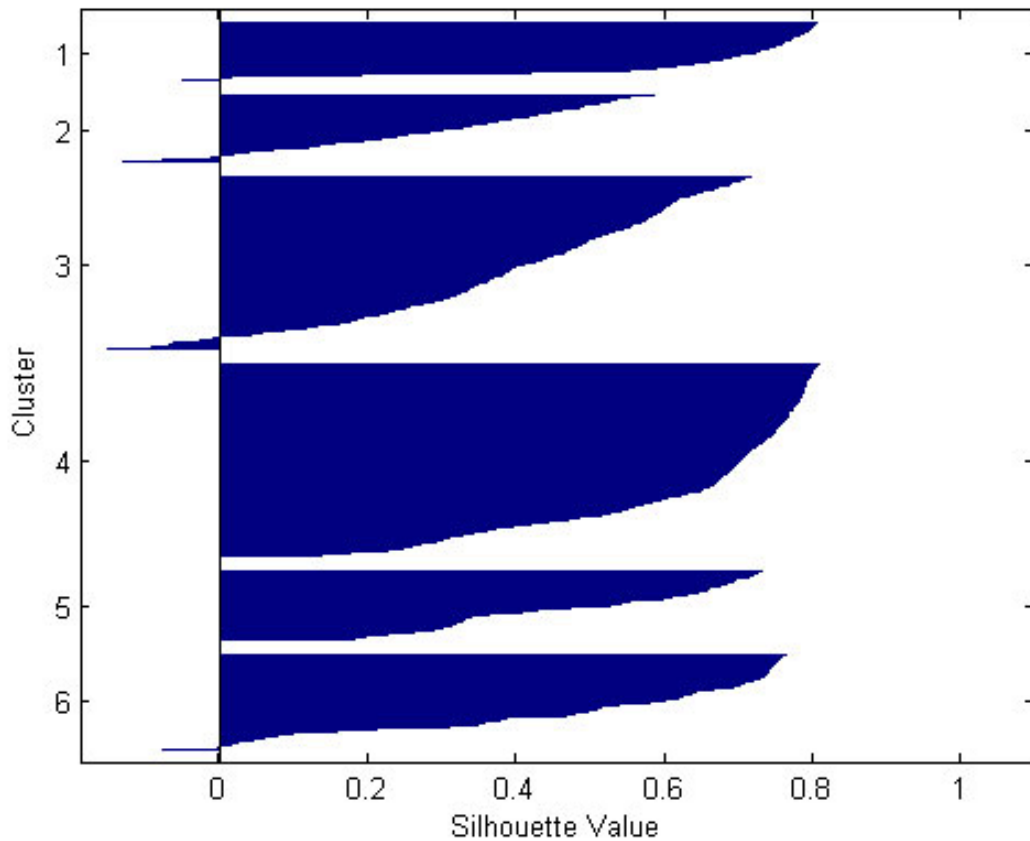


Figure 3.5: Silhouette Plot Example1.

by its member objects and by its centroid, or center. The centroid for each cluster is the point to which the sum of distances from all objects in that cluster is minimized. K-means computes cluster centroids differently for each distance measure, to minimize the sum with respect to the measure that has been specified.

According to the tested data, different Silhouettes showed that the Square Euclidean was the most suitable distance function in which it was the one that gave us the most accurate values in data clustering. The standard Euclidean distance is used for numeric attributes. The standard Euclidean distance is squared in order to place progressively greater weights on data points that are further apart. The standard Euclidean distance is

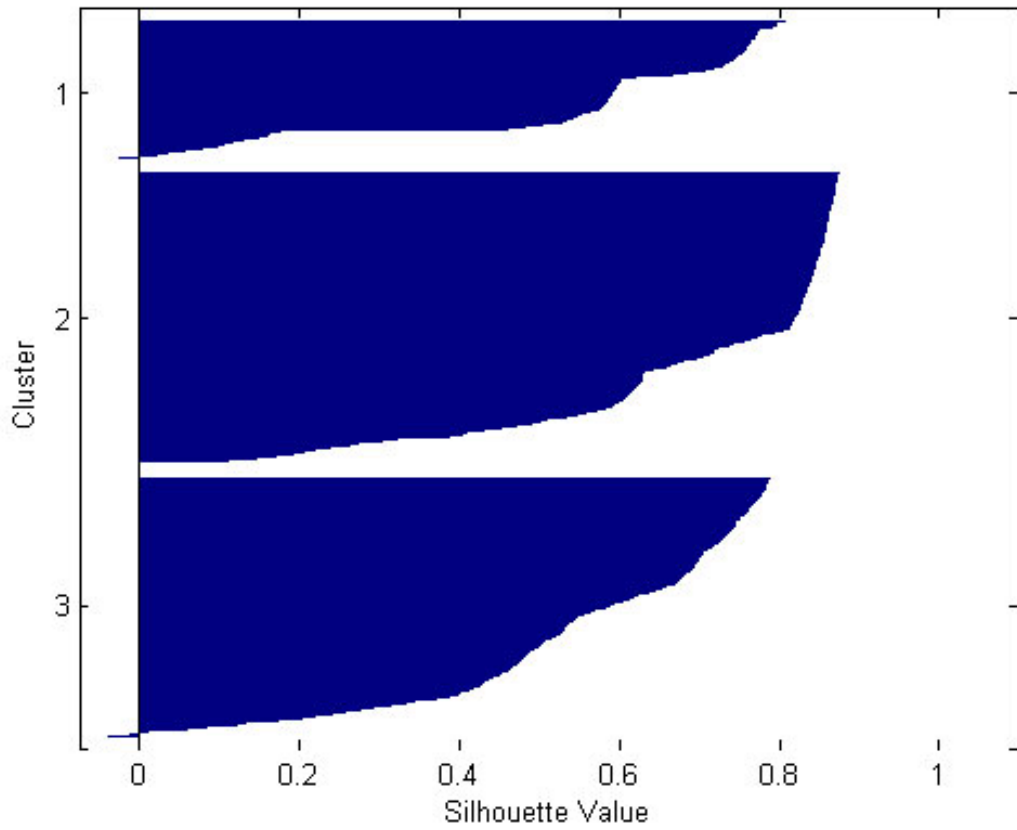


Figure 3.6: Silhouette Plot Example2.

shown in Formula (3.4).

$$\text{dist}(X_i, X_j) = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ir} - x_{jr})^2. \quad (3.4)$$

The most commonly used distance functions for numeric attributes are the Euclidean distance and Manhattan (city block) distance. Both distance measures are special cases of a more general distance function called the Minkowski distance. We use $\text{dist}(x_i, x_j)$ to denote the distance between two data points of r dimensions. The Minkowski distance is shown in Formula (3.5):

$$\text{dist}(X_i, X_j) = (|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \dots + |x_{ir} - x_{jr}|^h)^{\frac{1}{h}}, \quad (3.5)$$

where h is a positive integer.

If $h = 2$, it is the Euclidean distance which is shown in Formula (3.6),

$$\text{dist}(X_i, X_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ir} - x_{jr})^2}. \quad (3.6)$$

If $h = 1$, it is the Manhattan distance which is shown in Formula (3.7),

$$\text{dist}(X_i, X_j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ir} - x_{jr}|. \quad (3.7)$$

3.5.4. Suitable Cluster Identification Selection

After modeling our data, we need to Find a suitable naming and identification for each cluster. This will be explained in details in the Experimental and Theoretical Results chapter (Chapter 4), in the Clusters Representation section (Section 4.4).

3.6. Clusters Representation

Once a set of clusters is found, the next task is to find a way to represent the clusters. The resulting clusters need to be represented in a Comprehensible and reasonable way in which it facilitates the evaluation of the resulting clusters. There are many ways for the representation of clusters. We used the so-called "classification models". In this method, we treat each cluster as a class. That is, all the data points in a cluster are regarded as having the same class label, e.g., the cluster ID.

3.7. Cluster Evaluation

After finding the appropriate number of clusters, these clusters have to be assessed. When using data clustering algorithms, nobody knows what the correct clusters are given a dataset. Thus, the quality of a clustering is much harder to evaluate. That means that metrics must be used to measure how good the clustering techniques are and to evaluate the quality of a set of dataset clusters. To evaluate the clusters, some of the commonly

used evaluation methods were used, which are: User Inspection, Purity, Precision and Recall. Recall and precision are more useful measures that have a fixed range and are easy to compare across clusters. A combination of both precision and recall that measures the extent to which a cluster contains only objects of a particular class and all objects of that class.

3.7.1. User Inspection

User Inspection: users inspect the resulting clusters and score them. Since this process is subjective, we took the average of the scores from all users as the final score of the clustering. This manual inspection is obviously time consuming task. It is subjective as well. However, in most applications, some level of manual inspection is necessary because no other existing evaluation methods are able to guarantee the quality of the final clusters. It should be noted that direct user inspection may be easy for certain types of data, but not for others. In order to be sure of our evaluation, we've tried other evaluation methods; those will be explained in details in next subsection subsequently.

3.7.2. Purity

Purity: is the measure of the extent that a cluster contains only one class of data. The purity of each cluster is computed with equation (3.8), (Liu, 2007).

$$purity(D_i) = \max_j(Pr_i(c_j)), \quad (3.8)$$

where $Pr_i(c_j)$ is the proportion of class c_j data points in cluster i or D_i .

The total purity of the whole clustering (considering all clusters) is show in equation (3.9), (Liu, 2007).

$$purity_{total}(D) = \sum_{i=1}^k \frac{|D_i|}{|D|} \times purity(D_i). \quad (3.9)$$

3.7.3. Precision

Precision: can be seen as a measure of exactness or fidelity i.e. how many of the data points in this cluster belong there? Precision is the fraction of a cluster that consists of data points of a specific class. Precision is calculated as the portion of cluster j that is a member of class i , thus measuring how homogenous cluster j is with respect to class i , as seen in Equation (3.10), (Rosenberg and Hirschberg, 2007; Davis and Goadrich, 2006).

$$precision(c_i, l_j) = \frac{n_{ij}}{n_j}, \quad (3.10)$$

where c_i is the data points' class and l_j is the data points' cluster. The n_{ij} value is the count of class c_i 's data points in cluster l_j . And the n_j value is the total number of data points in cluster l_j . Precision has a fixed range: 0.0 to 1.0 (or 0% to 100%). The key in finding better clustering is to increase precision without sacrificing recall (Yang et al., 2003).

3.7.4. Recall

Recall: can be seen as a measure of completeness i.e. did all of the data points that belong in this cluster make it complete? Recall is the extent to which a cluster contains all objects of a specific class. Recall is calculated as the portion of items from class i that are present in cluster j , thus measuring how complete cluster j is with respect to class i , as seen in Equation (3.11), (Rosenberg and Hirschberg, 2007; Egghe, 2008).

$$recall(c_i, l_j) = \frac{n_{ij}}{n_i}, \quad (3.11)$$

where c_i is the data points' class and l_j is the data points' cluster. The n_{ij} value is the count of class c_i 's data points in cluster l_j . And the n_i value is the total number of data points in class c_i . Recall has a fixed range: 0.0 to 1.0 (or 0% to 100%). Good clusters

must have a high recall to be admissible in most applications (Yang et al., 2003).

4. Experimental and Theoretical Results

Experimental results can lead to more questions about the problem and issue under study. First, as results were found, additional insights often occurred. So in a way, the act of finding results led us to further analysis and interpretation. And second, results leave a permanent record of the experiment that can be used by others. The adopted Experimental and Theoretical steps are shown in Figure 4.1.

After data interpretation, the following was recognized:

1. Recognizing the number of clusters and identifying each one of them.
2. Recognizing the most important attributes that influence the student's section allocation.
3. Recognizing the most suitable Distance Function to be used.

As seen in Figure 4.1, The adopted Experimental and Theoretical steps are:

1. Data Preprocessing: This task was achieved first before testing. Data preprocessing step consists of three parts which are:
 - Dimension Reduction.
 - Reclassification.
 - Data Standardization.
2. Datasets Testing: Many experimental tests have been done on many datasets, using MATLAB ®2008b (The MathWorks, 2008).
3. Data Interpretation: After data interpretation, the following was recognized:
 - Significant Attributes: These are the attributes that have the most influence and impact on the sample data.

- Suitable Number of Clusters: This step was done in respect of Silhouette Values.
 - Suitable Distance Function: In this step the most appropriate grouping was found according to the findings.
4. Clusters Representation: This step is the task to find a way to represent the resulting clusters.
5. Clusters Evaluation: After Clusters Evaluation, the following were done:
- Suitable Grouping for CS2 Students: This step was done in respect of the Confusion matrices. Each cluster has its own Confusion matrix.
 - CS2 Students' Distribution: This distribution is done according to the pre-selected attributes; HSGE, CGPA, CS1 and English1.

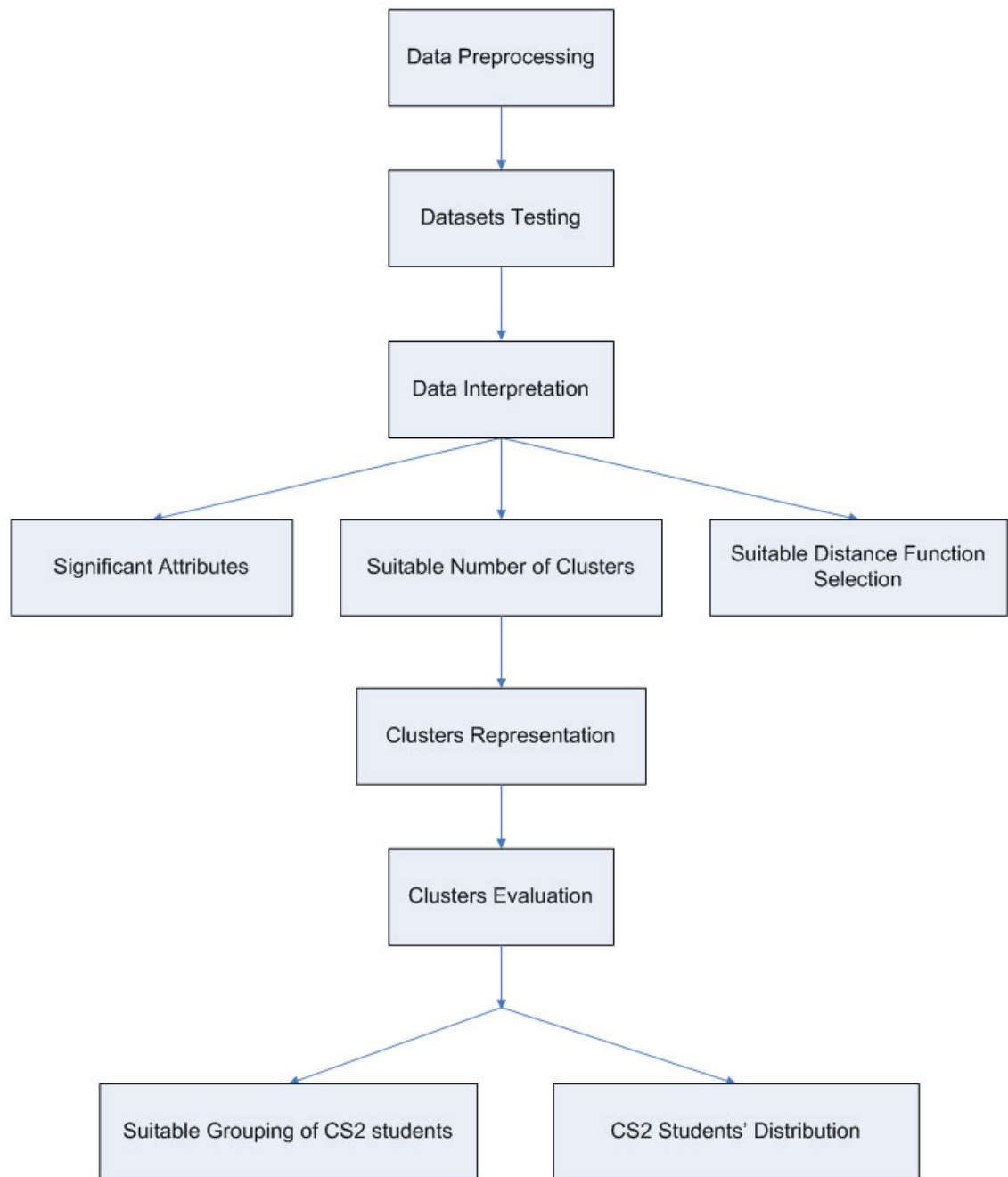


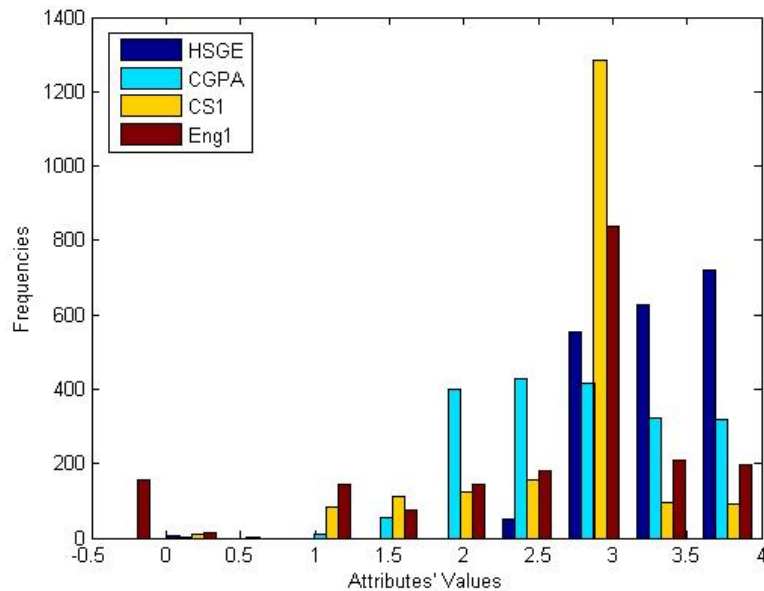
Figure 4.1: Experimental and Theoretical steps Flow Chart

4.1. Data Preprocessing

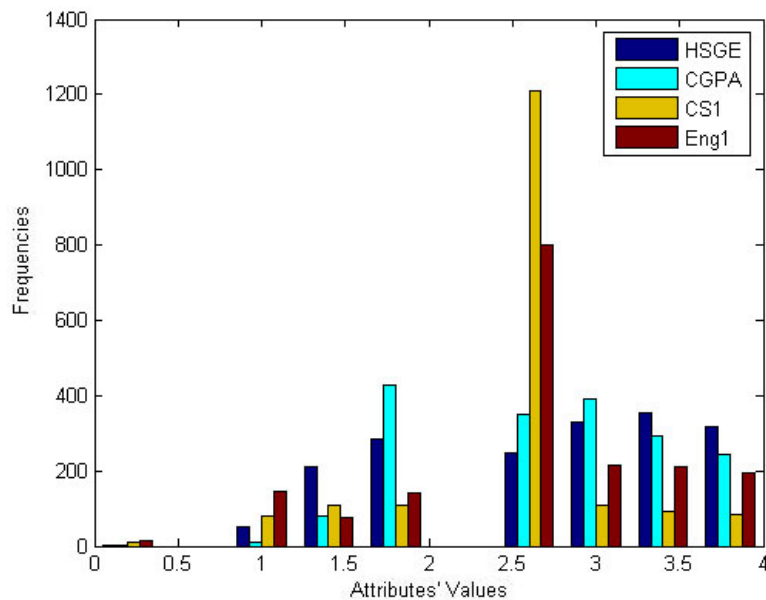
The preprocessing tasks were as follows:

1. Dimension Reduction.
2. Reclassification
3. Data Standardization

The first task was the dimension reduction. This task is needed in order to increase the efficiency of k-means Clustering algorithm. First, all unwanted records in the sample were removed. About 132 records in CS1 attribute that have no-data (null) values were found. We believe that the CS1 attribute is one of the most important attributes because it is the main prerequisite course of CS2, so we decided to remove such records from the sample in order to preserve data integrity. At that time, '-0.5' values to those records that have no-data (null) values in English1 attribute were given. After running many tests on a number of datasets, we found that English1 attribute has a big impact on data clustering. We decided to cross out the records that have no-data (null) values, just as we did in CS1 attribute in order to preserve the data standardization in the attributes, as seen in Figure 4.2(b). Those were about 156 records, as seen in Figure 4.2(a). Scaling data values in such similar range, [0 - 4], prevents outweighing attributes with large range to overpower the other ones.



(a) With '-0.5' values in English1 attribute.



(b) After preprocessing and discretization excluding '-0.5' values.

Figure 4.2: Different datasets before and after preprocessing and discretization

Another misleading value we found in CS1 and English1 attributes was the exempt ('free from' or 'pass') value, which denotes students who passed the Computer Skills Qualification or English Qualification exams respectively. Such records couldn't be crossed

out, because simply they were somehow an indicator on student levels in Computer Skills Qualification and English Qualification exams. Also those records were about 1172 records in CS1 attribute and 623 records in English1 attribute, which means that they were about 60% of the data in CS1 attribute, as seen in Figure 4.3.

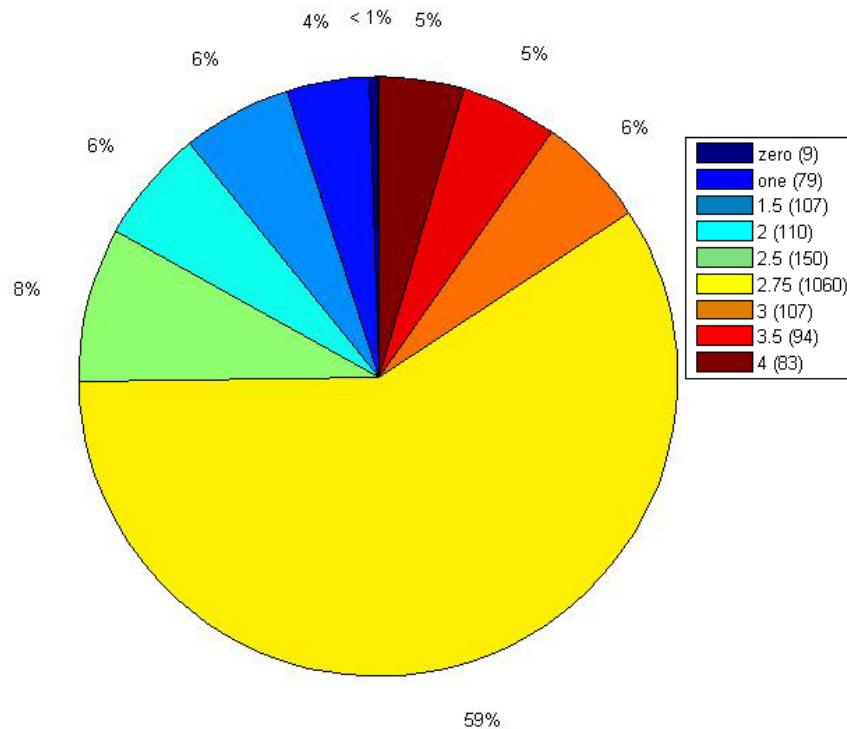


Figure 4.3: Pie Graph for the whole (CS1) Records.

At first, the value '5' was given to these records. It was noticed that '5' values is skewing data points toward high scores clusters, for more figures and information comparisons see Appendix A. So, the decision was to give those records the value '2.75'. This number was chosen in order not to affect CS1 and English1 attributes' intervals correspondingly. For example, if the value '2.5' was given, the differences between the real range and the pass values ones would not be differentiated. However the '2.75' value is a middle one between the high score 'A' and the failed one 'F' that is not affecting any of the grades.

4.2. Datasets Testing

After having performed data cleansing and achieving a collection of data of clean datasets, the datasets are ready now for testing. The total number of records, after data preprocessing and data cleansing is done, is 1799 records, a snapshot of the data is shown in Figure 4.4.

	A	C	F	G	H	J	L	M	N	O	Q	U	V	X
1	Student No.	HSGE	Cert. Nat.	Sem-GPA	GPA	CS-1	Faculty	Specialization	Sex	Nationality	CompQ	All Credit hrs.	Sem Crd hrs.	English 1
2		3.776	01	3.83	3.77	2.75	12	040	2	1	1	024	15	4.0
3		3.880	01	3.69	3.80	2.75	05	110	2	1	1	027	13	2.75
4		3.320	01	2.75	2.83	2.75	16	071	2	1	1	009	6	2.75
5		3.088	01	3.37	3.00	2.75	22	010	2	1	1	021	12	3.0
6		3.516	01	2.87	2.97	2.75	16	010	2	1	1	060	12	2.75
7		3.672	01	4.00	3.74	2.75	05	110	2	1	1	023	6	2.75
8		2.980	01	2.00	1.86	2.75	16	040	1	1	1	033	3	3.0
9		3.788	01	3.00	3.50	2.75	09	040	1	1	1	009	3	2.75
10		2.664	01	1.00	1.97	1.00	07	010	1	1	0	102	11	1.0
11		2.656	01	2.58	2.43	3.00	04	060	2	1	0	123	18	2.5
12		3.088	01	1.60	2.20	1.00	17	010	2	1	0	123	15	1.5
13		3.752	01	2.44	2.82	2.75	09	070	1	5	1	141	17	2.75
14		3.284	01	2.30	2.92	2.75	22	051	2	1	1	096	15	4.0
15		3.356	01	2.06	2.27	2.75	12	040	2	4	1	091	16	4.0
16		3.516	01	3.00	3.47	3.50	23	010	2	1	0	105	18	3.5
17		3.508	01	3.08	3.16	2.00	23	010	2	1	0	111	18	2.0
18		3.528	01	3.10	3.30	2.50	23	010	2	1	0	099	15	3.0
19		3.676	01	2.80	2.75	2.50	22	051	2	1	0	096	15	3.5
20		3.588	01	3.84	3.79	4.00	04	010	2	1	0	092	19	4.0
21		3.232	01	2.33	2.32	2.50	04	010	2	1	0	099	18	2.5
22		3.580	01	3.50	3.23	2.75	08	041	2	1	1	096	18	3.5
23		2.604	01	1.25	2.10	2.00	08	061	2	1	0	090	18	1.0
24		3.656	01	3.23	3.64	2.75	09	020	2	1	1	100	13	4.0
25		2.620	01	0.66	1.85	1.50	16	010	1	1	0	051	9	3.0
26		2.948	01	2.16	2.24	3.00	08	041	2	1	0	099	18	1.0
27		3.388	05	3.25	2.80	4.00	16	020	1	5	0	069	18	2.5
28		2.932	01	2.30	2.00	2.75	22	051	2	1	1	054	15	2.5
29		2.968	01	2.00	1.97	2.50	06	030	2	1	0	051	15	3.5
30		2.720	01	2.37	2.17	2.50	06	060	2	1	0	046	12	4.0
31		3.468	05	3.15	3.40	3.50	04	060	1	5	0	073	16	2.75

Figure 4.4: A snapshot of Students Data records after preprocessing task is completed.

Many experimental tests have been done on many datasets, using MATLAB ®2008b (The MathWorks, 2008). The purpose of running that much of tests on many datasets is proving the proposed solution and the proposed explanation about how things are. This can be done via repeated experimental observations that could lead us to good results. It is often unclear with experimental tests, how wide their scope is. Thus, it is impossible to tell from a small set of tests' results whether these results are particular or even are more widely applicable, even though the algorithm is completely specified and applied. For this reason, many experimental tests have been done on many datasets.

4.3. Data Interpretation

4.3.1. Significant Attributes

After Data preprocessing and data testing, some attributes were excluded while others were selected according to their influence on data clustering.

After running many tests on a number of datasets, we could find the most important attributes in the sample according to their influence and impact on data clustering, which were as follows:

1. High Schools General Exam (HSGE), which gives us an indicator on the background knowledge of each student.
2. Cumulative Grade Point Average (CGPA), which gives us a long term indicator (general) on students' levels in their specializations.
3. Computer Skills 1 (CS1), which gives us an indicator on each student's level in computer skills.
4. English1, which gives us an indicator on each student's level in English.

Many experiments have been done on many dataset with different attributes. Some of the important experiments are explained in details in Appendix A, B and C. Some of the excluded attributes were the Grade Point Average (GPA) and English II (English2). The GPA attribute was excluded because the CGPA attribute overpowers it. CGPA, as was mentioned, gives a long term indicator on students' levels in their specializations. So, there is no need for the GPA attribute to be included which gives us an indicator on students' levels at one semester that could be shown by the long term CGPA. Also we excluded the English2 attribute because many students haven't taken it yet and English1 attribute was enough to indicate each student's level in English. Also the influence of English1 attribute on data overpowers the English2 attribute's one.

4.3.2. Suitable Number of Clusters

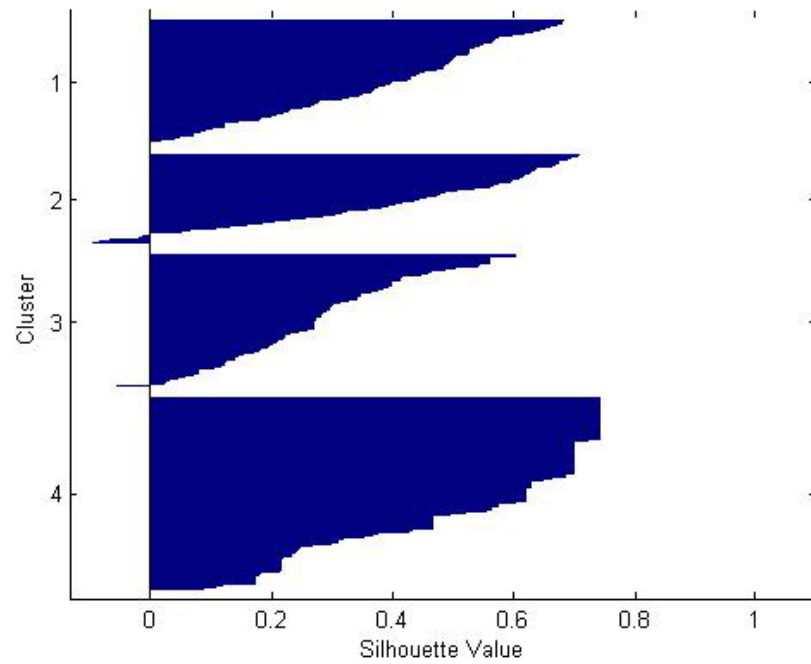
After running many investigations, the most appropriate grouping according to the findings was found. The last tested dataset, Experiment No.2, consists of 1799 records and four attributes (HSGE, CGPA, CS1 and English1). According to its Silhouettes, Silhouette was explained in details in chapter 3, the most appropriate number of clusters (groups) was $k=4$, as seen in Figures 4.11(a), 4.11(b), 4.6(a) and 4.6(b). Also $k=6$ clusters Silhouette was suitable, as seen in figure 4.6(a). Excluding this case ($k=6$) is explained in details in Appendix B.

The Silhouette value has three cases:

1. If silhouette value is close to 1, it means that the sample data is "well-clustered" and it was assigned to a very appropriate cluster.
2. If silhouette value is about zero, it means that the sample data could be assigned to another closest cluster as well, and the sample data lies equally far away from both clusters.
3. If silhouette value is close to -1, it means that the sample data is "misclassified" and is merely somewhere in between the clusters.

The overall average silhouette width for the entire plot is simply the average of the Silhouette values for all objects in the whole dataset. The largest overall average silhouette indicates the best clustering (number of cluster). Therefore, the number of cluster with maximum overall average silhouette width is taken as the optimal number of the clusters. Silhouettes that follows theses rules were those of $k=4$, as seen in Figure 4.11(a). That was the reason of choosing the most appropriate number of clusters (groups) which were 4.

DS11NK4

(a) $k=4$

DS11NK5

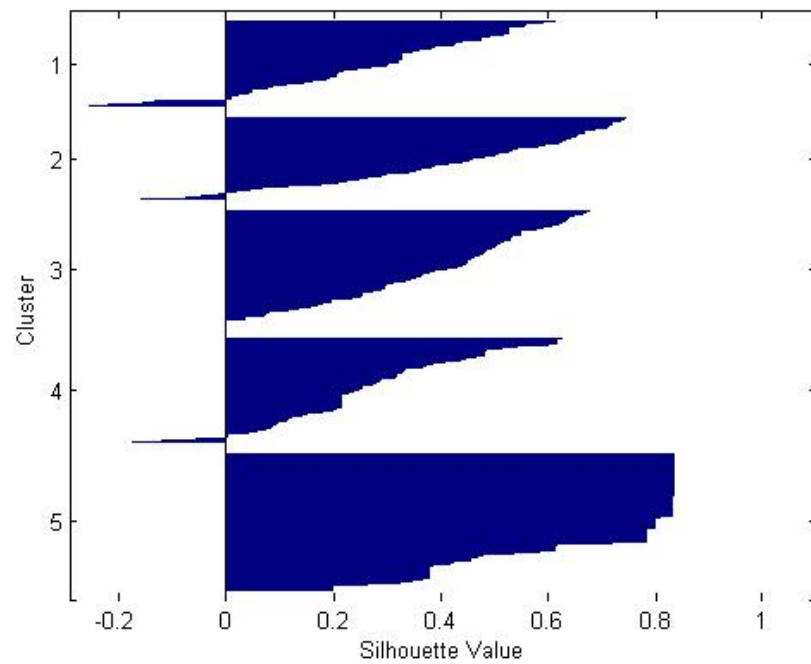
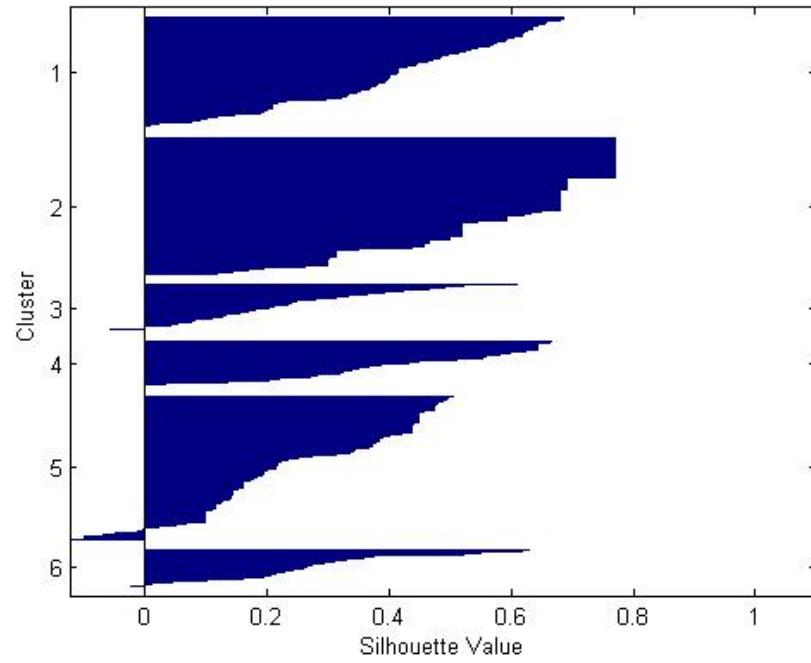
(b) $k=5$

Figure 4.5: Square Euclidean Distance Function Silhouettes for Experiment No.2 with different number of clusters (K).

DS11NK6

(a) $k=6$

DS11NK7

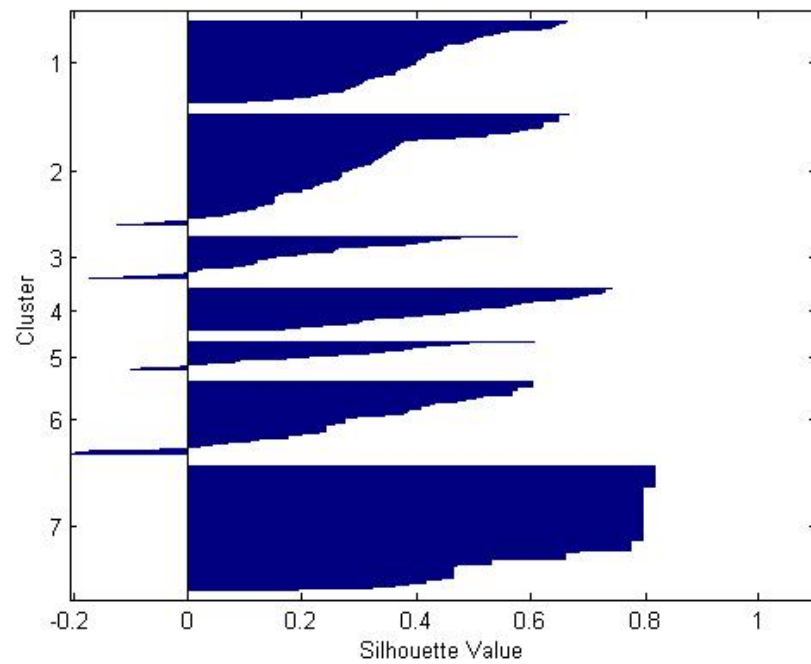
(b) $k=7$

Figure 4.6: Square Euclidean Distance Function Silhouettes for Experiment No.2 with different number of clusters (K).

4.3.3. Suitable Distance Function

Experiment No.1

One of the experiments, as an example, that was tested on a dataset that consist of 1955 records and three attributes which are: CGPA, CS1 and English1. Figure 4.7(a) and Figure 4.7(b) show us the Manhattan Distance Function for Experiment No.1 with $k=4$ and $k=5$ respectively.

Figure 4.8(a) and Figure 4.8(b) show us the Euclidean Distance Function for Experiment No.1 with $k=4$ and $k=5$ respectively.

Figure 4.9(a) and Figure 4.9(b) show us the Square Euclidean Distance Function for Experiment No.1 with $k=4$ and $k=5$ respectively. It is obvious from those figures that the Square Euclidean Distance Function for Experiment No.1 gave us the most accurate values in data clustering.

Experiment No.2

One of the other experiments that had proved the chosen of Square Euclidean Distance Function is Experiment No.2. It was tested on a dataset that consist of 1799 records and four attributes which are: HSGE, CGPA, CS1 and English1. Figure 4.10(a) and Figure 4.10(b) show us the Euclidean Distance Function for Experiment No.2 with $k=4$ and $k=5$ respectively. Figure 4.11(a) and Figure 4.11(b) show us the Square Euclidean Distance Function for Experiment No.2 with $k=4$ and $k=5$ respectively. It is obvious from those figures that the Square Euclidean Distance Function for Experiment No.2 gave us the most accurate values in data clustering. Silhouettes is used for comparisons on the selected distance function.

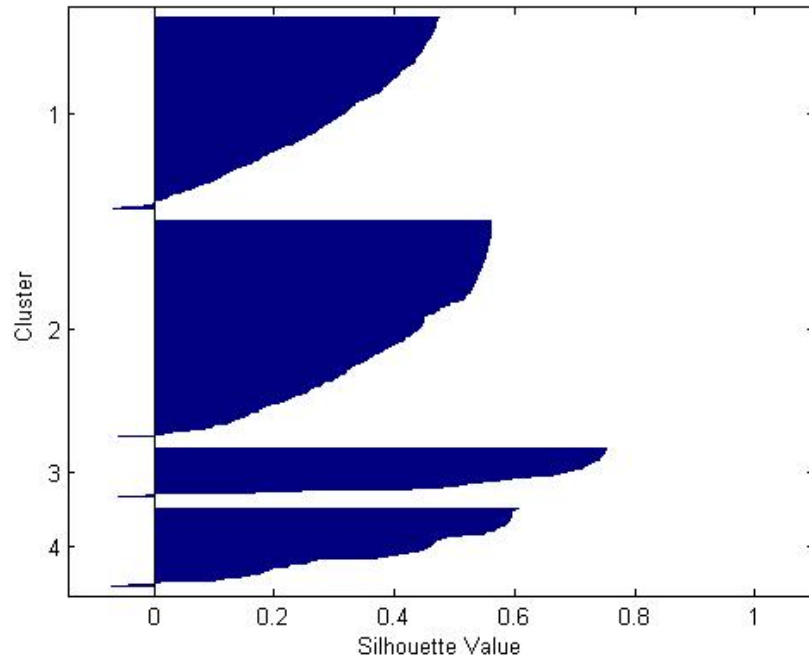
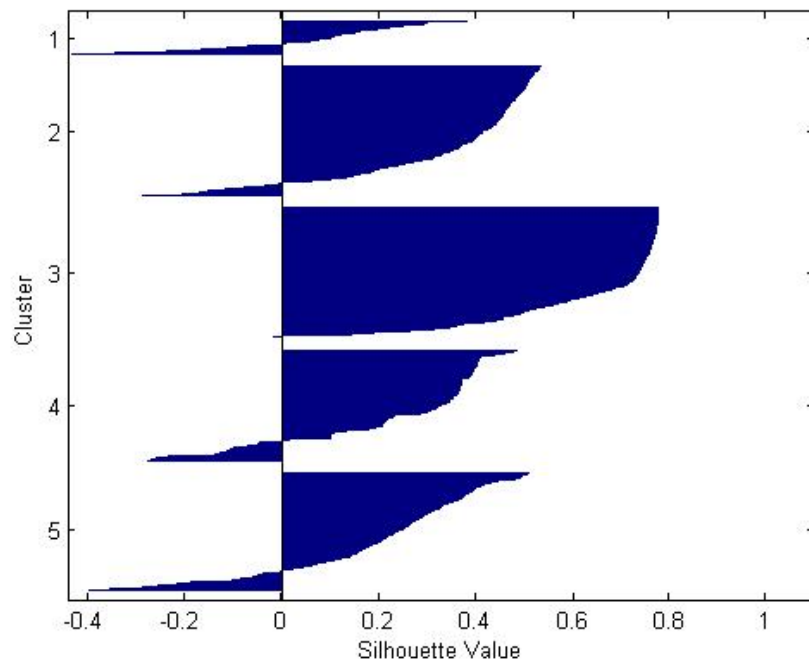
(a) Manhattan with $k=4$ clusters.(b) Manhattan with $k=5$ clusters.

Figure 4.7: Manhattan Distance Function Silhouettes for Experiment No.1.

4.4. Clusters Representation

Once a set of clusters is found, the next task is to find a way to represent the clusters. The resulting clusters need to be represented in a comprehensible and reasonable way in which it facilitates the evaluation of the resulting clusters. There are many ways for the representation of clusters. We used the so-called "classification models". In this method, we treat each cluster as a class. That is, all the data points in a cluster are regarded as having the same class label, e.g., the cluster ID.

4.5. Clusters Evaluation

After finding the appropriate number of clusters, these clusters have to be assessed. The quality of a clustering is hard to evaluate. That means that metrics must be used to measure how good the clustering techniques are and to evaluate the quality of a set of dataset clusters. To evaluate the clusters, some of the commonly used evaluation methods were used, which are: User Inspection, Purity, Precision and Recall. Recall and precision are more useful measures that have a fixed range and are easy to compare across clusters. A combination of both precision and recall that measures the extent to which a cluster contains only objects of a particular class and all objects of that class.

4.5.1. Evaluation of Experiment No.2 with k=4

Representing results is one of the appropriate ways of evaluation, as was mentioned in section 4.4. Figure 4.12 shows the 3-D representation of CS2 clusters with k=4 i.e. the total number of clusters is four. In this figure, the first cluster (Cluster1) shows a high score of the value (2.5), the second cluster (Cluster2) shows a high count in values below (2), the third cluster (Cluster3) illustrates the maximum count of the value (3), and finally, the fourth cluster (Cluster4) shows a high score of the value (3.5) and above. This result had guided us to the final grouping (categorization) results, this categorization will be explained in details in section 4.6. Distribution of students among clusters is shown

in Table 4.1. As shown in Table 4.1, the first cluster (Cluster1) has 411 data points, the second cluster (Cluster2) has 304 data points, the third cluster (Cluster3) has 441, and finally, the fourth cluster (Cluster4) has 643 data points. The last row of this table show the total number of data points which is 1799.

Table 4.1: Number of Students per Cluster in Experiment No.2 with k=4

Cluster	No. Of Students
1	411
2	304
3	441
4	643
Total	1799

The Confusion Matrix in table 4.2 shows the purity values of each cluster for (CS2) in Experiment No.2 with k=4 before categorization. It also shows the total purity of the whole clustering. Purity is the measure of the extent that a cluster contains only one class of data i.e. how pure is the cluster in respect of a class (value). As an example, the first cluster is 0.309 pure. That means that the maximum value's count, here is 127 which belongs to class (value) 2.5, is forming the purity of the first cluster (Cluster1) by dividing it by the total number of data points in this cluster (Cluster1) i.e. $\frac{127}{411}$.

Table 4.3 shows the Precision values for (CS2) in Experiment No.2 with k=4 before categorization. Precision is a useful measure that has a fixed range which is 0.0 to 1.0 (or 0% to 100%) and is easy to compare across clusters. The key in finding better clustering is to increase precision without sacrificing recall. The first row is the data points values (0, 1, 1.5, 2, 2.5, 3, 3.5 and 4) and the first column is showing the clusters ID's. Each cell in this table is the cross between the data points and their corresponding cluster. This cell's value is showing how many of the data points in this cluster belong there. As an example, in the first cluster, there are $(\frac{19}{411}) = 0.046$ data points out of 411 data points belong to class (value) zero.

Table 4.4 shows the Recall values for (CS2) in Experiment No.2 with k=4 before

categorization. Recall is a useful measure that has a fixed range which is 0.0 to 1.0 (or 0% to 100%) and is easy to compare across clusters. Good clusters must have a high recall. Each cell in this table is the cross between the data points and their corresponding cluster. This cell's value is showing if all of the data points that belong to this cluster make it complete. As an example, in the first cluster, there are $(\frac{19}{59}) = 0.322$ data points out of 59 data points which are part of class (value) zero.

Table 4.2: Confusion matrix with purity values for (CS2) in Experiment No.2 with k=4 before categorization.

Cluster	zero	1	1.5	2	2.5	3	3.5	4	Purity
1	19	13	31	83	127	84	36	18	0.309
2	19	84	65	79	42	12	2	1	0.276
3	11	8	16	49	114	99	89	55	0.259
4	10	1	2	7	41	68	132	382	0.594
Total	59	106	114	218	324	263	259	456	0.393

Table 4.3: Precision values for (CS2) in Experiment No.2 with k=4 before categorization.

Precision	zero	1	1.5	2	2.5	3	3.5	4
Precision of Cluster 1	0.046	0.032	0.075	0.202	0.309	0.204	0.088	0.044
Precision of Cluster 2	0.063	0.276	0.214	0.26	0.138	0.039	0.007	0.003
Precision of Cluster 3	0.025	0.018	0.036	0.111	0.259	0.224	0.202	0.125
Precision of Cluster 4	0.016	0.002	0.003	0.011	0.064	0.106	0.205	0.709

Table 4.4: Recall values for (CS2) in Experiment No.2 with k=4 before categorization.

Recall	zero	1	1.5	2	2.5	3	3.5	4
Recall of Cluster 1	0.322	0.123	0.272	0.381	0.392	0.319	0.139	0.039
Recall of Cluster 2	0.322	0.792	0.57	0.362	0.13	0.046	0.008	0.002
Recall of Cluster 3	0.186	0.075	0.151	0.225	0.352	0.376	0.344	0.121
Recall of Cluster 4	0.169	0.009	0.018	0.032	0.127	0.259	0.51	0.838

4.5.2. Evaluation of Experiment No.2 with k=7

Also the case of k=7 clusters was tested because CS2 course has seven passed grading categories which are: *D, D+, C, C+, B, B+* and *A*), in addition to the eighth failed category, which is (*F*). As seen in Figure 4.6(b), although k=7 is not a bad result but regarding the other K values, it is not the best. Figure 4.13 shows a 3-D representation of CS2 clusters with k=7. This representation will help us in evaluating Experiment No.2 with k=7.

Distribution of students among clusters is shown in table 4.5. As shown in Table 4.5, the first cluster (Cluster1) has 295 data points, the second cluster (Cluster2) has 393 data points, the third cluster (Cluster3) has 150, the fourth cluster (Cluster4) has 150 data points, the fifth cluster (Cluster5) has 106 data points, the sixth cluster (Cluster6) has 261 data points, and finally, the seventh cluster (Cluster7) has 444 data points. The last row of this table show the total number of data points which is 1799.

Table 4.5: Number of Students per Cluster in Experiment No.2 with k=7.

Cluster	No. Of Students
1	295
2	393
3	150
4	150
5	106
6	261
7	444
Total	1799

The Confusion matrix in table 4.6 shows the purity values of each cluster for (CS2) in Experiment No.2 with k=7. Purity is the measure of the extent that a cluster contains only one class of data i.e. how pure is the cluster in respect of a class (value). As an example, the first cluster is 0.315 pure. That means that the maximum value's count, here is 93 which belongs to class (value) 2.5, is forming the purity of the first cluster (Cluster1) by dividing it by the total number of data points in this cluster (Cluster1) i.e. $\frac{93}{295}$.

Also after extracting the confusion matrix of (CS2) in Experiment No.2 with k=7, the following were found:

1. No distinct categories could be concluded.
2. Each cluster has more than one maximum Recall value, as an example, cluster number (2) in table 4.6 has three maximum Recall values; zero, 2.5 and 3.
3. It is also obvious from Figure 4.13 that there are many clusters which have more than one high count in values, as an example, Cluster6 and Cluster7.

Table 4.7 shows the Precision values for (CS2) in Experiment No.2 with k=7. Precision is a useful measure that has a fixed range which is 0.0 to 1.0 (or 0% to 100%) and is easy to compare across clusters. The key in finding better clustering is to increase precision without sacrificing recall. The first row is the data points values (0, 1, 1.5, 2, 2.5, 3, 3.5 and 4) and the first column is showing the clusters ID's. Each cell in this table is the cross between the data points and their corresponding cluster. This cell's value is showing how many of the data points in this cluster belong there. As an example, in the first cluster, there are $(\frac{10}{295}) = 0.034$ data points out of 295 data points belong to class (value) zero.

Table 4.8 shows the Recall values for (CS2) in Experiment No.2 with k=7. Recall is a useful measure that has a fixed range which is 0.0 to 1.0 (or 0% to 100%) and is easy to compare across clusters. Good clusters must have a high recall. Each cell in this table is the cross between the data points and their corresponding cluster. This cell's value is showing if all of the data points that belong to this cluster make it complete. As an example, in the first cluster, there are $(\frac{10}{59}) = 0.169$ data points out of 59 data points which are part of class (value) zero.

Table 4.6: Confusion matrix with purity values for (CS2) in Experiment No.2 with k=7.

Cluster	zero	1	1.5	2	2.5	3	3.5	4	Purity
1	10	7	17	55	93	66	29	18	0.315
2	16	9	17	43	107	82	77	42	0.272
3	6	10	23	43	39	19	7	3	0.287
4	9	60	33	32	11	5	0	0	0.400
5	7	19	21	36	21	1	1	0	0.340
6	8	1	1	5	29	55	63	99	0.379
7	3	0	2	4	24	35	82	294	0.662
Total	59	106	114	218	324	263	259	456	0.407

Table 4.7: Precision values for (CS2) in Experiment No.2 with k=7.

Precision	zero	1	1.5	2	2.5	3	3.5	4
Precision of Cluster 1	0.034	0.024	0.058	0.186	0.315	0.224	0.098	0.061
Precision of Cluster 2	0.041	0.023	0.043	0.109	0.272	0.209	0.196	0.107
Precision of Cluster 3	0.040	0.067	0.153	0.287	0.260	0.127	0.047	0.020
Precision of Cluster 4	0.060	0.400	0.220	0.213	0.073	0.033	0.000	0.000
Precision of Cluster 5	0.066	0.179	0.198	0.340	0.198	0.009	0.009	0.000
Precision of Cluster 6	0.031	0.004	0.004	0.019	0.111	0.211	0.241	0.379
Precision of Cluster 7	0.007	0.000	0.005	0.009	0.054	0.079	0.185	0.662

Table 4.8: Recall values for (CS2) in Experiment No.2 with k=7.

Recall	zero	1	1.5	2	2.5	3	3.5	4
Recall of Cluster 1	0.169	0.066	0.149	0.252	0.287	0.251	0.112	0.039
Recall of Cluster 2	0.271	0.085	0.149	0.197	0.330	0.312	0.297	0.092
Recall of Cluster 3	0.102	0.094	0.202	0.197	0.120	0.072	0.027	0.007
Recall of Cluster 4	0.153	0.566	0.289	0.147	0.034	0.019	0.000	0.000
Recall of Cluster 5	0.119	0.179	0.184	0.165	0.065	0.004	0.004	0.000
Recall of Cluster 6	0.136	0.009	0.009	0.023	0.090	0.209	0.243	0.217
Recall of Cluster 7	0.051	0.000	0.018	0.018	0.074	0.133	0.317	0.645

4.6. Final Results

4.6.1. Suitable Grouping (Cluster Identification) for CS2 Students

According to the investigation of Experiment No.2 with k=4 that was explained in details in subsection 4.5.1, the appropriate categories were found for CS2 students, as shown in Table 4.9. And according to the Confusion matrix of (CS2) in Experiment

No.2 with $k=4$ after categorization Table 4.10, we found that the most suitable number of categories for the CS2 students was 4 which are: Weak, Good, V. Good and Excellent. Figure 4.14 shows the 3-D representation of CS2 categories for the results. As seen in Figure 4.14, the subdivision of the (CS2) values into such categories is as follows:

1. Weak; is the category that is giving for those values that are less than or equal 2 (≤ 2).
2. Good; is the category that is giving for values that are greater than 2 and less than 3 (> 2 and < 3).
3. V. Good; is the category that is giving for values that are greater than or equal 3 and less than 3.5 (≥ 3 and < 3.5).
4. Excellent; is the category that is giving for values that are greater than or equal 3.5 and less than or equal 4 (≥ 3.5 and ≤ 4).

Table 4.9: Categories of CS2 students with their ranges.

Category	Range
Weak	≤ 2
Good	> 2 and < 3
V. Good	≥ 3 and < 3.5
Excellent	≥ 3.5 and ≤ 4

Also the purity of this categorization was higher than the one of $k=4$ (before categorization), $k=6$ and $k=7$, as seen in table 4.10. This will be explained in details in the next subsection (subsection 4.6.2).

4.6.2. Evaluation of Experiment No.2 with $k=4$ after categorization.

Datasets classifications are used to evaluate clustering algorithms. Datasets classifications have several classes, and each data point is labeled with only one class. Using such a classification for cluster evaluation, we make the assumption that each class corresponds to a cluster. After clustering, we compare the cluster memberships with the class memberships to determine how good the clustering is.

In order to facilitate the evaluation, a confusion matrix from the resulting clusters was constructed. From the matrix, various measurements can be computed. Table 4.10 shows the Confusion Matrix with purity values for (CS2) in Experiment No.2 with $k=4$ after categorization. It also shows the total purity of the whole clustering. Purity is the measure of the extent that a cluster contains only one class of data i.e. how pure is the cluster in respect of a class (value). As an example, the first cluster is 0.355 pure. That means that the maximum value's count, here is 146 which belongs to the Weak category, is forming the purity of the first cluster (Cluster1) by dividing it by the total number of data points in this cluster (Cluster1) i.e. $\frac{146}{411}$.

Table 4.10: Confusion Matrix with purity values for (CS2) in Experiment No.2 with $k=4$ after categorization.

Cluster	Weak	Good	V. good	Excellent	Purity
1	146	127	84	54	0.355
2	247	42	12	3	0.813
3	84	114	99	144	0.327
4	20	41	68	514	0.799
Total	497	324	263	715	0.584

Table 4.11 shows the Precision values for (CS2) in Experiment No.2 with $k=4$ after categorization. Precision is a useful measure that has a fixed range which is 0.0 to 1.0 (or

0% to 100%) and is easy to compare across clusters. The key in finding better clustering is to increase precision without sacrificing recall. The first row is the data points values (Weak, Good, V. Good and Excellent) and the first column is showing the clusters ID's. Each cell in this table is the cross between the data points and their corresponding cluster. This cell's value is showing how many of the data points in this cluster belong there. As an example, in the first cluster, there are $(\frac{146}{411}) = 0.355$ data points out of 411 data points belong to the Weak category.

Table 4.12 shows Recall values for (CS2) in Experiment No.2 with k=4 after categorization. Recall is a useful measure that has a fixed range which is 0.0 to 1.0 (or 0% to 100%) and is easy to compare across clusters. Good clusters must have a high recall. Each cell in this table is the cross between the data points and their corresponding cluster. This cell's value is showing if all of the data points that belong to this cluster make it complete. As an example, in the first cluster, there are $(\frac{146}{497}) = 0.294$ data points out of 497 data points which are part of the Weak category.

CS2 attribute's values distribution all over the clusters can be seen in Figure 4.16.

Table 4.11: Precision values for (CS2) in Experiment No.2 with k=4 after categorization.

Precision	Weak	Good	V. good	Excellent
Precision of Cluster 1	0.355	0.309	0.204	0.131
Precision of Cluster 2	0.813	0.138	0.039	0.010
Precision of Cluster 3	0.190	0.259	0.224	0.327
Precision of Cluster 4	0.031	0.064	0.106	0.799

Table 4.12: Recall values for (CS2) in Experiment No.2 with k=4 after categorization.

Recall	Weak	Good	V. good	Excellent
Recall of Cluster 1	0.294	0.392	0.319	0.076
Recall of Cluster 2	0.497	0.130	0.046	0.004
Recall of Cluster 3	0.169	0.352	0.376	0.201
Recall of Cluster 4	0.040	0.127	0.259	0.719

4.6.3. CS2 Students' Distribution

After the final results' evaluation and after finding the suitable categorization of CS2, the CS2 students could be distributed over their suitable sections. Also similar groups of students were found depending on their previous HSGE, CGPA, CS1 and English1 attributes. This distribution is done according to the pre-selected attributes. And now the CS2 course could be adapted to the students according to their abilities. The four categories histograms are shown in Figures 4.15(a), 4.15(b), 4.16(a) and 4.16(b).

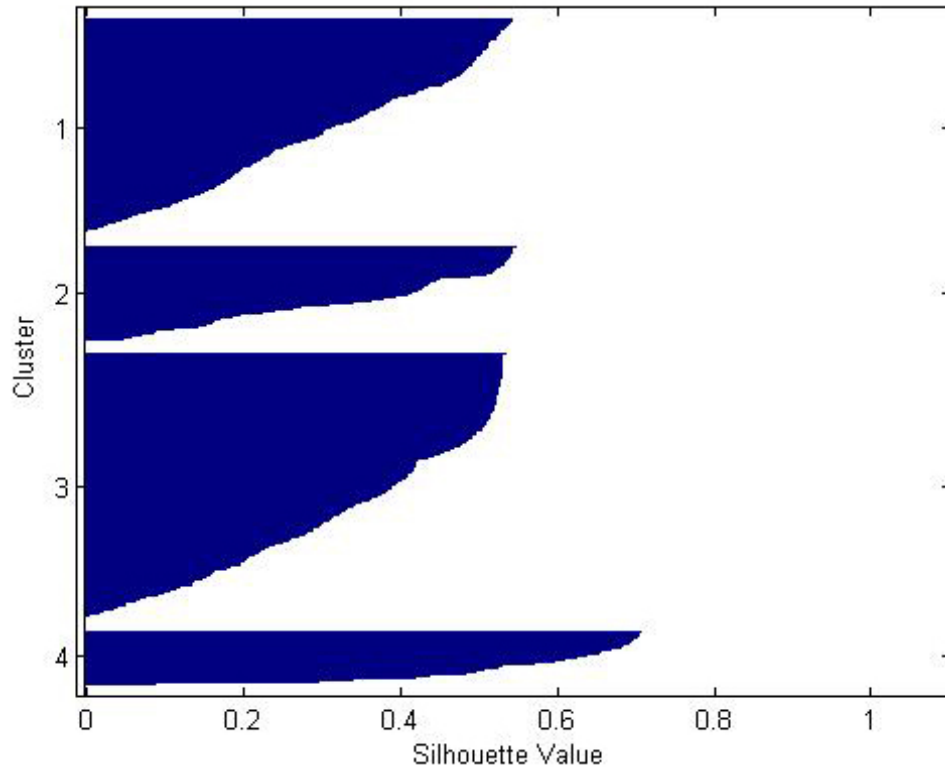
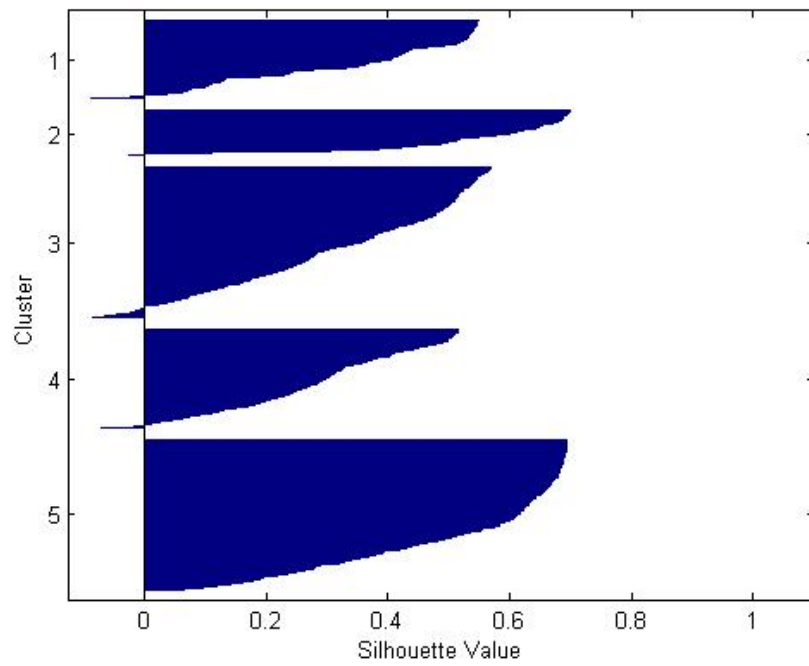
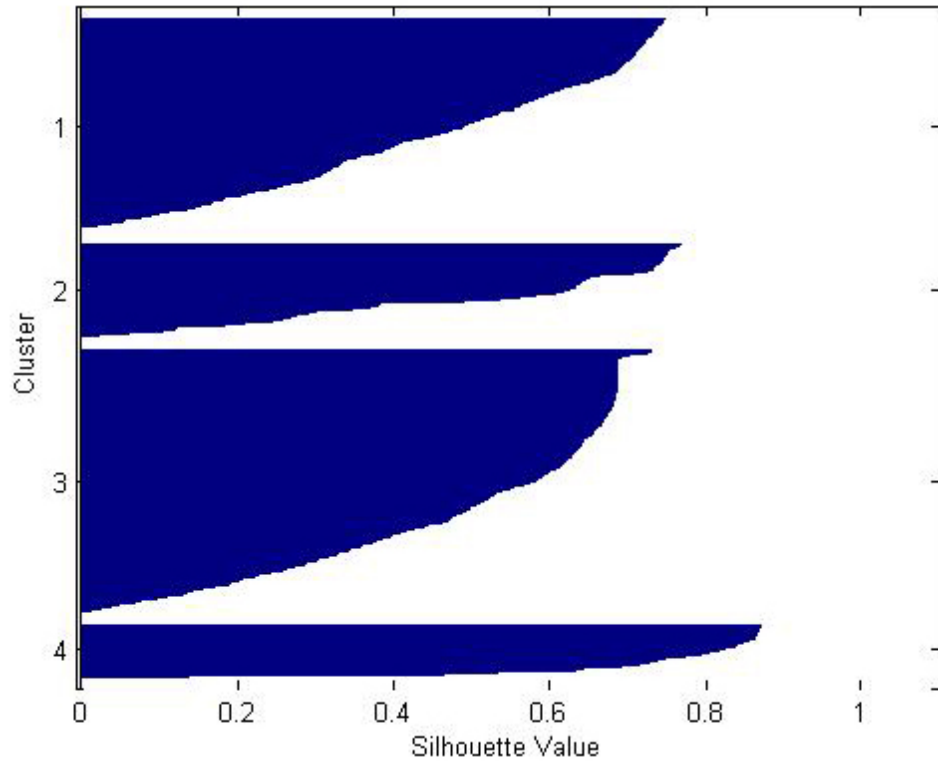
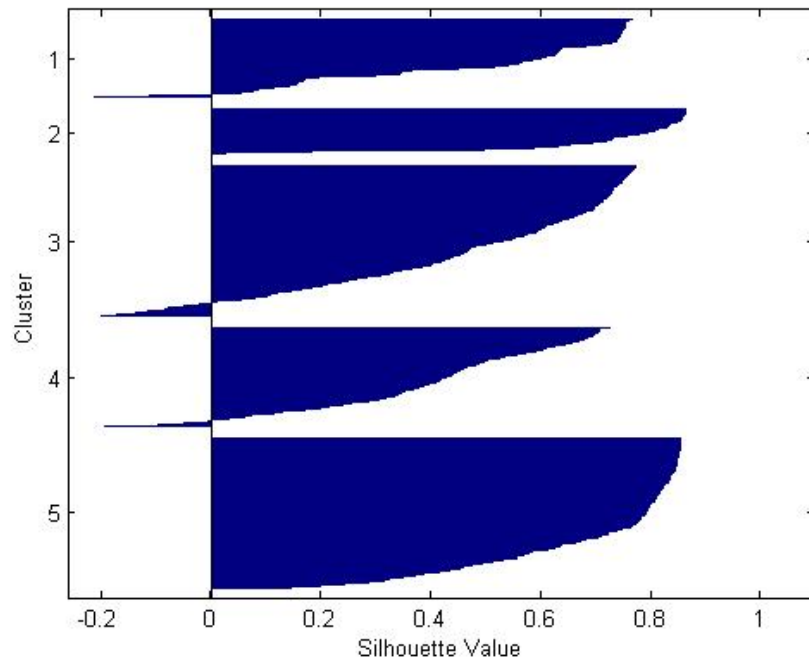
(a) Euclidean with $k=4$ clusters.(b) Euclidean with $k=5$ clusters.

Figure 4.8: Euclidean Distance Function Silhouettes for Experiment No.1.



(a) Square Euclidean with $k=4$ clusters.



(b) Square Euclidean with $k=5$ clusters.

Figure 4.9: Square Euclidean Distance Function Silhouettes for Experiment No.1.

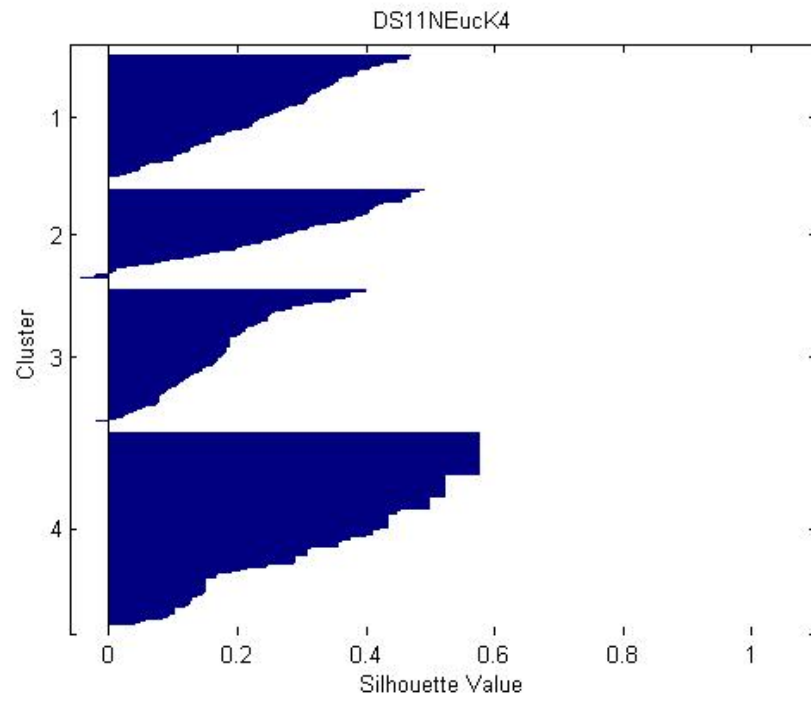
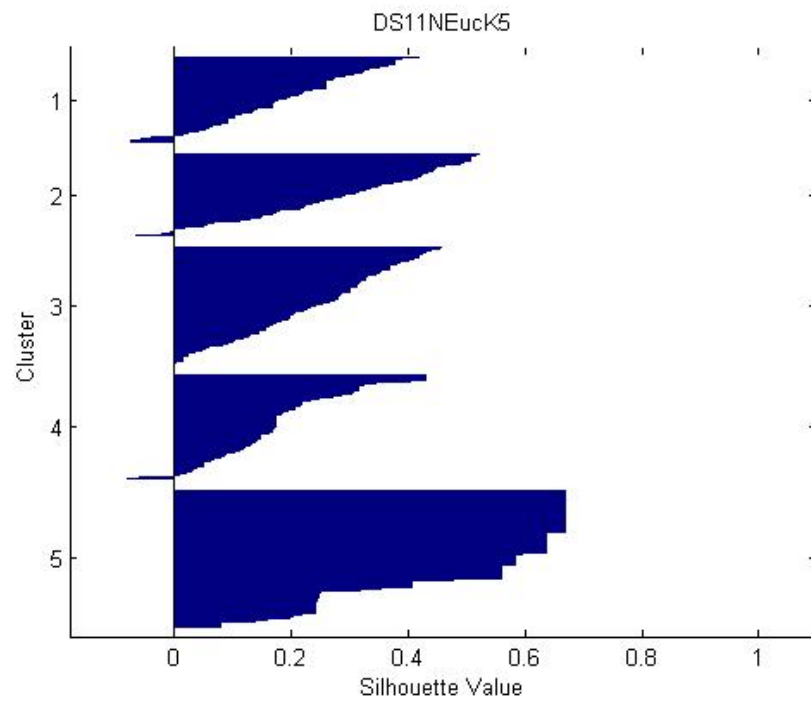
(a) Euclidean with $k=4$ clusters.(b) Euclidean with $k=5$ clusters.

Figure 4.10: Euclidean Distance Function Silhouettes for Experiment No.2.

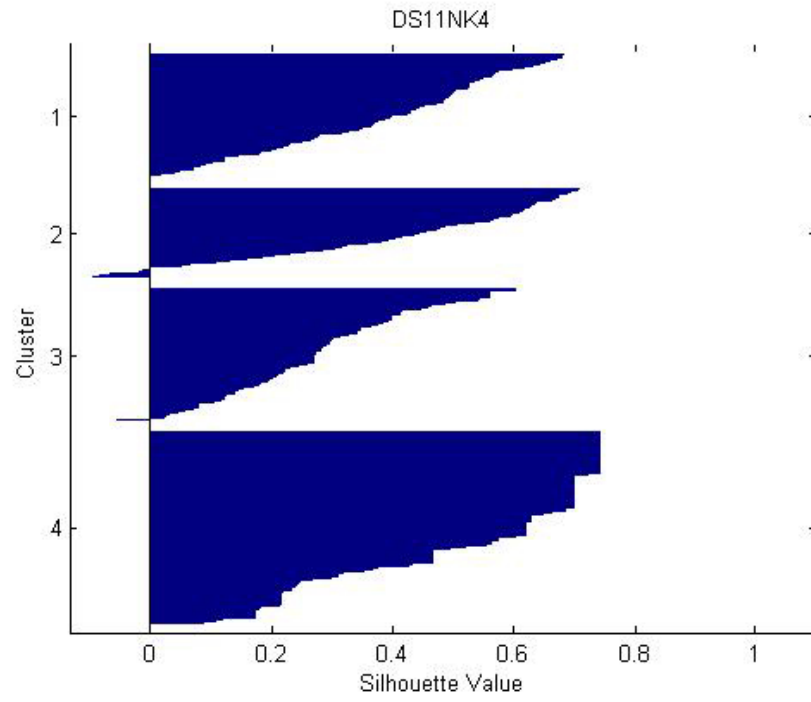
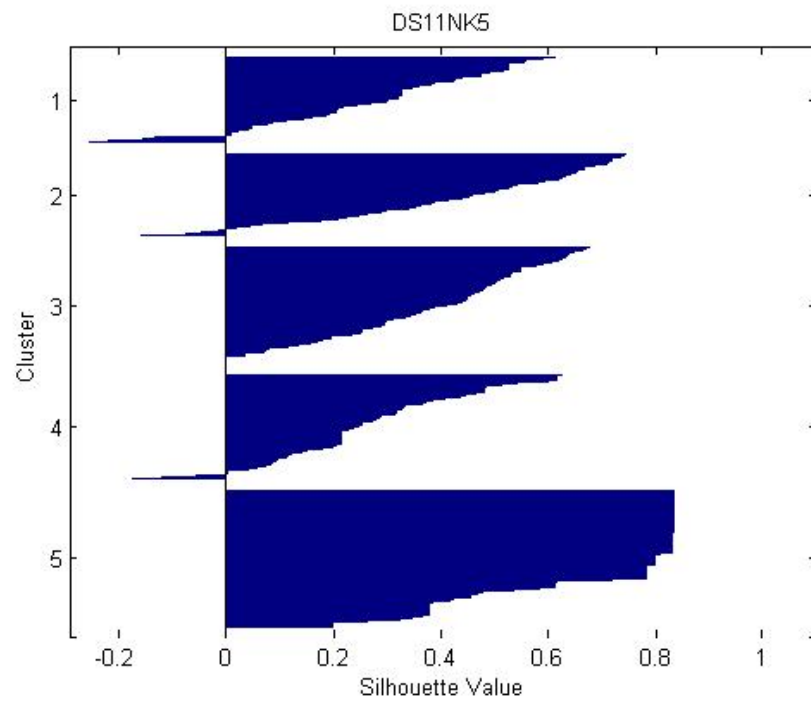
(a) Square Euclidean with $k=4$ clusters.(b) Square Euclidean with $k=5$ clusters.

Figure 4.11: Square Euclidean Distance Function Silhouettes for Experiment No.2.

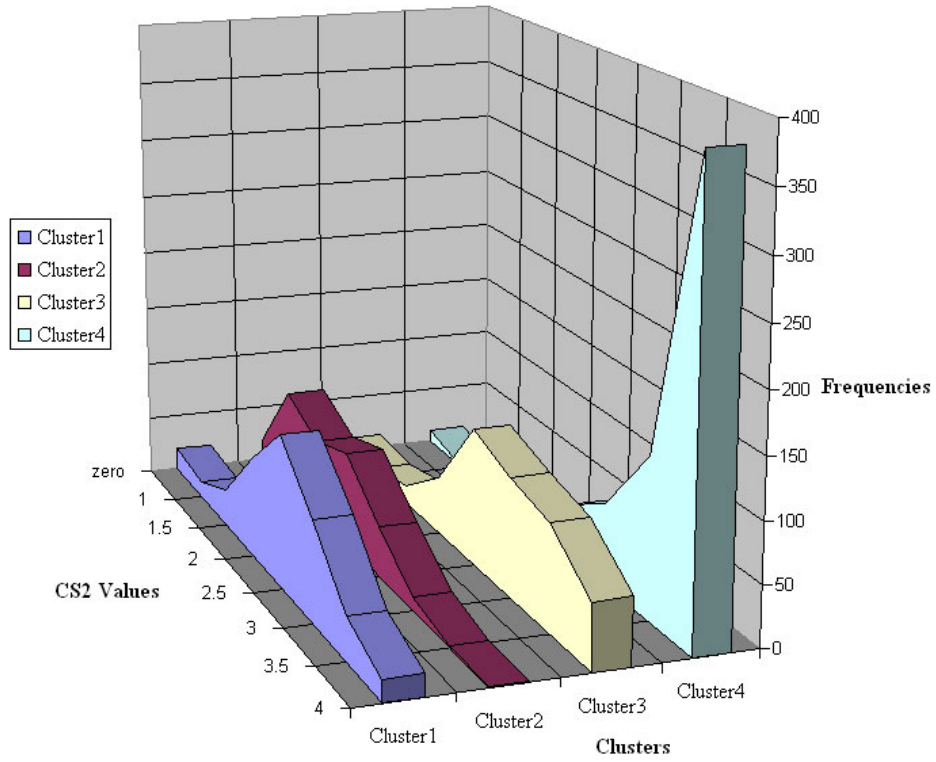


Figure 4.12: 3-D Diagram of CS2 clusters with $k=4$.

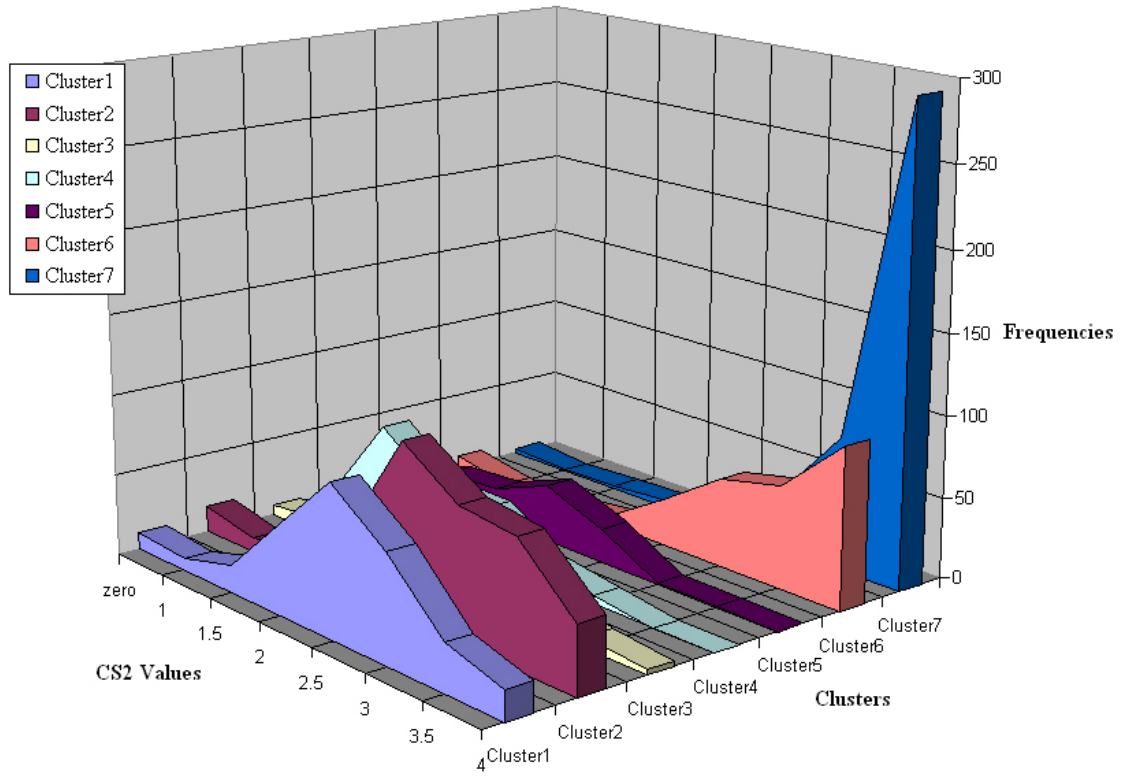


Figure 4.13: 3-D Diagram of CS2 clusters with $k=7$.

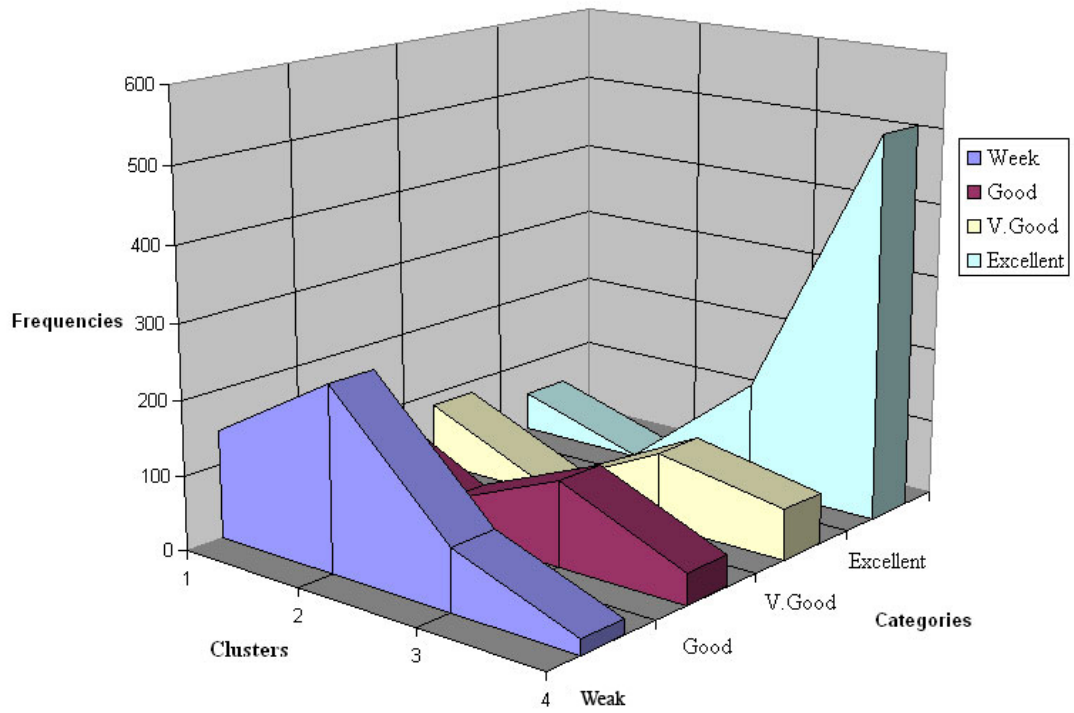
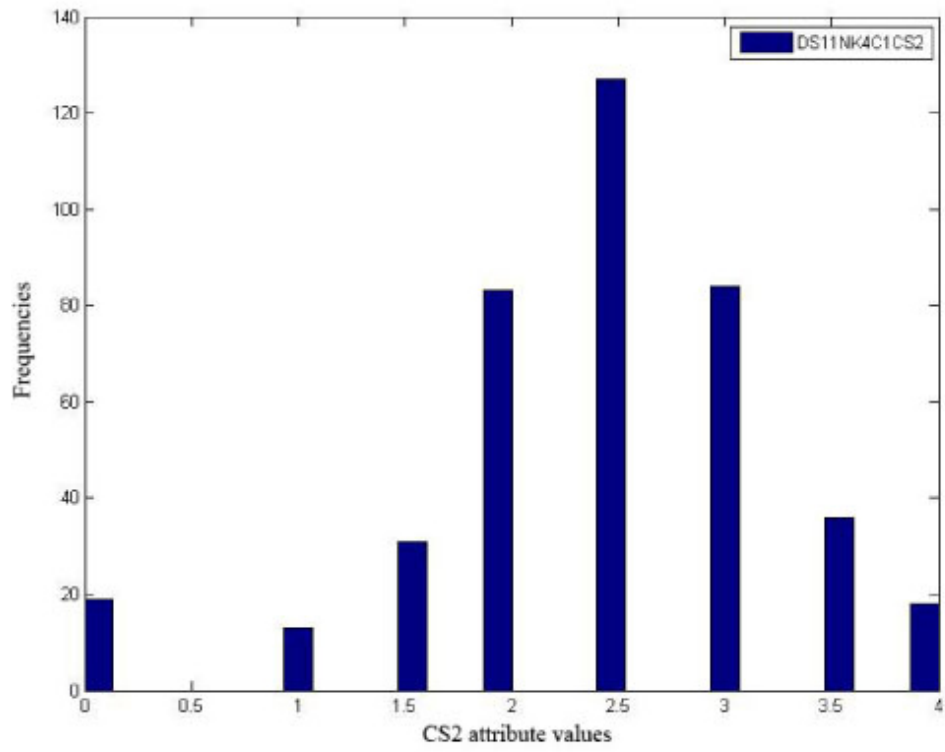
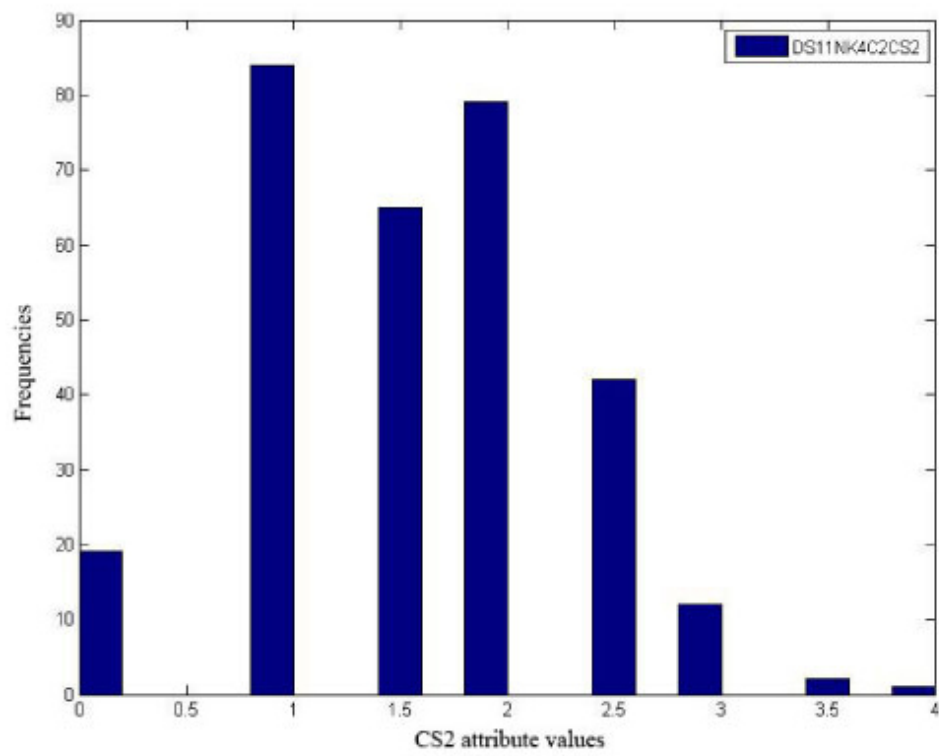


Figure 4.14: 3-D Diagram of CS2 categories for the results (Experiment No.2 with $k=4$).

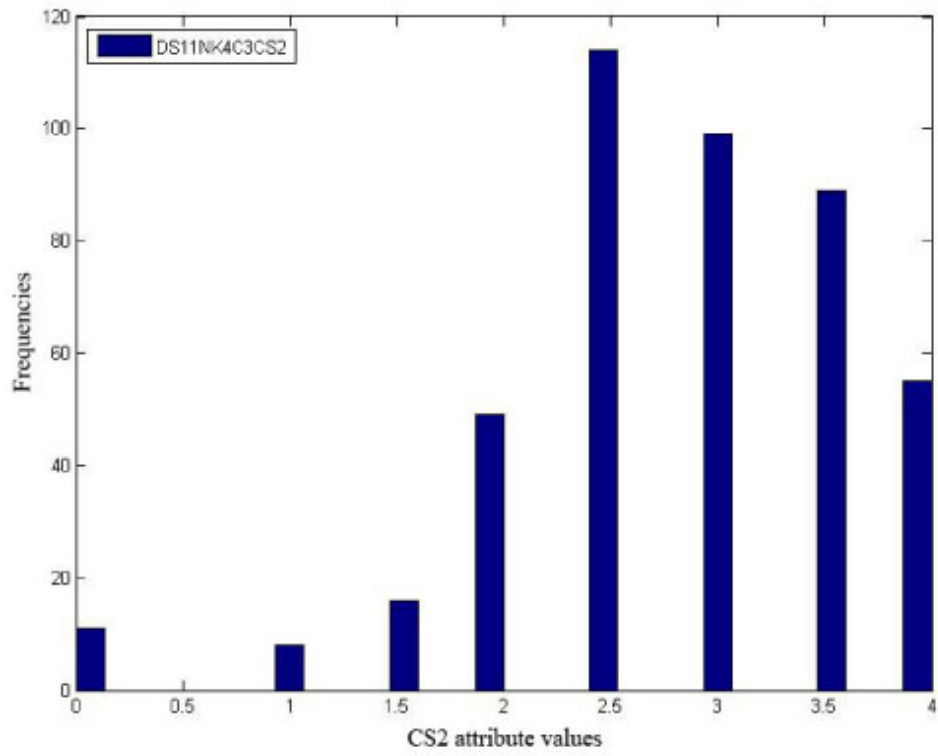


(a) Cluster (1)

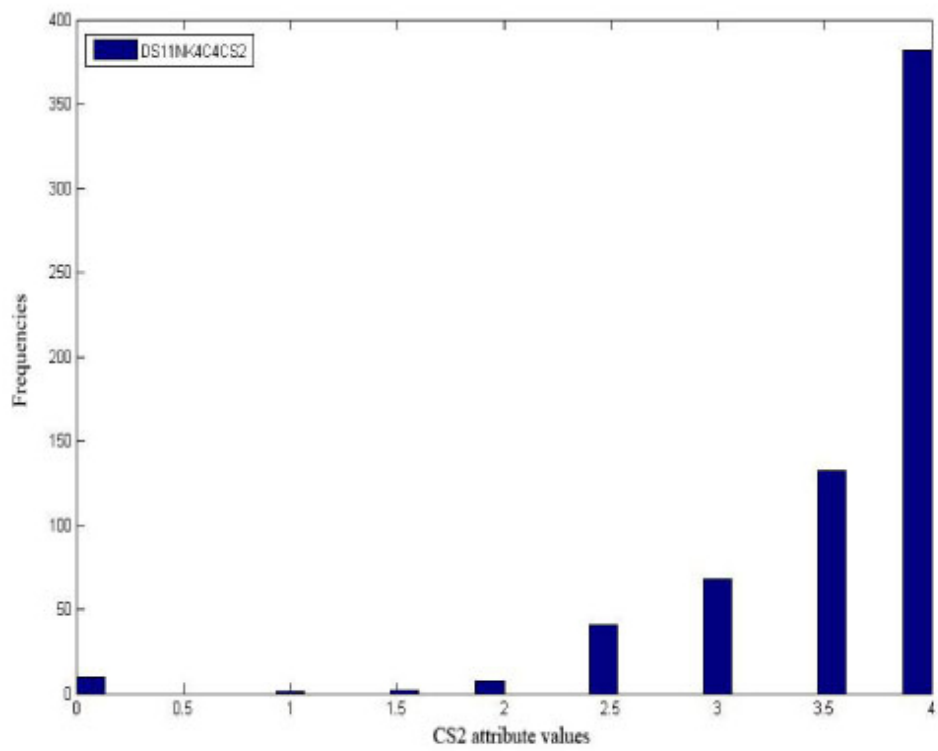


(b) Cluster (2)

Figure 4.15: CS2 attribute's values distribution all over the clusters (clusters 1 and 2).



(a) Cluster (3)



(b) Cluster (4)

Figure 4.16: CS2 attribute's values distribution all over the clusters (clusters 3 and 4).

5. Conclusions and Recommendations

This chapter will address the conclusions and the recommendations.

5.1. Conclusions

CS2 course is given to both medical and humanity colleges' students within the same sections i.e. medical and humanity students are taking the same course sections. This affects the academic achievement of both colleges' students and makes an achievement gaps which persist between medical and humanity colleges' students.

A suitable solution for the Computer Skills-2 (CS2) classes' classification problem was found. CS2 course problem resides in its grade scale at the end of the course semester. Because of the academic achievement gaps among students, the CS2 grade scale usually consists of two categories of grades. These categories are two levels. The first level is the high scores, usually consist of the medical colleges' students, and the other is the low scores, which is usually from humanity colleges' students.

According to the experimental results, four similar groups of students was found, which are: Weak, Good, V. Good and Excellent, depending on their previous backgrounds, knowledge, type of education, discipline, abilities, skills...etc. According to this categorization, the coordinator of the CS2 course will be able to allocate CS2 students to sections with similar interests. This allocation will enable instructors to provide students with specific topics, help, mentor, materials, exams, exercises...etc, according to a specific knowledge level that is suitable for each section. This may enable students to gain better knowledge, well understanding, get higher marks...etc.

According to the experimental results, the following were found:

1. CS2 Course students can be categorized in four classes:
 - Weak; is the category that is giving for those values that are less than or equal

2 (≤ 2).

- Good; is the category that is giving for values that are greater than 2 and less than 3 (> 2 and < 3).
- V. Good; is the category that is giving for values that are greater than or equal 3 and less than 3.5 (≥ 3 and < 3.5).
- Excellent; is the category that is giving for values that are greater than or equal 3.5 and less than or equal 4 (≥ 3.5 and ≤ 4).

2. The most significant attributes depending on their deterministic characteristics were:

- High Schools General Exam (HSGE).
- Cumulative Grade Point Average (CGPA).
- Computer Skills 1 (CS1).
- English1.

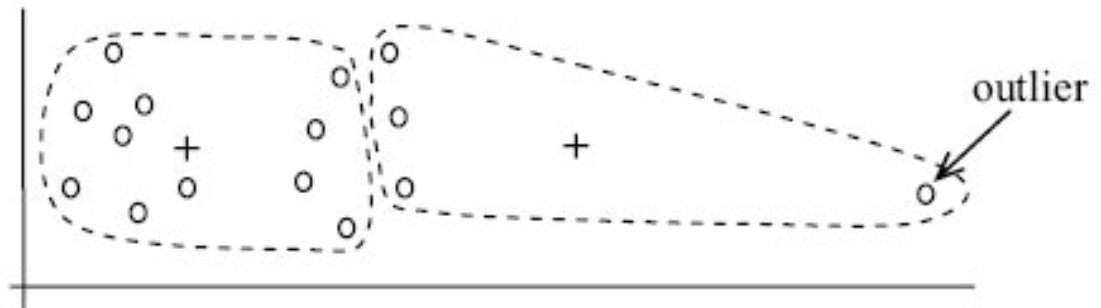
5.2. Recommendations

1. We recommend that the English 1 course be a prerequisite of CS2 course. English1 attribute was a deterministic one. It had a great impact on the data in experiments that were carried out. Also the influence of English1 attribute on data overpowers many attributes. CS2 course is taught in English; so English1 attribute was an indicator on each student's level in English.
2. We also recommend that the CS1 course must be taken with some restrictions before CS2 is registered by students. CS1 attribute was also another deterministic attribute. It had a great influence and impact on the data in experiments that were performed. Also the influence of CS1 attribute on data overpowers many attributes.

6. Future Work

6.1. Outliers Analysis

In statistics, an outlier is an observation that is numerically distant from the rest of the data. Outliers could be errors in the data recording or some special data points with very different values. Since the kmeans algorithm uses the mean as the centroid of each cluster, outliers may result in undesirable clusters. Effectiveness of outliers on clusters are shown in Figure 6.1(a) and 6.1(b).



(a) Undesirable Clusters



(b) Ideal Clusters

Figure 6.1: Clustering with and without the effect of outliers (Liu, 2007).

6.2. Dealing with Outliers

There are several methods for dealing with outliers. One simple method is to remove some data points in the clustering process that are much further away from the centroids

than other data points. To be safe, we may want to monitor these possible outliers over a few iterations and then decide whether to remove them. It is possible that a very small cluster of data points may be outliers. Usually, a threshold value is used to make the decision.

In the experiments, there were two kinds of outliers:

1. Ones were because of errors in data recording; those were obvious, so they were deleted.
2. Others were very far away from all other data points in the same clusters; those data points were noticed over many iterations and then the decision was either to remove them or reassigning them to other closest clusters. The process of removing or reassigning outliers is not that easy. This task will be the Future Work of this research.

A. Experiment No.1 Clusters

One of the experiments, as an example, that was tested on a dataset that consist of 1955 records and three attributes which are: CGPA, CS1 and English1.

One of the misleading values in CS1 and English1 attributes was the exempt ('free from' or 'pass') value, which denotes students who passed the Computer Skills or English Qualification exams respectively. Such records couldn't be crossed out, because simply they were somehow an indicator on student levels in Computer Skills and English Qualification exams. Also those records were about 1172 records in CS1 attribute and 623 records in English1 attribute, which means that they were about 60% of the data in CS1 attribute, as seen in Figure 4.3.

At first, the value '5' was given to these records. It was noticed that '5' values is skewing data points toward high scores clusters, as seen in the Figures A.2, A.3, A.4, A.5 and A.6. Figure A.2 shows the data points distribution in Cluster1. Figure A.3 shows the data points distribution in Cluster2. Figure A.4 shows the data points distribution in Cluster3. Figure A.5 shows the data points distribution in Cluster4. Figure A.6 shows the data points distribution in Cluster5.

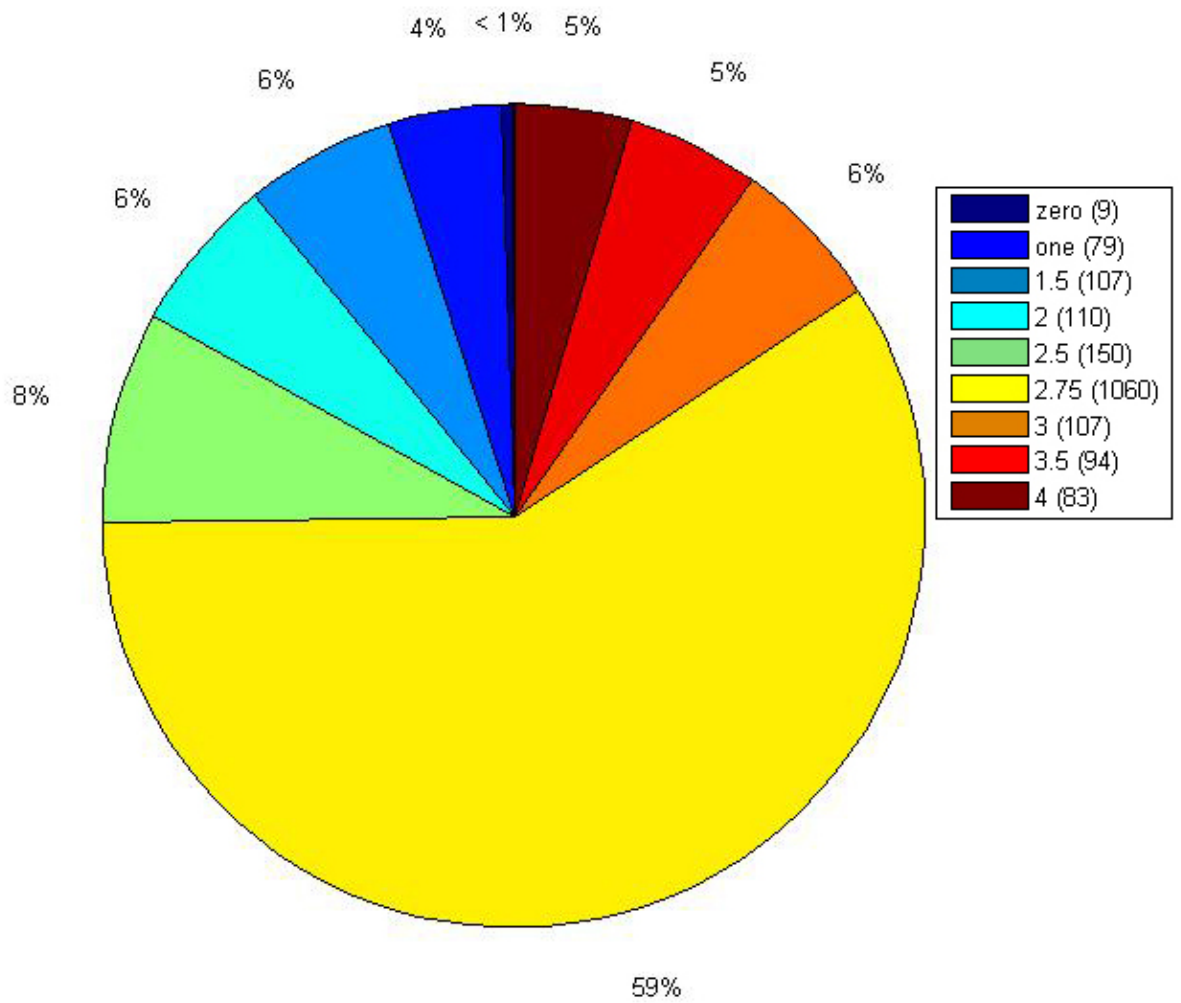


Figure A.1: Pie Graph for the whole (CS1) Records.

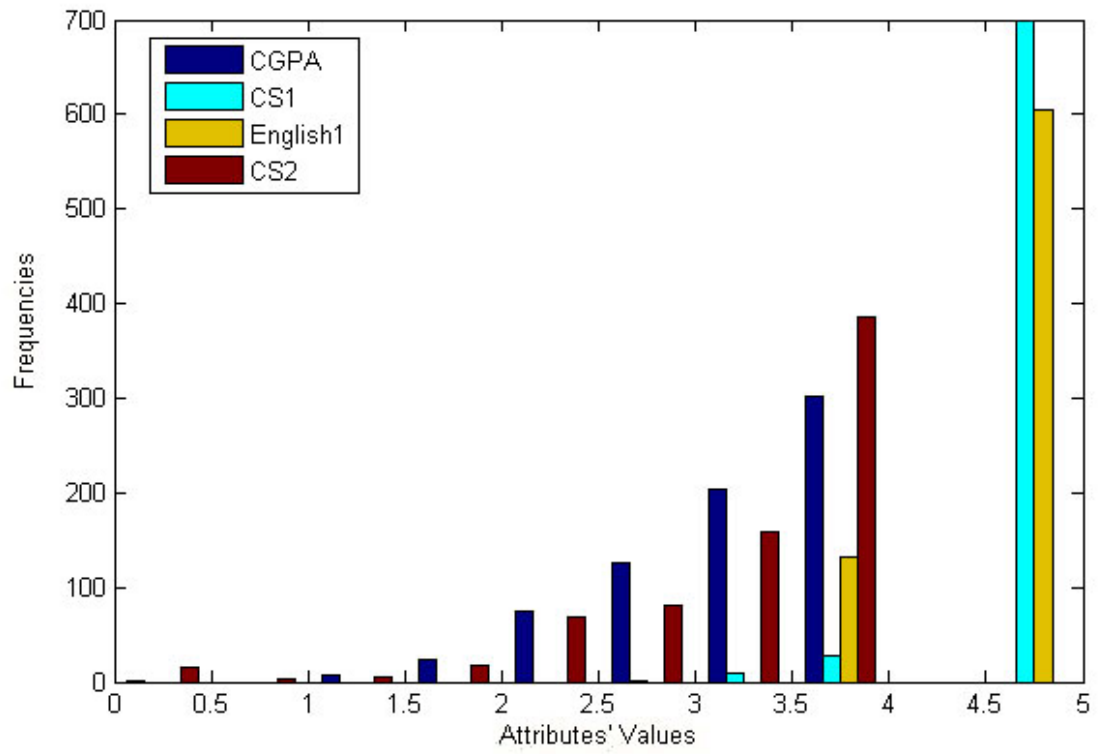


Figure A.2: Data points distribution in Cluster1.

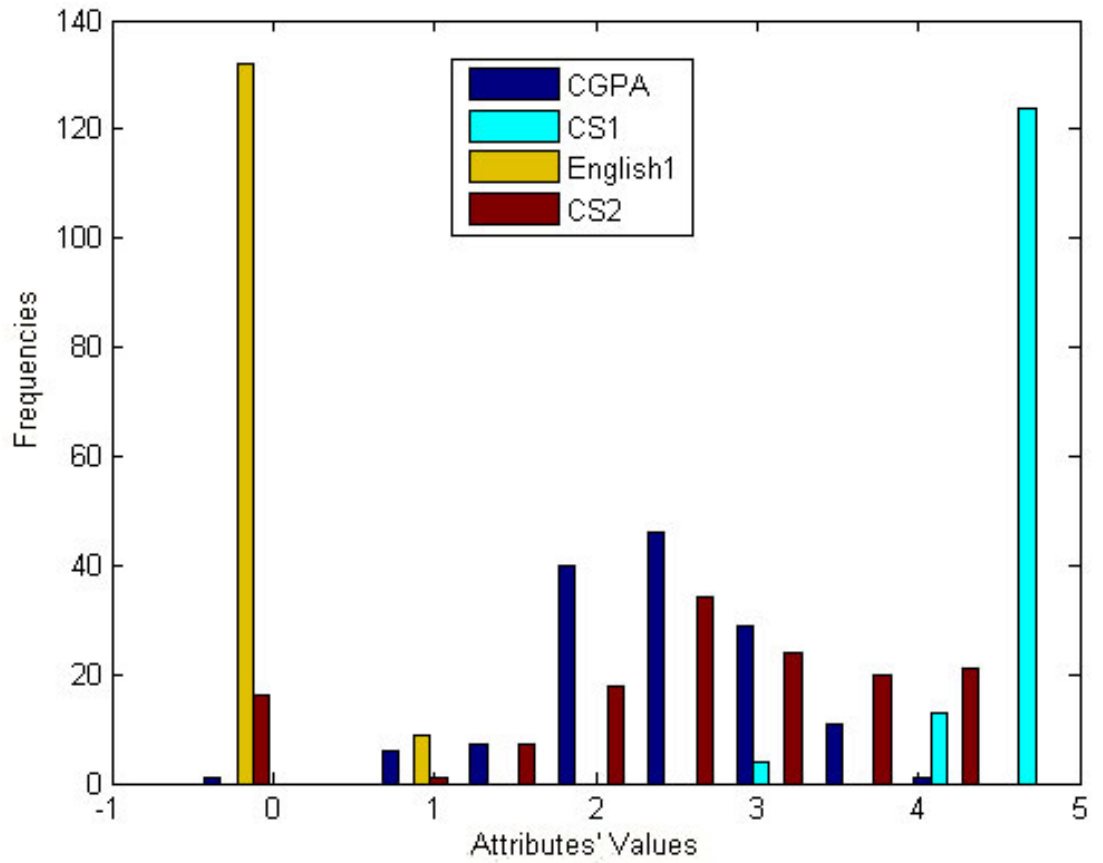


Figure A.3: Data points distribution in Cluster2.

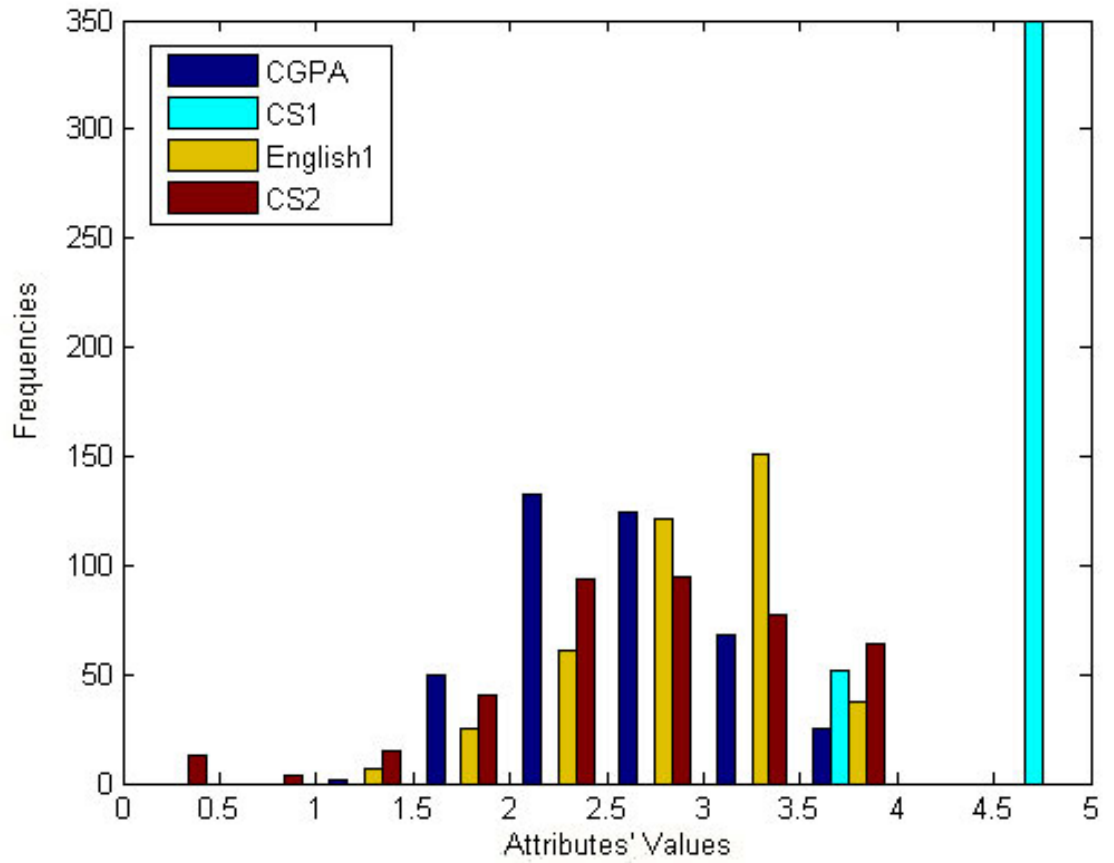


Figure A.4: Data points distribution in Cluster3.

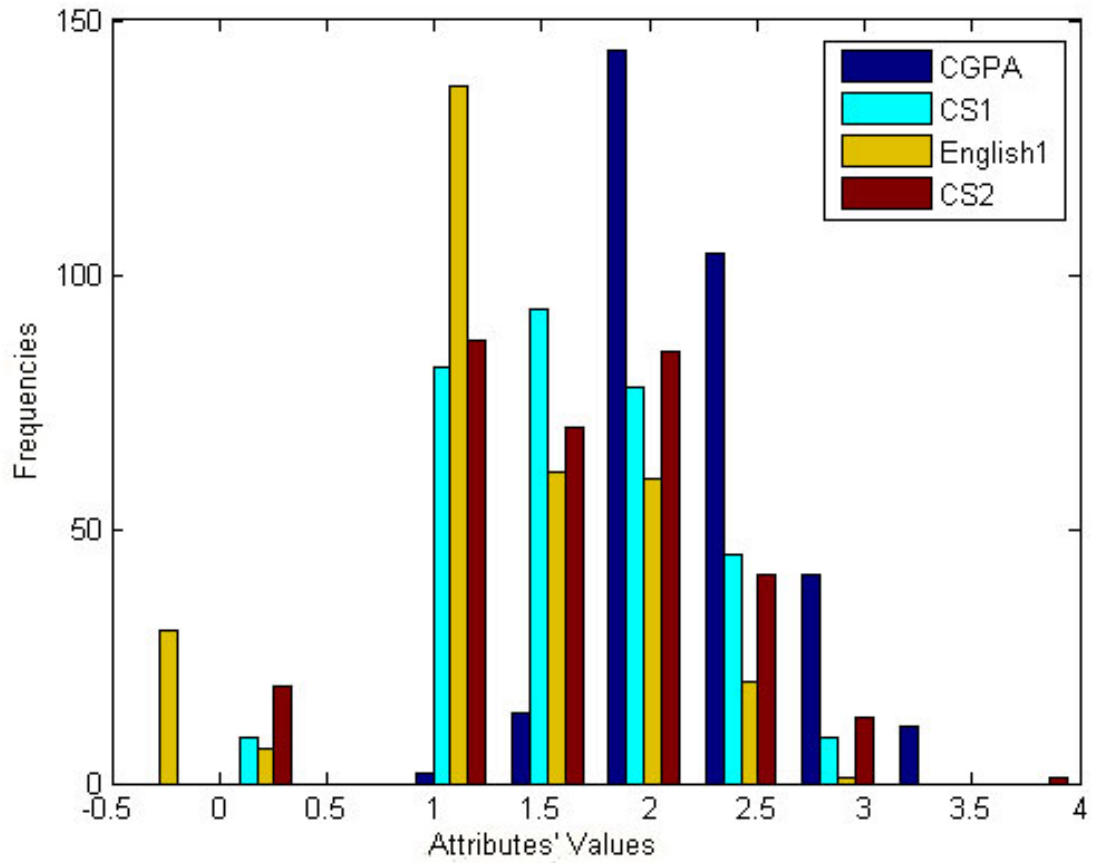


Figure A.5: Data points distribution in Cluster4.

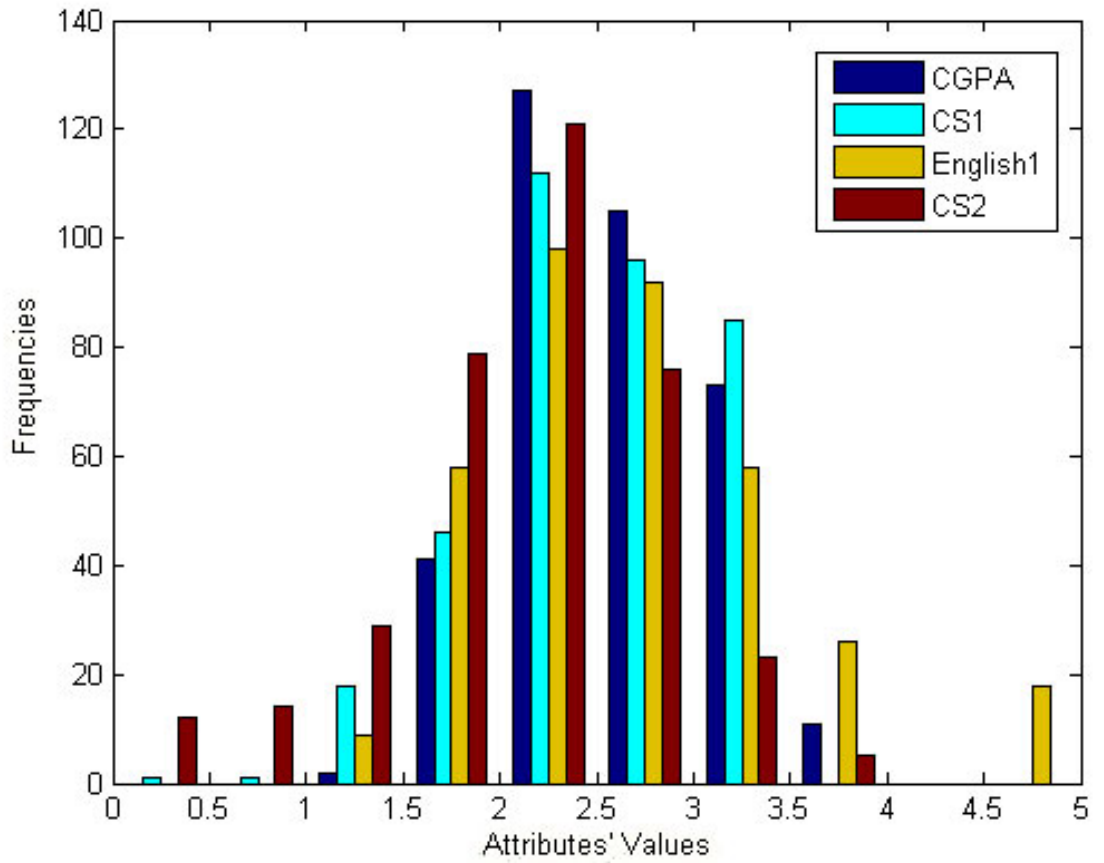


Figure A.6: Data points distribution in Cluster5.

B. Experiment No.2 with K=6

Many experimental tests has been done on many datasets, using MATLAB ®2008b (The MathWorks, 2008). The most important experimental results were mentioned and explained in details in Chapter 4.

One of the other experiments that had proved the chosen of Square Euclidean Distance Function is Experiment No.2. It was tested on a dataset that consist of 1799 records and four attributes which are: HSGE, CGPA, CS1 and English1. Representing results is one of the appropriate ways of evaluation. Figure B.1 shows the 3-D representation of CS2 clusters with k=6 and Figure B.2 shows the 3-D representation of CS2 clusters with k=6 but from the backside view.

The case of k=6 (six clusters) was tested because k=6 is not a bad result but regarding the other K values, it is not the best. Distribution of students among clusters is shown in table B.1. As shown in Table B.1, the first cluster (Cluster1) has 382 data points, the second cluster (Cluster2) has 475 data points, the third cluster (Cluster3) has 161, the fourth cluster (Cluster4) has 151 data points, the fifth cluster (Cluster5) has 500 data points, and the sixth cluster (Cluster6) has 130. The last row of this table show the total number of data points which is 1799.

Table B.1: Number of Students per Cluster in Experiment No.2 with k=6

Cluster	No. Of Students
1	382
2	475
3	161
4	151
5	500
6	130
Total	1799

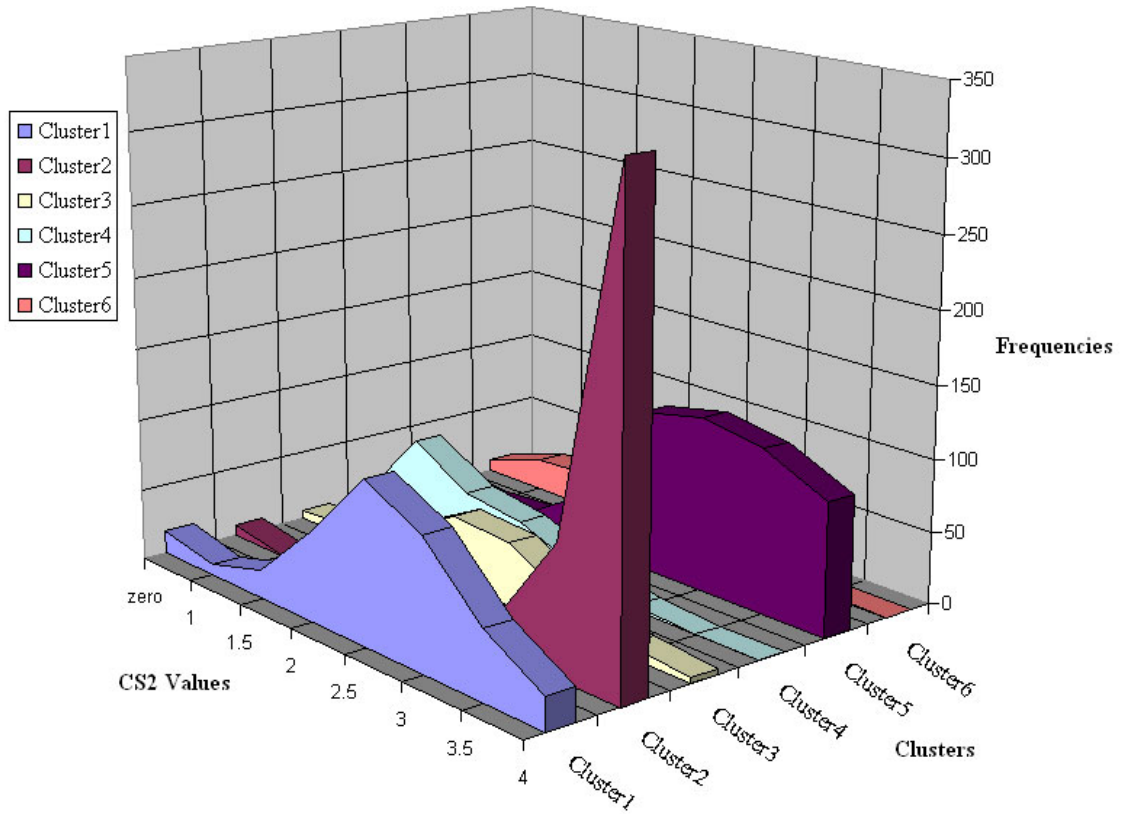


Figure B.1: 3-D Diagram of CS2 clusters with k=6.

The Confusion matrix in table B.2 shows the purity values of each cluster for (CS2) in Experiment No.2 with k=6. Purity is the measure of the extent that a cluster contains only one class of data i.e. how pure is the cluster in respect of a class (value). As an example, the first cluster is 0.296 pure. That means that the maximum value's count, here is 113 which belongs to class (value) 2.5, is forming the purity of the first cluster (Cluster1) by dividing it by the total number of data points in this cluster (Cluster1) i.e. $\frac{113}{382}$.

Also after extracting the confusion matrix of (CS2) in Experiment No.2 with k=6, the following were found:

1. No distinct categories could be concluded.
2. Each cluster has more than one maximum Recall value, as an example, as seen in

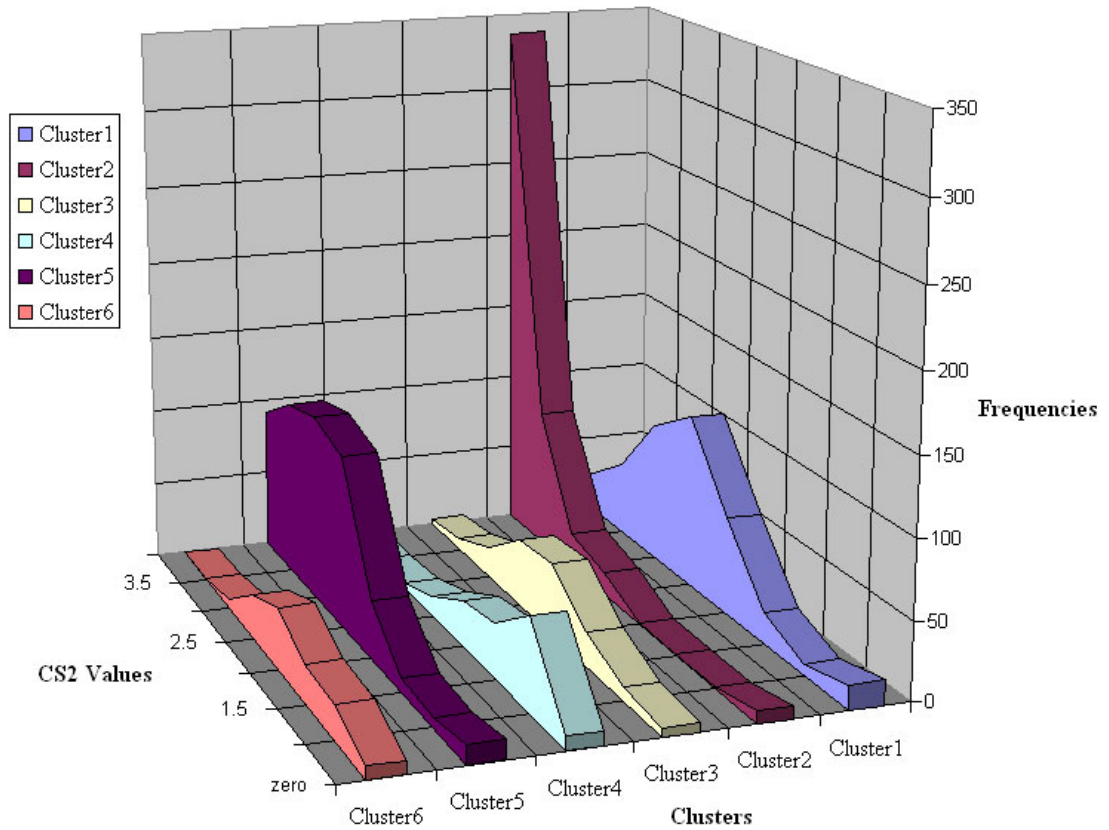


Figure B.2: 3-D Diagram of CS2 clusters with k=6 (Backside view).

table B.2.

3. It is also obvious from Figure B.1 that there are many clusters which have more than one high count in values, as an example, Cluster2 and Cluster5.

Table B.3 shows the Precision values for (CS2) in Experiment No.2 with k=6. Precision is a useful measure that has a fixed range which is 0.0 to 1.0 (or 0% to 100%) and is easy to compare across clusters. The key in finding better clustering is to increase precision without sacrificing recall. The first row is the data points values (0, 1, 1.5, 2, 2.5, 3, 3.5 and 4) and the first column is showing the clusters ID's. Each cell in this table is the cross between the data points and their corresponding cluster. This cell's value is showing how many of the data points in this cluster belong there. As an example, in the first

cluster, there are $(\frac{16}{382}) = 0.042$ data points out of 382 data points belong to class (value) zero.

Table B.4 shows the Recall values for (CS2) in Experiment No.2 with k=6. Recall is a useful measure that has a fixed range which is 0.0 to 1.0 (or 0% to 100%) and is easy to compare across clusters. Good clusters must have a high recall. Each cell in this table is the cross between the data points and their corresponding cluster. This cell's value is showing if all of the data points that belong to this cluster make it complete. As an example, in the first cluster, there are $(\frac{16}{59}) = 0.271$ data points out of 59 data points which are part of class (value) zero.

Table B.2: Confusion matrix with purity values for (CS2) in Experiment No.2 with k=6.

Cluster	zero	1	1.5	2	2.5	3	3.5	4	Purity
1	16	8	20	63	113	91	48	23	0.296
2	8	1	1	1	15	23	89	337	0.709
3	6	9	23	47	44	20	8	4	0.292
4	9	60	36	30	13	3	0	0	0.397
5	12	6	9	37	111	120	113	92	0.240
6	8	22	25	40	28	6	1	0	0.308
Total	59	106	114	218	324	263	259	456	0.399

Table B.3: Precision values for (CS2) in Experiment No.2 with k=6.

Precision	zero	1	1.5	2	2.5	3	3.5	4
Precision of Cluster 1	0.042	0.021	0.052	0.165	0.296	0.238	0.126	0.060
Precision of Cluster 2	0.017	0.002	0.002	0.002	0.032	0.048	0.187	0.709
Precision of Cluster 3	0.037	0.056	0.143	0.292	0.273	0.124	0.050	0.025
Precision of Cluster 4	0.060	0.397	0.238	0.199	0.086	0.020	0.000	0.000
Precision of Cluster 5	0.024	0.012	0.018	0.074	0.222	0.240	0.226	0.184
Precision of Cluster 6	0.062	0.169	0.192	0.308	0.215	0.046	0.008	0.000

Table B.4: Recall values for (CS2) in Experiment No.2 with k=6.

Recall	zero	1	1.5	2	2.5	3	3.5	4
Recall of Cluster 1	0.271	0.075	0.175	0.289	0.349	0.346	0.185	0.050
Recall of Cluster 2	0.136	0.009	0.009	0.005	0.046	0.087	0.344	0.739
Recall of Cluster 3	0.102	0.085	0.202	0.216	0.136	0.076	0.031	0.009
Recall of Cluster 4	0.153	0.566	0.316	0.138	0.040	0.011	0.000	0.000
Recall of Cluster 5	0.203	0.057	0.079	0.170	0.343	0.456	0.436	0.202
Recall of Cluster 6	0.136	0.208	0.219	0.183	0.086	0.023	0.004	0.000

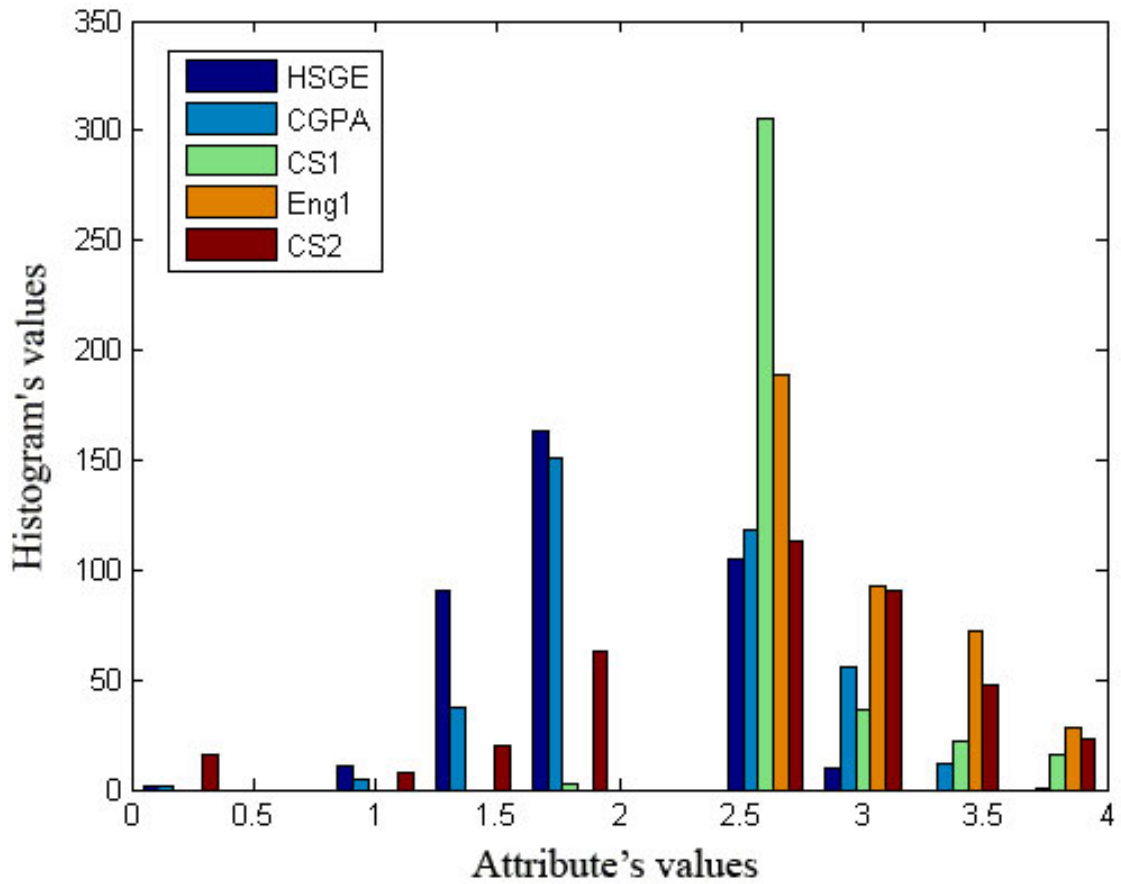


Figure B.3: Data points distribution's Histogram in Cluster1 (k=6).

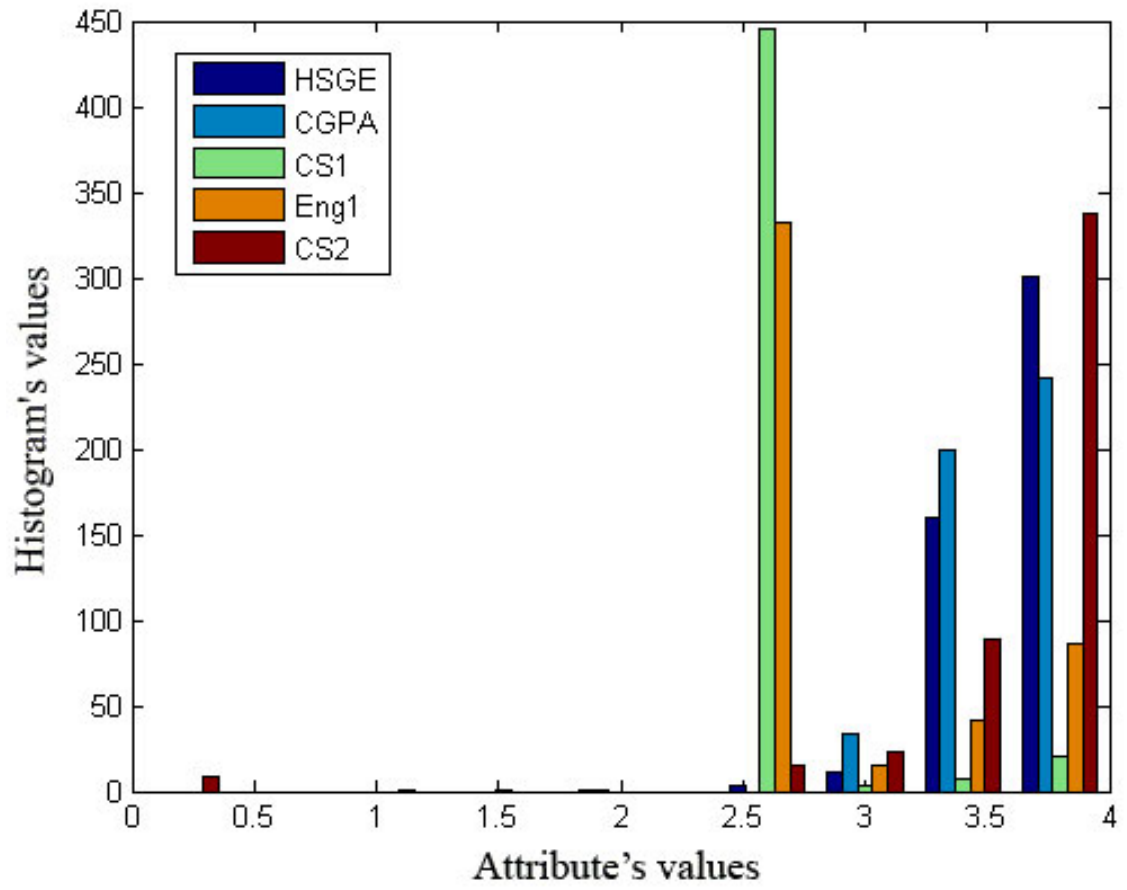


Figure B.4: Data points distribution's Histogram in Cluster2 (k=6).

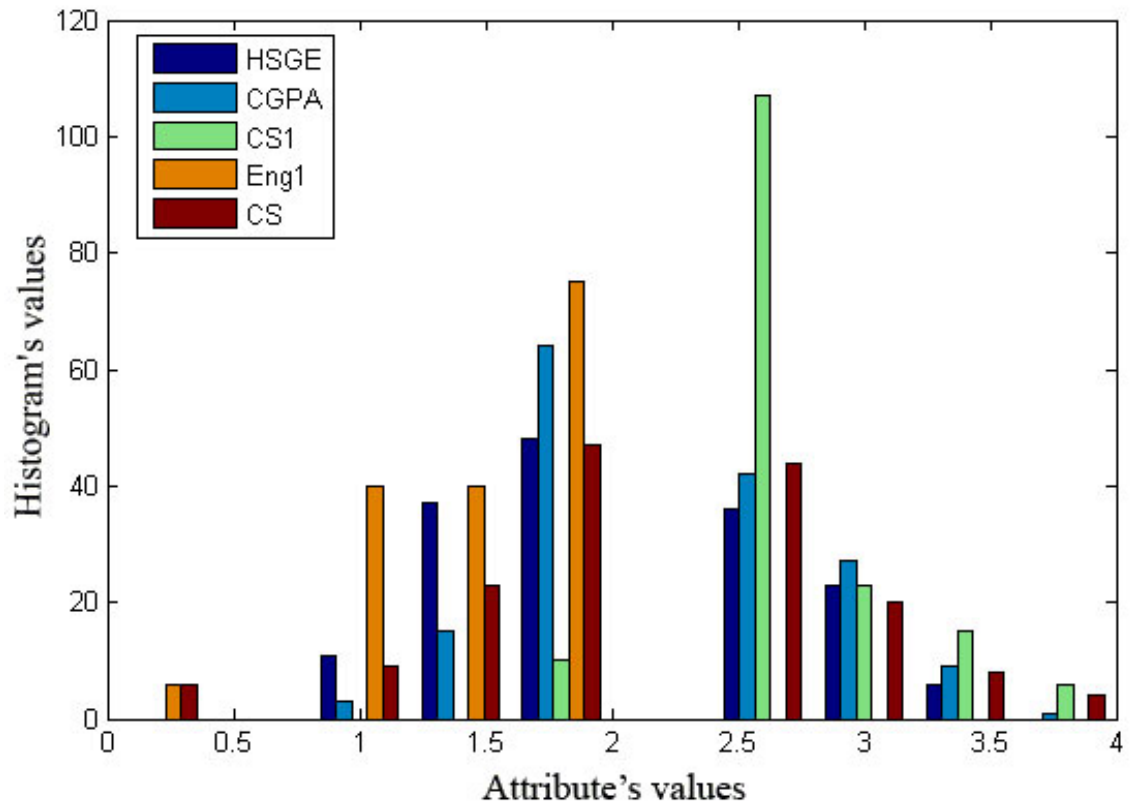


Figure B.5: Data points distribution's Histogram in Cluster3 (k=6).

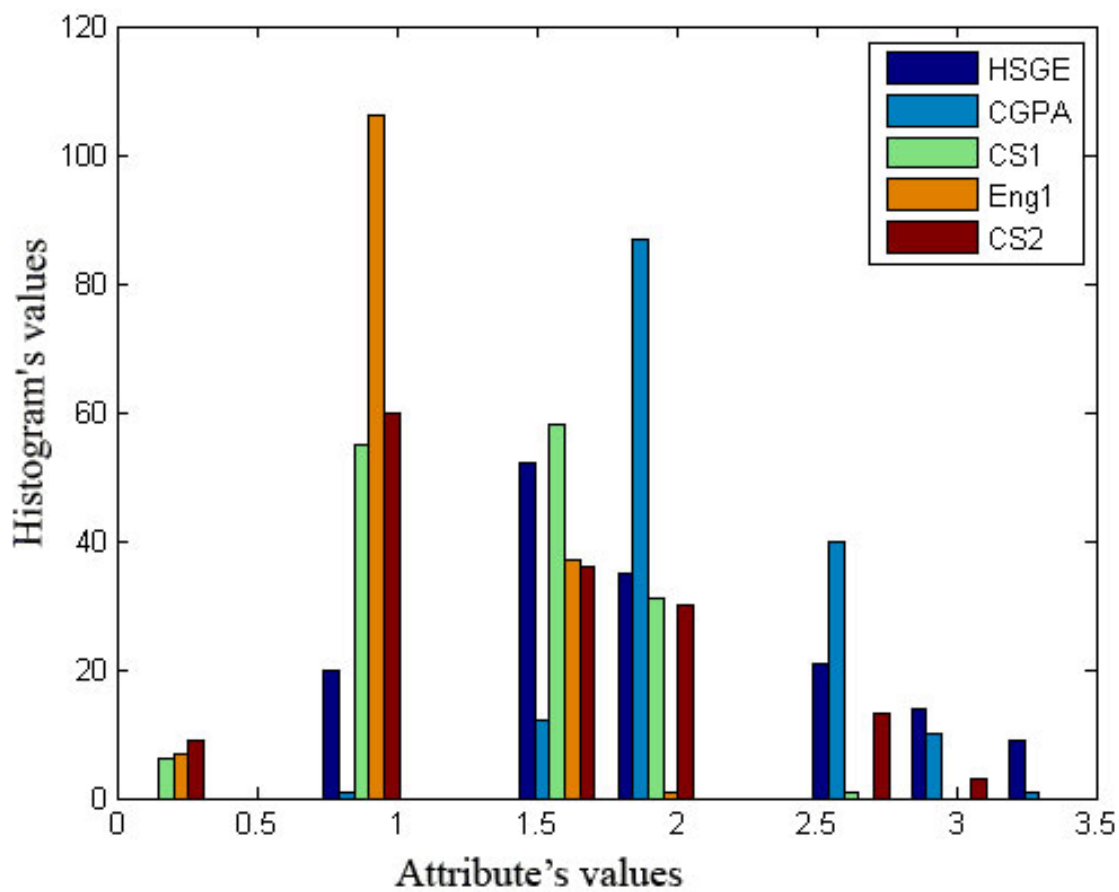


Figure B.6: Data points distribution's Histogram in Cluster4 (k=6).

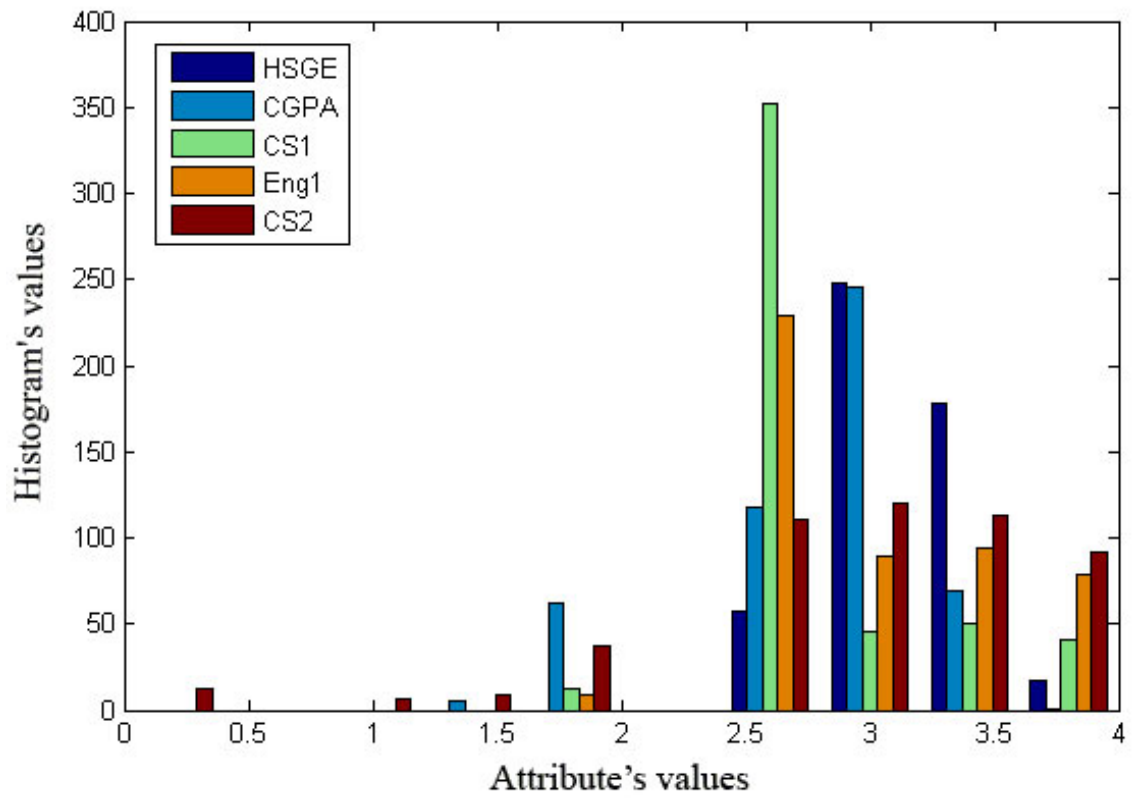


Figure B.7: Data points distribution's Histogram in Cluster5 (k=6).

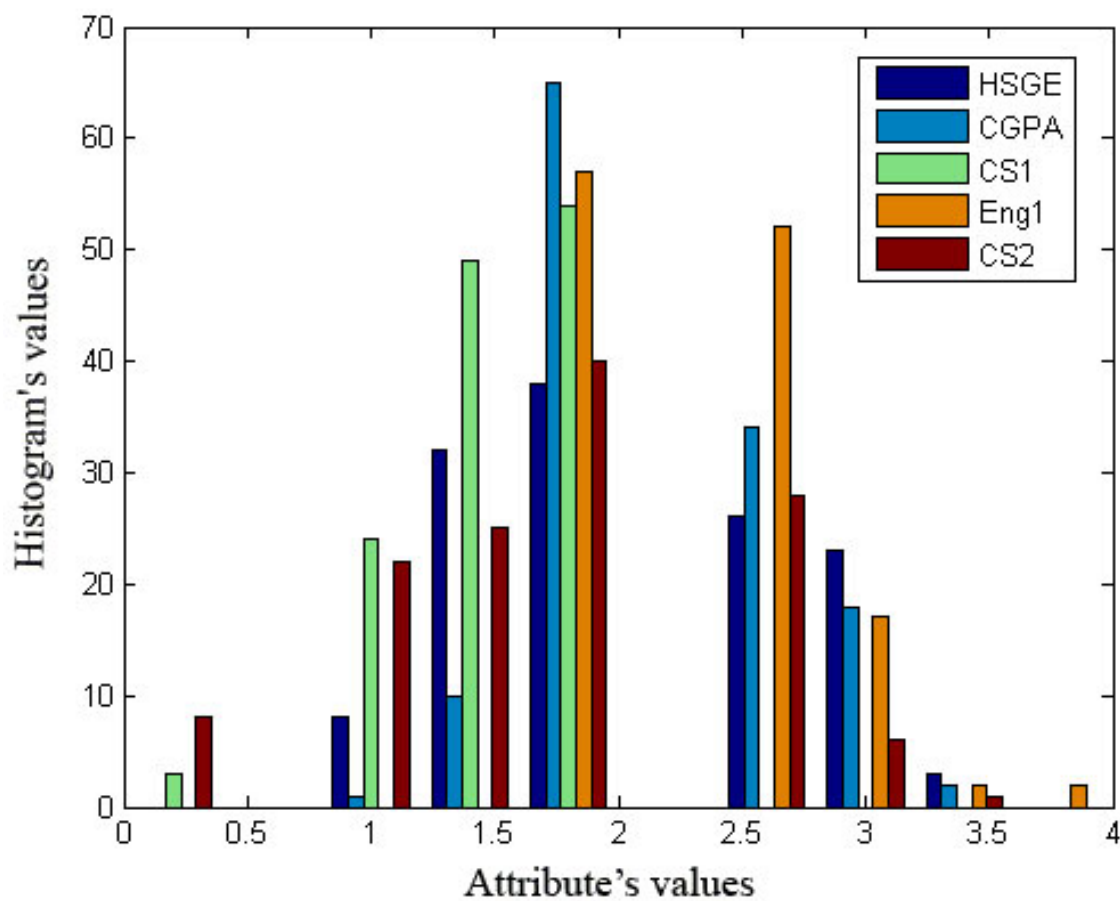


Figure B.8: Data points distribution's Histogram in Cluster6 (k=6).

C. Experiment No.2 Figures

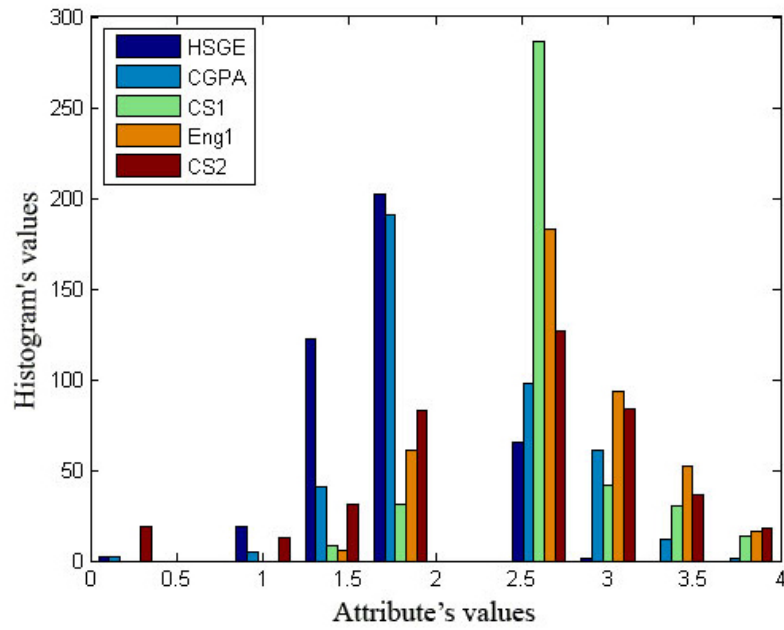


Figure C.1: Data points distribution's Histogram in Cluster1 (k=4).

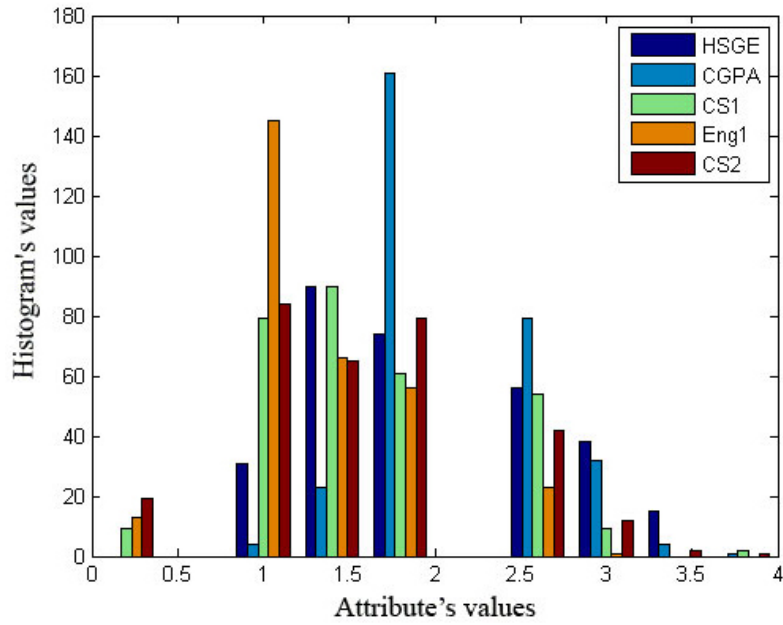


Figure C.2: Data points distribution's Histogram in Cluster2(k=4).

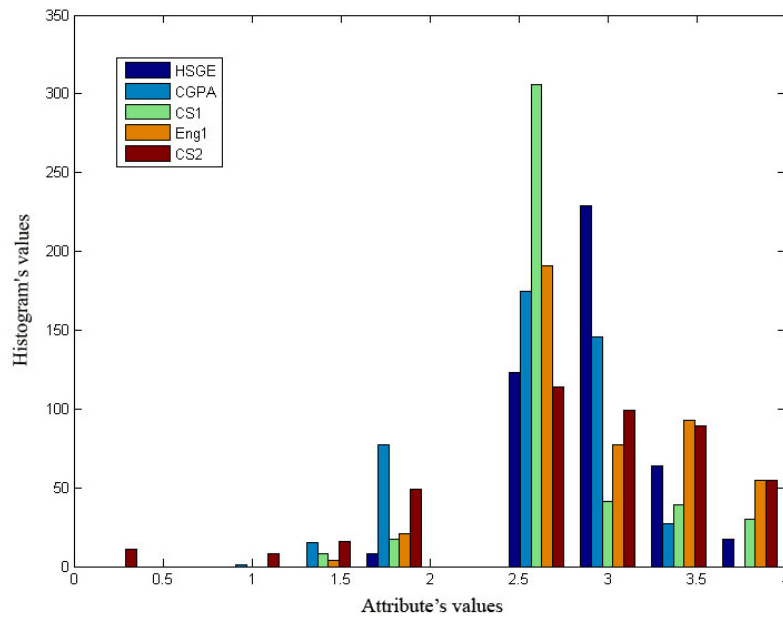


Figure C.3: Data points distribution's Histogram in Cluster3(k=4).

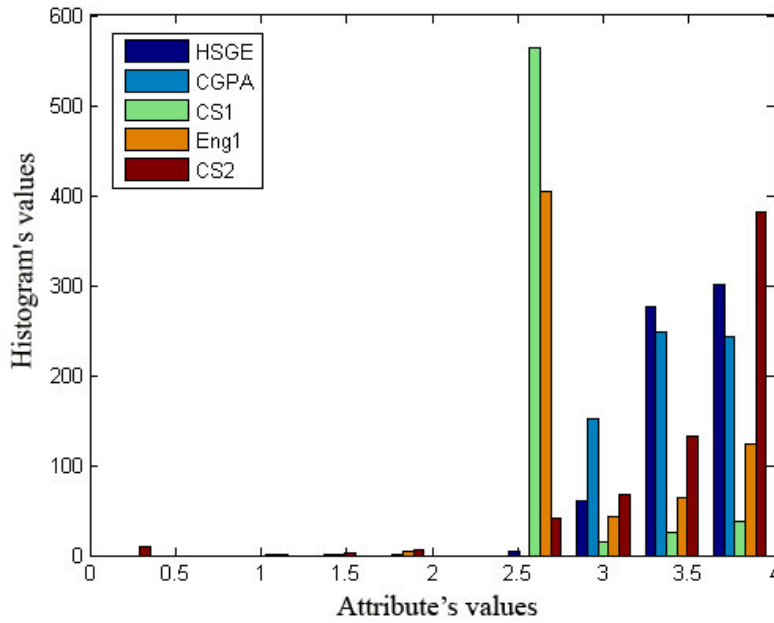


Figure C.4: Data points distribution's Histogram in Cluster4(k=4).

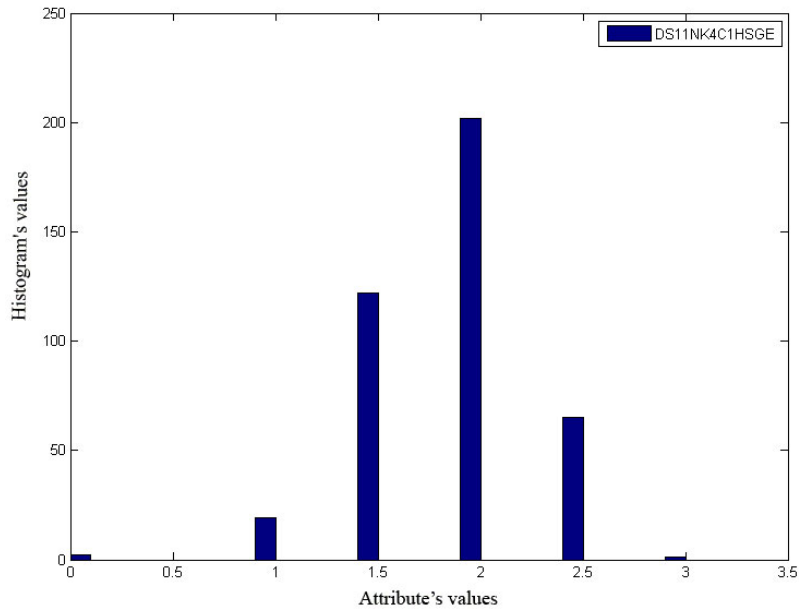


Figure C.5: HSGE Attribute's Histogram in Cluster1 (k=4).

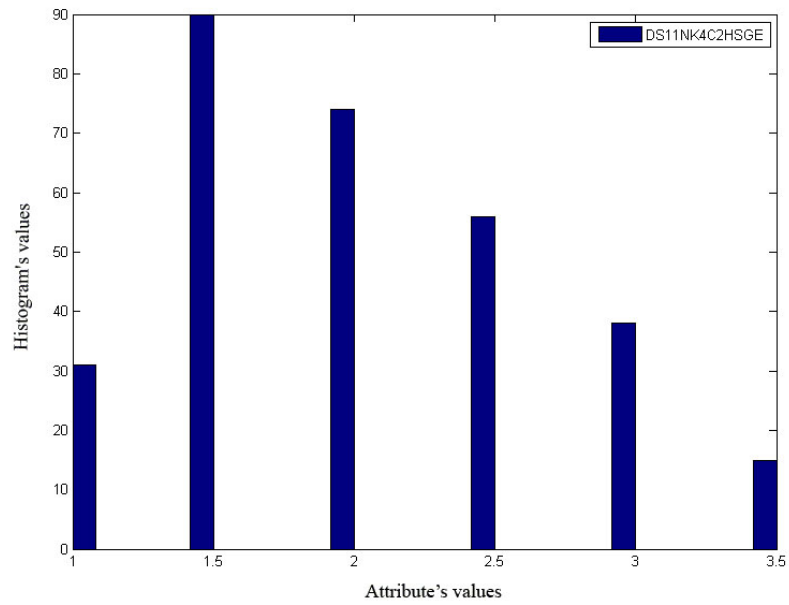


Figure C.6: HSGE Attribute's Histogram in Cluster2 (k=4).

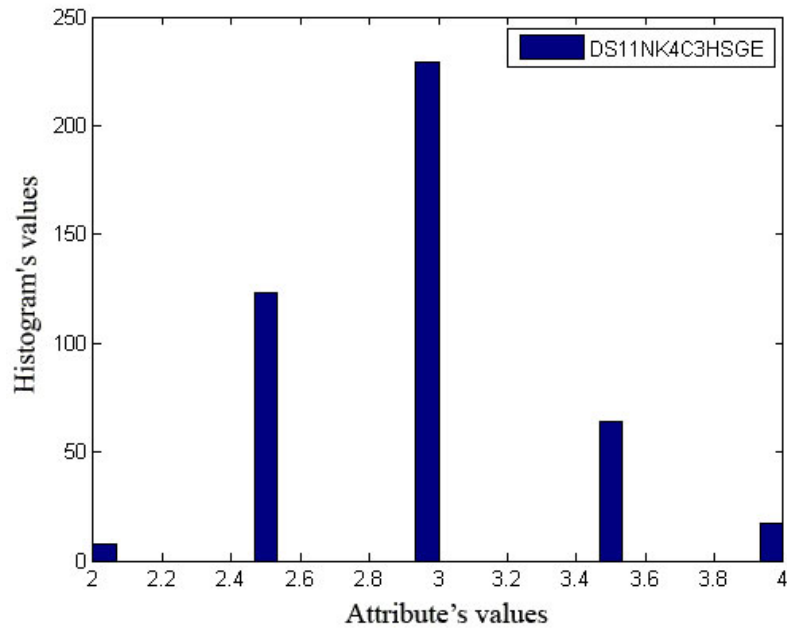


Figure C.7: HSGE Attribute's Histogram in Cluster3 (k=4).

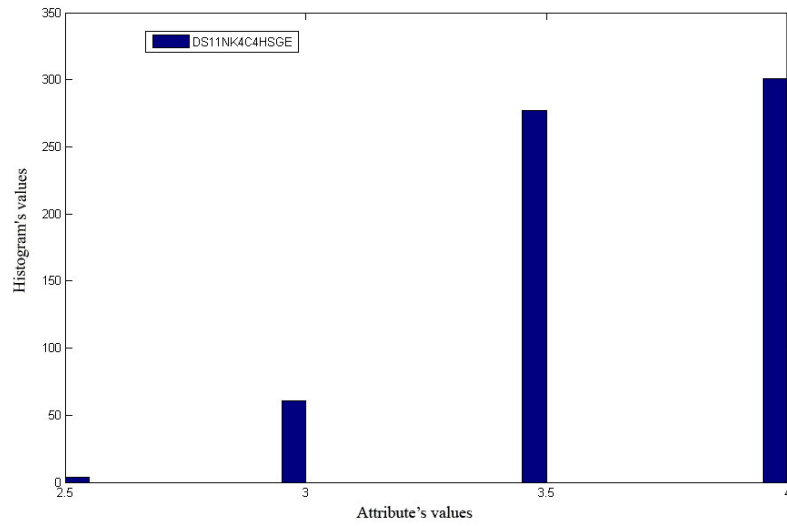


Figure C.8: HSGE Attribute's Histogram in Cluster4 (k=4).

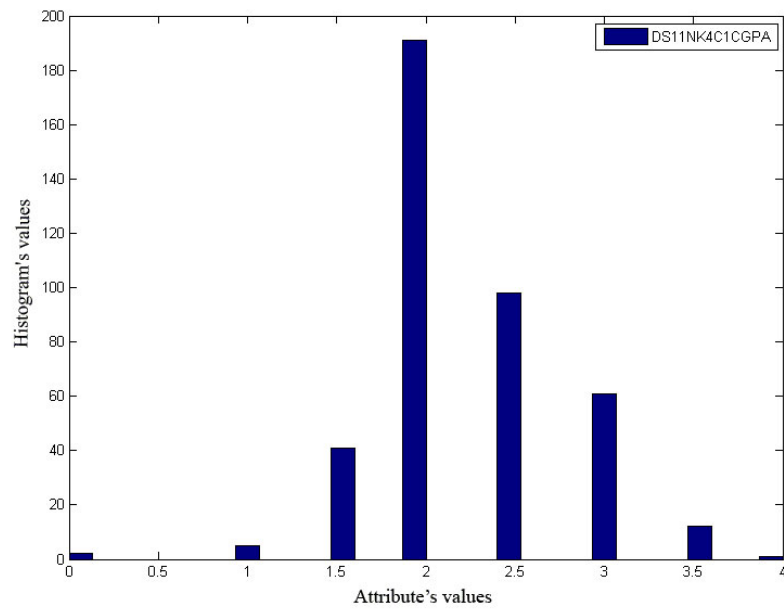


Figure C.9: CGPA Attribute's Histogram in Cluster1 (k=4).

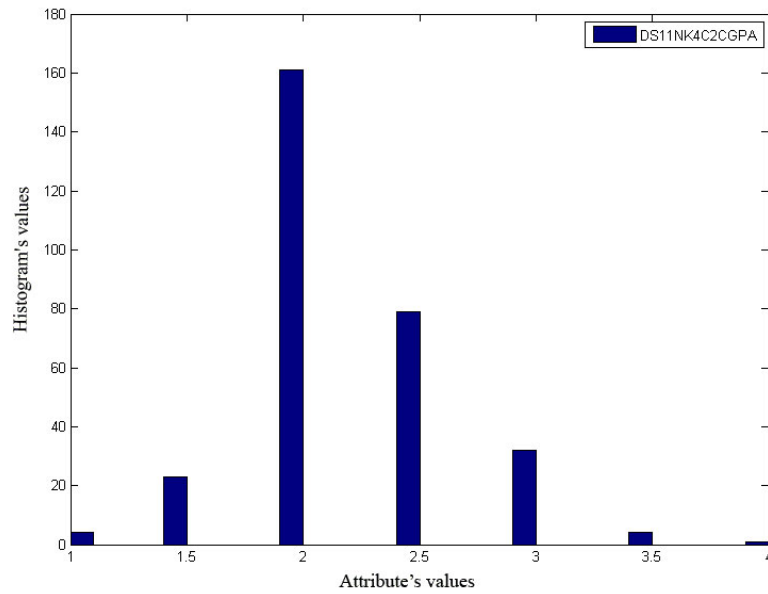


Figure C.10: CGPA Attribute's Histogram in Cluster2 (k=4).

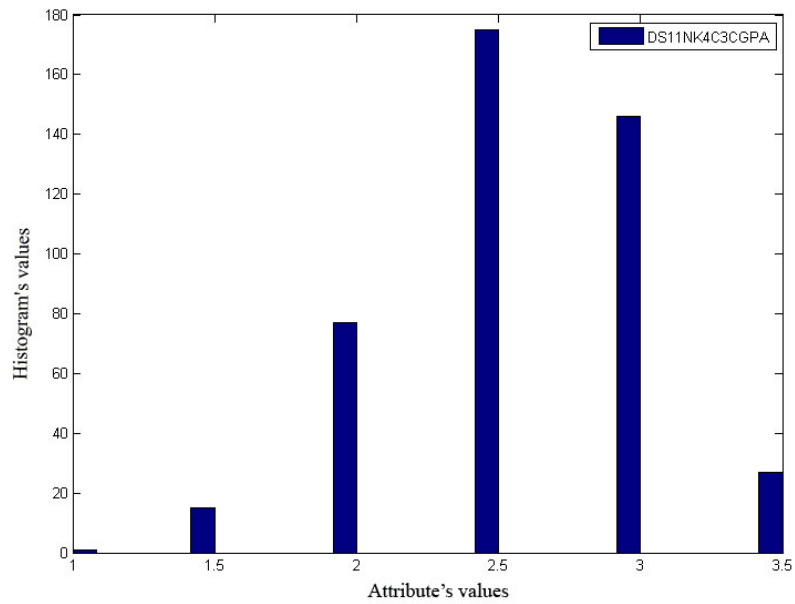


Figure C.11: CGPA Attribute's Histogram in Cluster3 (k=4).

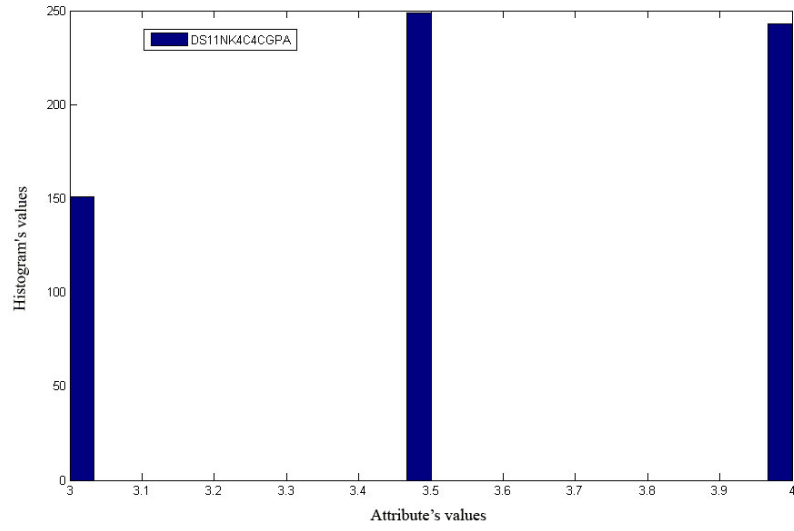


Figure C.12: CGPA Attribute's Histogram in Cluster4 (k=4).

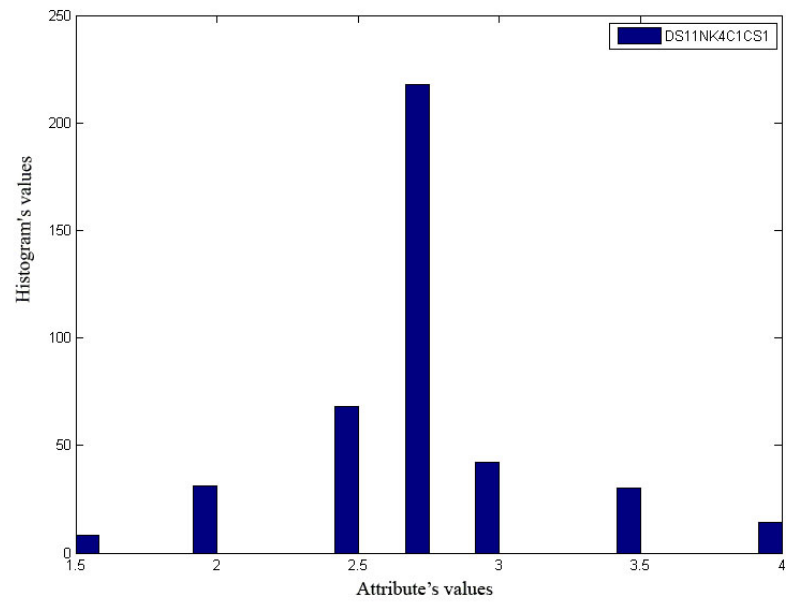


Figure C.13: CS1 Attribute's Histogram in Cluster1 (k=4).

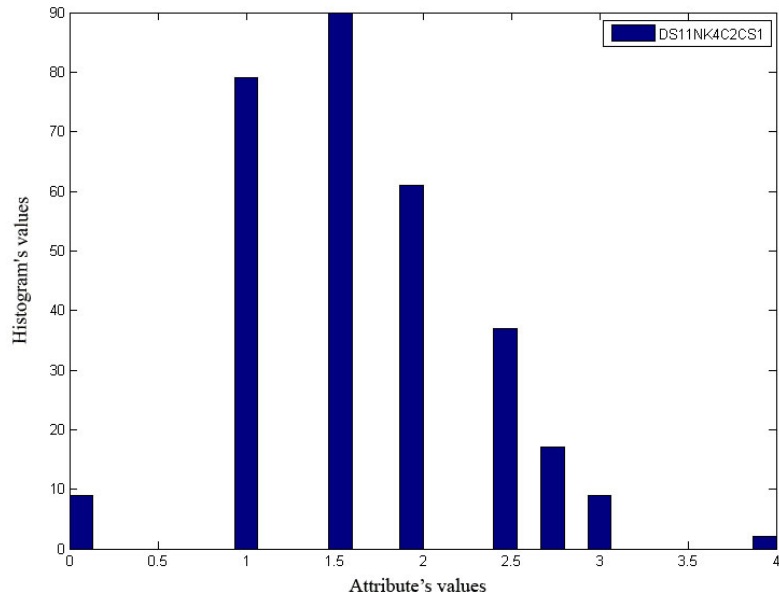


Figure C.14: CS1 Attribute's Histogram in Cluster2 (k=4).

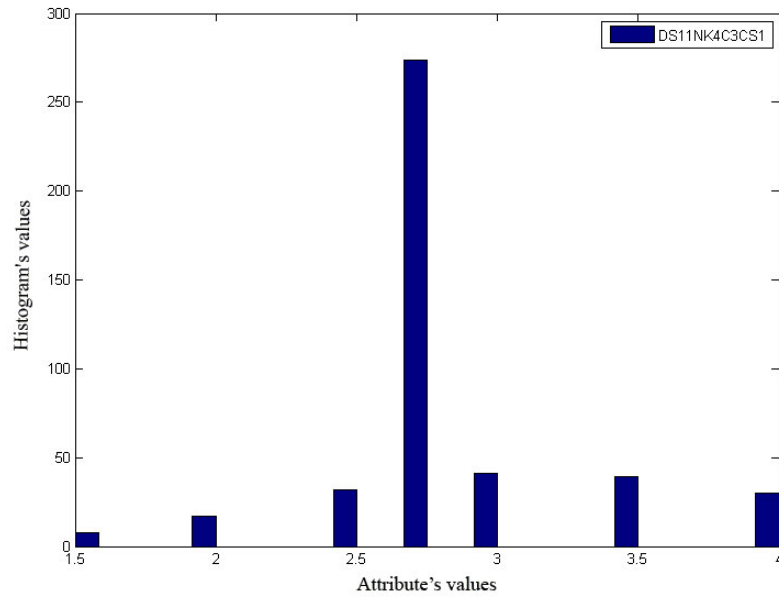


Figure C.15: CS1 Attribute's Histogram in Cluster3 (k=4).

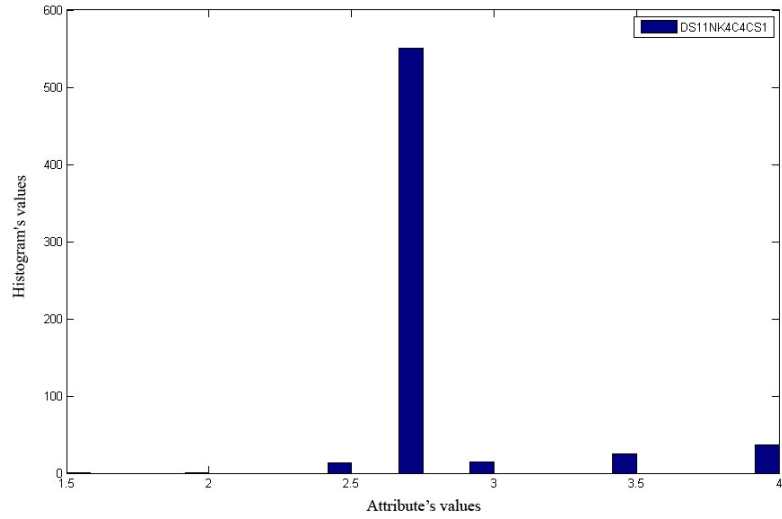


Figure C.16: CS1 Attribute's Histogram in Cluster4 (k=4).

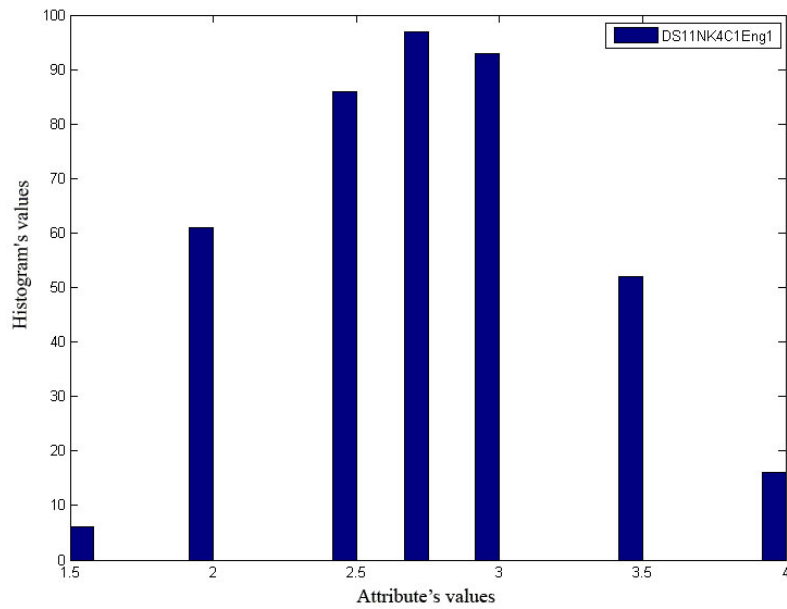


Figure C.17: English1 Attribute's Histogram in Cluster1 (k=4).

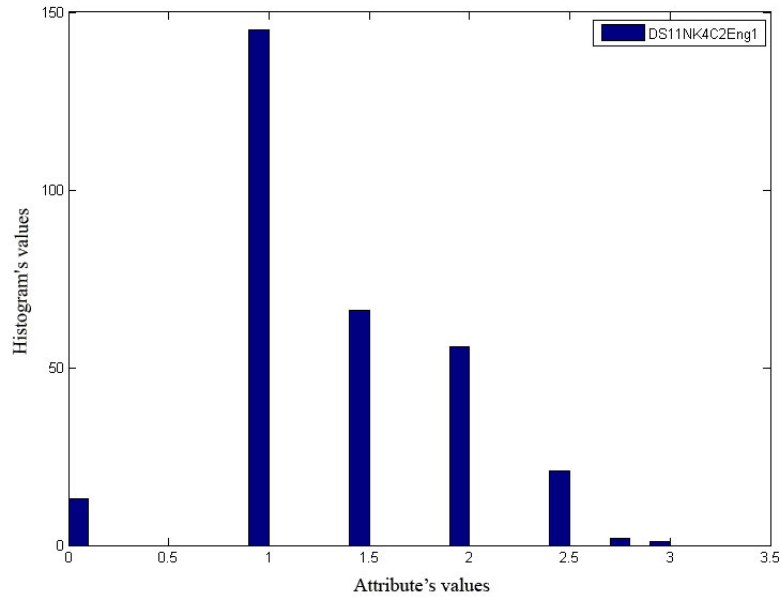


Figure C.18: English1 Attribute's Histogram in Cluster2 (k=4).

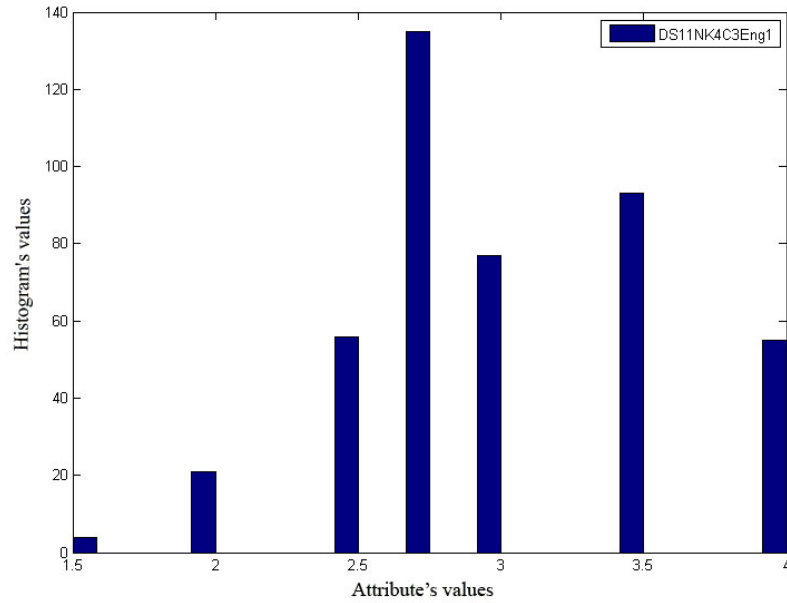


Figure C.19: English1 Attribute's Histogram in Cluster3 (k=4).

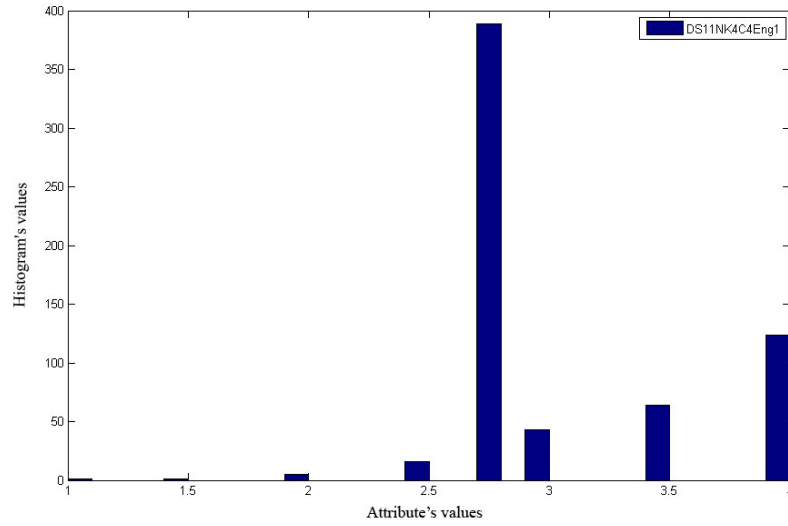


Figure C.20: English1 Attribute's Histogram in Cluster4 (k=4).

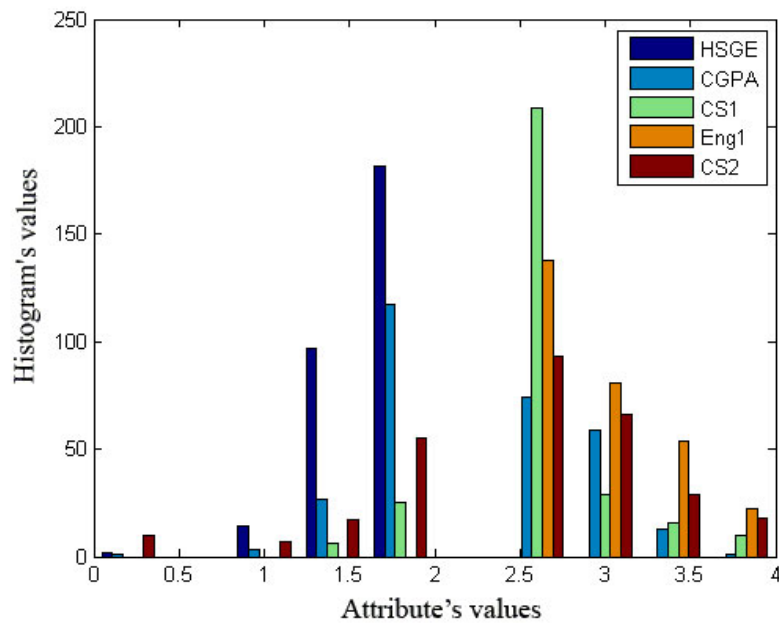


Figure C.21: Data points distribution's Histogram in Cluster1 (k=7).

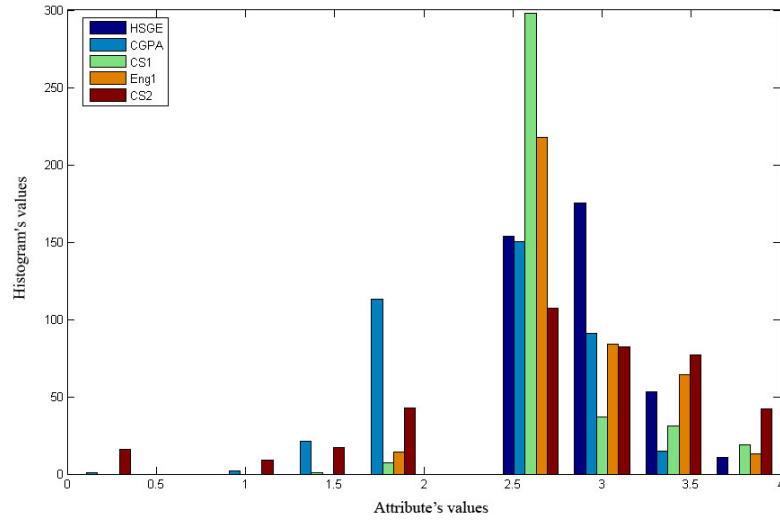


Figure C.22: Data points distribution's Histogram in Cluster2 (k=7).

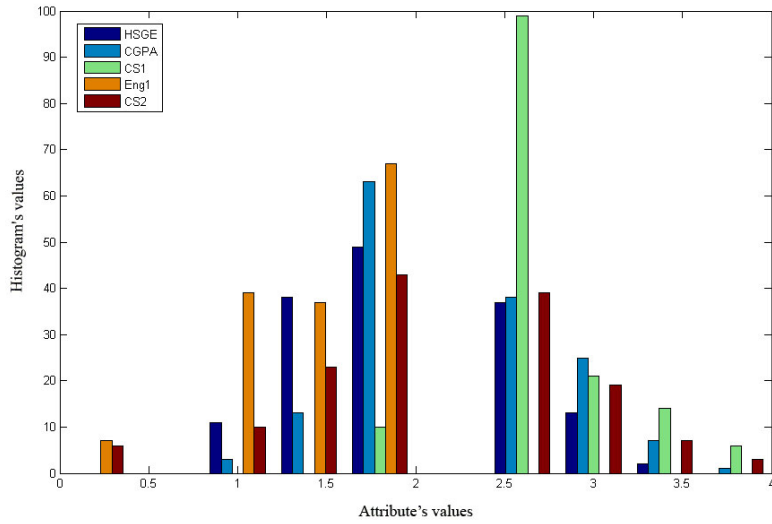


Figure C.23: Data points distribution's Histogram in Cluster3 (k=7).

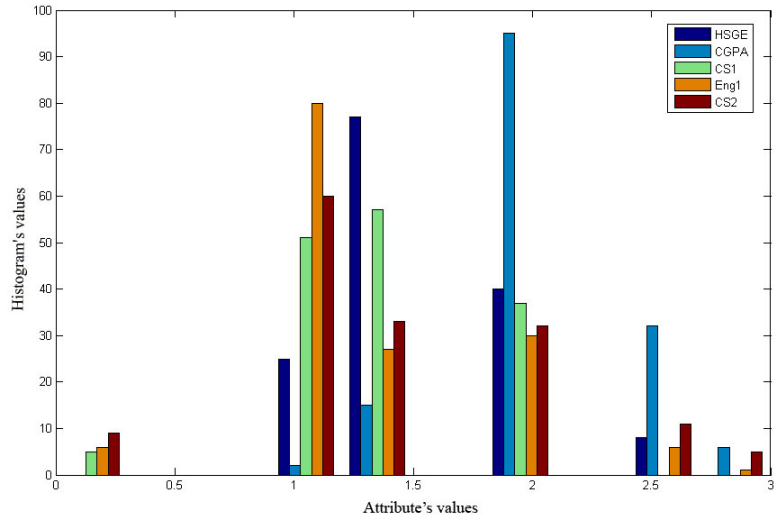


Figure C.24: Data points distribution's Histogram in Cluster4 (k=7).

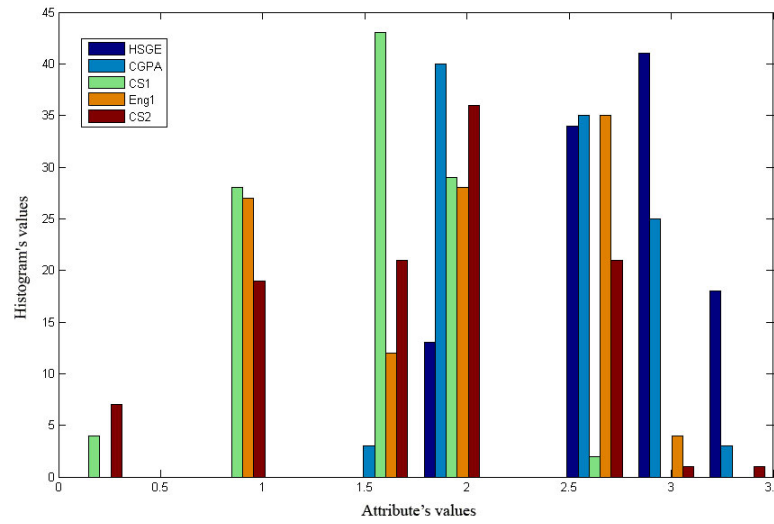


Figure C.25: Data points distribution's Histogram in Cluster5 (k=7).

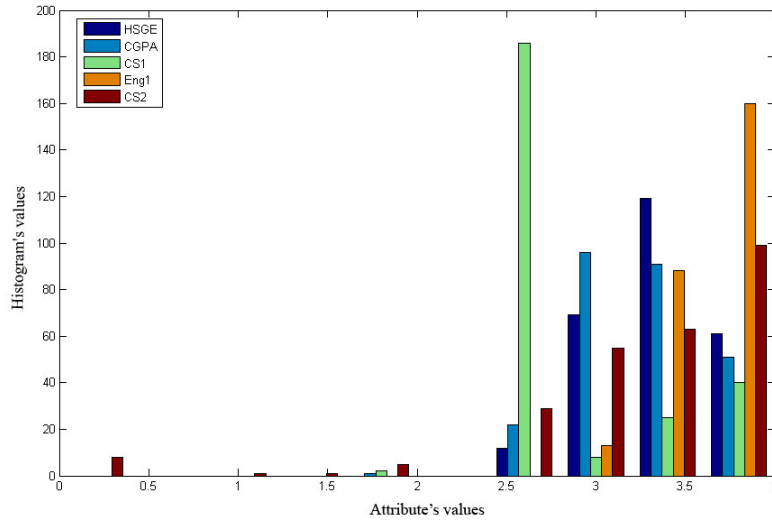


Figure C.26: Data points distribution's Histogram in Cluster6 (k=7).

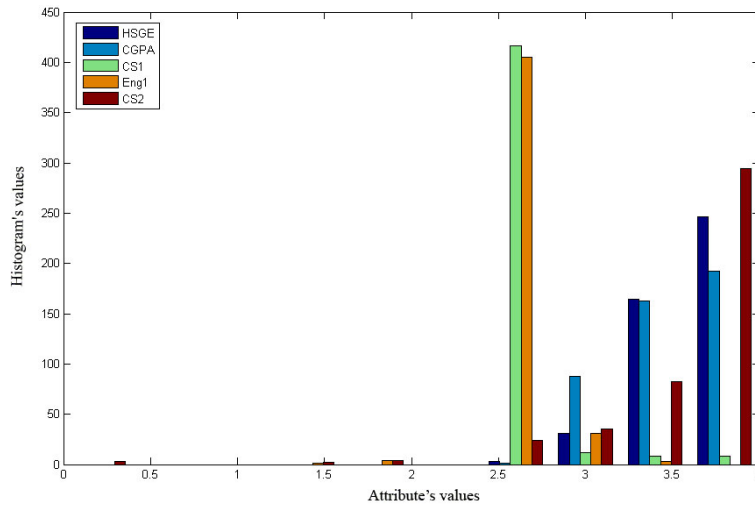


Figure C.27: Data points distribution's Histogram in Cluster7 (k=7).

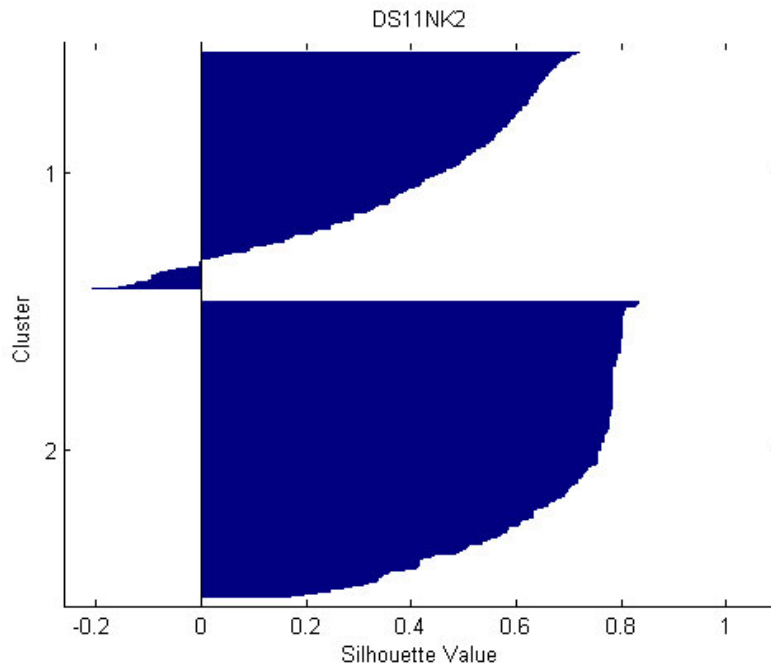


Figure C.28: Square Euclidean Distance Function Silhouettes for Experiment No.2 with $k=2$.

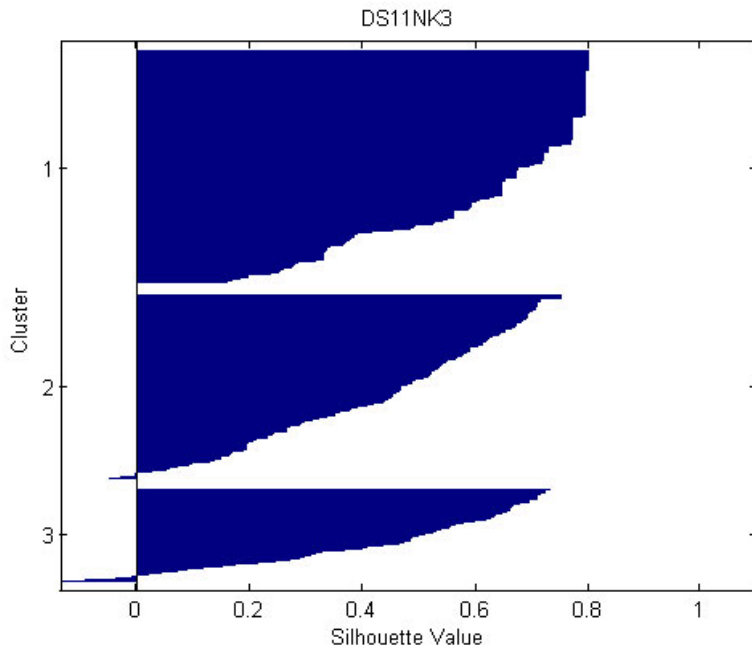


Figure C.29: Square Euclidean Distance Function Silhouettes for Experiment No.2 with $k=3$.

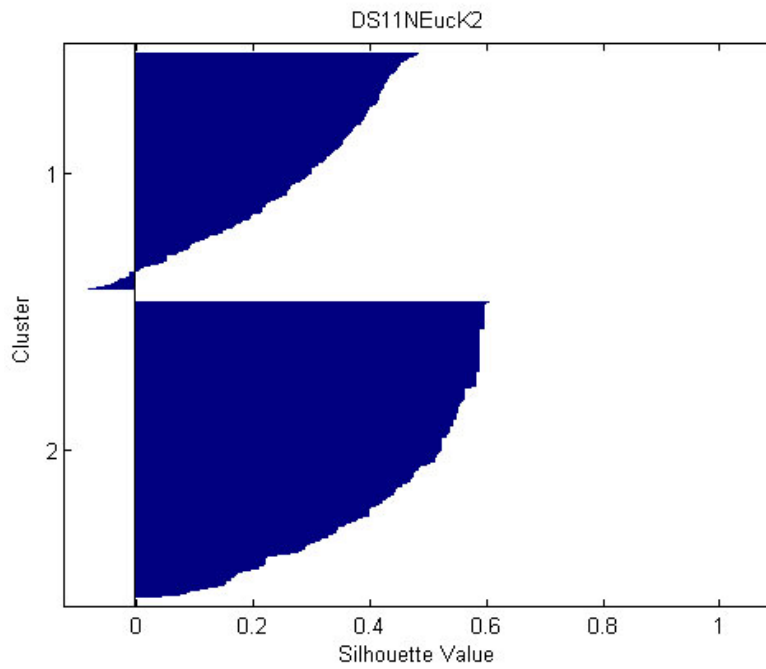


Figure C.30: Euclidean Distance Function Silhouettes for Experiment No.2 with $k=2$.

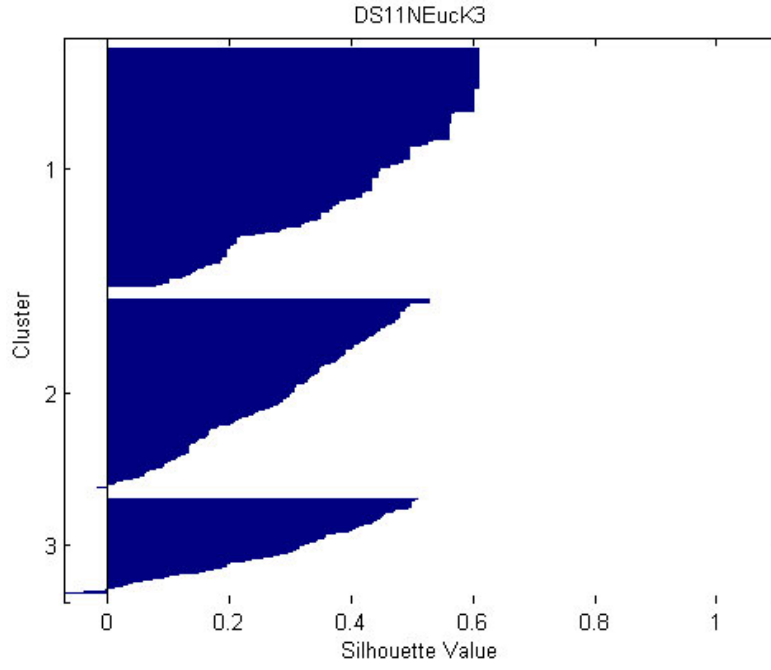


Figure C.31: Euclidean Distance Function Silhouettes for Experiment No.2 with $k=3$.

References

Abascal, E., Lautre, I. G., and Mallor, F. (2006). Data mining in a bicriteria clustering problem. *European Journal of Operational Research*, 173:705–716.

Achtert, E., Bohm, C., Kriegel, H.-P., Kroger, P., and Zimek, A. (2007). Robust, complete, and efficient correlation clustering. In *SDM*. SIAM.

Al-Zoubi, M., Salah, I., Sleit, A., Al-sharaeh, S., Huneiti, A., and Obeed, N. (2008). Efficient method for assigning student to proper groups. *European Journal of Scientific Research*, 21(3).

Amershi, S. and Conati, C. (2007). Unsupervised and supervised machine learning in user modeling for intelligent learning environments. In *IUI '07: Proceedings of the 12th international conference on Intelligent user interfaces*, pages 72–81, New York, NY, USA. ACM Press.

Arthur, D. and Vassilvitskii, S. (2005). On the worst case complexity of the k-means method. Technical Report 2005-34, Stanford InfoLab.

Ayaquica-Martinez, I., Martinez-Trinidad, J., and Carrasco-Ochoa, J. (2005). Conceptual k-means algorithm with similarity functions. In *Progress in Pattern Recognition, Image Analysis and Applications*, volume 3773/2005, pages 368–376. Springer Berlin / Heidelberg.

Bach, F. R. and Jordan, M. I. (2006). Learning spectral clustering, with application to

speech separation. *Journal of Machine Learning Research*, 7:1963–2001.

Bhatia, S. K. (2004). Adaptive k-means clustering. In *FLAIRS Conference*.

Borodin, A., Ostrovsky, R., and Rabani, Y. (2004). Subquadratic approximation algorithms for clustering problems in high dimensional spaces. *Machine Learning*, 56(1-3):153–167.

Bunn, J. and Carminati, F. (1988). *VAX Cluster User's Guide*. CERN, Geneva.

Cheng, S.-Y., Lin, C.-S., Chen, H.-H., and Heh, J.-S. (2005). Learning and diagnosis of individual and class conceptual perspectives: an intelligent systems approach using clustering techniques. *Comput. Educ.*, 44(3):257–283.

Chinrungrueng, C. and Sequin, C. H. (1995). Optimal adaptive k-means algorithm with dynamic adjustment of learning rate. *Neural Networks, IEEE Transactions on*, 6(1):157–169.

Chmielewski, M. R. and Jerzy (1996). Global discretization of continuous attributes as preprocessing for machine learning. *International Journal of Approximate Reasoning*, 15((no. 4):319–331.

Christen, P. (2007). Evaluation of a graduate level data mining course with industry participants. In *AusDM '07: Proceedings of the sixth Australasian conference on Data mining and analytics*, pages 233–241, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.

Ciriani, V., De Capitani di Vimercati, S., Foresti, S., and Samarati, P. (2008). k-anonymous data mining: A survey. In Aggarwal, C. and Yu, P., editors, *Privacy-Preserving Data Mining: Models and Algorithms*. Springer-Verlag.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 233–240, New York, NY, USA. ACM.

Egghe, L. (2008). The measures precision, recall, fallout and miss as a function of the number of retrieved documents and their mutual interrelations. *Inf. Process. Manage.*, 44(2):856–876.

Heather, S. (2006). Psychosocial risk clustering in high school students. *Social psychiatry and psychiatric epidemiology*, 41(6):498–507.

Ho, K. M. and Scott, P. D. (1997). Zeta: A global method for discretization of continuous variables. In *KDD*, pages 191–194.

Hoppner, F. and Klawonn, F. (2008). Clustering with size constraints. In Jain, L. C., Sato-Ilic, M., Virvou, M., Tsihrintzis, G. A., Balas, V. E., and Abeynayake, C., editors, *Computational Intelligence Paradigms*, volume 137 of *Studies in Computational Intelligence*, pages 167–180. Springer.

Huh, M.-H. and Lim, Y. (2009). Weighting variables in k-means clustering. *Journal of Applied Statistics*, 36(1):67–78.

Huijsmans, D. P. and Sebe, N. (2001). Extended performance graphs for cluster retrieval.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323.

Kim, H.-J. and Lee, S.-G. (2000). A semi-supervised document clustering technique for information organization. In *CIKM '00: Proceedings of the ninth international conference on Information and knowledge management*, pages 30–37, New York, NY, USA. ACM.

Klawonn, F. and Hoppner, F. (2006). Equi-sized, homogeneous partitioning. In Gabrys, B., Howlett, R. J., and Jain, L. C., editors, *KES (2)*, volume 4252 of *Lecture Notes in Computer Science*, pages 70–77. Springer.

Kogan, J., Nicholas, C., and Teboulle, M. (2006). *Grouping Multidimensional Data: Recent Advances in Clustering*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Lingras, P. and Yao, Y. Y. (2002). Time complexity of rough clustering: Gas versus k-means. In *TSCTC '02: Proceedings of the Third International Conference on Rough Sets and Current Trends in Computing*, pages 263–270, London, UK. Springer-Verlag.

Liu, B. (2007). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer.

Maimon, O. and Rokach, L. (2005). Introduction to knowledge discovery in databases. In Maimon, O. and Rokach, L., editors, *The Data Mining and Knowledge Discovery Handbook*, pages 1–17. Springer.

Mehlitz, M., Bauckhage, C., Kunegis, J., and Albayrak, S. (2007). A new evaluation measure for information retrieval systems. In *SMC*, pages 1200–1204. IEEE.

Mishra, N., Ron, D., and Swaminathan, R. (2004). A new conceptual clustering framework. *Mach. Learn.*, 56(1-3):115–151.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York.

Nakkrasae, S. (2004). R.edwards. fuzzy subtractive clustering based indexing approach for software components classification. *International Journal of Computer & Information Science*, 5.

Pan, J. J., Yang, Q., Yang, Y., Li, L., Li, F. T., and Li, G. W. (2007). Cost-sensitive data preprocessing for mining customer relationship management databases. *IEEE Intelligent Systems*, 22(1):46–51.

Pun, W. and Ali, A. (2007). Unique distance measure approach for k-means (udma-km) clustering algorithm. In *CD proceeding of The IEEE international conference*, pages 1 – 4, TENCON 2007 - 2007 IEEE Region 10 Conference. Piscataway, NJ, USA : IEEE Operations Center.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420.

Santos, A., Vaughn, B., and Bost, K. (2008). Specifying social structures in preschool

classrooms: descriptive and functional distinctions between affiliative subgroups. *Acta ethologica*, 11(2):101–113.

Savvion Incorporated (1999-2006). *Clustering Guide*. Savvion Incorporated, Sybase Inc.

Sharav-Schapiro, N., Palmor, Z. J., and Steinberg, A. (1999). Dynamic robust output min-max control for discrete uncertain systems. *J. Optim. Theory Appl.*, 103(2):421–439.

Shen, Q. and Chouchoulas, A. (2001). Rough set-based dimensionality reduction for supervised and unsupervised learning. *Applied Mathematics and Computer Science, Special Issue on Rough Sets and their Applications*, 11(3):583–601.

Song, M. and Rajasekaran, S. (2005). Fast k-means algorithms with constant approximation. In *ISAAC*, pages 1029–1038.

Tian, J., Zhu, L., Zhang, S., and Liu, L. (2005). Improvement and parallelism of k-means clustering algorithm. *Tsinghua Science & Technology*, 10:277–281.

Vrahatis, M. N., Boutsinas, B., Alevizos, P., and Pavlides, G. (2002). The new k-windows algorithm for improving the k-means clustering algorithm. *Journal of Complexity*, 18:375–391.

Wagstaff, K., Cardie, C., Rogers, S., and Schrodl, S. (2001). Constrained k-means clustering with background knowledge. In Brodley, C. E. and Danyluk, A. P., editors, *ICML*, pages 577–584. Morgan Kaufmann.

Weiss, G., editor (1997). *Distributed Artificial Intelligence Meets Machine Learning* -

Learning in Multi-Agent Environments, volume 1221 of *Lecture Notes in Computer Science*. Springer.

Wu, S., Crestani, F., and Bi, Y. (2006). Evaluating score normalization methods in data fusion. In *Information Retrieval Technology, AIRS 2006*, pages 642–648.

Yang, G., Mukherjee, S., and Ramakrishnan, I. V. (2003). On precision and recall of multi-attribute data extraction from semistructured sources. *Data Mining, IEEE International Conference on*, 0:395.

Yang, H., Wang, J., Shao, X., and Wang, N. S. (2007). Information system continuous attribute discretization based on binary particle swarm optimization. *Fuzzy Systems and Knowledge Discovery, Fourth International Conference on*, 3:173–177.

Yin, X., Han, J., and Yu, P. S. (2005). Cross-relational clustering with user's guidance. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 344–353, New York, NY, USA. ACM.

Zalmai, G. J. (2007). Parametric duality models for discrete minmax fractional programming problems containing generalized ρ -invex functions and arbitrary norms. *J. Appl. Math. Comput.*, 24(1):105–126.

Zhao, Y., Zhang, C., Zhang, S., and Zhao, L. (2006). Adapting k-means algorithm for discovering clusters in subspaces. In *APWeb*, pages 53–62.

Zighed, D. A., Rakotomalala, R., and Feschet, F. (1997). Optimal multiple intervals discretization of continuous attributes for supervised learning. In *KDD*, pages 295–298.

تحسين عملية توزيع طلبة مادة مهارات حاسوبية 2 على الشعب من خلال تقنيات تنقيب البيانات

إعداد
إسراء فواز الزغول

المشرف
الدكتور عمار محمد الحنيطي

المشرف المشارك
الدكتور عماد خالد صلاح

ملخص

تَجْمِيعُ البياناتِ أحد أهم الأدوات لَتَحْلِيلِ تركيب مجموعة المعلومات. قد طبقت في حقول مُخْتَلِفَة مثل التعلم الآلي، تنقيب البيانات تحليل الصور، استرجاع المعلومات. . . الخ. إن المشاكل الأكثر صعوبة في التحليل العنقودي تعريف عدد العناقيد في مجموعة المعلومات، و إختيار الفئة التي ينتمي إليها العضو وإيجاد طريقة استخدام المسافة المناسبة لقياس المسافة بين كل عنصر و وسيط العنقود. يتحرى هذا البحث مشكلة علامات الطلاب في مادة مهارات الحاسوب- 2 معتمدا على خوارزمية الوسيط "ك" التي تعتمد على تقنية التجميع العنقودي. طبقت هذه التقنية في الجامعة الأردنية على الطلاب الذين أخذوا مادة مهارات الحاسوب - 2 بشكل خاص.

خوارزمية الوسيط "ك" هي خوارزمية التجميع العنقودي التي تُسْتخدَم لتمييز مجموعات الطلاب الذين قد يشتركون في الإهتمامات المتشابهة.

إن الهدف الرئيسي لهذا البحث أن يجد الحل لمشاكل تصنيف الطلاب في مادة مهارات الحاسوب-2. المشكلة الرئيسية في هذه المادة تكمن في مقياس درجات الطلاب في نهاية الفصل الدراسي. إذا كان من الممكن إيجاد مجموعات مماثلة من الطلاب تعتمد على خلفياتهم المعرفية و معلوماتهم السابقة، نوع التعليم لديهم، الانضباط، القدرات، المهارات، فإن المنسق لهذه المادة بإمكانه توزيع الطلاب بحيث يكون هناك مستويات متقاربة من الطلاب. هذا التخصيص سيُمكن المدرسين من تزويد الطلاب بالمواضيع المعينة، المساعدة، النصيحة، المواد، الامتحانات، التمارين، طبقاً لمستوى المعرفة المناسب لكل شعبة. هذا قد يُمكن الطلاب من كسب معرفة أفضل، فهم أفضل، الحصول على العلامات الأعلى.

إن الأهداف الرئيسية لهذا البحث هي:

- 1- إيجاد العلاقة بين قدرات الطلاب وإنجازاتهم الأكاديمية في مادة مهارات الحاسوب-2.
- 2- إيجاد مجموعات من الطلاب لهم نفس الإهتمامات وبمعنى آخر: لهم نفس الخلفيات، المعرفة، نوع التعليم، القدرات، المهارات.
- 3- تكيف المادة إلى هذه المجموعات من الإهتمامات المتشابهة.
- 4- تحسين إنجازات الطلاب والفجوات الأكاديمية بين الطلاب ذوي الخلفيات المعرفية المتفاوتة. لحل مشكلة تصنيف الطلاب في مادة مهارات الحاسوب-2، فإن هذا البحث يسعى لإيجاد العلاقة بين اهتمامات الطلاب وإنجازاتهم الأكاديمية. و يُقدّم نظرة أيضاً في التزوّد و الدعم في إيجاد مجموعات الطلاب ذو الإهتمامات المشتركة. هذه النظرة ستساعد أيضاً في تكيف مادة مهارات الحاسوب-2 إلى هذه المجموعات من الإهتمامات المشتركة. تحسّين إنجازات الطلاب و التقليل من الفجوات الأكاديمية بين الطلاب. العديد من الاختبارات التجريبية عمّلت على العديد من مجموعات المعلومات.
- مسعى البحث سيّتحريّ مشكلة الدرجات في مادة مهارات الحاسوب-2 بالتحقيق مستندة على خوارزمية التجميع للوسيط "ك"، هذه التقنية ستطبق على الطلاب الذي تقدموا لمادة مهارات الحاسوب-2 لاكتشاف وإيجاد النتائج.
- التجميع قدّم لكي يكون أحد أكثر تقنيات تحليل البيانات المستعملة عموماً. و له تاريخ طويل أيضاً، وقد أستعمل تقريباً في كلّ الحقول والمجالات، ومثال على ذلك: في الطبّ، علم نفس، علم نبات، علم إجتماع، علم أحياء، التسويق، التأمين، علم مكتبات.