Academic Curriculum (Study Path) Mining of Mechanical Engineering Undergraduate Students

BY

MARYAM TEIMOORI

B.S. ELECTRICAL ENGINEERING, AMIRKABIR UNIVERSITY

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Masters of Science in Industrial Engineering
in the Graduate College of the
University of Illinois at Chicago, 2016

Chicago, Illinois

Defense Committee:

Houshang Darabi, Chair and Advisor
Michael J. Scott, Mechanical and Industrial Engineering
Mengqi Hu, Mechanical and Industrial Engineering

This thesis is dedicated to my husband, Anooshiravan and my daughter Shailene Sharabiani who have been a constant source of love and encouragement in my life.

ACKNOWLEDGMENTS

I would like to seize this opportunity and thank several people without whom this work would never be possible.

First and foremost, I would like to thank my advisor, Professor Houshang Darabi, who offered his continuous advice and encouragement throughout the course of this thesis. I thank him for his excellent guidance, understanding, and support. I doubt that I will ever be able to convey my appreciation fully, but I owe him my eternal gratitude.

I would like to express my gratitude to my committee members, Michael Scott and Mengqi Hu for generously offering their time throughout the preparation and review of this document.

I am also thankful to Ashkan Sharabiani, who was always willing to help and give his best suggestions.

Finally, my sincere thanks goes to my parents and to my husband for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis.

TABLE OF CONTENTS

1 INTRODUCTION 8
1.1 What is a Curriculum? 8
1.2 Chapter Synopsis 9
2 LITERATURE REVIEW 10
3 MATHEMATICAL BACKGROUND 12
3.1 K-Means Clustering 12
3.2 Logistic regression 12
3.3 Decision Tree and Random Forest 13
3.4 Neural Network 14
3.5 Cross Validation 15
3.6 Model Evaluation 15
3.7 Ensemble methods 16
3.8 Model selection 16
4 METHODS AND MODELS 17
4.1 Data sources and data preparation 17
4.2 Data visualization 21
4.3 Current official curriculum evaluation 22
4.4 Curriculum mining using Clustering 23
4.5 Developing prediction models to detect the study path 25
5 RESULTS AND DISCUSSION 30
5.1 Evaluation of clustering result 30
5.2 Analysis of study paths 32
5.3 Evaluation of study path prediction result 33
5.4 Courses overlap calculation 36
6 CONCLUSION AND FUTURE WORK 38
CITED LITERATURE 40

LIST OF TABLES

Table 1. Sample study path profile for a given student 19
Table 2. Equivalent semester according to transferred credit hours 20
Table 3. Study paths course-semester comparison 32
Table 4. The confusion matrix of step 1 prediction 34
Table 5. The confusion matrix of step 2 prediction 34
Table 6. The combination result of prediction in two steps 34

LIST OF FIGURES

Figure 1. An example of a proposed study path..... 2
Figure 2. Finding semester numbers that courses are taken..... 11
Figure 3. One snapshot of the visualization tools for reviewing students' behavior 14
Figure 4. ME proposed study path respect comparison 22
Figure 5. The centroid points of students study path clusters 23
Figure 6. Comparison of centroids and raw data distributions in each cluster for IE201 course..... 23
Figure 7. The overview of prediction steps..... 23
Figure 8. Input variables of prediction model in step 1 19
Figure 9. Input variables of prediction model in step 2 19
Figure 10. The Structure of prediction model..... 20
Figure 11. Clustering performance with different number of clusters (K) 23
Figure 12. Clusters detail -student entrance types histograms 33
Figure 13. Overlap ratio calculation example..... 30

LIST OF ABBEREVAIATIONS

ANN	Artificial Neural Network
DT	Decision Tree
LS	Logistic Regression
RF	Random Forest
DS	Decision System
UIC	University of Illinois in Chicago
MIE	Mechanical and Industrial Engineering Department
ME	Mechanical Engineering
IE	Industrial Engineering
EDM	Educational Data Mining
EDDIE	Enterprise Data Delivery Information Environment
TA	Teaching Assistance
GPA	Grade Point Average
WCSS	Within-Cluster Sum of Squares
CV	Cross Validation
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
AH	Academic History
UIN	University Identification Number

SUMMARY

Exploring the study paths of students in higher education is crucially important in order to maximize their likelihood of academic success. The study path of a student is the sequence of courses that the student takes in order to graduate. The main focus of this dissertation is to identify and analyze the study paths of Mechanical Engineering (ME) undergraduate students at the University of Illinois at Chicago (UIC). Using students' academic records, several machine learning techniques were applied in order to explore the behavior of students from admission to graduation. The major contributions of this thesis are as follows. First, it determines whether the current ME university approved study path is consistently followed by students. Using clustering methods, the study paths that are actually followed by students are derived. The graduation time and the Grade Point Average (GPA) of students in each study path are measured. Furthermore, classification techniques are used to predict the study path of new incoming students. Second, a new index in defined in order to calculate the overlap ratio of the courses taken by ME students. This index is used to measure students' tendency to take any given pair of courses in the same semester.

The contributions of this thesis can support both university students (through advising) and their academic departments (through scheduling). Students can benefit from receiving an enhanced advising (based on selected study path) that may result in a shorter graduation time and a higher GPA in each semester. Departments can achieve a more accurate course enrollment prediction and hence, provide a better course scheduling. Moreover, based on the study path clusters, students' graduation time can be projected which can help departments in assigning their classroom and teaching resources more efficiently.

1 INTRODUCTION

In this Chapter, the introduction of curriculum mining and the approaches used in educational data mining are presented.

1.1 What is a Curriculum?

An academic curriculum refers to a predetermined proposed sequence of courses for students to follow in each semester. This curriculum includes all courses required for a student to graduate. The curriculum is not usually mandatory; however, there are constraints on the study path in order to maintain a logical sequence of the prerequisites courses. An example of a study path is presented in Figure 1. In this example, the required courses for the mechanical engineering program are suggested through a proposed study path that is divided into 8 semesters. The objective of an ideal curriculum is to recommend a study path that improves a student's probability of accomplishing educational achievements.

Student data from 2005 through 2014, which included 521 students, was used. The objective is to first determine if the curriculum was always respected in the past in the MIE department. Next, it is to classify students according to their study path, and finally using that information to predict the path of new students. The results of this dissertation helps both students and the department in advising and scheduling, respectively.

- Students benefit from advising that can result in them experiencing a shorter graduation time frame, higher GPA and a balanced workload for each semester.

- Departments benefit from minimalized schedule conflict, reduced unused capacity of resources (classes, instructors, TAs, etc.), and a balanced workload for instructors.

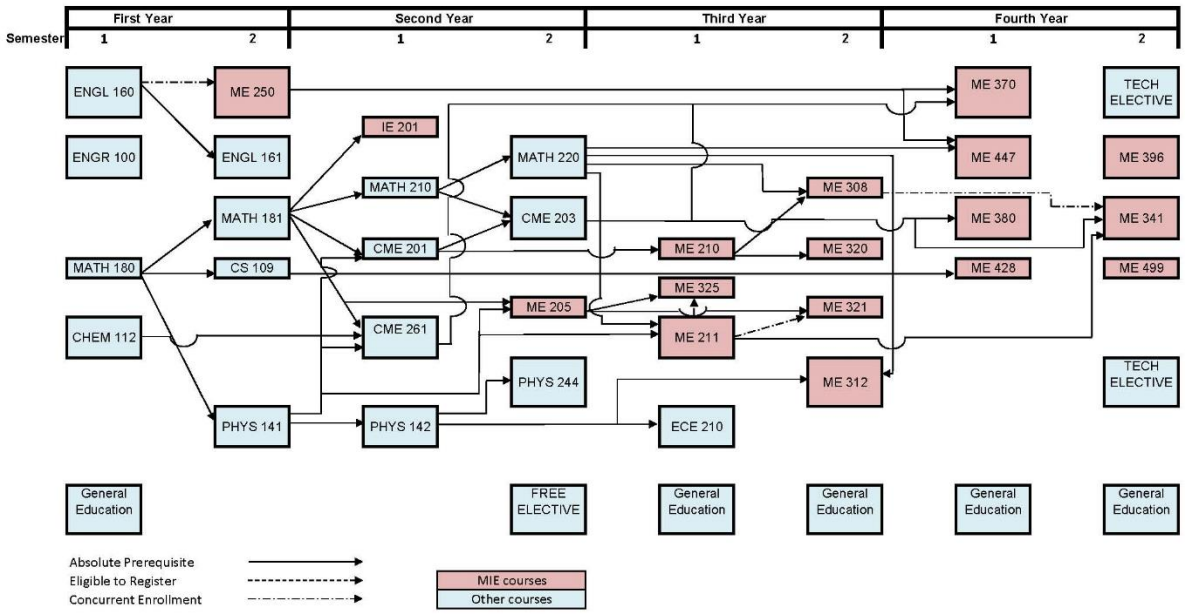


Figure 1 Example of a proposed study path

1.2 Chapter Synopsis

In Chapter 2, literature review of Curriculum Mining is presented. The data mining techniques used in this dissertation are K-means, logistic regression, random forest, and neural network methods. These methods are briefly described in Chapter 3. In Chapter 4, we discuss data resource, data preparation, data visualization, clustering, and prediction techniques on curriculum mining. Clustering is used to detect popular study paths of students, while prediction models are used to predict the study path that each student would fall into. The results of clustering and the predictions models are presented in Chapter 5 along with the analysis and evaluation of popular study paths

and course overlap measure. Finally, the conclusion and future works are presented in Chapter 6.

2 LITERATURE REVIEW

Educational data mining (EDM) and process mining techniques have been used to analyze and identify common study paths of students. [10] Describes, and defines different categories of users in educational environments along with the data they provide. It also provides a list of the most common tasks in the educational environment that has been applied using data and process mining techniques. In [8], the advance development in educational data mining (EDM) is summarized and the data and process mining results are reviewed.

[12] Illustrates an example of academic curriculum mining through a process mining framework including three main tasks: model discovery, curriculum model conformance checking, and curriculum model extensions. Curriculum model discovery, models academic curriculums to reproduce student behavior. Curriculum model conformance checking, verifies whether the behavior of the student reflects the expected behavior based on curriculum model. Finally, the curriculum model extension, projects information into the model in order to help understand of the academic processes.

[13] Proposes utilizing process and data mining on curriculum mining by using Colored Petri net and standard patterns. This study attempts to provide a mean to compare successful and less successful students, as well as develop recommendations for students to take courses based on their expected performances. First, a process

model of students taking courses is discovered. Then, paths in which successful and less successful students are likely to undergo are highlighted. Finally, the optimal path is recommended to students. This proposed method can be used in the analysis of students' study path patterns in order to enhance the design of a curriculum.

This dissertation primarily incorporates data mining rather than process mining techniques. K-means was used for clustering students based on their study paths, which resulted in three main classes of study paths: Fast Track, Regular Curriculum, and Extended-Time Curriculum. An ensemble of logistic regression, random forest, and neural network was also applied in order to predict the study path of students in semester 2 and semester 4.

3 MATHEMATICAL BACKGROUND

In this Chapter, K-Means Clustering and three classification techniques (Logistic Regression, Random Forest and Neural Network) are briefly presented.

3.1 K-Means Clustering

K-means clustering is an algorithm that attempts to discover categories in data [4]. K-means clustering aims to divide n observations $(x_1, x_2, .., x_n)$, into K $(\leq n)$ clusters. K clusters can be considered as a set $S = \{S_1, S_2, .., S_K\}$. Each observation is associated with the cluster that has the nearest mean to that observation. The algorithm attempts to minimize the within-cluster sum of squares (WCSSj):

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

Where μ_i is the mean of points in S_i

3.2 Logistic regression

Logistic regression is a classification method used when the target variable (response) is binary or multinomial (categorical). For example, in binary case if we consider Y as the binary target variable, it can be assumed that $P(Y=1)$ is dependent on the values of explanatory variables $\vec{x}$ $(x_1, x_2, .., x_p)$. There for the goal is to find $p(\vec{x})$ where $p(\vec{x}) \equiv P(Y \mid \vec{x})$.

Finding $p(\vec{x})$ is equivalent to modeling $E(Y \mid \vec{x})$, which can be done in ordinary least square (OLS) regression method, with a limitation on target variable $p(\vec{x})$.

The target variable should be between 0 and 1. Thus, a link function is defined to handle this limitation. Logit function is defined as $\log\left(p(\vec{x})/[1-p(\vec{x})]\right)$.

This function can be modeled as a linear function of explanatory variables:

$$\log\left(\frac{p(\vec{x})}{1-p(\vec{x})}\right)=\beta_0+\beta_1 x_1+...+\beta_p x_p$$

Where $x_1,...,x_p$ are explanatory variables. This model can be used to estimate the probabilities by:

$$p(\vec{x})=\frac{\exp(\beta_0+\beta_1 x_1+...+\beta_p x_p)}{1+\exp(\beta_0+\beta_1 x_1+...+\beta_p x_p)}$$

The coefficients of the model ($\beta_0,\beta_1,...,\beta_p$) are estimated by the Maximum Likelihood Estimation method (MLE), e.g., maximizing the likelihood of the following probability: $P(Y_1=y_1,...,Y_n=y_n\mid \vec{x}_1,...,\vec{x}_n)$ which is maximizing $\prod_{i=1}^{n}\left\{ p(\vec{x})^{\,y_i}\left[1-p(\vec{x}_i)\right]^{\,1-y_i}\right\}$. [1,2,5]

3.3 Decision Tree and Random Forest

A decision tree is constructed based on rules over the features in a data set that are configured as a tree. A node represents an attribute of the data set. Attributes can take on continuous or categorical values. In a decision tree (DT) algorithm, training dataset is split several times based on a predefined criterion, resulting in a structure, which resembles tree branches [9]. Gain is one of the most common criteria in DT. By using gain criteria, the reduction in entropy is maximized because of that particular split. The approximation of $P(Y\mid\vec{x})$ is the proportion of Y class features over all features of the node that encompasses $\vec{x}$.

Random forests are a mixture of tree classifiers such that each tree is created based on the values of a random vector sampled independently from the train dataset. [7]

3.4 Neural Network

An Artificial Neural Network (ANN) is a model, which was motivated by the configuration of biological neural networks. In a generic artificial neural network, processing elements, called neurons, process external or internal information. Inputs received by a processing element can be represented as an input vector $X\ (x_1, x_2, .., x_p)$, where x_i is the value from the *ith* input. A weight is linked with each related couple of neurons. Therefore weights related to the *jth* neuron can be symbolized of the form $W\ (w_1, w_2, .., w_{np})$, where w_{ij} symbolizes the weight related to the joining neurons of *i* and *j*. The output of each neuron is according to the weights connected with the neuron's inputs $(\sum_{i=1}^{n} x_i w_i)$. The following equation shows the output (y) of a neuron is a function of product of the weights and values of n+1 inputs:

$$y = f\left(\sum_{i=0}^{n} x_i w_i\right)$$

An output will be produced based on each input. An error, E, is defined as the accuracy of the response (difference of the predicted o_p and actual t_p output). The weights vary in order to maximize the accuracy (minimize the global error). [3]

$$E = \frac{1}{2}\sum_{k}\left(t_{pk} - o_{pk}\right)^2$$

3.5 Cross Validation

Cross-validation is a validation technique for evaluating how the results of a model will generalize to unseen data sets. It is mostly used when we want to evaluate how well a model will do in practice. In a classification problem, the prediction model is created based on a training data set. The model is then tested based on a data set that is not used in creating the model and is called testing data set. The objective of cross validation is to define a data set (i.e., the validation data set) to evaluate the model in the training phase. [6]

In 10-fold cross validation, the training data is split into 10 partitions. For 10 times, 9/10 of the data is used to make training sets (the validation datasets) and to build 10 models. These models are applied to the remaining 1 partition to calculate a performance estimate. Then the 10 performances are averaged and the end result is an average that is a reasonable estimate of the performance of a model on unseen data. [6]

3.6 Model Evaluation

A confusion matrix is usually used to evaluate a classification method. A confusion matrix (in a tabulated form) shows how many points in the data set are classified correctly or incorrectly. The definition of each cell in the confusion matrix is presented below:

- True positives (TP): the number of positive cases that were predicted correctly
- False positives (FP): number of positive cases that were predicted incorrectly
- True negatives (TN): number of negative cases that were predicted correctly
- False negatives (FN): number of negative cases that were predicted incorrectly

The measures that usually is used to pick the best model is total accuracy:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

3.7 Ensemble methods

In ensemble methods, multiple models (such as Logistic Regression, Random Forest, Artificial Neural Network, etc.) are generated and combined together to solve a problem. Ensemble methods are mainly used to improve the performance of a model. Different combination rules are defined in ensemble methods to extract the result of models combination. One of the combination rules is voting. Voting uses a popular vote of the results of applied models to provide the final prediction result. [11]

3.8 Model selection

In our classification problem (predicting the cluster of students) which is presented in section 4.5, the measure to evaluate the performance of the model was total accuracy of classification. Different classification methods were attempted and the best total classification accuracy was achieved by using the ensemble (voting) of 3 classification methods. The methods were Logistic Regression, Random Forest and Artificial Neural Network.

4 METHODS AND MODELS

In this Chapter, the methods that are used for data preparation and data visualization along with the models, which were developed, for data clustering and classification are presented.

4.1 Data sources and data preparation

The data source was from Enterprise Data Delivery Information Environment (EDDIE), which stores historical student data at the University of Illinois at Chicago (UIC). In this study, data of ME undergraduate students from 2005 until 2014 was used. The data included 521 students. The two datasets that were utilized in this project were the students-course history dataset, and students-major history dataset.

Students-course history data, displays courses that have been taken by each student in each semester, as well as student grades. Students-major history dataset displays which semesters each student has been enrolled in MIE department. Based on the students-major history dataset, we can discover semester numbers for each student (e.g., spring 2011 is the 3rd semester of a specific student). By linking the major-history table to the course history table for all students and based on the semester number, we can observe which courses were taken at each semester number. An example is presented in Figure 2.

In these data sets “AH Term Code” is Academic History Code which shows the year and the semester.

Major history table

De-identified UIN	AH Term Code	Student Level Description	Student Type Description	Semester
9756694	220098	Undergrad - Chicago	New First Time Freshman	1
9756694	220101	Undergrad - Chicago	Continuing	2
9756694	220105	Undergrad - Chicago	Continuing	2.5
9756694	220111	Undergrad - Chicago	Continuing	3

Course history dataset

De-identified UIN	Overall Course Detail Term Code	Overall Course Subject Code	Overall Course Number	Overall Course Grade Code	Overall Course Title	Semester
9756694	220098	ENGL	161	B	Academic Writing II	1
9756694	220098	ENGR	100	S*	Engineering Orientation	1
9756694	220098	HON	134	AH	Hon Core Nat World/US Society	1
9756694	220098	HON	222	SH	Honors Activity	1
9756694	220098	MATH	220	C	Differential Equations I	1
9756694	220101	CME	201	A	Statics	2
9756694	220101	HON	122	AH	Hon Core Ind & Soc / World Cul	2
9756694	220101	MGMT	340	B	Intro to Organizations	2
9756694	220105	AH	100	A	Intro to Art &Art History	2.5
9756694	220105	CME	203	A	Strength Of Materials	2.5
9756694	220105	ME	250	A	Engineering Graphics	2.5

Figure 2 Finding semester numbers that courses are taken

Calculating semester numbers instead of considering year/semester (AH Term Code) helps compare the study path of all students in different years. Summer semesters were included, but were classified as half semesters instead of main semesters. Since the semester numbers that courses are taken by each student are known, student study paths can easily be defined.

The path for each student is the vector of required courses, and the semesters in which the courses were taken. This vector was later used to measure similarities amongst the study paths of students. Table 1 shows a sample of a study path vector.

Table 1. Sample study path profile for a given student

student1	ENGL160	CHEM112	MATH180	ENGR100	ENGL161	ME250	MATH181	PHYS141	CS109	IE201	MATH210	CME201	PHYS142	IE342	MATH220	MATH310	CME203	IE365	IE442	IE471	ECE210	IE446	IE345	IE472	MGMT340	IE467	IE461	IE473	IE380	IE463	IE396	IE466	IE499
	1	1	2	2	3	3	3	3	4	4	4	4	5	5	5	5	6	6	6	6	7	7	7	8	6	9	9	9	9	10	10	10	10

Students are admitted to the college of engineering in two categories: freshmen, and transfer. Freshmen students are the students that have started studying mechanical engineering at MIE department at UIC. There are two types of transfer students, who are called Internal transfers and External transfers. Internal transfer students are the students, which have started their studies in any other major at UIC then transferred to MIE. The external transfer students are students that transferred from other university or college to MIE at UIC. Internal transfer students usually transfer several courses from their previous department to MIE department and External transfer students transfer several courses from their previous universities or colleges to MIE department.

For transfer students (both internal and external) the first semester within the MIE department is different from the freshmen students because the transferred courses are placed at the first semester of entrance to MIE department. In order to make all course-semester numbers comparable to one another, for transfer students we calculated the equivalent of semester numbers of the transferred courses through following steps:

Step 1: List transferred courses in the 1st semester

Step 2: Calculate the total credit hours of the main (required) courses, which are transferred and passed by each student

Step 3: Consider a range of credit hours for each semester, and assign equivalent semesters to each student. Table 2 reflects the equivalent semester numbers according to transferred and passed credit hours in the first semester, which are based on existing proposed curriculum.

Table 2. Equivalent semester according to transferred credit hours

Credit hour range	Semester
0-13	1
13-31	2
31-47	3
47-59	4
59-72	5
72-86	6
86-95	7
95-102	8

Step 4: Find the equivalent semester number for the first semester of each transfer student based on the total transferred credit hours. For example, if a student has 45 total transferred course credit hour, it means his first semester is equivalent to semester 4.

Step 5: Calculate the rest of course semester numbers (for the next semesters) based on the equivalent semester number.

The final step of data preparation was to conduct the missing values imputation. In order to assign the missing values of some student's study path vector, the K nearest

neighbor (KNN) method was employed. Each missing value in a study path vector was replaced by the average value of its 3 study path neighbors (K=3).

4.2 Data visualization

One visualization tool was developed in this study, which is called Students-Tree. Students-Tree is a dynamic tool that makes a user capable of reviewing the behaviors of students in different levels (major, year, entrance type, and leaving type) with different colors. Major, the first level of the tree, can be selected as either Mechanical engineering or Industrial engineering. The second level, year, is the range of years between 2000 and 2015. The third level is entrance type of students, which can be New, first time freshmen; External transferred; Internal transferred and Readmits. The fourth level, leaving type, which can be 1- graduated from MIE (MIE G) 2- graduated UIC from a major other than Mechanical or Industrial Engineering (Other G), 3-dropped out (Drop), or still current student (na).

One snapshot of this tool is presented in Figure 3. In this example, we observe that 136 students started studying Mechanical engineering in 2009. Out of those 136 students, 60 students were freshmen. Out of those 60 students, 23 graduated. Out of those 23 students; 2 students graduated after 7 semesters, 7 students after 8 semesters, 7 students after 9 semesters, 2 students after 10 semesters, 2 students after 11 semesters, and 3 students graduated after 12 semesters.

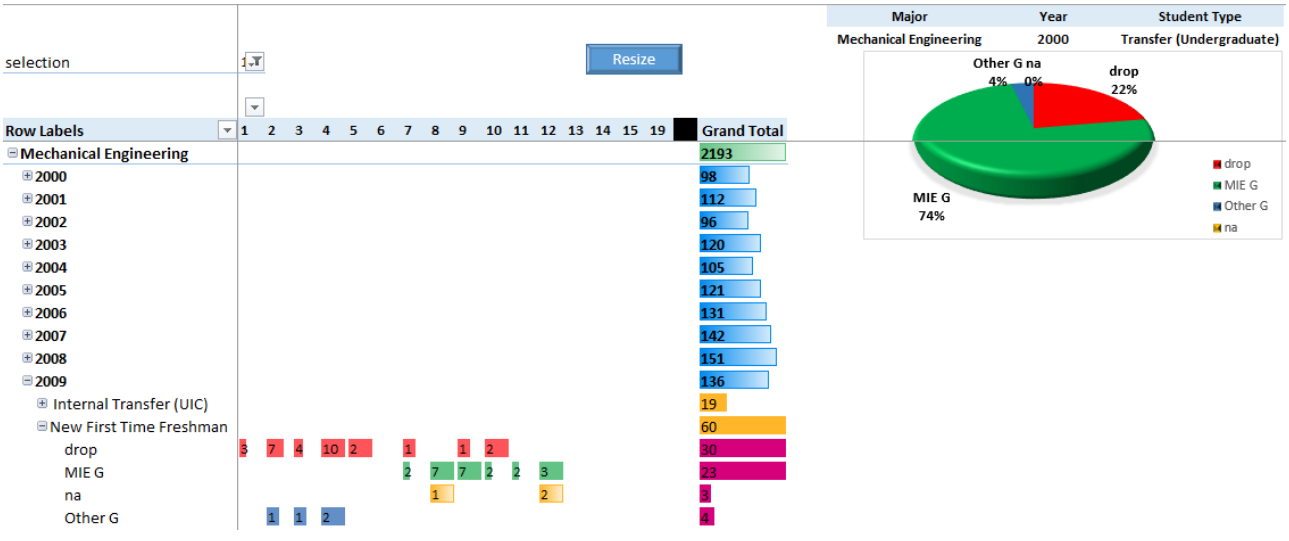


Figure 3. One snapshot of the visualization tools for reviewing students' behavior

4.3 Current official curriculum evaluation

By using the result of data preparation, it was determined if existing official study plan (curriculum) was always followed by undergrad students in the past 10 years. An example of the Mechanical Engineering proposed study path is presented in Figure 4. The colors in each circle represent the percentage of students who have followed the study path, in Green, taken the course sooner, in Blue, or taken the course later, in Red.

Figure 4. ME proposed study path respect comparison

The visualization results reflect that most of the students did not follow the official study plan in the past in MIE. Thus, the next question to answer was which study paths were the most popular? To address this question, we decided to use data

mining techniques to discover the popular paths. A clustering task was applied to find the main, popular study path of students.

4.4 Curriculum mining using Clustering

K-means clustering was used o cluster students based on their study path vector. Various K numbers were tried, but K equal to 3 was chosen based on the evaluation of clustering result (details are discussed in "Evaluation of clustering results "section"). The clustering result revealed three categories of students with different characteristics. The centroid points of each cluster for all the courses are presented in Figure 5. The centroid point represents the average of the semesters that students in each cluster have taken each course.

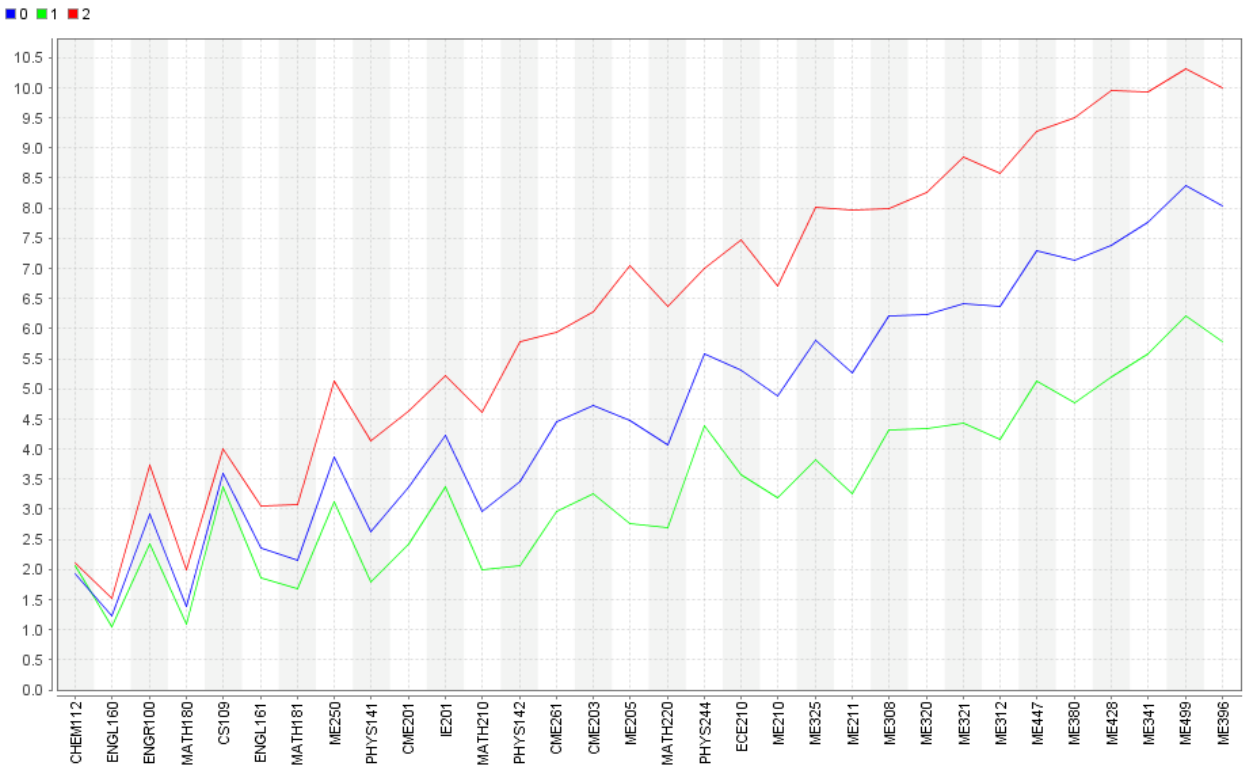


Figure 5. The centroid points of students' study path clusters

Based on the three clusters, three study paths were defined as Fast Track, Regular, and Extended-Time Curriculum. Fast Track-cluster (which is named 1 and

presented in green color in Figure5) is the study path of students that graduate in less than 6 semesters. In this path, students usually take more courses in each semester. Regular- cluster (which is named 0 and presented in blue color) is the study path of students that graduate around 8 semesters. Although this path is different from the proposed curriculum, it is still defined as a regular study path due to the duration similarity. Finally, Extended-Time Curriculum (which is named 2 and presented in red color) is the study path of students which are expected to have a longer studying duration, around 10 semesters. The different characteristics of these defined clusters are analyzed and presented in the "Analysis of Study Paths" in the next chapter.

The comparison of centroids and actual course-semester number distributions in each cluster can now be created for each course. As an example, Figure 6 shows the distribution of semesters that students have taken IE201 in Regular, Fast track, Extended-time curriculum clusters, and the centroids of this course. These graphs are created for all the required courses in Mechanical Engineering and they are presented in Appendix I.

Course:	IE201

	0	1	2
Clusters:	Regular	Fast track	Extended
Centroids:	**4.46**	**2.77**	**5.38**

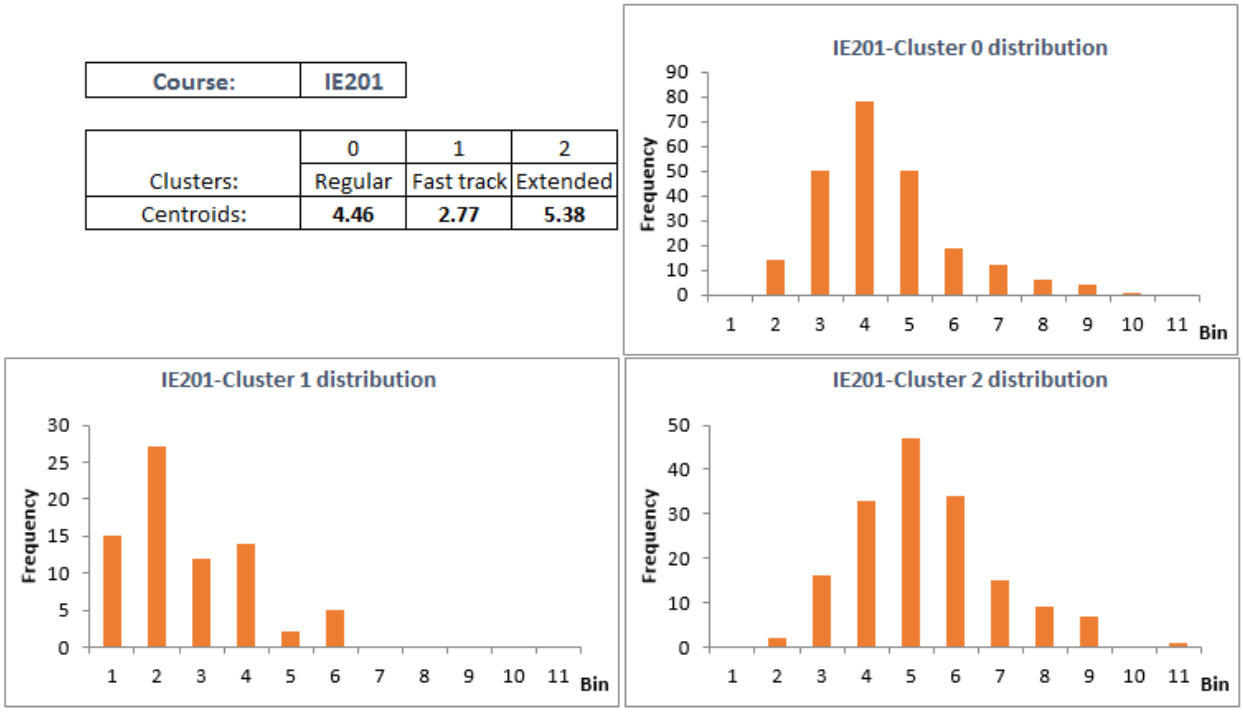


Figure 6. Comparison of centroids and raw data distributions in each cluster for IE201 course

4.5 Developing prediction models to detect the study path

Prediction models were created to forecast what the study path of a specific student would be in future upcoming semesters. This prediction is done in two steps. In the first step, we can predict and detect the students that fall in the Fast Track. This prediction (which is done for the students at the end of their second semester) is based on the course history of the first two semesters. In the second step, at the end of semester 4, we can predict all student clusters (Fast Track, Regular Curriculum, and Extended-Time Curriculum). The prediction overview is demonstrated in Figure 7. These results can be used in advising and informing - warning the students about the study path that they might be on. Students can review the study paths and the features of each study path and decide about the path that they want. They can use this result when they are selecting and taking courses in the next semesters.

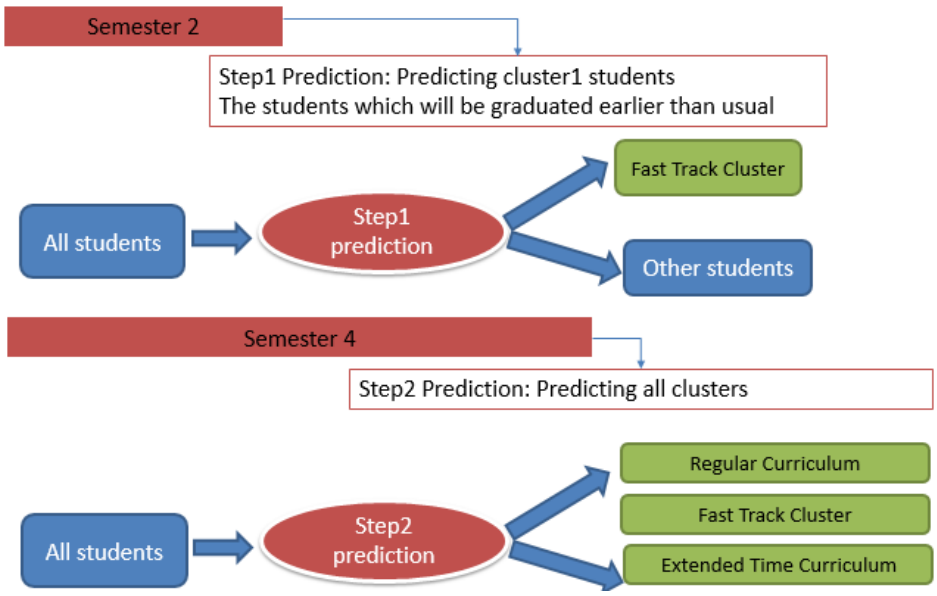


Figure 7. The overview of prediction steps

The input variables for prediction in Step 1 and Step 2 are the semester number of the courses, which are taken within first 2 semesters for Step 1, and first 4 semesters for Step 2. The target variable is the study path cluster. In Figure 8 and 9, the selected courses are attributes (input variables) of Step 1 and Step 2 prediction respectively.

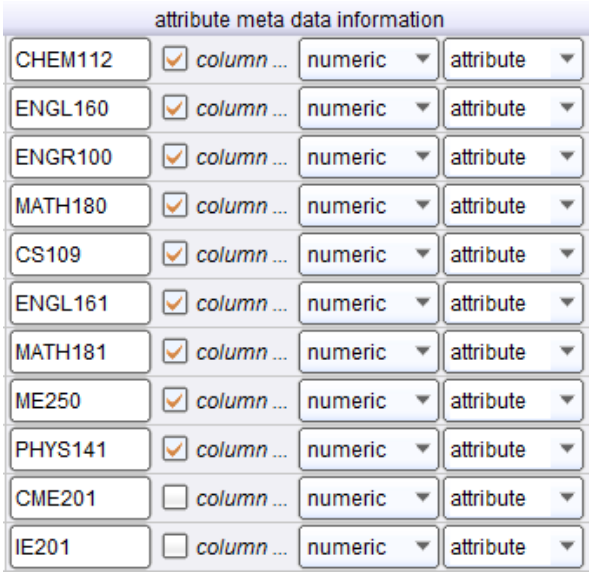


Figure 8. Input variables of prediction model in Step 1

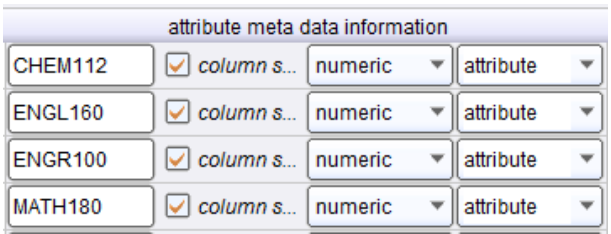


Figure 9. Input variables of prediction model in Step 2

The structure of prediction models is presented in Figure 10. In both prediction models (in step 1 and step 2), 10-fold cross validation was used for developing the prediction models and learning the coefficients. We used voting to find the result of three different classification techniques (Random forest, Logistic regression and Neural Network) to predict the class (cluster number) of each student.

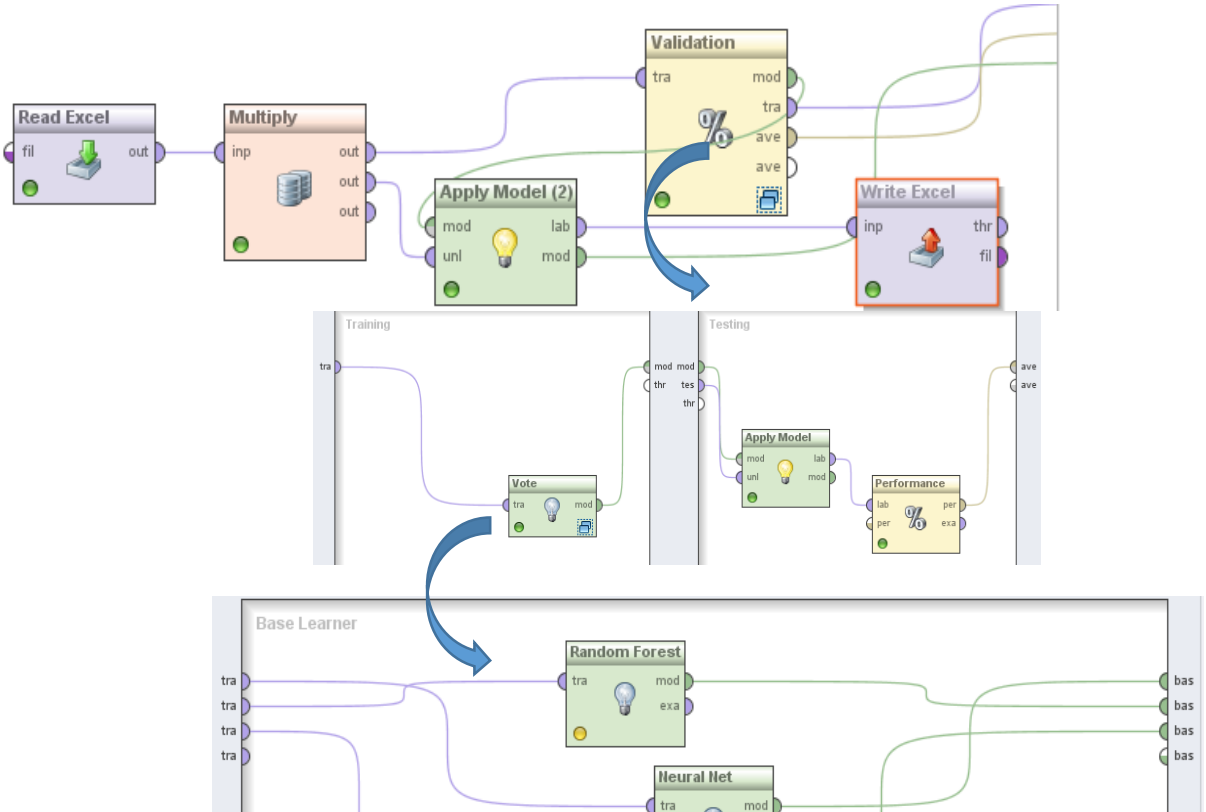


Figure 10. The Structure of a prediction model

The parameters in Random Forest model were set as following:

-Number of trees: 10

- Criterion: gain ratio

-Maximum depth: 20

-Confidence: 0.25

-Minimal gain: 0.1

-Minimal leaf size: 2

- Minimal size for split: 4

Along with applying pruning and pre-pruning.

The parameters in Logistic regression model were set as following:

-Kernel type: dot

-Kernel cache: 200

-Convergence epsilon: 0.001

-Max iterations: 100000.

The parameters in Neural Network model were set as following:

- Hidden layer: 1

- Number of nodes: 10

- Training cycles: 500

- Learning rate: 0.3

- Momentum: 0.2

- Error epsilon: 0.000001

5 RESULTS AND DISCUSSION

5.1 Evaluation of clustering result

In order to evaluate the clustering result, we reviewed the within_centroid_distance, which is calculated by averaging the distance between the centroid and all instances of a cluster. We also reviewed the Davies–Bouldin index for different number of clusters. We then selected the number of clusters based on low intra-cluster distances, high intra-cluster similarity, and high inter-cluster distances, low inter-cluster similarity and low Davies–Bouldin index. K means and different number of clusters produced a collection of clusters. Clusters with the smallest Davies–Bouldin index were considered the best. K=3 was the optimal number of clusters. In Figure 11, the comparison of clustering result based on different number of clusters is presented.

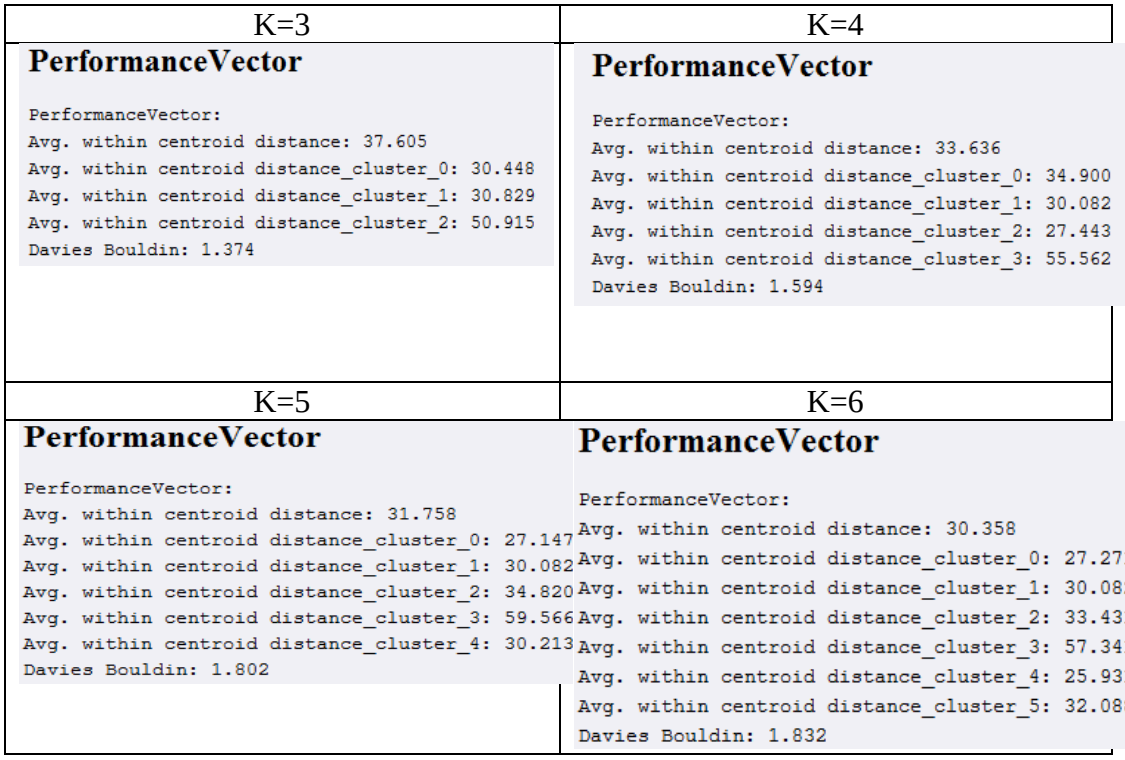


Figure 11. Clustering performance with different number of clusters (K)

An external evaluation for assessing the result of three main clusters was then used. In this method, clustering results are analyzed based on some part of the data, which were not used for clustering called benchmarks. These benchmarks involve a set of pre-classified objects created by experts, and can be used as a standard in order to evaluate the performance of clustering. The performance of clustering is high if the assigned class of benchmarks in the clustering process is close to the predetermined benchmark classes.

Benchmarks in this study were selected based on the number of the last semester for 46 students (approximately 10% of data). 10 students with the last semester of 12 or 13 were selected and assigned to the cluster 2 (Extended-Time Curriculum). 20 students were selected with the last semesters of 8 or 9, and were assigned cluster 0 (Regular Curriculum). 16 students with the last semester of 5 or 6 that were assigned cluster 1 (Fast Track Curriculum). The clustering procedure for all data sets were then run, and we discovered the assigned classes of all students based on their study paths. Finally, we compared clustering result with predefined classes of benchmarks. The result reflected a 97.8% accuracy of clustering performance on assigning classes to benchmarks.

5.2 Analysis of study paths

Out of 521 students in our dataset; there are 251 students in Regular study path, 170 students in Fast track study path, and 100 students in Extended-time study path. We measured the GPA of students in each study path cluster, and surprisingly noticed that the students in shorter study paths have better GPA (although they take more courses in each semester). The average GPA of the students in Fast Track, Regular and Extended-Time study path, respectively, are 3.16, 2.98 and 2.64.

Each study path is calculated based on rounded centroid semesters of related cluster. Comparisons of calculated course semesters with the current proposed study path (curriculum) is presented in Table 3.

Table 3. Study path course-semester comparison

Curriculum Type	CHEM112	ENGL160	ENGR100	MATH180	CS109	ENGL161	MATH181	ME250	PHYS141	CME201	IE201	MATH210	PHYS142	CME261	CME203	ME205	MATH220	PHYS244	ECE210	ME210	ME325	ME211	ME308	ME320	ME321	ME312	ME447	ME380	ME428	ME341	ME499	ME396
Current	1	1	1	1	2	2	2	2	2	3	3	3	3	3	4	4	4	4	5	5	5	5	6	6	6	6	7	7	7	8	8	8
Regular	2	1	3	1	3	2	2	3	2	4	4	3	3	4	5	4	4	6	5	5	6	5	6	6	6	6	7	7	7	8	8	8
Fast Track	2	1	2	1	2	1	1	3	1	2	3	1	1	3	3	2	2	4	3	3	3	3	4	4	4	4	5	4	5	5	6	5
Extended Time	3	1	4	2	5	2	3	5	3	5	5	4	4	6	6	6	5	8	7	6	7	7	8	8	8	8	9	9	9	9	10	10

Freshmen, Internal, and External transfer students have different distribution of course-semester in each defined study path. The histograms of course semesters in Regular, Fast Track, and Extended-Time study paths are represented in Figure 12. Internal transfer students take ENGR 100 course and the IE 201 course later than other students do in Regular study path. External transfer students take CHEM 112 course and the ENGR 100 course later than other students do in Extended-Time study path.

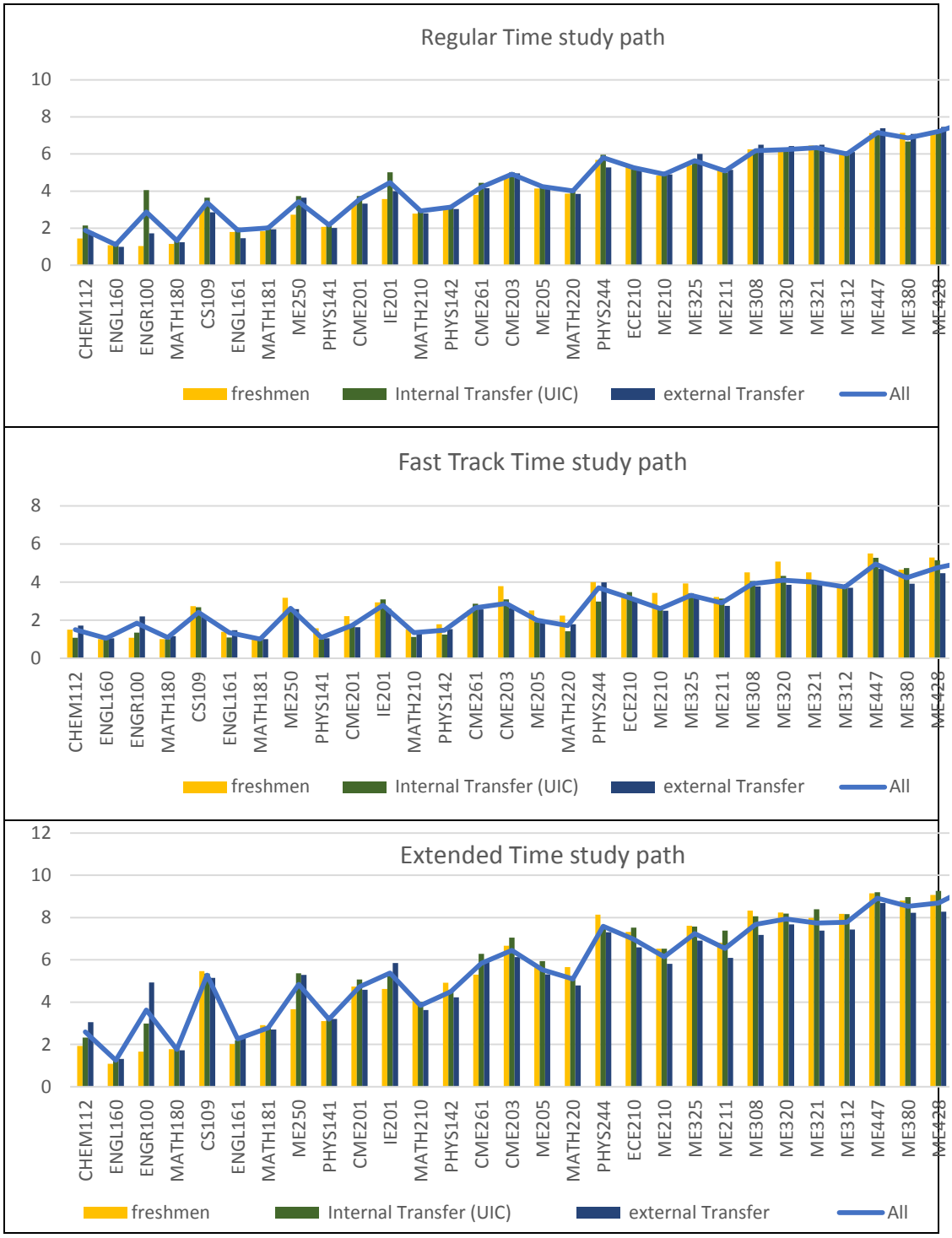


Figure 12. Clusters Detail -student entrance types histograms

5.3 Evaluation of study path prediction result

The first step prediction is used to detect the students that fall in the cluster 1 (Fast Track Cluster) and the accuracy is 90.59% (details are presented in Table 3). This prediction is done for the students in the end of semester 2.

Table 4. The confusion matrix of step 1 prediction

		Actual			
Prediction		0	1	2	Total
Student in other clusters (Regular or Extended)	0	242	24	100	366
Student in Fast Track Cluster	1	9	146	0	155
	Total	251	170	100	521

Actual cluster 0: Regular study path - Actual cluster 1: Fast track study path – Actual cluster 2: Extended study path.

Table 5 displays the second step prediction that can predict all the clusters of students, Fast Track, Regular Curriculum, Extended-Time Curriculum, with an accuracy of 83.69 % for the students in the end of semester 4.

Table 5. The confusion matrix of step 2 prediction

		Actual			
Prediction		0	1	2	Total
Regular study path	0	238	9	9	256
Fast Track study path	1	13	161	0	174
Extended study path	2	0	0	91	91
	Total	251	170	100	521

The combination result is presented in table 5.

Table 6. The combination result of prediction in two steps

			Actual			
Prediction	Step1	Step2	0	1	2	Total
Student in other clusters (Regular or Extended)	**0**		**242**	**24**	**100**	**366**
Regular		0	234	5	9	248
Fast Track		1	8	19	0	27
Extended		2	0	0	91	91
Student in Fast Track Cluster	**1**		**9**	**146**	**0**	**155**
Regular		0	4	4	0	8
Fast Track		1	5	142	0	148
		Total	**251**	**170**	**100**	**521**

5.4 Courses overlap calculation

In order to help the department in scheduling the courses we defined a measure, which calculates how often the courses are taken together. We called this measure the overlap ratio between each two courses. The overlap ratio of two courses shows the tendency of students to take both of the courses together in one semester. Therefore, in course scheduling point of view, when two courses have high overlap ratio it is better to avoid scheduling them in one time (same days and hours). To calculate the overlap of two courses, first, we normalize the distribution of the semester numbers that students have taken the course (students' enrollment over semesters). Then the overlap ratio would be calculated using the following formula:

$$Overlap\ ratio\ of\ course\ X\ and\ course\ Y = 1 - \frac{\sum_{t=1}^{n}(x_t - y_t)}{2}$$

Where *x* and *y* are the normalized enrolment of course X and Y, *t* is the semester number and *n* is maximum number of semester that the courses have been taken.

Since we have the distribution of students' enrollment distribution over semesters we can easily calculate overlap ratio measure for each pair of courses. One example is given in Figure 10 for calculating the overlap ratio for ME211 course and ME205 course. In the left table the first column is semesters and second column shows number of students that have taken ME210 in each semester (e.g., 12 students have taken ME210 at their 3rd semester, 74 students at their 4th semester, etc.). The third column shows the same thing for ME211 (distribution of students' enrollment over semesters).

In the right graph, the blue line shows the normalized distribution of students' enrollment over semesters for ME210 and the orange line shows the same thing for ME205. The difference between the two curves is calculated and then summed up. Half

of the summation result is the area between two curves. 1 minus the area between curves is defines as overlap ratio. The less area between two courses (the closer overlap ratio is to 1), means that most students take these courses in the same semesters and the courses are usually taken together.

Calculation of Overlap for ME2015 and ME210

Semester #	ME210	ME205	Normalized ME210	Normalized ME205	Distance
1	0	0	0.00	0.00	0.00
2	0	2	0.00	0.01	0.01
3	12	48	0.05	0.21	0.16
4	74	93	0.33	0.41	0.08
5	100	48	0.44	0.21	0.23
6	36	23	0.16	0.10	0.06
7	5	5	0.02	0.02	0.00
8	0	4	0.00	0.02	0.02
9	0	3	0.00	0.01	0.01
10	0	1	0.00	0.00	0.00
11	0	0	0.00	0.00	0.00
Total	227	227	1	1	0.57
				Overlap Ratio	**0.71**

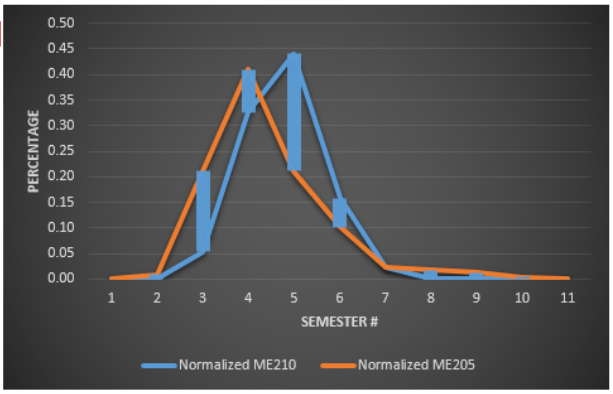


Figure 13. Overlap ratio calculation example

Overlap ratio of all required courses in ME department calculated and presented in Table 7. In order to visualize the overlap ratio percentage between courses, a color range is used

Table 7. ME courses overlap percentage comparison

	ME205	ME210	ME211	ME308	ME312	ME320	ME321	ME325	ME341	ME380	ME396	ME428	ME447
ME210	81%												
ME211	77%	85%											
ME308	55%	66%	78%										
ME312	58%	67%	80%	89%									
ME320	55%	63%	76%	94%	90%								
ME321	54%	65%	75%	89%	92%	89%							
ME325	64%	75%	87%	90%	86%	88%	84%						
ME341	36%	39%	54%	65%	70%	67%	71%	57%					
ME380	50%	54%	68%	77%	85%	79%	85%	72%	85%				
ME396	34%	36%	50%	61%	66%	63%	67%	54%	92%	81%			
ME428	45%	49%	62%	72%	79%	74%	80%	68%	87%	86%	82%		
ME447	42%	46%	59%	75%	78%	77%	81%	67%	89%	84%	84%	82%	
ME499	30%	33%	47%	54%	61%	56%	63%	50%	87%	76%	90%	89%	76%

6 CONCLUSION AND FUTURE WORK

In this dissertation, we determined if the current ME proposed study path is valid and always respected in the past, by using data evaluation and visualization techniques. The result shows that the current proposed curriculum (study path) is not followed by most of the students. We applied clustering techniques to categorize students according to three main popular study path clusters. We explored the characteristics of each study path cluster (e.g. graduation time and GPA). We compared the current proposed study path to these study path clusters. We used classification techniques to predict the study path cluster of students in two steps (at the end of second and forth semester). The accuracy of prediction in the first step and step 2 are 90.59% and 83.69%. Finally, we developed a method to measure the overlap of courses based on students' enrollment and we calculate the overlap ratio of all required courses in ME.

The results of this dissertation helps both students (through advising) and departments (through scheduling). Students benefit from reviewing the three popular study paths and characteristics and attributes of these study paths. They can select a study path that they think is more appropriate for them.

Selecting the right study path can result in a shorter graduation time, higher GPA, and a balanced challenging workload for each semester.

MIE department benefits include: minimum schedule conflict, reduced under-utilized capacity of resources (classes, instructors, TAs, etc.), as well as a balanced workload for instructors.

Future direction of this study can be the following items:

- Developing this project through other departments and majors at UIC,

- Including the general education and free elective courses in developing the proper study paths,
- Predicting the courses, which students of each cluster are going to get in the coming semesters (course enrollment prediction).

CITED LITERATURE

[1] Breiman L et al. Classification and regression trees. Belmont, CA: Wadsworth; 1984.

[2] Hosmer Jr, David W., and Stanley Lemeshow. Applied logistic regression. John Wiley & Sons, 2004.

[3] Hagan, Martin T., Howard B. Demuth, and Mark H. Beale. Neural network design. Boston: Pws Pub., 1996.

[4] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." ACM computing surveys (CSUR) 31, no. 3 (1999): 264-323.

[5] Kleinbaum, David G., and Mitchel Klein. Analysis of Matched Data Using Logistic Regression. Springer New York, 2010.

[6] Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." Ijcai. Vol. 14. No. 2. 1995.

[7] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2, no. 3 (2002): 18-22.

[8] Peña-Ayala, Alejandro. "Educational data mining: A survey and a data mining-based analysis of recent works." Expert systems with applications 41, no. 4 (2014): 1432-1462.

[9] Quinlan R. C4.5: programs for machine learning. Los Altos, CA: Morgan Kaufmann; 1993.

[10] Romero, Cristóbal, and Sebastián Ventura. "Educational data mining: a review of the state of the art." Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 40, no. 6 (2010): 601-618.

[11] Robi Polikar Ensemble learning. Scholarpedia, (2009) 4(1):2776.

[12] Trčka, Nikola, and Mykola Pechenizkiy. "From local patterns to global models: Towards domain driven educational process mining." In Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on, pp. 1114-1119. IEEE, 2009.

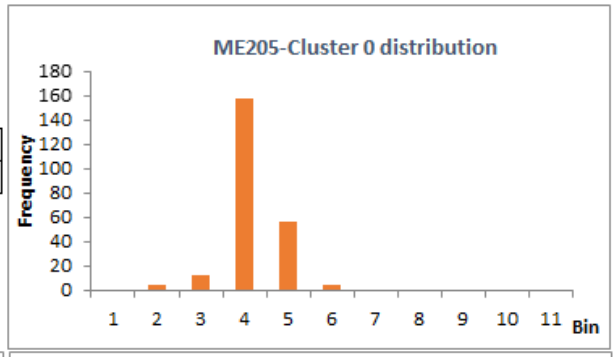
[13] Wang, Ren, and Osmar R. Zaïane. "Discovering Process in Curriculum Data to Provide Recommendation."

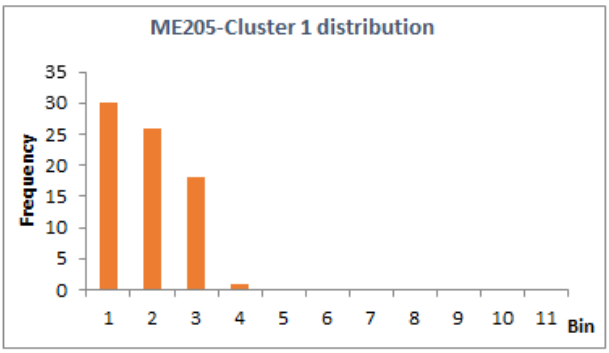
Appendix I

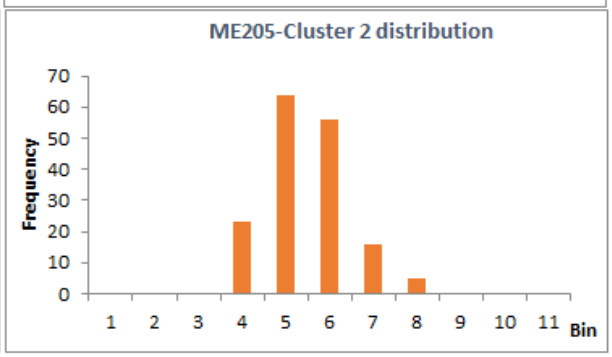
Following graphs show the comparison of centroids and actual course-semester number distributions of students in each cluster. The distributions can be used in calculating the probability of taking a course (by students in a cluster) in different semesters.

Course:	ME205

Clusters:	0	1	2
Centroids:	**4.24**	**1.99**	**5.53**

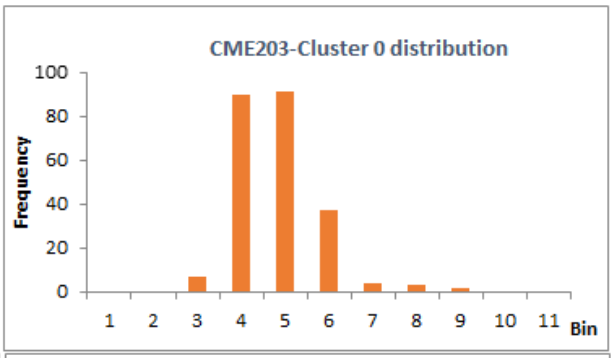


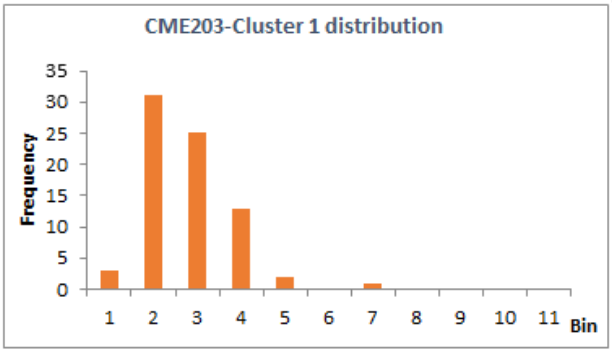


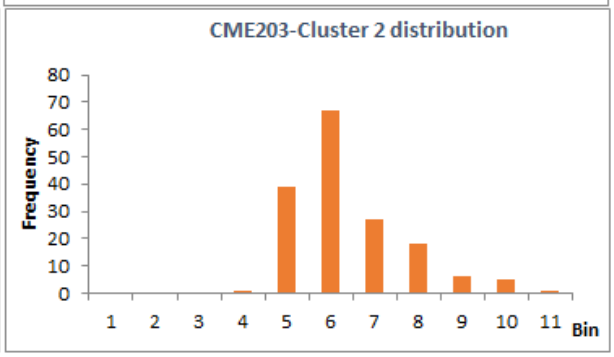


Course:	CME203

Clusters:	0	1	2
Centroids:	**4.91**	**2.87**	**6.46**

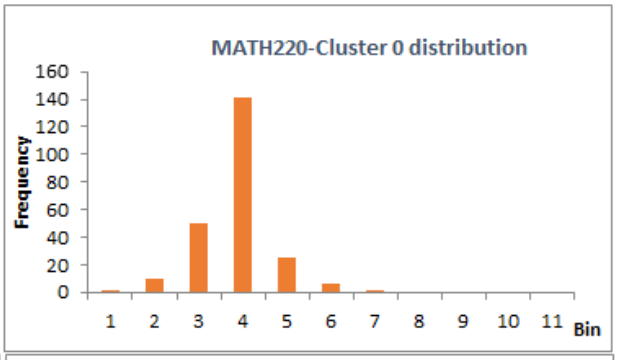


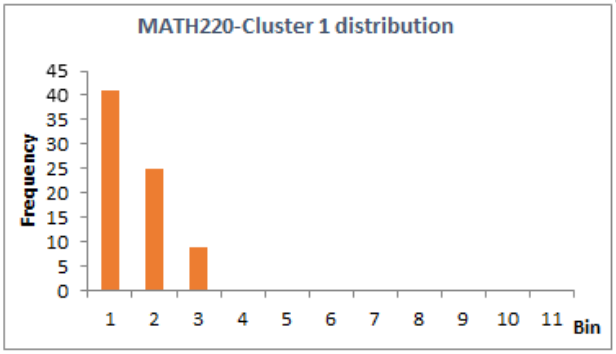


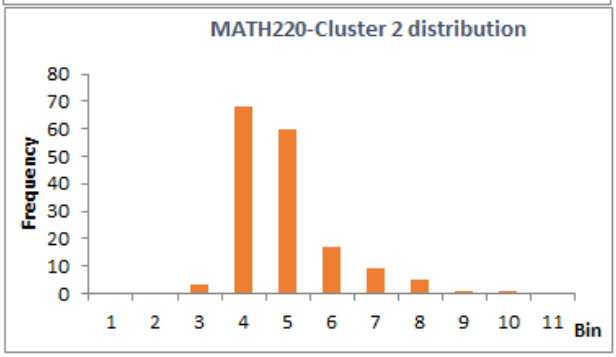


Course:	MATH220

Clusters:	0	1	2
Centroids:	**4.02**	**1.72**	**5.09**

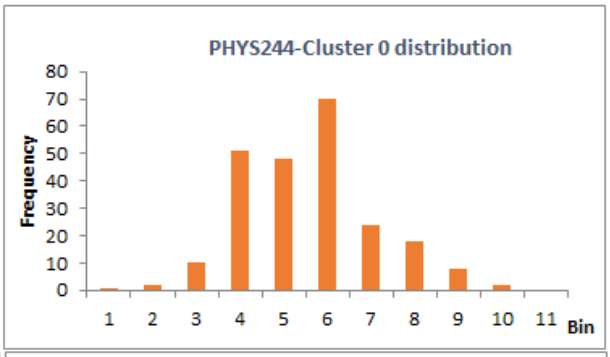


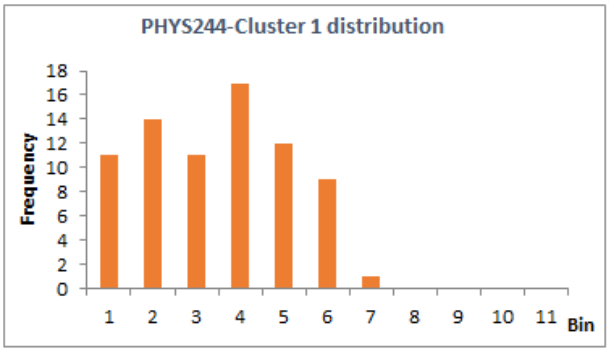


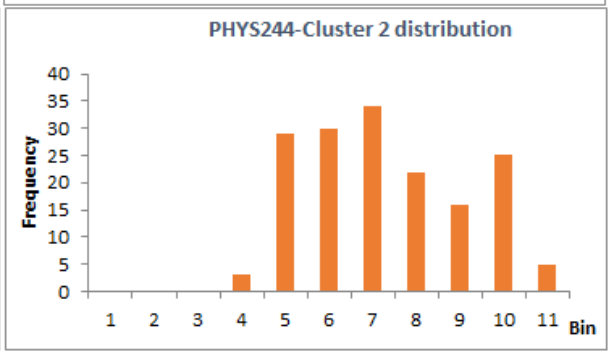


Course:	PHYS244

Clusters:	0	1	2
Centroids:	**5.80**	**3.69**	**7.59**

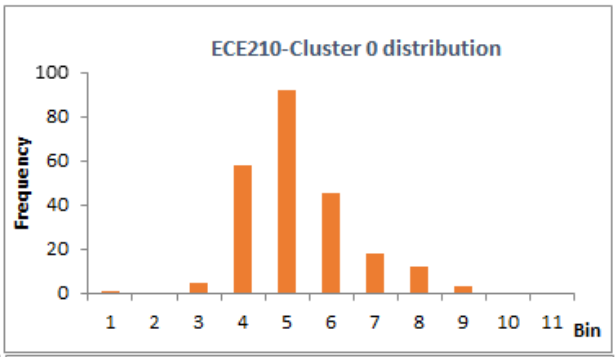


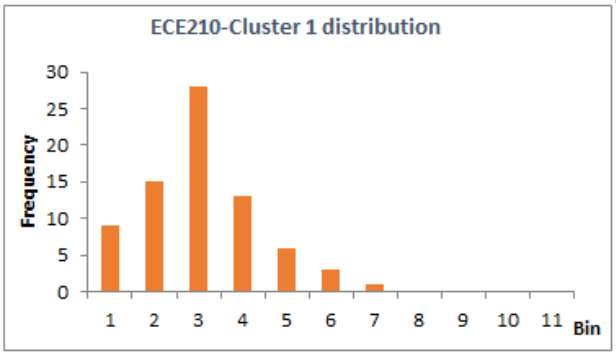


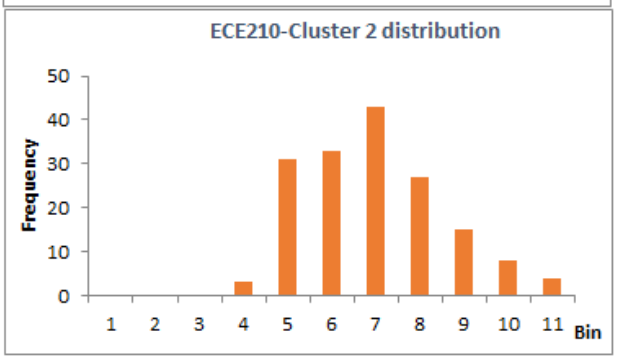


Course:	ECE210

Clusters:	0	1	2
Centroids:	**5.27**	**3.17**	**6.98**

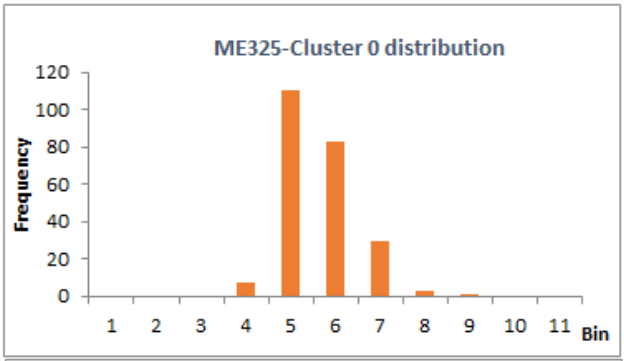


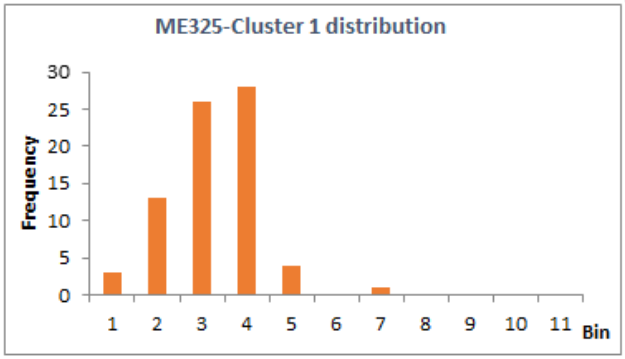


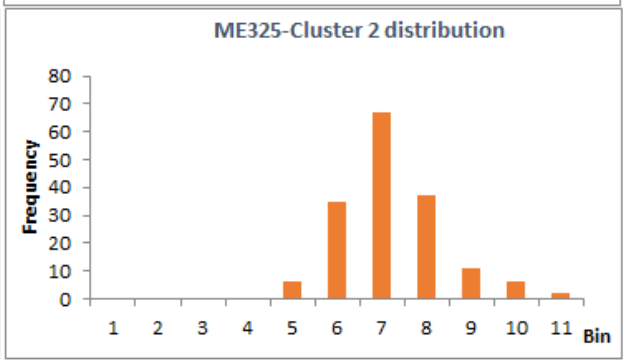


Course:	ME325

Clusters:	0	1	2
Centroids:	**5.64**	**3.30**	**7.24**

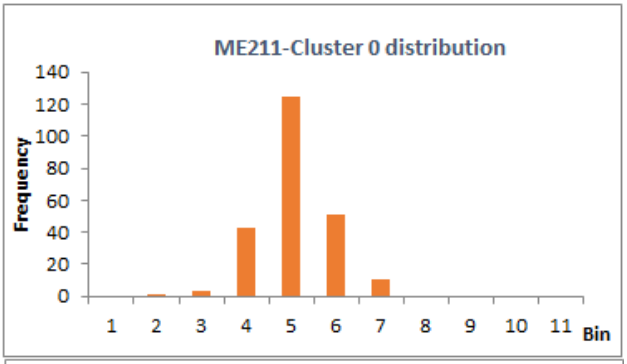


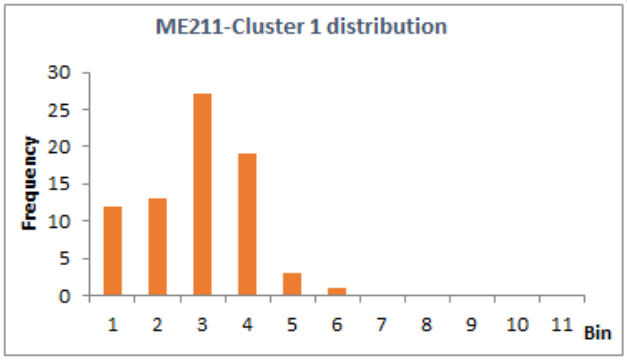


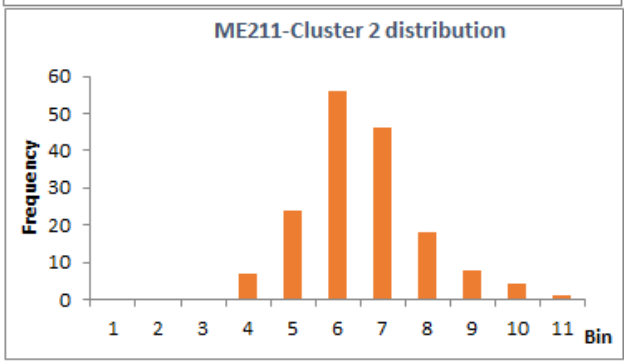


Course:	ME211

Clusters:	0	1	2
Centroids:	**5.09**	**2.90**	**6.54**

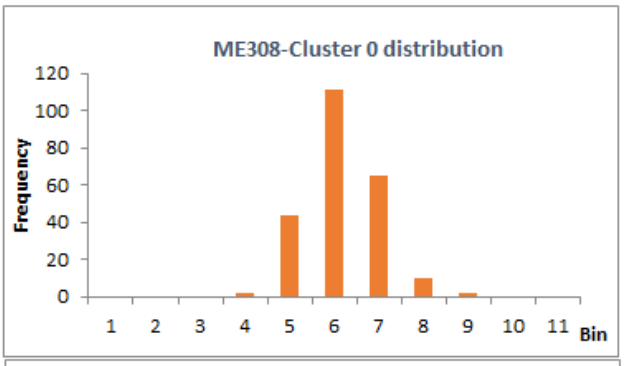


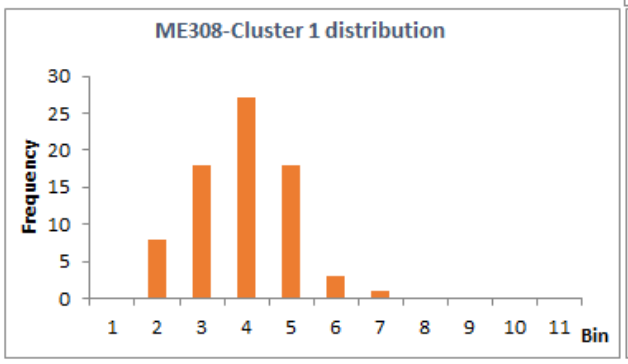


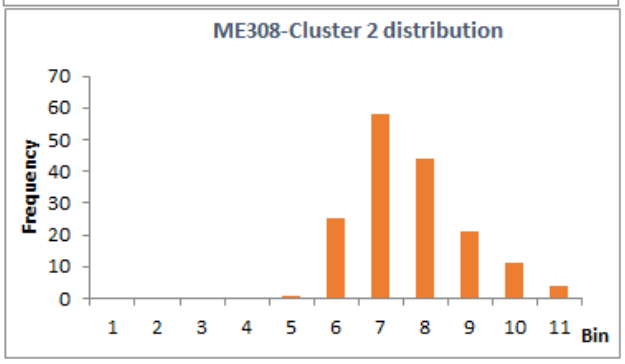


Course:	ME308

Clusters:	0	1	2
Centroids:	**6.18**	**3.91**	**7.67**

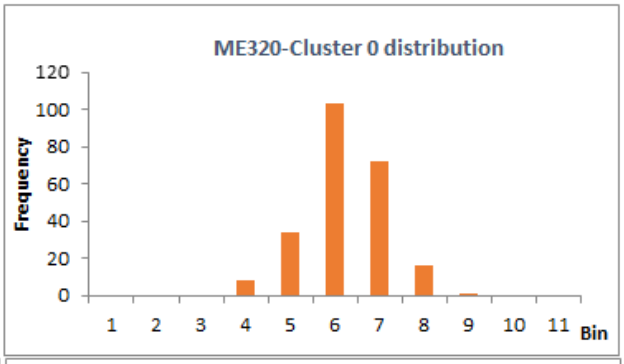


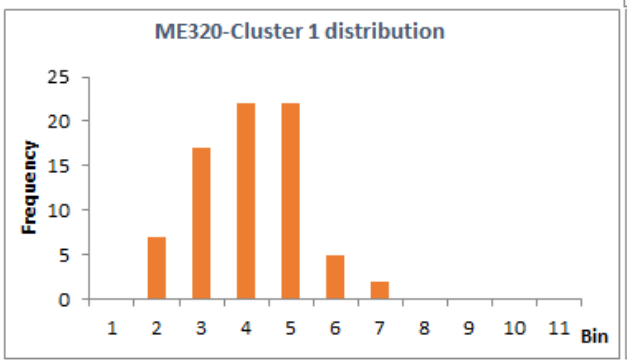


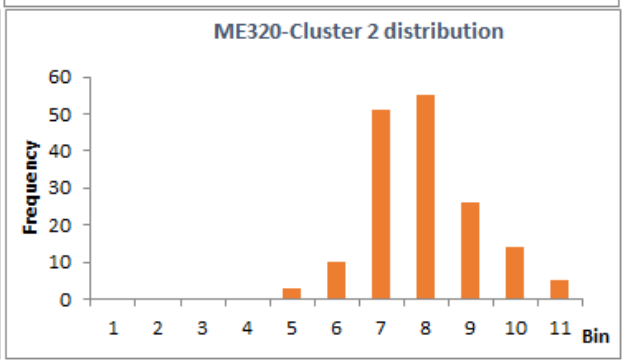


Course:	ME320

Clusters:	0	1	2
Centroids:	**6.25**	**4.10**	**7.93**

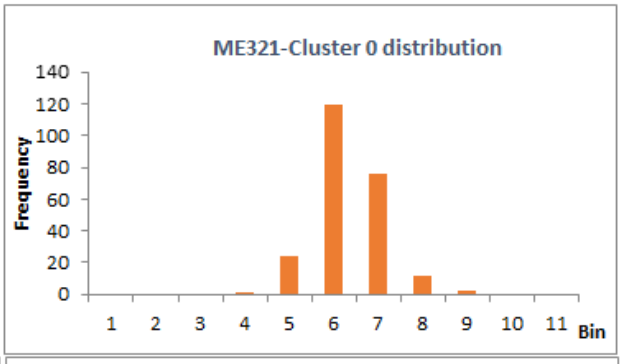


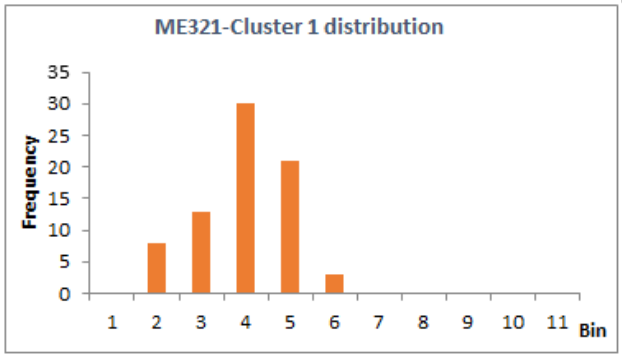


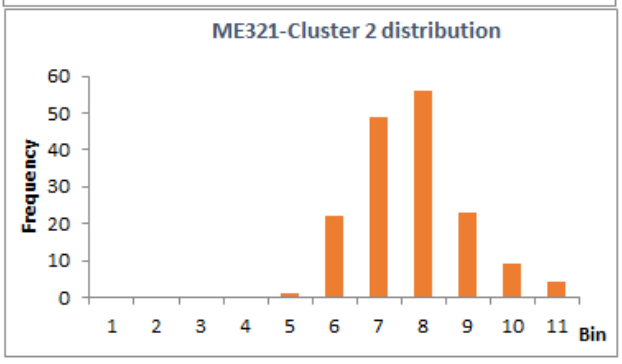


Course:	ME321

Clusters:	0	1	2
Centroids:	**6.34**	**3.99**	**7.74**

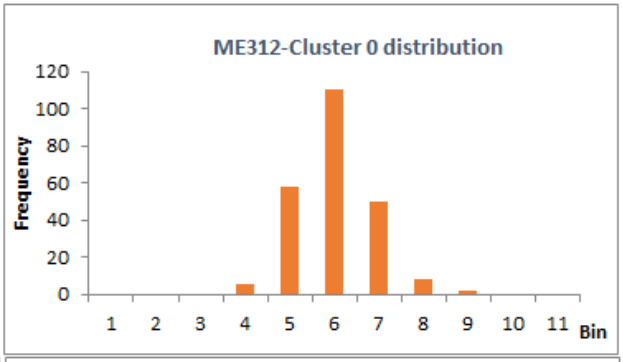


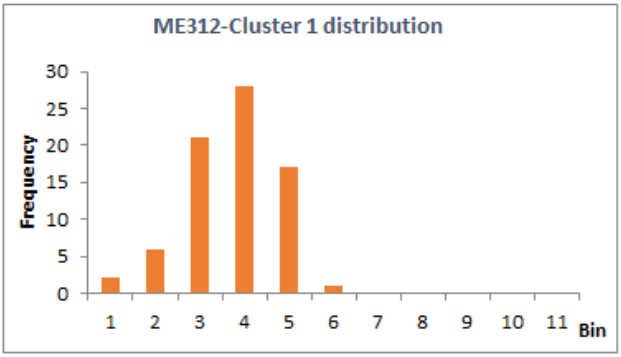


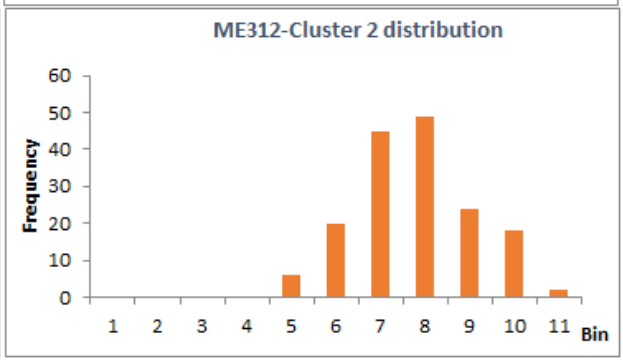


Course:	ME312

Clusters:	0	1	2
Centroids:	**6.01**	**3.74**	**7.78**

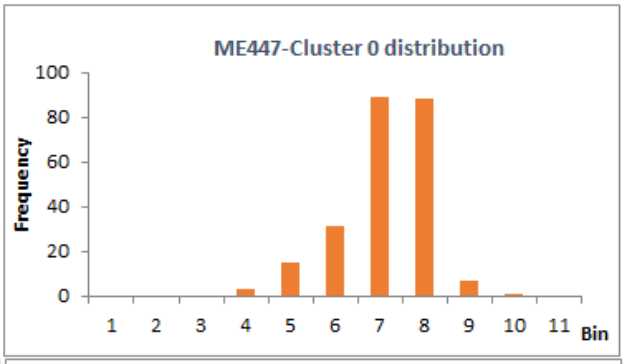


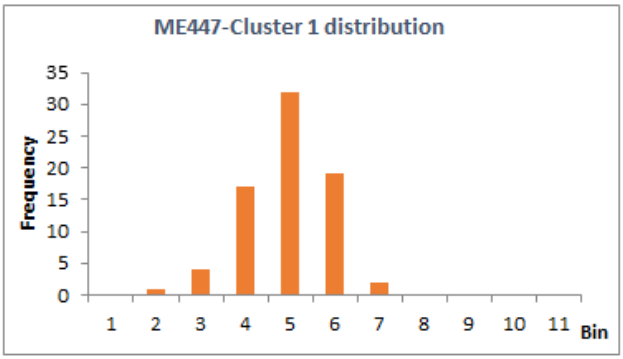


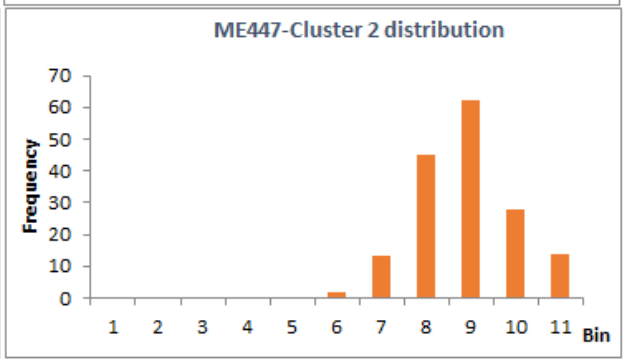


Course:	ME447

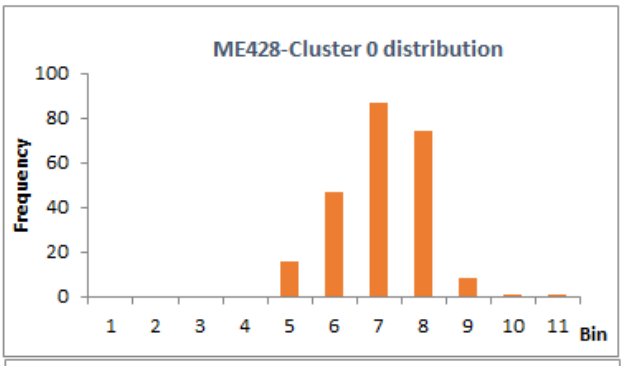
Clusters:	0	1	2
Centroids:	**7.15**	**4.94**	**8.91**

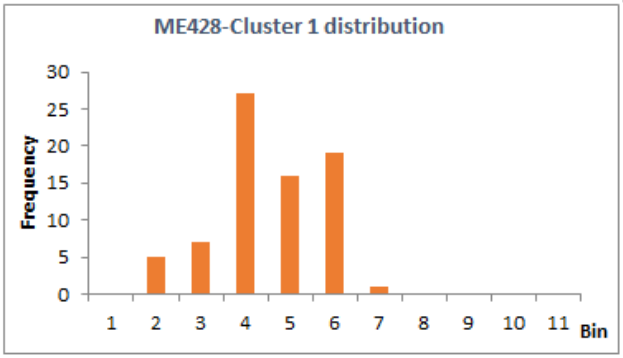


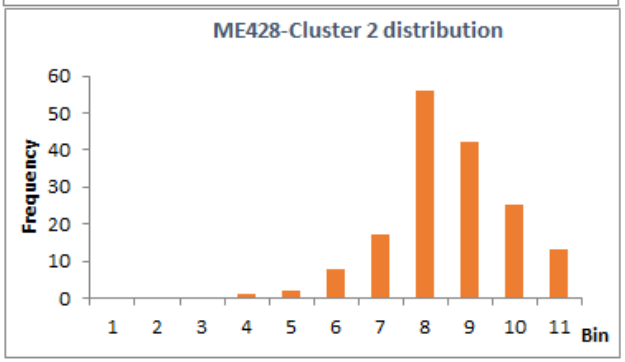




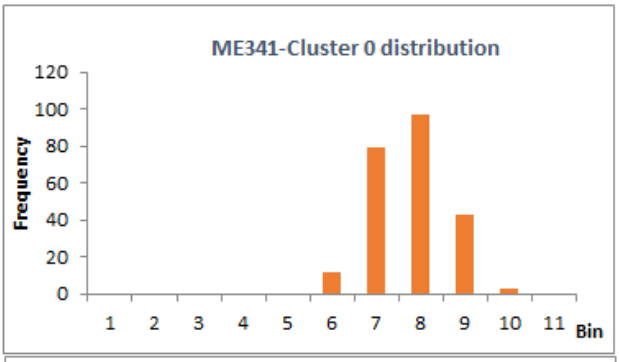
Course:	ME428		
Clusters:	0	1	2
Centroids:	**7.21**	**4.74**	**8.69**

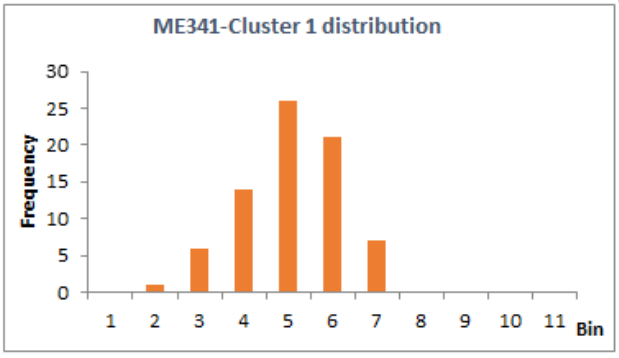


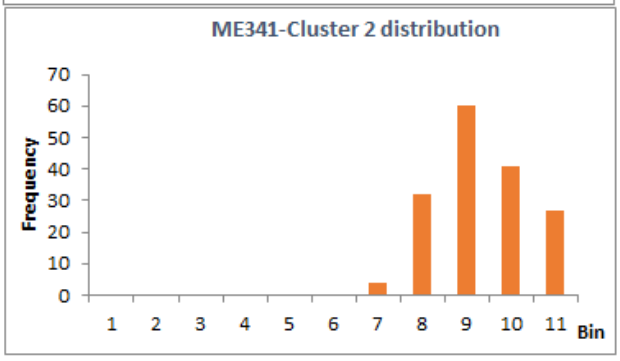




Course:	ME341		
Clusters:	0	1	2
Centroids:	**7.77**	**5.09**	**9.40**

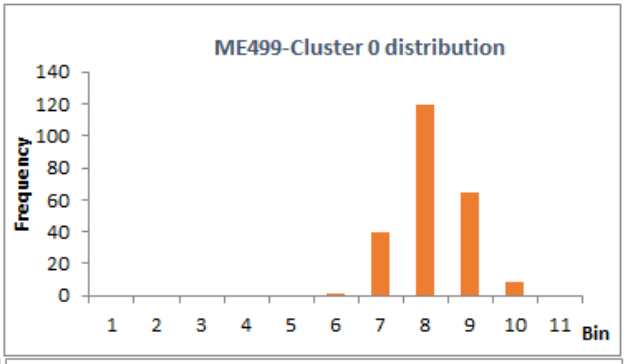


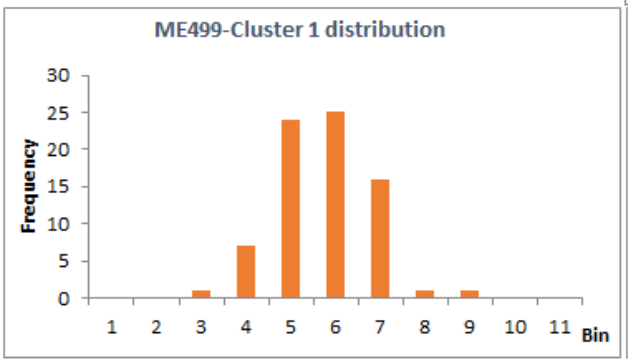


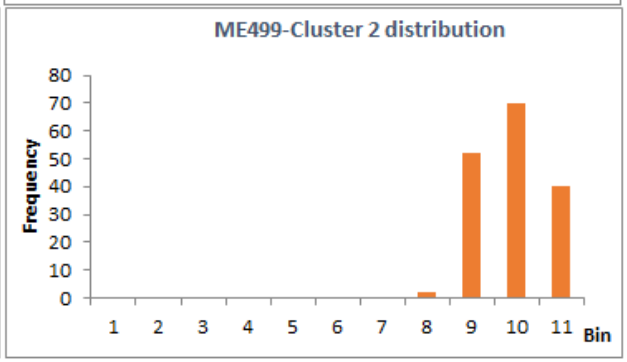


Course:	ME499

Clusters:	0	1	2
Centroids:	**8.22**	**5.82**	**10.03**

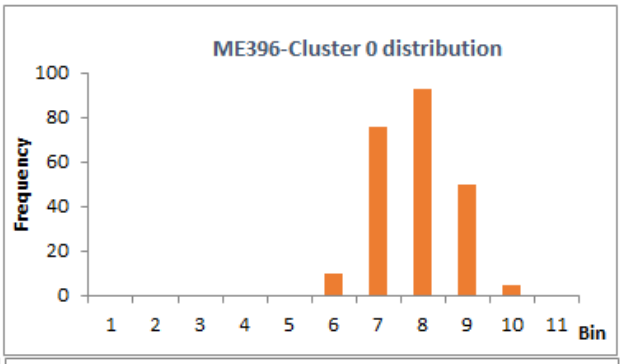


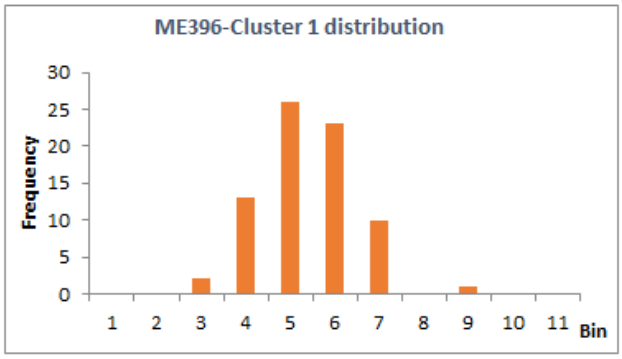


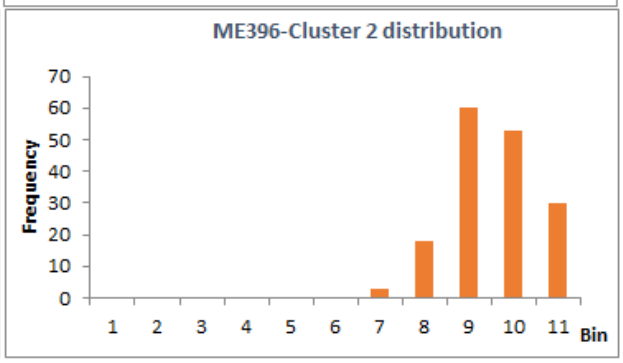


Course:	ME396

Clusters:	0	1	2
Centroids:	**7.90**	**5.44**	**9.72**







VITA

Maryam Teimoori

SEL 4209, 842 W Taylor, Chicago, IL 60607. Phone: (312) 545 7027
Email: mteimo3@uic.edu

Education

University of Illinois at Chicago(UIC) (GPA: 3.83/4) — **Chicago, IL** Aug. 2014-May.2016

MSc. Industrial Engineering

Thesis: "Academic Curriculum (Study Path) Mining of Mechanical Engineering Undergraduate Students"

Amirkabir University of Technology — **Tehran, Iran** Sept. 2002- June. 2006

BSc. Industrial Engineering

Technical Skills

Machine Learning, Process Mining, HV substation designing

Computer Skills

Mathematical Analysis:	R
Statistical and Data Mining Packages:	Rapid miner, XLminer , SPSS
Data Visualization:	R ggplot2, Tableau

Professional Experience

Research/Teaching Assistant at University of Illinois at Chicago — **Chicago, IL** 2014-2016

High Voltage Substation Design Engineer at Fulmen Co.
Fulmen is one of the main pillars of Iranian Electrical Industries, in domain of Electrical Engineering Services, High Voltage Turn-Key High Voltage Electrical Sub-Stations projects, and many other electrical sub-stations and electrical project. — **Tehran, Iran** 2004-2013

Electrical Utility designer at Rahshahr Co.
Rahshahr Co. is an Architect,Urban Design,Hydraulic&Energy Consaltants Group — **Tehran, Iran** 2003-2004

Tender expert at Mahtab Bargh CO.
Mahtab bargh is a Design&Engineering Co. — **Tehran, Iran** 2002-2003

Presentation

Maryam Teimoori, Ashkan Sharabiani, Anooshiravan Sharabiani, Fazle Karim, Houshang Darabi," Comparing trace-based and time series prediction modelling for estimating the enrollment in engineering courses" , ISERC Conference ,Nashville, TN, May,2015.