Data Mining Application Design Using K-MEANS and Exponential Smoothing Algorithm for Predicting New Student Registration

Fajar Masya
Information System, Computer Science Faculty, Mercu Buana University, Indonesia
fajar.masya@mercubuana.ac.id

Hendra Prastiawan
Information System, Computer Science Faculty, Mercu Buana University, Indonesia
hendra.prastiawan@mercubuana.ac.id

Krisna Cipta Raksa
Information System, Computer Science Faculty, Mercu Buana University, Indonesia
kciptaraksa@gmail.com

Abstract – Mercu Buana University Campus D is part of Mercu Buana University which began the operational in 2013. Since 2013 until 2017, Mercu Buana University Campus D still got less than a target about getting the new student. This can be due to various things including the lack of precisely marketing strategy undertaken. Therefore, in this study the authors make an application by implementing the concept of data mining using clustering and forecasting methods to obtain information from existing data registrants. So, the information can be used by decision makers to determine effective and efficient marketing strategies.

Keywords: Clustering, Data Mining, Forecasting, Registration

1. INTRODUCTION

Information is a message or collection of messages consisting of order sequences of symbols, or meaning that can be interpreted from a message or a collection of messages. Information can be said as knowledge gained from learning, experience, or instruction. Information can be obtained from some data collected, then processed with various techniques that will produce an output in the form of accurate information. In the world of Technology, Information is very important. Information is very useful for developing a business, expanding market reach, and so on. Mercu Buana University Campus D is part of Mercu Buana University which began the operational in 2013. Since 2013 until 2017, Mercu Buana University Campus D still got less than a target about getting the new student. This can be due to various things including the lack of precisely marketing strategy undertaken. Therefore, in this study the authors make an application by implementing the concept of data mining using clustering and forecasting methods to obtain information from existing data registrants. So, the information can be used by decision makers to determine effective and efficient marketing strategies.
1.2 Research Problems
Based on the above background, the author tries to identify some of the issues that exist among them as follows:
1) How can data mining predict the number of new students at the Mercu Buana University Campus D?
2) How to get new student prediction numbers for next academic year by utilizing the previous data?
3) How can the management make a decision based on data mining analysis technology?

1.3 Limitation of Research
The limitation in this research are:
1) The method used in this research is Prototype Method.
2) The Object location of this research is at Mercu Buana University.

1.4 Purpose And Objectives
The purpose of this research are:
1) Helping Mercu Buana University Campus D campus to be able to know the level of acceptance of new students in the next academic year.
2) Providing knowledge to the Marketing Manager of Mercu Buana University Campus D so as to facilitate in making a decision.
3) Use the past transaction data stored in order to provide various useful information for the marketing department of Mercu Buana University Campus D.

While the benefits derived from this research are:
1) The marketing department can determine marketing strategy effectively and efficiently, so the marketing part can meet the target of new student enrollment.
2) The management can contain a new policy based on the results of analysis of past transaction data that can increase business profits.

II. METHODOLOGY

2.1 Data Mining
Data Mining is a process that employs one or more computer learning techniques to analyze and extract knowledge automatically. Data Mining contains a search for the desired trends or patterns in a large database to help decision-making in the future. These patterns are controlled by certain devices that can provide a useful data analysis and can be studied more closely. Data mining in this context is a step in the Knowledge Discovery in Database (KDD) process.

Fig. 1 KDD Process

2.2 Clustering
Clustering or grouping is a method or function in data mining that aims to find a collection of objects until objects in a group have similarities with other objects that are in one group and not similar to objects in other groups. In addition, the clustering method can also minimize the distance within the cluster and maximize the distance between clusters. In this research, we will use Partitioning Methods. Partitioning Methods or Partition method is the simplest and most fundamental method in Clustering function. In partitioning methods, all objects are required to be within a group and they are binding. In applying partitioning methods there are several algorithms that can be used. Among the most popular are the k-means and k-medoids algorithms.

2.3 K-Means Algorithm
The k-means algorithm is one of the common algorithms used in applying clustering methods with partitioning methods.
The k-means algorithm will assess the quality of an object so that in one cluster have objects that are similar to each other but different from other objects that are in another cluster (Han et al, 2012: 451). The k-means algorithm is as follows:

**Fig. 2 K-Means Algorithm**

In terms of cluster determination of each object, the k-means algorithm uses the distance between objects as its pedestal. In calculating the distance between these objects, there are several methods used include the method of manhattan distance or also known as the city-block distance, euclidean distance, and minkowski distance. But in its implementation, k-means algorithm practitioners more often use the method of manhattan distance or euclidean distance. This is due to the advantages and the ease of both. Manhattan distance can detect and minimize the existence of outliers of datasets while euclidean distance will collect the objects with the smallest distance based. The euclidean distance formula is as follows:

\[ d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + \ldots + (z_1 - z_2)^2} \]

\(d\) = euclidean distance

\(x_1\) = first object

\(x_2\) = second object

While the manhattan distance formula is as follows:

\[ d = |a_1 - a_2| + |b_1 - b_2| + \ldots \]

\(d\) = manhattan distance

\(a_1\) = first object

\(a_2\) = second object

Then the formula for calculating centroid is as follows:

\[ \mu_i = \frac{1}{n_i} \sum_{x_j \in C_i} x_j \]

Where the centroid for the i-th cluster is derived from the sum calculation of all data in the i-th cluster divided by how much data there is in the i-th cluster. And this is an overview of the clustering process:

**Fig. 3 Clustering process**
III. ANALYSIS AND DESIGN SYSTEM

3.1 Problem Identification
In carrying out its duties and functions, the regular marketing team of the Mercu Buana University is required to be able to meet the target of the new students who have been determined. But in practice, every year the targets earned by regular marketing team 1 are always below the specified targets, such as for example by 2015, the target of fresh student is 750, but only 550 are enrolled. It is not without reason, many obstacles faced by the regular marketing team 1 when running promotional activities, such as when doing personal selling activities, the time given by the school does not match the needs of the marketing team, so the information submitted is less complete Accepted by the students. Then the constraints to the accommodation is the lack of transportation that can lead the marketing team to the location of the promotion caused by the lack of operational car campus D. Then the lack of understanding and the way the delivery of the presenter team so that the information submitted was not complete.

3.2 Analysis of Current System

![Fig. 4 Current Process](image-url)
3.3 Problem Solving

Based on the problems that have been described in the previous chapter, the authors propose to create an application called e-Marketing. Basically, this application serves to manage the data of new students who enroll in the University of Mercu Buana Campus D which includes data input, change data, and delete data. In addition, this application can also display the data of students who have enrolled in the university based on a certain period. And this application is also included function to analyze data by utilizing k-means clustering algorithm and also function for forecasting using single exponential smoothing algorithm.

3.4 Proposed System

![Use Case Diagram](image)

**Fig. 5 Use Case of the Proposed System**

Use case diagram contains the interaction between the end user with the application to be created. There are 3 users who can access the system and the three users have different access rights to the system. The first user is an employee or in this case is the marketing officer of Mercu Buana University Campus D, the employee can only access the data management function of the student and view the student registration report only. Then there is the user manager, ie in this case is the head of the marketing department of the University of Mercu Buana. Manager can perform data analysis and view student registration report but can not perform student registration data management. Then the last is the user admin, the user who can only manage the account contained in the system database and share the permissions to each account.

3.5 Class Diagram

![Class Diagram](image)

**Fig. 6 Class Diagram of the Proposed System**
3.6 Activity Diagram

1) Activity Diagram for Data Input

![Activity Diagram for Data Input](image1)

Fig. 7 Activity Diagram for Data Input

2) Activity Diagram for Analyst the Data

![Activity Diagram for Analyst the Data](image2)

Fig. 8 Activity Diagram for Analyst the Data

IV. IMPLEMENTATION

4.1 User Interface for Cluster Data Page

The data cluster page is the page where the k-means algorithm is executed. The data cluster page contains the results tables of the clustered data using the k-means algorithm.
On this page, users do not enter any data into the system. This system will retrieve data from the system database. Additionally, the value of k (cluster number) is determined by default when the system is created.

**4.2 User Interface for Forecast Data Page**

Forecast data page is a page that will run the exponential smoothing algorithm and will display the results as well. Just as the data cluster page, the data forecast page will also retrieve data from the system database, but the data retrieved by the data forecast page will be grouped first by year and month later after grouped, the data will be processed using the exponential smoothing formula. The results of the exponential smoothing algorithm will be displayed on the data forecast page. The results of the algorithm will be presented in the form of annual registration data and forecasting data for the next period. In addition, the result data from the exponential smoothing algorithm will also be presented in the form of a line diagram.
V. CONCLUSION

The conclusion that the author can be taken from the results of this research are as follows:

1) Implementation of data mining by k-means method to group data based on data distance with centroid can give knowledge to manager about pattern of data exist in system database, so that it can assist manager in determining a decision.

2) The application of the forecasting function with the exponential forecasting method can provide predictions of the number of applicants in the next academic year.

3) The system created can help the marketing department of Mercu Buana University Campus D in managing the registration data appropriately.

While this research needs further development, among others in terms of:

1) For the marketing of the Mercu Buana University, in order to keep the registration data properly so that later data can be utilized to improve the performance of marketing.

2) To determine centroid still using random way, forecasting method still not yield maximal result. It is hoped that the researchers can further develop the system better.

REFERENCES