Jurnal Intelektualita: Keislaman, Sosial, dan Sains
Vol. 9, No. 1, Juni 2020
Website: http://jurnal.radenfatah.ac.id/index.php/intelektualita
ISSN 2303-2952, e-ISSN 2622-8491

A Comparison Between Naïve Bayes and The K-Means Clustering Algorithm for The Application of Data Mining on The Admission of New Students

Nurhachita, Edi Surya Negara
Universitas Bina Darma, Palembang, Indonesia
Email: nurhachita@gmail.com

Abstract
The process of admitting new students at Universitas Islam Negeri Raden Fatah each year produces a lot of new student data. so that there is an accumulation of student data continuously. The purpose of this study is to compare the K-Means Clustering Algorithm and Naïve Bayes on the admission of new students as well as being one of the bases for making decisions to determine the promotion strategy of each study program. The data mining method used is Knowledge Discovery in Database (KDD). The tools used are Rapid Miner. The attributes used are national examination score, school origin, and study programs. The new student data used from 2016 to 2019 was an 18.930 item. The results of this study used the K-Means Clustering Algorithm to produce 3 clusters, while the Naïve Bayes results resulted in an accuracy value of 9.08%.

Keyword: Data Mining, Naïve Bayes, K-Means Clustering, New Student

Introduction

Information technology has an important role in most organization that manipulates and collects data in large databases. Stored data can be used to generate useful information for decision making. Data mining is an automatic data analysis process that helps users and administrators to discover and extract patterns from stored data[1]. Along with the development of the internet, the data stored, both in the form of text, images, sound, and video also increased very quickly and significantly. In Indonesia, internet users in 1998 were only 500,000 users whereas by 2015 it was projected that internet users had reached 139 million[2]. The large volume of data volume will become "garbage" in storage if it is not processed into useful information. Data mining technology provides a user-oriented approach to novel and hidden patterns in the data[3]. This is consistent with the definition of data that data is a fact that is recorded but has no meaning. Many universities have

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, 'From Data Mining to Knowledge Discovery in Databases', *AI Magazine* 17, no. 3 (1996): 37–53.

[2] Joko Suntoro, *Data Mining Algoritma Dan Implementasi Dengan Pemrograman PHP* (Jakarta, 2019).

[3] Jyoti Soni et al., 'Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction', *International Journal of Computer Applications* 17, no. 8 (2011): 43–48, https://doi.org/10.5120/2237-2860.

used Information Technology (IT) to support the admission process[4]. The application of information technology to education can also produce abundant student data and learning processes. At universities, data can be obtained from databases, data will continue to grow, such as student data. The use of a data mining techniques to analyze an educational database is expected to be of great benefit to the higher educational institutions[5].

The process of admitting new students at Universitas Islam Negeri Raden Fatah every year produces a lot of new student data. This happens continuously so that there is an accumulation of student data which will continuously increase in the search for student information. Based on the amount of new student data, by managing the data, information that can be seen can be done by the University. Based on the number of new student data, by organizing the data so that information can be accessed and accepted by the university, for example, a compilation of university promotions or outreach and study programs in schools to accept new students, universities access schools for promotion. This causes a waste of budget because too many schools will be visited, and not time efficient. This research will classify and clarify data on admission of new students at Universitas Islam Negeri Raden Fatah by utilizing the data mining process by applying Clustering and clarification techniques. By comparing the two algorithms, the K-Means Clustering algorithm, and Naïve Bayes. The tools used are Rapid Miner. The attributes used are national examination score, school origin, and study programs. Based on the results of the K-Means cluster Clustering Algorithm and Naïve Bayes can determine the promotion strategy of each study program. Based on the results of the cluster K-Means Clustering Algorithm and Naïve Bayes can see courses of interest in each school. The final results of the cluster can help the University.

Data Mining (DM) concept is to extract hidden patterns and to discover relationships between parameters in a vast amount of data[6]. Data Mining is the process of extracting data (previously unknown, implicit, and considered useless) into information or knowledge or patterns from large amounts of data. Data that is considered "garbage" because it is not patterned / not structured and is not useful, is processed (filter) so that it forms information or knowledge or new patterns that are useful[7]. Data mining is a series of processes to explore the added value of information that has not been known manually from a database. The information generated is obtained by extracting and recognizing important or interesting patterns from the data contained in the database[8]. From the explanation above it can be concluded that Data Mining is a step of analyzing the process of knowledge discovery in the database. Data mining is a process

[4] Flourensia Sapty Rahayu, Rangga Deputra Ginantaka, and Y Sigit Purnomo Wp, 'Analisis Manfaat Sistem Informasi Penerimaan Mahasiswa Baru Dengan Metode IT Balanced Scorecard', no. January 2019 (2017), https://doi.org/10.21460/jutei.2017.12.21.

[5] Wilairat Yathongchai et al., 'Factor Analysis with Data Mining Technique in Higher Educational Student Drop Out', *Latest Advances in Educational Technologies*, 2012, 111–16.

[6] Fadhilah Ahmad, Nur Hafieza Ismail, and Azwa Abdul Aziz, 'The Prediction of Students' Academic Performance Using Classification Data Mining Techniques', *Applied Mathematical Sciences* 9, no. 129 (2015): 6415–26, https://doi.org/10.12988/ams.2015.53289.

[7] Joko Suntoro, *Data Mining Algoritma Dan Implementasi Dengan Pemrograman PHP*.

[8] Tri Retno Vulandari, 'Pengertian Data Mining', in *Data Mining, Teori Dan Aplikasi Rapirminer*, 2017, 1.

that employs one or more machine learning techniques (machine learning) to analyze and extract knowledge automatically[9].

Clustering is also referred to as segmentation. This method is used to identify the natural group of a case based on an attribute group, grouping data that have similar attributes. Clustering is an unsupervised data mining method because there is not one attribute used to guide the learning process, so all input attributes are treated the same. Most clustering algorithms build a model through a series of repetitions and stop when the model has centered or gathered (the boundaries of this segmentation have stabilized)[10]. Clustering is data that does not have a label/class so it is often called the unsupervised learning technique. From the explanation above it can be concluded that Clustering is a grouping of data that does not have a class. Clustering is data that does not have a label/class so it is often called unsupervised learning techniques[11]. Grouping (grouping) is part of the science of data mining which is intended without direction (not supervised). Clustering is the process of dividing data into classes or clusters based on the agreed level[12].

K-Means algorithm entered into the application of data mining clustering. K-Means is a repetitive clustering algorithm. The K-Means algorithm sets cluster values (K) randomly, for the time being, they are the center of the cluster or commonly referred to as centroid, mean or "means". Each shelf counts data on each centroid. Clarify each data based on its proximity to centroids. Perform these steps until the centroid value does not change (stable)[13]. The k-means method is the oldest and most widely used clustering algorithm in a variety of small to medium applications because of the ease of implementation[14]. From the above explanation, it can be concluded that k-means is the oldest and easiest algorithm to use. Naïve Bayes algorithm is one of the clarification algorithms based on the Bayesian theorem in statistics. Naïve Bayes algorithm can be used to predict the probability of membership of a class[15].

Naïve Bayes algorithm can be used to predict the probability of membership of a class[16]. In the next explanation, The Naïve Bayes method will be described which is the basis for developing the proposed method, by utilizing the corpus that has been formed, then followed by a discussion of the results of the research and concluding with conclusions[17]. Naive Bayes which is also

[9] Hermawati, 'Analisis Faktor-Faktor Self-Care Terhadap Status Nutrisi Pada Pasien Hemodialisa Di RSUD Dr. Moewardi Surakarta', *Jurnal IImiah Rekamedis Dan Informatika Kesehatan* 7, no. 2 (2017): 29–35.

[10] Vulandari, 'Pengertian Data Mining'.

[11] Hong He and Yonghong Tan, 'Corrigendum to "A Two-Stage Genetic Algorithm for Automatic Clustering" [Neurocomputing 81 (2012) 49-59]', *Neurocomputing*, 2012, https://doi.org/10.1016/j.neucom.2012.02.009.

[12] Tutik Khotimah, 'PENGELOMPOKAN SURAT DALAM AL QUR﷽AN MENGGUNAKAN ALGORITMA K-MEANS', *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer* 5, no. 1 (2014): 83–88, https://doi.org/10.24176/simet.v5i1.141.

[13] Vulandari, 'Pengertian Data Mining'.

[14] Suyanto, *Data Mining Untuk Klasifikasi Dan Klasterisasi Data*, *SpringerReference*, 2017, https://doi.org/10.1007/SpringerReference_5414.

[15] Joko Suntoro, *Data Mining Algoritma Dan Implementasi Dengan Pemrograman PHP*.

[16] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques Second Edition*, *Morgan Kaufmann*, vol. 53, 2013, https://doi.org/10.1017/CBO9781107415324.004.

[17] Tata Sutabri, 'Improving Naïve Bayes in Sentiment Analysis For Hotel Industry in Indonesia', *2018 Third International Conference on Informatics and Computing (ICIC)*, 2016, 1–6.

called as Bayes' Rule is the basis for data mining methods and machine-learning. It creates a model with predictive capabilities. It provides new ways of understanding data and exploring it[18]. The Naïve Bayes classifier technique is used when the dimensionality of the inputs is high. This is a simple algorithm but gives good output than others[19].

Research Method

Knowledge Discovery and Data Mining (KDD) is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from data. The ongoing rapid growth of online data due to the Internet and the widespread use of databases have created an immense need for KDD methodologies. The challenge of extracting knowledge from data draws upon research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing, to deliver advanced business intelligence and web discovery solutions[20]. In this study, the method used for data processing is the admission data by using the stages of Knowledge Discovery in Database (KDD). Knowledge Discovery in Database (KDD) is the process of determining useful information and patterns in data. This information is contained in a large database that was previously unknown and potentially useful. Data mining is one step in a series of KDD iterative processes[21].

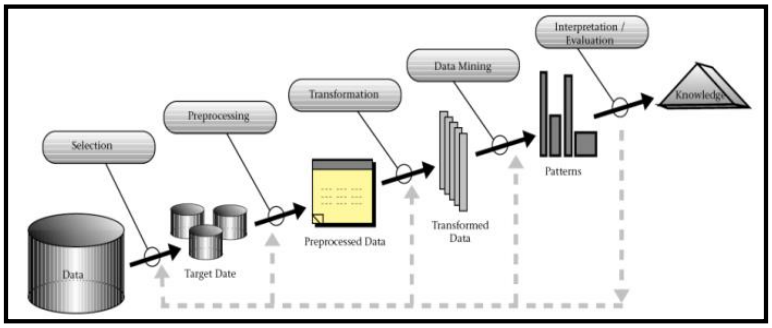

Figure 1. Stages in KDD

The stages of the Knowledge Discovery in Database (KDD) process consist of :

1. Data Selection

In this process the selection of data sets is done, creating a target data set, or focusing on a subset of variables (data samples) where the discovery will be performed. The results of the selection are stored in a separate file from the operational database. The attributes used are national

[18] Sairabi Mujawar and H. P. R. Devale, 'Prediction of Hear t Disease Using Modified K-Means and by Using Naive Bayes', *International Journal of Innovative Research in Computer and Communication Engineering* 3, no. 11 (2015): 0396–0400, https://doi.org/10.15680/IJIRCCE.2015.

[19] Mital Doshi and Setu K Chaturvedi, 'Correlation Based Feature Selection (CFS) Technique to Predict Student Perfromance', *International Journal of Computer Networks & Communications* 6, no. 3 (2014): 197–206, https://doi.org/10.5121/ijcnc.2014.6315.

[20] P.V.Praveen Sundar, 'A COMPARATIVE STUDY FOR PREDICTING STUDENT'S ACADEMIC PERFORMANCE USING BAYESIAN NETWORK CLASSIFIERS', *IOSR Journal of Engineering* 03, no. 02 (2013): 37–42, https://doi.org/10.9790/3021-03213742.

[21] Vulandari, 'Pengertian Data Mining'.

examination score, school origin, and selected study programs. The data in this study were sourced from Universitas Islam Negeri Raden Fatah where this data is secondary data consisting of new student data for 2016 up to 2019. The amount of data obtained was 18,930 consisting of Name, School Origin, National Examination, and Programs Studies. The following are examples of new students data from 2016 to 2019:

Table 1. New Student Data obtained

No	Name	Study Program	School Origin	National Examination Score
1	Nofrizal Ade Wijaya	Al-Quran and Tafsir Sciences	Vocational High School pgri 2 Plg	73
2	Siska Apriyanti	Hadith Science	Islamic Boarding School .Nurul Hikmah	77
3	Sifaul Hasanah	Al-Quran and Tafsir Sciences	Islamic Boarding School Nurul Hikmah	67
------	----------	----------	----------	------
18928	Meeya Maulina Ismala	Islamic studies	Senior High School N 1 Palembang	71
18929	Ali Joyo	Islamic studies	Islamic Boarding School Qodratullah Banyuasin	73
18930	Anita Silvia	Islamic studies	Islamic Boarding School 1 Musi	74

2. Pre-Processing and Cleaning

Data Pre-Processing and Data Cleaning is done by removing inconsistent data and noise, duplicating data, correcting data errors, and can be enriched with relevant external data.

3. Transformation

This process transforms or combines data into a more appropriate way to do the mining process by summarizing (aggregation). Data transformation is done to change the purpose of the data so that the data can be processed using the K-Means Clustering and Naïve Bayes Method. The variables used in the registration of new students are School Origin, National Examination, and Study Program. For the study program data grouped into 40 (forty) groups, school origin data grouped into 3 (three) groups, and National Examination score data grouped into 3 (three) groups. The results of data transformation can be seen in the table below :

Table 2. Transformation Results Data

Id	Study Program	School Origin	National Examination Score
k1	24	2	73
k2	25	3	77
k3	24	3	67
.....	.....	.....	.....
k18928	38	1	71
k18929	38	3	73
k18930	38	3	74

4. Data Mining

Data Mining Process is the process of finding interesting patterns or information in selected data using certain techniques, methods or algorithms under the objectives of the KDD process.

5. Interpretation/Evaluation

The process for translating patterns generated from Data Mining. Evaluate (test) whether the patterns or information found are by or contradictory to previous facts or hypotheses. Knowledge obtained from the patterns formed is presented in the form of visualization.

Results and Discussion

1. K-Means Clustering Algorithm

The data processing of new students using k-means clustering with Rapidminer software can be seen in the following figure:

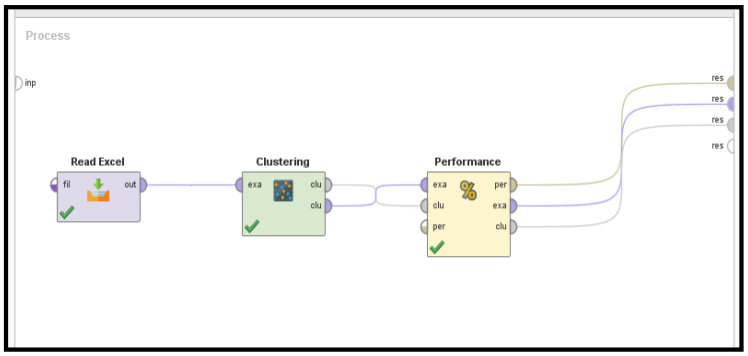

Figure 2. K-Means modeling on Rapidminer

By using the k-means clustering modeling as shown above, with the amount of data 18,930 and initializing the number of clusters of 3, according to the definition of k value with the number of cluster_0: 6927 items, number of cluster_1: 6569 items and number of cluster_2: 5434 items.

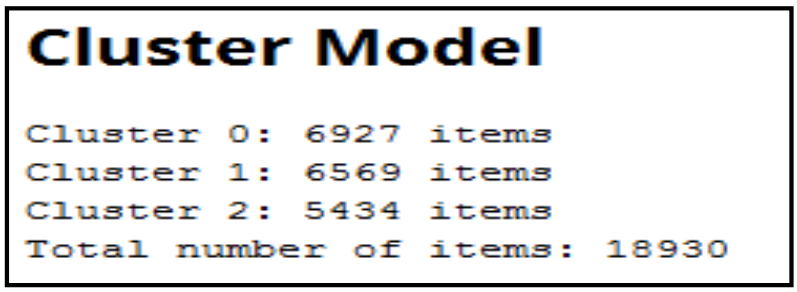

Figure 3. Cluster Model

The results of the spread of cluster_0, cluster_1 and cluster_2 of 18,930 in the k-means clustering modeling using Rapid Miner, for 3 groups of data can be seen in the following figure:

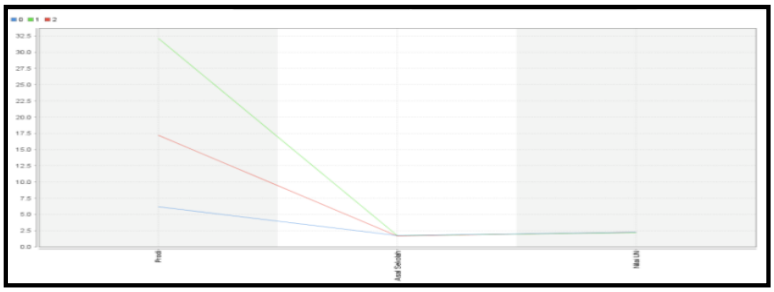

Figure 4. K-Means Clustering Modeling

The cluster analysis results in Figure 3. contain the results of grouping based on the proximity of the distance between the central point and student data on each attribute.

Table 3. Results of First Cluster Analysis (Cluster_0)

Study Program	Amount of Origin of Schools			Amount
	Senior High School	Vocational High School	Islamic Boarding School	
Islamic Family Law	280	48	218	546
Library Science	231	44	53	328
Communication Studies	406	68	85	559
Psychology	387	47	145	579
Sharia Economic Law	506	78	204	788
Information Systems	225	62	65	352
Islamic education	605	96	762	1463
Journalism	324	118	85	527
Islamic Economics	687	175	310	1172

Comparison of Schools	231	51	175	457
Biology	133	7	16	156
Amount				6927
Average of National Examination				66,9129

The results of the first cluster analysis in table 3 above, the highest number of students is in the Islamic Education study program.

Table 4. Results of Second Cluster Analysis (Cluster_1)

Study Program	**Amount of Origin of Schools**			**Amount**
	Senior High School	**Vocational High School**	**Islamic Boarding School**	
Hadith Science	90	25	116	231
Aqeedah and Islamic Philosophy	179	43	159	381
Study of Religions	232	55	101	388
Early Childhood Islamic Education	270	39	131	440
Chemistry Education	171	9	29	209
Biology Education	418	8	87	513
Mathematics education	345	33	96	474
English language education	359	68	145	572
Arabic Language Education	151	26	266	443
Islamic Education Management	481	90	280	851
Madrasa Teacher Education	596	77	320	993
Islamic Criminal Law	453	64	201	718
Islamic civilization	38	5	29	72
Islamic studies	36	3	33	72
Constitutional law	67	5	33	105
Sufism and Psychotherapy	71	15	21	107
Amount				6569
Average of National Examination				73,2559

The results of the second cluster analysis in table 4 above, the highest number of students is in the Madrasa Teacher Education study program.

Table 5. Results of Third Cluster Analysis (Cluster_2)

Study Program	Amount of Origin of Schools			Amount
	Senior High School	Vocational High School	Islamic Boarding School	
Political science	262	55	61	378
Syariah banking	1028	157	261	1446
Physical education	164	10	28	202
Chemistry	122	12	25	159
Zakat and Waqf Management	240	50	62	352
Islamic Community Development	132	36	57	225
Da'wah Management	226	76	1S43	445
Islamic Counseling Guidance	306	49	151	506
Islamic Broadcasting Communication	319	71	148	538
Islamic politics	170	26	66	262
History of Islamic Civilization	164	37	116	317
Arabic language and literature	83	11	76	170
Qur'anic Sciences and Exegesis	156	35	243	434
Amount				5434
Average of National Examination				65,9873

The results of the third cluster analysis in table 5 above, the highest number of students is in the Sharia banking study program.

2. Naïve Bayes

The data processing of new students using Naïve Bayes with Rapidminer software can be seen in the following figure:

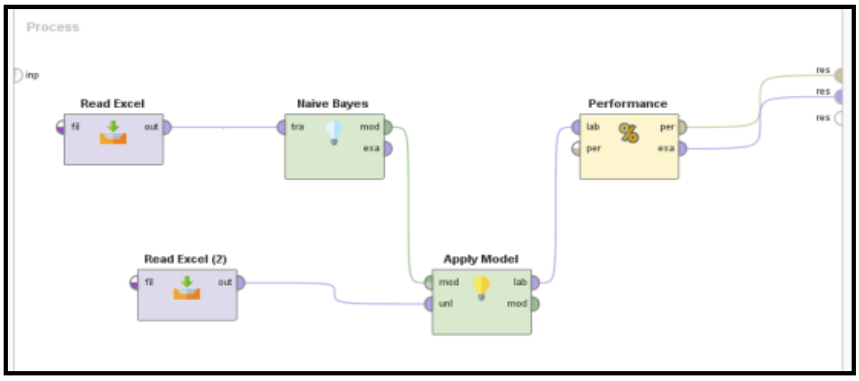

Figure 5. Performance Vector

Using Naïve Bayes modeling as shown above, with the amount of training data (new student admission data from 2016 to 2019) receiving 18,930 and testing data using 2017 new student admission data with a total of 4892. The accuracy of using Naïve Bayes is 9.08% like the picture above.

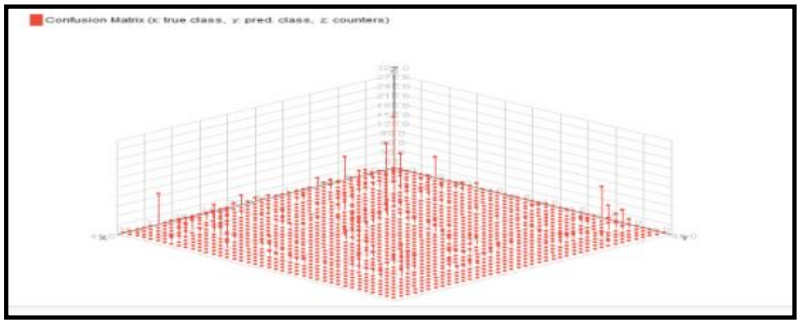

Figure 6. Plot View Accuracy

Data from study programs and prediction study programs from Naïve Bayes use Rapid Miner for new students who use testing data as in the following table:

Table 6. Naïve Bayes Results

Nu	Study Program	Prediction (Study Program)	Amount
1	Islamic Economics	Study of Religions	1
2	Islamic Economics	Biology	9
3	Islamic Economics	Islamic Economics	29
4	Islamic Economics	Mathematics education	43
5	Islamic Economics	Islamic education	72
6	Islamic Economics	Islamic politics	1
7	Islamic Economics	Syariah banking	46

Conclusion

Based on the research and discussion that has been carried out, it can be concluded that from the two methods of K-Means Clustering and Naïve Bayes, in determining the best student

recruitment promotion strategy at the Raden Fatah State Islamic University in Palembang and referring to the original data, the Naïve Bayes method. Data of new students used from 2016 to 2019 were 18930 items. The results of this study use the K-Means Clustering Algorithm to produce 3 clusters, namely the first cluster with a total of 6927 items, the second cluster with a total of 6569 items, and the third cluster with a total of 5434 items. Whereas for Naïve Bayes using data testing in 2017 a total of 4892 items produce an accuracy value of 9.08%.

References

Ahmad, Fadhilah, Nur Hafieza Ismail, and Azwa Abdul Aziz. 'The Prediction of Students' Academic Performance Using Classification Data Mining Techniques'. *Applied Mathematical Sciences* 9, no. 129 (2015): 6415–26. https://doi.org/10.12988/ams.2015.53289.

Doshi, Mital, and Setu K Chaturvedi. 'Correlation Based Feature Selection (CFS) Technique to Predict Student Perfromance'. *International Journal of Computer Networks & Communications* 6, no. 3 (2014): 197–206. https://doi.org/10.5121/ijcnc.2014.6315.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 'From Data Mining to Knowledge Discovery in Databases'. *AI Magazine* 17, no. 3 (1996): 37–53.

He, Hong, and Yonghong Tan. 'Corrigendum to "A Two-Stage Genetic Algorithm for Automatic Clustering" [Neurocomputing 81 (2012) 49-59]'. *Neurocomputing*, 2012. https://doi.org/10.1016/j.neucom.2012.02.009.

Hermawati. 'Analisis Faktor-Faktor Self-Care Terhadap Status Nutrisi Pada Pasien Hemodialisa Di RSUD Dr. Moewardi Surakarta'. *Jurnal IImiah Rekamedis Dan Informatika Kesehatan* 7, no. 2 (2017): 29–35.

Jiawei Han, and Micheline Kamber. *Data Mining: Concepts and Techniques Second Edition*. *Morgan Kaufmann*. Vol. 53, 2013. https://doi.org/10.1017/CBO9781107415324.004.

Joko Suntoro. *Data Mining Algoritma Dan Implementasi Dengan Pemrograman PHP*. Jakarta, 2019.

Khotimah, Tutik. 'PENGELOMPOKAN SURAT DALAM AL QUR﷽AN MENGGUNAKAN ALGORITMA K-MEANS'. *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer* 5, no. 1 (2014): 83–88. https://doi.org/10.24176/simet.v5i1.141.

Mujawar, Sairabi, and H. P. R. Devale. 'Prediction of Hear t Disease Using Modified K-Means and by Using Naive Bayes'. *International Journal of Innovative Research in Computer and Communication Engineering* 3, no. 11 (2015): 0396–0400. https://doi.org/10.15680/IJIRCCE.2015.

Rahayu, Flourensia Sapty, Rangga Deputra Ginantaka, and Y Sigit Purnomo Wp. 'Analisis Manfaat Sistem Informasi Penerimaan Mahasiswa Baru Dengan Metode IT Balanced Scorecard', no. January 2019 (2017). https://doi.org/10.21460/jutei.2017.12.21.

Soni, Jyoti, Ujma Ansari, Dipesh Sharma, and Sunita Soni. 'Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction'. *International Journal of Computer Applications* 17, no. 8 (2011): 43–48. https://doi.org/10.5120/2237-2860.

Sundar, P.V.Praveen. 'A COMPARATIVE STUDY FOR PREDICTING STUDENT'S ACADEMIC PERFORMANCE USING BAYESIAN NETWORK CLASSIFIERS'. *IOSR Journal of Engineering* 03, no. 02 (2013): 37–42. https://doi.org/10.9790/3021-03213742.

Sutabri, Tata. 'Improving Naïve Bayes in Sentiment Analysis For Hotel Industry in Indonesia'. *2018 Third International Conference on Informatics and Computing (ICIC)*, 2016, 1–6.

Suyanto. *Data Mining Untuk Klasifikasi Dan Klasterisasi Data*. *SpringerReference*, 2017. https://doi.org/10.1007/SpringerReference_5414.

Vulandari, Tri Retno. 'Pengertian Data Mining'. In *Data Mining, Teori Dan Aplikasi Rapirminer*, 1, 2017.

Yathongchai, Wilairat, Chusak Yathongchai, Kittisak Kerdprasop, and Nittaya Kerdprasop. 'Factor Analysis with Data Mining Technique in Higher Educational Student Drop Out'. *Latest Advances in Educational Technologies*, 2012, 111–16.