Journal of
Information Systems Engineering
and Business Intelligence

Vol.7, No.1, April 2021
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

Classification and Prediction of Students' GPA Using KMeans Clustering Algorithm to Assist Student Admission Process

Raden Gunawan Santosa[1)], Yuan Lukito[2)*], Antonius Rachmat Chrismanto[3)]

[1)2)3)] *Fakultas Teknologi Informasi, Universias Kristen Duta Wacana, Indonesia*
Jl. Dr. Wahidin Sudirohusodo 5-25, Yogyakarta
[1)]gunawan@staff.ukdw.ac.id, [2)]yuanlukito@ti.ukdw.ac.id, [3)]anton@ti.ukdw.ac.id

Abstract

Background: Student admission at universities aims to select the best candidates who will excel and finish their studies on time. There are many factors to be considered in student admission. To assist the process, an intelligent model is needed to spot the potentially high achieving students, as well as to identify potentially struggling students as early as possible.
Objective: This research uses K-means clustering to predict students' grade point average (GPA) based on students' profile, such as high school status and location, university entrance test score and English language competence.
Methods: Students' data from class of 2008 to 2017 are used to create two clusters using K-means clustering algorithm. Two centroids from the clusters are used to classify all the data into two groups: high GPA and low GPA. We use the data from class of 2018 as test data. The performance of the prediction is measured using accuracy, precision and recall.
Results: Based on the analysis, the K-means clustering method is 78.59% accurate among the merit-based-admission students and 94.627% among the regular-admission students.
Conclusion: The prediction involving merit-based-admission students has lower predictive accuracy values than that of involving regular-admission students because the clustering model for the merit-based-admission data is K = 3, but for the prediction, the assumption is K = 2.

Keywords: Accuracy, K-Means Clustering, Prediction, Student's GPA

Article history: Received 2 October 2020, first decision 11 December 2020, accepted 2 February 2021, available online 28 April 2021

I. INTRODUCTION

Students' admission process is one of the most important aspects in ensuring a higher education institution's quality. Higher education institutions, be it university, polytechnics or institute, always seek to produce high quality graduates. In Indonesia, high-achieving high school leavers usually prefer a public university to a private university. As such, private universities often struggle to attract high-achieving students. With applicants from a wide range of background and ability, private universities need to identify their student's ability in the admission process, as well as in the learning activities so they can provide adequate support when there is difficulty. Universitas Kristen Duta Wacana (UKDW) applies two approaches in its admission: the merit-based and the regular admission. These two admission processes have different requirements, so students could be grouped into two. Merit-based-admission students are selected based on the students' academic achievement record from high school. The regular-admission students are selected based on student's university enrolment test score.

To accelerate the admission process and to achieve the best results, there has to be a rigorous method. Data mining could be a suitable solution to predict the effectiveness. It is commonly used in many sectors, including education, for prediction, clustering, relationship mining, discovery within models and data distillation for judgement [1]. The goals of education data mining are, among others, to predict student's future learning behavior, discover or improving domain models, study the effect of educational support and advance scientific knowledge about learning and learners [2]. Data mining can be used to predict student's academic achievement as in previous studies [3], [4], [5] and [6]. Among the problems that can be solved using data mining are grouping, profiling, planning and scheduling, and detecting cheating in an online examination [7]. Research to predict student's academic trends and patterns using data mining has also been conducted [8]. Student's academic performance can be predicted early in the admission process

* Corresponding author

ISSN 2443-2555 (online) 2598-6333 (print)
doi: http://dx.doi.org/10.20473/jisebi.7.1.1-10

using artificial neural networks technique [9]. Naïve Bayes and decision tree can also be used to improve prediction of low academic performance [10].

Our previous research [11], we used logistics regression method to predict first semester student's GPA. We found logistics regression method to be more suitable to predict the performance of merit-based-admission students (73.73% accuracy) than that of among regular-admission students (56.76% accuracy). This result was improved in subsequent research [12] by using C4.5 and CART algorithms to predict first semester students' GPA. Those algorithms were more suitable to predict the performance of merit-based-admission students (86.86% accuracy) than regular-admission students (61.54% from C4.5 and 63.16% from CART). Although the prediction result for merit-based-admission students was in an acceptable range, some improvements were needed for predicting regular admission student's GPA. Therefore, this study aims: a) to look into the possibilities to improve students' GPA prediction using K-means clustering approach; and b) to find out how much improvement could be obtained in terms of accuracy compared to the previous research.

K-means clustering algorithm is an unsupervised learning algorithm [13] commonly used for creating groups of data based on their characteristics [14]. There has been extensive research using K-means clustering algorithm such as [15] to analyze students' performance by grouping them based on their academic achievements. Teachers can use this information to give support to students who are having difficulty learning in school. K-means clustering algorithm can also be used to group patients based on their medical record (the illness severity): patients with acute disease and patients with non-acute disease. The hospital will be able to provide more appropriate healthcare with this information [16]. K-means clustering algorithm is also commonly used to group customers in order to improve marketing and sales [17]. In higher education, K-means has been used to generate clusters to measure student's academic performance [18], [19], [20] and [21]. In this research, we use K-means clustering for its simplicity and because it has been tested in the previous research.

Research has explored student achievement [19] [20] [21], but only by using predictive models to all students regardless of the fact whether they are admitted based on merit or based on the enrolment test result. This research attempts to fill the gap and to create a predictive model for students' academic performance in high school. The aims of this research are: (1) to find out how K-means clustering algorithms is used to classify and predict student's GPA category into two groups: high GPA and low GPA; (2) to find out how the performance of the resulting predictions is measured using accuracy, F1-measure, precision and recall. This article is organized as follows: introduction, methodology, results, discussion and conclusion.

II. METHODS

This research is conducted in four steps: acquiring and pre-processing data, building clusters, making predictions using K-means cluster algorithm [9] and analyzing the predictions result.

A. Research Variables

UKDW admits students by using two approaches: merit-based admission and regular admission. The current research collected students' admission data from 2008 to 2018. The data collected is as follows: admission type (merit-based/regular), high school ownership (private/public high school), high school type (SMU/SMK), high school location (Java/outside Java), English capability level and first semester GPA. Regular-admission students are accepted based on their university enrolment test scores. This test covers the academic potential to measure numerical, verbal, spatial and analogy competence.

TABLE 1
RESEARCH VARIABLES

Variable Name	Variable Description	Possible Value
X1	High school status	Public = 1, Private = 2
X2	High school location	Java = 1, Outside Java = 2
X3	High school type	SMU = 1, SMK = 2
X4	English language capability	Level 1, 2, 3 dan 4
X5	First Semester GPA	0 – 4.0
X6	Numeric score	0 – 200
X7	Verbal score	0 – 200
X8	Spatial score	0 – 200
X9	Analogy/Logic	0 – 200

Table 1 describes the data used in this research. Variables X1 – X4 are used for merit-based-admission student and variables X1 – X9 are used for regular-admission student. We group student data into 10 groups based on their entrance years, from small group into large group as described in Table 2.

TABLE 2
STUDENT ENTRANCE YEAR GROUPPING

Group Number	Group	Count (Merit Based)	Count (Regular)
1	2008	63	305
2	2008 – 2009	74	554
3	2008 – 2010	140	763
4	2008 – 2011	278	870
5	2008 – 2012	398	989
6	2008 – 2013	523	1069
7	2008 – 2014	613	1150
8	2008 – 2015	806	1211
9	2008 – 2016	905	1322
10	2008 – 2017	1007	1421

K-means clustering algorithm needs a pre-determined number of cluster (k). We use k=2 (two clusters) in this research for every group as shown in Table 2. We assume that two clusters will represent high and low GPA group. We use Davies-Bouldine Index to find optimal k, and the results are displayed in Table 3 for merit-based-admission students and Table 4 for regular-admission students.

TABLE 3
DAVIES-BOULDINE INDEX FOR MERIT-BASED ADMISSION STUDENTS

Group	k=2	k=3	k=4	k=5	k=6	Optimal k
2008	0.882451	0.9821109	1.077107	1.105033	1.079057	2
2008-2009	0.8873462	1.024134	0.8972709	0.9428088	1.090413	2
2008-2010	0.9393635	1.017663	1.382461	1.162064	1.1633031	2
2008-2011	0.97452	1.041165	1.241645	1.13409	1.094575	2
2008-2012	1.057003	1.024022	1.333755	1.23103	1.268759	3
2008-2013	1.338011	1.053382	1.34979	1.234346	1.140603	3
2008-2014	1.291297	1.06031	1.362749	1.459068	1.163056	3
2008-2015	1.312711	1.100396	1.413251	1.303942	1.325148	3
2008-2016	1.310751	1.109655	1.433501	1.353649	1.362768	3
2008-2017	1.325247	1.100878	1.437655	1.366195	1,437855	3

TABLE 4
DAVIES-BOULDINE INDEX FOR REGULAR ADMISSION STUDENTS

Group	k=2	k=3	k=4	k=5	k=6	Optimal k
2008	1.191869	1.543355	1.317229	1.204394	1.162639	6
2008-2009	1.271391	1.581305	1.43425	1.437937	1.343446	2
2008-2010	1.298381	1.671398	1.513807	1.391293	1.277170	6
2008-2011	1.298381	1.616141	1.490241	1.408625	1.316753	2
2008-2012	1.280859	1.519721	1.501357	1.431998	1.369673	2
2008-2013	1.265046	1.510205	1.468167	1.432368	1.360189	2
2008-2014	1.252218	1.524241	1.436597	1.415155	1.390707	2
2008-2015	1.218866	1.415191	1.424077	1.407339	1.404628	2
2008-2016	1.189908	1.396142	1.432092	1.396043	1.374043	2
2008-2017	1.175776	1.388581	1.436559	1.389109	1.376963	2

B. Clusters Building

Cluster are formed based on the similarity of the data characteristics. K-means clustering algorithm uses a distance-based approach to determine an object's cluster. This algorithm starts by choosing centroid (center of cluster) randomly, which is followed by calculating centroid distance to all objects to determine each object cluster. This process will be repeated until all objects are in a stable condition, which means that all objects are already settled in one cluster and the centroids do not change from the previous iteration.

$$dist(c_i, o_i) = \sqrt{\sum_{i=1}^{n}(c_i - o_i)^2} \qquad (1)$$

We use Euclidean distance formula in Equation 1 to measure the similarity between an object (o) and the centroids (c). Variable n in Equation 1 denotes number of attributes.

$$\mu_r = \frac{1}{|n_r|}\sum_{i\epsilon C_r} \bar{x}_i \qquad (2)$$

The centroids will be updated, calculated with Equation 2 to determine new centroid for the cluster [22].Variable n_r represents the count of cluster r member, while $i\epsilon C_r$ is summation of all cluster member attributes (to calculate means). All clusters generated from this step are displayed in Table 5 and Table 6. Data groups used for cluster building are based on Table 2. Variable X1, X2, …, X9 follow conventions described in Table 1.

TABLE 5
CENTROID CLUSTER FOR MERIT-BASED ADMISSION STUDENTS

	Centroid Cluster 1 (High GPA)						Centroid Cluster 2 (Low GPA)					
Group	**X1**	**X2**	**X3**	**X4**	**X5**	**Member count**	**X1`**	**X2**	**X3**	**X4**	**X5**	**Member count**
2008	2	1	1	3	3.12	47	2	1	1	1	1.30	16
2008-2009	2	1	1	2	3.04	58	2	1	1	2	0.53	16
2008-2010	2	1	1	2	3.17	97	2	2	1	2	0.98	43
2008-2011	2	1	1	2	2.96	230	1	2	1	2	0.48	48
2008-2012	2	1	1	2	3.20	264	2	1	1	2	1.63	134
2008-2013	2	1	1	2	3.18	351	2	1	1	1	1.55	172
2008-2014	2	1	1	2	3.19	400	2	1	1	1	1.54	213
2008-2015	2	1	1	2	3.24	470	1	2	1	1	1.79	336
2008-2016	2	1	1	2	3.24	525	2	2	1	1	1.78	380
2008-2017	2	1	1	3	3.27	559	2	2	1	1	1.88	448
Average	2	1	1	2,2	3.16	300.1	1.8	1.5	1	1.4	1.35	180.6
Std. Dev.	0	0	0	0,4	0.09	181.43	0.4	0.5	0	0.49	0.49	151.5

TABLE 6
CENTROID CLUSTER FOR REGULAR ADMISSION STUDENTS

Group	**Centroid**	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**	**X9**	**Member count**
2008	Cluster_1	2	1	1	2	2.40	101	108	111	84	140
	Cluster_2	1	1	1	2	2.42	174	153	155	142	165
2008-2009	Cluster_1	2	1	1	2	2.33	99	100	105	88	243
	Cluster_2	2	1	1	2	2.53	168	153	152	141	311
2008-2010	Cluster_1	2	1	1	2	2.25	97	91	96	85	356
	Cluster_2	2	1	1	2	2.53	165	150	148	140	407
2008-2011	Cluster_1	2	1	1	2	2.16	92	85	92	85	413
	Cluster_2	2	1	1	2	2.49	164	148	146	137	457
2008-2012	Cluster_1	2	1	1	2	2.15	86	78	86	81	458
	Cluster_2	2	1	1	2	2.51	159	146	142	135	531
2008- 2013	Cluster_1	2	2	1	2	2.15	82	77	83	79	506
	Cluster_2	2	1	1	2	2.52	158	144	141	134	563
2008-2014	Cluster_1	2	2	1	2	2.13	81	74	82	78	562
	Cluster_2	2	1	1	2	2.52	157	143	141	134	588
2008-2015	Cluster_1	2	2	1	2	2.13	76	71	78	75	595
	Cluster_2	2	1	1	2	2.51	156	142	141	132	616
2008-2016	Cluster_1	2	2	1	1	2.12	70	65	75	72	647
	Cluster_2	2	1	1	2	2.52	153	139	138	130	675
2008-2017	Cluster_1	1	2	1	1	2.13	65	62	72	70	693
	Cluster_2	2	1	1	2	2.51	151	137	136	127	728
Average	Cluster_1	1.90	1.50	1.00	1.80	2.20	84.90	81.10	88.00	79.70	461.30
	Cluster_2	1,90	1.00	1.00	2.00	2.51	160.50	145.50	144.00	135.20	504.10
Std. Dev.	Cluster_1	0.32	0.53	0.00	0.42	0.10	12.37	14.90	12.86	6.00	177.17
	Cluster_2	0.32	0.00	0.00	0.00	0.03	7.14	5.52	6.11	4.87	172.24

C. Cluster Homogeneity

In this step, we calculate sum of square error (SSE) to measure cluster homogeneity. The SSE of the merit-based-admission students is displayed in Table 7 and the SSE of the regular-admission students is displayed in Table 8.

TABLE 7
SSE FOR MERIT-BASED ADMISSION STUDENTS CLUSTERS (K=2)

Group	SSE_{total}	SSE_{within}	$SSE_{Between}$	Percentage (%)	
				SSE_{Within}/SSE_{total}	$SSE_{between}/SSE_{total}$
2008	139.3668	83.35708	56.0097	59.2	40.2
2008-2009	183.8867	101.583	82.30367	55.2	44.8
2008-2010	316.8607	188.5683	128.2925	59.5	40.5
2008-2011	636.6747	401.2608	235.4128	63.0	37.0
2008-2012	911.4054	607.9039	303.5015	66.7	33.3
2008-2013	1211.606	791.009	420.5970	65.3	34.7
2008-2014	1475.592	935.695	539.8975	63.4	36.6
2008-2015	2035.182	1289.247	745.9349	63.3	36.7
2008-2016	2303.474	1456.033	847.4406	63.2	36.8
2008-2017	2584.679	1646.937	937.7421	63.7	36.3
Average				62.25	37.69

TABLE 8
SSE FOR REGULAR ADMISSION STUDENTS CLUSTERS (K=2)

Group	SSE_{total}	SSE_{within}	$SSE_{Between}$	Percentage (%)	
				SSE_{Within}/SSE_{total}	$SSE_{between}/SSE_{total}$
2008	2,282,015	1,339,343	942,671.5	58.7	41.3
2008-2009	4,482,341	2,769,807	1,712,534	61.8	38.2
2008-2010	7,047,879	4,427,478	2,620,400	62.8	37.2
2008-2011	8,654,840	5,428,938	3,225,903	62.7	37.3
2008-2012	10,382,275	6,451,393	3,930,881	62.1	37.9
2008-2013	11,546,397	7,114,039	4,432,359	61.6	38.4
2008-2014	12,706,946	7,770,251	4,936,695	61,1	38.9
2008-2015	13,936,703	8,343,054	5,593,649	59.9	40,1
2008-2016	15,812,879	9,278,629	6,534,250	59.7	41.3
2008-2017	17,339,785	10,064,450	7,275,335	58.0	42.0
Average				60.84	39.26

The SSE calculation results on Table 7 shows that $SSE_{total} > SSE_{within}$, which means that the two clusters have better homogeny than one cluster. Cluster homogeneity for two clusters is 62.25% on average. As for the regular-admission students cluster, the SSE calculation results also shows that $SSE_{total} > SSE_{within,}$ with average cluster homogeneity being 60.84%.

D. Making Predictions and Calculating Accuracy

We use students' data from class of 2018 as evaluation data., which consists of 109 students from the merit-based admission and 134 students from the regular admission. Each students' dataset from class of 2018 is classified into cluster 1 or cluster 2 using the smallest Euclidean distance between each data and cluster centroids. The evaluation is conducted using Crosstab to calculate the prediction accuracy.

TABLE 9
CROSSTAB FOR TWO CLASS PREDICTION ACCURACY

		Prediction		Sum(row)
		Class 1	Class 2	
Observation	Class 1	n_{11}	n_{12}	b_1
	Class 2	n_{21}	n_{22}	b_2
Sum (column)		k_1	k_2	n

Crosstab is built from the comparison of observation result and prediction result [23] described in Table 9. Variable n_{11}, and n_{22} represent the count of data that has been predicted correctly (belong to class 1 and 2). The accuracy is calculated with Equation 3.

$$accuracy\ (\%) = \frac{n11+n22}{n} \times 100 \quad (3)$$

Apart from using the accuracy formula, several related values are also used, such as precision, recall and F1-measure described in Equation 4-6 [24] [25].

$$precision(\%) = \frac{n_{11}}{n_{11}+n_{21}} \text{ x } 100 \quad (4)$$

$$recall(\%) = \frac{n_{11}}{n_{11}+n_{12}} \text{ x } 100 \quad (5)$$

$$F1 - measure(\%) = \frac{n_{11}}{2n_{11}++n_{12}+n_{21}} \text{ x } 100 \quad (6)$$

III. RESULTS

Prediction results for the merit-based-admission students can be viewed in Table 10 and 11. Based on the training data size (group), there is no correlation between training data size and prediction accuracy. The result is skewed to class 1 because of the asymmetrical data. The number of students with high GPA is more than those with low GPA. The prediction accuracy for the merit-based admission students is 78.259% on average, with 6.1717 standard deviation.

TABLE 10
PREDICTIONS ACCURACY FOR MERIT-BASED ADMISSION STUDENTS

Group	Crosstab				Matched (%)
	n_{11}	n_{12}	n_{21}	n_{22}	
2008	66	10	11	22	80.73
2008 – 2009	92	17	0	0	84.40
2008 – 2010	52	10	26	21	67.00
2008 – 2011	67	16	14	12	72.48
2008 – 2012	77	32	0	0	70.64
2008 – 2013	67	9	12	21	80.73
2008 – 2014	67	9	12	21	80.73
2008 – 2015	59	16	6	28	79.82
2008 – 2016	59	16	6	28	79.82
2008 – 2017	56	7	8	38	86.24
Average					78.259
Std. Dev.					6.1717

TABLE 11
PRECISION, RECALL AND F1-MEASURE FOR MERIT-BASED ADMISSION STUDENTS

Group	Precision (%)	Recall (%)	F1-Measure (%)
2008	85.71	86.84	86.27
2008-2009	100.00	84.40	91.54
2008-2010	66.67	83.87	74.29
2008-2011	82.72	80.72	81.71
2008-2012	100.00	70.64	82.80
2008-2013	84.81	88.16	86.45
2008-2014	84.81	88.16	86.45
2008-2015	90.77	78.67	84.29
2008-2016	90.77	78.67	84.29
2008-2017	87.50	88.89	88.19
Average	87.38	82.90	84.63
Std. Dev.	9.48	5.78	4.58

Based on Table 5, the characteristics of students with high GPA are: public school (X1=2), from Java (X2=1), SMU (X3=1), English level 2 (X4=2.2 ≈ 2) and GPA 3.16 on average. Meanwhile, the characteristics of students with low GPA are public school (X1=1.8 ≈ 2), outside Java (X2=1.8 ≈ 2), SMU (X3=1), English level 1 (X4=1.4 ≈ 1) and GPA 1.35 on average. Class high GPA and class low GPA have three most different characteristics: high school origin (Java or outside of Java), English level and GPA average. There is a high GPA disparity between the two classes.

High school academic record from those in Java is also correlated positively with high GPA. While academic record from high school located outside of Java correlated negatively with high GPA. For the merit-based admission, it is better to use English level as additional student selection criteria rather than depending only on high school academic record.

TABLE 12
PREDICTIONS ACCURACY FOR REGULAR ADMISSION STUDENTS

Group	Crosstab				Matched (%)
	n_{11}	n_{12}	n_{21}	n_{22}	
2008	106	14	0	14	89.55
2008 – 2009	112	6	0	16	95.52
2008 – 2010	106	8	0	20	94.03
2008 – 2011	105	24	0	5	82.09
2008 – 2012	104	3	0	27	97.76
2008 – 2013	103	2	0	29	98.51
2008 – 2014	103	2	0	29	98.51
2008 – 2015	99	5	0	30	96.27
2008 – 2016	97	7	0	30	94.78
2008 – 2017	96	1	0	37	99.25
Average					94.627
Std. Dev.					5.2511

TABLE 13
PRECISION, RECALL AND F1-MEASURE VALUE FOR REGULAR ADMISSION STUDENTS

Group	Precision (%)	Recall (%)	F1-Measure (%)
2008	100.00	88.33	93.81
2008-2009	100.00	94.92	97.38
2008-2010	100.00	92.98	96.36
2008-2011	100.00	81.40	89.74
2008-2012	100.00	97.20	98.58
2008-2013	100.00	98.10	99.04
2008-2014	100.00	98.10	99.04
2008-2015	100.00	95.19	97.54
2008-2016	100.00	93.27	96.52
2008-2017	100.00	98.97	99.48
Average	100.00	93.98	96.75
Std. Dev.	0.00	5.41	2.99

Prediction results among the regular-admission students can be viewed in Table 12-13. This results also do not show any correlation between group size and accuracy. The results are skewed to class 1 because of asymmetrical data. The number of students with high GPA is more than those who have low GPA. K-means clustering algorithm produces 94.627% accuracy on average.

Based on Table 4, the characteristics of low GPA students (cluster 1) are: public school (X1=1.9 ≈ 2), outside of Java (X2=1.5 ≈ 2), SMU (X3=1), English level 2 (X4=1.8 ≈ 2), GPA 2.2 on average, and academic potential test scores (X6=84.9, X7=81.10, X8=88, X9=79.7). Meanwhile, the characteristics of high GPA students (cluster 2) are public school (X1=1.9 ≈ 2), from Java island (X2=1), SMU (X3=1), English level 2 (X4=2), GPA 2.51 on average and academic potential test scores (X6=160.5, X7=145.5, X8=144, X9=135). Although there is a high disparity in academic potential test scores, the student's GPA from two classes is not significantly different.

IV. DISCUSSION

Warnilah [20] conducted a study by using the K-means clustering algorithm to map the achievements of students of SMP Negeri Sukahening. The attributes used in grouping students' achievements are name, extracurricular, knowledge and skill scores, attitude scores, and the number of student absences from class. The study used 173 students as samples with distance calculations performed by using Manhattan distance, Chebyshev distance and Euclidian distance, which resulted in an accuracy of 67%. Our research uses more student's data to improve the accuracy of student's academic prediction. The difference between Warnilah's study and the current study is that we used a cumulative data set for students between 2008 and 2018. In addition to the accuracy value, this study also presents other values on classification problems, namely precision, recall and F1-measure.

Asroni and Adrian [19] also used K-means algorithm to provide recommendations for choosing the best students based on the clusters. The selected students would have the right to participate in the competition. K-means involves the GPA and related courses to support academic skills. This study helps teachers select the best students to participate

in competitions. Our research use K-Means to cluster and predict student's academic performance based on their admission test results and high school characteristics.

Sya'iyah et al. [21] conducted a study by dividing student data clusters into three groups using K-means clustering. The purpose of this study was to obtain the characteristics of high-, medium- and low-achieving students. This study used 724 student data and four variables, namely the GPA, length of study (LS), English proficiency score (EP), and length of thesis assignment (LT). The results of this study were three student characteristics, namely cluster 1 students who have a GPA of 3.28 on a scale of 4, LS 4.52 years, EP 404, and LT 7.46 months. Cluster 2 students have a GPA of 3.29 on a scale of 4, LS 4.48 years, EP 481, and LT 7.26 months. Cluster 3 students have a GPA of 3.31 on a scale of 4, LS 4.50 years, EP 437, and LT 7.14 months. Sya'iyah et al. [21] conducted research using K = 3 clusters and got three characteristics of the cluster but did not make predictions. Our research took a different approach, we use K = 2 with a dataset in order to obtain cluster characteristics for students with a high GPA and students with a low GPA. By using these two clusters, the training data is predicted, and the accuracy is calculated.

Hossain et al. [26] proposed a new K-means clustering algorithm to performs dynamic data grouping. It calculates the threshold value as the centroid of K-means and based on this value the number of clusters is formed. In each K-means iteration, if the Euclidean distance between two points is less than or equal to the threshold value, then these two data points fall to the same group. Otherwise, the proposed method will create a new cluster with different data points. The results show that the proposed method outperforms the original K-means method. The research conducted by Hossain at al. is a theoretical study that aims to improve the performance of K-means clustering, whereas this research is an applicative research which applies K-means clustering to data mining. So, it includes the data mining section, which is often referred to as educational data mining (EDM).

In this research, regular-admission students are selected based on their academic potential scores, hence the students with higher scores will have higher GPA, although the differences between low GPA and high GPA is not significant. The finding supports the current method in the regular admission process that selects student with academic potential scores. Prediction result for regular admission process is better than our previous research [1] and [2]. The comparison is displayed in Fig. 1.

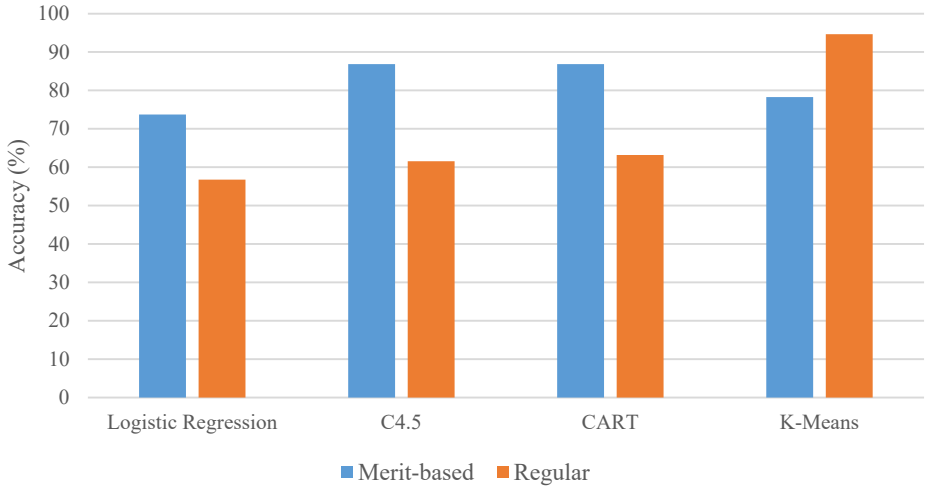

Fig. 1. Comparison of Accuracy from Previous Research

We have calculated the David-Bouldine index which shows that optimal k for our data is k=3 (merit-based-admission students) and k=2 (regular-admission students). The limitation of this research is that we used only k=2 for K-means clustering because we want to cluster students into two clusters based on their GPA (high and low GPA).

V. Conclusions

Based on the result of this research, we can conclude that the model has a lower accuracy in predicting the merit-based-admission students compared to the regular admission students because the clustering model for merit-based admission data is K = 3, but for the prediction, the assumption is K = 2. This research also recommends to use English level as additional criteria for merit-based admission, especially students from outside of Java island. This research can be improved in the future by adding more variables that are relevant to the student admission process depending on the selection process implemented in other higher education institutions.

Author Contributions: *Raden Gunawan Santosa*: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft. *Yuan Lukito*: Funding acquisition, Resources, Validation, Visualization, Writing – review & editing. *Antonius Rachmat Chrismanto*: Funding acquisition, Writing – review & editing.

Funding: This work was supported by Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) Universitas Kristen Duta Wacana under University Excellent Research Scheme, Grand number 061.D.01/LPPM/2020.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

[1] R. Baker, "Data Mining for Education," in *International Encyclopedia of Education*, Oxford, UK: Elsevier 7(3), 2010, pp. 112-118.

[2] R. Baker and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," *JEDM-Journal of Educational Data Mining 1 (1),* 2016.

[3] R. Asiif, A. Merceron, S. A. Ali and N. G. Haeder, "Analyzing Undergradute Students' Performance Using Educational Data Mining," *Computer & Education 113,* pp. 177-194, 2017.

[4] P. Gulati and S. Archana, "Educational Data Mining for Improving Educational Quality," *International Journal of Computer Science and Information Technology & Security (IJCSITS) Vol. 2 No. 3,* pp. 648-650, 2012.

[5] T. Thilagaraj and N. Sengottaiyan, "Review of Educational Data Mining in Higher Education System," in *Proceedings of The Second International Conference on Research in Intelligent and Computing in Engineering Vol. 10*, Gopeshwar, 2017.

[6] N. Bhagoriya and P. Pande, "Educational Data Mining in The Field of Higher Education - A Survey," *International Journal of Engineering Sciences & Research Technology,* pp. 697-699, 2017.

[7] H. Kaur, "A Review of Application of Data Mining in The Field of Education," *International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 4, April 2015,* pp. 409-412, 2015.

[8] S. Parack, Z. Zahid and F. Merchant, "Application of data mining in educational databases for predicting academic trends and patterns," in *2012 IEEE International Conference on Technology Enhanced Education (ICTEE)*, Kerala, 2012.

[9] H. A. Mengash, "Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems," *IEEE Access,* vol. 8, pp. 55462-55470, 2020.

[10] C. E. L. Guarin, E. L. Guzman and F. A. Gonzalez, "A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining," *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje,* vol. 10, no. 3, pp. 119-125, 2015.

[11] R. G. Santosa and A. R. Chrismanto, "Logistic Regression Model for Predicting First Semester Students GPA Category Based on High School Academic Achievement," *Researchersworld Journal of Arts, Science & Commerce,* vol. VIII, no. 2, pp. 58-66, 2017.

[12] D. Alverina, R. G. Santosa and A. R. Chrismanto, "Perbandingan Algoritma C4.5 dan CART Dalam Memprediksi Kategori Indeks Prestasi Mahasiswa," *Jurnal Teknologi dan Sistem Komputer,* vol. 6, no. 2, pp. 76-83, 2018.

[13] K. H. Esbensen, B. Swarbrick, F. Westad, P. Whitcomb and M. Anderson, Multivariate Data Analysis: An Introduction to Multivariate Analysis, Process Analytical Technology and Quality by Design, Oslo, Norway: CAMO Software AS, 2018.

[14] W. K. Hardle and L. Simar, Applied Multivariate Statistical Analysis Fifth Edition, Cham, Switzerland: Springer Nature Switzerland, 2019.

[15] M. K. Singh, A. Rani and R. Sharma, "An Optimised Approach For Student's Academic Performance By K-Means Clustering Algorithm Using Weka Interface," *International Journal of Advanced Computational Engineering and Networking,* vol. 2, no. 7, pp. 2-9, 2014.

[16] A. K. Wardhani, "K-Means Algorithm Implementation for Clustering of Patients Disease In Kajen Clinic of Pekalongan," *Jurnal Transformatika,* vol. 14, no. 1, pp. 30-37, 2016.

[17] K. R. Kashwan and C. M. Velu, "Customer Segmentation Using Clustering and Data Mining Techniques," *International Journal of Computer Theory and Engineering,* vol. 5, no. 6, pp. 856-861, 2013.

[18] S. D. Salam, P. Paul, R. Tabassum, I. Mahmud, M. A. Ullah, A. Rahman and R. M. Rahman, "Determination of Academic Performance and Academic Consistency by Fuzzy Logic," in *2018 International Conference on Intelligent Systems (IS)*, Funchal - Madeira, 2018.

[19] A. Asroni and R. Adrian, "Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik Dengan Weka Interface Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang," *Semesta Teknika,* vol. 18, no. 1, pp. 76-82, 2015.

[20] A. I. Warnilah, "Analisis Algoritma K-Means Clustering untuk Pemetaan Prestasi Siswa Studi Kasus SMP Negeri I Sukahening," *Indonesian Journal on Computer and Information Techology,* vol. 1, no. 1, pp. 83-95, 2016.

[21] K. Sya'iyah, H. Yuliansyah and I. Arfiani, "Clustering Student Data Based on K-Means Algorithms," *International Journal of Scientific & Techology Research (IJSTR),* vol. 8, no. 8, pp. 1014-1018, 2019.

[22] D. T. Larose and C. D. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, 2nd Edition, New Jersey, United States of America: John Wiley & Sons, Inc., 2014.

[23] G. K. Bhattachryya and R. A. Johnson, Statistical Principles and Methods 6th Edition, John Wiley & Sons, Inc, 2010.

[24] S. Bittrich, M. Kaden, C. Leberecht, F. Kaiser, T. Villman and D. Labudde, "Application of Interpretable Classification Model on Early Folding Residues During Protein Folding," *BioData Mining Methodology Open Acces 12: 1 ,* pp. 1-16, 2019.

[25] S. S. Alaoui, Y. Farhaoui and B. Aksasse, "Classification Algorithms in data Mining," *International Journal of Tomography and Simulation , August 2018,* 2018.

[26] M. Z. Hossain, M. N. Akhtar, R. Ahmad and M. Rahman, "A Dynamic K-Means Clustering for Data Mining," *Indonesian Journal of Electrical Engineering and Computer Sciences Vol. 13, No. 2, February 2019,* pp. 521-526, 2019.