EDUMASPUL
JURNAL PENDIDIKAN
Vol. 7 – No. 1, year (2023), page 958-969
| ISSN 2548-8201 (Print) | 2580-0469) (Online) |

Implementation of Feature Engineering and K-Mean Clustering Model in Predictive Analysis of Improving the Quality of Junior High School Education in Bogor Regency

Isbandi[1], Shonia J Kamila[2], Tri Wiwin[3]
[123] Universitas Islam Nusantara, Indonesia
* Corresponding Author. E-mail: [1]isbandis@gmail.com

Abstrak
Dalam rangka meningkatkan kualitas pendidikan. Pemerintah Republik Indonesia telah menetapkan Standar Nasional Pendidikan (SNP) yang merupakan kriteria minimum yang ditetapkan oleh pemerintah dalam sistem pendidikan. SNP merupakan standar yang harus dipenuhi oleh setiap sekolah dan seluruh pemangku kepentingan dalam mengelola dan menyelenggarakan pendidikan. Melalui analisis data menggunakan metode feature enginering dan K-mean untuk memprediksi kualitas pendidikan sekolah menengah pertama di Kabupaten Bogor menunjukkan bahwa standar pendidikan sekolah menengah pertama di Kabupaten Bogor dipengaruhi secara signifikan oleh nilai pencapaian 3 standar mutu pendidikan, yaitu standar pendidik dan tenaga kependidikan, standar sarana dan prasarana, serta standar pengelolaan. Sehingga Dinas Pendidikan Kabupaten Bogor perlu memfokuskan arah kebijakan untuk meningkatkan mutu pendidikan SMP pada indikator yang mempengaruhi 3 standar tersebut. Hasil klasterisasi K-mean mutu sekolah menunjukkan bahwa SMP di Kabupaten Bogor dibagi menjadi 2 kelompok besar tingkat mutu pendidikan, yaitu di bawah nilai tingkat mutu pendidikan 5,07 sebanyak 57,12% SMP dan sisanya di atas nilai tingkat mutu pendidikan 5,07 sebesar 42,88% SMP. Pengembangan lebih lanjut dari model tersebut perlu dikembangkan lebih lanjut ke tingkat sub-indikator sehingga mutu sekolah dapat lebih spesifik diukur, dan data yang digunakan sebagai bahan analisis diperluas dan diperkaya oleh data pendidikan lain dari berbagai sektor sehingga model yang dikembangkan dapat lebih komprehensif dalam mengukur pencapaian mutu pendidikan.

Kata kunci: sistem penjaminan mutu, K-mean, fitur enginering, pemodelan

Abstract
In order to improve the quality of education. The government of the Republic of Indonesia has established the National Education Standard (SNP) which is the minimum criterion set by the government in the education system. SNP is a standard that must be met by every school and all stakeholders in managing and delivering education. Through data analysis using feature enginering and K-meann methods to predict the quality of junior high school education in Bogor Regency shows that the junior high school education standards in Bogor Regency are significantly influenced by the achievement value of 3 education quality standards, namely educator and education staff standards, facilities and infrastructure standards, and management standards. So that the education office of Bogor Regency needs to focus the direction of policies to improve the quality of junior high school education in indicators that affect these 3 standards. The results of the k-men clustering of school quality showed that junior high schools in Bogor Regency were divided into 2 major groups of education quality level, namely below the value of the 5.07 education quality level as many as 57.12% of junior high schools and the rest above the value of the 5.07 education quality level of 42.88% of junior high schools. Further development of the model needs to be further developed to the level of sub-indicators so that school quality can be more specifically measured, and the data used as analysis material is expanded and enriched by other educational data from various sectors so that the model developed can be more comprehensive in measuring the achievement of education quality.

Keywords: quality assurance system, K-mean, feature engineering, modeling

INTRODUCTION

Education is one of the important gates to improve community welfare (Fadliyah et al., 2019). Education opens up opportunities for individuals and communities to develop themselves and realize their ideals. In this context, education is a means to obtain knowledge and is a basic right of every resident so that the fulfillment of this right is an obligation of the government (Inkiriwang et al., 2020). Considering the quality of education in Indonesia, the current condition of education in Indonesia is far from good. This is evidenced by the Education Index published by the Human Development Report – *United Nations Development Progrogmme* (UNDP), that Indonesia is ranked1.16 out of 1.91 countries with a Human Development Index (HDI) index of 0.705 in 2021/2 022 . The INDP data shows that the increase in education in Indonesia is relatively slow with an increasing rate of 0.95% HDI for the period 1990-2022 (United Nations Development Programme., 2022). Nevertheless, the Indonesian government continues to strive to improve the quality of education. One of these efforts is to establish a National Education Standard (SNP) which is the minimum criterion set by the government in the education system in all jurisdictions of Indonesia (Government Regulation No. 4 of 2022).

SNP is a standard that must be met by every school and all stakeholders in managing and delivering education. In an effort to achieve SNP compliance, each school must be able to evaluate the implementation of education based on article 1 number 21 of the National Education System Law concerning education quality control, education quality assurance, education quality determination, and education components (National Education System Law No. 20 of 2023). There are eight national education standards, namely graduate competency standards, content standards, process standards, educator and education personnel standards, facilities and infrastructure standards, management standards, financing standards, and assessment standards which are the pillars of education quality in Indonesia.

It can be seen from several educational conditions that occur in several remote areas of the country, education problems are the main thing in development as well as Bogor District. Currently in West Java, especially in kab. Bogor has 3,129 elementary, junior high, high school, and vocational levels whose management can be categorized into basic education groups, namely 1,854 elementary schools and 724 junior high schools managed under the authority of the district government. Bogor (Dapodik Ministry of Education and Culture, 2019).

At the primary and secondary education levels, all school data management is carried out through a system facilitated by the Ministry of Education, Culture, Research and Technology (Kemendikbudristek), one of which is education quality data collected through the Education Quality Assurance (PMP) system. The amount of education data managed by the education office, especially kab. Bogor collected in the database of the Ministry of Education and Culture provides information that refracts the condition of education in Bogor District. This study seeks to see the patterns and characteristics of information from various data sources will be identifiedas dominant factors that affect the quality of education in Bogor district, especially at the junior high school level. With the aim of knowing school groups that need more attention to improve the quality of education.

In the context of data science, machine learning, data mining, and data analytics, a feature is an attribute or variable used to represent or describe an aspect of a particular object (Utomo &; Mesran, 2020). Informative features are the basic fundamentals of data analysis. This feature describes the underlying object, and to distinguish and characterize different groups of objects (explicit or latent). Features are critical to generating prediction models that are accurate and easy to explain, and produce good results in a variety of data analytics tasks. the data reviewed represents a specific issue within a domain (Butcher &; Smith, 2020). According to Butcher&; Smith (2020), there are several main steps in managing features as the basis of analysis and model formation, namely:

- Feature Understanding, understanding data and problem domains with qualitative and quantitative information owned
- Feature Improvement, cleaning data, filling in blanks, transforming unstructured data, and normalizing data
- Feature Transformation, the formation of new features from existing features; This is often achieved using mathematical mapping.
- Feature Generation / Construction, Producing new features that are often not the result of feature transformation. In addition, it can be said that defined features of pattern/texture are one of the results of feature generation. Many domain-specific ways to define features also fall under the category of feature creation. Sometimes the term feature extraction is used for feature generation.
- Feature Selection: Selecting a small set of features from a very large set of features. The reduced feature set size makes it computationally feasible to use certain algorithms. The selection of features can also lead to improved quality in the results of the algorithm.
- Feature Analysis, concepts, methods, and measures to evaluate the usability of features and feature sets. Feature analysis is also often included as part of feature selection.
- General Automatic Feature Engineering methodology is about a generic approach to automatically generate a large number of features and select an effective subset of the resulting features.

Clustering is a method for finding subgroups in observations used widely in applications such as data segmentation to find the structure of data groups (Ray, S., 2019). K-Means is a clustering algorithm that divides observations into k clusters. Because it can determine the number of clusters, this algorithm can easily be used in classification where we divide data into clusters that can be equal to or more than the number of classes (Cady, 2017). To do clustering, several stages are needed, including:

- Determination of k value or cluster to be created
- Initialize centroid values randomly
- Centroid is the center value of a cluster.
- Setting each data point to the nearest centroid, at this stage will calculate the distance of each data to the centroid using Euclidean distance.
- Recalculate the centroid value of the newly formed cluster.
- optimize until criteria are met

This algorithm is good at clustering data that has a spherical form. However, identification needs to be initialized by determining the number of clusters through the elbow analysis method to determine the most optimal number of K (Yuan &; Yang, 2019).

METHODS

Problem solving is carried out by implementing several stages, namely:

a. Data Preparation, a process to prepare data so that it can have reliability as a basis for solving problems, through the process of cleaning and transforming raw data. This step is an important step before data processing by involving the process of reformatting the data, making corrections to the data, and combining data sets to enrich the data. This process is important as an initial prerequisite for placing data in context and providing insight into the data by eliminating bias generated due to poor data quality.
b. Statistical Analysis, this analysis is an initial perspective in analyzing and treating existing school quality information.
c. Correlation analysis and Feature density, the purpose of applying this method is to see the significance of the factors that affect each educational standard reviewed, as material for further analysis.
d. Clasification Method, a method applied to map form a classification model by implementing the k-mean clustering algorithm.

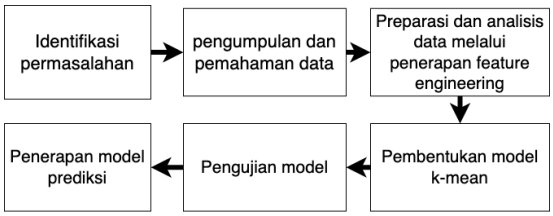


Figure 1 Prediction system model development stage

The selection and formation of the k-mean model will be based on the principle of constructive feedback from existing data. The process of building models, evaluating feedback, and making model improvements is carried out to obtain the desired level of accuracy. Evaluation and analysis of those metrics are used to explain model performance. In particular, the stage of development of the education quality recommendation system can be seen in figure 1.

RESULTS AND DISCUSSION

Taking into account the goal to be achieved, namely forming a model that can measure the performance of school education quality, the data needed for this study is school education quality data reported by schools through the education ministry system that is verified and validated by interested parties. This data is collected by the Ministry of Education and Culture and Technology through data input by schools once a year through the Education Quality Assurance (PMP) system provided by the Ministry of Education and Culture and Technology (Rahminawati, 2021). The primary and secondary education quality assurance information system is an information system that integrates all data and information about the quality of primary and secondary education in accordance with the National Education Standards. This education quality assurance information system aims to support the process of mapping the quality of education at the level of education units, regions and nationally (Ngabidin, 2020).

Education Quality Assurance Data is school education quality reporting data that can provide an overview of school quality in an education period through the Kemendikbudristek education quality assurance system. The main data in the prediction system analysis is Education Quality Assurance (PMP) data. obtained through:

- General information of a public nature, from direct access through rapormutu.pmp.kemdikbud.go.id systems,
- PMP detail data that is private is obtained through the rapormutu.pmp.kemdikbud.go.id system.
- PMP data reporting school profile, from dapodic data. Public dapodic data is obtained through dapo.kemdikbud.go.id addresses, while details of private dapodic data are obtained through the Ministry of Education and Culture's datamart specifically for West Java province.

Data Overview

Taking into account the applicable regulations, the data at the Ministry of Education

and Culture is centralized data that is only managed by the Ministry of Education and Culture and Technology, so the data used is only based on PMP data in the Ministry of Education and Culture and Technology. This PMP data will be the main focus in the formation of analysis and models, and contains data on indicators of each education standard that refers to the National Education Standard (SNP). These indicators can then be taken into account to see the achievement of school education quality. By paying attention to the information brought by PMP data, this PMP data is in accordance with the target of forming a performance prediction model that is the target of the data science project. In full PMP data, indicators and education standards and the relationship between the three are shown in figure 2.

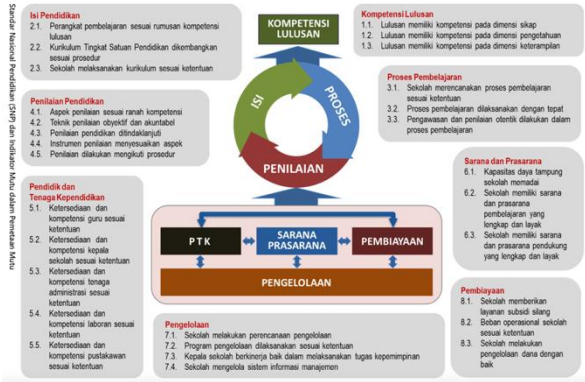


Figure 2. Relationships between attributes in PMP data

As a first step in understanding the data to be analyzed, the PMP data obtained is carried out by the data migration process into the form of a python dataframe. Specifically, this pmp data carries has data attributes as shown in figure 3.

```
dt_raw_pmp_2019.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18850 entries, 0 to 18849
Data columns (total 6 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   npsn     18850 non-null  int64
 1   nama     18850 non-null  object
 2   jenjang  18850 non-null  object
 3   nomor    18850 non-null  object
 4   uraian   18850 non-null  object
 5   pmp2019  9635 non-null   float64
dtypes: float64(1), int64(1), object(4)
```

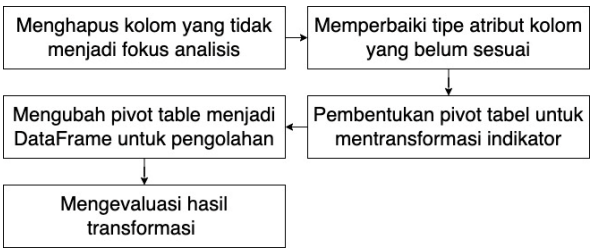
Figure 3. PMP data attribute data type

Based on the information it carries, PMP data has 5 types of data attributes, namely:

a. NPSN : National School Education Number. This NPSN attribute is a school identifier code in Indonesia that has a unique nature because this code distinguishes one school from another.
b. Name : the official name of the school that provides the education quality assurance data report
c. Level: The level of school education that provides PMP data reports, in this case is the Junior High School (SMP) level
d. Number : is the indicator code / sequence number for each educational standard reported by the school
e. Description: a full breakdown of indicators of each school's reported educational standard
f. PMP2019: The value of each education standard indicator reported by schools in 2019.

Data transformation and preparation

Taking into account the form of data above, the data needs to be transformed first so that the data can be understood better. The process of transforming the form of data structures is accompanied by changes and simplification of attributes/columns, especially the description columns in PMP data. This description column will then be stored in the reference column so that only attributes carried by the main structure: npsn and a series of indicator numbers are placed into the row structure.



From the results of data transformation, it shows that the pmp data table has 19 attributes/columns of education standard indicators, and the total number of schools that report pmp data is 355 schools (number of npsn rows). The indicator value of the results of statistical analysis and transformation contained in the indicator has a maximum range of 7 in accordance with the assessment range provided by the school, which is between 0 to 7. However, from all the education quality indicator data reported by the school, it turns out that many attributes contain null values. which needs to be cleaned first so that the data to be analyzed is more representative.

The PMP data representation is visualized using the msno.matrix nullity matrix graph shown by figure 4. The graphinforms that pmp data has null values and NaN which is an impurity in the data. How this NaN and Null data distribution can be. The msno graph of PMP data also provides information that NaN values are spread systematically in one row of school data, and randomly across various attributes/columns.

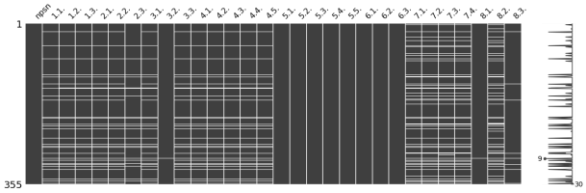


Figure 4. Matrix msno spread null PMP data

To solve the problem of null and Nan data in PMP data, the following handling steps are needed:

1. Cleaning data that systemically appears predominantly in school data rows. In this initial step, data deletion was carried out that had a NaN value of more than 90% for each school (npsn)
2. Because the threshold limit of NaN value is now less than 5%, the 2nd stage operation can be carried out, namely filling the Nan Value by the average method. This imputation process is implemented with the pandas command. After the results of data cleansing, this stage total row (npsn) is now 305 rows.

Figure 5 shows the results of data confectionery that PMP data no longer has null and NaN through filling in the average value.

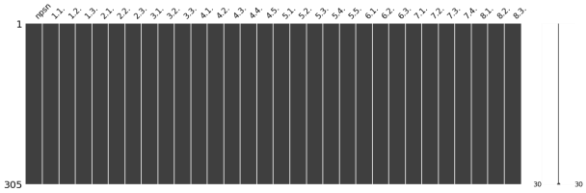


Figure 5. Matrix msno end of PMP data after filling in the average value

After cleansing the NaN and null values, then cleaning the indicator values that have the potential to become an outlayer in the data. To review the potential outlayer, PMP data analysis was carried out through visual analysis through boxplot graphs.

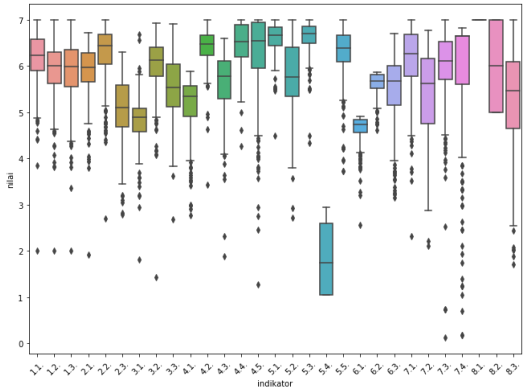


Figure 6. Boxplot graph of PMP data distribution

By looking at the box-plot graph of figure 6, the PMP 2019 data contains data that is considered an outlier because it is far beyond the range of low and high boundaries (points far from the boundary are represented by points far below and above it). Furthermore, the outlayer handling process will be carried out by taking into account the interquartile, low and high boundary values. This process is needed to identify the boundary of each boundary outlier to base winsorzing calculations in reducing/eliminating outliers. The implementation process of removing outliers is shown by the steps shown in figure 7 and the final result is shown in figure 8.

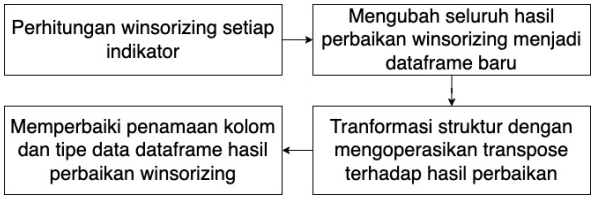


Figure 7. Steps to handle boundary outliers

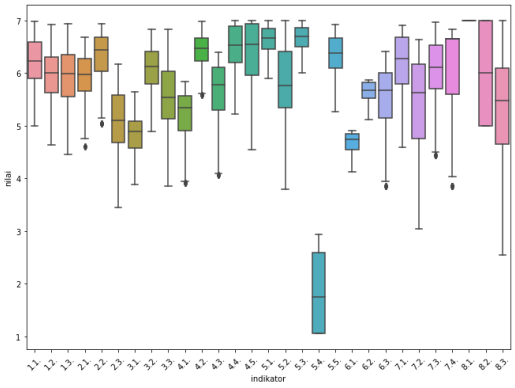


Figure 8. Boxplot chart of PMP data distribution after outlier repair

Engineering and Analysis of Attributes/Analysis Objects

By obtaining data that has been cleared of noise and outliers, then the calculation of the standard value of reporting for each school can be carried out. This standard value according to the Kemendikbudristek pmp guidelines is calculated based on the average of each indicator reported. To fulfill further analysis plus new attributes, as many as 8 attributes/columns are calculated based on the average of indicators for each education standard.

In addition to adding new attributes from S1 to S8, there is also an addition of column attributes that will become object labels, namely:

- SNP value, the value of which is the average of the entire standard of each object.
- Quality level value, this value is a categorization of the results of reporting each school education quality condition that reports PMP. The process of forming SNP labels is done by defining categoristic functions and applying them through operations to each row of the npsn/row object

By operating all the calculations above, it will provide new data conditions shown by table 1.

Table 1 Data snippet after adding PMP data labels.

...	S1	S2	S3	S4	S5	S6	S7	S8	SNP	level
...	5.881300	5.482167	5.038350	4.8540	5.530031	4.909500	5.541457	6.626667	5.482934	SNP_4
...	6.534967	6.081000	6.134933	6.6212	5.748414	5.512167	6.472320	6.860000	6.245625	SNP_4
...	5.179033	5.410833	4.901400	5.8864	5.037015	4.873500	4.532690	4.938667	5.094942	SNP_4
...	6.887167	6.565000	5.470383	6.4422	5.217048	4.991500	5.496065	6.194000	5.907920	SNP_4
...	6.693000	5.998167	6.281917	6.3172	6.038847	5.448833	5.967211	6.473333	6.152313	SNP_4
...	...	...	...	...	...	...	...	...	...	...
...	5.742033	5.451000	5.445300	6.4270	5.458532	5.263833	5.352336	5.843333	5.622921	SNP_4
...	6.852600	6.202000	6.102983	6.6018	5.769664	5.448833	6.406053	5.497333	6.110158	SNP_4
...	5.289633	5.402333	4.822167	5.5830	4.988715	5.185500	5.762011	7.000000	5.504170	SNP_4
...	6.655733	6.198333	5.950917	6.3642	5.470914	5.523500	5.764982	5.791333	5.964989	SNP_4
...	6.012667	6.271500	5.613300	6.4640	5.191881	5.396500	6.608262	6.333333	5.986430	SNP_4

The addition of this label obtained the standard value and level of each data object (npsn / school) that describes the standard value of the school. The results of the analysis show that the junior high school level in Bogor Regency is only at 2 SNP levels, namely SNP 3 and SNP 4, none are at SNP 1 and 2 levels.

When further analyzed against PMP 2019 data which is the focus of the data, it shows that the level of all objects at SNP level 4 is 295 objects and SNP level 3 is 10 objects. By looking at the data, it appears that there is an imbalance problem with the label data. Taking into account the number of SNP imbalance levels between levels (SNP 4 and SNP 3), an additional dataset derived from the previous year's PMP data for SNP level 3 in 2018 was added. In order for this data to be used as an additional data balancer for model formation data, the 2018 pmp data has been processed/prepared the same as the steps as handling the 2019 pmp data. The amount of dataset addition data from the 2018 data can be shown in table 2.

Table 2 proportion of total PMP data for dataset imbalance solution.

Data	PMP 2019	PMP 2018	Penyesuaian imbalance	%
SNP 4	295	352	295	55,24
SNP 3	10	229	239	44,76
Jumlah data	534			

Improvement and Analysis of Data Features

Taking into account the condition of all previously prepared data, the processed data shows that it only has two quality data level conditions for the junior high school level in Bogor District, namely SNP 3 and SNP 4, this condition can be considered PMP data at the junior high school level has binary characteristics by adjusting the meaning of the data to:

- SNP 4 is a school that has met the National Education Standard
- SNP 3 schools have not met the standards.

So that the school-level feature label can be corrected to a grade of 1 for schools that have met the standard and a grade of 0 for schools that have not met the standard. Improvement of the junior high school education level label feature is carried out by implementing the following functions:

```
def label_snp(level):
    if level=='SNP_3':
        return 0
    elif level=='SNP_4':
        return 1
```

This feature improvement can be viewed as a function that will replace the level label into numeric form, so that the attribute structure as shown by table 3 is obtained

Table 3. PMP data snippet after feature data repair.

	S1	S2	S3	S4	S5	S6	S7	S8	SNP	level
255	5.297533	5.990667	5.177800	6.287600	5.806782	4.765833	4.511620	6.114000	5.493979	1
405	5.621648	5.176946	6.570354	5.717226	1.471525	2.498791	5.013185	6.091436	4.770139	0
478	5.566090	6.315749	6.873637	6.291303	1.922321	2.081825	5.383784	5.774118	5.026103	0
116	6.209567	5.775500	5.487633	5.920200	5.518147	5.679833	6.357924	5.960000	5.863601	1
178	6.626467	6.139500	5.724783	6.254000	5.789197	5.232833	6.163770	5.687333	5.952235	1
349	5.461210	5.922908	6.538175	6.437867	0.829553	2.940832	5.500599	6.170465	4.975201	0
521	5.786347	5.508021	6.323077	5.628804	1.261763	2.595007	5.363903	6.023694	4.811327	0
321	5.417038	4.674719	6.153226	5.894991	2.447740	3.399647	4.978833	6.002507	4.871088	0
60	5.716467	5.716667	5.534150	6.204800	5.649731	5.231167	5.669282	6.766667	5.811116	1
267	6.613967	6.338833	5.813367	6.541000	5.438947	5.498833	6.386457	5.660000	6.036426	1

By correcting the feature to numeric, the data can be further processed for model formation material.

Furthermore, to see the relationship between quality standards, the term that characterizes the "relationship" between features is correlation. This area is most explored during PCA (Principal Component Analysis). The idea is that not all features are important or at least some of them will correlate strongly, or in other words: if two features correlate strongly it will manifest the same information and consequently one of the data can be deleted.

Figure 9. Hotplot correlation between standard data attributes in PMP data

Based on the data and heatmap graphs shown by figure 9, correlation 1 in the box that is in the diagonal position of the matrix is the relation of a feature itself. While the other box besides the diagonal shows a correlation between the 8 educational standards evaluated. Reference to determine the closeness of relationships, the following criteria are used:

- 0.00 - 0.19 Very low relationship
- 0.20 - 0.39 Low relationship
- 0.40 - 0.59 Quite meaningful relationships
- 0.60 - 0.79 Strong relationships
- 0.80 - 1.00 Very strong relationship

Here is considered feature-to-feature correlation and result-to-feature correlation. Among the features: higher correlation means that it can remove any of them. However, the high correlation between features and outcomes means they are important and retain a lot of information. In the graph, you can see the color and value correlation between the feature and the result. Thus, the highest values/most important features are 'S5' (0.95), 'S6' (0.94), 'S3' (0.6), and 'S7'

(0.69). Furthermore, the correlation between the two relatively lowest is 'S2' (0.24.

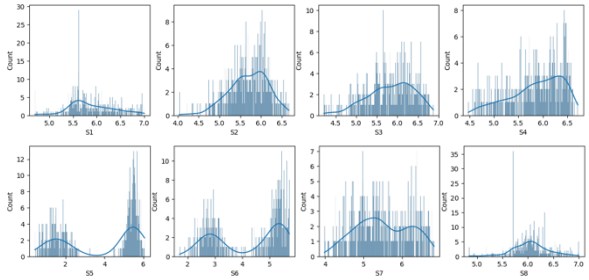


Figure 10. Histogram graph of each educational standard in PMP data

These results are then verified and analyzed using histogram graphs and density plots for each feature relevant to the results as shown in figure 10. The histogram graph shown informs the normality condition of each education quality standard data. Based on the PMP data histogram, the graduate competency standard (S1) has a positive skew value or leaning to the right, so that in this standard the mode value is greater than the average, and the median value is on the left of the average and right of the mode value.

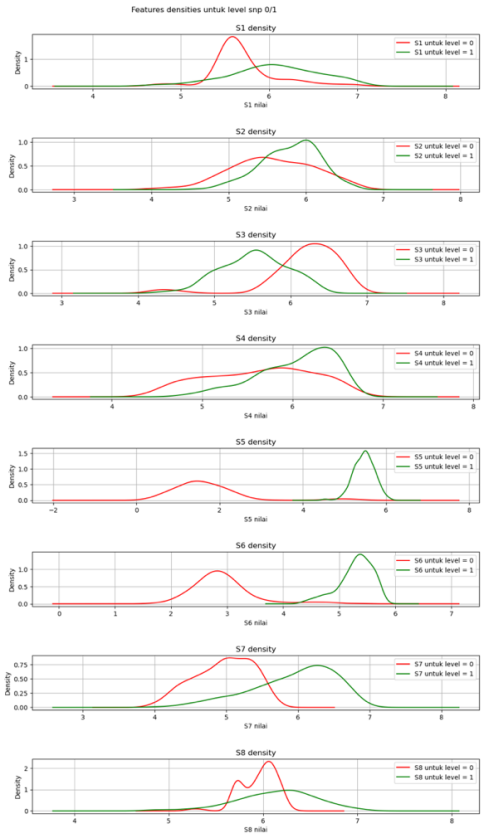


Figure 11. Histogram feature density of each educational standard in PMP data

In the content standard value (S2), process standard (S3), and financing standard (S8) the graph has an almost normal form, while in the assessment standard (S4) the skew histogram is negative or leaning left, so that in this assessment standard the mode value is to the right of the average, and the media. In particular, in the Educator and Education Personnel Standard (S5), facilities and infrastructure standards (S6) and management standards (S7) have abnormal forms and show as if there are two separate normality graphs, this condition shows sparity from level values below and above the standard that are significantly different. To further differentiate, the representation is more detailed into feature densities for each standard.

Plot figure 11. Informs that when the green and red lines are almost equal (overlapping), it means that the distinguishing feature does not distinguish output. Meanwhile, when the green and red lines shift far away, the feature can distinguish the output. In the case of educational standards in Bogor district, S2 or content standards appear to be slightly separated (horizontal shifts between curves) while in S5 or PTK standards there will be a clear distinction. This condition is in line with the results shown by the correlation value. based on the results of the picture, if 3 features are selected that can be distinguished by differentiating the levelity of educational standards are the standards of Educators and Education Personnel (S5), Facilities and Infrastructure standards (S6) and Management standards (S7).

Model K-mean

This clustering analysis model will be used as the main model in predicting school quality performance because basically the education quality data provided by schools does not / does not have object label attributes that distinguish the quality of the school directly. The process of forming a model with a clustering model is carried out by implementing a k-mean analysis.

Based on the dominant education standard data as a differentiator, each school standard object data can be reviewed by the forming group/cluster as a basis for grouping school quality assessment. In the initial process of k-mean analysis, tuning the k value was carried out using the WCSS graph to see the elbow of changes in the WSCC value formed by the data.

Based on the PMP dataset at the junior high school level of Bogor Regency, the WSCC value can be calculated as follows:

```
# mengimport library KMeans
from sklearn.cluster import KMeans

wcss = []
for k in range(1,11):
    kmeans = KMeans(n_clusters=k, init="k-means++")
    kmeans.fit(dtc_pmp)
    wcss.append(kmeans.inertia_)
plt.figure(figsize=(12,6))
plt.plot(range(1,11), wcss, linewidth=2, color="red", marker ="8")
plt.xlabel("K Value")
plt.xticks(np.arange(1,11,1))
plt.ylabel("WCSS")
plt.grid(b=True)
plt.show()
```

From the calculation above, the WSCC plot is obtained as shown in figure 12.

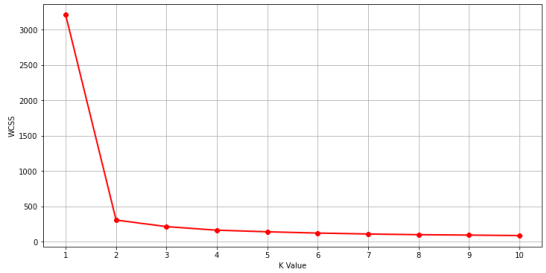


Figure 12. WCSS Graph of Junior High School Education Quality Standards

Based on the WCSS graph, the elbow value is at the value of k = 2, this k value is then used as the value of the clustering parameter using the k-mean method. With a model with a value of k = 2, the cluster value can be calculated as follows.

```
kmeans = KMeans(n_clusters = 2, init = 'k-means++', random_state = 42)
clusters = kmeans.fit_predict(dtc_pmp)
```

This value is then used as a reference for the formation of labels in the k-mean classification through the following python code implementation:

```
dc_pmp_kmean = dtc_pmp.copy()
dc_pmp_kmean['label']=clusters
dtc_pmp_ind['label']=clusters
```

So that clustering data is obtained as shown in table 5.1.

Table 4 data labeling based on k-mean clustering implementation results:

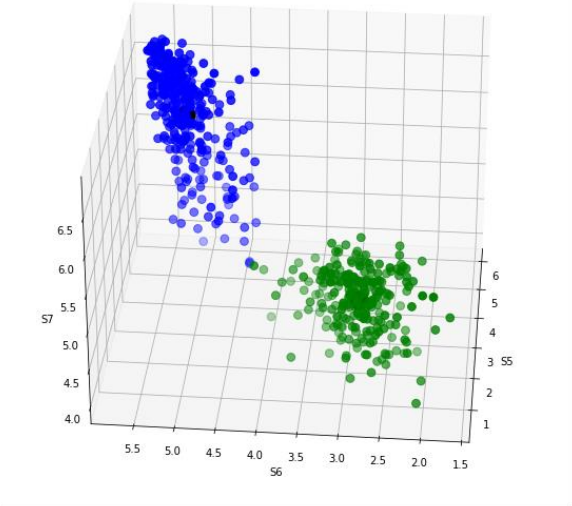
	5.1.	5.2.	5.3.	5.4.	5.5.	6.1.	6.2.	6.3.	7.1.	7.2.	7.3.	7.4.	label
47	6.940000	6.936000	6.9350	1.05	6.925820	4.847000	5.868500	6.369000	6.916000	6.364750	6.813347	6.650000	0
530	2.400000	1.055279	0.0000	0.00	0.000000	4.603846	1.983333	1.526389	6.319667	6.307285	2.064211	6.268000	1
376	3.147891	5.778442	2.5200	0.00	0.000000	3.815556	2.084142	1.146406	6.315061	5.850610	1.947067	4.342154	1
64	6.932500	6.602000	6.9650	1.05	6.859153	4.868000	5.257500	6.409000	6.841000	6.501700	6.976681	6.130000	0
310	3.417340	2.100065	0.0000	0.00	0.000000	4.512507	2.053333	1.875242	6.271293	6.220999	2.139398	7.000000	1
223	6.818750	5.470000	6.9625	2.94	5.266658	4.546000	5.146500	3.859000	6.916000	4.554700	6.206679	3.850000	0
391	3.126757	4.727173	0.0000	0.00	0.000000	3.862030	1.738333	2.581518	5.484769	6.115218	1.871313	4.512667	1
184	6.131250	5.657000	6.0000	2.94	5.834990	4.126000	5.118500	3.859000	5.379000	3.180050	4.756676	4.550000	0
418	2.026857	4.729171	1.2600	0.00	0.000000	4.709669	3.755892	1.410597	6.523103	5.693021	2.097537	5.220923	1
505	2.103542	4.725000	0.0000	0.00	0.000000	4.611492	2.824200	1.913333	6.082067	6.566095	2.085281	6.815200	1

Visually, the distribution and cluster of PMP data at the junior high school level in Bogor Regency are shown in figure 13. This visualization is obtained using the following python algorithm.

```
from mpl_toolkits.mplot3d import Axes3D

fig = plt.figure(figsize=(20,10))
ax = fig.add_subplot(111, projection='3d')
ax.scatter(dc_pmp_kmean["S5"][dc_pmp_kmean.label == 0],
           dc_pmp_kmean["S6"][dc_pmp_kmean.label == 0],
           dc_pmp_kmean["S7"][dc_pmp_kmean.label == 0], c='blue', s=60)
ax.scatter(dc_pmp_kmean["S5"][dc_pmp_kmean.label == 1],
           dc_pmp_kmean["S6"][dc_pmp_kmean.label == 1],
           dc_pmp_kmean["S7"][dc_pmp_kmean.label == 1], c='green', s=60)
ax.scatter(kmeans.cluster_centers_[0, 0],
           kmeans.cluster_centers_[0, 1],
           kmeans.cluster_centers_[0, 2], s = 300, c = 'black', label = 'Centroids')
ax.scatter(kmeans.cluster_centers_[1, 0],
           kmeans.cluster_centers_[1, 1],
           kmeans.cluster_centers_[1, 2], s = 300, c = 'black', label = 'Centroids')
ax.view_init(30, 185)
plt.xlabel("S5")
plt.ylabel("S6")
ax.set_zlabel("S7")
plt.show()
```

Figure 13. Visualization of PMP data clustering at junior high school level using K-mean



The clustering results using k-mean show that the data is divided into 2 groups with labels 1 and 0 where label 1 shows values above education standards and label 0 values below education standards. These two clusters have the characteristics of:

- Kluster label 1 :
  Centeroid coordinates : (5.45253261, 5.27576284, 5.89270723)
  Number of members : 305 (57.12% data)
- Cluster centroid coordinates label 0:
  Centeroid coordinates : (1.53145185, 2.83415636, 4.98306442)
  Number of members : 229 (42.88% data)

Performance evaluation of the k-mean cluster model is then measured using silhouette_score. The Silhouette coefficient is a measure of cluster cohesion and separation. This value will measure:

- How close the data point is to other points in the cluster

- How far the data point is from a point in another cluster

The value of the silhouette coefficient will range between -1 and 1. When the silhouette coefficient value is larger, it indicates that the sample is closer to its cluster than to other clusters.

Based on the results obtained from PMP data, the silhouette score is 0.7814. This value shows the characteristics of standard data that are dominant as data differentiators, so based on schools in Bogor Regency can be categorized into two groups / clusters. The value of the Silhouette coefficient resulting from group formation is 78.14% of each data point formed in the group. In detail, as many as 305 schools entered Cluster 1 and the remaining 229 schools entered Cluster 2. Although the value of these two clusters is not equal because cluster 1 is a group of schools that meet education standards, while cluster 2 is schools that are still below education quality standards.

The implementation of feature engineering and k-means clustering models in predictive analysis of education quality improvement can provide valuable insights in understanding the factors that influence education quality improvement(Hasanah et al., 2019) . Feature engineering involves transforming and adjusting relevant variables in educational datasets to improve the accuracy and predictability of models(Muliono &; Sembiring, 2019) . For example, variables such as the number of students per class, student participation rate, and educator qualifications can be transformed into new, more informative features such as student-teacher ratio, student attendance rate, or educator qualification index.

After performing feature engineering, the k-means clustering model can be used to group schools or regions with similar characteristics based on the features produced(Setyaningtyas et al., 2022) . By clustering, we can identify groups of schools or regions that have similar quality education or potential for similar quality improvement(Putra et al., 2022) . For example, schools in a particular cluster may show similar patterns in terms of student participation rates or academic achievement.

Through predictive analysis, we can apply the k-means clustering model to gain a better understanding of the factors that influence the improvement of education quality(Muliono &; Sembiring, 2019) . The results of this analysis can provide guidance for education policy makers to design appropriate programs to improve the quality of education in certain schools or regions. In addition, by using feature engineering techniques and k-means clustering models, we can identify trends and patterns that may be difficult to detect manually, thus enabling more effective and efficient improvement efforts in improving the quality of education(Priyambadha et al., 2020) .

CONCLUSIONS

Based on data analysis using data science methods to predict the quality of junior high school education in Bogor Regency, there are several things that can be concluded that the standard of education at the junior high school level of Bogor Regency is significantly influenced by the achievement value of 3 education quality standards, namely educator and education staff standards, facilities and infrastructure standards, and management standards. So that the education office of Bogor Regency needs to focus the direction of policies to improve the quality of junior high school education in indicators that affect these 3 standards.

The results of school quality clustering mapping show that junior high schools in Bogor Regency are divided into 2 major groups of education quality level, namely below the value of the 5.07 education quality level as much as 57.12% of junior high schools and the rest above the value of the 5.07 education quality level of 42.88% of junior high schools.

For further development of the system, the model needs to be further developed to the level of sub-indicators so that school quality can be more specifically measured, and the data used as analysis material is expanded and enriched by other educational data from various sectors so that the model developed can be more comprehensive in measuring the achievement of education quality.

REFERENCES

Butcher, B., & Smith, B. J. (2020). *Feature Engineering and Selection: A Practical Approach for Predictive Models: by Max Kuhn and Kjell Johnson. Boca Raton*, FL: Chapman

& Hall/CRC Press, 2019, ISBN: 978-1-13-807922-9.

Cady, Field, 2017, The data science handbook, Wiley, Hoboken, NJ.,

Dapodik Kemdikbud. Data Pokok Pendidikan Kementerian Pendidikan. 2019. Diakses tanggal 30 Desember 2022. dapo.dikdasmen.kemdikbud.go.id/

Direktorat jenderal pendidikan dasar dan menengah, 2016. *dokumen pedoman umum sistem penjaminan mutu pendidikan dasar dan menengah*, Kemdendikbudristekdikti

Fadliyah, Chairati, dan Triani, mike, 2019, *Pengaruh Pengeluaran Pemerintah Sektor Kesehatan, Pendidikan Dan Infrastruktur Terhadap Kesejahteraan Masyarakat Di Indonesia*, Jurnal Kajian ekonomi dan Pembangunan, Vol. No. 3.

Inkiriwang,Rizky Rinaldy., 2020., *Kewajiban Negara Dalam Penyediaan Fasilitas Pendidikan Kepada Masyarakat Menurut Undang-Undang Nomor 20 Tahun 2003 Tentang Sistem Pendidkan Nasional*, Jurnal Lex Privatum unsrat, Vol. 8 No. 2

Ngabidin, M. (2020). Budaya Mutu Wujudkan Sekolah Unggul: Kumpulan Praktik Baik Implementasi Sistem Penjaminan Mutu di Satuan Pendidikan. Penerbit Andi.

Peraturan Pemerintah No. 4 Tahun 2022, *Standar Nasional Pendidikan*, Kemendikbudristekdikti.

Priyambadha, B., Pradana, F., & Bachtiar, F. A. (2020). Penggalian Perilaku Pemain dalam Penentuan Tipe Permainan pada E-Learning Pemrograman Berbasis Gamification. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *7*(4), 765–772.

Putra, P. H., Hasibuan, A., & Marpaung, E. A. (2022). Perancangan Aplikasi Penentuan Minat Dan Bakat Anak Menggunakan Metode K-Mean. *Prosiding Seminar Nasional Sosial, Humaniora, Dan Teknologi*, 39–44.

Ray, S. (2019, February). A quick review of machine learning algorithms. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon) (pp. 35-39). IEEE.

Setyaningtyas, S., Nugroho, B. I., & Arif, Z. (2022). TINJAUAN PUSTAKA SISTEMATIS: PENERAPAN DATA MINING TEKNIK CLUSTERING ALGORITMA K-MEANS. *Jurnal Teknoif Teknik Informatika Institut Teknologi Padang*, *10*(2), 52–61.

Rahminawati, N. (2021). Sistem Penjaminan Mutu Internal Dalam Peningkatan Kualitas Sekolah Dasar. JAMP: Jurnal Administrasi dan Manajemen Pendidikan, 4(3), 212-219.

United Nations Development Programme., 2022, *Human development report 2021/2022 : uncertain times, unsettled lives : shaping our future in a transforming world.*

Undang-undang RI No. 20 tahun 2023 Sistem Pendidikan nasional. Kemendikbudristekdikti.

Utomo, D. P., & Mesran, M. (2020). Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. Jurnal Media InformatikaBudidarma, 4(2), 437-444.

Yuan, C., & Yang, H. (2019). Research on K-value selection method of K-means clustering algorithm. J, 2(2), 226-235.