Jurnal Ekonomi dan Pembangunan Vol. 29 No. 2 Tahun 2021
Journal homepage: https://jurnalekonomi.lipi.go.id/JEP
ISSN 0854-526x (print) | e-ISSN 2503-0272 (online) DOI: 10.14203/JEP.29.2.2021.183-198

The opportunity of data mining for macroeconomic data analysis: A case analysis of East Java Province

Gunawan

University of Surabaya, gunawan@staff.ubaya.ac.id

Abstract

The conventional data analysis in economics is based on a model derived from economic theories. In contrast, data mining is a data-driven analysis to extract data and find a pattern describing the empirical interaction between variables. The emerging area of data mining offers an opportunity for extracting information from macroeconomic data. However, it is still a challenge for economic researchers and policymakers to embrace data mining because it is closely related to the information technology discipline. This study responds to the limited use of data mining in the economic area by analyzing macroeconomic indicators published by the Indonesian Central Bureau of Statistics. The primary purpose of this study is to offer a case for using the data mining approach for macroeconomic indicators. The specific objectives were (1) to introduce the Cross-Industry Standard Process for Data Mining (CRISP-DM) as a process framework and Knime Analytics Platform as a data mining software for macroeconomic data analysis; and (2) to characterize East Java regencies/municipalities based on their macroeconomic indicators and region profiles. This study was categorized as secondary and quantitative research. The unit of analysis was the regency/municipality. Five macroeconomic indicators: Human Development Index (HDI), Gross Regional Domestic Products (GRDP), poverty rate, Gini Ratio, and open unemployment rate, were selected as the variables. Four region profiles: area, population, population density, and the number of villages were included in the analysis. The clustering model was implemented through Knime's workflow. The result of clustering grouped 38 regions into three. Its applicability and simplicity indicated the appropriateness of the CRISP-DM process framework for analyzing the structured official data. Furthermore, the predictive model, applied to past years' datasets, revealed the regions that experienced improvement and shifted their membership between clusters over three years. Moreover, the inclusion of region profiles has provided a better understanding of underlying factors explaining the association between macroeconomic indicators. This study suggests that the East Java Government considers different facilitation-focused programs based on the characteristics of three clusters for better budget efficiency. This research adds to the literature on economic development, particularly by introducing data mining, the CRISP-DM method, and Knime software to analyze macroeconomic indicators of regency/municipality.

Keywords : macroeconomics, data mining, CRISP-DM, cluster, East Java

JEL Classification: G33, C23, G41

Abstrak

Analisis data konvensional di bidang ekonomi didasarkan pada model yang diturunkan dari teori ekonomi. Sebaliknya, data mining adalah analisis berbasis data untuk mengekstrak data dan menemukan pola yang menggambarkan interaksi empiris antar variabel. Bidang data mining menawarkan peluang untuk mengekstraksi informasi dari data ekonomi makro. Namun, masih menjadi tantangan bagi para peneliti ekonomi dan pembuat kebijakan untuk menerapkan data mining karena terkait erat dengan disiplin teknologi informasi. Studi ini menjawab keterbatasan pemanfaatan data mining di bidang ekonomi dengan menganalisis indikator ekonomi makro yang diterbitkan oleh Badan Pusat Statistik (BPS). Tujuan utama dari penelitian ini adalah untuk memaparkan sebuah kasus dalam menggunakan pendekatan data mining pada indikator ekonomi makro. Tujuan khusus adalah (1) untuk memperkenalkan kesesuaian Cross-Industry Standard Process for Data Mining (CRISP-DM) sebagai kerangka proses dan Knime Analytics Platform sebagai perangkat lunak data mining untuk menganalisis data ekonomi makro; dan (2) mengkarakterisasi kabupaten/kota di Jawa Timur berdasarkan indikator ekonomi makro dan profil wilayahnya. Penelitian ini dikategorikan sebagai penelitian sekunder dan penelitian kuantitatif. Unit analisisnya adalah kabupaten/kota. Lima indikator ekonomi makro: Indeks Pembangunan Manusia (IPM), Produk Domestik Regional Bruto (PDRB), tingkat kemiskinan, Rasio Gini, dan tingkat pengangguran terbuka, dipilih sebagai variabel. Empat profil wilayah: luas wilayah, jumlah penduduk, kepadatan penduduk, dan jumlah desa dimasukkan dalam analisis. Model klasterisasi diimplementasikan melalui Knime workflow. Hasil klasterisasi mengelompokkan 38 daerah menjadi tiga. Penerapan dan kesederhanaannya menunjukkan kesesuaian kerangka

Dikirim : 29 Januari 2021
Direvisi akhir : 3 Desember 2021
Diterima : 27 Desember 2021

proses CRISP-DM untuk menganalisis data resmi ekonomi makro. Selanjutnya, model prediktif, yang diterapkan pada data tahun-tahun sebelumnya, mengungkapkan kabupaten/kota yang mengalami perbaikan dan pergeseran keanggotaannya antar klaster dalam periode tiga tahun. Selain itu, masuknya profil wilayah telah memberikan pemahaman yang lebih baik tentang faktor-faktor untuk menjelaskan hubungan antar indikator ekonomi makro. Studi ini menyarankan agar Pemerintah Jawa Timur mempertimbangkan program yang berfokus pada fasilitasi yang berbeda berdasarkan karakteristik tiga klaster untuk efisiensi anggaran yang lebih baik. Penelitian ini menambah literatur tentang perkembangan ekonomi, khususnya pengenalan data mining, metode CRISP-DM, dan software Knime untuk menganalisis indikator ekonomi makro kabupaten/kota.

Kata kunci: *ekonomi makro, data mining, CRISP-DM, klaster, Jawa Timur*

Klasifikasi JEL: *G33, C23, G41*

INTRODUCTION

Economic researchers and policymakers require information extracted from data to develop a theory, formulate recommendations, or make decisions. Therefore, the capability to analyze data was considered a skill to be mastered. In the last twenty years, algorithms, software, and technology development has made data analysis a new discipline named data science. Degree programs and related professional courses have been widely offered and named data science, data mining, data analytics, or big data. Data science is defined from a simple one as "the study of data" to a broad discipline (Cao, 2017). In general, data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract meaning and insights from structured or unstructured data to support a decision. Data science is related to data mining, which is defined as a process of extracting and discovering hidden information from data.

Data science has been used effectively in many areas (e.g., healthcare, logistics, marketing), but it has little been investigated in economics and finance (Barbaglia et al., 2021). Policymakers and business practitioners attempt to do forecasting or nowcasting (prediction of the present or the very near future) of economic indicators to make better decisions. Economic indicators contain data measured with high or low frequency. High-frequency data (e.g., stock price, social media content) generally produces big volume and unstructured data. The increasing volume of data occurs in many areas such as retail, banking, health care, manufacturing, and public service. On the other hand, the low-frequency data (e.g., yearly poverty rate, GDP) are primarily published as official statistics by the national statistics agency.

The Indonesian Central Bureau of Statistics (*Badan Pusat Statistik* - BPS) accumulates data from population censuses, statistical surveys, business and trade statistics, municipal statistics, and other government agency records. The data are called "official data" and processed to generate "official statistics," such as data about demographic, social, and economic figures of regions (Brito & Malerba, 2003). Provinces and regencies/municipalities rely on official statistics for formulating their regional development plan.

The official statistics become the backbone for analyzing economic conditions and estimating the effects of different economic policies. Provinces and regencies/municipalities rely on official statistics for formulating their regional development plan. Traditional economic analysis is based on an economic model describing a relationship between variables. On the other hand, data mining offers an opportunity to reveal the interaction between variables based on empirical data rather than pre-defined relationships. Data mining could work in a large and small volumes of numerical and non-numerical data, while traditional econometric methods work well in small and numerical data sets (Harding & Hersh, 2018). Moreover, adopting the new analytical approach in economics seemed slow (Taylor et al., 2014).

This study responds to the limited use of data mining for economic data (Barbaglia et al., 2021) by analyzing macroeconomic indicators published by BPS. The reason is that these structured data are well reported, covering comprehensive economic and social indicators, covering data for provinces and regencies/municipalities, published regularly and publicly available. Social and economic indicators have become a measure of

a society, region, or country is compared to the others. Macroeconomic indicators are considered critical information for policymakers (Mügge, 2016). Gross Domestic Product (GDP) growth, poverty rate, and unemployment rate are major economic indicators to determine economic development in a country or its regions. Data mining offers an opportunity to process official statistics, obtain relationships between data, and find patterns (Hassani et al., 2014). Implementing data mining techniques to official data could support public policymaking. The implementation should address the correct research method (Brito & Malerba, 2003), data quality, confidentiality, and timeliness (Hassani et al., 2010).

BPS publishes social and economic data about regency/municipality, provincial, or central government. The data are characterized as numerical, small volume, and structured data. The investigation of the Garuda Portal, the Indonesian journal database, indicated numerous studies using official statistics to analyze economic and social development among regencies/municipalities (e.g., Purnamasari et al., 2014; Wahyudi et al., 2016). However, those studies did not address data mining process methods, techniques, or software. While the traditional statistical data analysis technique could perform well on those kinds of data, the data mining approach offers more advantages.

This study emerged from a research question: "How could data mining be effectively implemented to macroeconomic indicators?" The primary purpose of this study is to present a case in using a data mining approach for extracting information from macroeconomic data. The specific objectives are (1) to introduce the appropriateness of the Cross-Industry Standard Process for Data Mining (CRISP-DM) as a process framework and Knime as a data mining software for analyzing macroeconomic data; and (2) to characterize East Java regencies/municipalities based on macroeconomic indicators and their region profiles. East Java province was selected first because its macroeconomic indicators among 34 provinces are about in the middle. For example, the Human Development Index is in on top 15, and the growth rate of gross regional domestic product is in the top 14 for data 2019. Therefore, it is intended that the case province reflect the other provinces. The second reason to choose East Java was its number of regencies/municipalities (38 regions) as the biggest among 34 Indonesian provinces (dataindonesia.id). From the statistical analysis perspective, the bigger the sample, the better the result of the analysis. The data source was official data from the East Java Province BPS (jatim.bps.go.id), and the unit of analysis was the regency/municipality.

LITERATURE REVIEW

Data mining concept and its role in macroeconomic study

Data refer to different types of information that are usually formatted and stored appropriately with a specific purpose. From the information technology perspective, data are numbers, text, or multimedia that a computer can process. Data may take the form of recorded transactional data such as customer purchases and ticket booking in business practices. Data will create value if it can be transformed into information to support a decision. Some people view data mining as one step of the knowledge discovery process consisting of seven steps: data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge presentation (Han et al., 2012). On the other hand, many people define data mining broadly, covering the entire knowledge discovery process from industry and academic perspectives. Thus, generally, data mining is defined as a process of discovering interesting patterns and knowledge from amounts of data (Han et al., 2012).

Economic discipline commonly deals with official data collected for some official purposes from censuses and formally conducted surveys. The economic theory is applied to specific problems through economic models, such as the aggregate demand–aggregate supply model describing price level and output through aggregate demand and aggregate supply. Empirical measurement of the economic model is performed through statistical inference. As a form of statistical inference, hypothesis testing uses sample data to test an assumption regarding a population parameter or distribution. This analysis is theory-driven.

On the other hand, data mining is an exploratory data analysis. A sample of data is analyzed and modeled to find a pattern describing an interaction between variables. As a data-driven analysis, data mining is considered hypothesis seeking, which leads to the idea of new theories (Feelders, 2002). The use of data mining for macroeconomic data creates an opportunity to understand the empirical interaction between variables, which might lead to a new hypothesis. Subsequently, the new hypothesis could lead to the idea for a new theory or model.

Data analysis in statistics and econometrics can be divided into four types: summarization, prediction, estimation, and hypothesis testing (Varian, 2014). Data mining concerns summarization, as its purpose is to extract interesting patterns in the data. Data mining also adopts machine learning techniques, and its capability also covers prediction (predictive modeling) and estimation. On the other hand, data mining is not aimed explicitly at hypothesis testing. However, some basic statistical tests (e.g., ANOVA test) are available in data mining software, such as the Knime Analytics Platform. Statistical data analysis, named inferential statistics, concerns hypothesis testing to deduce the characteristics of an entire population from a small but representative sample. This conventional approach is often criticized regarding, for example, the criteria of sample size, the definition of significant level, and the concept of confidence interval (Saporta, 2018).

Economic research relates highly to data analysis. Empirical economic research inductively develops generalizations from data, uses data to test competing models, evaluates policies, and forecasts the effects of new policies or modifications of existing policies (Heckman, 2001). As the volume and variety of data related to economic activities increases, economic researchers face data analysis challenges. The rising big data has reshaped economic analysis. A prior study analyzed the published articles on' big data research in economics' 2015-2019 (López-Robles et al., 2019). Based on the 1,034 articles found, it was noted that the number of articles showed evident growth.

The economic policy is made by looking at, for example, economic forecasting, which predicts the future economic condition. The classical steps of economic forecasting, which are having a theory, doing simulation, doing the calibration, and making predictions, would be highly reinforced by the emergence of big data (Taylor et al., 2014). The advancement in data analytic tools has contributed to delivering deep analysis. The predictive modeling solution of data mining works by analyzing past data and generating a model to help predict future outcomes.

Macroeconomic indicators

A macroeconomic indicator is a metric used to assess, measure, and evaluate a nation's economic performance. As the indicators can cover many aspects of economic development, there is no consensus to limit only a few indicators. Some of the considerable macroeconomic indicators are widely used, such as gross domestic product (GDP), inflation rate, unemployment rate, and poverty rate. The economic discipline categorizes the indicators into three based on the cycle of economic change: a leading indicator (e.g., stock market price), a coincident indicator (e.g., GDP), and a lagging indicator (e.g., unemployment level). Analyzing these indicators is critical to understanding current and predicting future economic activities. The predictive power of macroeconomic indicators, for example, could be used to predict the sign of the stock return of oil and gas industry stock (Liu & Kemp, 2019).

Macroeconomic indicators among countries are often published to compare economic performance among them, such as reported on the World Bank, OECD, or ASEAN websites. In Indonesia, the indicators are reported at the national, province, and regency/municipality levels. This study adopted five socio-economic indicators acquired from the Indonesian National Development Planning Agency (*Badan Perencanaan Pembangunan Nasional*) at its site sdgs.bappenas.go.id. Those indicators, named Macro Development Goal Indicators, comprise (1) Human Development Index, (2) Gini Ratio, (3) Poverty rate, (4) Open unemployment rate, and

(5) Economic growth. The reason for selecting these indicators was attributed to the Bappenas itself as the authorized agency that plans the national development. Each indicator is described as follows.

First, the Human Development Index (HDI) is defined by the UNDP as "a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable, and having a decent standard of living" (hdr.undp.org). It is a composite indicator of a human-oriented country's development achievements. HDI consists of 3 dimensions with four indicators, namely: the health dimension (life expectancy), the education dimension (the expected length of schooling, the average length of schooling), and the standard of living dimension (per capita expenditure or income) (BPS-Jatim, 2019). Second, the Gini ratio (index or coefficient) is a measure of income distribution calculated based on income class, and it indicates the level of income inequality of a region. The Gini ratio ranges from 0 (perfect equality) to 1 (perfect inequality). Third, the poverty rate is defined as the percentage of the population who live below the poverty line.

Furthermore, the fourth indicator, the open unemployment rate, is the total unemployed percentage against the entire labor force (sirusa.bps.go.id). A high unemployment rate indicates that a large labor force is not absorbed in the labor market. Finally, the fifth indicator, economic growth, was represented by the growth rate of gross regional domestic product (GRDP) at constant prices. GRDP growth shows the change in the production of goods and services in an economic area within a specific time interval, calculated based on the price in a particular year, determined as a reference.

Prior studies investigated the relationship between those five major indicators and some other economic indicators, as shown in Table 1. The first column presents the model with dependent (DV) and independent (IV) variables, while the second column indicates the findings. The third column shows the study objects, the fourth is the analysis method used, and the last is the reference.

Table 1 shows that some studies used HDI as a dependent variable, while others used poverty level. There was no consensus on which indicator became the dependent variable and which one became the independent variable. Linear regression was the most popular analysis method as it fits quantitative data of economic indicators and is simple to understand. However, as linear regression focused on the relationship between variables, it provided less information about the characteristic of the regions (provinces or regencies/municipalities). The use of data mining appears in the last two studies in Table 1, but both focused on HDI elements and did not include the region profiles. In addition, both studies did not use the model obtained from other datasets (model deployment). The relationship between variables shown in Table 1 indicates the propensity that (1) higher HDI is associated with higher GRDP, higher population, higher Gini ratio, and lower poverty rate; and (2) higher poverty rate is associated with higher unemployment.

Empirical economic research could inductively develop generalizations from the data, test the competing models, evaluate policies, and forecast the impact of new policies (Heckman, 2001). The research focusing on regions (province or regency) provides analysis that leads to consideration for policymaking. However, as presented in Table 1, some studies among areas in East Java were limited regarding the variables included or the analysis tool used. In addition, the exclusion of region profiles in the analysis makes the recommendation drawn from the result to regions limited.

RESEARCH METHOD

Methods

This study can be categorized as secondary research because of uses existing (secondary) data. Secondary research also requires a method with systematic steps and clear stages, namely: (1) research question formulation, (2) data identification, and (3) data evaluation (Johnston, 2014). Consistent with these three, this research adopted the Cross-Industry Standard Process for

Table 1. Some Related Studies on Macroeconomic Indicators

Model	Relationship found	Country – region	Analysis method	Reference
DV: Poverty IV: HDI, GRDP, minimum wage, investment	Poverty(+): HDI(-)	Indonesia – East Java	Linear regression	(Qurrata & Ramadhani, 2021)
DV: HDI IV: GRDP rate, population, unemployment	HDI(+): population (+)	Indonesia – East Java	Linear regression	(Wahyuningrum & Etty Soesilowati, 2021)
DV: HDI IV: population, GDP per capita, inflation, unemployment	HDI(+): population (-), GDP per capita (+)	10 ASEAN countries	Linear regression	(Arisman, 2018)
DV: HDI IV: GDP, Wealth Gini, Income Gini, poverty rate	HDI(+): poverty rate(+/-),Wealth Gini(+), Income Gini (+)	98 countries	Discriminant Analysis	(Georgescu et al., 2020)
DV: HDI IV: Gini, non-food expenditure, the dependency ratio	HDI(+): non-food expenditure(+)	Indonesia – East Java	Linear regression	(Imaningsih et al., 2020)
DV: Poverty IV: GRDP, Gini, unemployment, HDI	Poverty(+): unemployment(-)	Indonesia – 2 cities	Linear regression	(Sinaga, 2020)
Competitiveness index, GDP, HDI	Cluster, Mixed relationships	Latin America and Caribbean countries	Cluster analysis, linear regression	(Reyes & Useche, 2019)
HDI and its elements	Not available	Indonesia – East Java	Descriptive	(Hartanto et al., 2019)
HDI and its elements	Not available	Indonesia – 34 provinces	Data mining - clustering	(Rahmat et al., 2021)
HDI and its elements	Not available	Indonesia – 34 provinces	Data mining - clustering	(Saputra et al., 2020)

Note: DV = Dependent Variable, IV = Independent Variable, (+) or (-) indicates positive or negative relationship to dependent variable

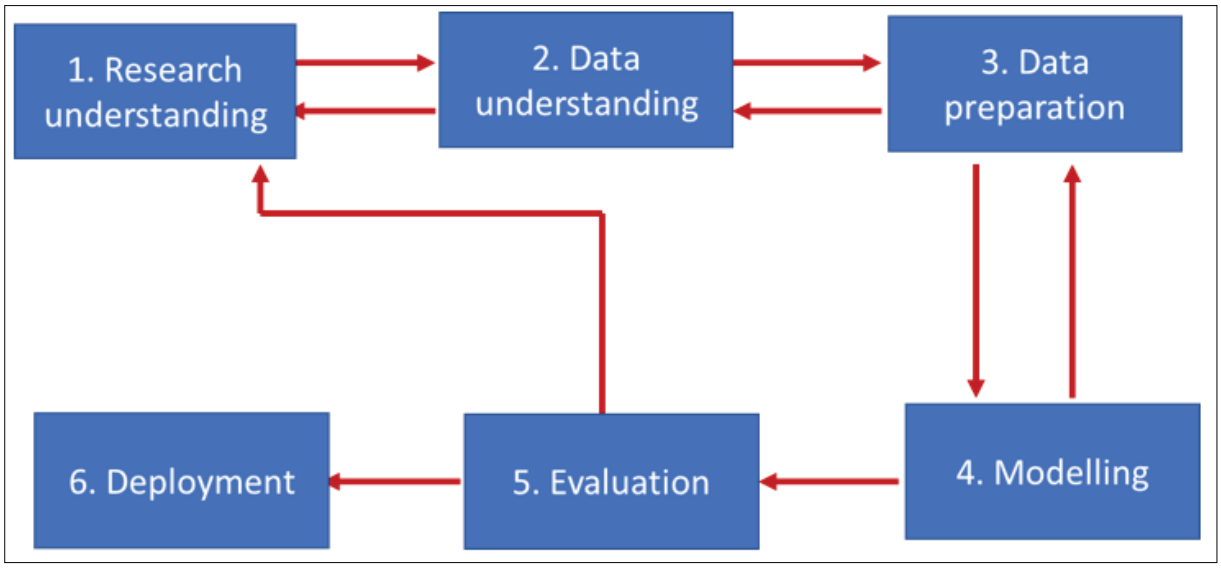
Data Mining (CRISP-DM) framework with six phases, as shown in Figure 1. Phases 1 to 3 are presented in this Research Method section, while the rest are in the Results and Discussion section.

Characteristics of data

In Phase 1 of CRISP-DM, the research understanding denotes research goals and data mining goals. Research goals have been presented in the Introduction section. Data mining goals were formulated as follows: (1) to implement k-means clustering methods for a dataset of macroeconomic indicators, and (2) to implement a predictive clustering model for other datasets.

As presented earlier, the data understanding phase (Phase 2) refers to the five Macro Development Goal indicators determined by Bappenas. These five socio-economic indicators are listed in Table 2. Government pursues the increase of HDI and GRDP through economic development. It means that higher HDI and higher GRDP are expected. Contrary, government attempts to reduce the Gini ratio (income inequality), the poverty level, and the open unemployment level. Table 2 indicates the expected trend.

Most of the official statistics published by BPS have a numerical type, with the ratio type (e.g., population) and interval type (e.g., human development index). The sample of this study was all 38 regions, consisting of 29 regencies and nine municipalities, in East Java Province. Official data was collected from jatim.bps.go.id. In addition to data available on the site, this study also used the report "East Java Province in the number Year 2020" (BPS-Jatim, 2020). This study sampled the data of five macroeconomics indicators for three years: 2017, 2018, and 2019. The reason was related to the Covid-19 pandemic starting in 2020. Three-year datasets before



Source (Colin Shearer, 2000)

Figure 1. CRISP-DM Process Framework

Table 2. Main Macroeconomic Indicators

Indicator	Definition	Expected trend
Human Development Index (HDI)	An index that measures human development from three fundamental aspects: longevity and healthy life; knowledge; and a decent standard of living.	Higher
Gini Ratio (GR)	A measure of income distribution is calculated based on income class.	Lower
Poverty rate (PR)	The percentage of the population has an average monthly per capita expenditure below the poverty line.	Lower
Open unemployment rate (UR)	The percentage of the population aged 15 years and over who are unemployed to the total labor force.	Lower
Gross Regional Domestic Product (GRDP) growth rate	The GRDP growth rate shows the growth in the production of goods and services in an economic area within a certain period.	Higher

the pandemic are expected to be comparable to one another.

The primary attribute of any region is its area and population. A regency/municipality consists of many villages, the smallest and lowest government structure directly related to citizens. The Indonesian government provides village funds to finance government administration, development implementation, community development, and village community empowerment. Therefore, four region attributes were selected, including a geographic measure (area), population, population density, and government administration (the number of villages). The indicator of income-per-capita was excluded because it has been included in HDI as a standard of living measure. Table 3 presents all nine variables.

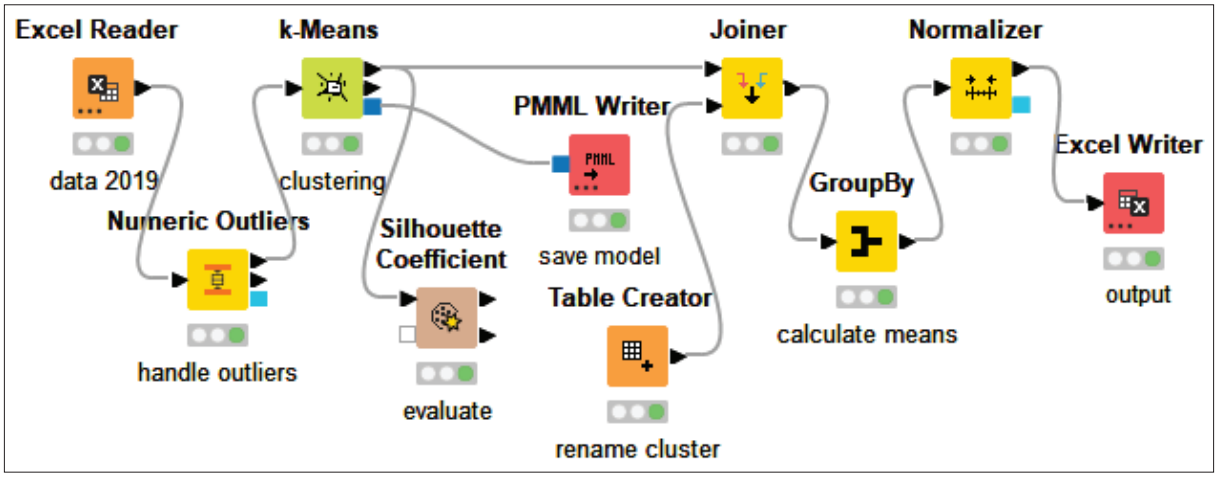
The raw data were cleaned and formatted (Phase 3: Data preparation) to be suitable for modeling analysis. Data were presented in one sheet where the list of regions was placed in the first column and variable names in the first row. Each data cell was checked to ensure all had the correct number format. Finally, the data file was saved in XLS or CSV format to input further modeling.

Data mining workflow

Many data mining software packages are available, either open-source or proprietary software. This study employed Knime, an open-source software. The popularity of Knime appeared from its advantage as a data mining application with a visual programming interface in which users do not require coding capability

Table 3. Macroeconomic Indicators and Region Profiles

Measure	Variable	Data period
Macroeconomic indicator	Human Development Index (HDI)	2017-2019
	Gini Ratio (GR)	2017-2019
	Poverty rate (PR)	2017-2019
	Open unemployment rate (UR)	2017-2019
	Gross Regional Domestic Product (GRDP) growth rate	2017-2019
Region profile	Area (km^2)	2019
	Percentage of the total population	2019
	Population density (people/km^2)	2019
	Number of villages	2019



Source: Author

Figure 2. Knime Workflow for k-Means Clustering

(Oliveira et al., 2019) so that non-programmers, such as economic researchers, could use the tools. With visual programming, the analysis process is represented by small icons arranged to indicate the flow and logic of the data mining process. Figure 2 shows the Knime workflow for the k-means clustering model. The workflow consists of nodes, from reading an input file on the left to writing the output file on the right. The 'PMML writer' node contains the k-means clustering model. PMML stands for Predictive Model Markup Language, a format for describing a data mining model containing input, transformations used to organize data, and the parameters that create the model itself.

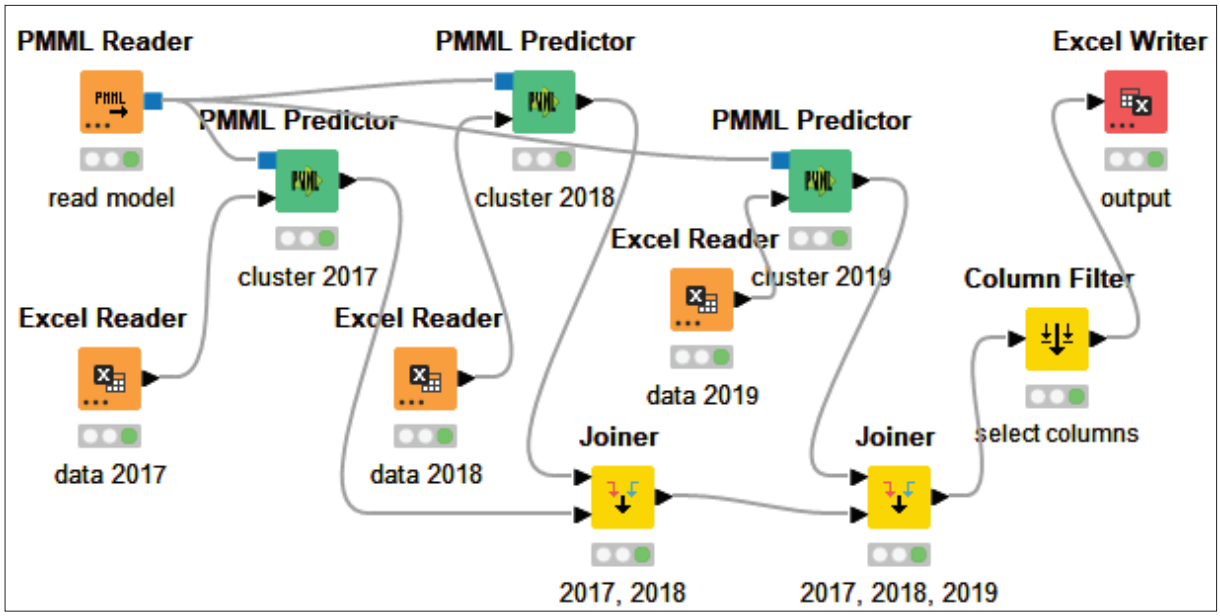
Furthermore, Figure 3 presents a process of calling the model with the 'PMML Reader' node, using it for data 2017-2018, and combining the result with data 2019 for easy comparison of cluster membership. The analysis revealed whether an individual regency/municipality in 2017 and 2018 belonged to the same cluster as in 2019 or not.

RESULTS AND DISCUSSION

Result of Clustering Model

Modeling was performed (Phase 4: Modeling) by selecting the appropriate approach based on the types and characteristics of data. If the data contain labeled output (e.g., performing and non-performing loans), classification and regression are appropriate. For numerical data types without labeled output, the appropriate approach is clustering. This study adopted the k-means algorithm. Knime has a node named "Numeric Outliers" to handle outliers, whether by removing the row or replacing the values. This study decides to return to the closest values. The number of clusters was determined considering the data characteristics and the practicality of the results. With 38 regions, the plausible alternative number of clusters (k) could be 2, 3, 4, or 5.

The evaluation (Phase 5: Evaluation) of the k-means modeling was performed through the 'Silhouette Coefficient' node. The node was



Source: Author

Figure 3. Knime Workflow for Predictive Model

to assess whether the number of clusters was optimal. The Silhouette coefficient has a score between 1 and -1; if the coefficient is close to 1, the objects are clustered well; if it is close to -1, the objects are not clustered well and are similar to objects in other clusters. Modeling was performed by assigning k for 2, 3, 4, and 5. The mean of Silhouette coefficients for each k was presented in Table 4. For k=2, the cluster size is 16 and 22 regions, in which the mean Silhouette coefficient of 16 regions is 0.43, and for 22 regions is 0.42. The overall mean of the Silhouette coefficient is 0.42. The highest overall Silhouette coefficient mean is at k=3, and the Silhouette coefficient for each cluster is adequate, with a value between 0.36 and 0.57. Therefore, clustering was decided for k=3. The k-means clustering produced three clusters, each member presented in table 5.

An investigation was carried out on each cluster's central point to obtain information on the differences between groups. The mean value of each variable was calculated (with Groupby node) and then normalized (with Normalizer node) to a range of 0 to 1 for easier comparison and interpretation. Table 6 displays the results. The value 0 indicates that the mean of a variable was the lowest among the three clusters, and the value 1 represents the highest value. Among five macroeconomic indicators, all moderate mean values stayed in cluster B.

Graph presentation provided good visualization of cluster characteristics. Figure 4 shows the difference in HDI distribution against poverty rate across three clusters. For example, regions in cluster A are likely to have a higher poverty rate but lower HDI. Furthermore, Figure 5 presents the scatter plot of HDI against population density. For example, regions in cluster C tend to have higher HDI and higher population density.

One of the valuable features of data mining is predictive modeling. The cluster model previously developed was applied to a new dataset (Phase 6: Deployment). In this study, a clustering process using dataset 2019 was performed. First, the clustering model from dataset 2019 was saved as a PMML file (with PMML Writer node, see Figure 3). PMML is an XML-based language and a standard to represent predictive and descriptive models and data pre and post-processing (Guazzelli et al., 2009). Then this file was deployed to datasets 2018 and 2017. The implementation of the past dataset was not aimed to predict the future but to investigate the changing cluster membership among regions over three years.

The analysis comparing the cluster members across 2017-2018-2019 found that six regencies experienced shifting in the cluster membership, as shown in Table 7. Those regions shifted toward better conditions from clusters A to B for

Table 4. Model Evaluation Through Silhouette Coefficient

k	Cluster size with mean SC	Overall SC Mean
2	16[0.43]; 22[0.42]	0.42
3	8[0.36]; 9[0.57]; 21[0.42]	0.45
4	6[0.18]; 7[-0.07]; 9[0.52]; 16[0.48]	0.34
5	4[0.41]; 7[0.24]; 8[0.23]; 8[0.47]; 11[0.24]	0.31

Note: SC = Silhouette coefficient

Table 5. Cluster Membership–Dataset 2019

Cluster	Region
Cluster A (8 regions)	Kab. Bangkalan, Kab. Bondowoso, Kab. Pacitan, Kab. Pamekasan, Kab. Probolinggo, Kab. Sampang, Kab. Sumenep, Kab. Tuban
Cluster B (21 regions)	Kab. Banyuwangi, Kab. Blitar, Kab. Bojonegoro, Kab. Gresik, Kab. Jember, Kab. Jombang, Kab. Kediri, Kab. Lamongan, Kab. Lumajang, Kab. Madiun, Kab. Magetan, Kab. Malang, Kab. Mojokerto, Kab. Nganjuk, Kab. Ngawi, Kab. Pasuruan, Kab. Ponorogo, Kab. Situbondo, Kab. Trenggalek, Kab. Tulungagung, Kota Probolinggo
Cluster C (9 regions)	Kab. Sidoarjo, Kota Batu, Kota Blitar, Kota Kediri, Kota Madiun, Kota Malang, Kota Mojokerto, Kota Pasuruan, Kota Surabaya

Note: The order of regions in each cluster is in alphabetical order

Table 6. Normalized Mean Values Among Clusters

Indicators	Cluster			Profile	Cluster		
	A	B	C		A	B	C
HDI	0	0.39	1	Area	0.93	1	0
GRDP	0	0.60	1	Percent Population	0.58	1	0
Poverty rate	1	0.42	0	Population density	0	0.10	1
Unemployment rate	0	0.56	1	No. Villages	0.91	1	0
Gini ratio	0	0.4	1				

Table 7. Regions Experienced Cluster Shifting

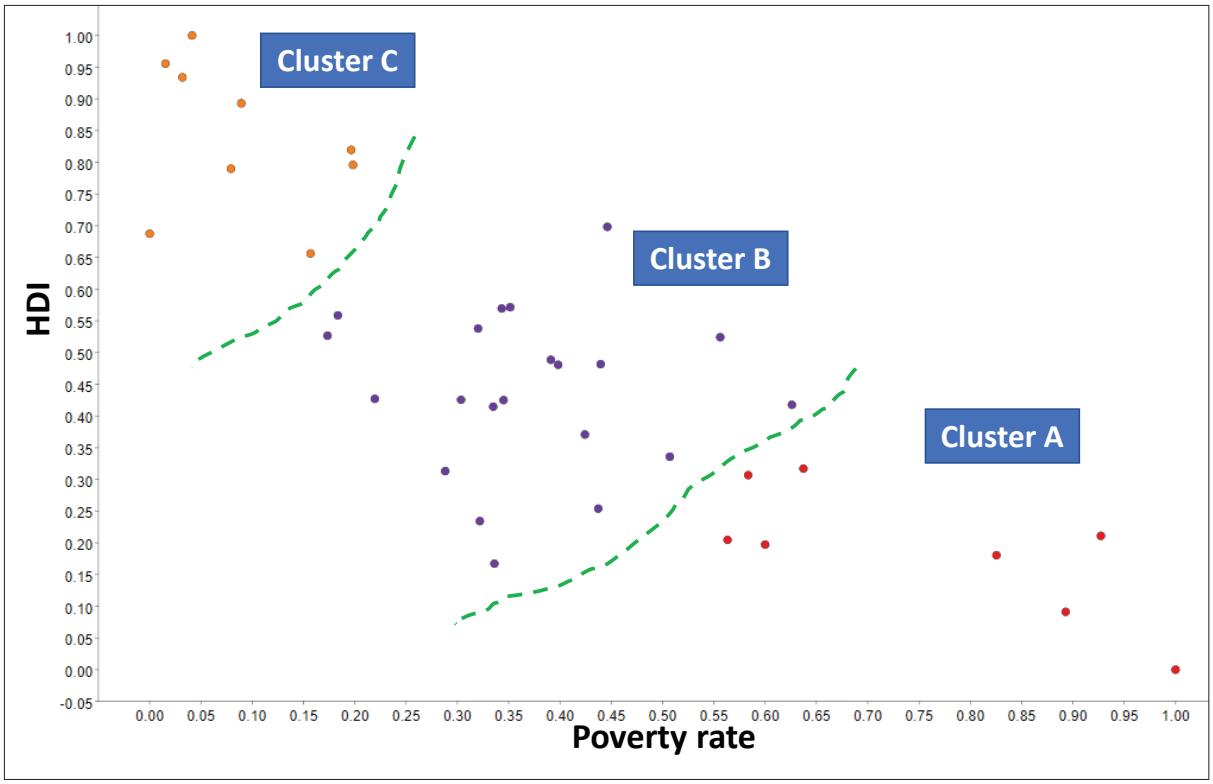
Region	2017	2018	2019	Region	2017	2018	2019
Kab. Situbondo	A	A	B	Kab. Lumajang	A	B	B
Kab. Ngawi	A	A	B	Kab. Jember	A	B	B
Kab. Bojonegoro	A	A	B	Kota Pasuruan	B	C	C

five regencies and B to C for one municipality. However, this analysis could not explain the factors causing the change, and further investigation is required.

Macroeconomic indicator and regional profiles

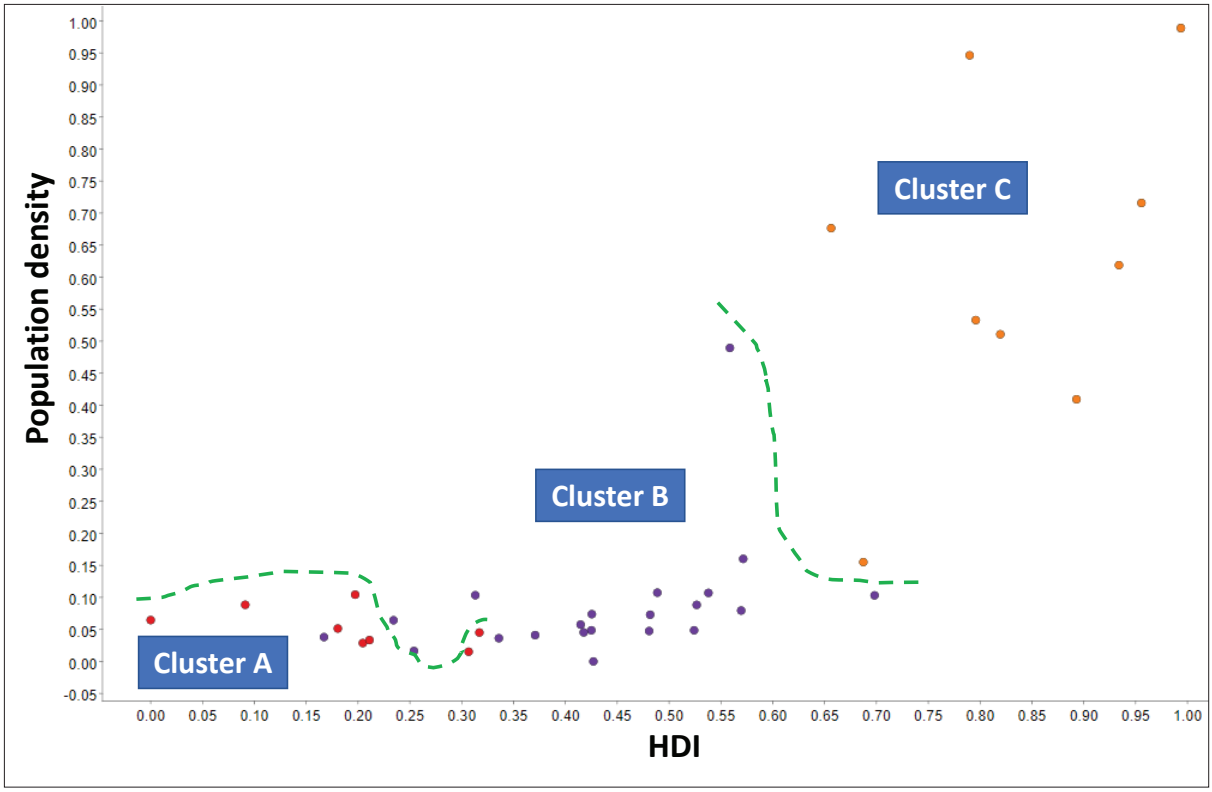
Linear regression is helpful to discover the relationship between macroeconomic indicators. However, it focuses on the variables rather than the regions. Instead, the clustering method focuses on classifying regions based on macroeconomic indicators. The region profiles could explain the different levels of macroeconomic indicators between groups. Based on the normalized mean values of each cluster shown in Table 6. The cluster characteristics addressing five macroeconomic indicators and region profiles are portrayed in Figure 6. The sign "up" and "down" refer to each variable's high or low value.

Figure 6 tells that a region with high HDI and GRDP has low poverty. However, that region has high-income inequality (Gini ratio) and a high unemployment rate, while the economic development expects low-income inequality and unemployment rate. The result could be interpreted that a region produces a higher level of outputs (GRDP), which leads to higher expenditure on health and education (improve HDI), and poverty mitigation. Then, this condition will increase the economic productivity leading to higher GRDP growth. An empirical study



Source: Author

Figure 4. Poverty Rate Versus HDI



Source: Author

Figure 5. HDI Versus Population Density

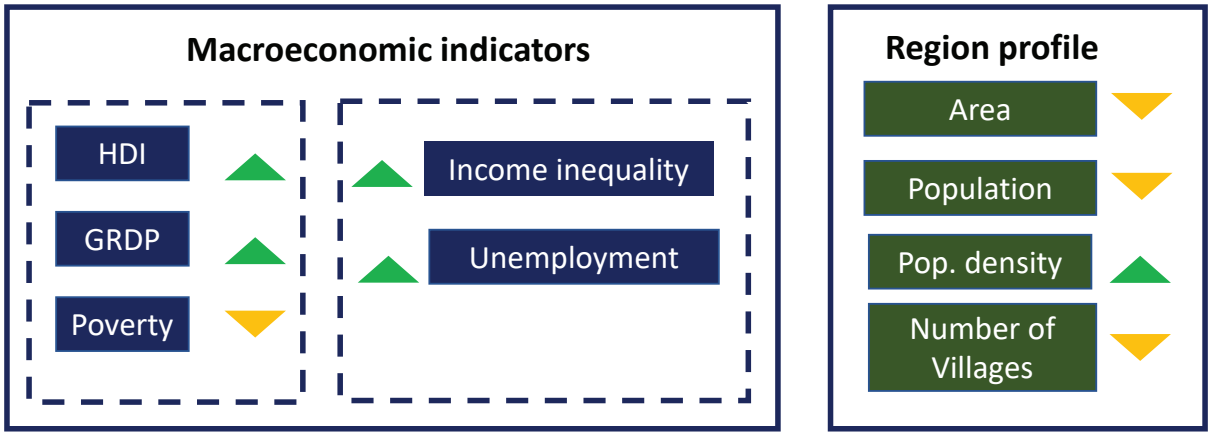
in East Java province supported the result that higher poverty is associated with lower HDI (Qurrata & Ramadhani, 2021). The result also supported a study in Brazil, indicating the positive contribution of HDI in reducing the poverty rate (Costa et al., 2018). The positive trend of HDI and GDP was supported by a study among ASEAN countries (Arisman, 2018) and Eastern European countries (Hudáková, 2017). However, the positive relationship between HDI and GDP is inconclusive; for example, a study in the Indian context showed the shallow impact of GDP on HDI (Khodabakhshi, 2011).

Furthermore, the positive direction between income inequality (Gini ratio) and the unemployment rate tells that the more unemployed people, the higher the income inequality. A prior study reported a positive correlation between income inequality and the unemployment rate whenever the unemployment was less than 15% (Cysne & Turchick, 2012). Based on the dataset 2019, the average unemployment rate was 10%. The finding that a higher Gini ratio is associated with higher HDI was not supported by a prior study in East Java province, which found a non-significant association (Imaningsih et al., 2020).

Figure 6 shows that a region with high GRDP tends to have high unemployment. The association between GRDP (or GDP in general) and the unemployment rate is explained by Okun's Law (Okun's rule of thumb), which states that the higher the economic growth (higher GDP), the lower the unemployment rate (Wen & Chen, 2012). Therefore, this study's finding seems to contradict Okun's Law. However, some prior studies also indicated inconclusive findings regarding the Law. For example, a study among seven G7 countries reported that Okun's Law was confirmed for five countries only but not the other two (Benos & Stavrakoudis, 2020). In addition, Okun's Law was intended for a country but not region level (regency/municipality) as used in this study.

Moreover, the finding of this study indicated high economic growth (GRDP) and high-income inequality. This finding is partially supported by the OECD report stating that income inequality could be good (positive correlation) or bad (negative correlation) for economic growth (Cingano, 2014). Furthermore, the report indicated that extensive empirical studies produced an inconclusive relationship between both indicators. Therefore, the simple interpretation could be that high inequality encourages people to work harder, save, invest, and undertake risks for high rates of return. Subsequently, these contribute to economic growth (GDP).

Figure 6 indicates that a cluster with higher HDI, higher GRDP, higher income inequality, higher unemployment, and lower poverty is related to a region with a higher population density, smaller area, lower population, and fewer villages. The region characteristics represent the cities. It is a fact that 8 of 9 regions in this cluster were municipalities. This finding proved that integrating macroeconomic indicators and regional profiles is essential to discovering the underlying reason for the relationship between macroeconomic indicators.



Source: Author

Figure 6. Cluster Characteristics on Macroeconomic Indicators and Profile

Cluster members

Clustering analysis grouped 38 regions into three clusters. Figure 7 presents the cluster membership of each region in the East Java map. Cluster A contained all four regencies in Madura plus four others: Tuban, Pacitan, Probolinggo, and Bondowoso. These regions were likely to have a high poverty rate with low HDI, GRDP, income inequality, and unemployment. These regencies have a big area and the lowest population density. Four regencies in Madura regularly contribute to the poverty level in East Java province. A prior study affirmed that Pamekasan experienced development, while Sumenep and Sampang were stagnant, and Bangkalan was degrading (Purnama Sari et al., 2017). The Suramadu Bridge that connects Surabaya and Madura was expected and predicted to have a multiplier effect on Madura's economic development (Efendi & Hendarto, 2013), but the expectations remain unfulfilled.

More than half of East Java regions belonged to cluster B, characterized as having a moderate value in HDI, poverty rate, income inequality, unemployment rate, and regional GDP. The demographic profile of regions in this cluster indicates a high area, population, number of villages, and low population density. Furthermore, cluster C was occupied by eight municipalities and the Sidoarjo regency. Sidoarjo is a part of greater Surabaya and is likely to share some similar characteristics to Surabaya. The profile of these regions was a small area, low population, high population density, and a low number of villages—these demographic and geographic profiles were associated with the city's characteristics. The high unemployment rate and income inequality seem to be understood due to the fast economic development.

CONCLUSION AND RECOMMENDATION

This study has presented a case of the data mining process in analyzing macroeconomic data. The analysis produces empirical relationships among macroeconomic indicators and region profiles. As a data-driven type of analysis, the empirical relationships or models could become a new hypothesis for advancing economic theory. The use of the CRISP-DM process framework and the Knime Analytics Platform is effective in implementing data mining for macroeconomic data. Data mining is not only for a big volume of data but also a small volume of social-economic indicators of regencies/municipalities. While conventional data analysis stops finding the result of modeling, the predictive model is the advantage of the data mining method. Moreover, the model could be deployed for other datasets of different timeframe or regions.

This study has characterized East Java regencies/municipalities based on five macroeconomic indicators and four region profiles. The clustering method classified 38 regions into three clusters with differentiated characteristics. The combination of macroeconomic conditions indicates that a region might not achieve an ideal situation of the five macro development indicators. Some indicators, such as the unemployment rate or income inequality, might inevitably impact the successful economic development indicated by the increasing Gross Regional Domestic Products. The use of the predictive model has provided information on which regencies/municipalities were shifting between clusters.

The government might put the finding of this study as additional information to support the East Java Regulation (Perda Provinsi Jawa Timur) No. 5/ 2012. This regulation indicates that 38 regions are classified into eight development areas (Pemprov_Jatim, 2012). This study might suggest that the East Java Government considers different facilitation-focused programs based on the characteristics of three clusters. The development program will be more effective and efficient by determining priority programs based on the cluster characteristics. Understanding the characteristics of those regional governments is vital to improving economic competitiveness (Sambodo, 2018). Furthermore, a collaboration between regencies/municipalities within the cluster should be stimulated to overcome their weakness.

This study has the following limitations. First, the data mining was applied to only one clustering method with a k-means algorithm. Further study might use different methods depending on the suitability of the data. Second, the case analysis

only covered one province. Further research may broaden the coverage by analyzing regencies/ municipalities of several provinces so that a comparison can be drawn.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their valuable suggestions and comments.

REFERENCES

Arisman. (2018). Determinant of human development index in Southeast Asia. *Jurnal Ilmu Ekonomi*, *7*(2), 118–137. https://doi.org/10.37950/jkpd.v2i2.44

Barbaglia, L., Consoli, S., Manzan, S., Saisana, D. R. R. M., & Pezzoli, L. T. (2021). Data science technologies in economics and finance: A gentle walk-in. In S. Consoli, D. R. Recupero, & M. Saisana (Eds.), *Data Science for Economics and Finance: Methodologies and Applications* (1–17). Springer. https://doi.org/10.1007/978-3-030-66891-4

Benos, N., & Stavrakoudis, A. (2020). Okun's Law: Copula-based evidence from G7 countries. *Quarterly Review of Economics and Finance*, *In Press*. https://doi.org/10.1016/j.qref.2020.10.004

BPS-Jatim. (2019). *Indeks pembangunan manusia Provinsi Jawa Timur 2019*.

BPS-Jatim. (2020). *Provinsi Jawa Timur dalam angka 2020*.

Brito, P., & Malerba, D. (2003). Mining official data. *Intelligent Data Analysis*, *7*(6), 497–500. https://doi.org/10.3233/ida-2003-7601

Cao, L. (2017). Data Science. *ACM Computing Surveys*, *50*(3), 1–42. https://doi.org/10.1145/3076253

Cingano, F. (2014). Trends in income inequality and its impact on economic growth. *OECD Social, Employment, and Migration Working Papers*, *163*, 0_1,5-59. https://doi.org/http://dx.doi.org/10.1787/5jxrjncwxv6j-en

Colin Shearer. (2000). The CRISP-DM Model: The new blueprint for data mining. *Journal of Data Warehousing*, *5*(4), 13. https://doi.org/10.1109/EITech.2016.7519646

Costa, G. O. T., Machado, A. F., & Amaral, P. V. (2018). Vulnerability to poverty in Brazilian municipalities in 2000 and 2010: A multidimensional approach. *EconomiA*, *19*(1), 132–148. https://doi.org/10.1016/j.econ.2017.11.001

Cysne, R. P., & Turchick, D. (2012). Equilibrium unemployment-inequality correlation. *Journal of Macroeconomics*, *34*(2), 454–469. https://doi.org/10.1016/j.jmacro.2011.12.009

Efendi, M., & Hendarto, R. M. (2013). *PEREKONOMIAN PULAU MADURA (Studi Kasus Kabupaten Bangkalan)*. *3*, 1–13.

Feelders, A. J. (2002). Data mining in economic science. Dealing with the Data Flood, 166–175. http://www.staff.science.uu.nl/~feeld101/dmecon.pdf

Georgescu, I., Androniceanu, A.-M., & Kunnunen, J. (2020). A discriminant analysis to the quantification of human development index under economic inequality. *The 14th International Management Conference*, 1053–1062. https://doi.org/10.24818/imc/2020/05.15

Guazzelli, A., Zeller, M., Lin, W. C., & Williams, G. (2009). PMML: An open standard for sharing models. *R Journal*, *1*(1), 60–65. https://doi.org/10.32614/rj-2009-010

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Elsevier.

Harding, M., & Hersh, J. (2018). Big data in economics. *IZA World of Labor*, *September*, 1–10. https://doi.org/10.15185/izawol.451

Hartanto, W., Islami, N. N., Mardiyana, L. O., Ikhsan, F. A., & Rizal, A. (2019). Analysis of human development index in East Java Province Indonesia. *IOP Conference Series: Earth and Environmental Science*, *243*(012061). https://doi.org/10.1088/1755-1315/243/1/012061

Hassani, H., Gheitanchi, S., & Yeganegi, M. R. (2010). On the application of data mining to official data. *Journal of Data Science*, *8*, 75–89.

Hassani, H., Saporta, G., & Silva, E. S. (2014). Data mining and official statistics: The past, the present and the future. *Big Data*, *2*(1), 34–43. https://doi.org/10.1089/big.2013.0038

Heckman, J. J. (2001). Econometrics and empirical economics. *Journal of Econometrics*, *100*(1), 3–5. https://doi.org/10.1016/S0304-4076(00)00044-0

Hudáková, J. (2017). Relationship between gross domestic product and human development index. *4th International Multidisciplinary Scientific Conferences on Social Sciences & Arts SGEM 2017*, 665–672. https://doi.org/10.5593/sgemsocial2017/14/S04.087

Imaningsih, N., Priana, W., Sishadiyati, S., Asmara, K., & Wijaya, R. S. (2020). Analysis of factors affecting human development index East Java. *The 2nd International Conference on Economics, Business, and Government Challenges*. https://doi.org/10.4108/eai.3-10-2019.2291908

Johnston, M. P. (2014). Secondary data analysis: A method of which the time has come. *Qualitative and Quantitative Methods in Libraries (QQML)*, *3*, 619–626.

Khodabakhshi, A. (2011). Relationship between GDP and human development indices in India. *International Journal of Trade, Economics and Finance*, *2*(3), 251–253. https://doi.org/10.7763/ijtef.2011.v2.111

Liu, J., & Kemp, A. (2019). Forecasting the sign of U.S. oil and gas industry stock index excess returns employing macroeconomic variables. *Energy Economics*, *81*, 672–686. https://doi.org/10.1016/j.eneco.2019.04.023

López-Robles, J. R., Rodríguez-Salvador, M., Gamboa-Rosales, N. K., Ramirez-Rosales, S., & Cobo, M. J. (2019). The last five years of big data research in economics, econometrics and finance: Identification and conceptual analysis. *Procedia Computer Science*, *162*(Itqm 2019), 729–736. https://doi.org/10.1016/j.procs.2019.12.044

Mügge, D. (2016). Studying macroeconomic indicators as powerful ideas. *Journal of European Public Policy*, *23*(3), 410–427. https://doi.org/10.1080/13501763.2015.1115537

Pemprov_Jatim. (2012). *Peraturan daerah Provinsi Jawa Timur No 5 Tahun 2012 tentang rencana tata ruang wilayah provinsi tahun 2011-2030*.

Purnama Sari, I., Riyono, B., & Supandi, A. (2017). Indeks pembangunan manusia di Madura: Analisis tipologi Klassen. *Journal of Applied Business and Economics*, *110*(9), 1689–1699.

Purnamasari, S. B., Yasin, H., & Wuryandari, T. (2014). Pemilihan cluster optimum pada Fuzzy C-Means pengelompokan kabupaten/kota di Provinsi Jawa Tengah berdasarkan indikator indeks pembangunan manusia. *Jurnal Gaussian*, *3*(3), 491–498.

Qurrata, V. A., & Ramadhani, N. (2021). The impact of HDI, minimum wages, investment and grdp on poverty in East Java in 2019. *KnE Social Sciences*, *2021*, 411–418. https://doi.org/10.18502/kss.v5i8.9393

Rahmat, A., Hardi, H., Syam, F. A., Zamzami, Z., Febriadi, B., & Windarto, A. P. (2021). Utilization of the field of data mining in mapping the area of the Human Development Index (HDI) in Indonesia. *Journal of Physics: Conference Series*, *1783*(012035). https://doi.org/10.1088/1742-6596/1783/1/012035

Reyes, G. E., & Useche, A. J. (2019). Competitiveness, economic growth and human development in Latin American and Caribbean countries 2006-2015: A performance and correlation analysis. *Competitiveness Review: An International Business Journal*, *29*(2), 139–159. https://doi.org/10.1108/CR-11-2017-0085

Sambodo, M. T. (2018). Tata kelola dan peningkatan daya saing ekonomi nasional: Suatu penulusuran konsep. *Jurnal Ekonomi Pembangunan*, *25*(2), 33–46. https://doi.org/10.14203/jep.25.2.2017.33-46

Saporta, G. (2018). From conventional data analysis methods to big data analytics. In M. Corlosquet-Habart & J. Janssen (Eds.), *Big Data for Insurance Companies* (Vol. 1, pp. 27–41). John Wiley & Sons. https://doi.org/10.1002/9781119489368.ch2

Saputra, F. A., Barakbah, A., & Rokhmawati, P. R. (2020). Data analytics of human development index (HDI) with features descriptive and predictive mining. *International Electronics Symposium*, 316–323.

Sinaga, M. (2020). Analysis of effect of GRDP (gross regional domestic product) per capita, inequality distribution income, unemployment and HDI (human development index) on poverty. *Budapest International Research and Critics Institute (BIRCI-Journal): Humanities and Social Sciences*, *3*(3), 2309–2317. https://doi.org/10.33258/birci.v3i3.1177

Taylor, L., Schroeder, R., & Meyer, E. (2014). Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same? *Big Data and Society*, *1*(2), 1–10. https://doi.org/10.1177/2053951714536877

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28. https://doi.org/10.1257/jep.28.2.3

Wahyudi, I., Nuryartono, N., & Rifin, A. (2016). Cluster of Indonesia Kabupaten-Kota Potential in Developing Food Crop and Horticulture Commodities. *Indonesian Journal of Business and Entrepreneurship*, *2*(3), 151–164. https://doi.org/10.17358/ijbe.2.3.151

Wahyuningrum, F., & Etty Soesilowati. (2021). The effect of economic growth, population and unemployment on HDI. *Indonesian Journal of Development Economics*, *4*(2), 1217–1229.

Wen, Y., & Chen, M. (2012). Okun's Law: A meaningful guide for monetary policy? In *Economic Synopses* (Vol. 2012, Issue 15). https://doi.org/10.20955/es.2012.15