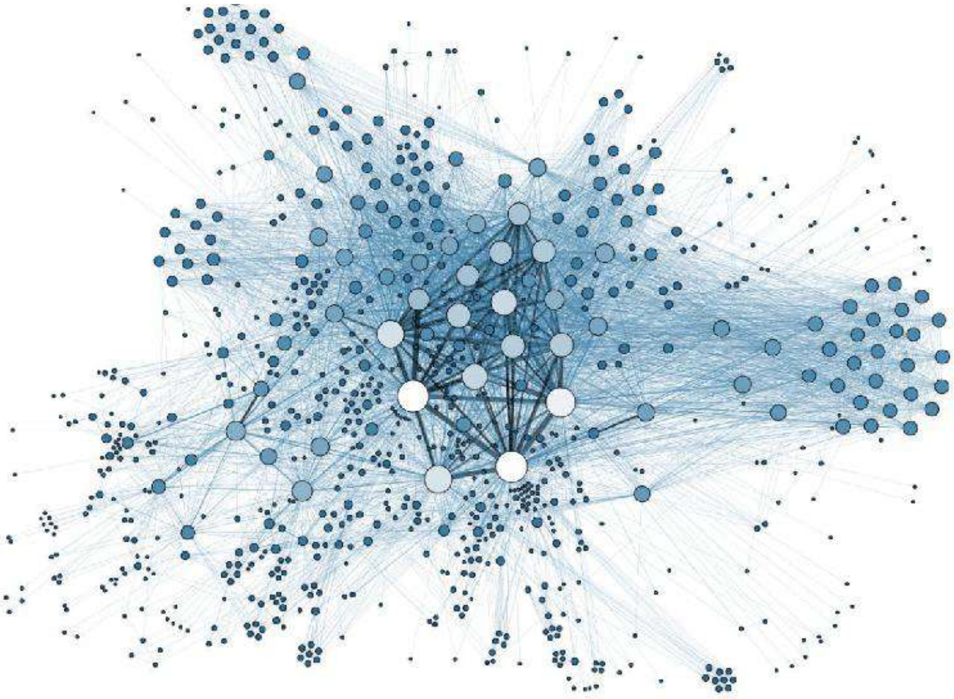
JURNAL APLIKASI SAINS DATA



Vol: 01 | No: 01 | Year: 2025

E-ISSN: 2025-xxxx
P-ISSN: 2025-xxxx
Doi: https://doi.org/10.33005/jasid.v1i1
https://jasid.upnjatim.ac.id/index.php/jasid

Editorial Team

E-ISSN: 2025-xxxx
P-ISSN: 2025-xxxx
Doi: https://doi.org/10.33005/jasid.v1i1
https://jasid.upnjatim.ac.id/index.php/jasid

Table of Content

E-ISSN: 2025-xxxx
P-ISSN: 2025-xxxx
Doi: https://doi.org/10.33005/jasid.v1i1
https://jasid.upnjatim.ac.id/index.php/jasid

Jurnal Aplikasi Sains Data

Journal homepage: https://jasid.upnjatim.ac.id/index.php/jasid

Comparison of ARIMA and SARIMA Methods for Non-Oil and Gas Export Forecasting in East Java

Dinda Galuh Guminta[1,*]

[1]Data Science Program Study of Universitas Negeri Surabaya

[1*]dindaguminta@unesa.ac.id

ABSTRACT

Abstract— Forecasting plays a pivotal role in economic planning, particularly in aligning supply with demand and informing production decisions. This study aims to compare the performance of the Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA) models in forecasting the non-oil and gas export values of East Java, a region known for its dynamic trade activity. Using monthly time series data spanning from January 2007 to January 2024, sourced from the Central Statistics Agency (BPS) of East Java Province, this research conducts an in-depth analysis of forecasting accuracy and model suitability. Before model implementation, the dataset underwent several preprocessing steps to ensure its quality, including the handling of missing values and outlier adjustments. Both ARIMA and SARIMA models were developed, calibrated, and evaluated using standard forecasting performance metrics, namely Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The ARIMA model exhibited consistently lower error rates across all three metrics, indicating its robustness in capturing the underlying patterns within the export data. In contrast, while the SARIMA model incorporated seasonal components, its performance did not surpass that of ARIMA in this specific case. The comparative findings suggest that, despite the seasonal nature of trade, the ARIMA model is more suitable for short-term forecasting of East Java's non-oil and gas exports. This research contributes to the broader literature on economic forecasting by emphasizing the importance of selecting appropriate models based on data characteristics. Furthermore, the results provide valuable insights for policymakers and stakeholders engaged in export planning and regional trade development In this result the ARIMA model overcome the SARIMA with MAPE 0.116 to 0.983.

Keywords: ARIMA, export, forecasting, gas, SARIMA.

I. INTRODUCTION

Forecasting is essential for a company because it relates to the balance between exports and product demand. Forecasting is a prediction of future data values based on relevant historical data. This process greatly influences decisions regarding the amount of production of goods to be exported [1]. The forecasting uses data recorded by the Central Statistics Agency (BPS) of East Java Province for 14 years.

This forecast aims to ensure that non-oil and gas exports in East Java remain regular and adequate. If demand is lower than the available stock, exports can be adjusted, and vice versa, if demand increases, production must be increased to be sufficient. The most significant decline in non-oil and gas exports in East Java occurred in July 2016, which was 37.68% from the previous month, and an increase occurred

* *Corresponding author.*

E-mail address: dindaguminta@unesa.ac.id

https://doi.org/10.33005/jasid.v1i1.2

in July 2018, by 53.34% from the previous month [2].

The solution can be found through forecasting or prediction to overcome this situation. Thus, the methods used to analyze data from January 2007 to January 2024 are Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average (SARIMA). Both methods are applied to predict future numbers and plan appropriate strategies. However, this study will compare the two methods and determine the most appropriate method for the existing data. In a previous study conducted by Herlena Bidi Astuti, et al. in an article entitled "ESTIMATION OF RETAIL PRICES OF CHICKEN EGGS IN BENGKULU CITY", it was concluded that the ARIMA model proved effective for forecasting with MSE and MAPE of 1,600 and 6.23, respectively [3]. Another study conducted by Laras Luthfiyyah Ibrahim and Eti Kurniati entitled "Forecasting the Number of Executive Train Passengers in Java Using the SARIMA Model" showed that the SARIMA(1,0,1)(1,1,0)12 model was used to estimate the number of train passengers in Java in March approaching the Eid al-Fitr holiday, with forecasting results reaching 4470 people [4]. The ARIMA method utilizes historical and current value data to produce accurate short-term predictions, allowing detailed analysis of data patterns and fluctuations, and providing reliable predictions for a limited period into the future [5]. Meanwhile, the SARIMA method forecasts data by considering seasonal components in its model, making it suitable for data analysis that shows specific patterns or cycles over a specific period.

Ref [6] using the ARIMA and SARIMA methods, tests were carried out to determine the stationarity of the variance and mean of the data, as well as to evaluate the significance of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). The next step is to perform differencing to make the data stationary. After that, the calculation of Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) values is carried out to validate the model. This approach can predict the amount of non-oil and gas exports in several future periods, providing the information needed to determine the optimal value of non-oil and gas exports.

II. RESEARCH METHODOLOGIES

The methodology in research refers to a series of systematic steps to achieve the stated research objectives. This methodology is designed with a structured and scientific flow. This study begins with collecting data to be used, pre-processing data, exploring time series data, performing data modeling, predicting testing data, and evaluating the model. Researchers use RSME, MAE, and MAPE to measure the model’s accuracy.

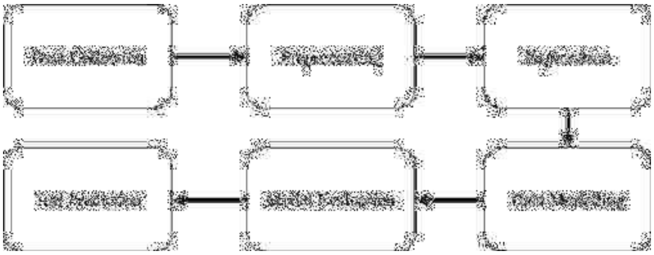


Figure 1. Research Flow

A. Data Collecting

The data we use in this study is monthly data from non-oil and gas exports in East Java, taken from the BPS East Java Province page. The data period used covers from January 2007 to January 2024. The selection of this period is based on the need to have a sufficient number of observations for time series analysis to identify patterns in the data.

B. Preprocessing

Data preprocessing is a crucial step in this study to ensure that the dataset is properly structured and ready for further analysis. The first task involves combining Excel files containing export data from the years 2007 to 2024 into a single dataset. This merging process creates a comprehensive and continuous data source. Following this, unnecessary rows—such as headers, footers, or empty rows that do not contribute meaningful information—are removed to clean the dataset.

Next, the column names are adjusted to represent the months of the year, from January to December, ensuring consistency and clarity. A new column labeled "Year" is then added to the far-left side of the dataset, containing the years 2007 to 2024. This "Year" column is subsequently set as the index of the DataFrame, which is essential for performing time-based operations and maintaining chronological order.

The dataset is then transposed to restructure the data, allowing the creation of a proper date range based on the number of columns. This transformation supports the conversion of the matrix-like format into a time series-friendly structure. The data is reshaped from a matrix into a series to facilitate easier manipulation and analysis. It is then transposed again to ensure the records are aligned in correct chronological order. This step helps preserve the temporal integrity of the dataset.

After restructuring, the data is reshaped into a columnar matrix format suitable for analytical models. A new DataFrame is created that focuses specifically on non-oil and gas export values, which are the primary subject of this study. To finalize the preprocessing, the index of the data is displayed to confirm that the date range is accurate, and the overall structure is verified. Finally, dataset information such as data types and missing values is reviewed, and any empty or null entries are removed to ensure the quality and completeness of the dataset.

C. Exploration

In the exploration of East Java's non-oil and gas export data from 2007 to 2024, the initial step is to create a time series plot to visualize trends throughout the time period. After that, a decomposition plot is carried out to identify the main components such as trends, seasonality, and residual components in the data. Furthermore, ACF and PACF analysis are used to evaluate the dependence between data values at previous times. A stationarity test is then carried out to verify whether the non-oil and gas export data has a constant mean and variance over time, so that the results indicate that the data is not stationary. Therefore, a differencing step is carried out to eliminate trends or seasonal patterns that may exist in the data, thus ensuring that the data is stationary and ready for further analysis using forecasting models such as ARIMA or SARIMA.

D. Data Modelling

Next, modeling is done using the ARIMA and SARIMA forecasting methods. Modeling using ARIMA will help identify patterns in non-stationary data, while SARIMA will consider seasonal effects that may affect export trends over time.

To carry out the modeling process, data that has gone through the pre-processing stage is required. Then, the researcher compares the results between the two methods that have been used. Then the data is separated into 2, namely as training data and testing data. After that, ARIMA and SARIMA modeling is carried out using training data until the best model is obtained.

E. Test Prediction

After performing data modeling using the ARIMA and SARIMA methods, the next step is to display the prediction results of the testing data from each method using a model that was previously created using training data.

F. Model Evaluation

Evaluation is an important stage to assess the performance or results of the developed model. In this context, metrics, RMSE, MAE, and MAPE are used to measure the level of model accuracy. After testing the model on the testing data, a comparison is made between the ARIMA and SARIMA methods. The best method will be selected based on the lowest MAPE results.

III. RESULT AND DISCUSSION

Researchers perform pre-processing using the help of the pandas and numpy libraries available in python used to combine excel then delete unnecessary rows and leave non-oil and gas data that will be used for modeling. The columns used are the non-oil and gas export report columns every month from January to December and the rows are the index years of non-oil and gas exports carried out, namely from 2007 to 2024. then transpose the data to create a date range that will later be used to sort the data by date. After that, change the data form into a column matrix until the data is shaped like the table below so that the data is ready to be carried out in the next stage.

TABLE I. EXPORT DATA NON-GAS IN EAST JAVA

Date	Data
2007-01-01	742.4
2007-02-01	793.83
2007-03-01	768.19
2007-04-01	955.19
2007-05-01	837.04
…	…
2023-08-01	1700.74
2023-09-01	1756.17
2023-10-01	1989.74
2023-11-01	2015.38
2023-12-01	2146.41

Based on the data obtained as in table 1, we will then process the data using the ARIMA and SARIMA methods to determine which method is most appropriate for use in forecasting non-oil and gas in East Java. The first step is to plot the data.

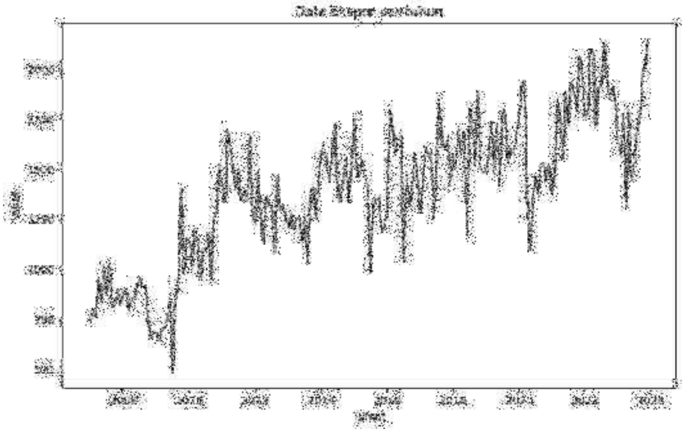


Figure 2. Data Plot

The sales data pattern obtained is a combination. In certain periods there is an increase and in certain periods there is also a decrease [7]. This data is not intermittent due to the graph from Figure 2, so it will be no problem to using an ARIMA or SARIMA model. A study comparing Croston, SES, and deep learning methods for intermittent demand forecasting found that RNN outperformed traditional models in empirical tests, with MAE proving to be a more robust evaluation metric than RMSSE [8].The highest increase occurred in July 2018 and the lowest decrease occurred in July 2016. The next stage is to display a decomposition plot containing trends, seasonality, and residual components in the data.

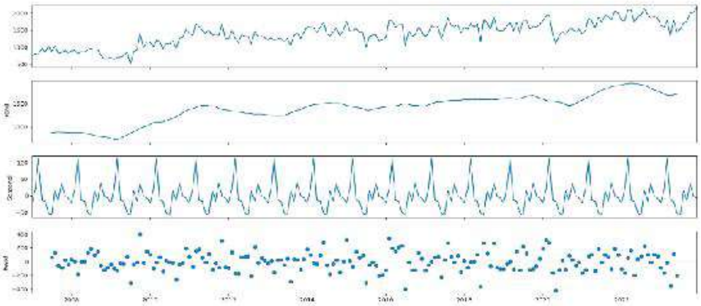


Figure 3. Decomposition Plot

From the illustration in Figure 3, it can be said that non-oil and gas export data in East Java has a long-term trend that increases over time, with significant annual and seasonal fluctuations. This indicates an increase in the value of non-oil and gas exports in East Java. The next step is to check the stationarity against the average value, using the ACF and PACF plots. This data look like can be solve using parametric instead non-parametric. Yet, a study proposed VGAMCV, a semi-parametric method combining Vector Generalized Additive Models with cross-validation, as a novel approach for multi-label classification, showing promising results compared to Random Forest, though without statistically significant improvements [9].

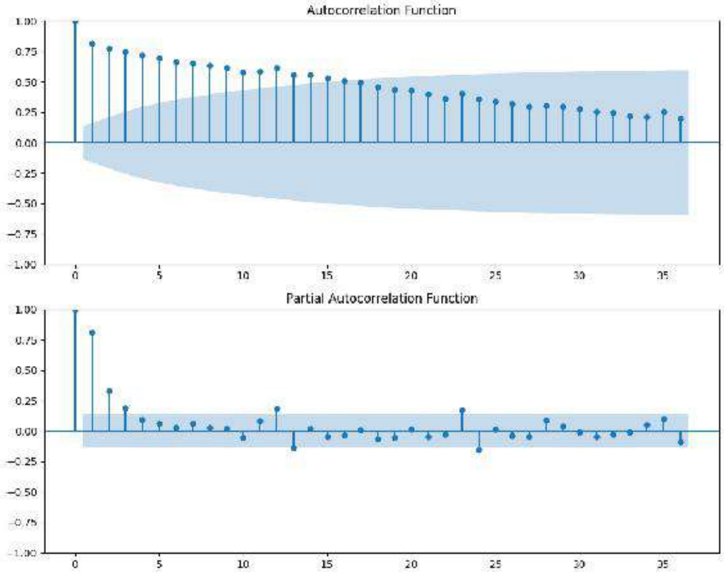


Figure 4. Plot ACF and PACF

It can be seen that in Figure 4 the ACF and PACF values at lag 0 are always 1 because they represent the correlation between the data value at lag 0 and itself. In the ACF plot, it can be seen that the ACF value decreases gradually (dies down) as the lag increases. This indicates that there is a dependency between the current value and the previous value in the data, but the dependency becomes weaker as the time interval increases. Data at lag 1 to lag 5 significantly affect the current data value. Data values at lag 1 and lag 2 have a significant partial effect on the current data value, regardless of the effect of data values at lag 0. Thus, data needs to be differencing so that the data becomes stationary.

TABLE II. DIFFERENCING RESULT

	adf	**kpss**
Test Statistic	-6.610145	0.056647
p-value	0.0	0.1
Numbers of lags	10	19
decision	Stationary	Stationary
Critical Value (1%)	-3.464875	0.739
Critical Value (5%)	-2.876714	0.463
Critical Value (10%)	-2.574859	0.347
Critical Value (2.5%)	NaN	0.574

Based on the results of differencing using the ADF test, it can be seen that the test statistic (-6.610145) is much lower than the critical value, as well as the p-value which is small from 0.05, this means that the data does not have a unit root (stationary) at that level of significance. While the KPSS test with a p-value of 0.1 and a test statistic (0.056647) which is much lower than the critical value, which means that the data does not have a trend that changes over time. So from the results of the stationarity test using the ADF and KPSS tests, it can be concluded that the non-oil and gas export data in East Java after differencing has become stationary.

Next, modeling is done using the ARIMA and SARIMA forecasting methods. Modeling using ARIMA will help identify patterns in non-stationary data, while SARIMA will consider seasonal effects that may affect export trends over time.

To perform modeling, data that has gone through the pre-processing stage is required. Researchers compare the results of the two models used. Then the data is separated into 2 types, namely used as training data and testing data. After that, ARIMA and SARIMA modeling is carried out using training data to obtain the best model.

When modeling data, the first thing to do is transform the data to make it stationary. This process is carried out by data differentiation, namely reducing each observation with the previous observation to eliminate trends and keep the variance constant. After that, the best parameter search is carried out for the ARIMA and SARIMA models using the auto_arima function.

TABLE III. BEST MODEL

Best Model	ARIMA(0, 0, 1)(0, 0, 1)[12]
Total Fit Time	16.235 seconds

From Table III, it is known that the output of the ARIMA model, the best parameters selected for the model are ARIMA(0,0,1)(0,0,1)[12]. This shows that the best ARIMA model has a seasonal component with an order of (0,0,1) and a period of 12 months. The recorded AIC value shows that the selected model has sufficient goodness of fit, with a low AIC value. Interpretation of this output helps to understand how the ARIMA model is selected based on the best parameters. Model coefficients, such as ma.L1 and ma.S.L12, show the influence of the previous period error and seasonal error on the value of non-oil and gas exports in East Java. The predictions generated from this model can provide insight into possible fluctuations in non-oil and gas exports in the future, although it should be noted that these predictions can vary from month to month. Thus, the output provides an overview of how the ARIMA model is selected and how the interpretation of these parameters helps in understanding data patterns and behavior.

TABLE IV. BEST MODEL

Best Model	ARIMA(0, 0, 1)(0, 0, 1)[12]
Total Fit Time	38.693 seconds

From Table IV, it is known that the SARIMA model output, the best parameters selected for the model are ARIMA (0,0,1) (0,0,1) [12]. This shows that the best SARIMA model has a seasonal component with an order of (0,0,1) and a period of 12 months. The recorded AIC value shows that the selected model has sufficient goodness of fit, with a low AIC value.

From this output, it helps to understand how the SARIMA model is selected based on the best parameters. The model coefficients, such as ma.L1 and ma.S.L12, show the influence of the previous period error and seasonal error on the value of non-oil and gas exports in East Java. The predictions generated from this model can provide insight into possible fluctuations in non-oil and gas exports in the future, although it should be noted that these predictions can vary from month to month. Thus, the output provides an overview of how the SARIMA model is selected and how the interpretation of these parameters helps in understanding data patterns and behavior.

After modeling the data using the ARIMA and SARIMA methods, the next step is to display the prediction results of the testing data from each method using a model that has been created previously using training data.

From the prediction results using the ARIMA model, the estimated value of non-oil and gas exports in East Java Province for the testing data period has been obtained. This prediction describes the data pattern that has been identified during the modeling process, taking into account the autoregressive and moving average effects in the data. Thus, this prediction provides an overview of the fluctuations in the value of non-oil and gas exports in the future based on the ARIMA model. This prediction is generated using the ARIMA code function, namely by modeling using the 'auto_arima' command on the training data and then making predictions for the testing data.

Meanwhile, the prediction using the SARIMA model also produces an estimate of the value of non-oil and gas exports in East Java Province for the testing data period. The SARIMA model considers seasonal effects that may affect export trends over time. Therefore, this prediction provides a more complete picture of the data pattern by taking seasonal factors into account.

To evaluate the performance of both models, plots of actual data and predictions from ARIMA and SARIMA have been presented. The plot of actual data and predictions generated from the ARIMA and SARIMA models provides a visual representation of how well the two models are able to predict the value of non-oil and gas exports in East Java.

The forecasting was conducted using data that has been collected by the BPS of East Java Province for 14 years.

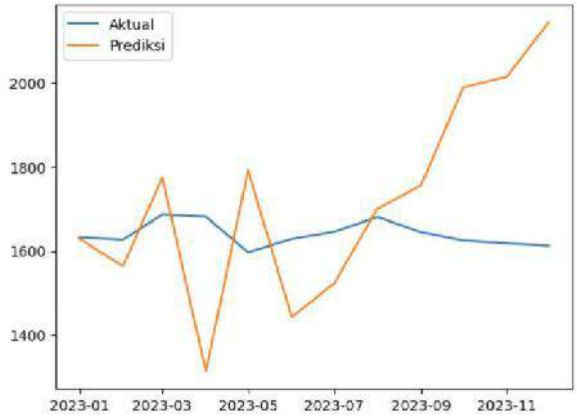


Figure 5. Example of a figure caption. *(figure caption)*

From Figure 5, it is known that the plot shows that the actual line (which represents the actual value) and the predicted line (which represents the value predicted by the model) have quite significant differences. This difference indicates that the two models may have different levels of accuracy in predicting the value of non-oil and gas exports.

In the ARIMA model, the difference between the actual and predicted lines may vary more from month to month. This shows that the ARIMA model may have fluctuating levels of accuracy in predicting the value of non-oil and gas exports at different times.

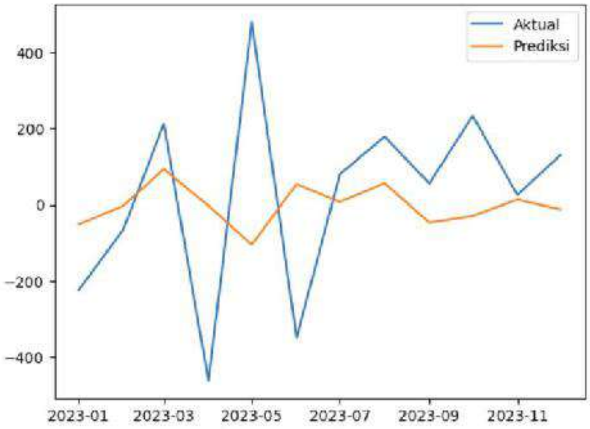


Figure 6. Example of a figure caption. *(figure caption)*

Meanwhile, in Figure 6, it is known that in the SARIMA model, it can be seen that the difference between the actual and predicted lines tends to be more stable or less variable from month to month. This may indicate that the SARIMA model has a more consistent level of accuracy in predicting the value of non-oil and gas exports in East Java. From the plot, a significant difference occurs between the actual and predicted lines. This illustrates that the models have different levels of accuracy in predicting the value of non-oil and gas exports in East Java Province. Further evaluation of the performance of the two models can be done by comparing evaluation matrices such as RMSE or MAE.

Evaluation is an important stage to assess the performance of the developed model. The best prediction or forecast is seen based on its level of accuracy, the smaller the error rate, the more accurate a model is in predicting. In this study, there are several evaluation metrics used to estimate the level of accuracy of the prediction results, such as RMSE, MAE, and MAPE. In this study, the two models can be compared based on the RMSE, MAE, and MAPE values to determine the best model. The smaller the RMSE, MAE and MAPE values mean that the model is the best model to use in predicting the value of non-oil and gas exports in East Java Province. Table V will show the results of the ARIMA and SARIMA model performance tests.

TABLE V. EVALUATION MODEL RESULTS

Model	RMSE	MAE	MAPE
ARIMA(0,0,1)	262.32	204.85	0.116
SARIMA(0,0,1)[12]	271.84	210.15	0.983

Based on Table V, it shows that the RMSE, MAE, and MAPE values of the ARIMA model are lower than the SARIMA model. So the ARIMA model is the best model that can be used for the analysis of non-oil and gas export values in East Java.

CONCLUSION

Based on the comparison of ARIMA and SARIMA forecasting methods in forecasting Non-Oil and Gas Exports in East Java for January 2007 to January 2024, the best method that can be used for forecasting non-oil and gas exports in East Java Province is the ARIMA method. This is because the ARIMA method shows lower RMSE, MAE, and MAPE values compared to the SARIMA method, namely RMSE: 262.32, MAE: 204.85, and MAPE: 0.116. Thus, this study recommends the use of the ARIMA model as the main forecasting tool for non-oil and gas exports in East Java, given its proven ability to produce predictions with smaller errors compared to the SARIMA model.

REFERENCES

[1] Erdin, "Peramalan Jumlah Penyediaan Air Bersih oleh Perusahaan Daerah Air Minum (PDAM) terhadap Masyarakat di Kabupaten Gowa Tahun 2020 dengan Metode ARIMA," Universitas Islam Negeri Alauddin Makasar, 2020.

[2] "Badan Pusat Statistik Provinsi Jawa Timur," 2024. [Online]. Available: https://jatim.bps.go.id/indicator/8/609/1/nilai-ekspor-jawa-timur-menurut-kategori-migas-dan-non-migas-bulanan.html. [Accessed 20 May 2024].

[3] H. B. Astuti, E. Fauzi, W. E. Putra, A. Alfayanti and A. Ishak, "Estimating Model Forecasting the Price of Chicken Eggs In the City of Bengkulu," AGRITEPA: Jurnal Ilmu Dan Teknologi Pertanian, vol. VIII, no. 2, pp. 137-147, 2021.

[4] L. L. Ibrahim and E. Kurniati, "Peramalan Jumlah Penumpang Kereta Api Eksekutif di Pulau Jawa Menggunakan Model SARIMA," Jurnal Riset Matematika, pp. 73-82, 2023..

[5] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti and M. Ciccozzi, "Application of the ARIMA model on the COVID-2019 epidemic dataset," Data in brief, vol. 29, p. 105340, 2020.

[6] I. L. M. S. a. C. D. S. Hakim, "Analisis Peramalan Harga Telur Ayam Ras Dengan Menggunakan Metode SARIMA," JURNAL MEDIA INFORMATIKA BUDIDARMA, vol. 8, no. 2, pp. 966-977, 2024.

[7] S. Suseno and S. Wibowo, "Penerapan Metode ARIMA dan SARIMA Pada Peramalan Penjualan Telur Ayam Pada PT Agromix Lestari Group," Jurnal Teknologi dan Manajemen Industri Terapan, vol. 2, no. I, pp. 33-40, 2023.

[8] A. Muhaimin, D. D. Prastyo and H. Horng-Shing Lu, "Forecasting with Recurrent Neural Network in Intermittent Demand Data," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 802-809, doi: 10.1109/Confluence51648.2021.9376880.

[9] A. Muhaimin, W. Wibowo, and P. A. Riyantoko, "Multi-label classification using vector generalized additive model via cross-validation," Journal of Information and Communication Technology, vol. 22, no. 4, pp. 657–673, 2023, doi: 10.32890/jict2023.22.4.5.

Jurnal Aplikasi Sains Data

Journal homepage: https://jasid.upnjatim.ac.id/index.php/jasid

Implementation of Content-Based Filtering in Tourist Destination Recommendation System in Central Java

Adigama Tri Nugraha[1,*]

[1]Statistics Department of Universitas Sebelas Maret

[1*]adigama@student.uns.ac.id

ABSTRACT

Abstract— Tourism is widely recognized as a strategic sector that significantly contributes to regional economic growth, not only through direct revenue streams but also via foreign exchange earnings, job creation, and the development of supporting industries. Central Java, a province in Indonesia rich in cultural heritage and natural beauty, possesses diverse tourist destinations ranging from natural landscapes and artificial attractions to cultural and special interest sites. However, despite this diversity, tourists often face challenges in discovering attractions that align with their personal preferences due to limited or non-personalized information services. To address this issue, this study proposes the development of a personalized recommendation system for tourist attractions in Central Java, leveraging content-based filtering techniques in combination with neural network machine learning. The system is designed to analyze both the intrinsic features of tourist sites and the explicit preferences of users to deliver highly relevant and individualized recommendations. The model is trained using the Adam optimization algorithm with a learning rate of 0.01 over 300 epochs to ensure stable and efficient learning. Evaluation results indicate that the system is capable of generating accurate recommendations with relatively low prediction error, as reflected by a Mean Squared Error (MSE) value of 0.1766.The outcomes of this research demonstrate that integrating artificial intelligence into tourism information services can significantly enhance the decision-making experience for tourists. By providing smarter and more user-centric recommendations, the proposed system not only helps travelers explore suitable destinations but also contributes to the broader objective of optimizing the tourism sector in Central Java through digital innovation.

Keywords: content-based filtering, machine learning, recommender system, tourism

I. INTRODUCTION

Tourism is a sector that has the potential to increase regional income [1]. Tourism development plays a crucial role in absorbing labor, encouraging equal business opportunities, supporting equitable national development, and making a significant contribution to state foreign exchange earnings. One of the areas with potential in the tourism sector is Central Java [2]. In 2018, there were 692 tourist attractions in Central Java, and this number increased to 1216 in 2022. In detail, tourist attractions in Central Java in 2022 include 454 natural destinations, 414 cultural destinations, 172 artificial destinations, 105 special interest destinations, and 71 other destinations. The tourism potential in Central Java continues to be explored through various efforts made to improve this sector. One of the steps taken is through tourism destination development programs, tourism development, and development of tourism and Creative

* *Corresponding author.*
E-mail address: adigama@student.uns.ac.id

https://doi.org/10.33005/jasid.v1i1.3

Economy (Ekraf) Human Resources (HR). The destination development program involves tourism area development activities, increasing tourist attractions, and developing the tourism industry. On the other hand, tourism development programs include efforts such as tourism market development, tourism promotion and information, and tourism image in Central Java[3].

In one of these tourism development programs, there are tourism promotion and information efforts to improve the tourism sector in Central Java. One of the things done to improve tourism information is to create a recommendation system related to tourist attractions in Central Java. This system can help users find tourist attractions that suit their preferences.

This study aims to create a tourist attraction recommendation system in Central Java through the application of machine learning with the content-based filtering method. By using a neural network, this system will learn the complex non-linear relationship between various tourist attraction features such as ratings, locations, and types of tourist attractions to produce more personal and accurate recommendations. Analysis of tourist attraction attributes and user preferences is the main focus, with Mean Squared Error (MSE) used to evaluate the performance of the recommendation system on the test set. As a result, this system is able to provide recommendations for tourist attractions that suit the user's profile and preferences, which can be sorted by rating, location, type of tourist attraction, or a combination of several factors.

II. RESEARCH METHODS

A. Data

This study utilizes two main data sources, namely tourist attraction data in Central Java, and User data that has provided reviews of tourist attractions in Central Java obtained through the Google Places API, which is an Application Programming Interface that provides location-based geographic data via the internet using HTTP [4]. Tourist attraction data in Central Java was collected using the midpoint coordinates of each city/district as a search reference. The search was conducted within a radius of 100,000 meters from the midpoint with search parameters set to obtain places with the category "tourist_attraction". The data taken initially amounted to 1400 tourist attractions, then filtered into 432 tourist attractions in Central Java based on 13 attributes used. The tourist attraction attributes used in this study can be seen in Table I.

TABLE I. DATA ATRIBUTE

ID	Attribute	Description	Data Type
1	Place_id	Unique ID of Tourist Attraction	*Numeric*
2	Place Name	Name of Tourist Attraction	*String*
3	formatted_phone_number	Phone Number of Tourist Attraction	*String*
4	formatted_address	Address of Tourist Attraction	*String*
5	city	City	*String*
6	website	Website	*String*
7	rating	Rating of Tourist Attraction	*Numeric*
8	user_ratings_total	Total Ratings	*Numeric*
9	photo_preference	Photo of Tourist	*String*
10	lng	Longitude	*Numeric*
11	url	Google Maps Website	*String*
12	url_photo	Google Maps Photo	*String*
13	type_tourist	Type of Tourist Attraction	*String*

User data (reviewers) who have given reviews on tourist attractions in Central Java were extracted from the Google Places API. The initial user data amounted to 7000, then filtered into 2132 user data with 3 variables used. The user data variables used can be seen in Table II.

TABLE II. USER DATA ATRIBUTE

No	Variabel	Keterangan	Jenis Data
1	place_name	Name of Tourist	*String*
2	reviewer_name	User Name	*String*
3	reviewer_rating	Rating Given by User	*Numeric*

B. Research Stagees

The data preprocessing stage is a crucial step in preparing data before it is used for modeling. Some of the preprocessing stages carried out are One Hot Encoding, Data Normalization, and dividing training data and test data.

One Hot Encoding is a technique in data processing, especially in categorical or classification cases, where categorical variables are converted into numeric variables. One Hot Encoding is done to avoid ambiguity of order or comparison in categorical variables and to ensure that the machine learning model can understand and process the information properly [5].

Data Normalization uses StandardScaler for tourist attraction and user features, and MinMaxScaler for the target variable, namely user ratings. StandardScaler performs two main transformations, namely calculating the mean and standard deviation of the data, and normalizing the data by subtracting the mean and dividing by the standard deviation [6]. StandardScaler is formulated in equation (1).

$$z = \frac{x-\mu}{\sigma} \tag{1}$$

MinMaxScaler is a preprocessing method that changes the feature transformation by adjusting each feature individually to a certain range [7]. Also we used the train and testing rule into 80% and 20%. The purpose of MinMaxScaler is to control the range of values of each sample in a feature so that it is not too large. MinMaxScaler is formulated as follows (2).

$$v' = \frac{v-\min(a)}{\max(a)-\min(a)}(range.max - range.min) + range.min \tag{2}$$

Perform neural network modeling. Neural networks or Artificial Neural Networks (ANN) are computational information processing systems that mimic the characteristics of human biological neural networks [8]. Neural networks involve the use of neurons as processing elements, where signals between neurons are sent through connectors. The model is built using a neural network architecture consisting of several layers, including input layers, hidden layers (sequential), L2 Normalization Layer, Dot Product Layer, and Output Layer (Dense). In the modeling process, the activation functions used are ReLU (Rectified Linear Unit) and Tanh in the hidden layers to handle non-linearity in the data [9]. The parameters used include the Adam optimizer, with a learning rate of 0.01. Each combination of parameters is run for 300 epochs, and model performance is evaluated using the Mean Squared Error (MSE) metric.

Model evaluation is performed on test data using the Mean Squared Error (MSE) metric. MSE measures the average of the squared differences between the values predicted by the model and the observed values [10]. Mathematically, MSE is calculated through equation (2.7). The MSE value is evaluated using the Adam optimizer with a learning rate of 0.01 to determine the best performance on the test data [11]. This evaluation aims to find the configuration that produces the most accurate predictions with minimal errors.

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left(x_{s,t} - x_{0,t}\right)^2 \quad (3)$$

After the evaluation process, we choose the best model. Implementing the model results into a tourist attraction recommendation system in Central Java.

III. RESULTS AND DISCUSSIONS

The construction of the neural network model in this study consists of several layers starting from input layers, hidden layers (sequential), l2 normalization layer, dot product layer, and ending with the output layer (Dense). Based on Table III, the total parameters obtained are 4082, the trained parameters are 4018, and 64 untrained parameters.

TABLE III. DATA ATRIBUTE

Layer (type)	**Output Shape**	**Param #**	**Connected to**
inputLayer_4 (InputLayer)	(*None*, 39)	0	-
*inputLayer_*5 (*InputLayer*)	(*None*, 41)	0	-
sequential_2 (*Sequential*)	(*None*, 8)	2008	*input_layer_*4[0]...
sequential_3 (*Sequential*)	(*None*, 8)	2072	*input_layer_*5[0]...
l2_normalize_layer... (L2NormalizeLayer)	(*None*, 8)	0	*sequential_*2[0][...]
l2_normalize_layer... (L2NormalizeLayer)	(*None*, 8)	0	*sequential_*3[0][...]
dot_1 (Dot)	(*None*, 1)	0	*l2_normalize_layer...*
			l2_normalize_layer...
dense_13 (Dense)	(*None*, 1)	2	*dot_1*[0][0]
Total Params		4082	
Trainable Params		4018	
Non-trainable Params		64	

The neural network architecture in this study uses two separate neural networks, namely users and tourist attractions. The model starts with two input layers that receive 39 attributes for users and 41 attributes for tourist attractions. Each input is processed through an identical neural network, consisting of a dense layer with 32 neurons and ReLU activation, followed by a dropout layer to reduce overfitting. Next, there is a second Dense layer with 16 neurons and ReLU activation, followed by batch normalization to improve training stability. The output of each neural network is normalized using L2 Normalization before calculating the dot product between the two normalized vectors. The dot product produces a match score between user preferences and tourist attraction characteristics. The dot product results are then fed into the final dense layer with one neuron and Tanh activation, producing a final prediction that reflects the level of match between users and tourist attractions. The results of the neural network architecture can be seen in Figure 1. The training process is carried out with optimizer, learning rate, and epoch parameters. The optimizer used is Adam with a learning rate of 0.01 and is run for 300 epochs and produces an MSE result of 0.1858.

Model evaluation is carried out on test data which aims to assess the model's ability to generalize to data that has never been seen before. This is important to ensure that the model not only works well on training data, but also has adequate performance when applied to real data. By measuring the Mean Squared Error (MSE) on the test data, the optimizer and parameters that provide the best performance can be identified, as well as understanding how the model adapts to changes in the input data.

After all predictions are calculated, the prediction results are compared with the actual values of the test data which have also been returned to their original scale. To evaluate model performance, the error metric is calculated using MSE. MSE is calculated by taking the average of the squared differences between the actual and predicted values. The MSE value provides an idea of how much average squared error the model makes in predictions.

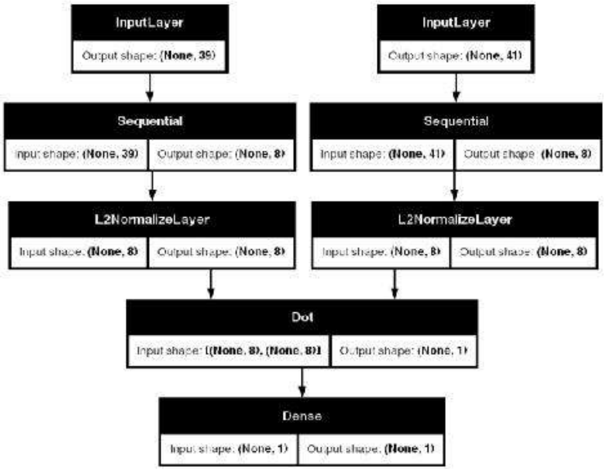


Figure 1. Neural Network Architecture

The results of this evaluation are then printed, with the MSE value being the main indicator for measuring model accuracy. Adam optimizer with a learning rate of 0.01 has an MSE of 0.1766 indicating that the model has better prediction performance, because the mean squared error between the prediction and the actual value is smaller. A comparison of the evaluation results of training data and test data can be seen in Figure 2.

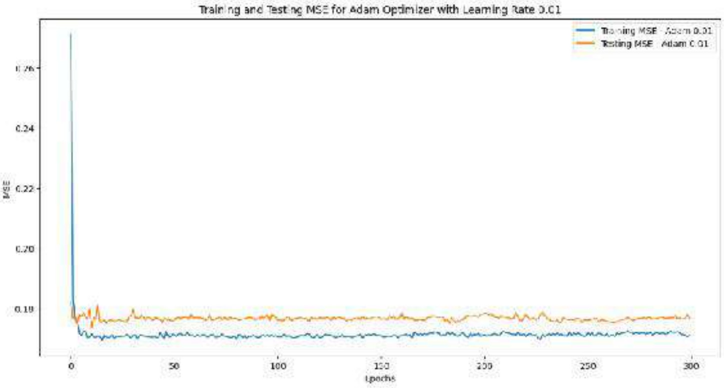


Figure 2. Loss Function Graph

The results of this study are recommendations for tourist attractions in Central Java based on user preferences who have provided reviews for tourist attractions in the region. The recommendation system used relies on historical review data from users to identify tourism patterns and preferences, so that it can provide recommendations that are more relevant and in accordance with user interests. Table IV shows users who have provided reviews for tourist attractions in Central Java.

TABLE IV. EXAMPLE OF USER PREFERENCE

ID Reviewer	Place Name	City	reviewer_rating	Tourist Type
1	Agro Salak Sodong Batang	Kabupaten Batang	4	Unnatural

Based on the preferences of users who have provided reviews for the Agro Salak Sodong Batang tourist attraction in Batang Regency with an artificial tourist type, the results of the recommendations for tourist attractions in Central Java are as in Table V.

TABLE V. EXAMPLE OF USER PREFERENCE

No	Place Name	City	Place Rating	User Ratings Total	Tipe Wisata	Predicted Rating
1	Curug Sewinong	Kabupaten Batang	5,0	3	Alam	5
2	Safari Beach Jateng	Kabupaten Batang	4,4	4474	Alam	5
3	Jembatan Sibiting	Kabupaten Batang	4,4	306	Buatan	5
4	Agro Salak Sodong Batang	Kabupaten Batang	4,3	54	Buatan	5
5	Agro Wisata Selopajang Timur	Kabupaten Batang	4,2	834	Buatan	5

The recommendation results show that the system successfully identified tourist attractions that users are most likely to like based on their preferences. The recommended attractions have artificial tourist types that match users' preferences, except for two attractions that have high popularity and predicted ratings even though they are not artificial tourist attractions. By providing relevant recommendations, users can enjoy a more personalized and satisfying tourist experience.

CONCLUSION

This study successfully developed a tourist attraction recommendation system in Central Java using a neural network based on content-based filtering. This system uses a neural network architecture consisting of input layers for user and tourist attraction attributes, sequential hidden layers for optimal attribute processing, L2 normalization layer for data normalization, dot product layer for suitability calculation, and dense output layer for recommendation prediction. From the evaluation results, the Adam optimizer with a learning rate of 0.01 provided the best performance with the lowest Mean Squared Error (MSE) on the test data, which was 0.1766. This system is able to provide relevant recommendations based on user preferences as reflected in their historical reviews of previous tourist attractions, covering various types of tourism that suit individual preferences.

REFERENCES (10PT, IEEE STYLE)

[1] W. Yudananto, S. S. Remi, and B. Muljarijadi, "Peranan Sektor Pariwisata Terhadap Perekonomian Daerah di Indonesia (Analisis Interregional Input-Output)," Jurnal, vol. 2, no. 4, 2012, Universitas Padjajaran, Bandung.

[2] U. Soebiyantoro, "Pengaruh Ketersediaan Sarana Prasarana, Sarana Transportasi Terhadap Kepuasan Wisatawan," Jurnal Manajemen Pemasaran, vol. 4, no. 1, pp. 16–22, 2009

[3] Badan Pusat Statistik, Kajian Dampak Pariwisata Terhadap Perekonomian Provinsi Jawa Tengah, Badan Pusat Statistik Provinsi Jawa Tengah, 2022.

[4] M. H. Satman and M. Altunbey, "Selecting Location of Retail Stores Using Artificial Neural Networks and Google Places API," International Journal of Statistics and Probability, vol. 3, no. 1, p. 67, 2014.

[5] R. K. Silviana, A. Nazir, E. Budianita, F. Syafria, and S. K. Gusti, "Pengklasteran Risiko COVID-19 di Riau Menggunakan Teknik One Hot Encoding dan Algoritma K-Means Clustering," Jurnal Informasi dan Komputer, vol. 10, no. 1, pp. 154–163, 2022.

[6] Z. Nabi, *Pro Spark Streaming: The Zen of Real-Time Analytics Using Apache Spark*, Apress, 2016.

[7] T. T. Hanifa, S. Al-Faraby, and F. Informatika, "Analisis Churn Prediction pada Data Pelanggan PT. Telekomunikasi dengan Logistic Regression dan Underbagging," vol. 4, pp. 3210–3225, 2017.

[8] L. V. Fausett, *Fundamentals of Neural Networks: Architectures, Algorithms and Application*, Pearson Education India, 2006.

[9] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," *arXiv preprint arXiv:1811.03378*, 2018.

[10] S. T. Alexander, "The Mean Squared Error (MSE) Performance Criteria," in *Adaptive Signal Processing*, Texts and Monographs in Computer Science, Springer, New York, NY, 1986.

[11] A. Muhaimin, D. D. Prastyo and H. Horng-Shing Lu, "Forecasting with Recurrent Neural Network in Intermittent Demand Data," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 802-809, doi: 10.1109/Confluence51648.2021.9376880.

Jurnal Aplikasi Sains Data

Journal homepage: https://jasid.upnjatim.ac.id/index.php/jasid

Application of Fuzzy Inference System for Quality Assessment of Formula Milk for Pregnant Women in Stunting Program

Wa Fijriyani R. Ganisi[1,*]

[1,*]Mathematics Department Universitas Pattimura

[1*]fijriyaniramliganisi@gmail.com

ABSTRACT

Abstract— Stunting remains a significant global public health challenge, affecting more than 149 million children under five years of age worldwide as reported by the United Nations in 2020. Indonesia alone accounts for approximately 6.3 million stunted children, highlighting the urgent need for effective intervention strategies. Stunting is primarily caused by chronic malnutrition during the first 1,000 days of life, which includes inadequate nutritional intake during pregnancy, poor infant feeding practices, and environmental factors such as inadequate sanitation. The 2022 Indonesian Nutrition Status Survey (SSGI) indicated a stunting prevalence of 21.6%, showing improvement from 24.4% in 2021, yet still significantly above the national target of 14% set for 2024. Given the critical role of maternal nutrition in reducing stunting risk, providing pregnant women with appropriate nutritional guidance is essential. This study aims to develop a decision support model using a Fuzzy Inference System (FIS) to assist pregnant women in selecting the most suitable formula milk based on nutritional value and affordability. The Mamdani FIS method was applied to analyze data from eight commercially available formula milk products. The evaluation measured the membership degrees corresponding to recommendation levels, factoring in both price and nutrition. The results identified Anmum Materna as the most favorable option, with a membership degree of 0.937, classified under the "Highly Recommended" category. This formula is priced at IDR 70,000 and contains a total nutritional value of 1024 grams, offering a balance of quality and affordability. This model demonstrates potential as a practical tool to support informed nutritional choices during pregnancy, contributing to stunting prevention efforts.

Keywords: *fuzzy, mamdani, maternal, milk, nutrition, stunting.*

I. INTRODUCTION

Based on UN 2020 data, more than 149 million children under five in the world are stunted, with 6.3 million of them in Indonesia. According to UNICEF, factors such as malnutrition in children in their first two years of life, lack of nutrition in mothers during pregnancy, and poor sanitation contribute to stunting. This highlights the urgency of addressing stunting with a holistic approach involving nutrition interventions, maternal and child health services, and improved sanitation to ensure optimal child growth and development [5].

Indonesia's high stunting rate, reaching 21.6% according to the 2022 Survey on the Status of Nutrition in Indonesia (SSGI), highlights the serious challenges in improving children's well-being. Despite a decrease from the previous year, which reached 24.4% in 2021, the rate is still far from the national

* *Corresponding author.*
E-mail address: author@institute.xxx

https://doi.org/10.33005/jasid.v1i1.8

target set at 14% by 2024. Stunting, which is a condition of failure to thrive in children, not only impacts physical height, but also has the potential to cause long-term health problems, including stunted cognitive development and weakened immunity [6].

The importance of fulfilling maternal nutrition since early pregnancy is an important highlight in stunting prevention. The prevalence of stunting in newborns, at 18.5% with a body length of less than 48 cm, shows that this condition can occur even before birth. Therefore, greater attention to maternal nutrition before and during pregnancy is crucial. A better understanding of the importance of adequate nutrition, easy access to quality maternal health services, and support in implementing healthy nutrition practices are important steps in reducing the prevalence of stunting and creating a healthier generation with more potential [6].

The pregnancy period affects the quality of future human resources because the condition of the fetus in the womb greatly affects its development [7]. The first trimester requires 2485 Kcal per day, while a normal adult woman only needs 2200 Kcal. The second and third trimesters require an additional 285 Kcal per day. The diet of pregnant women should follow the 4 healthy 5 perfect formula with additional consumption of milk as a source of animal protein that is rich in nutrients and easily digested. Fulfillment of calcium through milk as recommended is the best choice [8].

One important aspect of pregnant women's nutritional intake is milk consumption, which is a source of essential nutrients such as protein, calcium, and vitamins. However, the quality assessment of milk used by pregnant women today is often conventional and inaccurate [9]. This is where the importance of applying the fuzzy inference system method in assessing the quality of milk for pregnant women as part of the strategy to support the stunting prevention program.

The fuzzy inference system method allows the use of a more flexible and adaptive approach in assessing milk quality based on various complex and often ambiguous factors [10]. By utilizing data on the nutritional content of milk, the health condition of pregnant women, and other relevant factors, fuzzy inference systems can produce more accurate and reliable milk quality assessments..

II. RESEARCH METHODS

This research focuses on the mathematical analysis of dairy products using fuzzy logic method. Within the framework of this research, data was collected from infant formula products available in supermarkets and minimarkets in the research locations. In addition, data was also collected online from various sources such as e-commerce sites and official websites of manufacturers of infant formula. A total of 8 samples of infant formula, namely:

1. Anmun Materna
2. Lactamil
3. Frisomum Gold
4. SGM Bunda
5. Prenagen Emesis
6. Frisian Flag Mama
7. Vidoran Ibunda
8. Enfamama A+

The data analyzed included the content of nutrients such as protein (P), calcium (K), vitamin B9 (folic acid), iron, vitamin A, vitamin B6, vitamin B12, vitamin C, and vitamin D, which were then accumulated into total nutrients (NT).

This research method is an applied research that aims to build a model for determining the best formula milk drink for pregnant women based on price and nutrition variables. The steps used in this research include the fuzzification process, determination of fuzzy rules, fuzzy inference with the Mamdani method, and the defuzzification process.

1. Determine Input and Output Variables
   The first step is to determine the variables that will be used as input and output in the fuzzy model. Input variables may include price and nutritional content such as calories, fat, protein, and others, while the output variable is a recommendation for the quality of infant formula for pregnant women.
2. Define the Universal Set of Inputs and Outputs

After defining the input and output variables, the next step is to define the universal set for each variable. This includes determining the range of values that each input and output variable can take.

3. Fuzzification
   The fuzzification process involves converting the numerical values of the input variables into fuzzy values. This is done by using membership functions to convert numerical data into membership levels in fuzzy sets.
4. Determining Fuzzy Rules for Milk Selection
   The next step is to formulate fuzzy rules that will be used to determine recommendations for pregnant women's formula. These rules connect input variables with outputs based on certain knowledge and logic.
5. Determining Fuzzy Inference
   In the fuzzy inference stage, predefined rules are applied to obtain fuzzy results from the fuzzified inputs. This process uses fuzzy logic mechanisms to generate fuzzy outputs.
6. Defuzzification
   The final step is defuzzification, which converts the fuzzy results back into numerical values that can be interpreted as concrete recommendations. This process produces an output value that provides a recommendation for the best maternity formula based on fuzzy logic analysis.

This study analyzes price data and nutritional content of 8 samples of formula milk products for pregnant women. The goal is to produce optimal formula recommendations for pregnant women using fuzzy logic. Data analysis is carried out with the Mamdani model using the Fuzzy Inference System (FIS) in the MATLAB application.

III. RESULTS AND DISCUSSIONS

The nutrient content and price data of each sample are shown in Table 1 below:

Tabel 3. 1 Price Data and Nutricient Content of Samples

No	Price IDR	P (g)	K (g)	Vit B9 (g)	Iron (g)	Vit A (g)	Vit B6 (g)	Vit B12 (g)	Vit C (g)	Vit D (g)	NT
1	70.000	44	120	180	100	100	140	160	140	40	1024
2	87.000	48	100	120	100	80	120	120	120	40	848
3	95.000	52	80	80	100	80	100	100	140	40	772
4	35.000	33	60	60	75	60	60	90	75	24	537
5	86.625	44	140	180	80	100	160	120	120	40	984
6	45.000	22	60	40	50	40	60	60	70	70	472
7	35.000	31,5	70	122,5	105	70	105	70	105	35	714
8	135.890	52	100	100	80	80	120	80	120	24	756

This research uses the Mamdani fuzzy approach to analyze the data. This process involves five main steps: fuzzy set identification, fuzzification, fuzzy rule formation, inference, and defuzzification. The initial data was presented in percentage format. To convert percentages to grams, the percentage number is divided by one hundred and the result is multiplied by the weight of the dairy product in question. With this conversion, the data can be converted into grams, enabling further analysis regarding the optimization of the nutritional content of the dairy products.be identified, as well as understanding how the model adapts to changes in the input data.

3.1 Defining Fuzzy Variables and Value Ranges

Tabel 3. 2 Defining Fuzzy Variables and Value Ranges

Function	Variable Name	Conversation Set
Input	Harga	[0 140000]
	NT	[0 1025]
Output	TK	[0 1]

3.2 Perform Fuzzification

This process involves determining and calculating the membership function for each variable based on the collected data. For example, the price variable is divided into five membership levels: very cheap (SU), cheap (MU), medium (S), expensive (MA), and very expensive (SM). The same is true for the variables of total nutrition and recommendation level, each of which is classified into several membership levels according to the relevant categories.

Price fuzzy variable

$$\mu_{(SU)} = \begin{cases} \left(\frac{30.000-x}{30.000-0}\right); & 0 \leq x \leq 30.000 \\ 0; & x \geq 30.000 \end{cases}$$

$$\mu_{(MU)} = \begin{cases} 0\,; & x < 20.000 \; atau \; x > 60.000 \\ \left(\frac{x-20.000}{30.000-20.000}\right); & 20.000 \leq x \leq 30.000 \\ 1\,; & 30.000 \leq x \leq 50.000 \\ \left(\frac{60.000-x}{60.000-50.000}\right); & 50.000 \leq x \leq 60.000 \end{cases}$$

$$\mu_{(S)} = \begin{cases} 0\,; & x < 50.000 \; atau \; x > 90.000 \\ \left(\frac{x-50.000}{60.000-50.000}\right); & 50.000 \leq x \leq 60.000 \\ 1\,; & 60.000 \leq x \leq 80.000 \\ \left(\frac{90.000-x}{90.000-80.000}\right); & 80.000 \leq x \leq 90.000 \end{cases}$$

$$\mu_{(MA)} = \begin{cases} 0\,; & x < 80.000 \; atau \; x > 120.000 \\ \left(\frac{x-80.000}{90.000-80.000}\right); & 80.000 \leq x \leq 90.000 \\ 1\,; & 90.000 \leq x \leq 110.000 \\ \left(\frac{120-x}{120-110}\right); & 110.000 \leq x \leq 120.000 \end{cases}$$

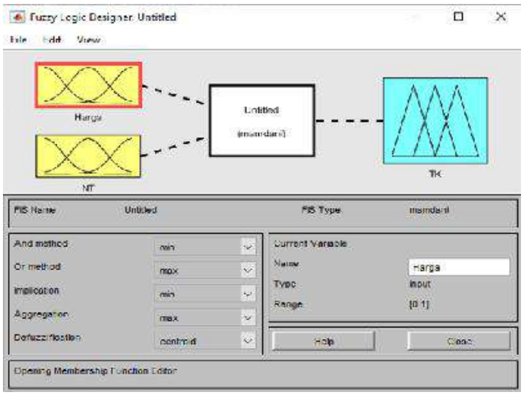
$$\mu_{(SM)} = \begin{cases} 0\,; & x < 110.000 \\ \left(\frac{x-110.000}{120.000-110.000}\right); & 110.000 \leq x \leq 120\,.000 \\ 1\,; & 120.000 \leq x \leq 140.000 \end{cases}$$

3.3 Implementation of Fuzzy Logic FIS with MATLAB

In this research, a method to determine the appropriate formula for pregnant women by considering the balance between price and nutrition will be examined. The approach used is fuzzy logic with Fuzzy Inference System (FIS) implemented using MATLAB software.

The steps to be executed are as follows:

3.3.1. Input and Output Identification

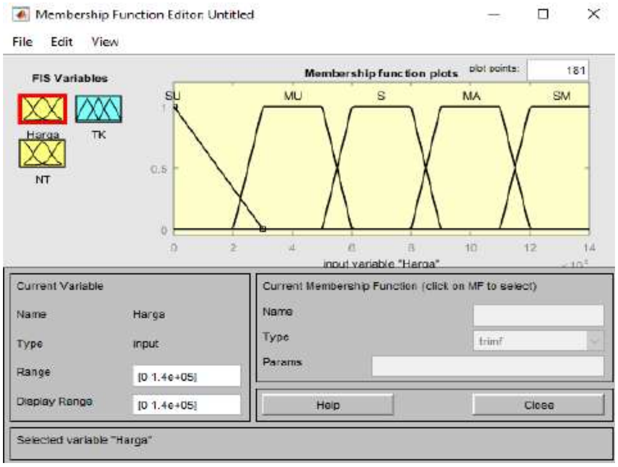


FIS Name: Untitled
FIS Type: mamdani
And method: min
Or method: max
Implication: min
Aggregation: max
Defuzzification: centroid
Current Variable
Name: Harga
Type: input
Range: [0 1]
Opening Membership Function Editor

Figure 3 1 Determination of Input and Output Parameters in FIS Editor

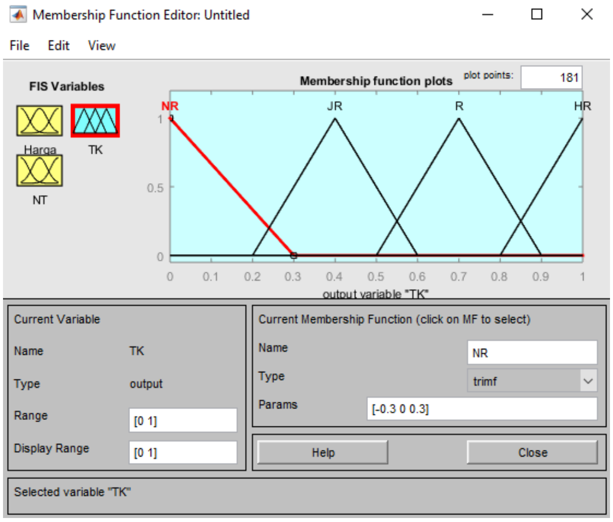
3.3.2. Fuzzification

The next step in this process is fuzzification to define fuzzy sets for the parameters of price, nutrition, and recommendation of infant formula for pregnant women. The value range for price is [0 - 140,000], while nutrition has a range of [0 - 1025] and the recommendation level has a range of [0-1]. This fuzzification is done to establish the membership level of each input and output, which is then implemented in MATLAB to produce graphs that demonstrate the fuzzification process visually.



plot points: 181
Current Variable
Name: Harga
Type: input
Range: [0 1.4e+05]
Display Range: [0 1.4e+05]
Current Membership Function (click on MF to select)
Name:
Type: trimf
Params:
Selected variable "Harga"

Figure 3 2 Price Membership Level Chart



Current Variable

Name TK

Type output

Range [0 1]

Display Range [0 1]

Current Membership Function (click on MF to select)

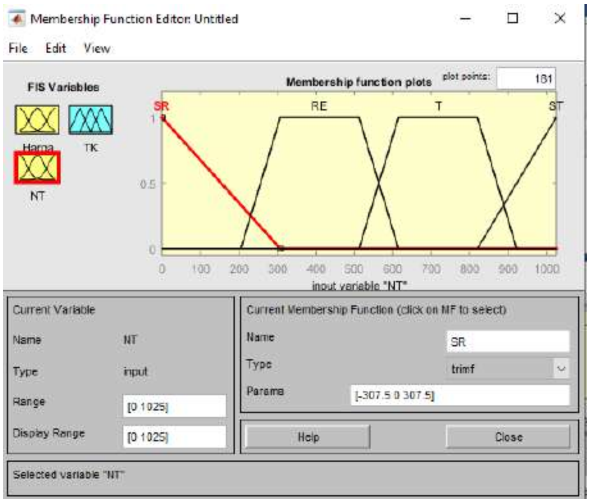
Name NR

Type trimf

Params [-0.3 0 0.3]

Selected variable "TK"

Figure 3 3 NT Membership Grade Chart



Current Variable

Name NT

Type input

Range [0 1025]

Display Range [0 1025]

Current Membership Function (click on MF to select)

Name SR

Type trimf

Params [-307.5 0 307.5]

Selected variable "NT"

Figure 3 4 TK Membership Grade Chart

The next step is to set up fuzzy rules to determine milk formula recommendations for pregnant women. For sample 1 in table 1, with a price of Rp 70,000 and NT 1024, we find the membership level of milk recommendation, namely μ(x) and μ(y), with x=1024 and y=70,000. The calculation result shows

$$\mu_{SR}(1024) = 0$$
$$\mu_{R}(1024) = 0$$
$$\mu_{T}(1024) = 0$$

$$\mu_{SU}(70.000) = 0$$
$$\mu_{MU}(70.000) = 0$$
$$\mu_{S}\,(70.000) = 1$$

$\mu_{ST}(1024) = \frac{1024-820}{1025-820} = 1$

$\mu_{MA}(70.000) = 0$
$\mu_{SM}(70.000) = 0$

Furthermore, calculating the predicate of each rule with the Min implication function in sample 1 of table 1, with μ_ST (1024)=1; μ_S (70,000)=1 is as follows:

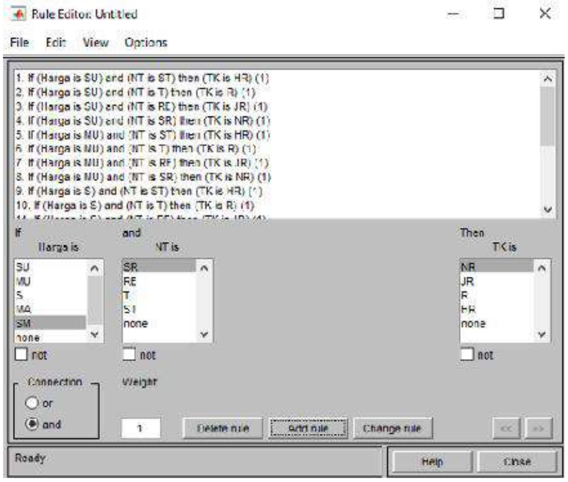
$$\alpha_1 = min[\mu_S(70.000), \mu_{ST}(1024)] = min[1\,; 1] = 1$$

3.3.3. Fuzzy Rule Formation

Based on the rules that have been created :

- [R1] IF the price is very cheap (SU) AND the nutritional content is very high (ST) THEN highly recommended (HR).
- [R2] IF the price is very cheap (SU) AND the nutritional content is high (T) THEN recommended (R).
- [R3] IF the price is very cheap (SU) AND the nutritional content is low (RE) THEN less recommended (LR).
- [R4] IF the price is very cheap (SU) AND the nutritional content is very low (SR) THEN not recommended (NR).
- R5] IF the price is low (MU) AND the nutritional content is very high (ST) THEN highly recommended (HR).
- R6] IF low price (MU) AND high nutritional content (T) THEN recommended (R).
- R7] IF low price (MU) AND low nutritional content (RE) THEN less recommended (LR).
- R8] IF low price (MU) AND very low nutritional content (ST) THEN not recommended (NR).
- R9] IF medium price (S) AND very high nutritional content (ST) THEN highly recommended (HR).
- [R10] IF the price is medium (S) AND the nutritional content is high (T) THEN recommended (R).
- R11] IF price is medium (S) AND nutritional content is low (RE) THEN less recommended (LR).
- [R12] IF price is medium (S) AND nutritional content is very low (SR) THEN not recommended (NR).
- R13] IF the price is expensive (MA) AND the nutritional content is very high (SR) THEN recommended (R).
- [R14] IF the price is expensive (MA) AND the nutritional content is high (T) THEN recommended (R).
- [R15] IF the price is expensive (MA) AND the nutrient content is low (RE) THEN not recommended (LR).
- [R16] IF the price is expensive (MA) AND the nutritional content is very low (SR) THEN not recommended (NR).
- R17] IF the price is very expensive (SM) AND the nutritional content is very high (ST) THEN recommended (R).
- R18] IF the price is very expensive (SM) AND the nutritional content is high (T) THEN recommended (R).
- R19] IF the price is very expensive (SM) AND the nutritional content is low (RE) THEN not recommended (LR).

- [R20] IF the price is very expensive (SM) AND the nutrient content is very low (SR) THEN not recommended (NR).



1. If (Harga is SU) and (NT is ST) then (TK is HR) (1)
2. If (Harga is SU) and (NT is T) then (TK is R) (1)
3. If (Harga is SU) and (NT is RE) then (TK is JR) (1)
4. If (Harga is SU) and (NT is SR) then (TK is NR) (1)
5. If (Harga is MU) and (NT is ST) then (TK is HR) (1)
6. If (Harga is MU) and (NT is T) then (TK is R) (1)
7. If (Harga is MU) and (NT is RE) then (TK is JR) (1)
8. If (Harga is MU) and (NT is SR) then (TK is NR) (1)
9. If (Harga is S) and (NT is ST) then (TK is HR) (1)
10. If (Harga is S) and (NT is T) then (TK is R) (1)

Figure 3 5 Fuzzy Rules

3.3.4. Defuzzification

The last step is the defuzzification process, where the membership degree of each sample is calculated based on the nutritional content and price. By entering the nutritional content and price values of the first sample, the membership degree in the recommendation column is obtained as shown in Figure 6. This result shows the membership degree for the recommendation in the first sample.

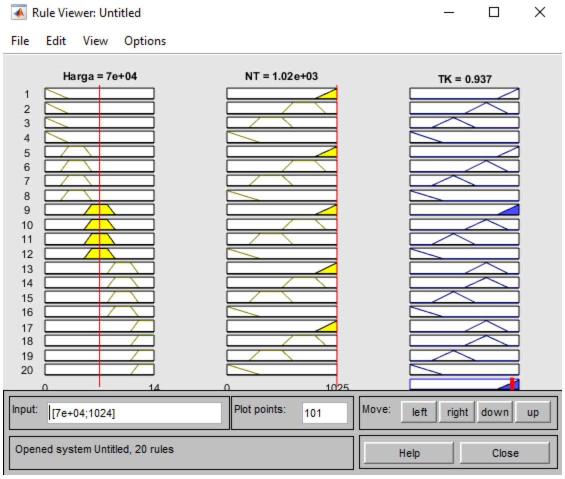


Figure 3 6 Memberhip Degree Canculation of Sample 1 Recommendation

Using the price and total nutritional value of sample 1 [70,000; 1024], a substitution is made in the input section. As a result, the membership degree for the recommendation of pregnant women's milk in sample 1 is 0.937. The same step is repeated for the other 7 samples of infant formula. The membership degree results of the 8 formula milk samples are organized in Table 4.3 based on the order of recommendation.

Tabel 3. 3 Membership Degree Sequence of Each Formula Milk Sample

No	Price (Rp)	NT(g)	Orde	Category
1	70.000	1024	0.937	HR
5	86.625	984	0.748	R
2	87.000	848	0.72	R
7	35.000	714	0.7	R
3	95.000	772	0.7	R
8	135.890	756	0.7	R
4	35.000	537	0.49	LR
6	45.000	472	0.4	LR

Description:

NR : Not recommended

LR : Less recommended

R : Recommended

HR : highly recommended.

The table above presents the results of the fuzzy logic analysis of 8 samples of infant formula for pregnant women, evaluated based on price and total nutritional value (NT). Focus was given to samples with high membership degrees and affordable prices. Anmum Materna formula milk has the highest membership degree (0.937) with a price of IDR 70,000 and NT of 1024 grams, belonging to the "Highly Recommended" category. There are also other recommended samples, such as samples 5 and 2, although at a slightly higher price.

Sample No. 7 has a lower price (IDR 35,000) because the package is 350 grams, which is smaller than the general size (400 grams). Nonetheless, this sample is still recommended with a degree of membership of 0.7, indicating this product is still in the recommended category. The low price does not lower the quality of the recommendation based on the calculated membership degree.

CONCLUSION

Based on the results obtained with the fuzzy logic method on 8 samples of formula milk for pregnant women, sample 1, Anmum Materna, stands out as the top choice with the highest membership degree of: 0.937 at a price of Rp 70,000 with a total nutrition of 1024 grams, falling into the “Highly Recommended” category. It was also found that other samples, namely 5 (Prenagen emesis) and 2 (Lactamil) were also recommended with fairly high membership degrees of 0.748 and 0.72, albeit at slightly higher prices of Rp. 86,625 and Rp. 87,000, respectively. Sample No. 7 (Vidoran Ibunda), with a lower price of IDR 35,000 and 350 gram packaging, is still recommended with a membership degree of 0.7, indicating good quality relative to its low price. Thus, Anmum Materna is the top choice for finding the best pregnancy milk with affordable price and high nutrition.

REFERENCES

[1] A. Khoirun Nisa, M. Abdy, dan Ahmad Zaki, J. Matematika, and F. Matematika dan Ilmu Pengetahuan Alam, “Penerapan Fuzzy Logic untuk Menentukan Minuman Susu Kemasan Terbaik dalam Pengoptimalan Gizi,” 2020. [Online]. Available: http://www.ojs.unm.ac.id/jmathcos

[2] N. Nurhidayah, Y. A. Lesnussa, and Z. A. Leleury, “FUZZY LOGIC APPLICATION ON

EMPLOYEE ACHIEVEMENT ASSESSMENT (CASE STUDY: EDUCATION QUALITY ASSURANCE INSTITUTE OF MALUKU PROVINCE),” BAREKENG: Jurnal Ilmu Matematika dan Terapan, vol. 16, no. 3, pp. 877–886, Sep. 2022, doi: 10.30598/barekengvol16iss3pp877-886.

[3] A. Kamsyakawuni, A. Riski, and A. B. Khumairoh, “APPLICATION FUZZY MAMDANI TO DETERMINE THE RIPENESS LEVEL OF CRYSTAL GUAVA FRUIT,” BAREKENG: Jurnal Ilmu Matematika dan Terapan, vol. 16, no. 3, pp. 1087–1096, Sep. 2022, doi: 10.30598/barekengvol16iss3pp1087-1096.

[4] R. Rumfot, Y. A. Lesnussa, and D. L. Rahakbauw, “PERBANDINGAN METODE FUZZY MAMDANI, SUGENO DAN TSUKAMOTO UNTUK MENENTUKAN JUMLAH PRODUKSI BATU PECAH,” 2024.

[5] Kementerian Pendidikan dan Kebudayaan., “14,9 Juta Anak di Dunia Alami Stunting Sebanyak 6,3 Juta di Indonesia, Wapres Minta Keluarga Prioritaskan Kebutuhan Gizi.” [Online]. Available: https://paudpedia.kemdikbud.go.id/berita/149-juta-anak-di-dunia-alami-stunting-sebanyak-63-juta-di-indonesia-wapres-minta-keluarga-prioritaskan-kebutuhan-gizi?do=MTY2NC01YjRhOGZkNA==&ix=MTEtYmJkNjQ3YzA=

[6] A. Muhaimin and K. Fithriasari, "Kohonen-SOM LOF Approach for Anomaly Detection," 2021 IEEE 7th Information Technology International Seminar (ITIS), Surabaya, Indonesia, 2021, pp. 1-6, doi: 10.1109/ITIS53497.2021.9791596.

[7] Kementerian Kesehatan Republik Indonesia., “Panduan Hari Gizi Nasional ke-64 tahun 2024.” Accessed: Apr. 30, 2024. [Online]. Available: https://ayosehat.kemkes.go.id/panduan-hari-gizi-nasional-ke-64-tahun-2024

[8] Kementerian Kesehatan Republik Indonesia, “Penyebab stunting anak. Sehat Negeriku.” Accessed: May 30, 2024. [Online]. Available: https://sehatnegeriku.kemkes.go.id/baca/umum/20180524/4125980/penyebab-stunting-anak/.

[9] A. Muhaimin, D. D. Prastyo and H. Horng-Shing Lu, "Forecasting with Recurrent Neural Network in Intermittent Demand Data," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 802-809, doi: 10.1109/Confluence51648.2021.9376880.

[10] A. Muhaimin, W. Wibowo, and P. A. Riyantoko, “Multi-label Classification Using Vector Generalized Additive Model via Cross-Validation”, *JICT*, vol. 22, no. 4, pp. 657–673, Oct. 2023.

Jurnal Aplikasi Sains Data

Journal homepage: https://jasid.upnjatim.ac.id/index.php/jasid

Comparative Analysis of Stochastic Gradient Descent Optimization and Adaptive Moment Estimation in Emotion Classification from Audio Using Convolutional Neural Network

Aldelia J. Tutuhatunewa[1], Dorteus L. Rahakbauw [2], and Zeth A. Leleury [3,*]

[1]Department of Mathematics, Pattimura University, [2] Department of Mathematics, Pattimura University, [3] Department of Mathematics, Pattimura University

[1]aldeliajoe@gmail.com, [2]dorteus.rahakbauw@lecturer.unpatti.ac.id, [3,*]zeth.arthur@lecturer.unpatti.ac.id

ABSTRACT

Abstract— Emotion is a fundamental aspect of human life that profoundly shapes behavior, social interactions, and decision-making processes. The ability to effectively communicate and foster mutual understanding between individuals relies heavily on accurately recognizing and expressing emotions. Among various channels of emotional expression, sound stands out as a powerful and direct medium that reflects and conveys human emotional states. This makes audio-based emotion recognition a critical and rapidly evolving field of study. With the rapid advancements in information technology and artificial intelligence, research focused on recognizing emotions through sound signals has gained significant momentum. Machine learning algorithms, particularly deep learning models like neural networks, have demonstrated remarkable capabilities in identifying and classifying emotions expressed through multiple modalities such as text, images, videos, and especially audio signals. Within the family of neural networks, Convolutional Neural Networks (CNNs) have been especially effective for audio emotion classification, due to their strength in extracting hierarchical and spatial features directly from raw input data. This study specifically investigates the comparative effectiveness of two popular optimization algorithms—Stochastic Gradient Descent (SGD) and Adaptive Moment Estimation (Adam)—in training CNN models for emotion classification from audio recordings. Utilizing the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset, experimental results indicate that CNNs trained with the SGD optimizer achieve an overall accuracy of 53%, surpassing the 48% accuracy achieved by Adam. These results underscore the potential advantages of SGD in fine-tuning deep learning models for audio-based emotion recognition. Consequently, researchers and practitioners are encouraged to consider SGD optimization to improve the performance and robustness of emotion classification systems based on audio data.

Keywords: Convolutional Neural Network (CNN), Audio classification, Mel Frequency Cepstral Coefficients (MFCC), Stochastic Gradient Descent (SGD), Adaptive Moment Estimation (Adam)

I. INTRODUCTION

Emotions are a fundamental aspect of human life that influence behavior, social interactions and decision-making. Successful communication and understanding between individuals largely depend on our ability to recognize and express emotions. In this domain, voice or audio plays a key role as a medium that reflects and expresses human emotions.

* *Corresponding author.*

E-mail address: zeth.arthur@lecturer.unpatti.ac.id

https://doi.org/10.33005/jasid.v1i1.5

In the era of information technology and artificial intelligence, emotion recognition through voice has become a growing research focus. Machines and machine learning models, especially neural networks, can be taught to understand and classify emotions contained in text, images, video, and audio. Three algorithms commonly used for image classification include Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Convolutional Neural Network (CNN). Among the three, CNN has the highest accuracy in model performance with parameter values Accuracy 0.942, Precision 0.943, Recall 0.942, and F1-Score 0.942 [1].

In building CNN models, the Optimizer function is often used to maximize model performance. An optimizer is an algorithm or method used to minimize or maximize the objective function during the training process. Many optimizers are used to manage the learning rate which is a parameter that regulates how much change will be applied to the weights during the training process. Some optimizers that are often used for CNN algorithms include Adam [2], Stochastic Gradient Descent (SGD) [3] [4], and Root Mean Square Propagation (RMSProp) [5]. Emotion classification in audio has been done using the Multilayer Perceptron method which classifies 8 types of emotions achieving an average accuracy of 96% [6]. However, this research only uses the Adam Optimizer.

Based on the background above, in this final project research, the researcher takes the research title "Comparative Analysis of Stochastic Gradient Descent Optimization and Adaptive Moment Estimation in Emotion Classification from Audio Using Convolutional Neural Network

II. RESEARCH METHODS

A. Data

This study utilizes two main data sources, namely tourist attraction data in Central Java, and User data that has provided reviews of tourist attractions in Central Java obtained through the Google Places API, which is an Application Programming Interface that provides location-based geographic data via the internet using HTTP [4]. Tourist attraction data in Central Java was collected using the midpoint coordinates of each city/district as a search reference. The search was conducted within a radius of 100,000 meters from the midpoint with search parameters set to obtain places with the category "tourist_attraction". The data taken initially amounted to 1400 tourist attractions, then filtered into 432 tourist attractions in Central Java based on 13 attributes used. The tourist attraction attributes used in this study can be seen in Table I.

TABLE I. DATA ATRIBUTE

ID	Attribute	Description	Data Type
1	Place_id	Unique ID of Tourist Attraction	*Numeric*
2	Place Name	Name of Tourist Attraction	*String*
3	formatted_phone_number	Phone Number of Tourist Attraction	*String*
4	formatted_address	Address of Tourist Attraction	*String*
5	city	City	*String*
6	website	Website	*String*
7	rating	Rating of Tourist Attraction	*Numeric*
8	user_ratings_total	Total Ratings	*Numeric*
9	photo_preference	Photo of Tourist	*String*
10	lng	Longitude	*Numeric*
11	url	Google Maps Website	*String*
12	url_photo	Google Maps Photo	*String*
13	type_tourist	Type of Tourist Attraction	*String*

B. Research Stagees

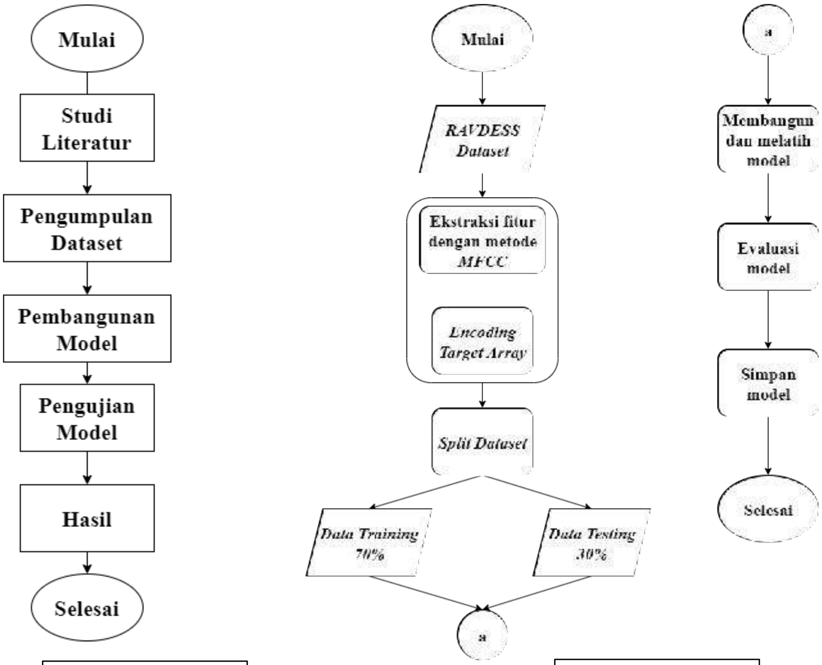


Figure 1a

Figure 1b

Figure 1. Flowchart of Research: (1a) Flowchart of Research Stages (1b) Flowchart of Model Building

This research is experimental research that compares two types of optimizers. Experimental research is used to test causal hypotheses about the relationship between two or more variables. The researcher will conduct a series of experiments to collect data, implement the model, and analyze the experimental results to draw conclusions about the effectiveness of two optimization techniques in each context.

The research phase begins with a data collection process conducted through the open-source platform Kaggle. The dataset used is RAVDESS [7], which is a collection of audio data containing expressions of emotion in the form of speech. This dataset is downloaded from the link available at kaggle.com.

Data pre-processing is performed to prepare the data for use by the model. This includes audio feature extraction using the Mel-Frequency Cepstral Coefficients (MFCCs) technique to convert the audio signal into a numerical representation that can be processed by the Convolutional Neural Network (CNN) model. Each audio data is also assigned a corresponding emotion label, and a label encoding process is performed to convert categorical labels into numerical form, namely 0 for neutral, 1 for calm, 2 for happy, 3 for sad, 4 for angry, 5 for fear, 6 for disgust, and 7 for surprise. The dataset is then divided into two subsets, namely 70% training data and 30% testing data.

CNN architecture specifically designed for emotion classification from audio data was built. This architecture consists of several layers, including convolution, pooling, and fully connected layers. The model training process is performed using a subset of training data and two types of optimizers, namely Stochastic Gradient Descent (SGD) and Adam. Training parameters such as learning rate and batch size are also adjusted to optimize model performance.

After the training process is complete, the model is evaluated using a subset of testing data with evaluation metrics such as training accuracy and validation accuracy. The last stage is result analysis, which compares the model performance between the use of SGD and Adam optimizers. Prior to the entire process, the dataset is first cleaned through incremental pre-processing, where data is selected and truncated using the MATLAB application. Data that is corrupted or defective after this process will be eliminated and not used in the study.

Analysis of the results is done by evaluating the model to determine the level of performance of the model in classifying various classes. One of the model evaluation techniques is to use confusion matrix. Confusion matrix is a matrix that shows and compares the actual or true value with the predicted value of the model that can be used to generate evaluation metrics such as Accuracy, Precision, Recall, and F1-Score.

There are 4 values generated in the confusion matrix table, namely True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). TN is the number of negative data detected correctly. FP is negative data but detected as positive data. TP is positive data that is detected correctly. FN is the opposite of True Positive, so it is positive data, but detected as negative data.

TABLE II. TABLE CONFUSION MATRIX

		True Values	
		True	***False***
Prediction	***True***	*TP* *Correct result*	*FP* *Unexpected result*
	False	*FN* *Missing result*	*TN* *Correct absence of result*

Precision is data that is retrieved based on less information. In binary classification, precision can be made equal to the positive predictive value. Precision is a measure of the model's accuracy in identifying positive samples. Precision shows how often the model's positive predictions are correct. The formula for finding precision [8]:

$$Precision = \left(\frac{TP}{(TP+FP}\right) \times 100\% \quad (1)$$

Recall is the successful removal of data relevant to the query. In binary classification, recall is known as sensitivity. The appearance that the retrieved relevant data is agreeing with the query can be seen by recall. Recall is a measure of the model's sensitivity in capturing all positive samples. Recall shows how good the model is at identifying all the samples that are truly positive. The formula for finding Recall [8]:

$$Recall = \left(\frac{TP}{(TP+FN)}\right) \times 100\% \quad (2)$$

F1-Score value or also known as the F Measure is obtained from the results of precision and recall between the predicted categories and the actual categories. F1-Score is a harmonized measure of precision and recall. F1-Score provides an overall picture of the model's performance by considering both. F1-Score provides a balance between precision and recall. The formula for finding F1-Score [8]:

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (3)$$

Accuracy is used to measure the performance of an algorithm in a way that can be interpreted. The accuracy of a model is usually determined after the model parameters and is calculated as a percentage. It is a measure of how accurate the model's predictions are compared to the actual data and accuracy is in train. The formula for finding accuracy [8]:

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (4)$$

II. RESULTS AND DISCUSSIONS

A. Pre-Processing Data

The data used in this study are human voice recordings in English, which have been pre-labeled. The dataset used is the Ryerson Audio-Visual Dataset of Emotional Speech and Song (RAVDESS), which was collected from the Kaggle website. The voice accents in this dataset are from North America. The dataset consists of two types of files, namely files containing individual conversations and files containing songs. In this study, 1440 audio files of recorded conversations were used. The duration of each audio in this dataset ranges from 3.5 to 5 seconds. The recorded conversations in this dataset involve male and female voices, each consisting of 12 individuals. Each audio file in this dataset has a 16 bit, 48kHz format with a .wav file extension.

TABLE III. TOTAL DATA

Emosi	**Pria**	**Wanita**
Netral	48	48
Tenang	96	96
Bahagia	96	96
Sedih	96	96
Marah	96	96
Takut	96	96
Jijik	96	96
Terkejut	96	96
Total	720	720

The audio data then goes through a silent removal process, which is the process of cutting the audio at the beginning and end to remove the silent parts before and after the actor speaks. This process is done with the help of MATLAB software. Sample data results that have been cut or have been done silent removal can be seen in Figure 2 and Figure 3 below.

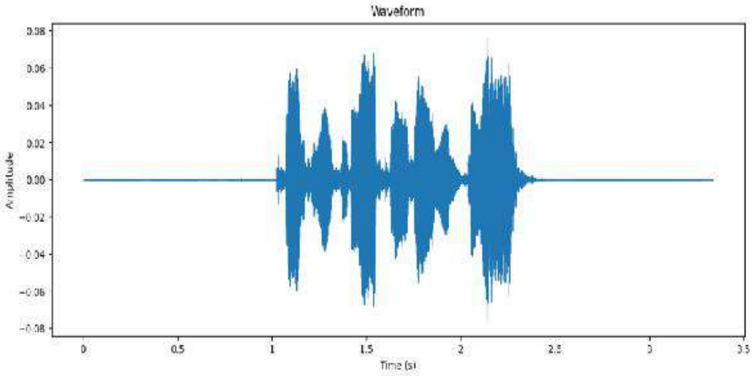


Figure 2. Happy Audio Visualization Before Silent Removal

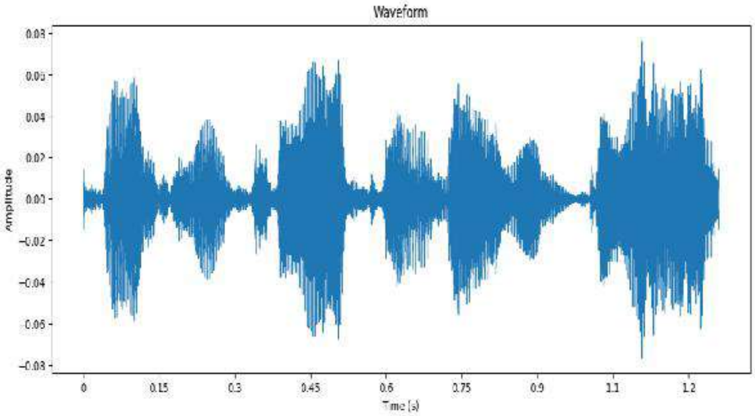


Figure 3. Happy Audio Visualization After Silent Removal

The data that has gone through the cutting process and is used as data in the study can be seen in table 3 below.

TABLE IV. AUDIO COUNT AFTER CUTTING

Kelas	Emosi	Jumlah
1	Netral	96
2	Tenang	190
3	Bahagia	191
4	Sedih	192
5	Marah	192
6	Takut	191
7	Jijik	192
8	Terkejut	191
	Total	1435

Audio signals are obtained with the help of the librosa library with python programming. Then these audio signals are used for the MFCC extraction process. After applying Discrete Cosine Transform (DCT) on the log energy of the Mel filter, each frame generates an MFCC vector. For example, if we use an MFCC count of 13 coefficients, each frame generates 13 MFCC coefficients. Since the audio signal uses MFCC for many frames, the output value of MFCC extraction is a matrix. Where each row represents one frame, and each column represents one MFCC coefficient. For example, for a signal divided into N frames and n_mfcc MFCC coefficients, the MFCC matrix has a size of $N \times n_{mfcc}$.

After the values are known, the data can be presented in the form of a table like the following table 4 and fill the non-numeric values (NaN) or empty values to 0. It can be found that NaN values occur due to the unequal length of the result vector. Replacing it with 0 will indicate that there is no information in the result to avoid loss of information [9].

TABLE V. SAMPLE MFCC EXTRACTION RESULTS FROM AUDIO

	1	**2**	**3**	**4**	**…**	**13**
1	-452,251	197,7236	-7,87114	-19,7107	…	0,542556
2	-458,808	204,3042	-23,309	-20,7977	…	-1,46148
3	-485,871	203,8673	-29,3221	-27,0689	…	-9,04032
4	-503,372	209,1617	-11,0603	-30,559	…	-7,3006
5	-533,569	196,6293	-1,95884	-20,9173	…	-5,19639
…	…	…	…	…	…	…
322	-481,685	148,3774	20,60678	2,111137	…	0,33366

Table V is a sample of data extracted from MFCC measuring 13 x 322. All data is then subjected to padding so that it has the same frame length and so that the size of each data is the same. The data is then divided into training data and testing data with 70% training data and 30% testing data [10].

TABLE VI. SPLIT DATA TRAINING AND TESTING

Kelas	**Emosi**	***Training***	***Testing***
1	Netral	67	29
2	Tenang	133	57
3	Bahagia	133	58
4	Sedih	134	58
5	Marah	134	58
6	Takut	133	58
7	Jijik	134	58
8	Terkejut	133	58
	Total	1001	434

After the data is divided, the data is normalized using the following formula (5).

$$x_{baru} = \frac{x_{lama} - \mu}{\sigma} \quad (5)$$

Where:

x_{baru} : Normalized value

x_{lama} : Old data value (not normalized)

μ : Mean

σ : Standard deviation

B. Convolutional Neural Network (CNN) Architecture Design

At this stage, the architecture design for the CNN model that will be used in the research is carried out. This research will consist of 2 convolution layers with ReLU activation function and followed by a fully connected layer consisting of a flatten layer. After that there is a hidden layer and an output layer. It is in this output layer where the Softmax activation function is used to classify the audio into the 8 existing emotion classes.

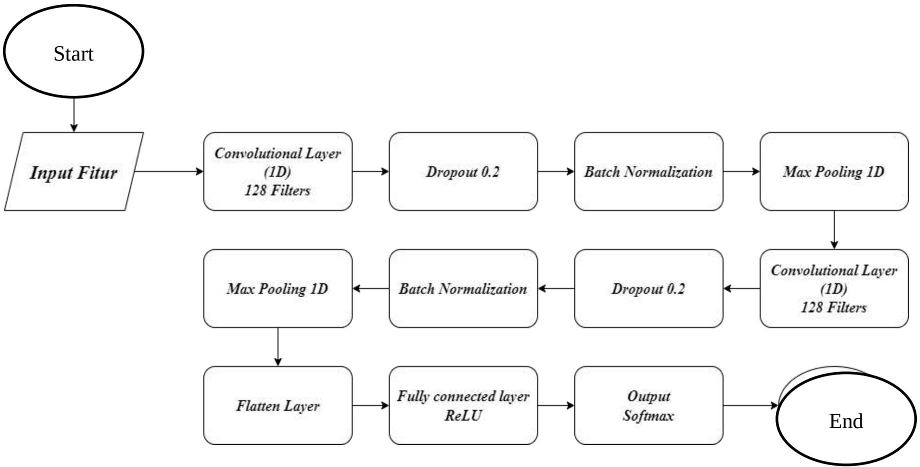


Figure 4. CNN Architecture Diagram

TABLE VII. SUMMARY MODEL CONVOLUTIONAL NEURAL NETWORK (CNN)

Model Sequential			
No.	***Layer (type)***	***Output shape***	***Param#***
1.	Conv1d_128 (Conv1D)	*(None, 4148, 128)*	*128*
2.	Dropout_128 (Dropout)	*(None, 4148, 128)*	*0*
3.	Batch_normalization_128 (BatchNormalization)	*(None, 4148, 128)*	*128*
4.	Max_pooling_128 (MaxPooling1D)	*(None, 2092, 128)*	*0*
5.	Conv1d_129 (Conv1D)	*(None, 2090, 128)*	*3104*
6.	Dropout_129 (Dropout)	*(None, 2090, 128)*	*0*
7.	Batch_normalization_128 (BatchNormalization)	*(None, 2090, 128)*	*128*
8.	Max_pooling_128 (MaxPooling1D)	*(None, 1045, 128)*	*0*
9.	Flatten_64 (Flatten)	*(None, 33440)*	*0*
10	Dense_128 (Dense)	*(None, 64)*	*2140224*
11.	Dense_129 (Dense)	*(None, 8)*	*520*
Total Params	2144232		
Trainable Params	2144104		
Non-trainable Params	128		

The next stage is training the convolutional neural network model with training data. Before the training process is carried out, it is necessary to initialize the parameters that will be used such as loss function, optimizer, number of epochs, number of batches, and callback. The model will go through the training process by being processed in google colab. In addition, 8 emotion classes have been represented into 8 labels representing 8 emotions. 0 = Neutral, 1 = Calm, 2 = Happy, 3 = Sad, 4 = Anger, 5 = Fear, 7 = Disgust, and 8 = Surprise. The following table 7 initializes the parameters for the CNN model.

TABLE VIII. PARAMETER INITIALIZATION

Parameter	Input
Output class	8
Output layer activation	Softmax
Hidden layer activation	ReLU
Epoch	50
Batch size	10 / 32
Optimizer function	SGD / Adam
Loss function	Categorical cross entropy

TABLE IX. ACCURACY DATA OF TRAINING RESULTS AND MODEL VALIDATION WITH SGD OPTIMIZER

Learning Rate	Batch_size = 10		Batch_size = 32	
	Accuracy	Val_ Accuracy	Accuracy	Val_ Accuracy
0.001	0.93	0.47	0.73	0.42
0.002	0.96	0.53	0.84	0.46
0.003	0.97	0.44	0.91	0.42
0.004	0.98	0.47	0.92	0.44
0.005	0.97	0.45	0.90	0.44
0.006	0.97	0.47	0.91	0.45
0.007	0.98	0.48	0.96	0.47
0.008	0.98	0.45	0.95	0.48
0.001	0.99	0.34	0.90	0.47

TABLE X. ACCURACY DATA OF TRAINING RESULTS AND MODEL VALIDATION WITH ADAM OPTIMIZER

Learning Rate	Batch_size = 10		Batch_size = 32	
	Accuracy	Val_ Accuracy	Accuracy	Val_ Accuracy
0.0001	0.91	0.43	0.93	0.44
0.0002	0.89	0.44	0.89	0.48
0.0003	0.96	0.45	0.83	0.40
0.0004	0.85	0.45	0.59	0.32
0.0005	0.49	0.35	0.61	0.32
0.0006	0.45	0.37	0.88	0.43
0.0007	0.92	0.42	0.77	0.38
0.0008	0.81	0.41	0.71	0.39
0.0009	0.73	0.36	0.73	0.40

Based on Table IX and Table X, the yellow cells indicate the optimal parameters for batch size and learning rate for the model used. The basis for this selection is to see the highest validation accuracy (Val_accuracy). For the SGD optimizer, the model works optimally with a learning rate of 0.007 and a batch size of 10. For the adam optimizer, the model works optimally with a learning rate of 0.0002 and a batch size of 10. However, due to the low validation accuracy, the model still needs further review. One of the symptoms of overfitting is the difference between training accuracy and validation accuracy, which can be seen in Table IX and Table X.

C. *Model Analysis with SGD Optimizer*

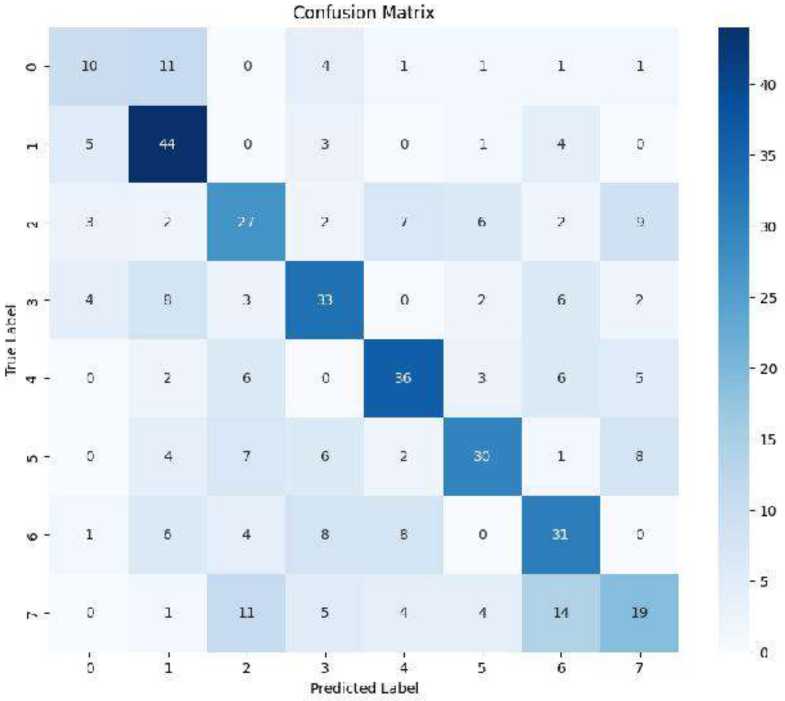


Figure 5. Confusion Matrix Model CNN Optimizer SGD

Based on Figure 4.6 True Labels are the rows of the matrix showing the original labels of the dataset and Predicted Labels are the columns showing the predicted labels made by the model. Diagonal elements (True Positives) (e.g., (0,0), (1,1), (2,2), etc.) represent the number of correct predictions for each class. For example, the cell at position (1,1) indicates that the model correctly predicted class 1 (the second class) 44 times. Then the Non-diagonal Elements (False Positives and False Negatives) represent incorrect predictions. For example, the cell at (1,2) shows the number of instances where class 1 was incorrectly predicted as class 2. The model correctly predicted class 0 for 10 times, but the model incorrectly predicted class 0 as class 1 for 11 times, class 3 for 4 times, and so on. The model correctly predicted class 1 for 44 times, but the model incorrectly predicted class 1 as class 0 for 5 times, class 3 for 3 times, and so on. The same goes for the other rows. Darker colors indicate larger numbers, while lighter colors indicate smaller numbers. The class with the darkest color on the diagonal is class 1 with 44 correct predictions. Next, we will calculate the precision, recall, and f1_score values. From the data above, it is known:

TABLE XI. TP, FP, FN, AND FN VALUES OF 8 EMOTION CLASSES MODEL WITH SGD OPTIMIZER

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
TP	10	44	27	33	36	30	31	19
FP	13	34	31	27	22	17	34	25
FN	19	13	31	25	22	28	27	39
TN	392	343	345	357	354	359	342	351

TABLE XII. PRECISION, RECALL, AND F1_SCORE VALUES WITH SGD OPTIMIZER

Label	Precision	Recall	F1-Score
0	43,47%	52,63%	0,38
1	56,41%	77,19%	0,65
2	46,55%	46,65%	0,47
3	55%	42,3%	0,55
4	62,06%	62,06%	0,62
5	63,82%	51,72%	0,57
6	47,69%	53,44%	0,5
7	43,18%	32,75%	0,37

For the model using the SGD optimizer, Label 1 performed best with an F1-Score of 0.65, precision of 56.41%, and recall of 77.19%. This shows that the model can identify most of the Label 1 samples accurately and consistently.Meanwhile, Label 7 has the lowest performance with F1-Score 0.37, precision 43.18%, and recall 32.75%. This indicates the model struggled to identify and capture samples for this category.

D. Model Analysis with ADAM Optimizer

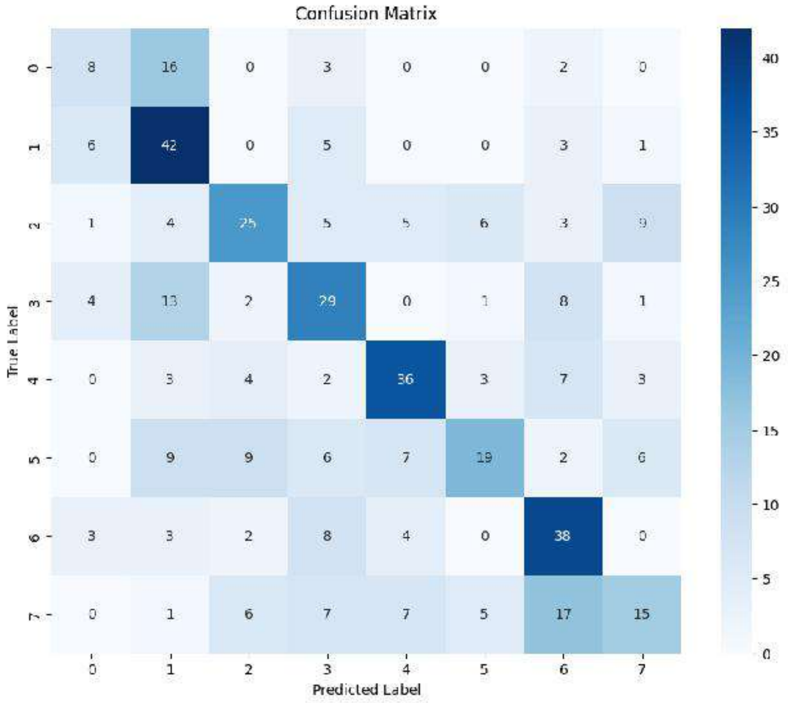


Figure 6. Confusion Matrix Model CNN Optimizer Adam

The cell at position (1,1) shows that the model correctly predicted class 1 (second class) 42 times. Then the Non-diagonal Elements (False Positives and False Negatives) represent incorrect predictions. In row 0 (Original Label = 0) the model correctly predicted class 0 for 8 times, but incorrectly predicted class 0 as class 1 for 16 times, class 3 for 3 times, and so on. Then in row 1 (Original Label = 1), the model correctly predicted class 1 for 42 times, but the model incorrectly predicted class 1 as class 0 for 6 times, class 3 for 5 times, and so on. The model still predicts class 1 more with 42 correct predictions.

Next, the precision, recall, and f1_score values will be calculated. From the data above, it is known:

TABLE XIII. TP, FP, FN, AND FN VALUES OF 8 EMOTION CLASSES MODEL WITH ADAM OPTIMIZER

	0	1	2	3	4	5	6	7
TP	8	42	25	29	36	19	38	15
FP	14	49	23	36	23	15	42	20
FN	21	15	33	29	22	39	20	43
TN	391	328	353	340	353	367	334	356

TABLE XIV. PRECISION, RECALL, AND F1_SCORE VALUES WITH ADAM OPTIMIZER

Label	Precision	Recall	F1-Score
0	36,36%	27,58%	0,31
1	46,15%	73,68%	0,57
2	52,08%	43,1%	0,47
3	44,61%	50%	0,47
4	61,01%	62,06%	0,62
5	55,88%	32,75%	0,41
6	47,5%	65,51%	0,55
7	42,85%	25,06%	0,32

The model using Adam's optimizer shows the best performance in detecting Anger emotion (Label 4), which is characterized by high precision, recall, and F1-Score.This means that the model can well identify and predict this emotion with good accuracy. In contrast, the model had the greatest difficulty in detecting the Neutral emotion (Label 0), which is characterized by the lowest precision and F1-Score values, and the Surprised emotion (Label 7), which has the lowest recall value. This suggests that the model needs further improvement to increase accuracy and consistency in detecting these emotions.

CONCLUSION

Based on the results of the research that has been done, using the Convolutional Neural Network (CNN) model which uses 9 types of learning rates and 2 types of batch sizes, it can be concluded that,

1. Convolutional Neural Network (CNN) training results using the SGD optimizer, have a validation accuracy value of 53% using a learning rate of 0.002 and batch size 10.
2. As for the Adam optimizer, the best training results are at a learning rate of 0.0002 and a batch size of 32. Where the validation accuracy obtained is 48%.
3. The results of the confusion matrix analysis of the CNN model with the SGD optimizer have a higher general accuracy, which is 53% compared to the Adam optimizer which is 48%.

In addition, the results of the training also show that the model is easier to recognize emotions labeled 1, namely audio with calm emotions and the most difficult to recognize audio labeled 0, namely audio neutral emotions. So, in this case of emotion classification from the RAVDESS audio dataset, the better optimizer is the Stochastic Gradient Descent (SGD) optimizer.

REFERENCES

[1] M. F. Naufal, "Analisis Perbandingan Algoritma SVM, KNN, dan CNN untuk Klasifikasi Citra Cuaca," Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 8(2), pp. 311-317, 2021.

[2] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2014.

[3] S. Ruder, "An Overview of Gradient Descent Optimization Algorithms," arXiv, 2016.

[4] L. Bottou, "Large-Scale Machine Learning with Stochastic Gradients Descent," In Proceedings of COMPSTAT'2010, pp. 177-186, 2010.

[5] F. Zou, L. Shen, Z. Jie, W. Zhang and W. Liu, "A Sufficient Condition for COnvergences of Adam and RMSProp," In Proceedings of the IEEE/CVF Confrence on computer vision and pattern recognition, pp. 11127-11135, 2019.

[6] D. Ardiyansyah and J. Jayanta, "Model Klasifikasi Emosi Berdasarkan Suara Manusia dengan Metode Multilayer Perceptron," Prosiding Seminar Nasional Mahasiswa Bidang Ilmu Komputer dan Aplikasinya, vol. 2, no. 1, pp. 689-702, 2021.

[7] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English.," 2018. [Online]. Available: www.kaggle.com. [Accessed April 2024].

[8] E. Alpaydin, Introduction to Machine Learning, Cambridge, Massachusetts: The MIT Press, 2014.

[9] S. Patil and D. G. K. Kharate, "Implementation of SVM with SMO for Identifying Speech Emotions Using FFT and Source Features," Turkish Journal of Computer and Mathematics Education, vol. 12, no. 6, 2021. S. T. Alexander, "The Mean Squared Error (MSE) Performance Criteria," in *Adaptive Signal Processing*, Texts and Monographs in Computer Science, Springer, New York, NY, 1986.

[10] A. Muhaimin, D. D. Prastyo and H. Horng-Shing Lu, "Forecasting with Recurrent Neural Network in Intermittent Demand Data," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 802-809, doi: 10.1109/Confluence51648.2021.9376880.

Jurnal Aplikasi Sains Data

Journal homepage: https://jasid.upnjatim.ac.id/index.php/jasid

Application of K-Prototypes Clustermix Algorithm for Clustering Risk Factors of Diabetes Disease

Martina Hildha[1,] Hasih Pratiwi[2] Etik Zukhronah[3]

[1]Statistics Department of Universitas Sebelas Maret

[1]mhildha.mh@student.uns.ac.id

ABSTRACT

Abstract— Diabetes mellitus (DM) is recognized as one of the most rapidly increasing chronic diseases worldwide, posing a significant public health challenge. According to the International Diabetes Federation (IDF), approximately 537 million people were living with diabetes mellitus globally, with projections estimating a rise to 643 million by 2030 and 783 million by 2045. Additionally, the World Health Organization (WHO) reported a 3% increase in mortality rates attributed to diabetes mellitus between 2000 and 2019, underscoring the urgent need for effective risk detection and management strategies. Early identification of risk factors is crucial to mitigating the impact of DM, and clustering analysis offers a promising method for stratifying patients based on risk profiles. This study employs the k-prototypes algorithm, which is particularly suited to clustering datasets with mixed numeric and categorical variables, to analyze DM risk factors. Utilizing data from the 2022 Behavioral Risk Factor Surveillance System (BRFSS) annual survey, the study examines a sample of 2,480 diabetes mellitus patients across the United States. The clustering analysis identified two optimal clusters (k=2) based on a high silhouette score of 0.821, indicating strong cluster cohesion and separation. Cluster 2, consisting of 77 patients, exhibited a higher risk profile for diabetes compared to Cluster 1, which included 2,403 patients. The clusters were characterized by significant differences in average values of key DM risk factors including weight, fruit and vegetable consumption, mental and physical health status, age, alcohol consumption, hypertension, smoking status, physical activity, mobility difficulties, sex, education level, income, and ethnicity. These findings highlight the utility of k-prototypes clustering in identifying high-risk DM subgroups to inform targeted prevention and intervention efforts.

Keywords: diabetes mellitus (DM), mixed data types, clustering, silhoutte score

I. INTRODUCTION

Diabetes mellitus has become one of the chronic diseases with the fastest growing number of patients in the world. According to data released by the International Diabetes Federation (IDF) in the 10th edition of the Atlas publication [1], 537 million people are estimated to have diabetes mellitus, this is predicted to continue to increase to reach 643 million by 2030, and 783 million by 2045. In 2019, diabetes mellitus was the direct cause of 1.5 million deaths, with 48% of total diabetes deaths occurring before the age of 70. Another 460,000 deaths from kidney disease were also attributed to diabetes mellitus, while deaths from elevated blood glucose levels accounted for 20% of total deaths from cardiovascular disease [2].

Diabetes Mellitus is a chronic disease due to the inability of the pancreas to produce enough insulin, as well as manage the use of insulin effectively [3]. Insulin has an important role in managing blood sugar levels in the body. When the body experiences an uncontrolled increase in blood sugar or called hyperglycemia, this can cause various kinds of damage to the body system, especially to vital parts of blood vessels and nerves. According to Ksanti, diabetes has two main types of factors: modifiable factors such as weight, physical activity, hypertension, diet, and dyslipidemia; and non-modifiable factors such as heredity.

* *Corresponding author.*

~~E-mail addre~~ss: mhildha.mh@student.uns.ac.id

https://doi.org/10.33005/jasid.v1i1.9

Grouping of risk factors can be done using clustering analysis. Clustering analysis is a part of data mining that functions to group an object based on characteristics that are similar and different from other objects. Clustering is divided into two types, namely hierarchical and non-hierarchical. Hierarchical is done by grouping an object based on a similarity measure, then it will continue on other objects until the cluster will be formed like a tree at the most similar to least similar level. Non-hierarchical clustering is done with initial cluster initiation, and continues without following the hierarchical process [5].

Diabetes risk factors are taken from diabetes patient data including demographic information, medical history, and lifestyle factors. Some of these factors may have different data types, which can be numeric and also categorical so that they can be categorized in mixed data. k-prototypes is a non-hierarchical clustering method that combines k-means (for numeric data) and k-modes (for categorical data) methods, making it suitable for use on datasets containing variables with mixed data types called ClusterMix, both numeric and categorical [6]. In comparison, k-prototypes is a non-hierarchical clustering method that focuses on data clustering applications specifically with mixed attributes (numeric and categorical).

Based on the background explanation above, researchers are interested in discussing mixed data cluster analysis or ClusterMix using the k-prototypes method on DM patient data in the US region. Determination of the k value (number of clusters) is done by measuring the similarity of the distance between clusters with the highest sillhoutte score. The purpose of this study is to obtain optimal clustering results and form patient clusters based on the similarity distance of each variable for the benefit of further handling of DM patients. The clustering results are also expected to be able to help make decisions and actions that will be taken by medical personnel to minimize the risk of DM. Meanwhile, the utilization of machine learning in this study is a learning or reference to the use of the ClusterMix k-prototypes method to handle mixed data.

II. RESEARCH METHODS

A. Data

The data used in this study are secondary data from the Behavioral Risk Factor Surveillance System in 2022 in 50 states in the District of Columbia and three regions in the US. This data is a health survey data regarding diabetes mellitus, which is 2480 data with 15 variables.

This study uses 15 variables that become risk factor variables that cause diabetes mellitus. The research variables used in the analysis are presented in Table I.

TABLE I. RESEARCH VARIABLES

No	Variable	Description	Data Type
1	*Weight*	Body weight (kg)	*Numeric*
2	Fruits	Total fruit consumption per day	*Numeric*
3	Veggies	Number of fruit vegetables per day	*Numeric*
4	MentHlth	Number of days that you feel mentally healthy in a month	*Numeric*
5	PhysHlth	Number of days that you feel physically well within one month	*Numeric*
6	Age	Age at first onset of diabetes (years)	*Numeric*
7	HvyAlcoholConsump	Average number of drinking glasses per day	*Numeric*
8	HighBP	0 : *No* 1 : *Yes*	*Kategoric*
9	*Smoker*	0 : *Never* 1 : *Former* 2 : *Someday* 3 : *Everday*	*Kategoric*
10	*PhysActivity*	0 : *No* 1 : *Yes*	*Kategoric*
11	*DiffWalk*	0 : *No* 1 : *Yes*	*Kategoric*

12	*Sex*	0 : *Female* 1 : *Male*	*Kategoric*
13	*Education*	0 : Tidak bersekolah 1 : SD 2 : SMP 3 : SMA 4 : D3/D4 5 : S1	*Kategoric*
14	*Income*	0 : < $10,000 1 : $10,000 < $15,000 2 : $15,000 < $20,000 3 : $20,000 < $25,000 4 : $25,000 < $35,000 5 : $35,000 < $50,000 6 : $50,000 < $75,000 7 : $75,000 < $100,000 8 : $100,000 < $150,000 9 : $150,000 < $200,000 10 : > $200,000	*Kategoric*
15	*Ethnic*	0 : *Other Race* 1 : *White* 2 : *Black* 3 : *Asian* 4 : *Amarican Indian* 5 : *Hispanic*	*Kategoric*

B. Research Stagees

This research uses the k-prototypes method to perform clustering analysis of diabetes mellitus patients. This clustering method is carried out using Python software. The clustering analysis steps using the k-prototypes method are described in Figure 1.

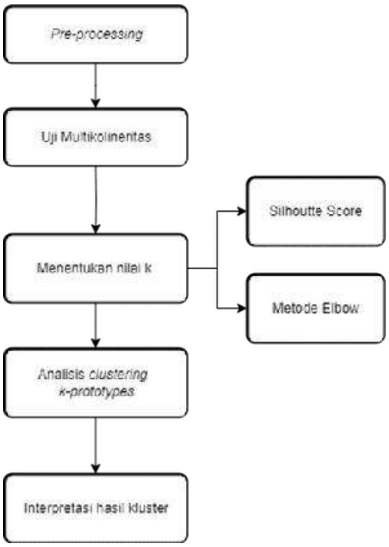


Figure 1. Flowchart

Figure 1 is a water diagram whose steps will be described in more detail as follows:

1. Pre-processing, data exploration which includes describing the data and checking for missing data.
2. Multicollinearity Test.
3. Determine the optimal number of clusters or k by finding the silhouette score and elbow angle in the elbow method.
   - Silhouette Score

     The silhouette score method, introduced by Rousseuw in 1987, is a tool for interpretation and validation of data clustering. The silhouette score has a range from -1 to 1, where values close to 1 indicate a very appropriate placement of objects in their clusters. Conversely, values close to -1 indicate less appropriate placement of objects [7]. Equation (1) is the formula used to calculate SC:

$$SC = \frac{b_i - a_i}{max\{a_i, a_i\}} \tag{1}$$

a_i: the average distance of the i-th data to all other data in a cluster;

b_i: the average distance of the i-th data to all other data in a cluster;

- Elbhow Method

  The Elbow method is a method of determining the optimal number of clusters can be done through the results of the comparison between the number of clusters that form an angle [8]. If a cluster A and the value in cluster B show a certain angle on the graph in the sense of experiencing a significant decrease, the number of cluster A values is considered the most appropriate. Value comparison is obtained through the calculation of Sum of Square Error (SSE) as in equation (2).

$$SSE = \sum_{k=1}^{k} \sum_{x_i} |x_i - c_k|^2 \tag{2}$$

k : cluster

x_i : distance of i-th object data

c_k : i-th cluster center

4. Clustering DM patient data based on risk factor variables using the k-prototypes algorithm with a mixed distance measure. The distance measure on mixed data types in the k-prototypes method is ideal for data that has two mixed types, namely numeric and categorical. This distance measure is shown in equation (3) [9]. The gamma coefficient (γ) in Equation (3) is calculated based on the average standard deviation (σ) of all numerical variables involved in the study.

$$d_{ij} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 + \gamma \sum_{s=p+1}^{m} \delta(x_{is}, x_{js}) \tag{3}$$

x_{ik} : i-th observation value on the k-th numerical variable,

x_{jk} : j-th cluster observation value on the kth numeric variable,

p : number of numerical variables

γ : Gamma coefficient

m : number of categorical variables

5. Interpreting the optimal cluster results.

III. RESULTS AND DISCUSSIONS

1. Preprocessing Data

TABLE II. DESCRIPTIVE STATISTICS OF NUMERICAL VARIABLES

VARIABEL	MISSING VALUE	MIN	MED	MAX	MEAN	STD
WEIGHT	105	31,75	90,26	244,94	92,75	24,3
FRUITS	202	0,00	1,00	100	1,5	6,00
VEGGIES	276	0,00	1,00	203	2,14	7,68
MENTHLTH	38	0,00	0,00	30,0	4,77	9,04
PHYSHLTH	65	0,00	0,00	30,0	7,12	10,8

Age	181	1,00	50,0	88,0	48,8	14,4
HvyAlcoholConsump	69	0,00	0,00	720,0	19,56	63,0

Table II. shows the data distribution of all numerical variables and the missing value of each variable, so it needs to be handled before further analysis is carried out. Handling missing value data on numerical variables is done by imputation.

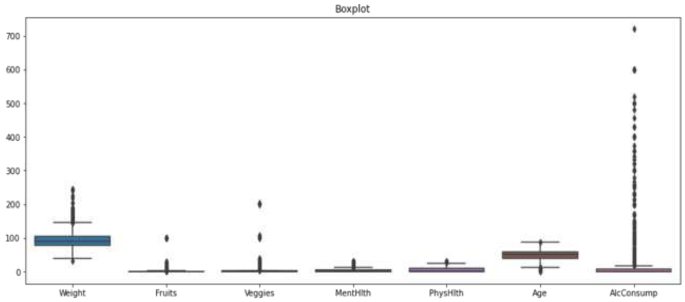


Figure 2. Boxplot of numerical variables

Figure 2 illustrates the distribution of data on numerical variables, where all variables have outlier data so it was decided to use the median value on each variable because the median value is robust to outliers.

TABLE III. DESCRIPTIVE STATISTICS OF NUMERICAL VARIABLES

Variabel	*Missing Value*	Keterangan
HighBP	10	0 = 687 1 = 1783
Smoker	31	0 = 1344 1 = 240 2 = 97 3 = 768
PhysActivity	1	0 = 1141 1 = 1338
DiffWalk	13	0 = 1556 1 = 911
Sex	0	0 = 1244 1 = 1236
Education	3	0 = 8 1 = 107 2 = 165 3 = 707 4 = 720 5 = 770
Income	256	0 = 130 1 = 150 2 = 144 3 = 187 4 = 344 5 = 318 6 = 341 7 = 238 8 = 223 9 = 82 10 = 67
Ethnic	0	0 = 106 1 = 1601 2 = 282 3 = 69 4 = 49 5 = 364

Table III shows the percentage or distribution of data in categorical variables including HighBP, Smoker, PhysActivity, DiffWalk, Sex, Education, Income, and Ethnic variables. In the categorical data used, there are missing values in all categorical variables except the Ethnic and Sex variables. Empty or null data is handled by entering the mode value in the missing value data.

2. Multicollinearity Test

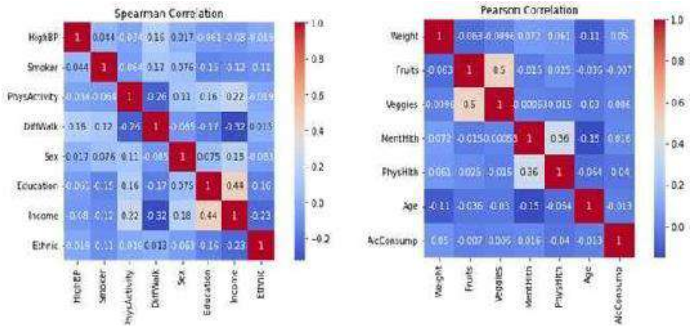


Figure 3. Cross Correlation Heatmap of Numeric Variables, and Categorical Variables

Multicollinearity test is a form of correlation test that exists in all factor variables in the data. If there is multicollinearity, the clusters formed will not be valid. The criterion for multicollinearity is if the relationship coefficient between variables exceeds 0.8. Meanwhile, correlation analysis in the relationship between numerical variables using Pearson correlation and categorical variables using Spearman correlation is used to determine the relationship of each variable. Figure 3 shows the relationship pattern between variables shown in the visual cross corelation heatmap. The redder the color, the stronger the relationship pattern between numerical variables. Figure 3 can be interpreted that between numeric and categorical variables do not have a strong relationship, so the analysis can continue.

3. Determination of Number of Clusters

The determination of the number of clusters (k) is done using the silhouette score and the elbow method. Many clusters with the highest silhouette score will be selected as the optimal cluster value. The following are the results of the Silhoutte Score calculation, namely the number k from 2 to 9 clusters shown in Table IV.

TABLE IV. DESCRIPTIVE STATISTICS OF NUMERICAL VARIABLES

k	*Silhoutte Score*
1	0,8201135
2	0,6439944
3	0,3100345
4	0,3208016
5	0,2311790
6	0,235277
7	0,233861
8	0,248135
9	0,246333

Table IV shows the results of the silhoutte score value of each k (cluster). The best number of clusters can be determined by looking at the highest silhoutte score value. The higher the Silhoutte Score value means that the distance of objects in the cluster is minimal, while the distance of objects with other cluster members is maximum. Based on Table IV, the highest Silhoutte score value is generated at k = 2, which is 0.8201, so that the diabetes mellitus patient data is grouped into 2 clusters.

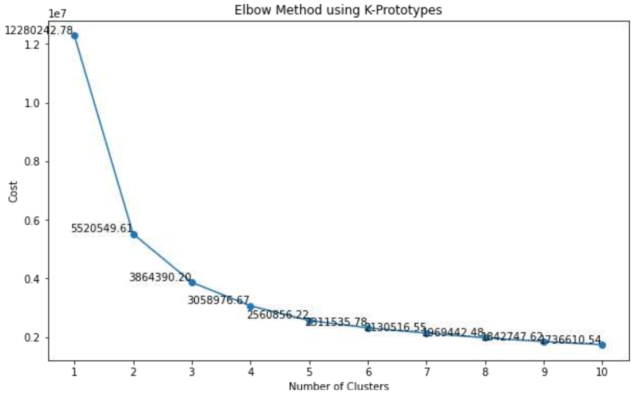


Figure 4. Elbow Method

Figure 4 shows that at k = 1, the SSE value reaches its peak and then begins to decrease significantly at k = 2. When k is at a point after 2, the SSE value decreases constantly. This indicates that right angle formation occurs at k=2.

4. Clustering Analysis with k-prototypes

   The application of the k-prototypes algorithm using Python obtained 2 clusters of diabetes mellitus patients with proportions as shown in Figure 5.

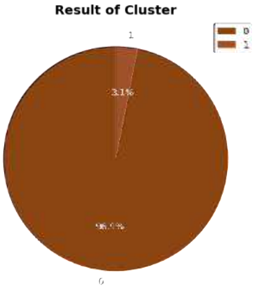


Figure 5. Proportion of Cluster Member Results

Figure 5 shows the results of the clusters formed from the number of k = 2. These results are visualized in the form of a pie chart which can be seen from Figure 4.8. The difference between cluster 1 and cluster 2 is quite significant. The distribution results show that the number of cluster 1 members is 2403 patients or 96.9%, while the number of cluster 2 members is only 77 patients or 3.1%. The Silhoutte Score obtained is 0.82011.

5. Interpretation of Clustering results

   The results of clustering analysis using the k-prototypes algorithm are displayed in the form of a comparison table of each cluster on each variable risk factor for diabetes mellitus. The results of the comparison of the characteristics of each cluster can be seen in Table V.

TABLE V. CHARACTERISTICS OF EACH CLUSTERING

Variabel	Cluster 1	Cluster 2	Rata-Rata/ Modus
Weight	92,54	95,96	92,64
Fruits	1,46	1,38	1,46
Veggies	2,01	2,09	2,01
MentHlth	4,68	5,14	4,70
PhysHlth	7,01	4,62	6,93
Age	48,91	48,55	48,9
HvyAlcoholConsump	9,67	310,63	19,02
HighBP	1	1	1
Smoker	0	3	0
PhysActivity	1	1	1
DiffWalk	0	0	0
Sex	0	1	0

Education	5	5	5
Income	5	6	4
Ethnic	1	1	1

Table V shows the salient characteristics of each cluster. Cluster 2 has an average body weight of 95.96, which is greater than the average body weight of all diabetes mellitus patients. While cluster 1, has an average that is almost the same as the average weight of all patients with diabetes mellitus, which is 92.54. The weight indicator is related to BMI where someone with an 'overweight' BMI category is more susceptible to the risk of causing diabetes mellitus.

A diet that involves regular consumption of fruits and vegetables with 2-3 servings per day can reduce the risk of diabetes mellitus [10]. Both clusters have an average of less than 2-3 servings of fruit consumption. Cluster 1 has an average daily fruit consumption of 1.46 servings which is greater than cluster 2 which only has an average of 1.36 servings. Meanwhile, the Veggies variable refers to the amount of daily vegetable consumption of diabetes mellitus patients. Both clusters have a normal average vegetable consumption equal to 2 servings. Cluster 1 has an average daily consumption of vegetables of 2 servings which is less than cluster 2 which only has an average of 2.09 servings.

There is a correlation between high physical and mental health risks for people with diabetes mellitus [11]. This is influenced by good blood sugar management and other factors such as blood pressure and lipids. Cluster 1 has an average number of days of 4.68 days which is less than cluster 2 which only has an average of 5.04 days. Meanwhile, the PhysHlth variable which refers to the number of days in a month where physical health conditions are in good condition shows a significant difference between clusters 1 and 2. Overall diabetes mellitus patients have an average of 6.93 close to 7 days in a period of 30 days (a month). Meanwhile, Cluster 1 has an average number of days of 7.01 or 7 days which is more than cluster 2 which only has an average of 4.62 or close to 5 days.

The Age variable refers to the age when the patient was diagnosed with diabetes mellitus. The results of this study show that the difference between clusters 1 and 2 is not far away and even almost the same. Cluster 1 has an average number of days of 48.91 years which is younger than cluster 2 which only has an average of 48.55 years. Age > 45 years has a risk factor of 1.4 times experiencing abnormal blood sugar levels [12]. This is due to lack of physical activity, increasing body weight, reduced muscle mass and progressive cell shrinkage.

The variable HvyAlcoholConsump which refers to the amount of daily alcohol consumption. The results of this study showed that the differences between clusters 1 and 2 were highly significant. The overall diabetes mellitus patients had an average of 19.02 servings of daily alcohol consumption. Cluster 1 had an average of 9.67 servings of alcohol consumption which was less than cluster 2 which had an average of 310.63 servings. Repeated consumption of large amounts of alcohol is associated with a higher risk for diabetes mellitus [13].

The HighBP variable refers to the history of high blood pressure in patients with diabetes mellitus. Both clusters have a majority of patients who share a history of high blood pressure. A history of elevated systolic and diastolic blood pressure in middle-aged and older adults can increase the risk of diabetes mellitus [14]. PhysActivity variable which refers to physical activity. In this variable, the majority of patients in both cluster 1 and cluster 2 both have physical activity activities. Meanwhile, the DiffWalk variable which refers to complaints of difficulty walking has the majority of patients in both cluster 1 and cluster 2 who do not have complaints of difficulty walking. The Education variable which refers to the last level of education shows that patients in both cluster 1 and cluster 2 are dominated by patients with the last level of education S1 and have the same Ethnic category in the Ethnic variable, namely white.

The Smoker variable or smoking history category shows that all diabetes mellitus patients are dominated by patients who do not have a history of smoking. This also applies to the majority of cluster 1 patients who do not smoke, in contrast to cluster 2, where the majority have a history of active smoking with a high frequency (often). Smoking itself causes resistance to insulin which has an impact on impaired glucose metabolism, thus increasing the risk of diabetes mellitus. The variable Sex or gender shows that overall diabetes mellitus patients are dominated by women. This also applies to the majority of cluster 1 patients who are predominantly female with an income of $25,000-$35,000, in contrast to cluster 2 which is dominated by males with higher incomes of $35,000-$50,000. According to Damayanti [15], in prevalence women are much more at risk of developing diabetes because physically women have a greater probability of increasing BMI than men.

Income variables show results that are not much different, the majority of cluster 1 patients have an income of $25,000-$35,000, in contrast to cluster 2 which has a higher income of $35,000-$50,000. The higher the income of an individual, the greater the level of lifestyle lived which has an impact on diet, especially the consumption of junk food.

CONCLUSION

The k-prototypes algorithm produced two clusters with a silhouette score of 0.821. Cluster 1 consists of 2403 patients and cluster 2 consists of 77 patients. Cluster 1 shows that the majority of patients are non-smoking females with an income of $25,000-$35,000 and low mean scores on Weight, Veggies, PhysHlth Menthlth, and HvyAlcoholConsump, and high mean scores on PhysHlth, Age, and Fruits variables. Cluster 2 shows that the majority of patients are men who have a smoking habit with high intensity followed by an income of $35,000-$50,000 and have a higher average value than cluster 1 on the variables Weight, Veggies, PhysHlth Menthlth, and HvyAlcoholConsump, and a lower average on the variables PhysHlth, Age, and Fruits. So that patients who are in cluster 2 have a higher risk of diabetes mellitus than patients in cluster 1.

REFERENCES

[1] W. Yudananto, S. S. Remi, and B. Muljarijadi, "Peranan Sektor Pariwisata Terhadap Perekonomian Daerah di Indonesia (Analisis Interregional Input-Output)," Jurnal, vol. 2, no. 4, 2012, Universitas Padjajaran, Bandung.

[2] International Diabetes Federation. *IDF Diabetes Atlas 10th Edition* ; International Diabetes Federation, 2021.

[3] ElSayed, N. A.; Aleppo, G.; Aroda, V. R.; Bannuru, R. R.; Brown, F. M.; Bruemmer, D.; Collins, B. S.; Cusi, K.; Das, S. R.; Gibbons, C. H.; Giurini, J. M.; Hilliard, M. E.; Isaacs, D.; Johnson, E. L.; Kahan, S.; Khunti, K.; Kosiborod, M.; Leon, J.; Lyons, S. K.; Murdock, L. Introduction and Methodology: Standards of Care in Diabetes—2023. *Diabetes Care* **2022**, *46* (Supplement_1), S1–S4.

[4] GBD 2021 Diabetes Collaborators. Global, Regional, and National Burden of Diabetes from 1990 to 2021, with Projections of Prevalence to 2050: A Systematic Analysis for the Global Burden of Disease Study 2021. *The Lancet* **2021**, *402* (10397).

[5] Kshanti, I. A. M. *Pedoman Pemantauan Glukosa Darah Mandiri*, 1st ed.; PB Perkeni: Jakarta, 2019.

[6] Yuan, C.; Yang, H. Research on K-Value Selection Method of K-Means Clustering Algorithm. *J* **2019**, *2* (2), 226–235.

[7] Madhuri, R.; Murty, M. R.; Murthy, J. V. R.; Reddy, P. V. G. D. P.; Satapathy, S. C. Cluster Analysis on Different Data Sets Using K-Modes and K-Prototype Algorithms. *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India- Vol II* **2014**, *249* (137), 137–144.

[8] Rousseeuw, P. J.; Leroy, A. M. *Robust Regression and Outlier Detection. Wiley Series in Probability and Mathematical Statistics*; John Wiley: New York, 1987.

[9] Bholowalia, P.; Kumar, A. EBK-Means: A Clustering Techiniques Based on Elbow Method and K-Means in WSN. *International Journal of Computer Application (0975-8887)* **2014**, *IX* (105), 17–24.

[10] Huang, Z. Extensions to the K-Means Algorithm for Clustering Large Data Sets with

Categorical Values. *Data Mining and Knowledge Discovery* **1998**, *2* (3), 283–304.

[11] Siregar, P. A. Analisis Karakteristik Dan Frekuensi Konsumsi Buah Dan Sayur Pada Penderita Diabetes Dan Non Diabetes. *Darussalam Nutrition Journal, Mei* **2021**, *2021* (1), 61–69.

[12] Kulzer, B. Physical and Psychological Long-Term Consequences of Diabetes Mellitus. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz* **2022**, *66* (4), 503.

[13] Rudi, A.; Kwureh, H. N. Faktor Risiko Yang Mempengaruhi Kadar Gula Darah Puasa pada Pengguna Layanan Laboratorium. *Wawasan Kesehatan* **2017**, *3* (2), 2087–4995.

[14] Wu, X.; Liu, X.; Liao, W.; Kang, N.; Dong, X.; Abdulai, T.; Zhai, Z.; Wang, C.; Wang, X.; Li, Y. Prevalence and Characteristics of Alcohol Consumption and Risk of Type 2 Diabetes Mellitus in Rural China. *BMC Public Health* **2021**, *21* (1). https://doi.org/10.1186/s12889-021-11681-0.

[15] Yang, X.; Chen, J.; Pan, A.; Wu, J. H. Y.; Zhao, F.; Xie, Y.; Wang, Y.; Ye, Y.; Pan, X.-F.; Yang, C.-X. Association between Higher Blood Pressure and Risk of Diabetes Mellitus in Middle-Aged and Elderly Chinese Adults. *Diabetes & Metabolism Journal* **2020**, *44* (3), 436.

[16] Damayanti, S. *Diabetes Melitus Dan Penatalaksanaan Keperawatan*; Nuha Medika; Yogyakarta, 2015.

E-ISSN: 2025-xxxx
P-ISSN: 2025-xxxx
Doi: https://doi.org/10.33005/jasid.v1i1
https://jasid.upnjatim.ac.id/index.php/jasid