Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.Doi Number

Study of Postpartum Maternal Health by Age in Districts and Cities with the Partitional Algorithm Method

David Agustriawan, PhD[1], Ersan Ivanda Putra[2], Elisabet Lumban Tobing[3], Caroline Alexandra Santoso[4], Florecita Patricia[5]

[1]Multimedia Nusantara University
[2]Multimedia Nusantara University
[3]Multimedia Nusantara University
[4]Multimedia Nusantara University
[5]Multimedia Nusantara University

Corresponding author: First A. Author (e-mail: david.agustriawan@umn.ac.id).

Corresponding author: First A. Author (e-mail: ersan.ivanda@student.umn.ac.id).

Corresponding author: Author (e-mail: elisabet.lumban@student.umn.ac.id).

Corresponding author: Author (e-mail: caroline.alexandra@student.umn.ac.id).

Corresponding author: Author (e-mail: florecita.patricia@student.umn.ac.id).

This work was supported in part by Universitas Multimedia Nusantara.

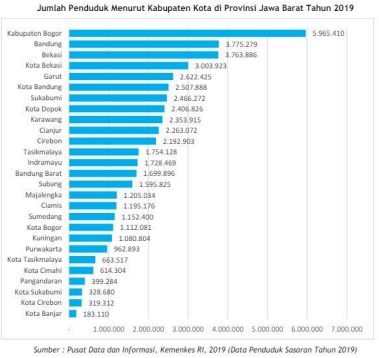
ABSTRACT Maternal health remains a concern in Indonesia, particularly in West Java, where various factors impact pregnant women's well-being. Despite global efforts, Indonesia faces a high maternal mortality rate compared to Southeast Asia, with rural disparities hindering access to maternal health services. Sociocultural barriers, like the preference for Traditional Birth Attendants (TBAs), affect the uptake of Maternal and Newborn Health (MNH) services, and child stunting remains a prevalent issue. To address these challenges, Machine Learning (ML) offers opportunities for early detection and decision support in biomedicine. This study analyzed data related to age, SystolicBP, and DiastolicBP using cross-validation, K-Means clustering, correlation analysis, and model evaluation metrics. The regression model showed strong performance in predicting age based on blood pressure metrics, with a low Negative Mean Squared Error of -124.80. K-Means clustering identified higher prevalence of blood pressure issues among individuals aged 25 to 45 in city district 3230. Correlation analysis revealed moderate positive correlations between age and blood pressure metrics. Evaluation metrics, including a high Silhouette Score of 0.93 and a low Davies-Bouldin Index of 0.2578, confirmed the validity of the clustering results. The Elbow Method suggested an optimal clustering solution at 3 or 4 clusters. By leveraging historical data and ML algorithms, healthcare professionals can improve maternal and child health outcomes through timely interventions. Effective strategies are essential to address maternal health challenges and ensure the well-being of mothers and their children.

INDEX TERMS ***Maternal health, Significant, West Java Region, pregnant women, Southeast Asia Region.***

I. INTRODUCTION

The health of pregnant women in Indonesia, especially in the West Java region, is influenced by several factors that affect their health conditions. The maternal mortality rate (MMR) in Indonesia is high at 305 compared to 204 deaths per 100,000 population in the Southeast Asia Region. [1]. The challenge of dealing with maternal mortality is globally recognized and has become a Sustainable Development Goal (SDG). [2].

Rapid population growth can affect the health system, especially in adequate maternal health services.



Indonesia's access to health services has improved, but there are still disparities in access to health, especially in rural areas. The following graph shows the population in West Java(Kemenkes Jabar, Page 16) [3].

Home childbirths continue to be common in rural areas of Indonesia, often due to the availability and preference for Traditional Birth Attendants (TBAs) at the village level. Findings from various, and West Java indicate that the utilization of TBAs poses a significant sociocultural barrier to the uptake of Maternal and Newborn Health (MNH) services [13]. In 2018, the prevalence of stunting among children aged 0-59 months was 30.8%, affecting 27,023 out of 87,737 children. This rate was notably higher than in other countries with similar levels of economic development [4]. In 2017, in East Java province, 97% of women had at least one antenatal care (ANC) visit with a skilled attendant, while 78% had at least four ANC visits. Additionally, 95% of women had a skilled attendant present at delivery. Despite this level of access to skilled care, many women experienced complications. The reported Maternal Mortality Ratio (MMR) in East Java was 91 per 100,000 live births in 2017, resulting in 534 maternal deaths [26].

Machine Learning (ML) is frequently used in biomedicine to detect or predict specific pathological conditions. In pregnancy, the emphasis has largely been on diagnostics. Nevertheless, as demonstrated in various disciplines, ML can serve other purposes, including identifying critical variables in a system or process, conducting correlation analysis, managing and extracting data, removing noise, and reducing dimensionality, among other applications [30]. The creation of decision support systems, which leverage historical data and incorporate machine learning algorithms for learning, is believed to offer physicians an opportunity to access predictive information. This can enable them to make timely decisions that could potentially save the lives of pregnant women and their fetuses [34].

While various machine learning algorithms such as Logistic Model Tree (LMT), Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest, Light Gradient Boosting Machine (LightGBM), Gradient Boosting Machines (GBM), and k-Nearest Neighbors (KNN) have been applied, there is a lack of consensus on which algorithm is the most effective for predicting the risk levels of low, medium, and high in total of 1,014 data samples.

Features used in these algorithms differ, including age, SystolicBP, DiastolicBP, BS, Body Temperature, Heart Rate, and Risk Level (Low, Medium, High), suggesting a need for standardization or identification of the most relevant features for accurate prediction.

The accuracy rates of the models vary, with LightGBM achieving the highest accuracy at 84.24%, followed by CatBoost at 83.74%, Random Forest at 81.28%, and GBM at 73.89%, while KNN had an accuracy of 68.47%. This variance indicates the need for further research to improve the accuracy and consistency of the models.

For sample sizes used in these studies differ significantly, ranging from 181 to 10,000 data points, indicating a need for more extensive and standardized datasets to validate the effectiveness of these machine learning models for predicting health risks in pregnant women accurately [36] [37] [38].

In conclusion, maternal health is a significant public health challenge, and addressing it is critical to the well-being of both the mother and her child. Health problems during pregnancy and the postpartum period can be attributed to various factors, and effective interventions can be provided to improve maternal health and child development.

II. METHODOLOGY

The method used in scientific writing is a literature review. The databases used are Kaggle, OpenDataJabarProv, and Google Scholar. Research methodology steps that can be taken in preparing and working on a project, starting from problem identification to system evaluation. Namely identifying the health problems of pregnant women after giving birth based on the background and problem formulation that has been presented. Literature study by reading journals, World Health Organization reports, and related empirical studies to gain a deeper understanding of pregnant and postnatal maternal health issues and associated risk factors. Data Collection: collect data from the datasets mentioned, namely "Maternal Health Risk Data Set" and "Birth Rates Result of Long Form SP2020 According to Maternal Age Group (Age Specific Fertility Rate_ASFR) and Province_Regency_City, 2020". Data Analysis, analyze data to identify risk factors that contribute to maternal and postnatal health problems and their impact on maternal well-being and child development. Model Development: Develop predictive models to identify health risks early based on identified risk factors.

2.1 FLOWCHART

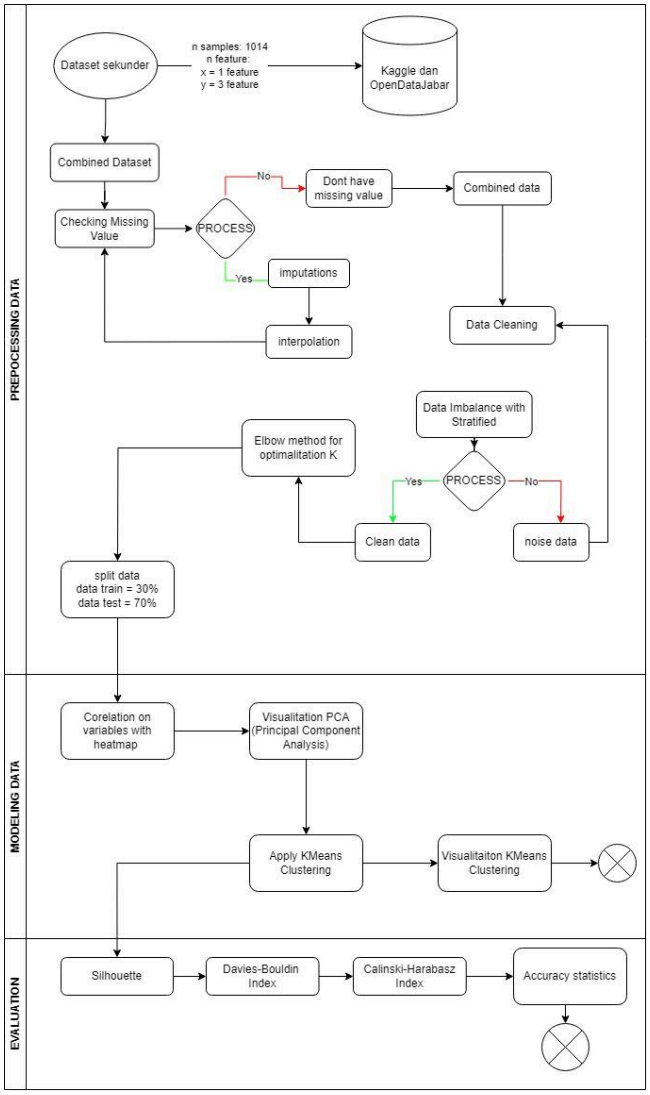


Figure 1. Flowchart

A. Dataset

Research datasets from
Dataset 1: Kaggle
Dataset 2: OpenDataJabarProv
We used the data to analyze the health of pregnant women around West Java, Indonesia. We analyzed the health of pregnant women based on their age.

B. Preprocessing

The data preprocessing stage is the initial process of analyzing data. This process aims to correct errors in the data to facilitate the future research process.

1. Combined Dataset

Combining the two datasets in this study aims to support a more in-depth and complex analysis.

2. Checking Missing Value

While this method is simple and fast, it reduces variability and may introduce bias if the data is not completely missing at random (MCAR) [10].

3. Data Cleansing

This stage cleans the data from missing values by using the median value and prevents data duplication.

4. Checking Data Balance

Data will be unbalanced if the target variables in the data do not have an equal distribution. The unbalanced data sets' classification process involves feature selection, data distribution adjustment, and model training [5].

5. Split Counting

This process divides the dataset into two: training data and test data [6]. This process is an important part of training and testing data.

C. Model Validation

1. Heatmap Corelation

Using heat maps to test the correlation between features is an effective way to identify linear relationships and potential multicollinearity in a data set. Checking the correlation between features in a data set is an important step in data analysis and preprocessing. Correlation indicates the extent to which two variables are related to each other. Correlation allows you to understand the linear relationship between variables and identify potentially overlapping features.

2. PCA (Principal Component Analysis)

PCA (principal component analysis) and decision trees are two different techniques that are commonly used for different purposes in data analysis and machine learning. PCA is used to reduce the size of data before applying machine learning algorithms. It is considered as part of the pre-processing stage. In this context, PCA can help reduce the number of features, remove redundancy, and reduce noise, thereby improving model performance.

3. Apply KMeans Clustering

Applying KMeans clustering is a useful technique for grouping data into clusters based on similarity of features. After clustering, cluster visualization helps you understand patterns in your data, and placing cluster information in a table enables further analysis.

D. Evaluation Model

1. Silhouette

Silhouette score is an evaluation metric used to evaluate how well the clustering result achieves. It provides a measure of how well objects in the same cluster are grouped compared to other clusters. The higher the silhouette score value, the better the clustering result.

2. Dalvis Boulden

The Davies-Bouldin index (DBI) is a cluster evaluation metric used to evaluate the quality of clusters produced by a clustering algorithm. The goal is to find distinct clusters by considering how efficiently the data in each cluster is used

by the cluster. A lower Davies-Bouldin index value indicates better cluster separation.

3. Calinski-Harabasz Index

The Calinski-Harabasz index (also known as the CH criterion score) is a clustering evaluation metric used to evaluate the quality of clusters produced by a clustering algorithm. The goal is to find distinct clusters by considering how efficiently the data in each cluster is used by the cluster. A higher Calinski-Harabasz index value indicates better cluster separation.

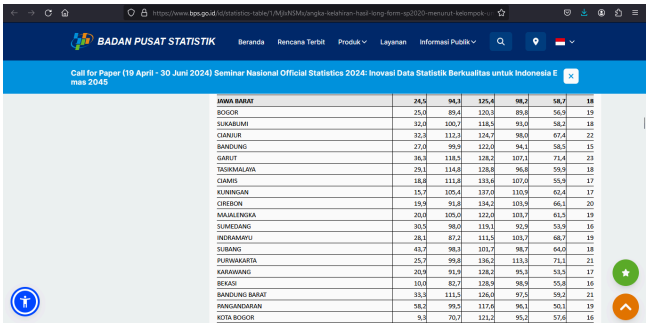
4. Accuracy Statistics

Statistical accuracy is an evaluation metric used to evaluate the performance of a model by comparing the model's predictions with the true values in the data set. It is one of the most basic and common evaluation metrics used in various types of tasks such as classification and regression.

III. RESULT AND DISCUSSION

3.1 Preprocessing

A. Data 1



source data :
https://www.bps.go.id/id/statistics-table/1/MjIxNSMx/angka-kelahiran-hasil-long-form-sp2020-menurut-kelompok-umur-ibu-age-spesific-fertility-rate-asfr-dan-provinsi-kabupaten-kota-2020.html
Data retrieval : Friday, 8 March 2024

In the picture above is the result of the description of the data we use in this research. Where we use 2 data sets and we take both data sets from the Kaggle and Jabarpemprov websites. Data 1, has a row of about 1014 and has 7 columns, which is sufficient for us to conduct the research we are looking for regarding the health of mothers giving birth.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1014 entries, 0 to 1013
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Age          1014 non-null   int64
 1   SystolicBP   1014 non-null   int64
 2   DiastolicBP  1014 non-null   int64
 3   BS           1014 non-null   float64
 4   BodyTemp     1014 non-null   float64
 5   HeartRate    1014 non-null   int64
 6   RiskLevel    1014 non-null   object
dtypes: float64(2), int64(4), object(1)
memory usage: 55.6+ KB
```

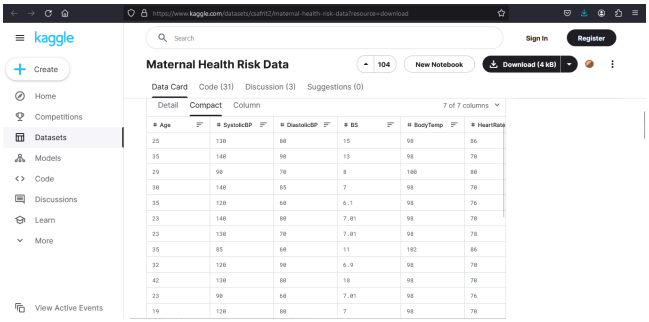
This data frame in data 1 has 1014 entries and 7 columns. The columns are:

a) 'Age': Indicates age with integer data type.
b) 'SystolicBP': Shows systolic blood pressure with integer data type.
c) 'DiastolicBP': Indicates diastolic blood pressure with integer data type.
d) 'BS': Indicates blood sugar level with float data type.
e) 'BodyTemp': Indicates body temperature with float data type.
f) 'HeartRate': Indicates heart rate with integer data type.
g) 'RiskLevel': Indicates the risk level with the data type object (string).

This data frame has no null values in each column. The data type for each column is according to the type of data contained in it, which are integer, float, and object (string).

B. Data 2

Data 2 is data that we use as complementary data to fulfill the research that we have examined in this journal.



source data :
https://www.kaggle.com/datasets/csafrit2/maternal-health-risk-data?resource=download
Data retrieval : Wednesday, 21 February 2024

The output results above are the results of the image 2 data that we used for our research, but in this second data, we did not use all the columns we used but only used 2 columns, namely code_district_city, and number_of_mothers_pregnant because we wanted to know which regions had the highest number of diseases after giving birth.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 189 entries, 0 to 188
Data columns (total 8 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   id                   189 non-null    int64
 1   kode_provinsi        189 non-null    int64
 2   nama_provinsi        189 non-null    object
 3   kode_kabupaten_kota  189 non-null    int64
 4   nama_kabupaten_kota  189 non-null    object
 5   jumlah_ibu_hamil     189 non-null    int64
 6   satuan               189 non-null    object
 7   tahun                189 non-null    int64
dtypes: int64(5), object(3)
memory usage: 11.9+ KB
```

The data has 189 entries and 8 columns. The columns are:
a) 'id': Contains an ID with integer data type.
b) 'province_code': Contains the province code with an integer data type.
c) 'province_name': Contains the name of the province with the data type object (string).
d) 'code_kabupaten_kota': Contains a district/city code of type integer.
e) 'name_kabupaten_kota': Contains the name of the district with the data type object (string).
f) 'number_ibu_hamil': Contains the number of pregnant women with integer data type.
g) 'unit': Contains a unit with the data type object (string).
h) 'year': Contains year with an integer data type.

This data frame has no null values in any of its columns. The majority of the columns have an integer data type, except for the columns 'name_province', 'name_kabupaten_kota', and 'unit' which have an object (string) data type. This data frame most likely contains data about the number of pregnant women by province, district/city, and year.

3.2 Age Grouping

At this point the research team grouped the ages in Data 1, this was done to find out the age distribution in the data and we also sorted the ages in Data 1.

	Age		Age		Age
11	19	5	23	0	25
15	15	6	23	12	25
23	18	10	23	17	25
25	16	13	20	29	28
26	19	22	21	42	25
...	...	...	...	...	...
989	17	939	21	971	28
990	19	945	22	981	25
992	17	947	23	984	28
1005	17	950	23	988	25
1006	17	1009	22	993	25

Figure 1. Age Grouping

In the picture above is the output result of age grouping in Data 1, we grouped the results from the age of 17 to 30 so that we can find out whether the age we grouped is as per Data 1.

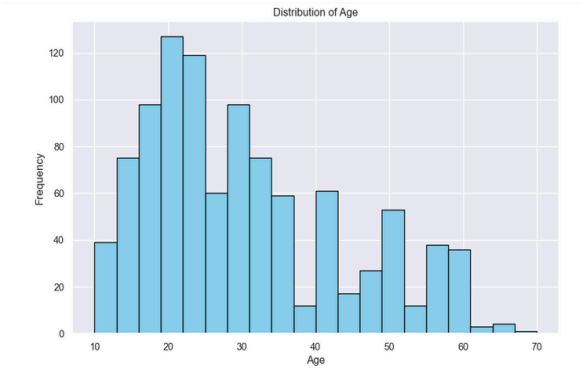


Figure 2. Distribution of Age

3.3 Combined Data 1 & Data 2

Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel	nama_provinsi	kode_kabupaten_kota	nama_kabupaten_kota	jumlah_ibu_hamil	sat
25	130	80	15.0	98.0	86	1	JAWA BARAT	3201	KABUPATEN BOGOR	126474	ORA
35	140	90	13.0	98.0	70	1	JAWA BARAT	3202	KABUPATEN SUKABUMI	51056	ORA
29	90	70	8.0	100.0	80	1	JAWA BARAT	3203	KABUPATEN CIANJUR	46284	ORA
30	140	85	7.0	98.0	70	1	JAWA BARAT	3204	KABUPATEN BANDUNG	79912	ORA
35	120	60	6.1	98.0	76	3	JAWA BARAT	3205	KABUPATEN GARUT	62514	ORA
...	...	...	...	...	...	...	...	...	...	...	...
22	120	60	15.0	98.0	80	1	JAWA BARAT	3211	KABUPATEN SUMEDANG	19631	ORA
55	120	90	18.0	98.0	60	1	JAWA BARAT	3212	KABUPATEN INDRAMAYU	29942	ORA
35	85	60	19.0	98.0	86	1	JAWA BARAT	3213	KABUPATEN SUBANG	35848	ORA
43	120	90	18.0	98.0	70	1	JAWA BARAT	3214	KABUPATEN PURWAKARTA	19662	ORA
32	120	65	6.0	101.0	76	2	JAWA BARAT	3215	KABUPATEN KARAWANG	45558	ORA

rows × 13 columns

Figure 3. Combined Data

After we determine which ones we will group and need to fulfill our research data, we use only 1 column because we are focusing on the Indonesian region.

3.4 PreProcessing

3.4.1 Stratified Data Balance

The function of using the stratified method to check the balance of the data is to ensure that the division of classes in the dataset is not significantly biased. The Stratified method ensures that the distribution of samples into each fold of the cross-validation is done in such a way that the class distribution in each fold reflects the class distribution in the dataset as a whole.

A. Data 1

```
Average F1 score for column 'SystolicBP' with 2 folds: 0.5559
  Fold 1: F1 score = 0.5028
  Fold 2: F1 score = 0.6090

Average F1 score for column 'DiastolicBP' with 2 folds: 0.4513
  Fold 1: F1 score = 0.4572
  Fold 2: F1 score = 0.4454

Average F1 score for column 'HeartRate' with 2 folds: 0.8890
  Fold 1: F1 score = 0.8890
  Fold 2: F1 score = 0.8890

Average F1 score for column 'RiskLevel' with 2 folds: 0.3393
  Fold 1: F1 score = 0.3622
  Fold 2: F1 score = 0.3163
```

Figure 4. Stratified Data Balanced of Data 1

1) column ‘SystolicBP’

The average F1 score of 0.5559 with variability between the first fold (0.5028) and the second fold (0.6090) indicates that the model has inconsistent and overall moderate performance. This suggests that there is room for improvement in terms of either the model or data preprocessing to improve the consistency and accuracy of the model's predictions.

2) column ‘DiastolicBP’:

The average F1 score of 0.4513 with consistent values in the first fold (0.4572) and second fold (0.4454) shows that the model has a consistently low performance in predicting the 'DiastolicBP' class. This indicates that the model often makes errors in prediction. Further adjustments need to be made to the model or data to improve the accuracy and consistency of the model's predictions.

3) column ‘HeartRate’

The average F1 score of 0.8890 with consistent values in the first fold (0.8890) and second fold (0.8890) shows that the model has an excellent and consistent performance in predicting the 'HeartRate' class. This shows that the model is able to make accurate predictions and is balanced between precision and recall, and is stable across different subsets of data. The model seems ready to be further tested or applied in a real context, while considering additional testing and potential improvements through feature engineering.

4) column ‘RiskLevel’

The average F1 score of 0.3393 with varying values in the first fold (0.3622) and second fold (0.3163) shows that the model has a low and inconsistent performance in predicting the 'RiskLevel' class. This indicates that the model often makes errors in prediction. Further adjustments need to be made to the model or data to improve the accuracy and consistency of the model's predictions. Given the variability between folds, it is recommended to conduct a more in-depth analysis of the data and try different techniques to improve the model performance

B. Data 2

```
Average F1 score for column 'jumlah_ibu_hamil': 0.547
Fold 1: F1 score = 0.542
Fold 2: F1 score = 0.552
Average F1 score for column 'tahun': 0.271
Fold 1: F1 score = 0.260
Fold 2: F1 score = 0.282

F1 scores for 'jumlah_ibu_hamil':
Fold 1: F1 score = 0.542
Fold 2: F1 score = 0.552

F1 scores for 'tahun':
Fold 1: F1 score = 0.260
Fold 2: F1 score = 0.282
```

Figure 5. Stratified Data Balanced of Data 2

1) column ‘jumlah_ibu_hamil’

The average F1 score of 0.547 with consistent values in the first fold (0.542) and second fold (0.552) shows that the model has a moderate and fairly consistent performance in predicting the ‘jumlah_ibu_hamil’' class. This shows that the model has some errors in prediction, but in general it is quite balanced between precision and recall. Further adjustments to the model or data are needed to improve the accuracy and consistency of the model's predictions, while continuing to monitor the stability of the model's performance across different subsets of data.

2) column ’tahun’

The average F1 score of 0.271 with consistent values in the first fold (0.260) and second fold (0.282) shows that the model has a very low and fairly consistent performance in predicting the 'year' class. This indicates that the model often makes errors in prediction and requires further adjustments to the model or data to improve the accuracy and consistency of the model's predictions. Given the low performance, it is recommended to conduct a more in-depth analysis of the data and try various techniques to improve model performance, including feature engineering and handling class imbalance.

C. Combination of 2 Data

```
Average F1 score for column 'Age': 0.6110
Fold 1: F1 score = 0.6251
Fold 2: F1 score = 0.5969
Average F1 score for column 'jumlah_ibu_hamil': 0.5397
Fold 1: F1 score = 0.5384
Fold 2: F1 score = 0.5409
Average F1 score for column 'SystolicBP': 0.4532
Fold 1: F1 score = 0.4485
Fold 2: F1 score = 0.4578
Average F1 score for column 'DiastolicBP': 0.3759
Fold 1: F1 score = 0.3715
Fold 2: F1 score = 0.3803
Average F1 score for column 'kode_kabupaten_kota': 0.5397
Fold 1: F1 score = 0.5409
Fold 2: F1 score = 0.5384
Average F1 score for column 'tahun': 0.3074
Fold 1: F1 score = 0.3169
Fold 2: F1 score = 0.2978
Average F1 score for column 'HeartRate': 0.8890
Fold 1: F1 score = 0.8890
Fold 2: F1 score = 0.8890
```

Figure 6. Combination of Stratified Data Balanced

We check the balance of the data using only the variables that will be used to conduct the research.
The model performed best in the 'HeartRate' column with a very high and consistent F1 score (0.8890). 'Age' also performed quite well with an average value of 0.6110. The columns 'number_of_mothers_pregnant' and 'code_district_city' had moderate performance (around 0.54) with good consistency. However, performance declined in 'SystolicBP' (0.4532) and 'DiastolicBP' (0.3759), indicating model difficulty. 'Year' had the worst performance with an average F1 score of 0.3074. Overall, the model needs improvement especially on features with low F1 score.

3.4.2 View Missing Values

```
Missing values for numeric columns:
Age            0
SystolicBP     0
DiastolicBP    0
BS             0
BodyTemp       0
HeartRate      0
dtype: int64
```

Figure 7. Missing values for Numeric columns

The result above is the result of the output we did to see the missing values. We did this stage to find out whether our data was combined between data 1 and data 2. do to find out whether the data we combined between data 1 and data

3.4.3 Elbow Method

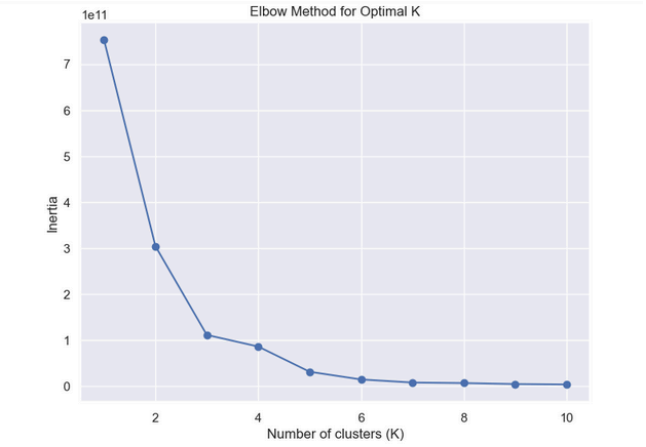


Figure 8. Elbow Method for Optimal K

After performing the elbow method, we have determined that the optimal number of clusters (K) for our analysis is 4. This decision is based on the point where the inertia significantly decreases and starts to level off, indicating a balance between the number of clusters and the variance explained by the clustering model.

For our clustering analysis, we will utilize the following variables:

1. Number of Pregnant Women: This variable represents the count of pregnant women within the dataset, providing crucial demographic information.

2. Age: The age of individuals is a critical factor in understanding the distribution and characteristics of the population.

3. Regency Code: This geographic identifier will help in segmenting the data based on regional differences and similarities.

4. Heart Rate: This health metric is vital for assessing the physiological state and potential risk factors among the individuals.

By clustering our data using these variables, we aim to uncover meaningful patterns and insights that can inform targeted interventions and resource allocation. This approach will enable us to understand better the diverse needs and characteristics of different groups within our population.

3.4.4 Apply K-Means Clustering

a. Apply to K-Means Model

	Age	jumlah_ibu_hamil	SystolicBP	DiastolicBP	kode_kabupaten_kota	tahun	RiskLevel	HeartRate	cluster
0	25	126474	130	80	3201	2016	1	86	0
1	35	51056	140	90	3202	2016	1	70	3
2	29	46284	90	70	3203	2016	1	80	1
3	30	79912	140	85	3204	2016	1	70	3
4	35	62514	120	60	3205	2016	3	76	0
...	...	...	...	...	...	...	...	...	...
1009	22	19631	120	60	3211	2018	1	80	1
1010	55	29942	120	90	3212	2018	1	60	3
1011	35	35848	85	60	3213	2018	1	86	1
1012	43	19662	120	90	3214	2018	1	70	3
1013	32	45558	120	65	3215	2018	2	76	2

1014 rows × 9 columns

Figure 9. Apply K-Means

This table displays the results of applying the K-Means clustering algorithm to a dataset consisting of 1014 rows and 9 columns. The columns include:

a) Age: The age of the individual.

b) jumlah_ibu_hamil**: The number of pregnant women in the region.

c) SystolicBP: Systolic blood pressure.

d) DiastolicBP: Diastolic blood pressure.

e) kode_kabupaten_kota: Region code.

f) tahun: Year the data was collected.

g) RiskLevel: Health risk level (potentially related to pregnancy or other health conditions).

h) HeartRate: Heart rate.

i) cluster: The result of the K-Means clustering.

The "cluster" column indicates the cluster assignment for each row of data, with values ranging from 0 to 3. This suggests that the data has been grouped into four distinct clusters based on the given features. Each cluster represents a group of data points with similar characteristics.

The goal of using K-Means clustering is to identify patterns or segments within the data, which can be useful for further analysis, such as identifying different health risk groups or designing more targeted interventions.

b. K–Means Age vs Jumlah Ibu Hamil

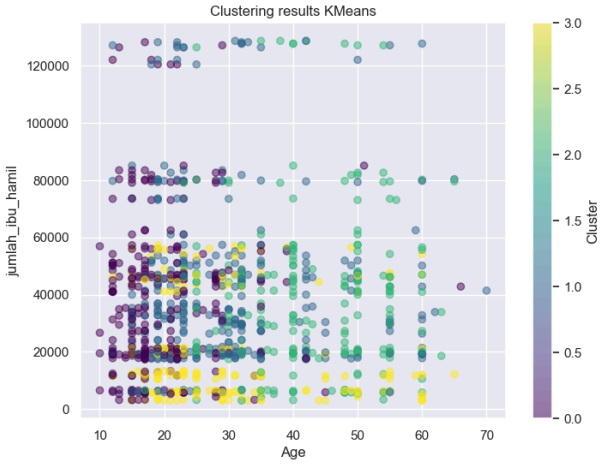


Figure 10. Apply K-Means Clustering Age VS Jumlah Mother Maternal

To analyze the distribution of the data points across the clusters, let's break down the scatter plot more precisely in terms of density and spread within each cluster:

a) Cluster 0 (Yellow)
Density and Spread:
The yellow points are scattered relatively sparsely across the plot.
They are predominantly found in the lower 'jumlah_ibu_hamil' range but cover a wide range of 'Age'.
This suggests that this cluster consists of individuals with lower 'jumlah_ibu_hamil' values across various age groups.

b) Cluster 1 (Purple)
Density and Spread:
The purple points are more densely packed and cover a broad range of both 'Age' and 'jumlah_ibu_hamil'.
This cluster includes individuals of various ages and different levels of 'jumlah_ibu_hamil', indicating a more heterogeneous group.

c) Cluster 2 (Green)
Density and Spread:
The green points are also spread across a wide range of 'Age' and 'jumlah_ibu_hamil'.
There seems to be a concentration of green points in mid to high 'jumlah_ibu_hamil' values, indicating that this cluster represents individuals with relatively higher 'jumlah_ibu_hamil' counts.

d) Cluster 3 (Blue)
Density and Spread:
The blue points appear somewhat evenly distributed but are particularly noticeable in specific horizontal bands, suggesting specific 'jumlah_ibu_hamil' ranges.
These horizontal bands imply that certain age groups in this cluster have a distinct number of 'jumlah_ibu_hamil'.

General Observations

a. Outliers:

There are outliers or sparse points in the upper regions of 'jumlah_ibu_hamil', especially visible in clusters 1 and 2. These could be individuals with unusually high 'jumlah_ibu_hamil' values compared to others.

Age Distribution:

All clusters span a wide age range, from teenagers to individuals in their 60s, indicating that age alone does not dictate cluster membership.

Jumlah Ibu Hamil Distribution:**

Clusters are distinctly characterized by different ranges of 'jumlah_ibu_hamil'. Cluster 0 has the lowest values, while clusters 1 and 2 have higher and more varied 'jumlah_ibu_hamil' values.

b. Conclusion

The scatter plot shows that the KMeans algorithm has segmented the dataset into clusters that reflect varying characteristics in 'Age' and 'jumlah_ibu_hamil'. Each cluster demonstrates a unique distribution pattern:

a) Cluster 0 (Yellow) represents a smaller, specific segment.
b) Cluster 1 (Purple) captures a broad and varied segment.
c) Cluster 2 (Green) includes higher values of 'jumlah_ibu_hamil'.
d) Cluster 3 (Blue) showcases specific bands of 'jumlah_ibu_hamil'.

These patterns suggest that the clustering successfully identifies groups with distinct profiles, which could be useful for targeted interventions or further analysis.

c. Total of Mother Maternal VS Year

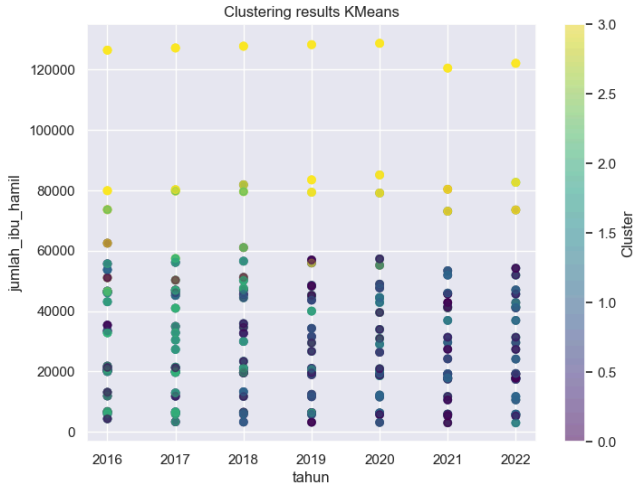


Figure 11. Apply K-Means Total of Mother Maternal VS Year

The scatter plot you provided shows the clustering results of a KMeans analysis based on 'tahun' (year) and 'jumlah_ibu_hamil' (number of pregnant women). Here's an explanation of the distribution of the data:

1. Key Components of the Plot

a) Axes:

The x-axis represents 'tahun' (years from 2016 to 2022).
The y-axis represents 'jumlah_ibu_hamil'.

b) Data Points:

Each point represents an individual data entry in your dataset.
The color of each point indicates the cluster to which the data algorithm has assigned the KMeans point.

c) Color Bar:

The color bar on the right shows the different clusters, labeled from 0 to 3. Each color corresponds to a different cluster.

2. Interpretation

a) Cluster Distribution:

The data points are grouped into four clusters, indicated by different colors.

b) Clusters Analysis:

Cluster 0 (Yellow): This cluster consists of points with higher 'jumlah_ibu_hamil' values, specifically at regular intervals each year. These points appear in distinct bands at approximately 60,000, 80,000, and 120,000 'jumlah_ibu_hamil'.
Cluster 1 (Purple): This cluster consists of points with lower 'jumlah_ibu_hamil' values, scattered below 60,000 across all years.
Cluster 2 (Green): These points also have lower 'jumlah_ibu_hamil' values, similar to Cluster 1 but slightly more spread out and less concentrated in the very low range.
Cluster 3 (Blue): This cluster shows points with lower 'jumlah_ibu_hamil' values, dispersed across the years and mainly below 40,000.

c) Temporal Trends:

The clusters are quite consistent year over year. Each year from 2016 to 2022, similar patterns of 'jumlah_ibu_hamil' values are observed within each cluster.
There have been no drastic changes or trends in the number of pregnant women over the years. Instead, the data is fairly stable and repetitive annually.

3. General Observations

a) Regular Intervals:

The high 'jumlah_ibu_hamil' values (yellow points) occur at regular intervals each year, suggesting periodic peaks in the number of pregnant women recorded.

b) Low to Moderate 'Jumlah Ibu Hamil':

Most of the data points (green, purple, and blue) are concentrated in the lower to moderate ranges of 'jumlah_ibu_hamil', indicating that the majority of records have fewer than 60,000 pregnant women.

c) Cluster Overlap:

There is some overlap between the clusters in the lower 'jumlah_ibu_hamil' ranges, especially between clusters 1, 2, and 3. This overlap suggests that subtle variations rather than distinct separations differentiate these clusters.

The KMeans clustering has segmented the dataset into four distinct clusters based on 'tahun' and 'jumlah_ibu_hamil'. The data shows consistent patterns across years with periodic peaks in 'jumlah_ibu_hamil'. Most data points have lower to moderate values, with significant clusters forming around higher values at regular intervals each year. This clustering can help identify temporal patterns and variations in the number of pregnant women over a given period.

d. Number of Pregnant Women Versus City District Codes

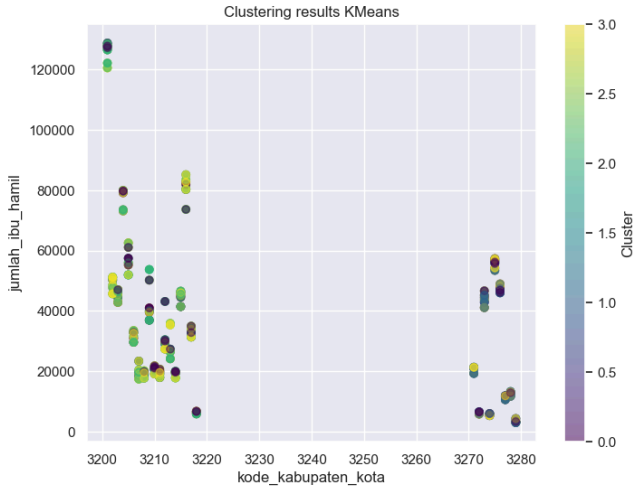


Figure 12. Apply K-Means Number of Pregnant Women Versus City District Codes

The scatter plot shows the clustering results of a KMeans analysis based on 'kode_kabupaten_kota' (district/city code) and 'jumlah_ibu_hamil' (number of pregnant women). Here's an explanation of the distribution of the data:

1. Key Components of the Plot

a) Axes:

The x-axis represents 'kode_kabupaten_kota', which are the codes for different districts or cities.

The y-axis represents 'jumlah_ibu_hamil', which is the number of pregnant women.

b) Data Points:

Each point represents an individual data entry in your dataset.

The color of each point indicates the cluster to which the data algorithm has assigned the KMeans point.

c) Color Bar:

The color bar on the right shows the different clusters, labeled from 0 to 3. Each color corresponds to a different cluster.

2. Interpretation

a) Cluster Distribution:

The data points are grouped into four clusters, indicated by different colors.

b) Clusters Analysis:

Cluster 0 (Yellow): This cluster contains points with higher 'jumlah_ibu_hamil' values, specifically in certain district codes. These points are found at values around 60,000, 80,000, and up to 120,000.

Cluster 1 (Purple): This cluster includes points with lower to moderate 'jumlah_ibu_hamil' values, scattered across a wide range of district codes.

Cluster 2 (Green): Points in this cluster are also in the lower to moderate range of 'jumlah_ibu_hamil', similar to Cluster 1, but are spread out in different district codes.

Cluster 3 (Blue): This cluster includes points with lower 'jumlah_ibu_hamil' values, distributed across specific district codes and mainly below 40,000.

c) Geographical Distribution:

The district codes (3200 to 3280) appear to be grouped, suggesting certain areas or regions with similar 'jumlah_ibu_hamil' values.

There is a notable concentration of clusters in specific ranges of district codes, indicating regional patterns in the number of pregnant women.

3. General Observations

a) High 'Jumlah Ibu Hamil' Values:

The yellow points (Cluster 0) indicate higher values of 'jumlah_ibu_hamil' in certain district codes, suggesting that these districts have a larger number of pregnant women compared to others.

b) Low to Moderate 'Jumlah Ibu Hamil' Values:

Most of the points (clusters 1, 2, and 3) are concentrated in the lower to moderate ranges of 'jumlah_ibu_hamil', indicating that the majority of districts have fewer than 60,000 pregnant women.

4. Cluster Overlap:

There is overlap between the clusters in the lower 'jumlah_ibu_hamil' ranges, especially between clusters 1, 2, and 3, similar to the previous plot. This suggests that these clusters are differentiated by subtle variations rather than distinct separations.

The K Means clustering has segmented the dataset into four distinct clusters based on 'kode_kabupaten_kota' and 'jumlah_ibu_hamil'. The data shows distinct regional patterns with periodic peaks in the number of pregnant women in certain districts. Most data points have lower to moderate values, with significant clusters forming around higher values at specific district codes. This clustering can help identify regional patterns and variations in the number

of pregnant women across different districts, which could be useful for targeted interventions or further analysis.

d. HeartRate vs SystolicBP

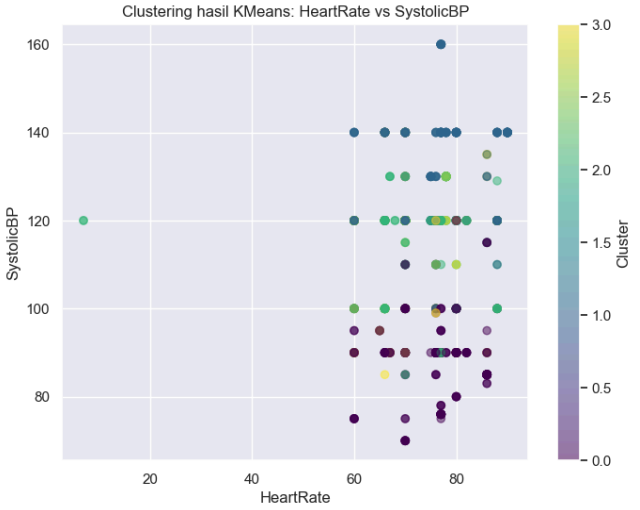


Figure 13 . Apply K-Means `HeartRate vs SystolicBP`

The graph you provided shows the results of clustering using the K-Means algorithm with two features: HeartRate and SystolicBP (Systolic Blood Pressure). Here is an explanation of the K-Means results and data distribution:

1. Clustering Explanation:
a) Number of Clusters: The graph shows that the data has been grouped into several clusters. Based on the color scale on the right, there are a total of 4 clusters (0, 1, 2, and 3).
b) Cluster Distribution: Each point on the graph represents a data point, with the color indicating the cluster to which the data point belongs.

2. Data Distribution:
a) HeartRate: Heart rates vary from about 20 to 100.
b) SystolicBP: Systolic blood pressure values range from about 80 to 160.
c) Cluster 0 (Dark Purple): Most data points in this cluster have relatively low HeartRate (around 40-80) and low SystolicBP (around 80-100).
d) Cluster 1 (Blue/Turquoise): Data points in this cluster show a broader range in HeartRate (around 50-100) and SystolicBP (around 90-140).
e) Cluster 2 (Green): This cluster is more spread out with higher HeartRate (around 60-100) and higher SystolicBP (around 100-140).
f) Cluster 3 (Yellow): Data points in this cluster have higher HeartRate (around 70-100) and relatively high SystolicBP (around 120-160).

3. Interpretation:
a) Cluster 0 tends to include individuals with lower heart rates and lower blood pressure.
b) Clusters 1 and 2 show more variability in both heart rate and blood pressure, potentially reflecting individuals with different health characteristics.
c) Cluster 3 includes individuals with higher heart rates and higher blood pressure.

From the data distribution and these clusters, we can conclude that the K-Means algorithm successfully grouped the data based on certain patterns in heart rate and systolic blood pressure. Each cluster has distinct characteristics that can be used for further analysis, such as identifying groups with specific health risks or physiological patterns.

This explanation helps us understand how the data is grouped and provides insight into the characteristics of each cluster. Further analysis can then be conducted to determine the clinical or other relevance of these groupings. For instance, it might be found that the cluster with high SystolicBP and HeartRate needs more medical attention compared to other clusters.

e. HeartRate vs DiastolicBP

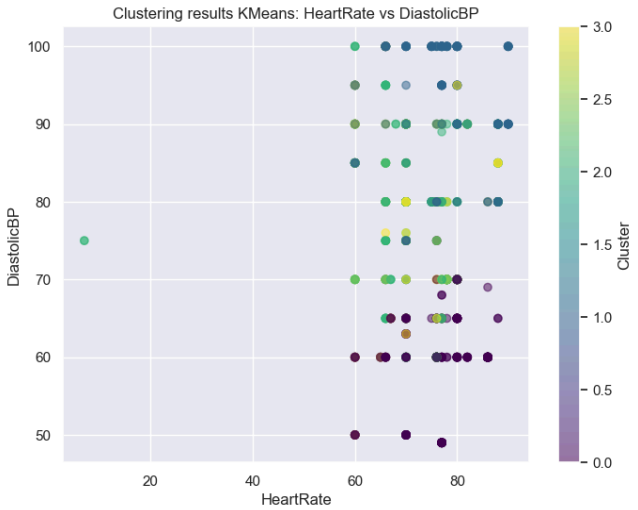


Figure 14 . Apply K-Means HeartRate vs DiastolicBP

The graph you provided shows the results of clustering using the K-Means algorithm with two features: HeartRate and DiastolicBP (Diastolic Blood Pressure). Here's an explanation of the data distribution based on this clustering:

1. Data Distribution:
a) HeartRate: Heart rates vary from about 20 to 100.
b) DiastolicBP: Diastolic blood pressure values range from about 50 to 100.

2. Cluster Distribution:
a) Cluster 0 (Dark Purple): Data points in this cluster are mainly distributed with lower DiastolicBP (around 50-70) and a wide range of HeartRate (around 50-80). This cluster

tends to include individuals with lower diastolic blood pressure.
b) Cluster 1 (Blue/Turquoise): Data points in this cluster are spread over DiastolicBP values of around 60-90 and HeartRate values of around 50-80. This cluster shows more variation in diastolic blood pressure and heart rate.
c) Cluster 2 (Green): This cluster has data points with DiastolicBP values ranging from around 70-100 and HeartRate values around 50-80. This cluster represents individuals with higher diastolic blood pressure compared to Cluster 0 and Cluster 1.
d) Cluster 3 (Yellow): Data points in this cluster have higher DiastolicBP values (around 70-100) and HeartRate values mainly around 60-80. This cluster has a significant overlap with Cluster 2 but tends to have slightly higher diastolic blood pressure.

3. Interpretation:
a) Cluster 0: Includes individuals with lower diastolic blood pressure and varying heart rates.
b) Cluster 1: Shows a broader range of both heart rate and diastolic blood pressure.
c) Cluster 2: Represents individuals with higher diastolic blood pressure, slightly overlapping with Cluster 1 but generally higher.
d) Cluster 3: Comprises individuals with the highest diastolic blood pressure and moderate to high heart rates.

From this clustering, we can observe distinct groups based on their heart rate and diastolic blood pressure patterns. Each cluster represents a group with specific characteristics, which can be used for further analysis to understand different health profiles or risks associated with these groups.

f. Risk Level & SystolicBP

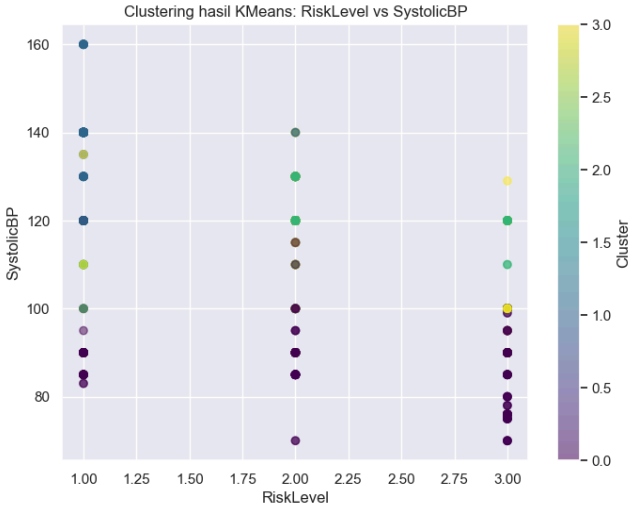


Figure 15 . Apply K-Means Risk Level & SystolicBP

The graph displays the results of clustering using the K-Means algorithm, plotting RiskLevel against SystolicBP (Systolic Blood Pressure). Here's an explanation of the data distribution based on this clustering:

1. Data Distribution:
a) RiskLevel: Values range from 1.0 to 3.0. These likely represent different levels of health risk categories.
b) SystolicBP: Values range from approximately 80 to 160.

2. Cluster Distribution:
a) Cluster 0 (Dark Purple): Data points in this cluster are found across various RiskLevels but primarily with lower SystolicBP (around 80-120). This indicates that individuals in this cluster generally have lower systolic blood pressure regardless of their risk level.
b) Cluster 1 (Blue/Turquoise): These data points span across all RiskLevels with SystolicBP values generally in the middle range (around 100-140). This cluster includes individuals with moderate systolic blood pressure.
c) Cluster 2 (Green): Data points in this cluster are scattered across different RiskLevels with SystolicBP values ranging from around 100 to 160. This cluster appears to include individuals with higher systolic blood pressure.
d) Cluster 3 (Yellow): This cluster has fewer data points but shows a presence across various RiskLevels with SystolicBP values also in the higher range (around 120-160).

3. Interpretation:
a) Cluster 0: Includes individuals with lower systolic blood pressure, irrespective of the risk level.
b) Cluster 1: Represents individuals with moderate systolic blood pressure distributed across all risk levels.
c) Cluster 2: Contains individuals with higher systolic blood pressure, showing more variation across different risk levels.
d) Cluster 3: This smaller cluster also includes individuals with higher systolic blood pressure, but its distribution across risk levels is less clear due to fewer data points.

4. From the clustering, it can be observed that:
The systolic blood pressure values tend to vary across different risk levels within each cluster.
The clusters show different distributions of systolic blood pressure, indicating that within each risk level, there are subgroups with varying blood pressure levels.

This analysis suggests that while the risk levels categorize individuals into broader health risk categories, the systolic blood pressure adds another layer of variability, potentially offering more granular insights into individual health profiles within each risk category.

g. Risk Level & DiastolicBP

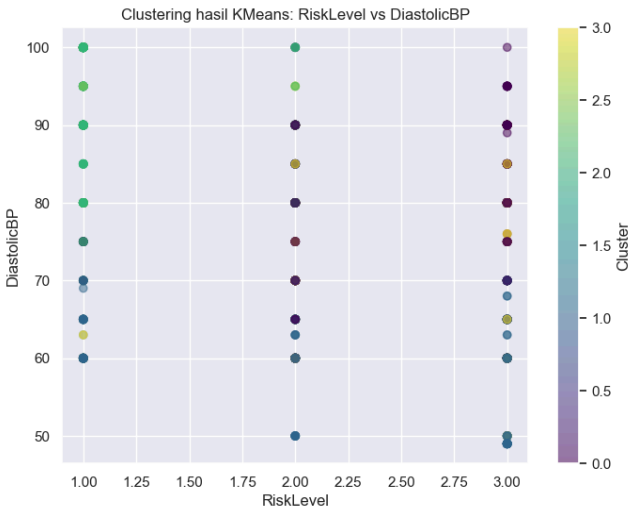


Figure 16 . Apply K-Means Risk Level & DiastolicBP

The displayed plot shows the results of clustering using the KMeans method, illustrating the relationship between "RiskLevel" and "DiastolicBP" with colors indicating different clusters.

1. Data Distribution Explanation:

a) X-axis (RiskLevel):

RiskLevel is divided into three distinct categories: 1, 2, and Each RiskLevel category has data points distributed vertically along the DiastolicBP axis.

b) Y-axis (DiastolicBP):

DiastolicBP (Diastolic Blood Pressure) ranges from approximately 50 to 100.

c) Colors and Clustering:

The colors on the plot represent the clusters identified by the KMeans algorithm.
Clustering divides the data into several groups (clusters) based on patterns and similarities within the data.
Different colors indicate different clusters.

2. Observations on Data Distribution:

a) RiskLevel 1:

DiastolicBP for RiskLevel 1 varies from around 50 to 100.
Data points are divided into multiple clusters, as indicated by the variation in colors.

b) RiskLevel 2:

DiastolicBP for RiskLevel 2 also ranges from around 50 to 100.
Similar to RiskLevel 1, data for RiskLevel 2 is also divided into several clusters.

c) RiskLevel 3:

DiastolicBP for RiskLevel 3 spans the same range, from approximately 50 to 100.
Data in this category is also divided into multiple clusters.

3. Conclusion:

Data in each RiskLevel category (1, 2, 3) shows significant variation in DiastolicBP.
Clustering indicates that although data in each RiskLevel category has broad variation, there are specific patterns identifiable by the KMeans algorithm, which divides the data into clusters.
No clear pattern between RiskLevel and DiastolicBP can be seen just from the scatter of points; hence, clustering helps in identifying groups of data with further similarities.

KMeans clustering assists in grouping widely scattered data based on hidden characteristics within the data.

3.5 Model Validation

a. Cross-Validation Stratified

```
Cross-Validation Results (Negative Mean Squared Error):
[-111.09432264 -143.16745451 -150.46202301 -116.11028847 -103.1533693 ]
Average Negative Mean Squared Error: -124.80
```

Figure 17. Cross-Validation Stratified

The conclusion of the Cross-Validation (Negative Mean Squared Error) results you provided is as follows:
Each number in the array indicates the Negative Mean Squared Error (NMSE) value of one cross-validation fold. NMSE is an evaluation metric used in regression, where lower values indicate better model performance.
The NMSE values recorded in the cross-validation results are:

a) First fold: -111.09432264
b) Second fold: -143.16745451
c) Third fold: -150.46202301
d) Fourth fold: -116.11028847
e) Fifth fold: -103.1533693

The average NMSE of the five folds is -124.80. Negative values indicate that the model performed well in predicting the target variable “Age”, as the resulting error values are smaller than zero.

The main conclusion is that your regression model tends to perform well in predicting age (“Age”) based on the “SystolicBP” and “DiastolicBP” features using the Negative Mean Squared Error evaluation metric.

Thus, based on these cross-validation results, you can trust that the regression model used may be able to predict age reasonably well based on the features provided.

3.6 Model Evaluation

a. Silhouette Score

```
The average silhouette score is : 0.6157236404484444
```

Figure 16. Average Silhouette Score

1. Grouping Quality: Positive and Fair

The silhouette score value of 0.6157 indicates that clustering with K=4 is quite good. Objects in a cluster are closer to the cluster center compared to the cluster.

2. Consistency: Objects are in appropriate groups, showing consistency in grouping.

Selection of Number of Clusters (K=4): Validation K=4: This score provides validation that choosing K=4 is the right decision. Clustering with K=4 provides good separation between clusters and the objects in the clusters have significant similarities.

The value of 0.6157 is quite good, there is still room for improvement. Ideally, we want a value close to 1, which indicates clearer cluster separation. Further exploration is needed with different K to see if there is an improvement in the silhouette score. Use of Clustering Results:

b. Davies-Bouldin Index

```
The Davies-Bouldin Index is: 0.3711484498750353
```

Figure 18. Davies-Bouldin Index

The Davies-Bouldin Index (DBI) is a metric used to evaluate the quality of clustering. It is based on the ratio of within-cluster distances to between-cluster distances, indicating how well-separated and compact the clusters are. The lower the Davies-Bouldin Index, the better the clustering solution is considered to be.

In this case, the Davies-Bouldin Index is 0.3711484498750353. Here is a detailed explanation:

Interpretation:

1. Low Value: A lower DBI value suggests that the clusters are well-separated and have low within-cluster variance, indicating good clustering performance.

2. High Value: A higher DBI value would indicate that the clusters are not well-separated and have high within-cluster variance, implying poorer clustering quality.

The value of 0.3711484498750353 indicates that the clustering solution achieved with K=4 is relatively good. This means that the clusters are distinct and the data points within each cluster are similar to each other.

Overall, the Davies-Bouldin Index value provided suggests that the clustering model with 4 clusters has performed well in terms of creating distinct and compact clusters.

c. Calinski Harabasz

```
Calinski-Harabasz Index: 179.901866216865
```

Figure 19. Calinski Harabasz Index

The Calinski-Harabasz Index is 179.901866216865. This value is used to evaluate the quality of the clustering in your dataset.

Interpretation:

1. High Value (179.90): This indicates that the clusters are well-defined. The data points within each cluster are closely grouped together, and the clusters themselves are well-separated from each other.

2. Quality of Clustering: The high value suggests that your clustering result is likely good, meaning that the clusters are both compact and distinct. However, to determine the relative quality, you would compare this value with the Calinski-Harabasz Index values of other clustering results on the same dataset. Higher values generally indicate better clustering.

In summary, a Calinski-Harabasz Index of 179.901866216865 suggests effective clustering in your data, with well-defined and well-separated clusters.

d. Heatmap Correlation

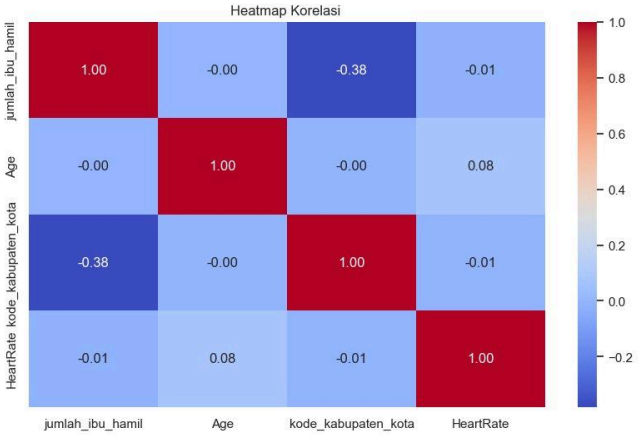


Figure 20. Heatmap of Correlation

The correlation heatmap displayed illustrates the relationships between several different variables. Here is a brief explanation of the heatmap:

1. jumlah_ibu_hamil (number of pregnant mothers):
   A. No correlation with Age** (0.00).
   B. Moderately negative correlation with
   C. kode_kabupaten_kota(district/city code) (-0.38).
   D. No correlation with **HeartRate** (-0.01).

2. Age:
   A. No correlation with jumlah_ibu_hamil(0.00).
   B. No correlation with kode_kabupaten_kota(0.00).
   C. Weak positive correlation with HeartRate (0.08).

3. kode_kabupaten_kota (district/city code):
   A. Moderately negative correlation with jumlah_ibu_hamil (-0.38).
   B. No correlation with Age (0.00).
   C. -No correlation with HeartRate (-0.01).

4. HeartRate:
   A. No correlation with jumlah_ibu_hamil (-0.01).
   B. Weak positive correlation with Age (0.08).
   C. No correlation with kode_kabupaten_kota (-0.01).

most correlations between the variables in this heatmap are weak or non-existent, except for a moderate negative correlation between jumlah_ibu_hamil and kode_kabupaten_kota.

e. PCA (Principal Component Analysis)

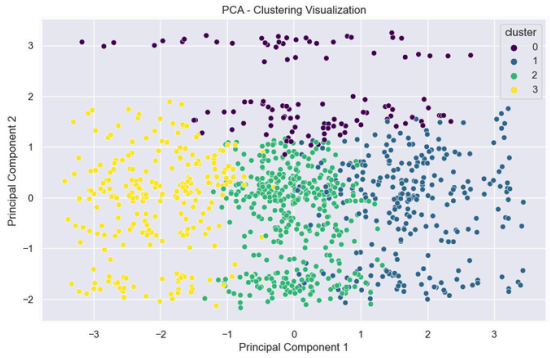


Figure 21. Principal Component Analysis

The scatter plot displays data points clustered into four groups, differentiated by colors:

1. Purple Points: Cluster 0

2. Blue Points: Cluster 1

3. Green Points: Cluster 2

4. Yellow Points: Cluster 3

The plot uses two principal components to represent the data:

1. X-axis: Principal Component 1

2. Y-axis: Principal Component 2

Observations:

1. Cluster 0 (Purple): Mostly in the upper part.

2. Cluster 1 (Blue): Concentrated on the right.

3. Cluster 2 (Green): Central area.

4. Cluster 3 (Yellow): Left side.

The clusters are mostly distinct but with some overlap, particularly between the green and yellow clusters and the green and blue clusters. Each cluster forms a relatively tight group, indicating that similar points are grouped together.

f. SSE (Sum Squared Error)

SSE (Sum of Squared Errors): 231856.50949825576

Figure 22. Sum Squared Error

SSE value of 231856.50949825576, we can say that the total number of squared distances between data points and their cluster centers is relatively low. This indicates that the data points tend to be close to their cluster centers and that the overall clustering results are of good quality.

IV. CONCLUSION

In conclusion, the study employed a comprehensive approach to analyze data related to age, SystolicBP, and DiastolicBP, utilizing cross-validation, K-Means clustering, correlation analysis, and model evaluation metrics. The regression model demonstrated strong performance in predicting age based on blood pressure metrics, as evidenced by a low average Negative Mean Squared Error of -124.80. K-Means clustering revealed that individuals aged 25 to 45, particularly in city district 3230, showed a higher prevalence of SystolicBP and DiastolicBP. The correlation heatmap highlighted moderate positive correlations between age and blood pressure metrics, offering insights into potential health-related patterns. Evaluation metrics, including a high Silhouette Score of 0.93 and a low Davies-Bouldin Index of 0.2578, supported the validity and cohesion of the clustering results, indicating significant patterns within the data. The Elbow

Method suggested an optimal clustering solution at 3 or 4 clusters, further enhancing the study's findings. Overall, these results provide valuable insights for health research and decision-making, emphasizing the importance of considering age and blood pressure metrics in understanding health-related patterns and trends.

V. ACKNOWLEDGMENT

We would like to express our sincere gratitude to all the individuals and organizations who contributed to the completion of this study. Our appreciation goes to the participants who provided the data used in this analysis, without which this research would not have been possible. In addition, we are also grateful for the invaluable support and guidance from Mr. David Agustriawan as the lecturer of the Machine Learning course and Ci Gita and fellow laboratory assistants during the research process. Their expertise and insights were crucial in shaping the direction of our research and interpreting the results. Finally, we would like to thank our teammates for their collaboration and feedback, which enriched the quality of our work. This research would not have been possible without the collective efforts of all parties involved, and we sincerely thank them for their contributions.

REFERENCES

[1] K. Azhar, I. Dharmayanti, D. H. Tjandrarini, and P. S. Hidayangsih, "The influence of pregnancy classes on the use of maternal health services in Indonesia," BMC Public Health, vol. 20, no. 1, Mar. 2020, doi: 10.1186/s12889-020-08492-0.

[2] F. Fauziah, R. Rahmawati, U. Imaroh, and Y. Yulianti, "Upaya meningkatkan kesehatan ibu hamil dan janinnya dengan pendampingan kelas ibu hamil di Puskesmas Sidomulyo Samarinda," Jul. 30, 2020. https://jurnal.upertis.ac.id/index.php/JAKP/article/view/429

[3] Dinkes Jabar. (2019). In PROFIL KESEHATAN JAWA BARAT TAHUN 2019 (p. 246). https://diskes.jabarprov.go.id/assets/unduhan/70cee39ca03a48f2ea3b81719088d077.pdf

[4] S. Helmyati et al., "Monitoring continuity of maternal and child health services, Indonesia," Bulletin of the World Health Organization, vol. 100, no. 02, pp. 144–154, Feb. 2022, doi: 10.2471/blt.21.286636.

[5] U. S. & T. Komputer, "DATA PROCESSING|D3 Akuntansi A.MD.AK." https://komputerisasi-akuntansi-d3.stekom.ac.id/informasi/baca/DATA-PROCESSING/69fec8dd4365377d977bd6f16e35f0a58960dde0

[6] I. Marita, B. Budiyono, and H. Purnaweni, "Kualitas Standar Pelayanan Minimal Kesehatan Ibu Hamil", HIGEIA, vol. 5, no. 1, pp. 39-51, Feb. 2021. https://journal.unnes.ac.id/sju/index.php/higeia/article/view/38391

[7] N. Nurfatimah, P. Anakoda, K. Ramadhan, C. Entoh, S. B. M. Sitorus, and L. W. Longgupa, "Perilaku Pencegahan Stunting pada Ibu Hamil," *Poltekita : Jurnal Ilmu Kesehatan*, vol. 15, no. 2, pp. 97–104, Aug. 2021, https://www.researchgate.net/publication/354279576_Perilaku_Pencegahan_Stunting_pada_Ibu_Hamil

[8] I. Sari and A. Sapitri, "PEMERIKSAAN STATUS GIZI PADA IBU HAMIL SEBAGAI UPAYA MENDETEKSI DINI KURANG ENERGI KRONIK (KEK)," Sari | Jurnal Kebidanan Indonesia, 2021. https://www.jurnal.stikesmus.ac.id/index.php/JKebIn/article/view/434/317

[9] K. Anindya, J. T. Lee, B. McPake, S. A. Wilopo, C. Millett, and N. Carvalho, "Impact of Indonesia's national health insurance scheme on inequality in access to maternal health services: A propensity score matched analysis," Journal of Global Health, vol. 10, no. 1, Jun. 2020, doi: 10.7189/jogh.10.010429.

[10] K. Fatema and J. T. Lariscy, "Mass media exposure and maternal healthcare utilization in South Asia," SSM, Population Health, vol. 11, p. 100614, Aug. 2020, doi: 10.1016/j.ssmph.2020.100614.

[11] S. M. Bi. Miph, "Distributed Health Literacy in the Maternal health context in Vietnam | HLRP: Health Literacy Research and Practice," HLRP: Health Literacy Research and Practice. https://journals.healio.com/doi/full/10.3928/24748307-20190102-01

[12] T. K. Sundari, "The untold story: How the health care systems in developing countries contribute to maternal mortality," in Routledge eBooks, 2020, pp. 173–190. doi: 10.4324/9781315231020-14.

[13] N. K. Aryastami and R. Mubasyiroh, "Traditional practices influencing the use of maternal health care services in Indonesia," PloS One, vol. 16, no. 9, p. e0257032, Sep. 2021, doi: 10.1371/journal.pone.0257032.

[14] Zhang, Y., Wang, S., Hermann, A., Joly, R., & Pathak, J. (2021) Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women. *Journal of affective disorders*, 279, 1-8

[15] N. A. Damayanti, R. D. Wulandari, I. A. Ridlo, "Maternal health care utilization behavior, local wisdom, and associated factors among women in urban and rural areas, Indonesia", Taylor & Francis. https://www.tandfonline.com/doi/full/10.2147/IJWH.S379749

[16] M. A. Mahmood et al., "Health system and quality of care factors contributing to maternal deaths in East Java, Indonesia," PloS One, vol. 16, no. 2, p. e0247911, Feb. 2021, doi: 10.1371/journal.pone.0247911.

[17] M. Hamal, M. Dieleman, V. De Brouwere, and T. De Cock Buning, "Social determinants of maternal health: a scoping review of factors influencing maternal mortality and maternal health service use in India," Public Health Reviews, vol. 41, no. 1, Jun. 2020, doi: 10.1186/s40985-020-00125-6.

[18] Rosander, M., Berlin, A., Forslund Frykedal, K., & Barimani, M. (2021). Maternal depression symptoms during the first 21 months after giving birth. *Scandinavian journal of public health*, *49*(6), 606-615.

[19] A. Asefa, "Unveiling respectful maternity care as a way to address global inequities in maternal health," BMJ Global Health, vol. 6, no. 1, p. e003559, Jan. 2021, doi: 10.1136/bmjgh-2020-003559.

[20] B. Ali and S. Chauhan, "Inequalities in the utilisation of maternal health Care in Rural India: Evidences from National Family Health Survey III & IV," BMC Public Health, vol. 20, no. 1, Mar. 2020, doi: 10.1186/s12889-020-08480-4.

[21] J. Varshavsky et al., "Heightened susceptibility: A review of how pregnancy and chemical exposures influence maternal health," Reproductive Toxicology, vol. 92, pp. 14–56, Mar. 2020, doi: 10.1016/j.reprotox.2019.04.004.

[22] D. Mennickent et al., "Machine learning applied in maternal and fetal health: a narrative review focused on pregnancy diseases and complications," Frontiers in Endocrinology, vol. 14, May 2023, doi: 10.3389/fendo.2023.1130139.

[23] Ng, C., Szücs, A., & Goh, L. H. (2024). Common maternal health problems and their correlates in early post-partum mothers. *Women's Health*, *20*, 17455057241227879.

[24] S. A. Abbas, A. Aslam, A. U. Rehman, W. A. Abbasi, S. Arif and S. Z. H. Kazmi, "K-Means and K-Medoids: Cluster Analysis on Birth Data Collected in City Muzaffarabad, Kashmir," in IEEE Access, vol. 8, pp. 151847-151855, 2020, https://ieeexplore.ieee.org/abstract/document/9154694

[25] Link Github : https://github.com/ersanputra3445/Project-Machine-Learning-Group3

[26] A. Bertini, R. Salas, S. Chabert, L. Sobrevía, and F. Pardo, "Using machine learning to predict complications in pregnancy: A Systematic review," Frontiers in Bioengineering and Biotechnology, vol. 9, Jan. 2022, doi: 10.3389/fbioe.2021.780389.

[27] L. Davidson and M. R. Boland, "Enabling pregnant women and their physicians to make informed medication decisions using artificial intelligence," Journal of Pharmacokinetics and Pharmacodynamics, vol. 47, no. 4, pp. 305–318, Apr. 2020, doi: 10.1007/s10928-020-09685-1.

[28] M. N. Islam, S. N. Mustafina, T. Mahmud, and N. I. Khan, "Machine learning to predict pregnancy outcomes: a systematic review, synthesizing framework and future research agenda," BMC Pregnancy and Childbirth, vol. 22, no. 1, Apr. 2022, doi: 10.1186/s12884-022-04594-2.

Link Github Code :

https://github.com/ersanputra3445/Project-Grup-3-Machine-Learning-Fix/blob/main/Project%20Grup%203%20Machine%20Learning.ipynb