Transforming Numerical Data to Categorical Using Clustering K-Means for Naive Bayes Application to Diabetes Prediction

Desvania Tirta Izzati[1], Syifa Rahmadani Hemi Syafitri[2], and Nabila Winanda Meirani[3]

[1,2,3]Informatics, Universitas Jenderal Soedirman, Indonesia
[1]H1D022088_Desvania, [2]H1D022096_Syifa, [3]H1D022108_Nabila

Email: [1]desvania.izzati@mhs.unsoed.ac.id@, [2]syifa.syafitri@mha.unsoed.ac.id,
[3]nabila.meirani@mhs.unsoed.ac.id

Received : Jun 9, 2024; Revised : Jun 9, 2024; Accepted : Jun 9, 2024; Published : Jun 9, 2024

Abstrak

Diabetes is a chronic condition that significantly impacts kidney, eye, and heart health over time. Among its various types, type 2 diabetes, or diabetes mellitus, is particularly prevalent and characterized by chronic hyperglycemia due to impaired insulin secretion or action. Data mining techniques have become essential in medical diagnostics and analysis, facilitating decision-making processes through the discovery of patterns and relationships within large datasets. This study focuses on enhancing the performance of the Naive Bayes classifier, which traditionally performs better with categorical data, by transforming numerical data into categorical data. The research demonstrates that the accuracy of the Naive Bayes model is lower when handling numerical data compared to categorical data. To achieve this transformation, clustering methods were employed, specifically utilizing the elbow method to determine the optimal number of clusters. The elbow method, a well-established technique, helps estimate the ideal cluster number for the dataset under analysis. Following the determination of optimal clusters, the K-means clustering method was applied to the dataset. K-means clustering, a non-hierarchical analysis technique, groups objects with similar characteristics into clusters. This process results in a categorical dataset that enhances the Naive Bayes classifier's performance. The study's findings reveal a significant improvement in the model's accuracy, increasing from 77% with numerical data to 95% with categorical data post-clustering. These results underscore the importance of pre-processing techniques, such as clustering, in optimizing classification models for better predictive accuracy in medical data analysis.

Keywords : *Clustering, Diabetes Mellitus, Diabetes Prediction, Elbow Method, Gaussian Naive Bayes, K-Means Clustering, Medical Data Analysis, Naive Bayes Classifier, Predictive Accuracy, Transforming Data.*

1. INTRODUCTION

Diabetes is a chronic condition defined by an elevated blood glucose level that directly affects the pancreas,and the body is incapable of producing insulin[1][2]. Diabetes is growing faster among the young adults and children because of the hectic schedule due to which they have less amount of time to perform physical activity[3]. Diabetes has several types, one of which is type 2 diabetes, also known as diabetes mellitus is a signifcant public health problem globally[4]. Diabetes mellitus (DM) is a group of metabolic diseases characterized by chronic hyperglycemia, which results from impaired insulin secretion, insulin action, or both[5]. Prediction for an accurate diagnosis of diabetes mellitus, especially in the early days of its development, is a challenge for medical professionals[2].

Machine learning is an artificial intelligence (AI)-based application that automatically builds analytical models that can learn from data, identify patterns, and make decisions with minimal latency [6]. in this study to predict diabetes using various machine learning methods and identify the most efficient and accurate one[7]. Numerous organizations utilize data mining to analyze enormous datasets, to enhance the decision-making process, and to obtain better long-term results[1]. The machine learning method used in this research is clustering and classification.

Clustering used in this study to determine the accuracy and processing time of a non-hierarchical kernel technique for unsupervised clustering, namely K-means[8]. In this study, we aim to transform numerical data into categorical data, as Naive Bayes performs better with categorical data. It is demonstrated that the accuracy is lower when the data type is numerical compared to the accuracy after the data type is converted to categorical. In the process of converting numerical data to categorical data, clustering methods are employed, including the elbow method. The elbow method, the oldest visual method for estimating the potential optimal cluster number for the analyzed dataset[9],the elbow method is utilized to determine the optimal number of clusters before implementing the clustering technique.

After determining the optimal number of clusters using the elbow method, the next step is to implement the K-means clustering method on the dataset. A non-hierarchical cluster analysis technique called K-means clustering aims to divide existing items into one or more clusters. aims to combine objects with similar qualities together into one or more clusters or groups of objects, in order to achieve the goal of grouping similar objects together[10].

2. METHOD

The research method used in this study can be seen in Figure 1.

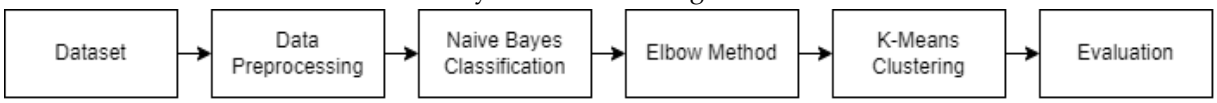

Figure 1. Method

2.1. Diabetes Disease Dataset

The dataset of people with diabetes in this study uses a genereal dataset that is open to public which can be accessed through the link https://www.kagle.com/datasets/uciml/pima-indians-diabetes-database. The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, glucose, blood pressure, skin thickness, their BMI, insulin level, and age. The description of each column is in Table 1 and the data type of each column is in Table 2.

Table 1. Dataset Description

No	Column	Description
1	Pregnancies	Pregnancy age in months
2	Glucose	Plasma glucose level or the amount of sugar (glucose) in the blood measured in mg/dL.
3	BloodPressure	Diastolic blood pressure, a measure of the pressure experienced by bloo in the arterial blood vessels, measured in mm Hg.
4	SkinThickness	Triceps skinfold thickness, measured in mm. This is a measurement o subcutaneous fat.
5	Insulin	Body insulin levels, measured in μU/mL.

6	BMI	Body Mass Index (BMI) is a medical screening tool that estimates th amount of body fat based on height and weight measurements.
7	DiabetesPedigreeFunction	Diabetes pedigree function, this value indicates the probability or risk o diabetes based on family history and genetic factors.
8	Age	Age of the individual in years
9	Outcome	A diabetes diagnostic result, where 1 indicates the individual has diabete and 0 indicates the individual does not have diabetes.

2.2. Data Preprocessing

At this stage, several processes are carried out to process data before classification using the Naïve Bayes algorithm and the K-Means algorithm. The process includes:

a. Missing Values
b. Feature Scaling

2.3. Naïve Bayes Classification

At this stage we use Naïve Bayes Classifier (NBC) which is a machine learning algorithm that can predict and calculate the posterior probability of a class based on the distribution of words in the document discussed in probability theory. According to the literature, NB is one of the best performing classifiers used in data mining[11][12]. The results of the algorithm was used to support the level of accuracy using the Confusion Matrix. The confusion matrix was obtained by calculating the value of precision, recall, and F-Measure[13]. Labels are represented in rows and predicted classes or labels are represented in columns [14]. The confusion matrix is shown in Figure 2

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Figure 2. Confusion Matrix

2.4. Elbow Method

This research was conducted using K-means clustering with the elbow method to determine an optimal number of clusters[15]. Elbow is believed to find the best number of clusters in the K-means method.

2.5. K-Means Clustering

This stage we use the K-Means Clustering method, which is a technique used to group data objects into groups that have certain similarities. K-means clustering algorithm is used to improve the accuracy of feature selection and Sum Squared Error (SSE) can be used to determine the appropriate number of centroids by implementing the elbow method[16][17]. This algorithm first randomizes k centers, calculates the distance from all data points in the sample data set to each center point, and assigns each data point to its nearest centroid classification, the set of points collected by the same centroid is a cluster, then, the centroid of each cluster is updated according to this classification result [18].

2.6. Evaluation

At this stage, it contains the evaluation results between naive bayes before clustering and naive bayes after clustering.

3. RESULT

This section is a discussion of the research that has been done. Starting from the preprocessing stage, Naive Bayes Classification, K-Means Clustering and evaluation.

3.1. Data Preprocessing

Prior to the data mining process, the dataset was first subjected to data preprocessing which included checking for missing value and feature scaling using StandardScaler from scikit-learn.

3.1.1. Missing Values

Researchers conducted data cleaning to clean up the missing value data[19]. This result shows that there are no missing values in any column of the dataset. As seen in the output, all columns have a value of 0 which indicates the number of missing values in each column of the dataset.

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

Figure 3. Result of Missing Value Calculation

3.1.2. Feature Scaling

Feature Scaling used in this research is normalization. Normalization is an additional step between the two possibilities in the dataset to be tested normal and non-normal distribution. The normal data is ready to be processed. Normalization rescales the data between 0 and 1[20][21].

This result shows the first five rows of the dataset that have been scaled. All features (except the Outcome column) have been processed using StandardScaler, which transforms the data to have a mean of 0 and a standard deviation of 1. This ensures that all features are similarly scaled, which is important for many machine learning algorithms. The Outcome column remains unchanged as this is the label to be predicted.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	0.639947	0.848324	0.149641	0.907270	-0.692891	0.204013	0.468492	1.425995	1
1	-0.844885	-1.123396	-0.160546	0.530902	-0.692891	-0.684422	-0.365061	-0.190672	0
2	1.233880	1.943724	-0.263941	-1.288212	-0.692891	-1.103255	0.604397	-0.105584	1
3	-0.844885	-0.998208	-0.160546	0.154533	0.123302	-0.494043	-0.920763	-1.041549	0
4	-1.141852	0.504055	-1.504687	0.907270	0.765836	1.409746	5.484909	-0.020496	1

Figure 4. Result of Feature Scaling

3.2. Naïve Bayes Classification

The purpose of this research is to improve the accuracy of Naive Bayes implementation by transforming the dataset type from numeric to categorical using k-means clustering technique. In order to achieve this goal, it is necessary to compare the results of Naive Bayes accuracy before and after optimization using clustering. Therefore, classification with Naive Bayes was performed using a dataset that had not gone through the k-means clustering process. Thus, we can evaluate the accuracy improvement obtained after applying clustering techniques as part of the optimization process.

In this study we used the Gaussian Naive Bayes algorithm because when applied to Gaussian Naive Bayes classification, it will produce high accuracy, precision, recall, and F1-score[22].

3.2.1. Data Preparation

The data preparation process starts with cleaning and splitting the dataset into features and labels. Features are columns that represent diabetes measurement parameters while labels are columns that show the results of diabetes diagnosis.

At first, unnecessary columns 'Outcome' were removed from the main dataset, resulting in Data Frame X containing only the relevant features for the model. Next, the 'Outcome' column is separated and stored in the y variable, which serves as the target label for classification.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	0.639947	0.848324	0.149641	0.907270	-0.692891	0.204013	0.468492	1.425995
1	-0.844885	-1.123396	-0.160546	0.530902	-0.692891	-0.684422	-0.365061	-0.190672
2	1.233880	1.943724	-0.263941	-1.288212	-0.692891	-1.103255	0.604397	-0.105584
3	-0.844885	-0.998208	-0.160546	0.154533	0.123302	-0.494043	-0.920763	-1.041549
4	-1.141852	0.504055	-1.504687	0.907270	0.765836	1.409746	5.484909	-0.020496

Figure 5. Features Column

```
0     1
1     0
2     1
3     0
4     1
5     0
6     1
7     0
8     1
9     1
10    0
Name: Outcome, dtype: int64
```

Figure 6. Outcome Column

After that, the dataset is divided into two parts: training set and test set. Train or test split in which the training dataset is used to construct the model whereas, the testing dataset is used to assess the model's predictive capability[23]. This is done using a function that divides X and y into X_train, X_test, y_train, and y_test with a proportion of 80% for training and 20% for testing, ensuring that model evaluation is performed on data not seen during training.

```
X_train =  614
X_test =  154
y_train =  614
y_test =  154
```

Figure 7. Numbers of each Data Split

The variables X_train and y_train store the training data, which consists of 614 samples. The training data is used to train the machine model so that the model can learn the patterns in the data. The variables X_test and y_test store the test data, which consists of 154 samples. The test data is used to test the performance of the model that has been trained on the training data. Using the test data, we can measure how well the model can generalize the patterns it has learned from the training data to new data that it has never seen before.

3.2.2. Naïve Bayes Model Training

After preparing the data, the next step is to train the Naive Bayes model. The Naive Bayes model used here is Gaussian Naive Bayes, implemented by the GaussianNB class. The model object is created by calling the GaussianNB() constructor. Then, the model is trained using the prepared training data X_train and y_train using the fit method.

The fit method in the Gaussian Naive Bayes model is a function used to train the model with the given data. It calculates the statistical parameters (mean and variance) required to make predictions based on the Gaussian distribution of each feature in the dataset for each class. This training process allows the model to learn the relationship between the features in X_train and the corresponding labels in y_train, so that it can be used to make predictions on new data.

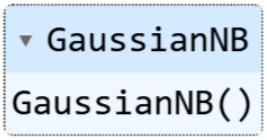

Figure 8. Gaussian Naive Bayes

3.2.3. Prediction Generation

Once the Gaussian Naive Bayes model is trained with the data training, the next step is to test the ability of the model by making predictions on the test data. This prediction process is done using the predict method from the trained model object. The predict method works by taking the features of each sample in X_test and predicting the corresponding label based on the knowledge gained from the training data. The prediction result is then stored in the variable y_pred, which contains the predicted label for each sample in X_test.

```
array([0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0,
       1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1,
       0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1,
       0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1,
       0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1,
       0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0])
```

Figure 9. Label Prediction Result

3.2.4. Classification Naïve Bayes Report Before Clustering

The last step is to evaluate the performance of the model. The predicted labels in y_pred can then be compared with the actual labels in y_test to evaluate the model performance. This model performance evaluation can be done with various metrics such as accuracy, precision, recall, and F1-score. This function generates a classification report that includes evaluation metrics such as precision, recall, f1-score, and support for each class. This report provides a comprehensive overview of the model's performance, including how accurately the model classifies each class and the balance between precision and recall. By evaluating the predicted results against the original labels, we can gauge how well the model classifies new samples and understand its strengths and weaknesses.

```
Accuracy: 0.7662337662337663
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.80      0.81        99
           1       0.66      0.71      0.68        55

    accuracy                           0.77       154
   macro avg       0.75      0.75      0.75       154
weighted avg       0.77      0.77      0.77       154
```

Figure 10. Classification Report Before Clustering

3.3. Elbow Method

Elbow method is one method that is often used[24]. Elbow method is one method that is often usedthe elbow method is effective in getting the desired number of clusters, and the k-means technique is also effective in clustering the dataset according to width, height, and width[25]**.** An empty SSE (Sum of Squared Errors) list was created to store the inertia values of each trained K-means model. The range of k values (number of clusters) tested was set from 1 to 10.

Furthermore, for each k value within the range, a K-means model was created with the corresponding number of clusters. The K-means algorithm runs 10 different initializations and selects the best result based on inertia. The random state is set to 42 to ensure consistent results. The K-means models are trained on the scaled data except for the last column, the Outcome column.

After that, the inertia value of each trained model is stored in the SSE list and an Elbow diagram is created by plotting the SSE value against the k number of clusters. This plot is used to identify the "Elbow" point, which is the point at which the decrease in SSE value starts to slow down, indicating the optimal number of clusters.

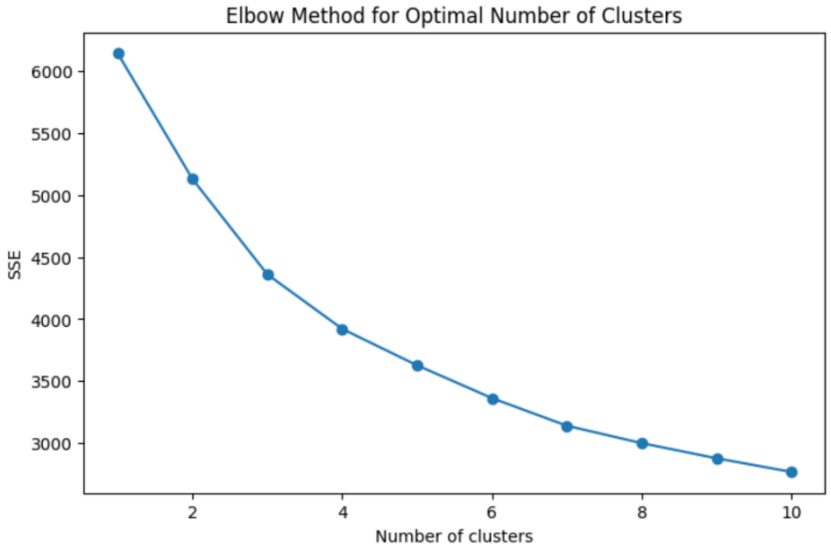

Figure 11. Elbow diagram

3.4. Optimization with K-Means Clustering

The objective of this research is to provide optimized clusters for high accuracy using K-means.

3.4.1. Clustering K-Means

In order to optimize the use of Naive Bayes classification and produce high accuracy, it is necessary to change the dataset from numeric data type to categorical data type. To achieve this, we chose to use the K-means clustering method.

The K-means algorithm is applied to group the data into three clusters. The K-means model is trained on the scaled data, and the resulting cluster labels are added as new columns in the dataset. The number of samples in each cluster was then calculated and printed to give an idea of the distribution of data within the formed clusters. The clustering results are visualized using a scatter plot, where each data point is colored according to the cluster determined by K-means, using the colormap 'viridis'.

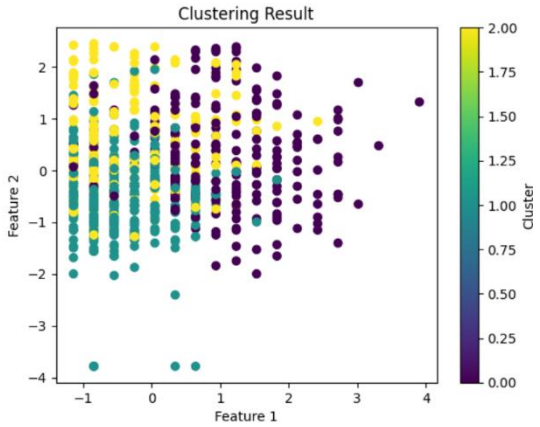

Figure 12. Scanner Plot of Clustering Results

Eventually, the clustered dataset was saved into a new CSV file for further use. The output shows the number of samples in each cluster: cluster 1 has 337 samples, cluster 2 has 216 samples, and cluster 0 has 215 samples. Visualization of the clustering results helps in understanding how the data is organized in different clusters.

```
Number of samples in each cluster:
Cluster
1    337
2    216
0    215
Name: count, dtype: int64
```

Figure 13. Scanner Plot of Clustering Results

In this case, the data that is clustered is each row that represents a patient. Each row in the dataset contains information about a particular patient, including the values of the diabetes symptoms observed in that patient. When we use a clustering algorithm such as K-Means, we group patients based on the similarity of features or symptoms observed in them. Each row of data (each patient) is considered as a point in the feature space, where the dimensions of the space are given by the observed attributes or features (e.g., blood glucose level, blood pressure, body mass index, etc.). Each cluster will contain patients who have similar symptom profiles, which can help in further understanding of the characteristics of different patient populations.

This process results in the addition of one additional column in the dataset, which contains information regarding the cluster category assigned to each row of patient data. Each row in this dataset represents one patient with a number of features that characterize their condition. After the application of the clustering algorithm, the additional column added indicates the cluster or group to which the

patient is assigned. In other words, each patient is classified into one of several clusters based on the similarity of their features.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	Cluster
0	0.639947	0.848324	0.149641	0.907270	-0.692891	0.204013	0.468492	1.425995	1	0
1	-0.844885	-1.123396	-0.160546	0.530902	-0.692891	-0.684422	-0.365061	-0.190672	0	1
2	1.233880	1.943724	-0.263941	-1.288212	-0.692891	-1.103255	0.604397	-0.105584	1	0
3	-0.844885	-0.998208	-0.160546	0.154533	0.123302	-0.494043	-0.920763	-1.041549	0	1
4	-1.141852	0.504055	-1.504687	0.907270	0.765836	1.409746	5.484909	-0.020496	1	2

Figure 14. Dataset after Clustering

3.4.2. Implementation Naive Bayes after Clustering K-Means

After the k-means clustering process, the dataset is again tested using Naive Bayes with a similar approach as before clustering. However, in splitting the data into train data and test data, two columns are dropped, namely the 'Outcome' and 'Cluster' columns. As before, the feature columns are separated into X variables, while the separated columns, which is the target or label column, is stored in the y variable. Data division is done with a ratio of 80% training data and 20% testing data, which will then be further processed in the evaluation of model performance.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.98      0.96        48
           1       0.98      0.94      0.96        65
           2       0.90      0.93      0.92        41

    accuracy                           0.95       154
   macro avg       0.94      0.95      0.95       154
weighted avg       0.95      0.95      0.95       154
```

Figure 15. Classification Report After Clustering

The results of the Naive Bayes classification report after the optimization process using clustering showed solid performance. With accuracy reaching 95%, the model successfully predicted most of the classes correctly. Precision of classes 0, 1, and 2 were 0.94, 0.98, and 0.90 respectively, indicating that the model rarely gave wrong predictions for each class. Recall of classes 0, 1, and 2 are 0.98, 0.94, and 0.93, indicating the model's ability to find most samples that actually belong to each class. The F1-score of classes 0, 1, and 2 were 0.96, 0.96, and 0.92, indicating a balance between precision and recall. With a support of 48, 65, and 41 respectively, it reflects the contribution of each class to the overall evaluation of the model. Overall, the Naive Bayes model after optimization with clustering is able to provide accurate and consistent predictions for each class in the dataset.

3.5. Evaluation

As shown in Figure 10, the results of precision, recall, f1-score, and accuracy of Naive Bayes before clustering have smaller values compared to naive bayes after clustering as shown in Figure 16. This shows that Naive Bayes works well for categorical data.

4. DISCUSSIONS

The discussion on the comparison of Naive Bayes accuracy before and after the K-Means clustering process highlights the positive impact of pre-processing techniques on the performance of classification models. The comparison of optimization results before and after using K-Means clustering is shown in Table 3.

Table 3. Results Comparison

Metric	Naïve Bayes	Naïve Bayes + Clustering K-Means
Precision 0	0.83	0.94
Precision 1	0.66	0.98
Precision 2	-	0.90
Recall 0	0.80	0.98
Recall 1	0.71	0.94
Recall 2	-	0.93
F1-score 0	0.81	0.96
F1-score 1	0.68	0.96
F1-score 2	-	0.92
Accuracy	0.77	0.95

Prior to K-Means clustering, the Naive Bayes model tended to encounter challenges in handling complex numerical data, which could result in less satisfactory classification performance. For instance, before clustering, the Naive Bayes model had an accuracy of 77%. However, following the clustering process, the dataset originally comprising numerical features was converted into a simpler categorical form, enabling the model to more effectively capture patterns within the data. As an illustration, post-clustering, the Naive Bayes model's accuracy increased to 95%, indicating a significant improvement.

Evaluation results demonstrate a substantial enhancement in model accuracy after the K-Means clustering process. With reduced data complexity and improved model capability in understanding underlying patterns, Naive Bayes can provide more accurate predictions post-clustering. This discussion underscores the importance of preprocessing techniques such as clustering in enhancing classification model performance, emphasizing the potential of K-Means clustering as an effective step in preparing data for Naive Bayes applications.

5. CONCLUSION

The application of clustering, specifically K-Means, to transform numerical data into categorical forms has shown promising results in enhancing the performance of Naive Bayes classifiers. This project aimed to explore the potential benefits of utilizing clustering techniques as a preprocessing step for Naive Bayes applications. By categorizing numerical features into distinct clusters, the complexity of the dataset was effectively reduced, enabling Naive Bayes classifiers to capture underlying patterns more accurately.

Through experimentation and evaluation, it was observed that the Naive Bayes model trained on clustered data exhibited improved performance compared to the model trained on raw numerical data. The classification reports revealed higher precision, recall, and F1-scores for each class, as well as an overall increase in accuracy. This enhancement in performance underscores the effectiveness of clustering as a feature engineering technique for Naive Bayes applications.

Furthermore, the comparison between the Naive Bayes models before and after clustering optimization provided valuable insights into the impact of preprocessing on classification outcomes. The significant improvements observed in various performance metrics highlight the importance of data transformation techniques, such as clustering, in enhancing the predictive capabilities of machine learning models.

In conclusion, the utilization of clustering, particularly K-Means, for transforming numerical data into categorical forms has proven to be beneficial for Naive Bayes applications. This approach not only simplifies the dataset but also enhances the discriminative power of Naive Bayes classifiers, ultimately leading to more accurate predictions.

REFERENCES

[1] R. Rastogi and M. Bansal, "Diabetes prediction model using data mining techniques," *Meas. Sensors*, vol. 25, no. December 2022, p. 100605, 2023, doi: 10.1016/j.measen.2022.100605.

[2] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthc. Technol. Lett.*, vol. 10, no. 1–2, pp. 1–10, 2023, doi: 10.1049/htl2.12039.

[3] C. Krishna Suryadevara, "Issue 4 DIABETES RISK ASSESSMENT USING MACHINE LEARNING: A COMPARATIVE STUDY OF CLASSIFICATION ALGORITHMS," *Int. Eng. J. Res. Dev.*, vol. 8, no. 4, pp. 1–10, 2023, [Online]. Available: www.iejrd.com

[4] A. Mansoori *et al.*, "Prediction of type 2 diabetes mellitus using hematological factors based on machine learning approaches: a cohort study analysis," *Sci. Rep.*, vol. 13, no. 1, pp. 1–11, 2023, doi: 10.1038/s41598-022-27340-2.

[5] D. Iverson, "No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title," vol. 4, no. 02, pp. 7823–7830, 2024.

[6] M. S. Ali, M. K. Islam, A. A. Das, D. U. S. Duranta, M. F. Haque, and M. H. Rahman, "A Novel Approach for Best Parameters Selection and Feature Engineering to Analyze and Detect Diabetes: Machine Learning Insights," *Biomed Res. Int.*, vol. 2023, 2023, doi: 10.1155/2023/8583210.

[7] K. Abnoosian, R. Farnoosh, and M. H. Behzadi, "Prediction of diabetes disease using an ensemble of machine learning multi-classifier models," *BMC Bioinformatics*, vol. 24, no. 1, pp. 1–24, 2023, doi: 10.1186/s12859-023-05465-z.

[8] S. S. Kavitha and N. Kaulgud, "Quantum K-means clustering method for detecting heart disease using quantum circuit approach," *Soft Comput.*, vol. 27, no. 18, pp. 13255–13268, 2023, doi: 10.1007/s00500-022-07200-x.

[9] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *EURASIP J. Wirel. Commun. Netw.*, 2021, doi: 10.1186/s13638-021-01910-w.

[10] M. Annas and S. N. Wahab, "Data Mining Methods: K-Means Clustering Algorithms," *Int. J. Cyber IT Serv. Manag.*, vol. 3, no. 1, pp. 40–47, 2023, [Online]. Available: https://iiast.iaic-publisher.org/ijcitsm/index.php/IJCITSM/article/view/122

[11] A. Salma and W. Silfianti, "Sentiment Analysis of User Review on COVID-19 Information Applications Using Naïve Bayes Classifier, Support Vector Machine, and K-Nearest Neighbors," *Int. Res. J. Adv. Eng. Sci.*, vol. 6, no. 4, pp. 158–162, 2021.

[12] K. Lemons, "A Comparison Between Naïve Bayes and Random Forest to Predict Breast Cancer," *Int. J. Undergrad. Res. Creat. Act.*, vol. 12, no. 1, p. 1, 2020, doi: 10.7710/2168-0620.0287.

[13] P. Subarkah, W. R. Damayanti, and R. A. Permana, "Comparison of Correlated Algorithm Accuracy Naive Bayes Classifier and Naive Bayes Classifier for Classification of heart failure," *Ilk. J. Ilm.*, vol. 14, no. 2, pp. 120–125, 2022, doi: 10.33096/ilkom.v14i2.1148.120-125.

[14] A. V. D. Sano, A. A. Stefanus, E. D. Madyatmadja, H. Nindito, A. Purnomo, and C. P. M. Sianipar, "Proposing a visualized comparative review analysis model on tourism domain using Naïve Bayes classifier," *Procedia Comput. Sci.*, vol. 227, pp. 482–489, 2023, doi: 10.1016/j.procs.2023.10.549.

[15] V. A. Permadi, S. P. Tahalea, and R. P. Agusdin, "K-Means and Elbow Method for Cluster Analysis of Elementary School Data," *Prog. Pendidik.*, vol. 4, no. 1, pp. 50–57, 2023, doi:

10.29303/prospek.v4i1.328.

[16] K. E. Setiawan, A. Kurniawan, A. Chowanda, and D. Suhartono, "Clustering models for hospitals in Jakarta using fuzzy c-means and k-means," *Procedia Comput. Sci.*, vol. 216, no. 2022, pp. 356–363, 2022, doi: 10.1016/j.procs.2022.12.146.

[17] X. Li, J. Zhang, and F. Safara, "Improving the Accuracy of Diabetes Diagnosis Applications through a Hybrid Feature Selection Algorithm," *Neural Process. Lett.*, vol. 55, no. 1, pp. 153–169, 2023, doi: 10.1007/s11063-021-10491-0.

[18] H. Zhao, "Design and Implementation of an Improved K-Means Clustering Algorithm," *Mob. Inf. Syst.*, vol. 2022, 2022, doi: 10.1155/2022/6041484.

[19] A. Alam, D. A. F. Alana, and C. Juliane, "Comparison Of The C.45 And Naive Bayes Algorithms To Predict Diabetes," *Sinkron*, vol. 8, no. 4, pp. 2641–2650, 2023, doi: 10.33395/sinkron.v8i4.12998.

[20] W. A. Arifin, I. Ariawan, A. A. Rosalia, L. Lukman, and N. Tufailah, "Data scaling performance on various machine learning algorithms to identify abalone sex," *J. Teknol. dan Sist. Komput.*, vol. 10, no. 1, pp. 26–31, 2022, doi: 10.14710/jtsiskom.2021.14105.

[21] G. KARATAŞ BAYDOĞMUŞ, "The Effects of Normalization and Standardization an Internet of Things Attack Detection," *Eur. J. Sci. Technol.*, no. 29, pp. 187–192, 2021, doi: 10.31590/ejosat.1017427.

[22] N. Rismayanti, A. Naswin, U. Zaky, M. Zakariyah, and D. A. Purnamasari, "Evaluating Thresholding-Based Segmentation and Humoment Feature Extraction in Acute Lymphoblastic Leukemia Classification using Gaussian Naive Bayes," *Int. J. Artif. Intell. Med. Issues*, vol. 1, no. 2, pp. 74–83, 2023, doi: 10.56705/ijaimi.v1i2.99.

[23] I. O. Muraina, "Ideal Dataset Splitting Ratios in Machine Learning Algorithms: General Concerns for Data Scientists and Data Analysts," *7th Int. Mardin Artuklu Sci. Res. Conf.*, no. February, pp. 496–504, 2022.

[24] I. F. Ashari, E. Dwi Nugroho, R. Baraku, I. Novri Yanda, and R. Liwardana, "Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta," *J. Appl. Informatics Comput.*, vol. 7, no. 1, pp. 89–97, 2023, doi: 10.30871/jaic.v7i1.4947.

[25] A. Maravillas, "International Journal of Advance Research in Integration of K-means Algorithm and Elbow Method in Clustering the Bivalve Species," no. January, 2023.

SUBMISSION CHECK-LIST

All items must be YES before submitting this task in order to make sure a manuscript well-prepared within a high standard. Please put X in column Yes after checking each item

General			
No	**Item**	**Yes**	**No**
1	I have prepared the manuscript using "DataMining-Template.docx"		
2	The manuscript is written in single or double column using Palatino Linotype10 pt font		
3	Title is matching with the focus of the manuscript		
4	I have complete authors information:		

	a. Affiliation of all authors		
	b. Indicate one of the authors is the corresponding author		
	c. Email of corresponding author		
5	The length of the manuscript is not less than 2500 words (excluding abstract, author information and reference list)		
6	The manuscript contains the core contents: introduction with literature review, method, results, discussion, conclusion, references		
7	Contains maximum 2 level of sections		
Abstract			
No	**Item**	**Yes**	**No**
1	The abstract is written between 150 – 250 words		
2	contains the purpose		
3	contains the method / research approach		
4	contains the important findings / conclusion		
5	contains novelty/originality		
6	The abstract is completed with keywords		
Figures (if applicable)			
No	**Item**	**Yes**	**No**
1	All figures are clear and readable (image, text and legend) with high resolution		
2	All figures are in formal style, without redundant title		
3	All information in the figure are in English and all decimal written in international standard, using point (.) not comma (,)		
4	All figures have captions (at the bottom) with consecutive numbers		
5	All figures are cited in the text using consistent citation style		
6	Never citing figure using below and above, but using the number of figure		
Tables (if applicable)			
No	**Item**	**Yes**	**No**
1	All texts in tables are clear and readable		
2	All tables are in formal style		
3	All information in the table are in English and all decimal written in international standard, using point (.) not comma (,)		
4	All tables are captioned on top with consecutive numbers		
5	All tables are cited in the text using consistent citation		
6	Never citing table using below and above, but using the number of table		
Equations (if applicable)			
No	**Item**	**Yes**	**No**
1	All equations are written using editor tool (editable), not a cropped		
2	All equations are captioned on top with consecutive numbers		
3	All equations are cited in the text using consistent citation		
4	Never citing equation using below and above, but using the number of equation		
References			
No	**Item**	**Yes**	**No**
1	Reference list and citation consistently follows the IEEE style		

2	Journal names in the reference list are not in abbreviated format, but it should be in full name format		
3	All references are cited in the text		
4	Citation in the text follows general consistent citation rules		
5	Paper cites at least than 25 references		
6	Book sources are not more than 20% of the reference list		
7	Self-citation in the list is not more than 2		
8	Write all authors in the reference list unless authors more than 7		