Jurnal Teknologi Informasi dan Komunikasi

Vol: 15 No 02 2024 **E-ISSN: 2477-3255**

Accepted by the Editor: 08-10-2024 | Revision: 08-11-2024 | Published: 30-11-2024

Comparative Analysis of KNN and Neavy Bayes Algorithms in Socio-Economic Data Classification in Indonesia

Kiki Buhori[1], Andrianingsih[2]

[1,2]Program Studi Magister Teknologi Informasi Fakultas Teknologi Komunikasi dan Informatika Universitas Nasional

[1,2]Jl. Sawo Manila No.61, RT.14/RW.7, Pejaten Bar., Ps. Minggu, Kota Jakarta Selatan, Daerah Khusus Ibukota Jakarta 12520

e-mail: [1]kikibuhori2022@student.unas.ac.id, [2]andrianingsih@civitas.unas.ac.id

Abstract

The global economy continues to recover as trade flows, employment, and incomes improve. However, the economic recovery is uneven across countries and business sectors. The economic recovery has also resulted in structural changes, meaning that some sectors, jobs, technologies and behaviors will not return to pre-pandemic trends. Future developments depend on local economic conditions. The economy has the most important aspect in a country where the economy makes a country capable of meeting its needs by utilizing limited resources. This study aims to compare two data mining classification algorithms, namely Naïve Bayes and K-Nearest Neighbor, in analyzing socio-economic data in Indonesia. Based on this problem, the data mining classification method is used in determining the algorithm that is suitable for predicting socio-economic data in Indonesia. The two algorithms used are K-NN and Naive Bayes. After testing the two algorithms using confusion matrix and K-Fold Cross Validation, the results obtained from the two models have an accuracy of Naïve Bayes 98.25% and K-NN 97.78% and the results of K-Fold Cross Validation Naïve Bayes 98% and K-NN 96%. Naïve Bayes is superior to K-NN in this context of socioeconomic data classification in Indonesia, especially in terms of accuracy. Although K-NN shows good consistency, Naïve Bayes provides more accurate results.

Keywords: *Socio-economics, Data Mining, Classification, K-NN, Naïve Bayes, Python*

1. Introduction

The global economy continues to recover as trade flows, employment, and incomes improve. However, the economic recovery is uneven across countries and business sectors. The economic recovery has also resulted in structural changes, meaning that some sectors, jobs, technologies and behaviors will not return to pre-pandemic trends. Future developments depend on local economic conditions. The economy has the most important aspect in a country where the economy makes a country capable of meeting its needs by utilizing limited resources. from limited resources, economic problems arise due to unlimited human needs, however, Indonesia also faces various challenges in achieving socio-economic welfare for all its people [1]. On the other hand, the increase in international commodity prices is still supporting Indonesia's export

https://doi.org/10.31849/digitalzone.v15i2. 23337

performance which again recorded the highest value. Even so, the increase in imports this time was higher than exports, resulting in a trade balance surplus [2]. Indonesia's domestic economic development is considered quite stable in strengthening and industrial expansion continues. To understand and analyze the socio-economy in Indonesia, accurate, relevant, and up-to-date data and information that can be accessed by various parties, including the government, private sector, academics, and the public are also needed. This socioeconomic data is a valuable asset that can be processed and analyzed for various purposes, including understanding the socioeconomic conditions of the community. In the literature study, there are several theories that researchers use for the purposes of analyzing problems and also solving them. Data mining is the process of extracting useful patterns or knowledge from large and complex data. Data mining can be used for various purposes, such as classification, clustering, association, prediction, estimation, description, and visualization. Data mining can be applied to various fields, such as business, education, health and security [3]. One of the most commonly used data mining techniques is classification. Classification is the process of categorizing data into certain classes based on their attributes [4]. Classification can be used to identify characteristics, recognize patterns, make decisions, and predict outcomes. There are many classification algorithms that have been developed, such as Decision Tree, Naïve Bayes, K-Nearest Neighbor, Neural Network, Support Vector Machine, and others [5]. Naïve Bayes and K-Nearest Neighbor (KNN) algorithms were chosen because they offer different approaches to the classification process, which are relevant for various dataset scenarios. Naïve Bayes was chosen for its simplicity, speed and efficiency, especially on large datasets with the assumption of independence between features. This makes Naïve Bayes often used in classification problems, while KNN was chosen for its ability to capture non-linear patterns through a distance-based approach [6]. This algorithm can provide good classification results on datasets with complex data distributions. The selection of these two algorithms is also based on the consideration that they have different working methods Naïve Bayes is based on probability and statistical assumptions, while KNN is based on instance learning so that it provides room for a more in-depth comparison regarding the advantages and limitations of each in handling the dataset used.

The methodology used by the Cross-Industry Standard Process for Data Mining (CRIDP-DM), this method includes stages ranging from Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment [7]. In this context, data mining can be used to classify socio-economic data in Indonesia based on certain categories. The Naïve Bayes algorithm is a classification method based on Bayes theory and probability methods. This algorithm is used to classify data into predetermined categories or groups based on the characteristics or attributes possessed by the data concerned [8]. K-Nearest Neighbor algorithm, on the other hand, is used to classify data based on predefined labels or categories. This algorithm belongs to the supervised learning category, which is learning that uses labeled data as input [9].

The selection of Naïve Bayes and K-Nearest Neighbor (KNN) algorithms in this research is in line with the stages of the Cross-Industry Standard Process for Data Mining (CRISP-DM), which includes the steps of Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. In this context, data mining is used to classify socioeconomic data in Indonesia based on certain categories, with the selected algorithm having specific advantages in the classification task [10].

The Naïve Bayes algorithm, which is based on Bayes theory and probability methods, is able to classify data into predefined categories based on the characteristics or attributes possessed by the data. Its simple and efficient nature makes it suitable for high-dimensional data, such as text or categorical data. Meanwhile, the KNN algorithm, which belongs to the supervised learning category, is used to classify data based on pre-existing labels or categories. With its distance-based approach, KNN provides flexibility in handling non-linear data distribution patterns, making it an ideal choice for datasets that require sensitivity to local patterns [11]. By combining these two algorithms, the classification process of socio-economic

data can be done effectively, utilizing the advantages of each method to provide accurate and relevant results.

This research aims to evaluate the effectiveness of K-Nearest Neighbors (KNN) and Naive Bayes algorithms in classifying socioeconomic data in Indonesia. Specifically, this research will measure the performance of both algorithms, compare their prediction accuracy, and analyze how different training and test data split ratios (80/20, 70/30, 60/40, and 50/50). This research is important because socio economic data classification in Indonesia still faces various challenges, such as data heterogeneity, information gaps, and lack of application of effective data mining methods. To date, KNN and Naïve Bayes-based approaches have not been widely applied or developed in the context of socio economic analysis in Indonesia. These approaches offer the potential to improve accuracy and efficiency in processing complex and diverse data, thereby generating deeper insights.

2. Research Methods

This research was conducted with the methodology used in the research to classify socio-economic data using the K-Nearest Neighbors (KNN) and Naive Bayes algorithms. The methodology used is Cross-Industry Standard Process for Data Mining (CRIDP-DM), this method includes stages ranging from Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment [12]. The steps taken in this research are shown below:

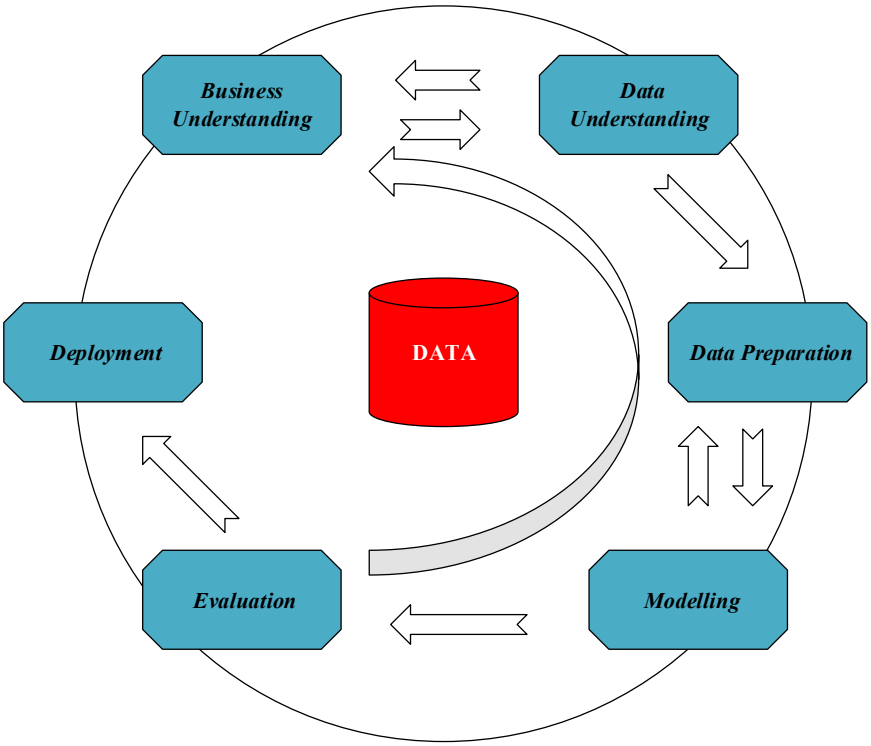


Figure 1. CRIDP-DM Method

CRISP-DM (Cross Industry Standard Process for Data Mining) is a standardization of data mining processing that has been developed where existing data will go through each structured and defined phase clearly and efficiently. In addition to applying a model in the data mining process, algorithm selection greatly affects the performance comparison of data mining methods. [13]. CRISP-DM is used to analyze strategies that are useful in solving research problems or business and corporate problems. The CRISP-DM (Cross-Industry Standard Process Model for Data Mining) method explains the data mining process which consists of six

https://doi.org/10.31849/digitalzone.v15i2. 23337

stages. The stages in the CRISP-DM method include business understanding, data understanding, data preparation, modeling, evaluation, and deployment [14].

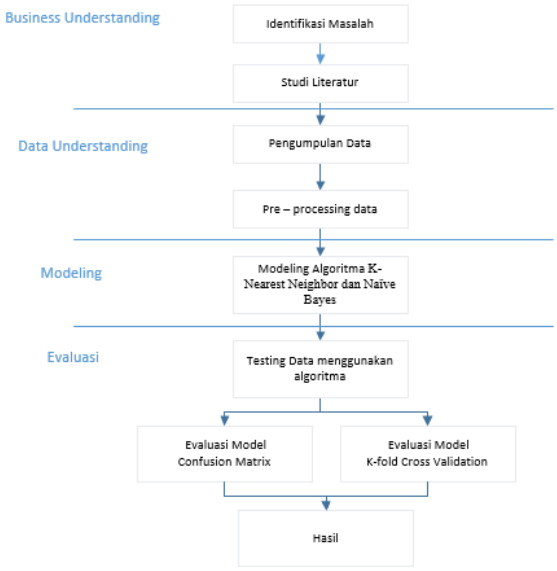


Figure 2. Flow of Research

2.1. Business Undertanding

In this stage, this objective is to identify the problem and literature study in comparing the performance of KNN and Naive Bayes algorithms in classifying socio-economic data in Indonesia. This objective will be broken down into several specific data analysis objectives, such as measuring the performance of KNN and Naive Bayes algorithms in the classification of socio-economic data, comparing the classification results of both algorithms and using linear regression to predict relevant variables and evaluating the results.

2.2. Data Understanding

This stage of the study includes various indicators that describe the socio-economic conditions of regions in Indonesia. These indicators help in analyzing and understanding the differences and similarities between the regions based on various socio-economic aspects consisting of:

1. Percentage of poor people, showing the proportion of people living below the poverty line in a region.
2. Human development index (HDI), measures human development based on three basic dimensions: health (long and healthy life), education (knowledge), and decent standard of living.
3. Labor force participation rate, shows the percentage of the population that is actively in the labor force either working or looking for work. This is how to start another subsection.

2.3. Data Preparation

This study uses data obtained for this study as many as 514 records of socio-economic data that have a percentage of the increase in human resources to be tested.

1. Data Normalization
   Normalization is done to standardize the features in the dataset to have a uniform scale. The method used is StandardScaler from Scikit-Learn, which ensures each feature has a mean of 0 and a standard deviation of 1. This is important for the classification algorithm to work efficiently and accurately.

2. Labeling Process
   In this stage, the labeling process uses the K-means Clustering method, where the *k* value is set to 5. The normalized socio-economic data is used as input for the K-means algorithm. The columns used for clustering include; Percentage of Poor Population, Human Development Index and Labor Force Participation Rate.

2.4. Modeling

At this stage, the training data is processed so that it will produce several rules and will form an accurate decision. There are two models that will be used, namely the Naïve Bayes classification algorithm and the K-Nearest Neighbor classification algorithm [15].

2.5. Evaluasi

In the evaluation stage, it is called the classification stage because at this stage the test for accuracy will be determined. The testing stage is to see the accuracy results in the classification process on the two algorithms and evaluate with Confusion Matrix and K-fold Cross Validation [16].

3. Results and Discussion

In the initial data processing stage, the experiments used in this study used the Cross-Standard Industry for Data Mining (CRISP-DM) model.

	No	Provinsi	Kab/Kota	Persentase Penduduk Miskin (P0)	Rata-rata Lama Sekolah Penduduk	Pengeluaran per Kapita Disesuaikan	Indeks Pembangunan Manusia	Umur Harapan Hidup (Tahun)	Persentase rumah tangga yang memiliki akses terhadap sanitasi layak	Persentase rumah tangga yang memiliki akses terhadap air minum layak	Tingkat Pengangguran Terbuka	Tingkat Partisipasi Angkatan Kerja	PDRB atas Dasar Harga Konstan menurut Pengeluaran (Rupiah)
0	1	ACEH	Simeulue	18.98	9.48	7148	66.41	65.28	71.56	87.45	5.71	71.15	1648096
1	2	ACEH	Aceh Singkil	20.36	8.68	8776	69.22	67.43	69.56	78.58	8.36	62.85	1780419
2	3	ACEH	Aceh Selatan	13.18	8.88	8180	67.44	64.40	62.55	79.65	6.46	60.85	4345784
3	4	ACEH	Aceh Tenggara	13.41	9.67	8030	69.44	68.22	62.71	86.71	6.43	69.62	3487157
4	5	ACEH	Aceh Timur	14.45	8.21	8577	67.83	68.74	66.75	83.16	7.13	59.48	8433526
...	...	...	...	...	...	...	...	...	...	...	...	...	...
509	510	PAPUA	Puncak	36.26	2.16	5412	43.17	65.86	11.43	85.03	0.94	89.43	831070
510	511	PAPUA	Dogiyai	28.81	4.94	5415	55.00	65.85	12.11	71.24	5.68	78.20	906904
511	512	PAPUA	Intan Jaya	41.66	3.09	5328	48.34	65.69	0.36	35.01	1.43	75.75	767101
512	513	PAPUA	Deiyai	40.59	3.25	4673	49.96	65.36	0.00	85.23	0.79	85.01	841296
513	514	PAPUA	Kota Jayapura	11.39	11.57	14937	80.11	70.52	85.31	97.10	11.67	63.75	22852202

514 rows × 13 columns

Figure 3. Socio-economic Data

Based on Figure 3. is a socio-economic dataset used in this analysis covering 514 regencies and cities from various provinces in Indonesia. This data provides insight into the socio-economic conditions in each region through several key variables. dataset display used as a comparison of testing using the KNN and Naive Bayes algorithms where testing is divided into 2 (two), namely testing training data and test data.

3.1. Business Understanding

Socio-economic conditions in Indonesia are diverse and influenced by various factors such as the percentage of poor people, human development index and labor force participation rate. Clustering regions based on socioeconomic indicators can help in understanding and addressing inequality as well as designing more effective and focused policies. In this research, analyzing socioeconomic data using machine learning algorithms such as K-Nearest Neighbor (KNN) and Naïve Bayes can provide a deeper and more accurate insight into the socioeconomic conditions in Indonesia.

3.2. Data Understandig

The main focus in understanding the socioeconomic data used in the study. This data includes a wide range of indicators that reflect the social and economic conditions of regions in Indonesia. A good understanding of the structure and characteristics of the data is essential to ensure accurate and relevant analysis.

Data Preview:

	No	Provinsi	Kab/Kota	Persentase Penduduk Miskin	Pengeluaran per Kapita Disesuaikan	Indeks Pembangunan Manusia	Umur Harapan Hidup Tahun	Persentase rumah tangga yang memiliki aks
0	1	ACEH	Simeulue	18,98	7,148	66,41	65,28	71,56
1	2	ACEH	Aceh Singkil	20,36	8,776	69,22	67,43	69,56
2	3	ACEH	Aceh Selatan	13,18	8,180	67,44	64,40	62,55
3	4	ACEH	Aceh Tenggara	13,41	8,030	69,44	68,22	62,71
4	5	ACEH	Aceh Timur	14,45	8,577	67,83	68,74	66,75
5	6	ACEH	Aceh Tengah	15,26	10,780	73,37	68,86	90,58
6	7	ACEH	Aceh Barat	18,81	9,593	71,67	67,99	89,60
7	8	ACEH	Aceh Besar	14,05	9,644	73,58	69,79	87,40
8	9	ACEH	Pidie	19,59	9,860	70,70	66,95	54,10
9	10	ACEH	Bireuen	13,25	8,867	72,33	71,26	81,89

Figure 4. Research Dataset Variables

In the figure this dataset is a dataset used to conduct research that provides valuable insights. This data can be used by policymakers to improve the quality of life of the population, alleviating poverty. With a good understanding of each indicator, local governments can make more informed decisions for the well-being of the community.

Normalisasi Data

Normalized Data:

	Provinsi	Kab/Kota	Persentase Penduduk Miskin	Indeks Pembangunan Manusia	Tingkat Partisipasi Angkatan Kerja
0	ACEH	Simeulue	0.4226	0.6178	0.3553
1	ACEH	Aceh Singkil	0.4577	0.6695	0.1555
2	ACEH	Aceh Selatan	0.2749	0.6367	0.1074
3	ACEH	Aceh Tenggara	0.2808	0.6735	0.3185
4	ACEH	Aceh Timur	0.3073	0.6439	0.0744
5	ACEH	Aceh Tengah	0.3279	0.7459	0.4793
6	ACEH	Aceh Barat	0.4183	0.7146	0.0881
7	ACEH	Aceh Besar	0.2971	0.7497	0.1271
8	ACEH	Pidie	0.4381	0.6967	0.0939
9	ACEH	Bireuen	0.2767	0.7267	0.2292

Figure 5. Data Normalization Results

Normalizing the data with MinMaxScaler helps ensure that all features are on the same scale, which is critical to the performance and accuracy of machine learning models. This normalized data is ready to be used in the training and testing process of the model, ensuring that the model can learn and make predictions more effectively and efficiently.

	Provinsi	Kab/Kota	Persentase Penduduk Miskin	Indeks Pembangunan Manusia	Tingkat Partisipasi Angkatan Kerja	Cluster	Cluster_Label
0	ACEH	Simeulue	0.4226	0.6178	0.3553	4	Mid
1	ACEH	Aceh Singkil	0.4577	0.6695	0.1555	4	Mid
2	ACEH	Aceh Selatan	0.2749	0.6367	0.1074	4	Mid
3	ACEH	Aceh Tenggara	0.2808	0.6735	0.3185	4	Mid
4	ACEH	Aceh Timur	0.3073	0.6439	0.0744	4	Mid
5	ACEH	Aceh Tengah	0.3279	0.7459	0.4793	1	Mid
6	ACEH	Aceh Barat	0.4183	0.7146	0.0881	4	Mid
7	ACEH	Aceh Besar	0.2971	0.7497	0.1271	4	Mid
8	ACEH	Pidie	0.4381	0.6967	0.0939	4	Mid
9	ACEH	Bireuen	0.2767	0.7267	0.2292	4	Mid

Figure 6. Labeling Process Result

In the picture above, is the clustering process in determining the group label into 5 (five) and what data will be used as validation test material. From 5 groups, labels are made into 3 (three) consisting of mid, low and high. This also determines the variables that will be taken, namely the percentage of poor people, the human development index and the labor force participation rate, which are 10 variables.

3.3. Modeling dan Evaluation

```
# Define models
ModelNB = GaussianNB()
ModelKNN = KNeighborsClassifier()
```

Figure 7. Modeling Naïve Bayes dan KNN

From the figure above, we can see how the Naive Bayes and KNN models work with various train-test data splits. Each model is tested with different proportions of training data to measure the model's performance under different data conditions.

3.4. Modeling dan Evaluation

In this test using socio-economic data where the results of the confusion matrix evaluation of each algorithm have accuracy results with testing data Train-Test Split 80:20, 70:30, 60:40 and 50:50:

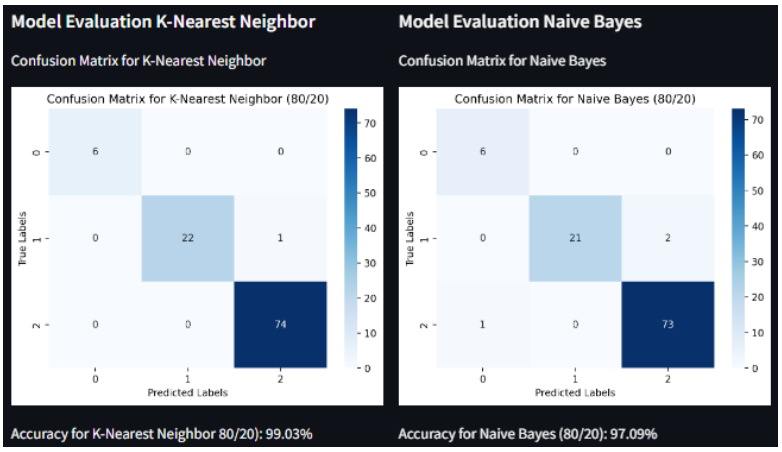
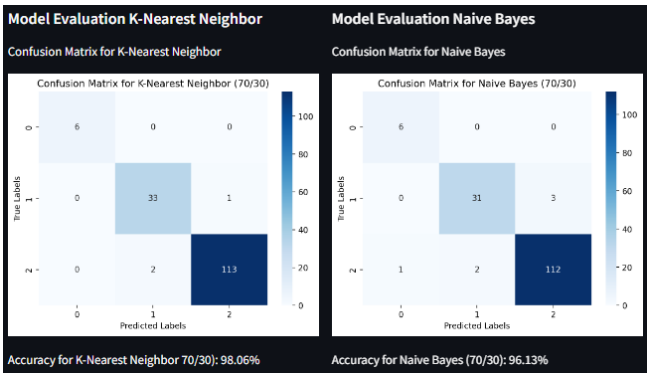


Figure 8. Result Confusion Matrix 80:20

Both models perform very well in classification, as evidenced by the high values on the main diagonal of the confusion matrix (dark blue color). This indicates that most of the predictions match the actual labels. K-Nearest Neighbor had almost all accurate predictions; there was only one misclassification, when the label was first predicted to be 1 but ended up being 2. Although Naive Bayes also performed well, there were some misclassifications. There were two cases where the clear label was 0 but predicted to be 1, and one case where the clear label was 1 but predicted to be 2. 99.03% was K-Nearest Neighbor. It has a very high accuracy, indicating that this model is very good at predicting labels accurately. The accuracy of Naive Bayes (97.09%) is also quite high, although somewhat lower than K-Nearest Neighbor.



https://doi.org/10.31849/digitalzone.v15i2. 23337

Figure 9. Confusion Matrix 70:30

This figure shows the evaluation comparison between two classification models, K-Nearest Neighbor (KNN) and Naive Bayes. Both models were evaluated using Confusion Matrix and accuracy metrics, with 70% data split for training and 30% for testing. As a result, KNN showed slightly superior performance with 98.06% accuracy, having only one misclassification. Naive Bayes also performed well with 96.13% accuracy, but there were slightly more misclassifications. Although KNN is slightly superior in this case, the selection of the best model still depends on the context of the problem and other considerations such as complexity and interpretability. Additional information such as the dataset used and other evaluation metrics will provide a more complete picture of the performance of both modes.

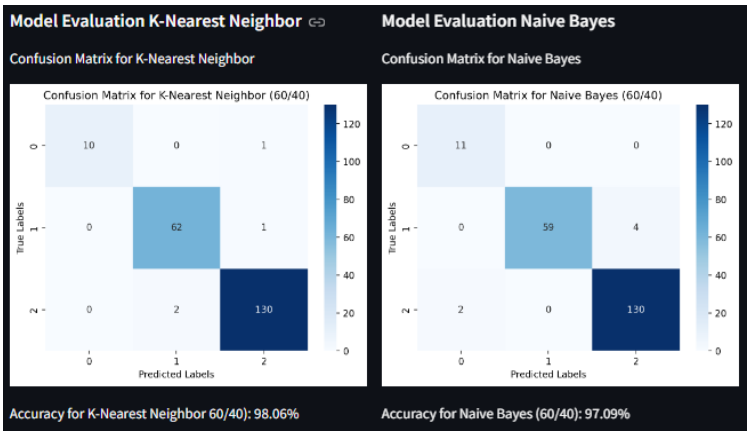


Figure 10. Confusion Matrix 60:40

This figure compares the performance of two classification models, K-Nearest Neighbor (KNN) and Naive Bayes, using Confusion Matrix and accuracy on a 60/40 data split. KNN shows slightly higher accuracy (98.06%) than Naive Bayes (97.09%), with fewer misclassifications in the Confusion Matrix. Although KNN had a slight edge, the selection of the best model depends on the context of the problem and other factors such as complexity and interpretability. Additional information on other datasets and evaluation metrics will provide a more comprehensive understanding of the performance of both models.

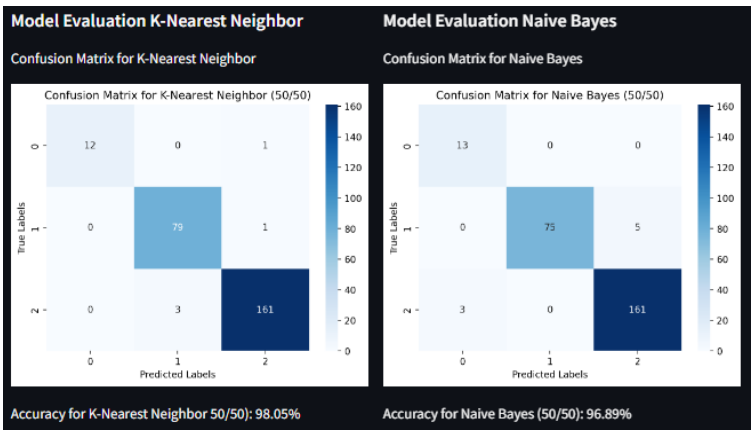


Figure 11. Confusion Matrix 50:50

The images of both models show excellent performance with accuracy above 96%. KNN is slightly ahead with an accuracy of 98.05%, compared to Naive Bayes which has an accuracy of

96.89%. The Confusion Matrix provides more details on the types of errors made by each model. KNN had only a few misclassifications, especially in predicting the "1" class. Naive Bayes also performed well, but had slightly more errors, especially in predicting the "0" and "1" classes. Although KNN is slightly superior in terms of accuracy, Naive Bayes is still a strong model.

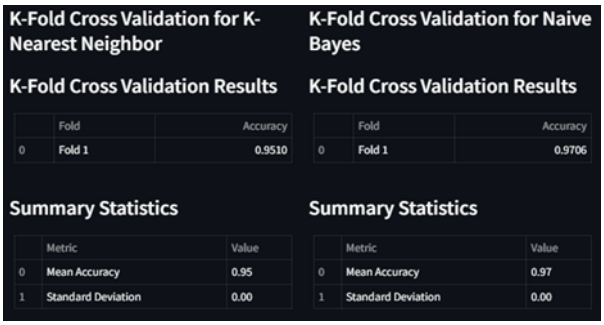


Figure 12. K-Fold Validation Evaluation Results

Based on the K-Fold Cross Validation results, it can be seen that both classification models, K-Nearest Neighbors (KNN) and Naive Bayes, performed very well. Naive Bayes is slightly superior with an average accuracy of 97% compared to KNN which reaches 95%. The stability of the performance of both models is also guaranteed, as evidenced by the standard deviation value of 0.00 which shows the consistency of accuracy in each fold. However, it should be noted that these results are only from one fold, so further analysis with more folds is required to get a more comprehensive picture. In addition, consideration of other evaluation metrics and model complexity are also important in selecting the best model for the specific problem at hand. This study uses socio-economic data from 514 districts/cities in Indonesia to analyze regional welfare using the K-Nearest Neighbor (KNN) and Naive Bayes algorithms. KNN showed high accuracy in classifying data with clear patterns, especially in the class "Mid," while Naive Bayes was more stable and efficient, although less optimal on correlated data. Comparison with the baseline data shows that the model's predictions are generally in line with the distribution of socioeconomic indicators, such as HDI and the percentage of poor people [17][18]. Confusion matrix and statistical analysis were used to measure the agreement of the predicted results with the actual data. KNN is superior in high precision, while Naive Bayes is suitable for large datasets with efficiency requirements. Both have the potential to support data-driven policies, with opportunities to improve accuracy through model optimization and data.

4. Conclusion

The results of the accuracy comparison between the Naive Bayes and K-Nearest Neighbor (KNN) algorithms with various training and test data splits (train-test split), show that both have different performance depending on the data split. In the 80/20 split, both Naive Bayes and KNN achieved the same highest accuracy of 99.03%. This shows that both algorithms are able to utilize larger training data very well. However, as the proportion of training data decreases, the performance difference between the two algorithms starts to show. At a 70/30 split, the accuracy of both algorithms remained equal at 98.71%.

However, when the training data was reduced to 60/40, Naive Bayes experienced a decrease in accuracy to 97.57%, while KNN showed better performance with an accuracy of 98.06%. This shows that KNN is still able to maintain a relatively good performance despite the reduced portion of training data. However, in the 50/50 split, KNN experienced a significant decrease in accuracy to 95.33%, while Naive Bayes again outperformed with an accuracy of 97.67%. Overall, Naive Bayes showed better consistency across various data splits, especially on a more balanced split between training and test data. On the other hand, KNN tends to perform very well on larger training data shares, but experiences a significant drop in performance when the

share of training data is reduced. This suggests that Naive Bayes may be more suitable for use in conditions where training data is limited, while KNN is more optimal when a sizable amount of training data is available. The k-fold validation evaluation results of both the K-Nearest Neighbor (KNN) and Naïve Bayes algorithms using different train-test split data have the same accuracy results in each data split, namely 98% KNN accuracy and 96% Naïve Bayes. From the results of this evaluation, it shows that the KNN algorithm is superior in evaluating using k-fold validation than naïve bayes.

Reference

[1] A. H. Wibowo dan T. I. Oesman, "The comparative analysis on the accuracy of k-NN, Naive Bayes, and Decision Tree Algorithms in predicting crimes and criminal actions in Sleman Regency," dalam *Journal of Physics: Conference Series*, Institute of Physics Publishing, Mar 2020. doi: https://doi.org/10.1088/1742-6596/1450/1/012076.

[2] Badan Pusat Statistik Provinsi DKI Jakarta, "Badan Pusat Statistik Provinsi DKI Jakarta," 2022.

[3] J. K. Maranatha, "Perbedaan Algoritma C4.5 dan Naïve Bayes dalam Kepuasan Mahasiswa terhadap Pelayanan di Perguruan Tinggi", Diakses: 6 Desember 2024. [Daring]. Tersedia pada: https://www.researchgate.net/publication/351271819

[4] E. Firasari, N. Khasanah, U. Khultsum, D. N. Kholifah, R. Komarudin, dan W. Widyastuty, "Comparation of K-Nearest Neighboor (K-NN) and Naive Bayes Algorithm for the Classification of the Poor in Recipients of Social Assistance," dalam *Journal of Physics: Conference Series*, IOP Publishing Ltd, Nov 2020. https://doi.org/10.1088/1742-6596/1641/1/012077.

[5] S. Y. Irianto. S. Imaniar Ikko Mulya Rizky*, "Perbandingan Kinerja Algoritma Naive Bayes, Support Vector Machine dan Random forest untuk Prediksi Penyakit Ginjal Kronis," 2023.

[6] E. Ozturk Kiyak, B. Ghasemkhani, dan D. Birant, "High-Level K-Nearest Neighbors (HLKNN): A Supervised Machine Learning Model for Classification Analysis," *Electronics (Switzerland)*, vol. 12, no. 18, Sep 2023, https://doi.org/10.3390/electronics12183828.

[7] V. Bayu Anwari dan Yuliazmi, "Implementasi Algoritma K-Nearest Neighbors Pada Analisis Sentimen Masyarakat Terhadap Penerapan Pemberlakuan Pembatasan Kegiatan Masyarakat." https://doi.org/10.36080/skanika.v5i1.2912.

[8] Utomo, "Perbandingan Algoritma Machine Learning Untuk Penentuan Klasifikasi Kemiskinan Multidimensi Di Provinsi Nusa Tenggara Timur," *BPS Provinsi NTT: JURNAL STATISTIKA TERAPAN*, vol. V2I01.24, no. 10.5300/JSTAR., hlm. 1–12, 2022, https://doi.org/10.5300/JSTAR.V2I01.24.

[9] S. S. R. H. A. T. M. S. Eghi Ditendra, "Comparison of Classification Algorithms for Sentiment Analysis of Islam Nusantara in Indonesia," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 2, hlm. 71–77, 2022, https://doi.org/10.57152/malcom.v5i1.

[10] C. Schröer, F. Kruse, dan J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," dalam *Procedia Computer Science*, Elsevier B.V., 2021, hlm. 526–534. doi: https://doi.org/10.1016/j.procs.2021.01.199.

[11] V. Plotnikova, M. Dumas, dan F. Milani, "Adaptations of data mining methodologies: A systematic literature review," *PeerJ Comput Sci*, vol. 6, hlm. 1–43, 2020, https://doi.org/10.7717/peerj-cs.267.

[12] N. Cholifah Sastya dan D. I. Nugraha, "Penerapan Metode CRISP-DM dalam Menganalisis Data untuk Menentukan Customer Behavior di MeatSolution," *Jurnal Pendidikan Dan Aplikasi Industri*, vol. 10, 2023, https://doi.org/10.33592/unistek.v10i2.3079.

[13] M. A. Hasanah, S. Soim, dan A. S. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," *Journal of Applied Informatics and Computing (JAIC)*, vol. 5, no. 2, hlm. 103, 2021, [Daring]. Tersedia pada: http://jurnal.polibatam.ac.id/index.php/JAIC

[14] Walim, "Analisis Perbandingan Algoritma Naive Bayes, Random Forest Dan C.45 Dalam Klasifikasi Kelayakan Masyarakat Untuk Mendapatkan Bantuan Dana Desa," 2018.

[15] D. Normawati dan S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *Jurnal Sains Komputer & Informatika (J-SAKTI*, vol. 5, no. 2, hlm. 697–711, 2021, http://dx.doi.org/10.30645/j-sakti.v5i2.369.

[16] A. Nugroho dan A. Amrullah, "Evaluasi Kinerja Algoritma K-Nn Menggunakan K-Fold Cross Validation Pada Data Debitur Ksp Galih Manunggal," *JINTEKS*, vol. 5, no. 2, 2023, https://doi.org/10.51401/jinteks.v5i2.2506.

[17] M. F. Essy Rahma Meilaniwati, "Klasifikasi penduduk miskin penerima PKH menggunakan metode naïve bayes dan KNN," 2022, [Daring]. Tersedia pada: http://journal.student.uny.ac.id/ojs/index.php/jktm

[18] M. Iksan Maulana dan U. Hayati, "Perbandingan Algoritma Naïve Bayes Dan K-Nearest Neighbors Untuk Klasifikasi Topik Berita Pada Situs Detik.Com," *Jurnal Mahasiswa Teknik Informatika*, vol. 8, no. 3, Jun 2024, https://doi.org/10.36040/jati.v8i3.9779.