IPTEK The Journal of Technology and Science, 37(1), 2026 (e/pISSN:2088-2033, 0853-4098)
DOI: 10.12962/j20882033.v37i1.9263 | Received 13 July, 2023; Revised 17 November 2025; Accepted 17 November 2025

ORIGINAL RESEARCH

Selection of Feature Data in KNN Classification Datasets

M. Iman Nur Hakim*[1,2] | Iwan Setyawan[2] | Danny Manongga[2] | Hindriyanto Dwi Purnomo[2] | Hendry[2]

[1]Department of Automotive Engineering Technology, Politeknik Keselamatan Transportasi Jalan, Tegal, 52125, Indonesia

[2]Departement of Computer Science, Universitas Kristen Satya Wacana, Salatiga, 50711, Indonesia

Correspondence

*Email: m.iman@pktj.ac.id

Present Address

Department of Automotive Engineering Technology, Politeknik Keselamatan Transportasi Jalan, Jl. Perintis Kemerdekaan No.17, Slerok, Kec. Tegal Tim., Kota Tegal, Jawa Tengah 52125, Indonesia

Abstract

Feature data in a dataset can affect the data processing, either for the better or for the worse. In addition, feature data can also affect the time of data processing. Selection of the right feature data is necessary so that the feature data can represent the entire dataset. In this study, a search for feature data will be carried out that can result in better data processing. The classification process will be carried out on an Iris dataset with the KNN algorithm. The iris dataset has 4 features data (Sepal Length, Sepal Width, Petal Length, Petal Width) and the exact feature data variation will be determined in this classification. The dataset will be broken down into 7 variations of data and tested with a comparison of the training data and test data, namely 90:10, 80:20, 70:30, 60:40, 50:50, 40:60, 30:70, 20:80 and 10:90. The KNN algorithm used has parameters with the number of n neighbors 5 and the Minkowski metric. In this study, the highest accuracy value was 96% and the lowest accuracy value was 71%. The highest accuracy value is obtained from the variation of the Petal Length and Petal Width data features while the lowest accuracy value is obtained from the variation of the Sepal Length and Sepal Width data features.

KEYWORDS:

Accuracy; Feature data; KNN.

1 | INTRODUCTION

The feature data contained in the dataset is a part of processing the data. However, a large amount of feature data in a dataset will overload data processing. Selection of feature data from a dataset can be done, where only feature data will be used that can describe the entire dataset [1]. Performing feature selection can improve performance and simplify the dataset [2]. Most of the attributes (or features) may be insignificant or may even burden the classification

process and result in poor classification accuracy [3]. Therefore, it is necessary to select important features in the dataset.

Feature selection is a process applied to a dataset to reduce the number of attributes [4]. Feature selection is an important stage in the classification process, because the selected features greatly affect the accuracy of the classification [5] which in turn can shorten data processing time [6]. Classification data processing can be easily done through machine learning [7].

Machine Learning is generally divided into three approaches, namely supervised, unsupervised and semi- supervised learning. Classification, also known as supervised or inductive learning, has been widely used and yields good results in its application [8, 9]. K-Nearest Neighbor (KNN) is the most widely used algorithm among the ten other algorithms for data mining research [10]. In the KNN algorithm, the distance between data points and the majority class among its neighbors is the basis for making classification decisions [11]. The results of the classification using the KNN algorithm are quite good and accurate in solving classification problems [12].

2 | PREVIOUS RESEARCHES

The research titled 'Feature Data Selection for Improving the Performance of Entity Similarity Searches in the Internet of Things' [1] proposes a selection mechanism for the entity main features (SMEF). The SMEF is a feature data selection method based on the quantitative dynamic sensor data. It uses the feature matrix to remove the irrelevant entity features. it was found that the similarity search algorithm based on feature data selection can improve the average search accuracy by more than 10%, as well as increase the search speed and reduce the data transmission and storage costs. Meanwhile, in Feature Selection: A Data Perspective research [2], In this study, the data was grouped into 4 groups: similarity based, information theoretical based, sparse learning based and statistical based methods. it was found that Feature selection is effective in preprocessing data and reducing data dimensionality. Furthermore, it is essential for successful data mining and machine learning applications. It has been a challenging research topic with practical significance inmany areas such as statistics, pattern recognition, machine learning, and data mining. In this research, we will determine the exact data features that can be used to process a dataset without reducing the accuracy results obtained through the KNN algorithm. The dataset will be broken down into 7 different datasets and tested with various comparisons of training data and test data.

3 | METHOD

The research begins by breaking down the datasets used, then testing the classification accuracy of each dataset. Accuracy testing is done by comparing different ratios of training data and test data.

The dataset used in this study is the Iris dataset which contains 4 feature data and 1 data label. In this study, we will look for which feature data can provide a fairly good accuracy value in KNN classification. Testing will be carried out on the original dataset (dataset 0) and on the new dataset (dataset 1 – 6) which contains 2 feature data each (Table 1). The following dataset will be used in testing.

Classification accuracy testing uses the KNN algorithm with the number of n neighbors 5 and the Minkowski metric. Each dataset will be tested based on a comparison of different training data and test data. The comparison of training data and test data is 90:10, 80:20, 70:30, 60:40, 50:50, 40:60, 30:70, 20:80, 10:90. The tools used in this research are jupyter lab with the python programming language.

4 | RESULT AND DISCUSSION

Testing the classification accuracy with the KNN algorithm from seven datasets was carried out with several comparisons of training data and test data. The accuracy of each data set shows various values.

TABLE 1 The breakdown of the dataset used in the research

Dataset	Feature Data			
	Sepal Length	**Sepal Width**	**Petal Length**	**Petal Width**
Dataset 0	✓	✓	✓	✓
Dataset 1	✓	✓		
Dataset 2	✓		✓	
Dataset 3	✓			✓
Dataset 4		✓	✓	
Dataset 5		✓		✓
Dataset 6			✓	✓

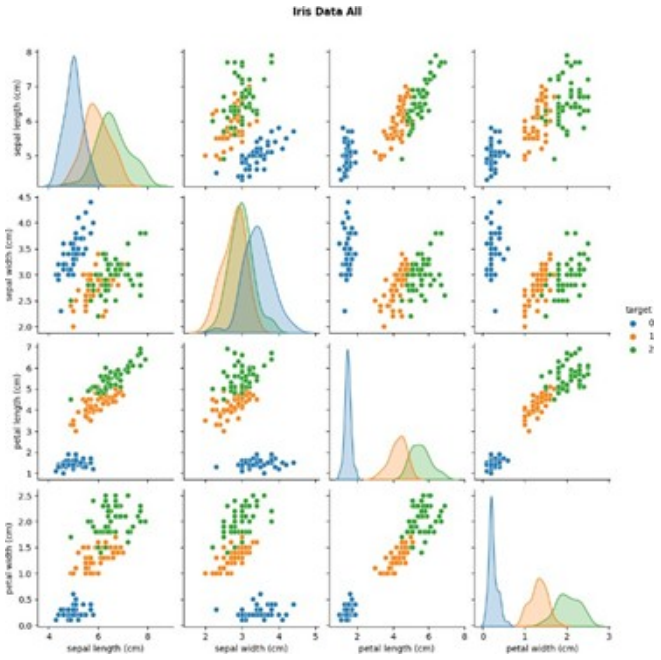


FIGURE 1 Distribution of feature data in dataset 0

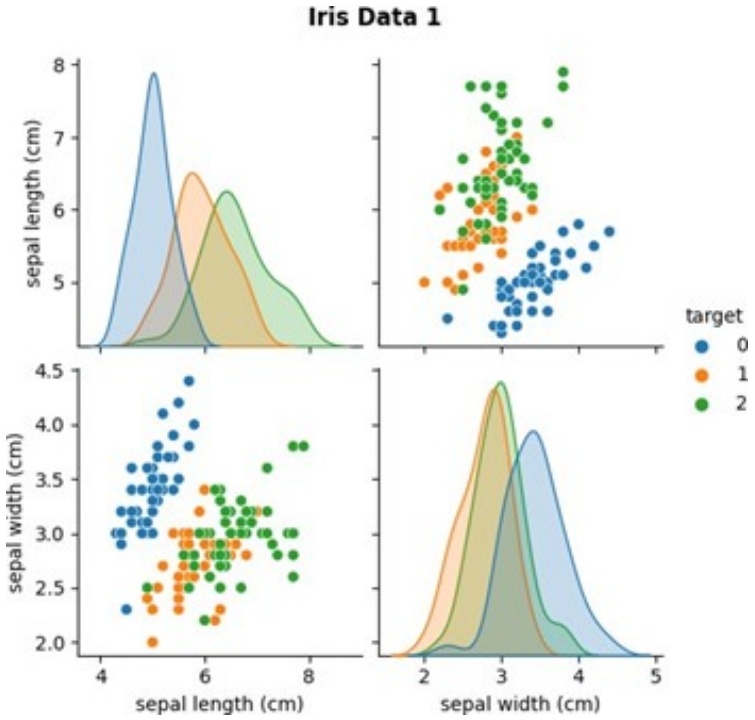
TABLE 2 Accuracy results of dataset 0

Training : Testing Data (%)	Accuracy of Dataset 0 (%)
90 : 10	100
80 : 20	96.67
70 : 30	97.78
60 : 40	95.00
50 : 50	96.00
40 : 60	92.22
30 : 70	91.43
20 : 80	93.33
10 : 90	90.37

4.1 | Data Distribution and Accuracy Value

Dataset 0 is the original dataset from Iris which contains 4 feature data (Sepal Length, Sepal Width, Petal Length, Petal Width). The distribution of the dataset is depicted in Figure 1 . The results of the classification accuracy test on dataset 0 are shown in Table 2 .

Dataset 1 is a dataset that has Sepal Length and Sepal Width data features with the data distribution in Figure 2 and the results of testing the classification accuracy in dataset 1 are shown in Table 3 . Dataset 2 is a dataset that has Sepal Length and Petal Length data features with the data distribution in Figure 3 The results of the classification accuracy test on dataset 2 are shown in Table 4 .

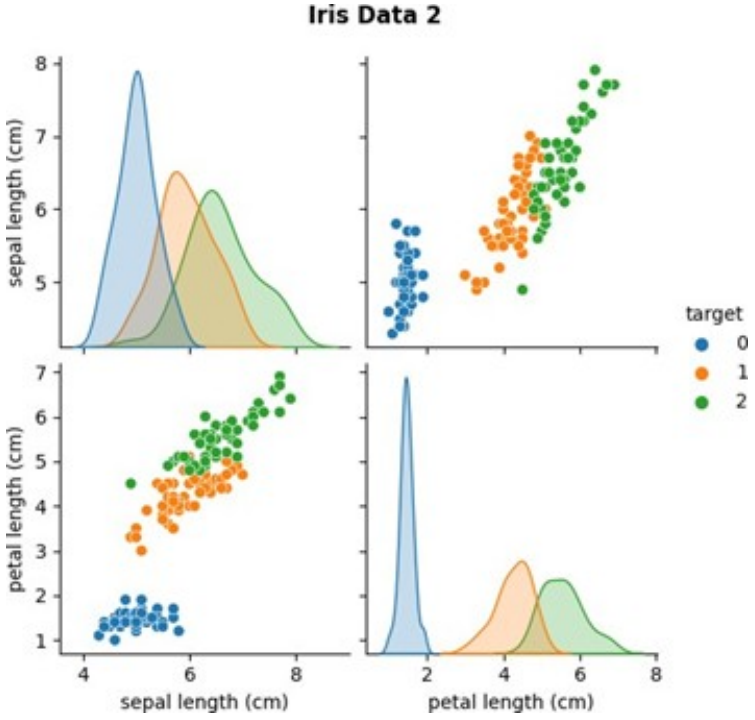


[t!]

FIGURE 2 Distribution of feature data in dataset 1

TABLE 3 Accuracy results of dataset 1

Training Data : Testing Data (%)	Accuracy of Dataset 1 (%)
90 : 10	100
80 : 20	96.67
70 : 30	97.78
60 : 40	95.00
50 : 50	96.00
40 : 60	92.22
30 : 70	91.43
20 : 80	93.33
10 : 90	90.37



[t!]

FIGURE 3 Distribution of feature data in dataset 2

TABLE 4 Accuracy results of dataset 2

Training Data : Testing Data (%)	Accuracy of dataset 2 (%)
90 : 10	86.67
80 : 20	100
70 : 30	97.78
60 : 40	93.33
50 : 50	94.67
40 : 60	95.56
30 : 70	90.48
20 : 80	91.67
10 : 90	89.63

Dataset 3 is a dataset that has Sepal Length and Petal Width data features with the data distribution in Figure 4 . The results of testing the classification accuracy in dataset 3 are shown in Table 5 . Dataset 4 is a dataset that has Sepal Width and Petal Length data features with the data distribution in Figure 5 . The results of the classification accuracy test on dataset 4 are shown in Table 6 .

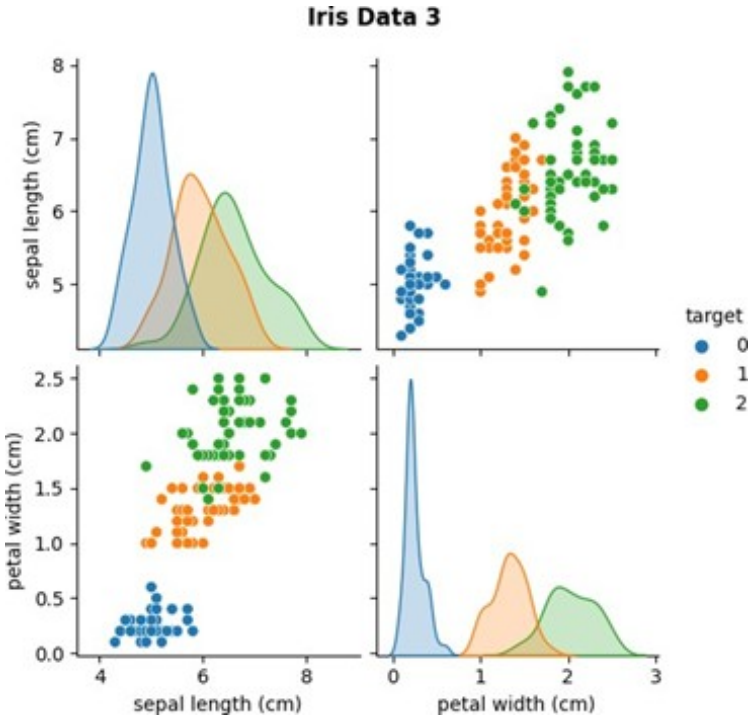


FIGURE 4 Distribution of feature data in dataset 3

TABLE 5 Accuracy results of dataset 3

Training Data : Testing Data (%)	Accuracy of dataset 3 (%)
90 : 10	93.33
80 : 20	96.67
70 : 30	95.56
60 : 40	93.33
50 : 50	94.67
40 : 60	91.11
30 : 70	88.57
20 : 80	86.67
10 : 90	86.67

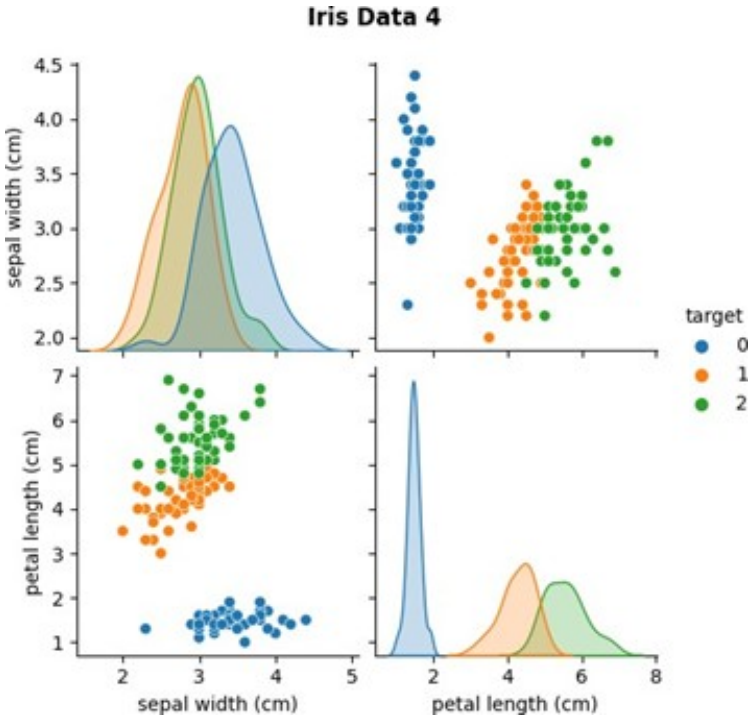


FIGURE 5 Distribution of feature data in dataset 4

TABLE 6 Accuracy results of dataset 4

Training Data : Testing Data (%)	Accuracy of dataset 4 (%)
90 : 10	86.67
80 : 20	96.67
70 : 30	93.33
60 : 40	93.33
50 : 50	93.33
40 : 60	94.44
30 : 70	93.33
20 : 80	91.67
10 : 90	91.85

Dataset 5 is a dataset that has feature data Sepal Width and Petal Width with the data distribution in Figure 6 . The results of the classification accuracy test on dataset 5 are shown in Table 7 . Dataset 6 is a dataset that has

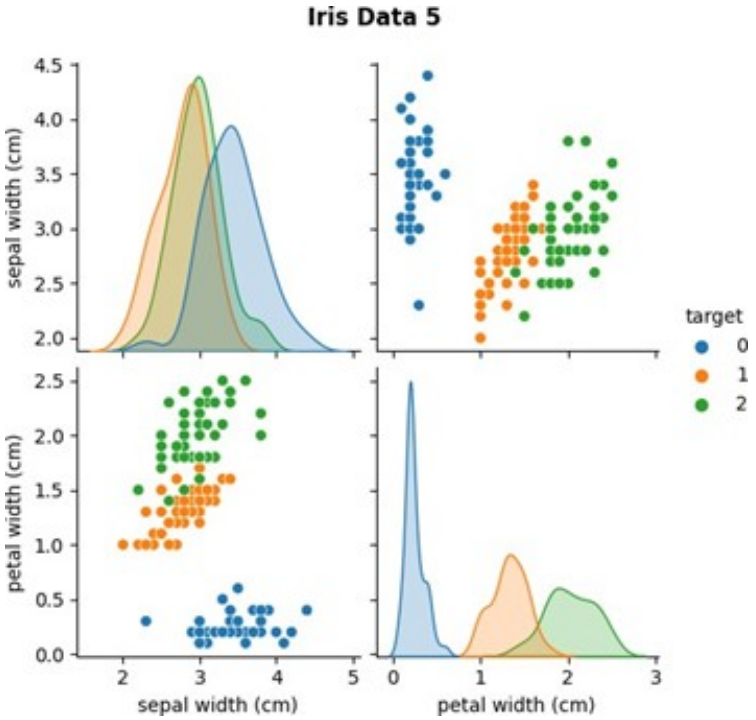


FIGURE 6 Distribution of feature data in dataset 5

TABLE 7 Accuracy results of dataset 5

Training Data : Testing Data (%)	Accuracy of dataset 5 (%)
90 : 10	93.33
80 : 20	96.67
70 : 30	95.56
60 : 40	93.33
50 : 50	94.67
40 : 60	95.56
30 : 70	93.33
20 : 80	95.00
10 : 90	87.41

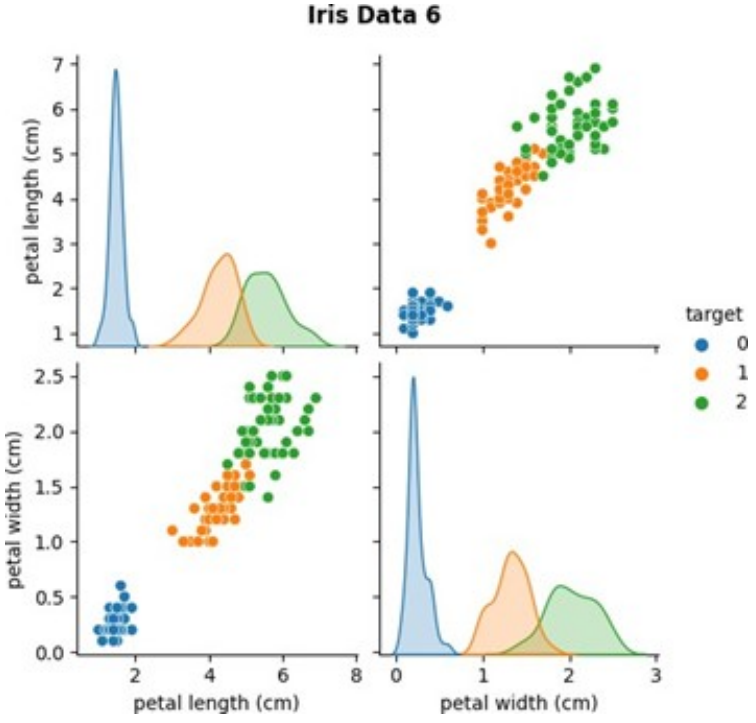


FIGURE 7 Distribution of feature data in dataset 6

TABLE 8 Accuracy results of dataset 6

Training Data : Testing Data (%)	Accuracy of dataset 6 (%)
90 : 10	100.00
80 : 20	100.00
70 : 30	97.78
60 : 40	95.00
50 : 50	96.00
40 : 60	96.67
30 : 70	92.38
20 : 80	91.67
10 : 90	94.81

Petal Length and Petal Width data features with the data distribution in Figure 7 . The results of the classification accuracy test on dataset 6 are shown in Table 8 .

TABLE 9 Accuracy of each dataset

Training Data :	Accuracy (%)						
Testing Data Ratio(%)	**Dataset 0**	**Dataset 1**	**Dataset 2**	**Dataset 3**	**Dataset 4**	**Dataset 5**	**Dataset 6**
90 : 10	100	73.33	86.67	93.33	86.67	93.33	100.00
80 : 20	96.67	66.67	100	96.67	96.67	96.67	100.00
70 : 30	97.78	75.56	97.78	95.56	93.33	95.56	97.78
60 : 40	95.00	66.67	93.33	93.33	93.33	93.33	95.00
50 : 50	96.00	66.67	94.67	94.67	93.33	94.67	96.00
40 : 60	92.22	65.56	95.56	91.11	94.44	95.56	96.67
30 : 70	91.43	68.57	90.48	88.57	93.33	93.33	92.38
20 : 80	93.33	77.50	91.67	86.67	91.67	95.00	91.67
10 : 90	90.37	77.04	89.63	86.67	91.85	87.41	94.81
Average	**95**	**71**	**93**	**92**	**93**	**94**	**96**

TABLE 10 Descriptive Statistics

Dataset	**n**	**Mean (%)**	**Std Dev**	**Min**	**Max**	**Median**
Dataset 0	9	94.76	3.18	90.37	100	95
Dataset 1	9	70.84	4.96	65.56	77.5	68.57
Dataset 2	9	93.31	4.18	86.67	100	93.33
Dataset 3	9	91.84	3.78	86.67	96.67	93.33
Dataset 4	9	92.74	2.7	86.67	96.67	93.33
Dataset 5	9	93.87	2.69	87.41	96.67	94.67
Dataset 6	9	96.03	2.96	91.67	100	96

4.2 | Comparison of Accuracy Results

After the entire dataset has been tested, the accuracy results for each dataset are obtained, the accuracy values are compared and the average value is determined (Table 9). Looking at the results of the comparison of the accuracy values of each dataset, the highest average accuracy value is obtained in dataset 6 while the lowest accuracy value is obtained in dataset 1.

Looking at the distribution of data in dataset 1 and dataset 6 (Figure 8), the distribution of data in dataset 1 is very intersecting for each data target/label. In contrast to dataset 6 which has smaller data distribution slices. This greatly affects the performance of the KNN algorithm which is a distance-based algorithm and KNN will determine the data label based on the majority class of the nearest neighbors. Thus, if there are intersections between data that have different data labels, it can cause errors in determining data labels/classifications. High errors result in low accuracy values obtained. In contrast to dataset 6 which has smaller slices, the labeling for new data will be more accurate because the neighbors of the data will show more uniform data.

We validate that the accuracy improvement is indeed happening using statistical tests (t-test and ANOVA) to validate the significance of the accuracy difference. Table 10 shows the descriptive statistics, and Figure 9 shows the accuracy graphs for each dataset.

In the Independent Samples T-test, 3 comparison tests were carried out between datasets 0 and 6, datasets 1 and 0, and datasets 1 and 6. Figure 10 shows a comparison of the results of the three tests carried out, there was a significant difference between datasets 1 and 0, and datasets 1 and 6, while in datasets 0 and 6 the difference was not significant.

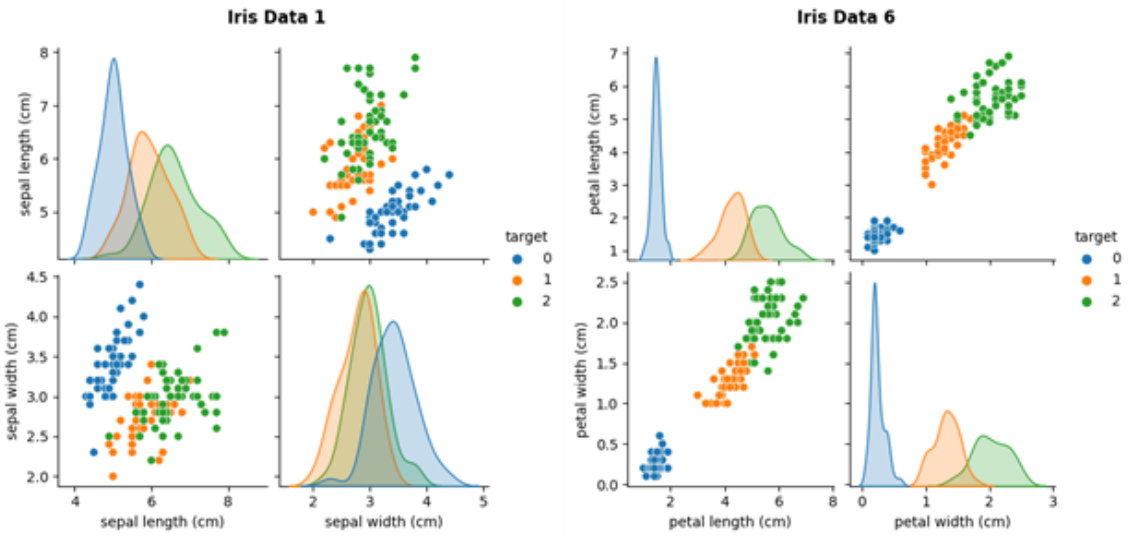


FIGURE 8 Distribution of feature data between dataset 1 and dataset 6.

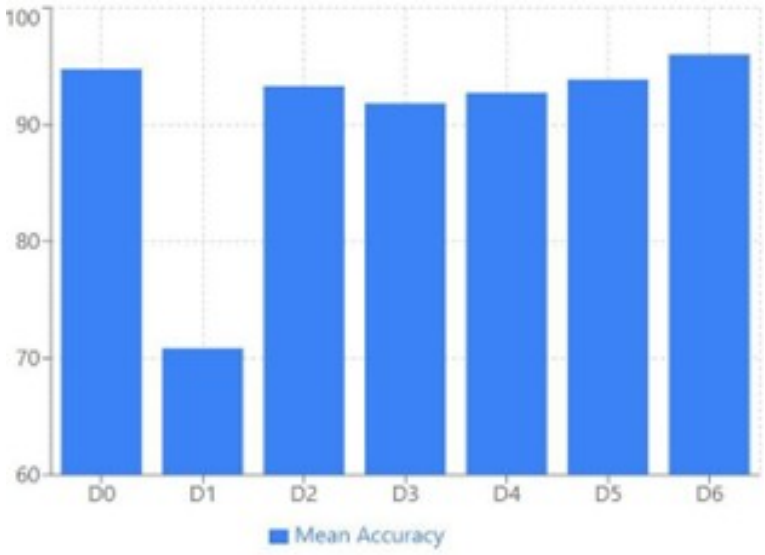


FIGURE 9 Accuracy of Each Dataset.

TABLE 11 ANOVA Table

Source	SS	df	MS	F	p-value
Between Groups	4151.56	6	691.93	53.9747	< 0.05
Within Groups	717.89	56	12.82	-	-

In the one-way ANOVA test, the results were obtained as in table 11 . The results of the ANOVA test obtained an F Statistic value of 53.9747, Critical F of 2.25 and p-value <0.05 and it can be concluded that there are significant differences between the datasets.

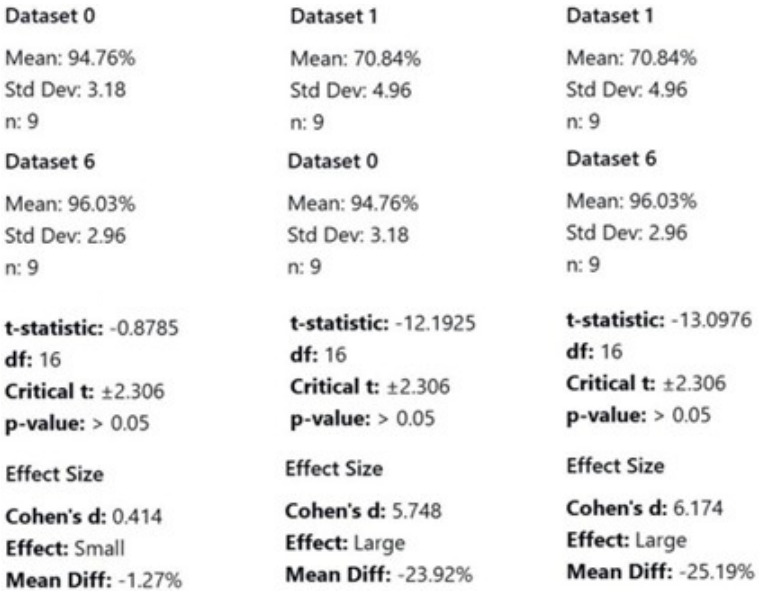


FIGURE 10 Independent Samples T-test Result.

TABLE 12 Comparison Result

Parameter	Result		
	Dataset A	**Dataset B**	**Dataset C**
Number of data features	4	5	4
Accuracy	83,5 %	88,2 %	92,3 %
Selected data features	rfc, loc, amc, cam	wmc, rfc, lcom 3, cam, max_cc	wmc, rfc, lcom3, cam

The results of the two statistical tests showed that the ANOVA F value = 53.9747, indicating a significant difference between datasets ($p < 0.05$) and the best dataset was Dataset 6 (Petal Length + Petal Width) with a mean accuracy of 96.03%. The worst dataset was Dataset 1 (Sepal Length + Sepal Width) with a mean accuracy of 70.84%. The difference between Dataset 1 and 6 was 25.19% (Cohen's d = 6.174, Large effect). The ANOVA test confirmed that feature selection significantly affected classification accuracy. The petal features (petal length and width) provided superior classification performance. The sepal features (sepal length and width) showed consistently low performance. A large effect size ($d > 0.8$) indicated a practically meaningful difference, not just a statistical one. Analysis of the statistical test results provided strong statistical validation that the right feature selection significantly affected the accuracy of KNN classification on the Iris dataset. To reinforce the fact that selecting the right data features can have a positive effect on accuracy results, feature selection was repeated on another dataset. The dataset used was a software defect prediction dataset containing 20 data features and 1 data label. A search was conducted for the best data feature composition that could be used and provide good accuracy values in KNN classification. Testing was carried out on 3 datasets (dataset A, B, C) with dataset A containing 339 data, dataset B containing 339 data, and C containing 567 data. Table 12 shows the results of feature selection and the accuracy obtained. From the overall test, dataset C obtained the best accuracy of 92.3% with 4 data features (out of 20 data features) this indicates that data feature selection affects the accuracy results.

5 | CONCLUSION

Feature data from a dataset can affect the performance of the KNN classification. Selection of the right feature data in a dataset can represent the entire dataset and can be used to classify without reducing accuracy. Proper feature selection can increase accuracy by up to 25%. Furthermore, dimension reduction from 4 features to 2 optimal features does not reduce performance. Determining which dataset to use can be seen from the distribution of the dataset by selecting the distribution of data that has the smallest / fewest slices. It can be seen from the seven datasets tested that dataset 6 has a higher accuracy value than the other datasets, even compared to the original dataset.

Accuracy can be increased by finding the value of n neighbors and using the right distance metric so that the labeling of the majority of neighboring data can be more accurate.

References

1. Liu S, Liu Y, Wu F, Fan W. Feature Data Selection for Improving the Performance of Entity Similarity Searches in the Internet of Things. IEEE Access 2019;7.
2. Li J, et al. Feature selection: A data perspective. ACM Computing Surveys 2017;50(6).
3. Saxsena D, Sathiyanarayanan M. Feature Selection Using Heterogeneous Data Indexes: A data science perspective. In: Proceedings of the 4th International Conference on Contemporary Computing and Informatics (IC3I 2019) Institute of Electrical and Electronics Engineers Inc.; 2019. p. 204–210.
4. Pratama I, Chandra AY, Presetyaningrum PT. Seleksi Fitur dan Penanganan Imbalanced Data menggunakan RFECV dan ADASYN. Jurnal Eksplora Informatika 2022 Jan;11(1):38–49.
5. Saputra J, S W, Sujatmika AR, Arifin AZ. Seleksi Fitur Menggunakan Random Forest Dan Neural Network. In: The 13th Industrial Electronics Seminar 2011 (IES 2011), vol. 1; 2011. p. 93–97.
6. AlShboul R, Thabtah F, Abdelhamid N, Al-diabat M. A visualization cybersecurity method based on features' dissimilarity. Computers & Security 2018 Aug;77:289–303.
7. Sugihartono T. Implementasi Sistem Pendukung Keputusan Penerima Bantuan Rumah Tidak Layak Huni Berbasis Web. Jurnal Sisfokom (Sistem Informasi dan Komputer) 2018 Mar;7(1):52–56.
8. Nugraha W, Sasongko A. Hyperparameter Tuning on Classification Algorithm with Grid Search. SISTEMASI 2022 May;11(2):391.
9. Han MKJ, Pei J. Data Mining: Concept and Techniques. San Francisco: Elsevier; 2012.
10. Wu X, Kumar V. The Top Ten Algorithms in Data Mining. Boca Raton, London, New York: CRC Press Taylor & Francis Group; 2009.
11. Catal C, Diri B. A systematic review of software fault prediction studies. Expert Systems with Applications 2009;36(4):7346–7354.
12. Adeniyi DA, Wei Z, Yongquan Y. Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. Applied Computing and Informatics 2016 Jan;12(1):90–108.

How to cite this article: M. Iman Nur Hakim, Iwan Setyawan, Danny Manongga, Hindriyanto Dwi Purnomo, Hendry, Selection of Feature Data in KNN Classification Datasets, *IPTEK The Journal for Technology and Science*, *37(1): 7-16.*