Jurnal Mantik, 7 (3) (2023), ISSN 2685-4236 (Online)

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik

Optimization of cross-validation testing on the decision tree and k-nearest neighbor in classifying election data

Desilia Selvida[1], Purwa Hasan Putra[2]

[1]Faculty of Computer Science and Information Technology, Computer Science, University North Sumatra, Indonesia

[2]Computer and Informatics Engineering, Computer Engineering, Politeknik Negeri Medan, Indonesia

A R T I C L E I N F O

Article history:

Received Sep 9, 2023
Revised Sep 12, 2023
Accepted Sep 20, 2023

Keywords:

Classifieds;
Cross Validation'
Decision Tree;
Election Data;
K-Nearest Neighbor.

ABSTRACT

General elections are the process of choosing someone who represents the people to occupy a government seat. Polemics regarding the postponement of the 2024 General Elections are widely discussed by Indonesian people. However, the fact is that the position of the government (executive) is currently the majority. This condition is caused by the support of a strong party coalition in the legislature (parliament) in a presidential system. This problem can be solved by data mining. Data mining is one way that can be used to predict and detect a case, including predicting the winning party. There are various kinds of algorithms. The results of the study are positive value predictions (class precision), namely 94.88% with 19 data suitability and 352 data discrepancies, for negative value predictions, namely 60.42% with 29 data suitability and 19 data discrepancies. Meanwhile, the true negative class recall was 94.88% and the true positive was 60.42%. The results of the accuracy of testing with a decision tree is 90.92%. While the results of the K-Nearest Neighbor optimization, it is known that the prediction of positive value (class precision) is 93.98% with 23 data suitability and 352 data discrepancy, for negative prediction value is 67.57% with 25 data suitability and 12 data discrepancy. While the true negative class recall was 96.77% and true positive was 52.08%. The results of the accuracy of testing with a decision tree is 91.65%.

Corresponding Author:

Desilia Selvida,
Computer Science, Faculty of Computer Science and Information Technology,
University North Sumatra,
Jl. Dr. T. Mansur No.9, Padang Bulan, Medan, Sumatera Utara, 20155 Indonesia
Email: desilia.selvida@usu.ac.id

1. INTRODUCTION

The polemic related to the postponement of the 2024 General Election is very much discussed by the Indonesian people (Samponu & Kusrini, 2018). However, the fact is that the position of the government (executive) is currently in the majority. This condition is caused by the support of a strong coalition of parties in the legislature (parliament) in the presidential system. This is called a unified government. The unified government will smooth out every existing discourse to be made into something real. Another example of

this is the implementation of constitutional amendments, as an agreement between the executive and legislative bodies (Jimmy et al., 2023).

General elections in Indonesia have experienced several changes from one election period to another. During the New Order elections, we were familiar with a closed list proportional election system (Zarti et al., 2023). The election of legislative candidates is not determined by voters, but becomes the authority of political party elites in accordance with the composition of the list of candidates along with serial numbers (RI Law No. 10 of 2008) (Badrul et al., 2015).

This problem can be solved by data mining. Data Mining is the extraction of important and interesting information or patterns from very large database data (Triyansyah & Fitrianah, 2018). Data mining has one technique, namely a classification technique which is useful for predicting the value of the categorical target variable (Zarti et al., 2023). Data mining is one way that can be used to predict and detect a case, including predicting the winning party (Karo et al., 2018). There are various kinds of algorithms or methods in data mining that can be used in predicting them, including linear regression methods, decision trees, k-means algorithms, and so on (Rahayu et al., 2022).

Previously, there was an additional cross validation technique for optimizing the K-nearest neighbors' algorithm by evaluating the optimal algorithm model. This technique will optimize the results of accuracy and concepts that are important in data science and data analysis. And even then, used to prevent or at least minimize overfitting. Overfitting means that the model is adjusted too much to the training data (Prasetyo & Laksana, 2022).

The Decision tree method is often used to classify an image or image and is the method most frequently used, in the use of a decision tree method. items will be grouped with a decision, so that it will be easy to understand (Robianto ; Sampe Hotlan Sitorus ; Uray Ristian, 2021). The Decision Tree is one of the most popular supervised learning-based methods for classification. It is an iterative top-down based approach (Az-zahra et al., 2021). A decision tree comprises a root node, decision nodes, and leaf nodes. The root node represents the most important attribute of the dataset used to obtain the best prediction (Putra et al., 2023).

Using the K-Nearst Neighbor (KNN) method is a method of classifying a set of data based on the majority of categories and the goal is to classify new objects based on attributes and sample samples from training data (Hasan Putra et al., 2022).

Based on previous research where the decision tree and k-nearest neighbor methods can be used in data classification. Therefore, the authors will implement a system to determine Optimization of Cross Validation Testing on the Decision Tree and K-Nearest Neighbor in Classifying Election Data.

2. RESEARCH METHOD

2.1 Research Stages

This framework is the steps that will be taken in order to solve the problem to be discussed. Figure 1. below is the framework (framework) used in this study.

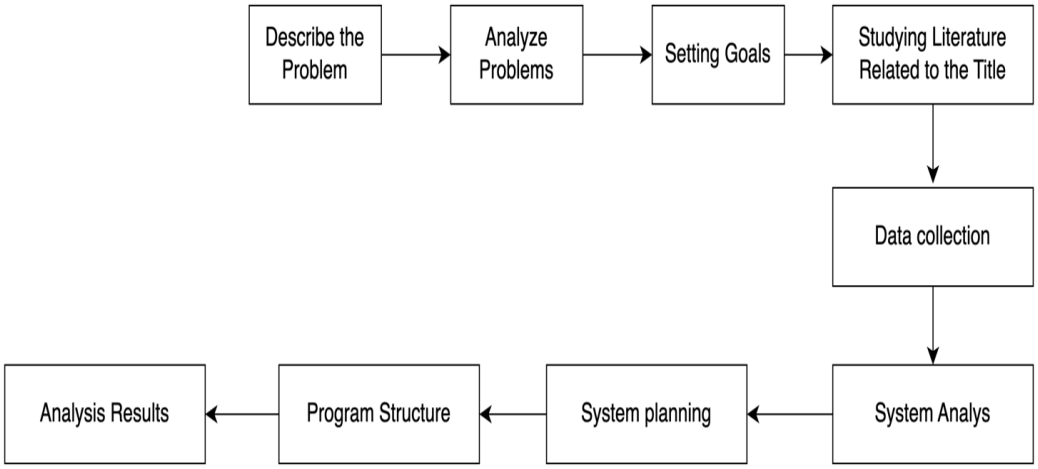

Figure 1. Research Procedure Framework

Based on the framework in Figure 1, each step can be described as follows: Describe the problem clearly to get the results of an Optimization of Cross Validation Testing on the Decision Tree and K-Nearest Neighbor in Classifying Election Data. The problem analysis step is a step to understand the problem whose scope or boundaries have been determined. By analyzing the predetermined problem, it is hoped that the problem can be well understood. Based on the understanding of the problems of the problems, the objectives to be achieved in this study were determined. This goal determines the targets to be achieved, especially those that can overcome existing problems. To achieve the goal, then studied some of the literature that is expected to be used. Then the literature studied is selected which will be used in this study. Literature sources were obtained from books and journals. The data needed is data that will be used as material for research, namely data on types of payment deals from Kaggle. System analysis is quite important to do, because here the author must know the weaknesses of the system, obstacles, constraints and opportunities that are not able to be achieved by the current system in order to find alternative solutions to the problem. The system is designed using rapid miner tools to manage payment type data by looking at the accuracy and percentage of calculations for each method. Program Structure Design is a design that describes the relationship between a communication system with other communication systems. At this stage it will provide the results of the research analysis of the naïve Bayes classifier and decision tree models which produce the percentage model of each method.

2.2 Cross Validation

Cross validation or can be called rotational estimation is a model validation technique to assess how the results of statistical analysis will generalize to independent data sets (Azis et al., 2020). K fold cross validation is used to estimate prediction error in evaluating model performance. The data is divided into k subsets of almost equal number. Models in classification trained and tested k. In each iteration, one of the subsets will be used as training data and testing data (Mardiana et al., 2022). Cross validation or rotation estimation is a model validation technique to assess and find out

how the statistical analysis results will be generalize independent data sets (Tuntun et al., 2022)

Cross validation is a method of taking a training data and test data randomly to ensure the similarity of the number of observed datasets and training datasets as well as one appearance in the testing dataset (Fauzi & Yunial, 2022).

At stage the training dataset is divided into training data and validation data. The model will be trained using training data and validated using k-fold validation data. In this research used 5-fold cross validation in evaluating the performance of the model (Martini et al., 2022). Cross validation is one of the statistical methods implemented to evaluate the performance of the model or algorithm that has been designed. At the training stage the dataset is divided into training data and validation data (FUADAH et al., 2022).

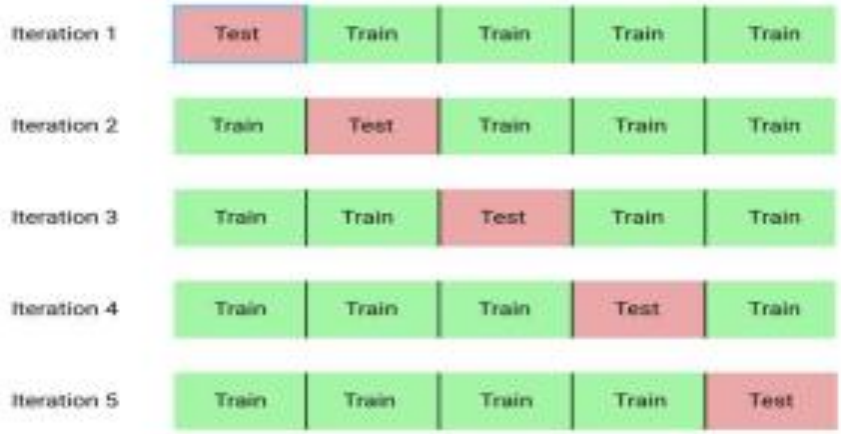

Figure 2. Cross Validation Simulation

2.3 Decision Tree

The Decision Tree Algorithm is a classification method that uses a tree example, stating the nodes that describe each attribute, where the leaves describe each class, also each branch describes the value of each class (Estian Pambudi et al., 2022). The root node represents the node at the top of the tree. Each of these nodes represents a dividing node, where each of these nodes is one input and has at least two outputs. Leaf node is the last node, has only one input, and has no output (Solehuddin et al., 2022). The decision tree at each leaf node represents the label for each class. The decision tree in each branch states the conditions that must be filled in and each tree top describes the value of the data class (Robianto ; Sampe Hotlan Sitorus ; Uray Ristian, 2021). Decision Tree adalah struktur flowchart yang menyerupai Tree (pohon), dimana setiap simpul internal menandakan suatu tes pada atribut, setiap cabang merepresentasikan hasil tes,dan simpul daun merepresentasikan kelas atau distribusi kelas (Salasa & Maharani, 2022).

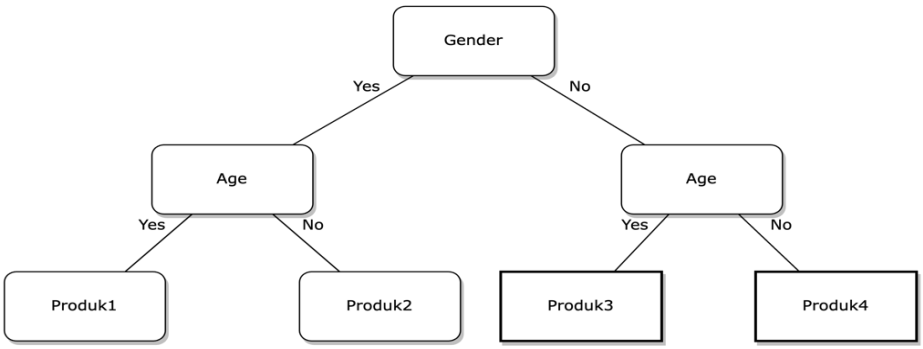

Figure 3. The concept of the Decision Tree

2.4 K-Nearest Neighbor

K-Nearest Neighbor (K-NN) is an instance-based learning group. This algorithm is also a lazy learning technique. K-NN is done by finding groups of k objects in the training data that are closest (similar) to the objects in the new data or testing data (Zulaikhah Hariyanti Rukmana et al., 2022). Case in point, for example, it is desired to find a solution to a new patient's problem by using a solution from an old patient (Purwani et al., 2022). To find solutions from these new patients, proximity to old patient cases is used, solutions from old cases that are close to new cases are used as a solution. (Azis et al., 2020).

There are many ways to measure the proximity of new data to old data (training data), including the Euclidean distance and the Manhattan distance (city block distance), the most commonly used is the Euclidean distance. The euclidean equations are shown in Eq (Azis et al., 2020).

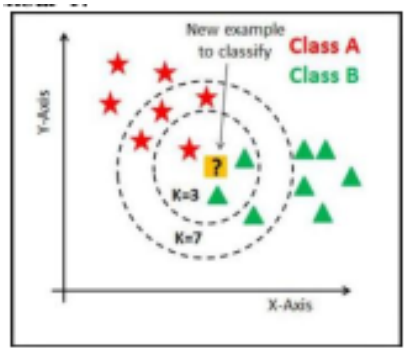

Figure 4. KNN Workflow

3. RESULTS AND DISCUSSIONS

The data mining process is carried out using the Optimization of Cross Validation Testing on the Decision Tree and K-Nearest Neighbor classification (Dollen et al., 2023). In this research, the main research approach will be carried out, namely qualitative and quantitative approaches. The research methodology carried out consisted of 4 stages, which can be seen as follows:

Figure 5. Research Methodology

3.1 Data Understanding

At this stage the researcher begins by collecting initial data and then combining them into one dataset, the data that has been collected will be examined and the results of data collection activities in order to identify problems with the data that has been collected. This stage provides an analytical foundation for a study by summarizing and identifying potential problems in the data (Rahayu et al., 2022). This stage must also be carried out carefully and not in a hurry by data practitioners. Examples of data visualization performed by data practitioners. Usually, if you are not careful, insight or conclusions cannot be found immediately. The data in this study were obtained from the Election Commission dataset with 9 attributes and a total of 425 data.

Row No.	TERPILIH ATAU TIDAK	NAMA PARTAI POLITIK	JENIS KELA...	KECAMATAN	NO.URUT P...	SUARA SAH...	JUML.PERO...	DAERAH PE...	NO.URUT C...	SUARA SAH...
1	TIDAK	HATI NURANI RAKYAT	L	LEBAKSIU	1	18578	1	1	1	594
2	TIDAK	HATI NURANI RAKYAT	L	SLAWI	1	18578	1	1	2	943
3	TIDAK	HATI NURANI RAKYAT	P	SLAWI	1	18578	1	1	3	1730
4	YA	HATI NURANI RAKYAT	L	DUKUHWARU	1	18578	1	1	4	2508
5	TIDAK	HATI NURANI RAKYAT	L	SLAWI	1	18578	1	2	1	923
6	TIDAK	HATI NURANI RAKYAT	P	TARUB	1	18578	1	2	2	308
7	TIDAK	HATI NURANI RAKYAT	L	TARUB	1	18578	1	2	3	54
8	TIDAK	HATI NURANI RAKYAT	L	BOJONG	1	18578	1	3	1	1682
9	TIDAK	HATI NURANI RAKYAT	P	JATINEGARA	1	18578	1	3	2	918
10	TIDAK	HATI NURANI RAKYAT	L	SLAWI	1	18578	1	3	3	87
11	TIDAK	HATI NURANI RAKYAT	L	BALAPULANG	1	18578	1	4	1	728
12	TIDAK	HATI NURANI RAKYAT	L	MARGASARI	1	18578	1	4	2	346
13	TIDAK	HATI NURANI RAKYAT	P	LEBAKSIU	1	18578	1	4	3	184
14	TIDAK	HATI NURANI RAKYAT	L	LEBAKSIU	1	18578	1	5	1	381
15	TIDAK	HATI NURANI RAKYAT	L	ADIWERNA	1	18578	1	5	2	148
16	TIDAK	HATI NURANI RAKYAT	P	SLAWI	1	18578	1	5	3	128
17	TIDAK	HATI NURANI RAKYAT	L	SLAWI	1	18578	1	6	1	441
18	TIDAK	HATI NURANI RAKYAT	L	BOJONG	1	18578	1	6	2	88
19	TIDAK	HATI NURANI RAKYAT	P	BALAPULANG	1	18578	1	6	3	74
20	TIDAK	PARTAI KARYA PEDULI ...	P	PAGERBARA...	2	12373	0	1	1	351

Figure 6. KPU election data

3.2 Data Preprocessing

Preprocessing is the process of preparing raw data for use in the data transformation process into the data format required by the user. The next process before the algorithm model is made is data preprocessing. In this study, preprocessing techniques were used, namely: cleansing, data aggregation, checking for missing values. The data generated from the KPU dataset still contains redundancies that will interfere with the classification process, so preprocessing is needed to filter and clean it. The following is the result of the preprocessing process.

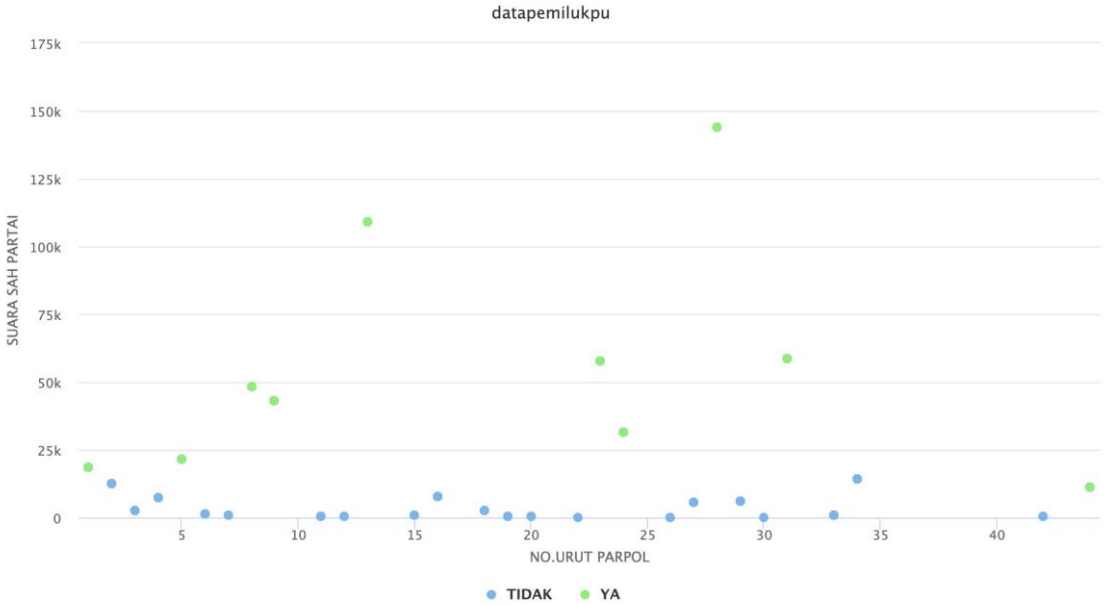

Figure 7. Graph of KPU Election Data

Classification is the process of grouping an idea, data or a new group into a group that has something in common with one of the class characteristics. The classification process includes the training process and data testing. The dataset for the training and testing process uses data from the KPU which continues to the prediction stage. Data from the predicted KPU is processed and compared with the stored training results. The training process aims to create a machine learning model that can classify election data. Testing is carried out to see the performance of the Cross Validation model that has been made. The output of the classification results in the election data classification system by testing the Decision Tree and K-Nearest Neighbor.

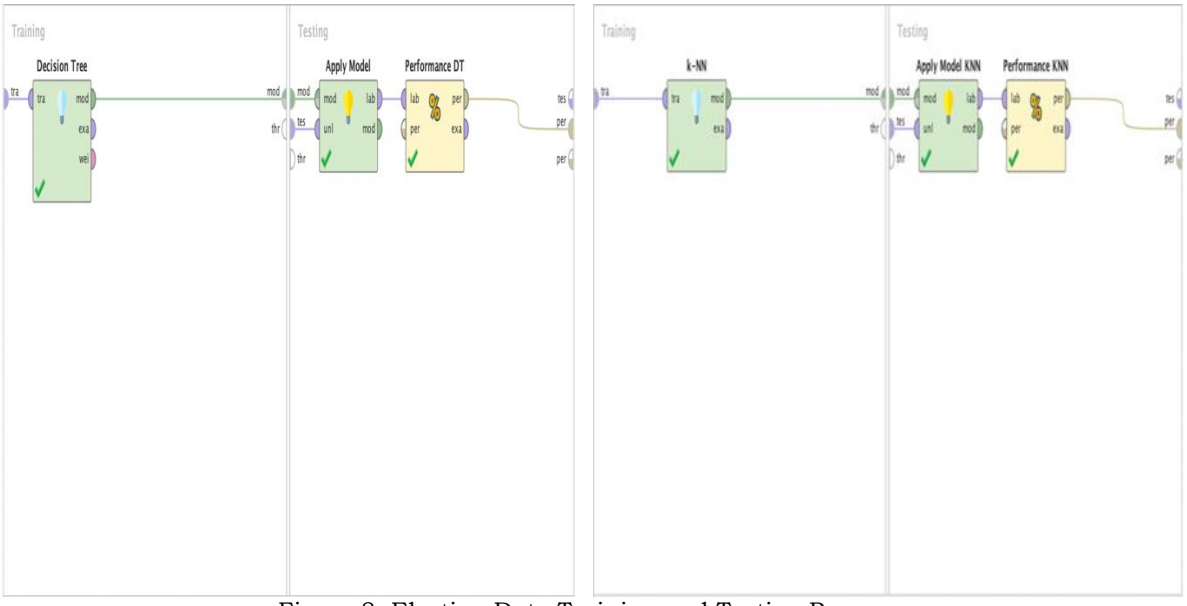

Figure 8. Election Data Training and Testing Process

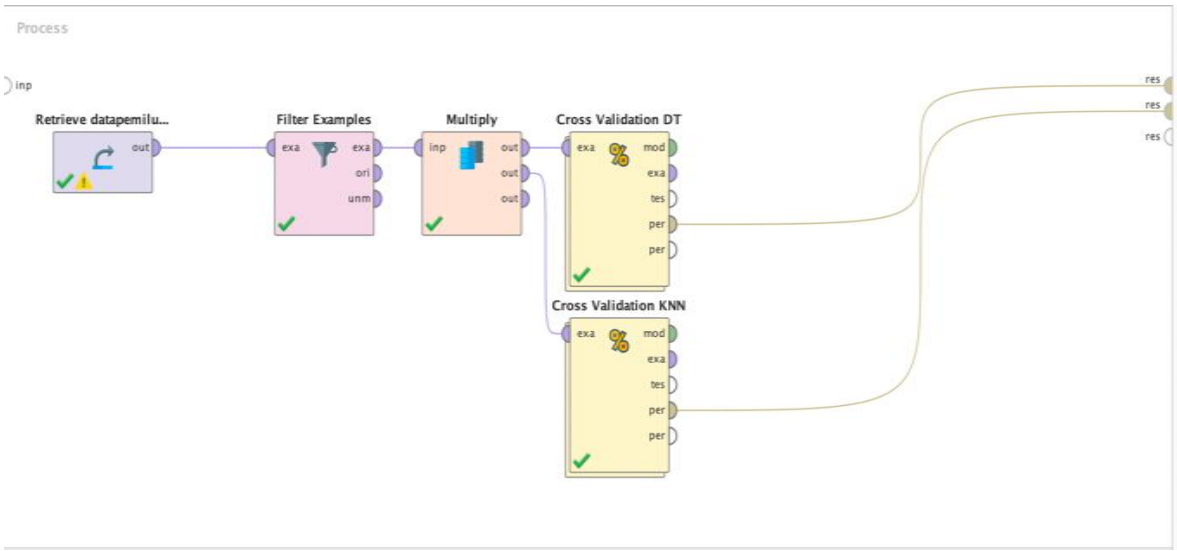

Figure 9. Algorithm Optimization Process

In table 1. It is known that the prediction of positive values (class precision) is 94.88% with 19 data suitability and 352 data discrepancies, for predictions of negative values, namely 60.42% with 29 data suitability and 19 data discrepancies. Meanwhile, the true negative class recall was 94.88% and the true positive was 60.42%. The results of the accuracy of testing with a decision tree is 90.92%.

Table 1. Decision Tree Accuracy Results

	True No	True Yes	Class precision
Pred. No	352	19	94.88%
Pred Yes	19	29	60.42%
Class Recall	94.88%	60.42%	

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$
$$\text{accuracy} = \frac{352+29}{352+29+19+19}$$
accuracy: 90.92%

In table 2. It is known that the prediction of positive value (class precision) is 93.98% with 23 data suitability and 352 data discrepancy, for negative prediction value is 67.57% with 25 data suitability and 12 data discrepancy. While the true negative class recall was 96.77% and true positive was 52.08%. The results of the accuracy of testing with a decision tree is 91.65%

Table 2. K-Nearest Neighbor Accuracy Results

	True No	True Yes	Class precision
Pred. No	352	23	93.98%
Pred Yes	12	25	67.57%
Class Recall	96.77%	52.08%	

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$
$$\text{accuracy} = \frac{352+25}{352+25+23+12}$$
accuracy: 91.65%

4. CONCLUSION

The conclusions generated are based on the research that has been done, namely Optimization of Cross Validation Testing on the Decision Tree and K-Nearest Neighbor in Classifying Election Data. The results of the study with Decision Tree optimization are known to predict positive values (class precision), namely 94.88% with 19 data suitability and 352 data discrepancies, for negative prediction values, namely 60.42% with 29 data suitability and 19 data discrepancies. Meanwhile, the true negative class recall was 94.88% and the true positive was 60.42%. The results of the accuracy of testing with a decision tree is 90.92%. While the results of the K-Nearest Neighbor optimization, it is known that the prediction of positive value (class precision) is 93.98% with 23 data suitability and 352 data discrepancy, for negative prediction value is 67.57% with 25 data suitability and 12 data discrepancy. While the true negative class recall was 96.77% and true positive was 52.08%. The results of the accuracy of testing with a decision tree is 91.65%. In this study the authors found that the Decision Tree and K-Nearest Neighbor methods can be combined with very good accuracy results. From this research the author found that suggestions for further research are needed by applying more fund records to get the results of value accuracy with a better percentage. Research can also be developed by adding algorithms for unbalanced data or imblanced data in the dataset.

REFERENCES

Az-zahra, A. A., Marsaoly, A. F., Lestyani, I. P., Salsabila, R., & Madjida, W. O. Z. (2021). Penerapan Algoritma K-Modes Clustering Dengan Validasi Davies Bouldin Index Pada Pengelompokkan Tingkat Minat Belanja Online Di Provinsi Daerah Istimewa Yogyakarta. Jurnal MSA (Matematika Dan Statistika Serta Aplikasinya), 9(1), 24. https://doi.org/10.24252/msa.v9i1.18555

Azis, H., Purnawansyah, P., Fattah, F., & Putri, I. P. (2020). Performa Klasifikasi K-NN dan Cross Validation Pada Data Pasien Pengidap Penyakit Jantung. ILKOM Jurnal Ilmiah, 12(2), 81–86. https://doi.org/10.33096/ilkom.v12i2.507.81-86

Badrul, M., Studi, P., & Informasi, S. (2015). Prediksi Hasil Pemilu Legislatif Dengan Menggunakan Algoritma K-Nearest Neighbor. Jurnal Pilar Nusa Mandiri, 11(2), 152–160.

Dollen, D. Von, Neukart, F., Weimer, D., & Bäck, T. (2023). Predicting vehicle prices via quantum - assisted feature selection. International Journal of Information Technology. https://doi.org/10.1007/s41870-023-01370-z

Estian Pambudi, R., Sriyanto, & Firmansyah. (2022). Klasifikasi Penyakit Stroke Menggunakan Algoritma Decision TreeC.45. Ijccs, x, No.x(x), 1–5.

Fauzi, A., & Yunial, A. H. (2022). JEPIN (Jurnal Edukasi dan Penelitian Informatika) Optimasi Algoritma Klasifikasi Naive Bayes, Decision Tree, K-Nearest Neighbor, dan Random Forest menggunakan Algoritma Particle Swarm Optimization pada Diabetes Dataset. (JEPIN) Jurnal Edukasi Dan Penelitian Informatika, 8(3), 470–481.

FUADAH, Y. N., UBAIDULLAH, I. D., IBRAHIM, N., TALININGSING, F. F., SY, N. K., & PRAMUDITHO, M. A. (2022). Optimasi Convolutional Neural Network dan K-Fold Cross Validation pada Sistem Klasifikasi Glaukoma. ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika, 10(3), 728. https://doi.org/10.26760/elkomika.v10i3.728

Hasan Putra, P., Syahputra Novelan, M., & Rizki, M. (2022). Analysis K-Nearest Neighbor Method in Classification of Vegetable Quality Based on Color. Journal of Applied Engineering and Technological Science, 3(2), 126–132.

Jimmy, Hermaliani, E. H., & Kurniawati, L. (2023). Analisis Klasifikasi Sentimen Pengguna Media Sosial Twitter Terhadap Penundaan Pemilu Presiden Tahun 2024. Jurnal Indonesia : Manajemen Informatika Dan Komunikasi, 4(2), 570–579. https://doi.org/10.35870/jimik.v4i2.243

Karo, I. M. K., Huda, A. F., & MaulanaAdhinugraha, K. (2018). A cluster validity for spatial clustering based on davies bouldin index and Polygon Dissimilarity function. Proceedings of the 2nd International Conference on Informatics and Computing, ICIC 2017, 2018-Janua, 1–6. https://doi.org/10.1109/IAC.2017.8280572

Mardiana, L., Kusnandar, D., & Satyahadewi, N. (2022). Analisis Diskriminan Dengan K Fold Cross Validation Untuk Klasifikasi Kualitas Air Di Kota Pontianak. Buletin Ilmiah Mat. Stat. Dan Terapannya (Bimaster), 11(1), 97–102.

Martini, M., Anwar, R. S., & Masshitah, S. (2022). Analisa Decision Tree Untuk Menentukan Jadwal Kerja Karyawan Restoran Pada Hari Libur. JURNAL PETISI (Pendidikan Teknologi Informasi), 3(1), 5–14. https://doi.org/10.36232/jurnalpetisi.v3i1.2041

Prasetyo, A. B., & Laksana, T. G. (2022). Optimasi Algoritma K-Nearest Neighbors dengan Teknik Cross Validation Dengan Streamlit (Studi Data: Penyakit Diabetes). Journal of Applied Informatics and Computing (JAIC), 6(2), 194. http://jurnal.polibatam.ac.id/index.php/JAIC

Purwani, F., Wahyudi, R. T., & Jaya, I. D. (2022). Penerapan Algoritma K-Nearest Neighbor dengan Euclidean Distance untuk Menentukan Kelompok Uang Kuliah Tunggal Mahasiswa. Edumatic: Jurnal Pendidikan Informatika, 6(2), 344–353. https://doi.org/10.29408/edumatic.v6i2.6547

Putra, P. H., Purba, B., & Dalimunthe, Y. A. (2023). Random forest and decision tree algorithms for car price prediction. 1(2), 81–89.

Rahayu, W. I., Anindita, A., & Fauzan, M. N. (2022). PENENTUAN VALIDASI DATA PEMILIH DAN KLASIFIKASI HASIL PEMILU DPRD KAB.BONE UNTUK MEMPREDIKSI PARTAI PEMENANG MENGGUNAKAN METODE NAIVE BAYES Program Studi D4 Teknik Informatika 123 Politeknik Pos Indonesia 123. Jurnal Teknik Informatika, 14(1), 32–39.

Robianto ; Sampe Hotlan Sitorus ; Uray Ristian. (2021). Penerapan Metode Decision Tree Untuk Mengklasifikasikan Mutu Buah Jeruk BerdasarkanFitur Warna Dan Ukuran. Jurnal Komputer Dan Aplikasi, 9(01), 76–86.

Salasa, S. A., & Maharani, W. (2022). Personality Detection of Twitter Social Media Users using the Support Vector Machine Method. Jurnal Sistem Komputer Dan Informatika (JSON), 4(2), 263. https://doi.org/10.30865/json.v4i2.5345

Samponu, Y. B., & Kusrini, K. (2018). Optimasi Algoritma Naive Bayes Menggunakan Metode Cross Validation Untuk Meningkatkan Akurasi Prediksi Tingkat Kelulusan Tepat Waktu. Jurnal ELTIKOM, 1(2), 56–63. https://doi.org/10.31961/eltikom.v1i2.29

Solehuddin, M., Syafei, W. A., & Gernowo, R. (2022). Metode Decision Tree untuk Meningkatkan Kualitas Rencana Pelaksanaan Pembelajaran dengan Algoritma C4.5. Jurnal Penelitian Dan Pengembangan Pendidikan, 6(3), 510–519. https://doi.org/10.23887/jppp.v6i3.52840

Triyansyah, D., & Fitrianah, D. (2018). Analisis Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing. Jurnal Telekomunikasi Dan Komputer, 8(3), 163. https://doi.org/10.22441/incomtech.v8i3.4174

Tuntun, R., Kusrini, K., & Kusnawi, K. (2022). Analisis Perbandingan Kinerja Algoritma Klasifikasi dengan Menggunakan Metode K-Fold Cross Validation. Jurnal Media Informatika Budidarma, 6(4), 2111. https://doi.org/10.30865/mib.v6i4.4681

Zarti, M. N., Sahputra, E., Sonita, A., & ... (2023). Application Of Data Mining Using The Naïve Bayes Classification Method To Predict Public Interest Participation In The 2024 Elections. Jurnal Komputer ..., 3(1), 105–114. https://penerbitadm.com/index.php/KOMITEK/article/view/1192%0Ahttps://penerbitadm.com/index.php/KOMITEK/article/download/1192/1648

Zulaikhah Hariyanti Rukmana, S., Aziz, A., & Harianto, W. (2022). Optimasi Algoritma K-Nearest Neighbor (Knn) Dengan Normalisasi Dan Seleksi Fitur Untuk Klasifikasi Penyakit Liver. JATI (Jurnal Mahasiswa Teknik Informatika), 6(2), 439–445. https://doi.org/10.36040/jati.v6i2.4722