Plagiarism Checker X Originality Report

Similarity Found: 9%

Date: Saturday, December 03, 2022
Statistics: 326 words Plagiarized / 3813 Total words
Remarks: Low Plagiarism Detected - Your Document needs Optional Improvement.

IJISCS | 1 COMPARISON OF K-NEAREST NEIGHBOR AND NAÏVE BAYES FOR BREAST CANCER CLASSIFICATION USING PYTHON Irma Handayani1, Ikrimach2 1Depaetment of Informatics, University of Technology Yogyakarta 2Department of Information System, University of Technology Yogyakarta 1,2 Siliwangi (Ringroad Utara) Street, Jombor, Sleman, D.I.Yogyakarta, Indonesia *Corresponding author irma.handayani@staff.uty.ac.id ikrimach@staff.uty.ac.id Article history: Received October 10, 2020 Revised November 3, 2020 Accepted November 16, 2020 Keywords: data mining; classification; k-nn; naïve bayes; breast cancer.

Abstract Classification is widely used to determine decisions according to new knowledge gained from processing past data using algorithms. The number of attributes can affect the performance of an algorithm. Several data mining methods that are widely used for classification include the K-Nearest Neighbor and naïve Bayes algorithm. The best algorithm for one data type is not necessarily good for another data type. It is even possible that a good algorithm will be horrendous for other data types. To overcome this issue, this study will analyze the accuracy of the K-Nearest Neighbor and Naïve Bayes algorithms for the classification of breast cancer.

So that patients with existing parameters can be predicted which are malignant and benign breast cancer. This pattern can be used as a diagnostic measure so that the cancer can be detected earlier and is expected to reduce the mortality rate from breast cancer. The test results using k- fold (k-10) cross validation, followed by confusion matrix of 455 data consisting of 284 data on cases of benign cancer, 171 data on malignant cancer cases, on the K-NN classifier were able to correctly classify 441 data with an accuracy rate of 0.97%, while the Naïve Bayes classifier was able to correctly classify 428 data with an accuracy rate of 0.94%. 1.0

INTRODUCTION Breast cancer (Carcinoma mammae) is an uncontrolled growth of cells in the milk-producing glands (lobular), glandular ducts from the lobular to the nipple (ductus), and the supporting tissue of the breast that surrounds the lobules, ducts, blood vessels and lymph vessels, but excludes breast skin [1]. Cancer or often referred to as a tumor is generally divided into two types, namely benign and malignant. At a benign level, the tumor will have a noncancerous condition and progression, where the disease can be detected but does not spread and damage surrounding tissue.

Meanwhile, at a malignant level, the tumor will spread and damage the surrounding tissues and organs [2]. Recording of cancer is often done to anticipate and analyze patients from an early age so that prevention can be done. By knowing cancer early, the handling will be easier because cancer cells have not developed further [3]. In general, the detection of the level of malignancy of breast cancer is by means of prognosis. Full Paper eISSN : 2598-246X pISSN : 2598-0793 International Journal Information System and Computer Science (IJISCS) IJISCS | 2 The prognosis is the medical team's "best guess" in determining whether or not a patient is cured of breast cancer.

Apart from prognosis, another way is the use of bioinformatics using data mining techniques, because it has been proven to be able to detect the level of malignancy of breast cancer [4]. As information technology advances, especially in the field of artificial intelligence, machine learning techniques were introduced to help improve automatic detection capabilities. With the help of this system, the possibility of misdiagnosis made by medical professionals can be avoided, and medical data can be checked in a short time and in more detail [5].

An extraction process to find information in previously unknown data is known as data mining [6]. Data mining uses pattern recognition techniques such as statistics and mathematics to find patterns from old data or cases [7]. One of the main roles of data mining is classification. Classification is widely used to determine decisions according to new knowledge gained from processing past data using algorithms. In the classification dataset, there is one objective attribute or it can also be called the label attribute. This attribute will be searched from new data on the basis of other attributes in the past.

The number of attributes can affect the performance of an algorithm. This results in if the classification process is inaccurate, the researcher needs to double-check at each previous stage to look for errors. Several data mining methods that are widely used for classification include the K-Nearest Neighbor and Naïve Bayes algorithms. In previous research, many data mining methods have been used to diagnose diseases using the K-NN and Naïve Bayes algorithms. As in [8] using the K-NN algorithm for kidney stone

classification, the classification is performed using K-NN machine learning algorithm with the 10 nearest neighbors, where the accuracy as high as 98.17% is achieved.

Researcher [9] performed optimization of the k parameter in the K-NN algorithm for breast cancer detection, the results showed that the KNN with a k value of 13 had the best accuracy rate of 97.28% with an error value of 1.5% and a micro value of 97,28%. Researcher [10] Naive Bayes classification to predict colon cancer, it achieves up to 95.24% classification accuracy, thus this model can be an efficient analysis tool. Researchers [11] Detection of tumor types using naïve bayes, the five tumor types of Gene Expression Cancer RNA-Seq Data Set on WEKA tool classified using Naive Bayes algorithm with a 98.7516% accuracy, 98.5% accuracy, 98.7526% accuracy and 98.5294% accuracy by 10 -fold cross validation, 50 – 50% train-test, 40- 60% train-test, 66-34% train-test data partition, respectively. Based on previous studies, it shows that the K-NN and naïve Bayes algorithms are proven to be good at classifying.

This study applies the K-NN and Naïve Bayes algorithms for the classification of breast cancer by analyzing the results of the algorithm's accuracy. The K-NN algorithm takes a data classification approach by optimizing sample data which can be used as a reference for training data to produce data classification for breast cancer based on the learning process. Meanwhile, the Naïve Bayes method performs a classification based on probability and the Bayesian theorem. 2.0 THEORETICAL 2.1.

Data Mining Data mining is the activity of extracting important information or knowledge from a large data set using certain techniques. Information or knowledge generated from data mining can be used to improve decision making. Some preliminary steps before we enter ready-made data into certain data mining techniques are [12]: 1) Data selection: selection of data sets that will be used from the existing database in accordance with the desired purpose 2) Data cleansing: cleaning data from noise or outliers or data with missing value 3) Data transformation: performs certain transformations so that data sets are ready for processing or can produce better analysis. 2.2.

Classification Classification means grouping objects based on existing groups. This classification requires training data that has been labeled as a group or class [12]. collecting the same object or entity and separating objects or unequal entities. In general it IJISCS | 3 can be said that classification is the process of calculating data existing or also called training data with new data or testing data. This process will generate possibilities in testing data [13]. 2.3. K-Nearest Neighbor K-Nearest Neighbor is a learning based algorithm where data sets training is stored, so the classification for new record that are not classified are obtained by comparing in to the record that is

most similar to the training set.

K-NN algorithm besides being used for classification, it also be used for estimation and prediction [7]. The steps K-NN algorithm: 1) Determine the parameter k (number of closest neighbors) 2) Calculate the distance (similarity) between all training records and new objects 3) Sorting data based on distance value from the smallest to largest value 4) Retrieving data of a number of k values 5) Determining the frequency labels most often among k training records closest with object. 2.4.

Numerical Attribute Similarity In the numeric attribute there is a calculation of the distance (distance between two objects) that can be done with using Euclidean distance, Manhattan distance and Minkowski distance calculation. In this research, the written uses Euclidean distance to calculate the distance between two object with nominal attributes. The neighbor proximity or distance is calculated based on Euclidean distance with equation 1 [14]. Information: distance between data x and data y attribute between to k from the test data (x), with k attribute value to k from the training data (y), with k After the distance or dissimilarity (d) is calculate then it is converted into similarity (s) with an interval between 0 and 1 1 (s [0, 1]) with equation 2. 2.5. Confusion Matrix Confusion Matrix is a table to evaluate the performance of the identification model.

Confusion Matrix shows the result of identification between the amount of correct prediction data and the number of incorrect predictive data compared to the facts produced. Table 1 shows the Confusion Matrix [15]. Table 1. Confusion Matrix Actual Prediction Negative Positive Negative a b Positive c d There are several terms based on Table 1. a) True Positive (TP) is positive data correctly indicated on the model. Calculation TP values can be calculated using equation 3. ???? = ?? ?? + ?? (?) IJISCS | 4 b) False Positive (FP) is positive data incorrectly indicated on the model. Calculation FP values can be calculated using equation 4. c) True Negative (TN) is negative data that is correctly indicated in the model.

Calculation TN value can be calculated using equation 5. ???? = ?? ?? + ?? (?) d) False Negative (FN) is negative data that is incorrectly indicated in the model. Calculation FN values can be calculated using equation 6. 2.6. Measurement Accuracy Measurement of accuracy is a step to prove the level of performance of an algorithm dataset used. In this research, confusion matrix is used as a performance measurement tool classification algorithm. Confusion matrix is a calculation that compares datasets with the results of the classification in accordance with the actual data with the total amount of data. The final result of this matrix is the level of accuracy with units of percent (%).

This level of accuracy will be used later the researchers' reference to the performance of

the classification algorithm. Confusion matrix contains information comparison of classification labels with actual labels. From table 1 you can calculate the level of accuracy from an algorithm model using equation 7 [16]. ???? + ???? (?) Information: a: the classification result is positive with the class actually positive b: negative classification results with positive actual class c: the result of the classification is positive with the class actually negative d: negative classification results with the actual class positive 2.7. Naïve Bayes Classifier The Naïve Bayes algorithm can be used for binary and multi-class classification problems.

Naïve Bayes is also a classification that represents each object class based on probabilistic conclusions or recapitulation and finds the most likely class that is suitable for each object whose class will be determined from existing test objects based on attributes or variables whose values have been known [17]. The use of NB for classification is considered important for several reasons, such as: a) It is very easy to build because it does not require a complicated iterative parameter estimation scheme and this method can be directly implemented into a very large amount of data. b) Easy to interpret so that users who are less skilled in classification techniques can easily understand the final result obtained.

Calculation of Probability and Classifier of Test Data: - After the data is divided into training data and test data, the standard deviation and mean will be calculated for each target parameter class (Diagnosis) for each attribute. - After the standard deviation and mean per each target parameter class (Diagnosis) per each attribute, will be used for the classification for test data 1. IJISCS | 5 - Naïve Bayes classification calculates the probability of the value of the diagnosis parameter based on the value of other parameters.

Calculation of probability using the Gaussian Naïve Bayes formula: ?? ^ (?? ?? | ?? ??) = 1 v 2 ???? ???? ?????? (- (?? ?? - ?? ????) 2) 2 ?? ???? 2) (8) ?? ???? : mean (average) of feature values ?? ?? of examples for which c = ?? ?? ?? ???? : standard deviation of feature values ?? ?? of examples for which c = ?? ?? After the probability of each attribute is calculated, it will be multiplied into the probability of the diagnosis value. 2.7. Machine Learning Machine learning is an area in artificial intelligence that deals with the development of techniques that can be programmed and learned from past data.

Pattern recognition, data mining and machine learning are often used to describe the same thing. This field intersects with the science of probability and statistics, sometimes optimization. Machine learning is an analytical tool in data mining [12]. 2.7. Python Python is a computer programming language, just like other programming languages, for example C, C ++, Java, PHP, and others. As a programming language, python

certainly has its own dialect, vocabulary or keywords, and rules that are clearly different from other programming languages [18]. Python excellence: 1) Python code is designed to be easy to read, learn, reuse and maintain.

In addition, python also supports object-oriented programming and visual programming. 2) Python can increase productivity and save programmers' morning time. To obtain the same result, Python code is much less than code written in other programming languages such as C, C ++, Java, PHP and others. 3) Supports multi platforms. 4) Through a specific mechanism, python code can be integrated with applications written in other programming languages. 3.0 METHODOLOGY The methodology used in this study is divided into several stages as shown in Figure 1 [19]. Fig. 1.

Research Methodology Flow Diagram IJISCS | 6 A) Literature Study and Problem Analysis In the initial stage, it is done by searching for and studying library materials from journals, digital libraries, papers, literature books, e-books, or scientific works that can support the writing process. This stage is carried out obtain information related to data mining, classification, machine learning, python, the K-NN algorithm and the Naïve Bayes algorithm. Information obtained from observing problems related to the factors needed to be used in this study and observing studies related to data classification using the K-NN and the Naïve Bayes method.

B) Data Collection The next stage is to prepare training and testing data taken from the Breast Cancer Wisconsin (Diagnostic) UCI (University of California, Irvine) Machine Learning Repository dataset. The system designed is a system used to classify existing breast cancer data into several classes, namely benign and malignant classes. The class division is used based on the value of each patient, namely radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimensions.

C) Implementation and Testing Based on the data that has been obtained and also various references that are complete, the following steps are implementation and testing. The results of the system design are outlined in the form of program implementation which results in writing program code to obtain test data results. The test carried out is testing the classification accuracy generated by the system using the K-NN algorithm and the Naïve Bayes algorithm. Accuracy is measured using the k-fold cross validation method 4.0 RESULANTS AND DISCUSSION 4.1.

Cancer Data Compilation Process The data used as training data and test data are data about breast cancer. There are 455 training data consisting of 284 data on cases of

benign cancer, 171 data on cases of malignant cancer. Based on the data obtained, the attributes used to classify are radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimensions. The data stored for each attribute is the measurement average (mean), standard error of measurement (se), and the minimum value (worst). 4.2. Input Data and Output Data Input data is data that will be used as input to the system.

This input data will then be processed using the K-NN and Naïve Bayes classification methods to determine the class of patients. The data used include: 1. Radius is the distance from the edge of the cancer to the center of the cancer. The radius data collected are the average radius measurement (radius_mean), standard error measurement radius (radius_se), and minimum value of radius (radius_worst). 2. Texture is the standard deviation of the grayscale value of the cancer x-ray results. The texture data collected were the average texture measurement (texture_mean), the standard error for measuring the texture (texture_se), and the minimum value for the texture (texture _sean). 3. Perimeter is a measure of the size of the cancer core.

The perimeter data collected were the average perimeter (perimeter_mean) measurement, the perimeter measurement error standard (perimeter_se), and the perimeter minimum value (perimeter_worst). 4. Area is a measurement of the surface area of the cancer. The area data collected are the area measurement average (area_mean), the area measurement standard error (area_se), and the minimum area value (area_worst). 5. Smoothness is the average local variation in the length of the cancer radius. Smoothness data collected are the average smoothness measurement (smoothness_mean), standard error of smoothness measurement (smoothness_se), and the minimum value of smoothness (smoothness_sean). 6.

Compactness is the average ratio of the volume and surface area of the cancer. The compactness data collected were the average compactness measurement IJISCS | 7 (compactness_mean), the compactness measurement standard error (compactness_se), and the minimum compactness value (compactness_worst). 7. Concavity is the average level of concavity of the cancer contour. The concavity data collected were the average concavity measurement (concavity_mean), the concavity measurement standard error (concavity_se), and the minimum concavity value (concavity_worst).

8. Concave points is the average number of the sunken portion of the cancer contour. The data for concave points collected were the average measurement of concave points (concave points_mean), standard error of measurement of concave points (concave points_se), and the minimum value of concave points (concave points_worst). 9. Symmetry is the level of symmetry of the cancer. The data on the symmetry that is

collected are the average symmetry_mean measurement, the symmetry_se measurement standard error, and the symmetry_worst minimum value. 10.

Fractal_dimension is [No explanation from dataset]. The fractal_dimension data collected were the average fractal dimension (fractal_dimension_mean) measurement, the fractal dimension (fractal_dimension_se) measurement standard, and the minimum fractal dimension (fractal_dimension_worst) value. Output Data Output data is output data processed from the system based on input data. The cancer grade states the result of the classification system against the patient. The patient class is divided into 2 classes, namely benign and malignant. 4.2.

Python Implementation The data set that has been collected is then implemented in programming using python. The data set in csv format is then imported into the Jupiter Framework to be executed until it gets the classification results. The next step is to import the libraries needed in the classification process for the result information from python. The import libraries used are numpy and pandas. Import library and dataset call, shown in Figure 1. Figure 1. Import Library Then, the next steps are as follows: - Calls the SKLearn library for train_test_split, K Fold, Cross_Validation, K-NN Classifier, Confusion matrix, Gaussian Naïve bayes.

- Displays CSV data into tables - Divide the data set into 80 percent train dataset and 20 percent test data - Print a score of the results of training and testing - Create confusion matrix and print graphs. The source code process can be seen in Figure 2. IJISCS | 8 Figure 2. K-NN Implementation In the same way, here is an implementation of Naïve Bayes. Figure 3. Naïve Bayes Implementation The results of the implementation of each method and confusion matrix are as follows: Figure 4.

The Results of the K-NN classification and Confusion Matrix IJISCS | 9 Based on the calculation of accuracy using the k-fold cross validation for k = 10, the K-NN algorithm accuracy result is 0.97%. Based on the confusion matrix, it can be seen that the system is able to correctly classify each type of cancer, for as many as 282 benign cancers and as many as 159 malignant cancers. Figure 5. Classification Results and Confusion Matrix Naïve Bayes Based on the calculation of the accuracy using the k-fold cross validation for k = 10, the accuracy of the Naïve Bayes algorithm is 0.94%.

Based on the confusion matrix, it can be seen that the system is able to correctly classify each type of cancer, for 276 benign and 152 malignant cancers. 5.0 CONCLUSION The test results of the system using the K-NN classifier were able to correctly classify 441 data while the Naïve Bayes classifier was able to correctly classify 428 data. The Naïve Bayes classifier gets a lower accuracy value than the K-NN classifier, this is due to NB's

performance which allows for correlation and / or irrelevance of its attributes.

If two or more attributes have very strong correlation, they receive too much weight in the final decision as to which class an object will occupy. This results in decreased accuracy in the domain with the correlation of its attributes. Whereas the K-NN classifier is based on the similarity possessed by objects based on the distance between the object to be determined and the object that already exists. REFERENCES [1] D. B. Abrams et al. Encycl. Behav. Med., pp. 79 81, 2013. [2] 2015 Int. Conf. Emerg. Res. Electron. Comput. Sci. Technol. (pp. 108 113), 2015. [3] B. Aisyah and Y. Sulistyo, J. Tek. Elektro, vol. 8, no. 2, pp. 43 46, 2016. [4] Proc. - ICCES 2012 2012 Int. Conf. Comput. Eng. Syst., no. November, pp.

180 185, 2012. [5] Simetris J. Tek. Mesin, Elektro dan Ilmu Komput., vol. 7, no. 1, p. 283, 2016. [6] I. H. Witten, E. Frank, and M. a Hall, Data Mining: Practical Machine Learning Tools and Techniques, 2011. IJISCS | 10 [7] D.r "scoverig Itrodti taMiig," Discov. Knowl. Data, 2005. [8] et al. ed classification of renal calculi: Comput. Biol. Med., vol. 112, no. January, 2019. [9] IC-Tech, vol. XII, no. 2, 2017. [10] IOP Conf. Ser. Mater. Sci. Eng., vol. 546, no. 5, 2019. [11] Expression Cancer RNA- . [12] B. Santoso and A. Umam, Data Mining dan Big Data Analytics, 2nd ed. Yogyakarta: Penebar Media Pustaka, 2018. [13] B. Sulistyo, Pengantar Ilmu Perpustakaan. Jakarta: PT. Gramedia Pustaka Utama, Jakarta, 1991.

[14] 28, 2012. [15] Digit. Signal Process. A Rev. J., vol. 17, no. 4, pp. 694 701, 2007. [16] F. Gorunescu, Data Mining: Concept, Model and Techniques. Heidelberg, Berlin: Springer, 2011. [17] M. Arhami and M. Nasir, DATA MINING Algoritma dan Implementasi, I. Yogyakarta: ANDI. [18] B. Raharjo, Python Untuk Aplikasi Desktop dan Web, Revisi. Bandung: Informatika Bandung, 2019. [19] -Nearest Neighbor Algorithm on Classification of Disk Indones. J. Inf. Syst., vol. 2, no. 1, p. 57, 2019.

INTERNET SOURCES:

3% - https://ojs.stmikpringsewu.ac.id/index.php/ijiscs/article/download/953/pdf
<1% - https://pubmed.ncbi.nlm.nih.gov/32030658/
<1% - https://www.sciencedirect.com/topics/medicine-and-dentistry/breast-cancer
<1% - https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html
<1% - https://www.coursehero.com/file/90562816/947-2124-1-PBpdf/
<1% - https://www.wgu.edu/blog/what-ai-technology-how-used2003.html
<1% - http://etd.repository.ugm.ac.id/penelitian/detail/131221

<1% -
https://www.researchgate.net/publication/326866871_Classification_algorithms_in_Data_Mining
<1% - https://byjus.com/commerce/meaning-and-objectives-of-classification-of-data/
<1% -
https://medium.com/analytics-vidhya/machine-learning-knn-algorithm-4130f799d697
<1% - https://www.ibm.com/topics/knn
<1% - https://www.proglobalbusinesssolutions.com/what-is-data-mining/
<1% - https://www.aionlinecourse.com/tutorial/machine-learning/k-nearest-neighbor
<1% -
https://towardsdatascience.com/knn-algorithm-what-when-why-how-41405c16c36f
<1% - https://www.exceldemy.com/excel-calculate-distance-between-two-coordinates/
<1% - https://en.wikipedia.org/wiki/Confusion_matrix
<1% - https://machinelearningmastery.com/confusion-matrix-machine-learning/
<1% -
https://medium.com/swlh/artificial-intelligence-machine-learning-and-deep-learning-whats-the-real-difference-94fe7e528097
<1% - https://www.softwaretestinghelp.com/data-mining-vs-machine-learning-vs-ai/
<1% -
https://archive.ics.uci.edu/ml/datasets/Breast%20Cancer%20Wisconsin%20(Diagnostic)
<1% -
https://text-id.123dok.com/document/y4er510q-keywords-radial-basis-probabilistic-neural-network-classification-breast-cancer-abstrak-penerapan-metode-rbpnn-untuk-klasifikasi-kanker-payudara.html
<1% -
https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_training_test_data.htm
<1% - http://klik.ulm.ac.id/index.php/klik/article/view/51
<1% - https://www.statology.org/standard-error-of-measurement/
<1% - https://www.statology.org/standard-deviation-vs-standard-error/
<1% - https://www.w3schools.com/python/python_ml_train_test.asp
<1% - https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning
1% - http://www.ojs.stmikpringsewu.ac.id/index.php/ijiscs/article/view/953
<1% -
https://www.semanticscholar.org/paper/Data-mining-dan-big-data-analytics%2C-ed.-2-Santosa-Umam/d36bdef2a5c478be87e547348d47e9438110a923
<1% - https://ojs.trigunadharma.ac.id/index.php/jsk/issue/view/51