Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2022. Vol. 1. No. 3

Knearst Algorithm Analysis – Neighbor Breast Cancer Prediction Coimbra

I Gusti Prahmana[1], Kristina Annatasia Br Sitepu[2]*

[1,2]STMIK Kaputama

Igustiprahmana4@gmail.com[1]; kannatasia88@gmail.com[2]*

Abstract

A process to explain the results of the KNN algorithm analysis with the prediction of Breast Cancer Coimbra disease (Breast Cancer). The prediction output of the KNN algorithm will be added with the Simple Linear Regression algorithm modeling to measure the predictive data through a straight line as an illustration of the correlation relationship between 2 or more variables. Linear regression prediction is used as a technique for the relationship between variables in the prediction process of the Breast Cancer Coimbra data set (Breast Cancer). for the value of K in analyzing the KNN algorithm, take the nearest neighbor with the ranking results with K = 5 nearest neighbors which are taken in the KNN calculation. Which is where the output of the KNN algorithm classification will be analyzed with the Simple Linear Regression algorithm with Dependent (Cause) and Independent (effect) variables. The test results determine that the patient has breast cancer and the number of predictions based on age with glucose means that the patient is predicted to have breast cancer. analyze the KNN algorithm with Simple Liner Regression modeling with Python programming language.

Keywords: *K – Nearest Neighbor, Breast Cancer Coimbra*

1. Introduction

Data mining is a process that uses mathematical, statistical, artificial intelligence in machine learning to identify highly utilized information and knowledge that is used a variety of databases [1], [2].
K-Nearest Neighbors or commonly abbreviated as KNN is a method nonparametric to classify data in an unknown class and select the closest k data located in a data. In general, k is set as an odd number to avoid appearing the same number of distances in the classification process [3]. This method provides a more flexible approach with an explicit form for f(k). This method is often more complex to understand and interpret. For multiple observations on predictor data, the parametric method works better. Analyze to classify the most popular in pattern recognition.

In this study, analyzing the KNN algorithm with the prediction of Breast Cancer Coimbra (Breast Cancer) the prediction output of the KNN algorithm will be added to the Simple Linear Regression algorithm modeling to estimate predictions on the data in a straight line as things that have a relationship between 2 different variables. Linear regression was applied to the technique on related variables when predicting the Breast Cancer Coimbra data set. The analysis carried out with the KNN algorithm is by modeling Simple Linear Regression on a number of data sets. The source of the data sets is the UC Irvine Machine learning Repository (UCI Machine learning Repository) which has different data sets (instances) and the number of attributes. Accuracy measurement results in the prediction of Breast Cancer Coimbra disease (Breast Cancer).

Some definitions, data mining to process a value in doing the addition of a large database of unknown scientific knowledge systemized. A decision result that is processed by data mining to make decisions in the future that develop for the future [4],[5] . Data mining can be called science in the form of a database, is a grouping of data, which uses in the form of patterns or dataset networks that very large. Issued in data mining as a decision-making system [6], [7]. In analyzing a data mining in a large amount of data, has the aim of obtaining and retrieving information that is substantiated in the form of accurate. science for people who involve work in take a knowledge system to solve the problem. Process collecting data is analyzed in the form of knowledge of decision-making short.[7], [5].

Classification Method is a process that illustrates that aims to be used to predict in class object whose class label is not known. Classification is part of data mining, where data mining is to explain the discovery of knowledge in data [8]. K-Nearest Neighbor is a way and regulation but nonparametric techniques are very effective on the classification pattern, but the classification performance depending on the value of k [9]. K is used in every class, which can cause high local sensitivity with k value. If k is too small, the classification information useful may not be sufficient, while large values of k can easily causes the outlier to be included in the k nearest neighbors of the true class [10], [9].

Using the Coimbra Breast Cancer dataset, totaling 116 data with 9 numerical attributes obtained from the UCI Machine Learning Repository. Analyzing the KNN algorithm by calculating the proximity of the distance to the data and the final process of classification with KNN Knowing the comparison and improving the evaluation results between methods calculation of the proximity of the nearest neighbor so that it can be provide input in science for further research in the development of science in solving algorithms with addition of other models in the KNN algorithm.

2. Research Method

The research methods used are:

2.1 Literature study

Regarding matters related to the K-Nearst Neighbor algorithm and the Simple Linear Regression algorithm from various books, journals, articles and several other references.

2.2 Research analysis

Analyzing the K-Nearst Neighbor algorithm, the predicted output results to be analyzed with the Simple Linear Regression algorithm modeling can affect the variables as a predictive quality test, so you must use relevant variables in testing the Dependent and Independent variables that can affect the predicted output results.

The first workflow is to determine the initialization of Distance K, then after determining the initialization of Distance K, the next step is to calculate the distance between the sample data and the testing data. After that determine the final value of the prediction of the majority of the KNN Algorithm. The block diagrams are made below:

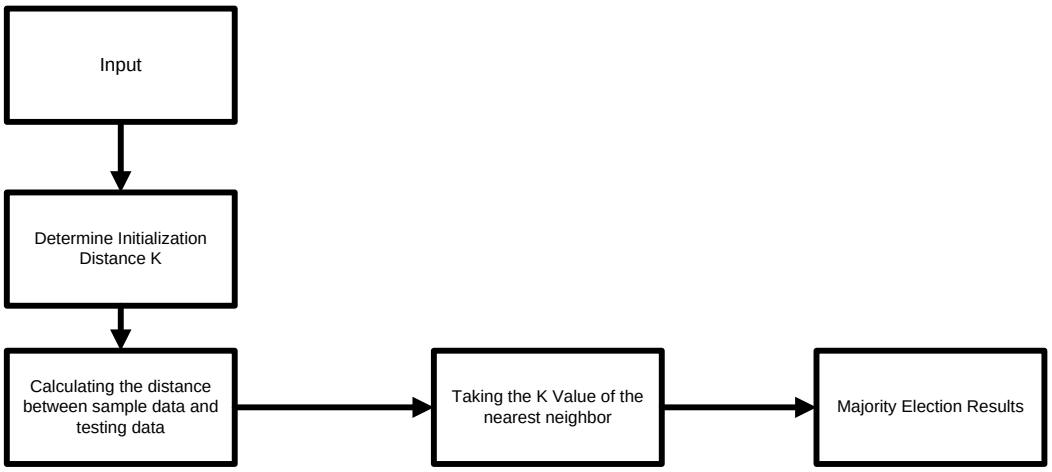

Figure 2.1. Research Workflow Block Diagram

3. Results And Discussion

3.1 Research Data

In this study using 9 attributes which then 9 attributes must be analyzed to produce a predictive result for Breast Cancer Coimbra disease (Breast Cancer) 116 data sets. In this study data that could be collected in routine blood analyzes – in particular, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP-1, Age and Body Mass Index (BMI) – might be used to predict the presence of breast cancer. Given training data in the form of 9 attributes with a classification of 1 (patients without breast cancer) and 2 (patients with breast cancer) to classify a data whether it is classified as 1 or 2, the following are the data:

1. Data Set Breast Cancer Coimbra (Breast Cancer)
   The table of data sets used in the calculation of the algorithm analysis:

Table 3.1 Data Set Breast Cancer Coimbra (Breast Cancer)

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
48	23,5	70	2,707	0,46740	8,8071	9,7024	7,99585	417,114	**1**
83	20,690	92	3,115	0,706893	8,8438	5,429285	4,06405	468,786	**1**

68	35,56	131	8,15	2,63353	17,87	11,9	4,19	198,4	**2**
75	30,48	152	7,01	2,62828	50,53	10,06	11,73	99,45	**2**
72	25,59	82	2,82	0,570392	24,96	33,75	3,27	392,46	**2**
...	...	...	...	...	...	...	...	...	**...**
...	...	...	...	...	...	...	...	...	**...**
86	27,18	138	19,91	6,777364	90,28	14,11	4,35	90,09	**2**

3.2. Calculation of the KNN Algorithm

Provide new data with a new classification, namely:

Table 3.2 New Data Classification

	X1	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**	**X9**
No	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
117	56	26,8	98	15,45	3,777765	87,99	12,78	3,78	78,899

3.2.1 Finding K with parameter (Nearest Neighbor)

Table 3.2 Determining Parameter K (number of nearest neighbours)

$$dis(t_i, t_j) = \sqrt{\sum_{h=1}^{k} (t_{ih} - t_{jh})^2}$$

Euclidean Distance **(56,26.8,98,15.45,3.777765,87.99,12.78,3,78.78,899)**

$$\sqrt{\begin{matrix}(48-56)^2+(23{,}5-36{,}8)^2+(70-98)^2+(2{,}707-15{,}45)^2 \\ +(0{,}467408667-3.777765)^2+(8{,}8071-87.99)^2 \\ +(9{,}7024-12{,}78)^2+(7{,}99585-3{,}78)^2 \\ +(417{,}114-78{,}899)^2\end{matrix}} = 348{,}88$$

3.2.2 Determine the ranking of the nearest neighbors

Sort distance Sorting those in the group, only the smallest Ecludien is sorted second by ascending method.

Table 3.3 Sorting the distance to the nearest neighbour

Euclidean Distance	**Urutan Jarak**	**Apakah Termasuk K-NN**	**Classification**
51,59	1	Ya (K < 3)	2
72,52	2	Ya (K < 3)	2
78,47	3	Ya (K = 3)	1
143,4	4	Tidak (K > 3)	2
148,64	5	Tidak (K > 3)	2

Determine the nearest K<=3 value, so row one includes classification 1 and the remainder is 2.

Table 3.4 Determining the ranking of the nearest neighbors

Euclidean Distance	**Urutan Jarak**	**Apakah Termasuk K-NN**	**Classification**
51,59	1	Ya (K < 3)	2
72,52	2	Ya (K < 3)	2

78,47	3	Ya (K = 3)	1
143,4	4	Tidak (K > 3)	2
148,64	5	Tidak (K > 3)	2

3.2.3 Determine the majority of nearest neighbors .

The closest K assessment as a new data prediction assessment. The data in rows one, two and three we have 2 classifications of patient categories without breast cancer and 2 categories of patients with breast cancer. on the addition of the majority (2 > 1), new data will be concluded:

X1	X2	X3	X4	X5	X6	X7	X8	X9
Age	**BMI**	**Glucose**	**Insulin**	**HOMA**	**Leptin**	**Adiponectin**	**Resistin**	**MCP.1**
56	26,8	98	15,45	3,777765	87,99	12,78	3,78	78,899

Then we will predict the patient as a patient who is included in category 2 with breast cancer.

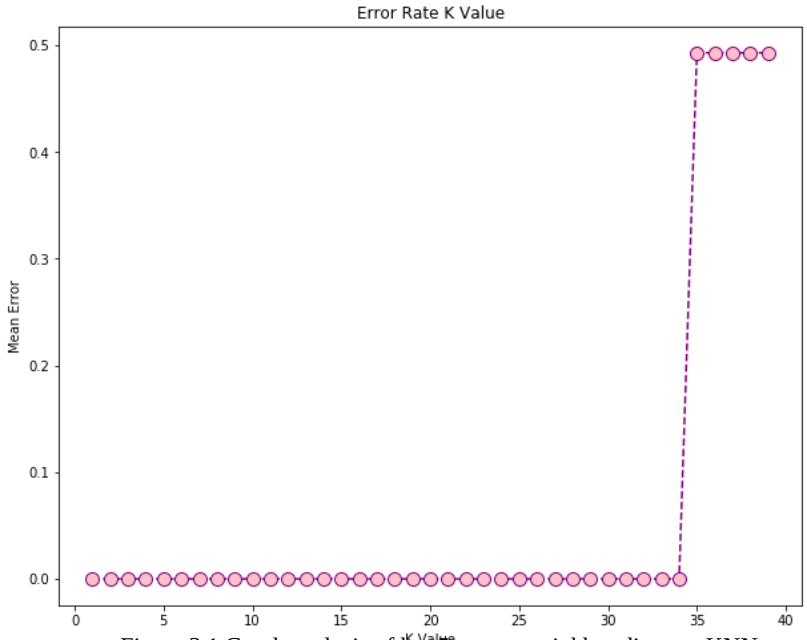

Figure 3.1 Graph analysis of k = 5 nearest neighbor distance KNN

4. Conclusions And Suggestions

With this research, which has succeeded in analyzing the KNN algorithm calculations, the results of the KNN will be analyzed using the Simple Linear Regression algorithm. for the value of K in conducting the KNN analysis, take the nearest neighbor with the ranking results with K = 5, the nearest neighbor takes in the calculation of KNN. The output results from the classification of the KNN algorithm will be analyzed with a Simple Linear Regression algorithm with Dependent (cause) and Independent (consequences) variables. The results of testing the test data accuracy value are 97%. in using the analysis of the Simple Linear Regression algorithm in determining patients with breast cancer. and the number of predictions based on age with glucose then the patient is predicted to develop breast cancer. analyze the KNN algorithm with Simple Liner Regression modeling with Python programming language.

References

[1] M. Dunhan, “Data Mining: Introductory and Advanced Topics. Prentice Hall,” *Engineering*, 2003.
[2] J. Han and M. Kamber, “Data Mining : Concepts and Techniques (2nd edition) Bibliographic Notes for Chapter 11 Applications and Trends in Data Mining,” *SIGKDD Explor.*, 2006.
[3] Eko Prasetyo, *Data Mining : Konsep Dan Aplikasi Menggunakan Matlab*. 2013.
[4] F. Yunita, “Penerapan Data Mining Menggunkan Algoritma K-Means Clustring Pada Penerimaan Mahasiswa Baru (STUDI KASUS : UNIVERSITAS ISLAM INDRAGIRI),” *Sistemasi*, vol. 7, no. 3, 2018.
[5] A. M. H. Pardede *et al.*, “Implementation of Data Mining to Classify the Consumer’s Complaints of Electricity Usage Based on Consumer’s Locations Using Clustering Method,” 2019, doi: 10.1088/1742-6596/1363/1/012079.

[6] S. R. Kumaran, M. S. Othman, and L. M. Yusuf, "Data mining approaches in business intelligence: Postgraduate data analytic," *J. Teknol.*, vol. 78, no. 8–2, 2016, doi: 10.11113/jt.v78.9544.
[7] C. Vercellis, *Business Intelligence: Data Mining and Optimization for Decision Making*. 2009.
[8] A. Danades, D. Pratama, D. Anggraini, and D. Anggriani, "Comparison of accuracy level K-Nearest Neighbor algorithm and support vector machine algorithm in classification water quality status," 2017, doi: 10.1109/FIT.2016.7857553.
[9] Y. Wang, Z. Pan, and Y. Pan, "A Training Data Set Cleaning Method by Classification Ability Ranking for the k -Nearest Neighbor Classifier," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 5, 2020, doi: 10.1109/TNNLS.2019.2920864.
[10] J. Gou, Y. Zhan, Y. Rao, X. Shen, X. Wang, and W. He, "Improved pseudo nearest neighbor classification," *Knowledge-Based Syst.*, vol. 70, 2014, doi: 10.1016/j.knosys.2014.07.020.