International Journal of Psychosocial Rehabilitation, Conference Special Issue
ISSN: 1475-7192

KNEARST - NEIGHBOR ALGORITHM ANALYSIS USING SIMPLE LINIER REGRESSION MODELING

*[1]I Gusti Prahmana, [2] Dr. Herman Mawengkang, [3] Dr. Muhammad Zarlis

ABSTRACT-- *A process to explain the results of the KNN algorithm analysis with prediction of Breast Cancer Coimbra (KNN algorithm) output results will be added with the modeling of the Simple Linear Regression algorithm to measure predictive data through a straight line as an illustration of the correlation between 2 or more variables. Linear regression prediction, is used as a technique for the relationship of variables in the prediction process of the Breast Cancer Coimbra data set. for the K value in analyzing the KNN algorithm take the nearest neighbor with the results of the ranking with K = 5 the nearest neighbor taken in the KNN calculation. Which is where the results of the KNN algorithm classification results will be analyzed by the Simple Linear Regression algorithm with Dependent (Cause) and Independent (effect) variables. The test results are 97% accurate. It is that by using algorithmic analysis by modeling Simple Linear Regression to determine patients affected by breast cancer and the number of predictions based on age with glucose, 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, Simple Linear Regression, Breast Cancer Coimbra*

I. INTRODUCTION

In this study analyzing the KNN algorithm with prediction of Breast Cancer Coimbra (KNN algorithm) output results will be added by modeling the Simple Linear Regression algorithm to estimate predictions on data in a straight line as having a relationship between 2 different variables. Linear regression is applied to techniques on related variables when the prediction process of the Breast Cancer Coimbra data set. By using the Simple Linear Regression modeling algorithm can influence variables for predictive quality testing, so the need for relevant variables to test Dependent and Independent variables can provide predictive output results from the KNN method.

The analysis performed with the KNN algorithm is by modeling Simple Linear Regression on a number of data sets. The source data set is the UC Irvine Machine learning Repository (UCI Machine learning Repository) has a different data set (instance) and number of attributes. Accuracy measurements yield prediction results for Breast Cancer Coimbra.

[1*]*Megister Computer Science Program, Faculty of Computer Science,University of North Sumatra, Medan, Indonesia, igustiprahmana27@gmail.com.*

[2] *Prof., Faculty of Mathematics Science,University of North Sumatra, Medan, Indonesia*.

[3]*Prof., Faculty of Computer Science,University of North Sumatra, Medan, Indonesia*.

Received: 02 Jan 2020 | Revised: 12 Feb 2020 | Accepted: 17 Mar 2020

II. RESEARCH METHOD

The research methods used are:

2.1 Literature study

Regarding matters relating 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 prediction output results that will be analyzed by modeling the Simple Linear Regression algorithm can affect variables as predictive quality test then, must use relevant variables in testing Dependent and Independent variables can influence the results of predictive output.

III. RESULTS AND DISCUSSION

3.1 Research Data

In this study using 9 Attributes which then 9 attributes must be analyzed to produce the prediction results of Breast Cancer Coimbra 116 data set. In this study data that can be collected in routine blood analysis - specifically, glucose, insulin, HOMA, leptin, adiponectin, resistin, MCP-1, age and body mass index (BMI) - may be used to predict the presence of breast cancer.

Given training data in the form of 9 attributes with classification 1 (patients not affected by breast cancer) and 2 (patients affected by breast cancer) to classify a data whether classified as 1 or 2, the following is the data:
The data set table used for algorithmic calculation calculations:

3.2 Calculation of the Algorithm KNN Algorithm

Provide new data with new classifications, namely:

Table 3.1: Classification of new data

	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

3.2.1Find K with parameters (Nearby Neighbors).

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

Determine the closest neighbor ranking.

Determine the closest value K <= 3 So row one includes classification 1 and the remainder 2.

Table 3.3: Determines the ranking of the nearest neighbors

Euclidean Distance	Order Of Distance	Is included K-NN	Classification
51,59	1	Yes (K < 3)	2
72,52	2	Yes (K < 3)	2
78,47	3	Yes (K = 3)	1
143,4	4	No (K > 3)	2
148,64	5	No (K > 3)	2

3.2.3Determine the majority of the nearest neighbors.

Nearest K evaluation as a new predictive data assessment. Data that has in line one, two and three we have 2 classification categories of patients not affected by breast cancer and 2 categories of patients affected by breast cancer. in the sum of the majority (2 > 1) new data will be concluded:

Then we will predict the paisen as a patient who falls into category 2 with breast cancer.

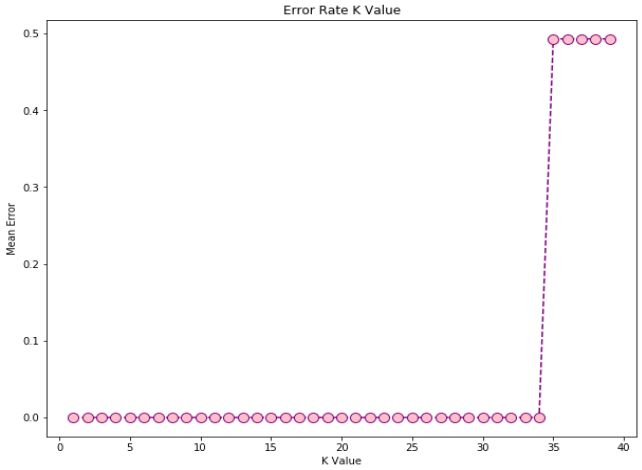

Figure 4.1: Graph of analysis k = 5 distance nearest KNN neighbors

3.3 Calculation of the Simple Linear Regression Algorithm

By using the Simple Linear Regression modeling algorithm can affect variables as predictive quality test, it must use the relevant variables to test the Age as Dependent and Independent variables can influence the predictive output of the KNN algorithm.

Identify the Cause and Predictor Variable Factors (Response)

Variable Cause Value X = Age

Variable Due to Y Value = Glucosa

3.3.1 Collection of data variables X and Y

The following is the data successfully collected by Uchi Mechine Learning Data Set about Breast Cancer Coimbra data

3.3.2 Calculates the values of X, Y and XY respectively

Table 3.5: Data calculation algorithm simple linear regression

X1	Y1	X2	Y2	XY
Age	**Glucose**			
48	70	$48^2 = 2304$	$70^2 = 4900$	2304 * 4900 = 3360
83	92	$83^2 = 6889$	$92^2 = 8464$	6889 * 8464 = 7636
82	91	$82^2 = 6724$	$91^2 = 8281$	6724 * 8281 = 7462
68	77	$68^2 = 4624$	$77^2 = 5929$	4624 * 5929 = 5236
86	92	$86^2 = 7396$	$92^2 = 8464$	7396 * 8464 = 7912
6703	**11442**	**413877**	**1177318**	**665123**

3.3.1 Calculates the values of a and b in linear regression using the formula:

3.3.3.1 Menghitung Konstanta (a) :

$$\frac{(\sum y)(\sum x^2)-(\sum x)(\sum xy)}{n(\sum x^2)-(\sum x)^2} \quad (1)$$

$$a = \frac{(11442)\,(413877) - (6703)\,(665123)}{117\,(413877) - (6703)^2}$$

$$a = 79{,}37$$

3.3.3.2 Calculating the Number of Regression Coefficients (b)

$$\frac{n(\sum xy)-(\sum x)(\sum y)}{n(\sum x^2)-(\sum x)^2} \quad (2)$$

$$b = \frac{117\,(665123) - (6703)\,(11442)}{117\,(413877) - (6703)^2}$$

$$b = 0{,}32$$

3.3.3.4 Creating a Regression Model

Y value = A value + BX value

Value of Y = 79.37 + 0.32X

3.3.4 Get predictions from cause and effect variables.

Results Predict the number of patients affected by breast cancer if the age of the patient is 40 years old as vasiable X.

Y Value = 79.37 + 0.32 (40)

Value of Y = 92.23

age of patient 40 years, it will be predicted there are 92.23 Total Glucose from 117 data sets that will be affected by breast cancer produced by patient

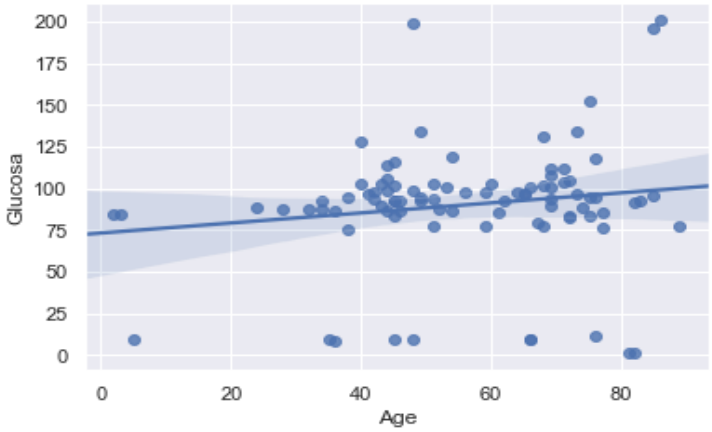

Figure 3.2: Graph Analysis Simple Linier Regression

IV. CONCLUSIONS AND SUGGESTIONS

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

REFERENCES

1. Okfalisa, Ratika, F. & Yelfi, V. 2018. The Comparison of Linear Regression Method and K-Nearest Neighbors in Scholarship Recipit. ISSN : 978-1-5386-5889-5 IEEE SNPD 2018, June 27-29, 2018, Busan, Korea.
2. Danades, A., Pratama, D, Anggraini, D, Anggriani, D. 2016. Comparison of Accuracy Level K of the nearest neighbor Algorithm and Support Vector Machine Algorithm in Classification Water Quality Status.International Conference on System Engineering and Technology, pp. 137-141.
3. The Pan, Yidi Zhibin Wang, Weiping Me A new k-harmonic nearest neighbor classifier based on the multi-local means. China Expert Systems With Applications (2016).
4. Pu, Y. , Peng, J. & Huang, L. 2015. An efficient KNN algorithm implemented on FPGA based heterogeneous computing system using OpenCL. IEEE. 978-1-4799-9969-9. IEEE DOI 10.1109/FCCM.2015.
5. Taneja, S., Gupita, C., & Goyal, K., Gureja, D. 2014. An Enhanced K-Nearest Neighbor Algorithm Using Information Gain and Clustering. IEEE. ISSN: 978-1-4799-4910-6. IEEE . DOI 10.1109/ACCT.2014.
6. Wu, C.H., Li, J.B, & Chang, T.Y. 2013. A Simple Linear Regression Analysis Assisting System. IEEE. ISSN: 978-0-7695-5111-1. IEEE DOI 10.1109/ICEBE.2013.
7. Mihaescu, M.C. 2011. Classification of Learners Using Linear Regression. IEE. ISSN: 978-83-60810-39-2. 2011