Global Scientific JOURNALS

# IDENTIFICATION OF BRACHYPELMA GENAR TARANTULA REPRODUCTION APPARATUS USING KNN AND SINGLE DECISION THRESHOLD METHOD

Apriandy Angdresey, Ivana Valentine Masala
Informatics Engineering Department
De La Salle Catholic University, Manado - Indonesia
Email: aangdresey@unikadelasalle.ac.id, imasala@unikadelasalle.ac.id

***Abstract*— In tarantula breeding, it takes a long time to wait for the tarantula to reach adulthood. If it have entered adulthood, we can distinguish the sex of the tarantula. In this study, a system was built which could predict the reproduction apparatus of the tarantula based on the parameters, a value that was determined according to the size of the tarantula. Based on several studies using the K-Nearest Neighbor (K-NN) which is an algorithm to classify these objects, also Single Decision Threshold method. We apply calculations based on learning data that are the closest distance to the object, and then will be proven using the rapidminer tool. The final result that will be obtained in the form of a prediction of the reproduction apparatus of the tarantula is the sex of male, female and unsex, with the accuracy rate is 96.35%.**

***Index Term*—Tarantula, Prediction; KNN; Single Decision Threshold;**

## I. INTRODUCTION

Tarantula is an eight-legged animal belonging to the species of spider in the Theraphosidae family and there are at least 883 species that have been identified and possibly more [1], the name of the tarantula comes from the Italian city of Taranto. Brachypelma genus is a sub-family of tarantulas included in the new-world category ¬ which has the main characteristics, namely in the defense that uses urticating hair and has a terrestial type (living on the surface of the ground), this genus was discovered in 1891 and is widespread throughout Central America, starting from Mexico, Costa Rica, Guatemala and Panama.

In the category of exotic animals, especially tarantulas, many hobbyists and breeders also have difficulty in determining the sex of tarantulas that still have a size of 1cm to 4cm through the lower abdomen or called the ventral and outer skin (exoskeleton) because in tarantulas with male sex have a lifetime between 3 years and 5 years and tarantulas with female sex have a life span of up to 20 years. Tarantulas that have female sex are more desirable to be maintained because of their longer life. Therefore it takes quite a long time to wait for the tarantula to develop to a minimum size of 5cm to identify through the ventral and exoskeleton.

Sexual dimorphism or systematic differences in the two sexes of the tarantula in the genus Brachypelma can be determined when entering adulthood ie the female tarantula tends to be larger and heavier than the male tarantula, therefore it is necessary to know how to distinguish the sexes in the tarantula that still have size under the size of 5cm.

K-Nearest Neighbor is the right method for classification. The way this method works is to find the closest distance from the data to be evaluated with the k-nearest neighbor data in the test data and then it will be represented using Single Decision Threshold (One Feature) with the features in the form of the sex of the tarantula.

## II. RELATED WORK

According to research in paper [2] to determine the feasibility of a car used the KNN algorithm. The process of determining the results of the KNN algorithm is started by entering the attribute values, namely: Car Prices, Maintenance Prices, Doors, Loads of People, Baggage Size, Security then determine the value of K = 11 in the classification of car feasibility by using 1728 training data and testing data as much as 1, then calculate the distance of each object's euclid against the training data provided as much as 25 data and after all have been calculated then sorted by the smallest euclid value then class y labels are collected for the classification of nearest neigborhood so that the resulting eligibility category is not good as much as 8, the feasibility Not good 2, Very good 1. Then in using data as much as 1728, obtained satisfactory accuracy results as much as 95.7755%.

This paper explain [3] the tarantula identification process with another method which is Linear Discriminant Analysis (LDA) method. Their purposes to identifying the tarantula's reproductive apparatus with time accuracy to show the fastest execution time result. The parameters are using which are age, leg span, weight, and lower abdomen of tarantula.

In paper [4] explains that a person's nutritional status can be determined through influential variables with calculations using one of the classification methods used in decision making and can be done by a computer, namely K-Nearest Neighbor (KNN) and then will be represented using a Single Decision Threshold (One Feature) and then the value of the system validity will be calculated. In this study it is assumed that input provided by end-users is done by sending an SMS containing variables used in the calculation, namely gender, height (cm), weight (kg), abdominal circumference (cm), pelvic circumference (cm) , and fat (%).

Another reference [5], used the classification method KNN to mining the web usage data. They present to find the right products information by time consuming task. The system provide some relevant information with data train to be on-line and real time condition. The KNN model shown the result classifier. Their approaches classification problem the objective of the system to get the model and put for each labels to map those models. The simplest model of KNN based on the distance and text recommendation model. Some formulas combined, and the simulation analysis process used by Matlab to get the result.

Determination of end-user nutritional status is distinguished according to sex and is done by involving a minimum of 25 data samples for each sex. In this study, nutritional status was divided into 3 classes, namely "Thin", "Normal", and "Obesity". The data that has been obtained from the SMS parsing is calculated by the KNN method to get the results in the form of end-user nutritional status decisions. From the results of data collection of a number of K values, namely 5 data, the nutritional status results are obtained: Normal = 4, Obesity = 1, After obtaining the results of nutritional status and compared to the amount, the decision is obtained that the end-user's nutritional status is "Normal" .

For paper [6] the results of system validity testing are represented using a single decision threshold (one feature) between child development data in training data and testing data with attributes of the test results using real / test data, Naive Bayes classifier results and results if appropriate or not using test data of 110 test data. Then the validity of the system is assessed by calculating the value of TP, TN, FP and FN after that to produce accuracy of system validation then the value of the system is calculated so that the accuracy rate of 83.1% is obtained.

## III. SYSTEM MODEL

In this study the system handles the determination of decisions based on the sex of the tarantula through the attributes of sexual dimorphism. The variables of sexual dimorphism are taken: age (based on month), legspan (length of leg span), leg length, weight (gram) and ventral width (front side of abdomen). In this study, the determination of the sex of tarantula can be divided into 3 types, namely "male", "female" and "unsex".

$$d(x,y) = \sqrt{\sum_{j=1}^{p} (a_i - a_j)^2 + (ls_i - ls_j)^2 + (ll_i - ll_j)^2 + (w_i - w_j)^2 + (vw_i - vw_j)^2}$$

Annotation,

$d(xi,xj)$ : Eucledean Distance

$a$: age (month)

$ls_i$: leg span

$ll_i$: leg lenght

$w_i$: weight (gr)

$vw_i$: ventral width (front side of the stomach)

$i_j$: 1,2,3,…n

Furthermore, the results of testing the validity of the system for the sexes of Male, Female and Unsex are represented using Single Decision Threshold (onefeature) with features in the form of tarantula sex, then it can be explained as follows:

| | | ***Fact*** | | |
|---|---|---|---|---|
| | | + | - | |
| ***Decision Model*** | + | ***TP*** | ***TN*** | ***100%*** |
| | - | ***FP*** | ***FN*** | ***100%*** |

Annotation,

TP: *True Positive*,

TN: *True Negative*,

FP: *False Positive*,

FN: *False Negative*,

Then the system validity is assessed by calculating the value of TP, TN, FP, and FN by using the results of the system validity testing using the bellow formula:

$$\text{System performance} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

From these formulas the resulting system performance values can produce a higher level of accuracy.

## IV. METHODS

This research conducted was a study of tarantulas with the genus Brachypelma. The study was conducted by taking data from the characteristics of sexual dimorphism in each tarantula that will be included in the sample data. Sampling of data was carried out by observing 85 tarantulas of the genus Brachypelma and taking data from the Indonesian Tarantula Keeper community of 27 data with a measurement value for each part of the body needed starting from age, leg length, leg length, weight, ventral found in the stomach the bottom list and identify the sex that has been identified that is marked "Male" or "Female" or that has not been identified "Unsex".

After all data has been collected, cleaning is done by eliminating tarantula data that has incomplete attributes or has a different genus. Data that has been cleaned and then entered in a folder that contains all data that has been cleared.

From the cleansed results the classification is done using the KNN method which is a method created to classify objects based on training data with the closest distance to the object. The first process used in the KNN is to use the K parameter and then calculate the distance between the data to be evaluated with all training then sorted the distance formed then the closest distance will be determined to the sequence K after that pair the corresponding class and look for the number of classes from the nearest neighbors and set the class as the data class to be evaluated.

Furthermore, the system validity is calculated using Single Decision Threshold (One Feature) which is assessed by calculating the value of TP, TN, FP and FN after that to produce system validation accuracy, the work value of the system is calculated so that a higher level of accuracy is obtained.

## V. EXPERIMENT

In this research, we used 85 data testing are obtained from direct observations on tarantulas, that have been maintained since December 2016 or approximately for 1 year. The data are stored in Microsoft Excel with *.csv* format, and we using rapidminer tools, then calculate the validity of the system using Single Decision Threshold for one feature.

| *a* | *ls* | *ll* | *w* | vw | jk |
|---|---|---|---|---|---|
| 12 | *9* | *4* | 4 | 0.6 | Female |
| 12 | *6* | *4* | 3 | 0.3 | Male |
| 10 | *3.5* | *3* | 2.2 | 0.3 | Unsex |
| 25 | *12* | *4.1* | 16.1 | 1 | Female |
| … | | | | | … |
| 12 | *8.5* | *4.5* | 4 | 0.6 | Female |
| 12 | *7* | *4.4* | 5 | 0.4 | Female |
| 12 | *7.5* | *4.1* | 3.1 | 0.42 | Male |
| …. | | | | | … |
| 10 | *4.5* | *2.5* | 1.4 | 0.3 | Unsex |
| 10 | *3.5* | *3* | 2.5 | 0.3 | Male |
| 10 | *3.6* | *3* | 2.51 | 0.3 | Male |
| … | | | | | … |
| 12 | *9* | *4* | 4.2 | 0.6 | Female |
| 12 | *5.9* | *4* | 3 | 0.4 | Male |
| ….. | ….. | ….. | ….. | ….. | ….. |

**Table 1.** Tarantula's Data Training

| *age* | *legspan* | *leglength* | *abdomen* | weight | ventral | sex |
|---|---|---|---|---|---|---|
| 11 | *8* | *4* | medium | 3 | 0.6 | ? |

**Table 2.** Tarantula's Data Testing

```
PerformanceVector:
accuracy: 83.33%
ConfusionMatrix:
```

| True: | female | Male | unsex |
|---|---|---|---|
| female: | 9 | 1 | 0 |
| male: | 2 | 10 | 1 |
| unsex: | 0 | 0 | 1 |

```
classification_error: 16.67%
confusionMatrix:
```

| True: | female | male | unsex |
|---|---|---|---|
| female: | 9 | 1 | 0 |
| male: | 2 | 10 | 1 |
| unsex: | 0 | 0 | 1 |

**Figure 1.** Vector of Performances by RapidMiner

In Figure 1 show the result of performance vector as we get by rapidminner. Where the level of accuracy is 83.33% with classification error 16.67%. Furthermore, in Figure 2 show the simulation of the tarantulas data consist of 3 kind categorize of reproductive apparatus score, and calculated by the simulation system to get the result. Result of the timing accuracy between two algorithms applied, and shown the accurate one.

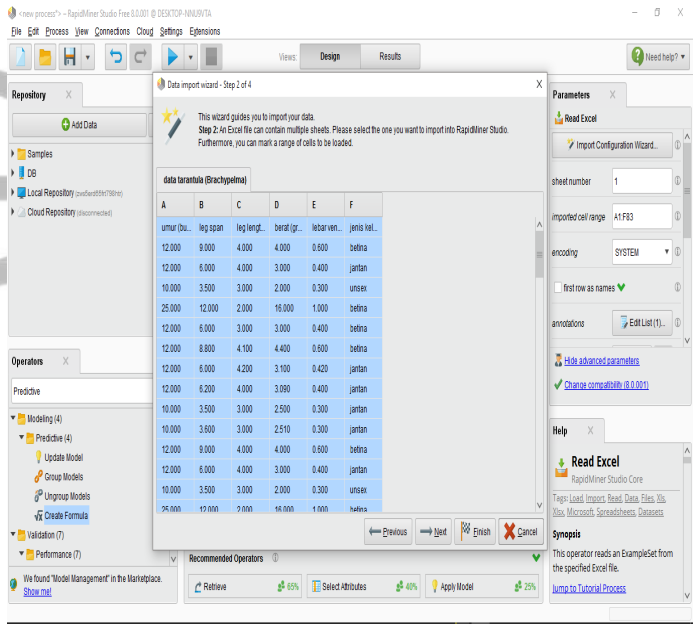

**Figure 2**. Simulation in RapidMiner

In Figure 2 shown the data of the parameters are chosen by the tarantula's data training according to the tarantula's species. If the data have been input in the system, it will continue to another step, in this case should be according to the simulation system procedures.

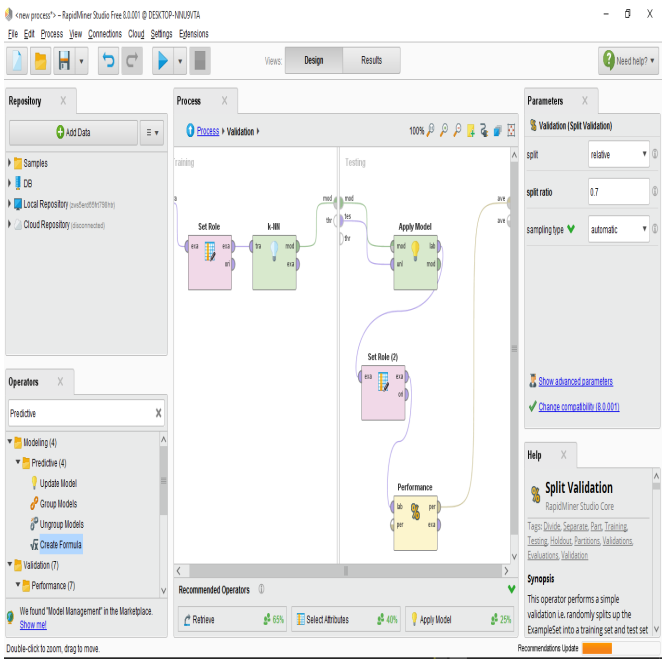

**Figure 3**. K-NN Algorithm Model and Tool Performance Model Applied.

In Figure 3, shown the two models applied into the simulation systems according to the procedures. The relation should be in the right way to get the right probability result. If the relation has a different links, the score result will be incorrect. The data has to be valid by the system and adjust the calculation to the parameters, next will show the performance.

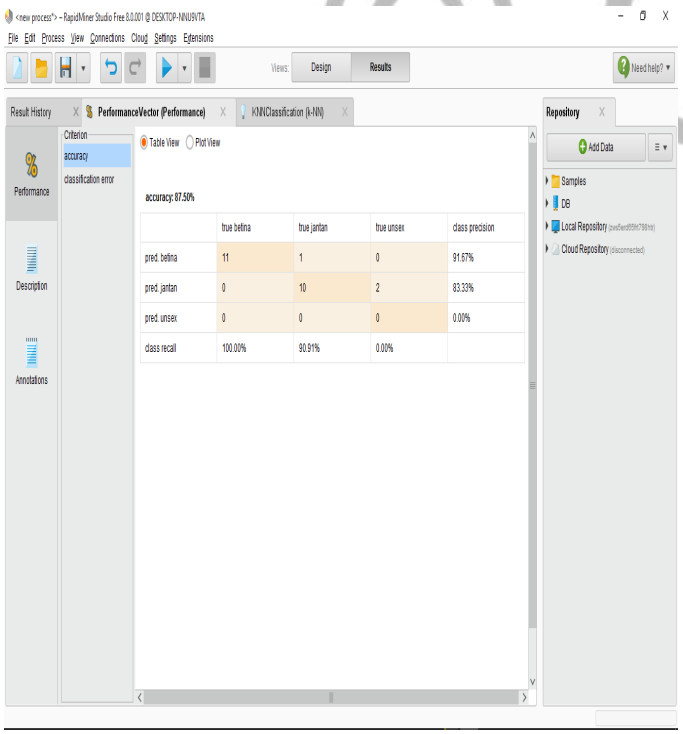

**Figure 4**. The Result of Accuracy

The results of testing of the system validity are represented using a single decision threshold (one feature) with features in the form of gender, then it can be explained as follows:

The following table provides a comparison of the tarantula's apparatus between reality and the system:

| | ***True* Female** | ***True* Male** | ***True Unsex*** |
|---|---|---|---|
| **Female Prediction** | *10* | *1* | *0* |
| **Male Prediction** | *1* | *10* | *1* |
| **Unsex Prediction** | *0* | *0* | *1* |

**Tabel 3.** Comparison of Tarantula's Reproductive Apparatus

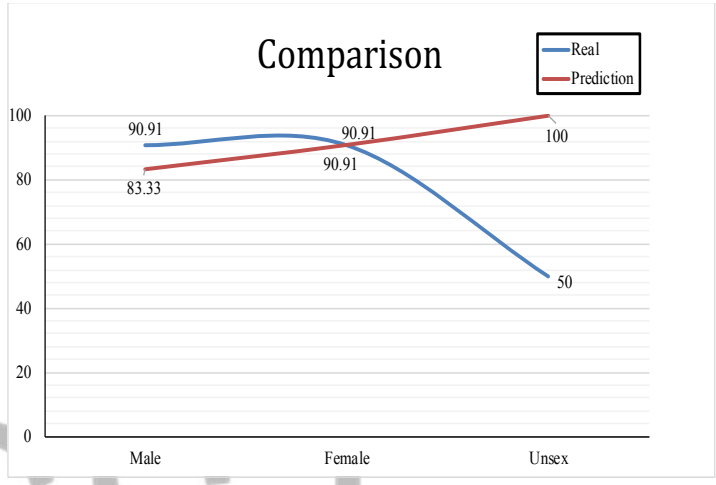


**Figure 5.** The Comparison Result

Reality graphs in the percentage value of true female, true male and true unsex. In male and female comparison appartus, a positive positive value of 90.91% and true positve value of the unsex were 50%. Prediction value produced in males is 90.91%, females 83.33% and unsex is 100%. Furthermore, calculating the value of system validity by calculating the value of TP, TN, FP, and FN in Table 2. So as to produce a system performance value of 96.35%. As shown on the graph, between the real an prediction percentages value of the female and male apparatus but the unsex species had some distinguish value till 50%.

## VI. CONCLUSIONS

This work concluded that the system can be used as a tool to determine the Gender of Tarantula using the K-NN method according to the physical condition parameters of the tarantula, using training data totaling 85 data with a total of k = 15 obtained values accuracy 83.33% and if the greater the amount of training data the system will become more accurate. Furthermore, the validity of the system using Single Decision Threshold (One Feature) is assessed by calculating the value of TP, TN, FP, and FN so as to produce an accuracy value of 96.35%.

For the future works, collect more data and can be developed into web-based applications or other platforms and can be combined with other methods so that accuracy can be more accurate in all conditions.

## VII. REFERENCES

[1] R. C. West, "Brachypelma of Mexico," *Journal of the British Tarantula Society,* pp. 108-119, 2005.

[2] A. Nouvel, "Klasifikasi Kendaraan Roda Empat Berbasis KNN," *Jurnal Bianglala Informatika ,* vol. III, no. 2, pp. 66-69, 2015.

[3] A. Angdresey, M. Wongkar, "Identification of The Reproductive Apparatus of Tarantula Genus Brachypelma Using Linear Discriminant Analysis Method", *in Internation Conference on Electrical Engineering and Computer Science,* 2018.

[4] S. Hermaduanti, N. Hermaduanti, "Sistem Pendukung Keputusan Berbasis SMS untuk Menentukan Status Gizi dengan Metode K-Nearest Neighbor", *in Seminar Nasional Aplikasi Teknologi Informasi*, 2008.

[5] D.A. Adeniyi, Z. Wei, Y. Yongquan, "Automated Web Usage Data Mining and Recommendation System Using K-Nearest Neighbor (KNN) Classification Method", *Applied Computing and Informatics*, Volume 12, pp.90-108, 2016.

[6] N. Mariana, "Penerapan Algorima k-NN (Nearest Neighbor) Untuk Deteksi Penyakit (Kanker Serviks)", *Dinamika Informatika,* Vol. VII, no. 1, pp. 26-34, 2015.