Journal of Knowledge Information Technology and Systems

ISSN 1975-7700

http://www.kkits.or.kr

Identification of Environmental Factors in Fruit Disease by Logistic Regression

Vasanth Ragu[1], Myeongbae Lee[1], Meonghun Lee[2], A B M Salman Rahman[1], Jangwoo Park[1], Yongyun Cho[1], Changsun Shin[*1]

[1]*Department of Information & Communication Engineering, Sunchon National University, Suncheon-si*
[2]*National Institute of Agricultural Sciences*

ABSTRACT

Plant diseases are one of the most significance issues in agriculture fields. The diseases destroy fruits and are decreased the production rate. They also decrease the economic growth in worldwide and increase the demand about preventing plant diseases. The major cause of plant diseases is climate changes, weather conditions and environmental factors. To prevent the diseases, we need to identify the plant diseases about corresponding environmental factors. In this study, we analysis the plum data and identify the plant diseases with corresponding environmental factors by using Logistic Regression model. The first process is to identify the number of diseases and its type, and then implement Logistic Regression to predict the diseases by using environmental factors as inputs. Finally, we compare the actual value of diseases with predicted value of diseases, and need to check the accuracy of diseases by using correlation method. We also find the environmental factors for the reasons to forms more diseases. In result and discussion, it would be clearly explained about the plum diseases and their environmental factors.

KEYWORDS: Plum data, Environmental factors, Logistic regression, Correlation method, Disease type.

ARTICLE INFO: Received 3 August 2018, Revised 15 September 2018, Accepted 12 October 2018.

---

[*] Corresponding author is with the Department of Information & Communication Engineering, Sunchon University, Suncheon-si, Republic of Korea, 57922.
E-mail address: csshin@sunchon.ac.kr

1. Introduction

Plants diseases are one of the most significance issues in agriculture fields. Plant diseases destroy any type of fruits and decrease the production in agriculture fields. It also decreases the economic growth and increases the demand in worldwide. The major part of cause plant diseases is climate changes, weather changes and environmental changes in worldwide. To be solve, we need some machine learning algorithm to find out which environmental factor to causes the plants diseases. So, in this study examine plum diseases and identify the corresponding environmental factors by using logistic regression and correlation method.

In plum data, consist of different categories they are Month, Date, Temperature, Rainfall, Windspeed, Humidity, Daylight or Sunlight, Bud, Flowering, Perfect Flowering, Average Plum Fruit width, Average Plum Fruit Length, and the diseases are X0, X1, X2, X3, X4, X5, X6, X7, X8, & X9. So, this data shortly denoted as environmental factors, fruit size, and the series of X's are variety of diseases. First, we analyse the number of disease (count) and different type of diseases in plum fruits. In this process, easily recognize diseases types and it counts. Based on this result, we split the different categories of disease and then implement the logistic regression model to predict the plum fruit diseases one-by-one. In this case, use a different combination of input factors and produce the single output. After predicting the plum diseases, need to compare the actual disease and predicted disease by the help of correlation method. The purpose of the correlation method is defined as the relationship between two variables. Depends on the correlation result, its effortless to identify the combination of environmental factors in plum data. In this result, we concluded which factors cause more diseases in plum fruits.

2. Related Works

Sindhuja Sankaran, and et al, had been published a review of advance techniques for detecting plant diseases [1]. Jayme Garcia Arnal Barbedo was examined a review on the main challenges in automatic plant disease identification based on visible range images [2]. Jayme Garcia Arnal Barbedo and et al, were analysed and identified the multiple plant diseases using digital image processing [3]. Jiang Lu and et al, were published an in-field automatic wheat disease diagnosis system [4]. Konstantinos P. Ferentinos was examined deep learning models for plant disease detection and diagnosis [5]. Andreas Kamilaris, and Francesc X. Prenafeta-Boldu, were published the deep learning in agriculture: a survey [6]. Alexander Johannes and et al, were analysed automatic plant disease diagnosis using mobile capture devices, applied on a wheat use case [7]. Monalisa Ray and et al, examined the fungal disease detection in plants: Traditional assays, novel diagnostic techniques and biosensors [8]. Wayne Goodridge and et al, investigated the intelligent diagnosis of diseases in plants using a hybrid multicriteria decision making technique [9]. Savita Kolhe and et al, had been published a web-based intelligent disease-diagnosis system using a new fuzzy-logic based approach for drawing the inferences in crops [10]. Ilaria Pertot and et al, examined identificator: a web-based tool for visual plant disease identification, a proof of concept with a case study on strawberry [11]. Shanwen Zhang and et al, were published the leaf image based cucumber disease recognition using sparse representation classification [12]. Shanwen Zhang and et al, studied the plant diseased leaf segmentation and recognition by fusion of superpixel, K-means and PHOG [13]. Santanu Phadikar and et al, examined the rice diseases classification using feature selection and rule generation techniques [14]. Roberto Oberti and et al,were analysed the automatic detection of powdery mildew on grapevine leaves by image analysis: Optimal view-angle range to increase the sensitivity [15]. Based on the related works, this study analysed and identify the environmental factors in fruit disease by using a logistic regression model and the correlation method.

2.1 Logistic Regression Model

In regression analysis, logistic regression is estimating the parameters of a logistic model [16]. The two possible dependent variable values are often labelled as “0” and “1”, which represent outcomes such as pass/fail, win/lose, healthy/sick [16]. Peter Peduzzi and et al, were published a simulation study of the number of events per variable in logistic regression analysis [17]. Based on this reference and many of the authors detailed explained the logistic regression model.

2.2 Pearson Correlation Coefficient

The Pearson correlation coefficient or bivariate correlation is a measure of the linear correlation between two variables X and Y [18]. It has a value between +1 and -1, where +1 is total positive linear correlation, -1 is total negative linear correlation and 0 is no linear correlation [18].

$$\rho X, Y = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

where, cov is the covariance, σ_X is the standard deviation of X, σ_Y is the standard deviation of Y.

3. Materials and Methods

The plum data consist of several categories, they are month, date, temperature, rainfall, windspeed, humidity, daylight or sunlight, bud, flowering, perfect flowering, average of plum fruit width, average of plum fruit length, and the number of diseases are X0, X1, X2, X3, X4, X5, X6, X7, X8, and X9. This plum data is 7 months' data from Jan 01, 2018 to July 31, 2018and the data are collected from the area is HaeRyong in Korea. In this plum data, first essentially to analyse the number of plum diseases and variety of plum diseases.

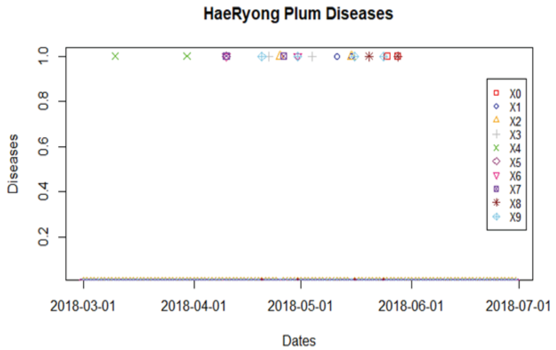

Figure 1. Diagram shows for Identifying the number of diseases in plum data.

<Figure 1> diagram shows for identifying the number of diseases in plum data. In <figure 1> X-axis has number of days from 01-03-2018 to 01-07-2018 and Y-axis has disease point from 0 to 1. If 0 means no disease and 1 means diseases occur in plum data. In <figure 1> contains different colors and various symbols shows for many diseases are occurred in plum data from X0 to X9. <Table 1> elaborately shows for different types of diseases and the number of diseases occurred in plum data.

Table 1. shows for types of diseases and the number of diseases.

Types of Diseases	Number of Diseases (Counts)
X0	2
X1	2
X2	2
X3	2
X4	2
X5	3
X6	3
X7	2
X8	2
X9	4

4. Result and Discussion

In this section, discussed the result of logistic regression and disease correlation values. Initially implemented the logistic regression model for the purpose of predicting the plum diseases. Afterward, necessary to compare the actual diseases and predicted diseases because need to find the accuracy of the result. In this case, the Pearson correlation coefficient method is used to find the accuracy of disease result. <Table 2> shows for the result of diseases correlation values from X0 to X4. The correlation is defined as a relationship between two values, one is the actual value of X and another is predicted value of X. <Table 2> shows the result of X0, X1, X2, X3, X4 diseases and the corresponding environmental factors (Input factors) used in the logistic regression model.

If the correlation value is 1 or close to 1 that value is perfect matching to actual disease and the corresponding environmental factors are caused the diseases. In <table 2> the highest correlation values are in bold number. In <table 2> and <table 3> T for Temperature, R for Rainfall, W for Windspeed, H for Humidity, D for Daylight, W1 for average of fruits width, and L for average of fruits Length.

Table 2. shows for result of diseases correlation values from X0 to X4.

Different Combination of Input Factors	**Disease Correlation Values**				
	X0	**X1**	**X2**	**X3**	**X4**
T	0.12	0.14	0.13	0.06	0.02
T + R	0.12	0.14	0.14	0.06	0.06
T + R + W	0.16	0.19	0.19	0.08	0.06
T + R+ W + H	0.16	0.20	0.20	0.09	0.07
T + R + W + H + D	0.18	0.20	0.22	0.13	0.12
T + R + W + H + D + W1	0.40	0.75	0.66	0.13	0.13
T + R + W + H + D + W1 + L	**0.41**	**0.76**	**0.68**	0.13	0.13

<Table 3> shows for the result of diseases correlation values for X5, X6, X7, X8, and X9. In <table 3> the highest correlation values are in bold number. The highest correlation values with their corresponding environmental factors and their plots are shown one by one.

Table 3. shows for the result of diseases correlation values from X5 to X9.

Different Combination of Input Factors	**Disease Correlation Values**				
	X5	**X6**	**X7**	**X8**	**X9**
T	0.01	0.01	0.01	0.10	0.17
T+ R	0.06	0.06	0.06	0.10	0.18
T + R + W	0.06	0.06	0.06	0.14	0.19
T + R + W + H	0.29	0.29	0.44	0.20	0.21
T + R + W + H + D	0.12	0.12	0.54	0.44	0.23
T + R + W + H + D + W1	0.10	0.11	0.65	**0.61**	0.25
T + R + W + H + D +W1 + L	**0.82**	**0.81**	**0.65**	0.60	**0.49**

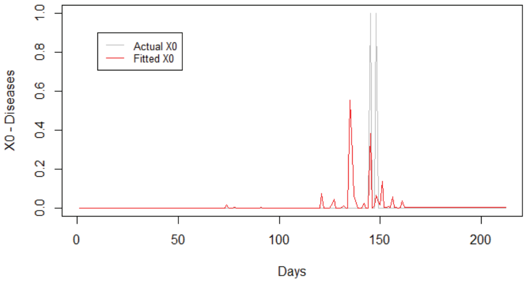

Figure 2. The diagram shows for the result of comparison of actual X0 disease and predicted X0 disease.

<Figure 2> diagram shows for the result of comparison of actual X0 disease and predicted X0 disease. X-axis has number of days and Y-axis has diseases point. If Y value is 0 then, it has no diseases and if Y value is 1 then, it has diseases occurred. The grey line shows for actual disease value and the redline shows for predicted disease value. The X0 diseases are formed on 150th days (5th month) and the number of X0 diseases (count) is 2. The maximum correlation value of X0 disease is 0.41 and their corresponding environmental factors are temperature, rainfall, windspeed, humidity, daylight or sunlight, an average of plum width, an average of plum length. Therefore, these environmental factors are caused to formed X0 diseases.

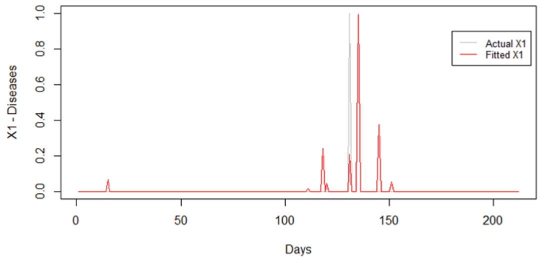

Figure 3. The diagram shows for the result of comparison of actual X1 disease and predicted X1 disease.

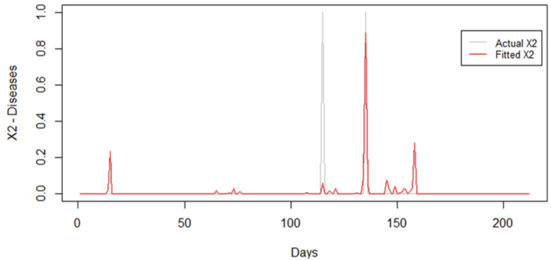

Figure 4. The diagram shows for the result of comparison of actual X2 disease and predicted X2 disease.

<Figure 3> diagram shows for the result of comparison of actual X1 disease and predicted X1 disease. The X1 diseases are formed on above 120th days (after 4th month) and number of X1 diseases (count) is2. The maximum correlation value of X1 disease is 0.76 and their corresponding environmental factors are temperature, rainfall, windspeed, humidity, daylight or sunlight, an average of plum width, an average of plum length. So, these environmental factors are caused to formed X1 diseases.

<Figure 4> diagram shows for the result of comparison of actual X2 disease and predicted X2 disease. The X2 diseases are created on above 100th days (after 3rd month) and number of X2 diseases (count) is2. The highest correlation value of X2 disease is 0.68 and their corresponding environmental factors are temperature, rainfall, windspeed, humidity, daylight or sunlight, an average of plum width, and an average of plum length. Therefore, these environmental factors are caused to created X2 diseases.

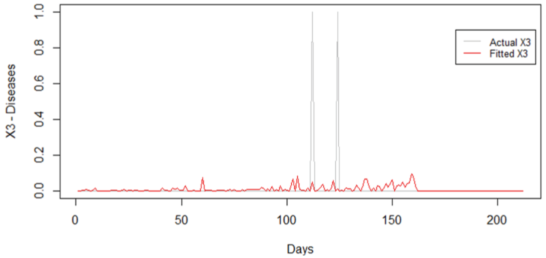

Figure 5. The diagram shows for the result of comparison of actual X3 disease and predicted X3 disease.

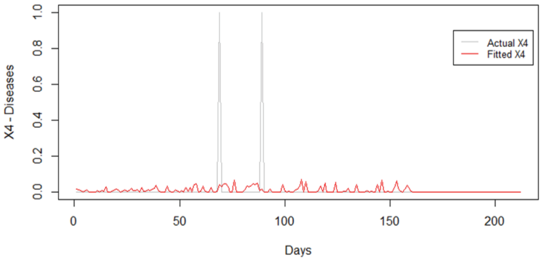

Figure 6. The diagram shows for the result of comparison of actual X4 disease and predicted X4 disease.

<Figure 5> diagram shows for the result of comparison of actual X3 disease and predicted X3 disease. The actual diseases are occurred on above 100th days (after 3rd month) and number of diseases (count) is 2. In this result, didn’t get the predicted diseases, because of the input factors are not suited to predicted X3 disease.

<Figure 6> diagram shows for the result of comparison of actual X4 disease and predicted X4 disease. The actual diseases are occurred on above 60th days (after 2nd month) and number of diseases (count) is 2. In this result, didn’t get the predicted diseases, because of the input factors are not suited to predicted X4 disease.

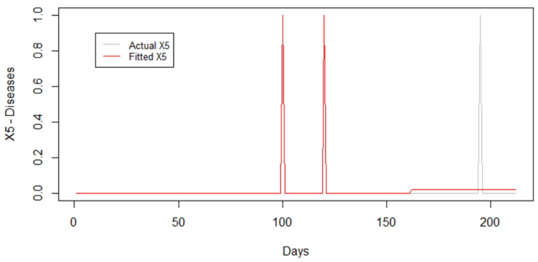

Figure 7. The diagram shows for the result of comparison of actual X5 disease and predicted X5 disease.

<Figure 7> diagram shows for the result of comparison of actual X5 disease and predicted X5 disease. <Figure 8> diagram shows for the result of comparison of actual X6 disease and predicted X6 disease. The X5 and X6 diseases are occurred on 100th days (after 3rd month) and the number of diseases (count) is 3. The maximum diseases correlation value of X5 disease is 0.82 and the X6 disease is 0.81 and their corresponding environmental factors are temperature, rainfall, windspeed, humidity, daylight or sunlight, average plum width, average plum length. Therefore, these environmental factors are caused to formed X5 diseases and X6 diseases.

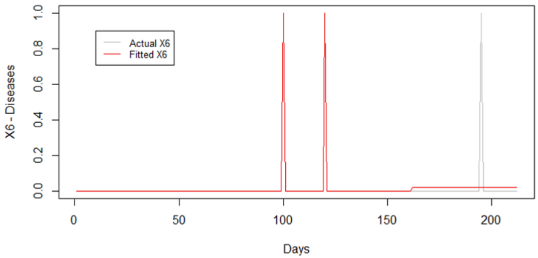

Figure 8. The diagram shows for the result of comparison of actual X6 disease and predicted X6 disease.

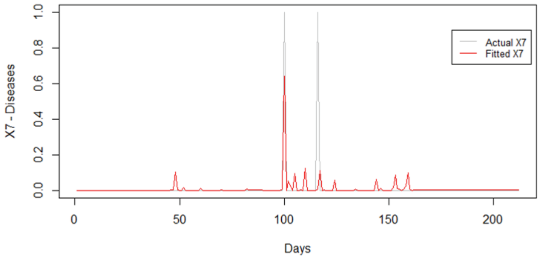

Figure 9. The diagram shows for the result of comparison of actual X7 disease and predicted X7 disease.

<Figure 9> diagram shows for the result of comparison of actual X7 disease and

predicted X7 disease in plum data. The actual X7 diseases are presented on 100th days (after 3rd month) and the number of diseases (count) is 2. The highest disease correlation value is 0.65 and their parallel environmental factors are temperature, rainfall, windspeed, humidity, daylight or sunlight, average plum width, and average plum length. Therefore, these input factors are caused to simulated the X7 diseases.

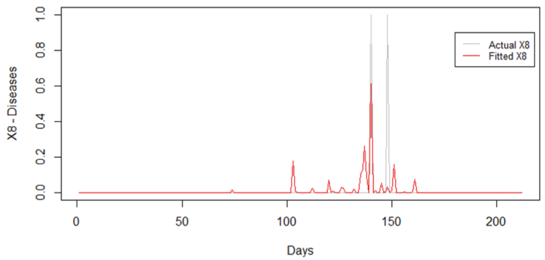

Figure 10. The diagram shows for the result of comparison of actual X8 disease and predicted X8 disease.

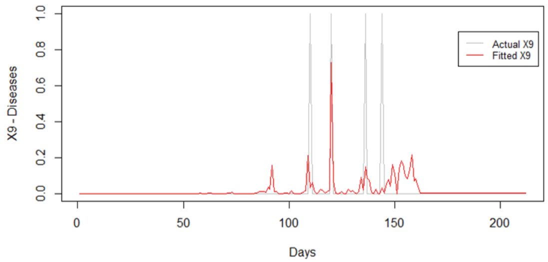

Figure 11. The diagram shows for the result of comparison of actual X9 disease and predicted X9 disease.

<Figure 10> diagram shows for the result of comparison of actual X8 disease and predicted X8 disease in plum data. The actual X8 diseases are presented on close to 140th days (after 4th month) and the number of diseases (count) is 2. The highest disease correlation value is 0.61 and their parallel environmental factors are temperature, rainfall, windspeed, humidity, daylight or sunlight, average plum width. Therefore, these input factors are caused to simulated the X8 diseases.

<Figure 11> diagram shows for the result of comparison of actual X9 disease and predicted X9 disease in plum data. The actual X9 diseases are presented on close to 110th days (after 3rd month) and the number of diseases (count) is 4. The maximum disease correlation value is 0.49 and their parallel environmental factors are temperature, rainfall, windspeed, humidity, daylight or sunlight, average plum width, and average plum length. Therefore, these input factors are caused to simulated the X9 diseases.

5. Conclusion

In this study, environmental changes are one of the most important reason to formed plum plant diseases or plum fruits diseases. To be solved, need to identify the environmental factors by using a logistic regression model and Pearson correlation coefficient. <Table 1> shows types of diseases and the number of diseases occurred in plum data. <Table 2> shows for result of diseases correlation values for X0, X1, X2, X3, & X4 diseases and their maximum diseases correlation value of X0 is 0.41, X1 is 0.76, X2 is 0.68. <Table 3> shows for result of diseases correlation values for X5, X6, X7, X8, & X9 and their highest diseases correlation values are X5 is 0.82, X6 is 0.81, X7 is 0.65, X8 is 0.61 and X9 is 0.49. For the reason, each disease is the different type, based on the diseases the correlation values are varying from to others. In result, elaborately explained about each and every diseases and their plots. Therefore, based on this results, we identified the most corresponding environmental factors are temperature, rainfall, humidity, windspeed, daylight or sunlight, an average of plum fruit width, an average of plum fruit length. In the future, this study will improve more and implements new sophisticated machine learning algorithms in plum diseases. This study will be extended into strawberry data and also identify the environmental factors in strawberry diseases.

References

[1] S. Sankaran, A. Mishra, R. Ehsani, and C. Davis, A review of advanced techniques for detecting plant diseases, Computers and Electronics in Agriculture, Vol. 72, Issue 1, pp. 1-13, 2010.

[2] J. G. A. Barbedo, *A review onthe main challenges in automatic plant disease identification based on visible range images*, Biosystems Engineering, Vol. 144, pp. 52-60, 2016.

[3] J. G. A. Barbedo, L. V. Koenigkan, and T. T. Santos, *Identifying multiple plant disease using digital image processing*, Biosystems Engineering, Vol. 147, pp. 104-116, 2016.

[4] J. Lu, J. Hu, G. Zhao, F. Mei, and C. Zhang, *An in-field automatic wheat disease diagnosis system*, Computer and Electronics in Agriculture, Vol. 142, pp. 369-379, 2017.

[5] K. P. Ferentinos, *Deep learning models for plant disease detection and diagnosis*, Computers and Electronics in Agriculture, Vol. 145, pp. 311-318, 2018.

[6] A. Kamilaris, and F. X. Prenafeta-Boldu, *Deep learning in agriculture: A survey*, Computers and Electronics in Agriculture, Vol. 147, pp. 70-90, 2018.

[7] A. Johannes, A. Picon, A. Alvarez-Gila, J. Echazarra, S. R. Vaamonde, A. D. Navajas, and A. Ortiz-Barredo, *Automatic plant disease diagnosis using mobile capture devices, applied on a wheat use case*, Vol. 138, pp. 200-209, 2017.

[8] M. Ray, A. Ray, S. Dash, A. Mishra, K. G. Achary, S. Nayak, and S. Singh, *Fungal disease detection in plants: Traditional assays, novel diagnostic techniques and biosensors*, Biosensors and Bioelectronics, Vol. 87, pp. 708-723, 2017.

[9] W. Goodridge, M. Bernard, R. Jordan, and R. Rampersad, *Intelligent diagnosis of diseases in plants using a hybrid MultiCriteria decision making technique*, Computers and Electronics in Agriculture, Vol. 133, pp. 80-87, 2017.

[10] S. Kolhe, R. Kamal, H. S. Saini, and G. K. Gupta, *A web-based intelligent disease-diagnosis system using a new fuzzy-logic based approach for drawing the inferences in crops*, Computers and Electronics in Agriculture, Vol. 76, Issue 1, pp. 16-27, 2011.

[11] I. Pertot, T. Kuflik, I. Gordon, S. Freeman, and Y. Elad, *Identificator: A web-based tool*

for visual plant disease identification, a proof of concept with a case study on strawberry, Computers and Electronics in Agriculture, Vol. 84, pp. 144-154, 2012.

[12] S. Zhang, X. Wu, Z. You, and L. Zhang, *Leaf image based cucumber disease recognition using sparse representation classification*, Computers and Electronics in Agriculture, Vol. 134, pp. 135-141, 2017.

[13] S. Zhang, H. Wang, W. Huang, and Z. You, *Plant diseased leaf segmentation and recognition by fusion of superpixel, K-means and PHOG*, Optik, Vol. 157, pp. 866-872, 2018.

[14] S. Phadikar, J. Sil, and A. K. Das, *Rice diseases classification using feature selection and rule generation techniques*, Computers and Electronics in Agriculture, Vol. 90, pp. 76-85, 2013.

[15] R. Oberti, M. Marchi, P. Tirelli, A. Calcante, M. Iriti, and A. N. Borghese, *Automatic detection of powdery mildew on grapevine leaves by image analysis: Optimal view-angle range to increase the sensitivity*, Computers and Electronics in Agriculture, Vol. 104, pp. 1-8, 2014.

[16] Logistic Regression from Wikipedia, https://en.wikipedia.org/wiki/Logistic_regression, 10 Jan. 2018.

[17] P. Peduzzi, J. Concato, El. Kemper, T. R. Holford, and A. R. Feinstein, *A simulation study of the number of events per variable in logistic regression analysis*, Journal of Clinical Epidemiology, Vol. 49, Issue 12, pp. 1373-1379, 1996.

[18] Pearson Correlation Coefficient from Wikipedia, https://en.wikipedia.org/wiki/Pearson_correlation_coefficient, 10 Jan. 2018.

로지스틱 회귀 분석에 의한 과실질병의 환경요인 규명

바산트 라구[1], 이명배[1], 이명훈[2], A.B.M. 살만 라흐만[1], 박장우[3], 조용윤[3], 신창선*[3]
[1]*순천대학교 정보통신공학과 대학원생*
[2]*국립농업과학원 스마트팜개발과 연구사*
[3]*순천대학교 정보통신공학과 교수*

요 약

식물질병은 농업 분야에서 가장 중요한 문제 중 하나이다. 식물질병은 과일을 파괴하고 생산율을 감소시키며, 세계적 경제성장을 감소시키고 식물질병 예방에 대한 수요를 증가시킨다. 식물질병의 주요원인은 기후변화, 기상조건과 환경요인이 있다. 식물질병을 예방하기 위해서는 식물질병에 따른 해당 환경요인을 규명해야한다. 본 연구에서는 로지스틱 회귀 모델을 사용하여 매실데이터를 분석하고 해당 환경요인에 대한 식물질병을 규명한다. 이를 위해 우선 질병의 수와 유형을 확인한 후 환경요인을 입력으로 하여 질병을 예측하는 로지스틱 회귀 분석을 구현한다. 마지막으로, 식물질병의 실제값과 예측값을 비교하고, 상관관계를 이용하여 질병예측의 정확성을 확인한다. 또한 질병을 발생시키는 추가적인 환경적 요인들을 제시한다. 마지막으로, 매실질병에 대한 환경적 요인을 명확하게 제시한다.

Acknowledgments

This research was supported by IPET (Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries) through Agri-Bio-industry Technology Development Program, funded by MAFRA (Ministry of Agriculture, Food and Rural Affairs) (No. 315018-3).

Vasanth Ragu completed Master of Computer Application (MCA) degree at Indo-American College from India. He is pursuing a Doctorate degree in information and Communication Engineering at Sunchon National University in South Korea. His area of interest includes Forecast Model, Ubiquitous Computing, Machine Learning, and Big Data Processing.

E-mail address: vasanth4224@scnu.ac.kr

Myeongbae Lee completed Bachelor degree in Computer Engineering from Korea. He received Master degree on Computer Science in South Korea. And currently pursuing Doctorate degree in the Information and Communication Engineering. He area of interest includes Advanced Agriculture Technology, IT Convergence, Cloud and Ubiquitous Computing.

E-mail address: lmb@scnu.ac.kr

Meong Hun Lee is a researcher of National Institute of Agricultural Sciences, Republic of Korea. He received his M.S. (2006) and Ph.D. (2011) degrees from Sunchon National University. His work includes a wide variety of wireless communication, big data analytics, and applied researches. His research interests are in the fields of mobile and wireless networks, ICT convergence, and standards.

E-mail address: jblee@jeongbo.ac.kr

A B M Salman Rahman received the Bachelor degree on Electronics and Telecommunication Engineering from Southeast University, Dhaka, Bangladesh. Currently pursuing Masters leading Doctorate degree in information and Communication Engineering at Sunchon National University in South Korea. His area of interest includes Forecast Model, Ubiquitous Computing, and Big Data Processing.

E-mail address: salman01717@gmail.com

Jangwoo Park received the B.S., M.S. and Ph.D. degrees in Electronic engineering from Hanyang University, Seoul, Korea in 1987, 1989 and 1993, respectively. In 1995, he joined the faculty member of the Sunchon National University, where he is currently a professor in the Department of Information & Communication engineering. His research focuses on Localization and SoC and system designs and RFID/USN technologies.

E-mail address: jwpark@sunchon.ac.kr

Yongyun Cho received the Ph.D. degree in computer engineering at Soongsil University. Currently, he is an assistant professor of the Department of Information & communication engineering in Sunchon National University. His main research interests include System Software, Embedded Software and Ubiquitous Computing.

E-mail address: yycho@sunchon.ac.kr

Changsun Shin received the Ph.D. degree in Computer Engineering at Wonkwang University. Currently, he is a professor of the Dept. of Information & Communication Engineering in Sunchon National University. His researching interests include Distributed Computing, IoT, Machine Learning, and Agriculture/ICT Convergence.

E-mail address: csshin@sunchon.ac.kr