Clasificación de hongos por medio de aprendizaje automático

Melany M. Molina E1315027456@LIVE.ULEAM.EDU.EC
Extensión El Carmen
Universidad Laica Eloy Alfaro de Manabí
El Carmen, Ecuador
https://github.com/Melany1523/IAWorks.git

Editor: Melany M. Molina

Abstract

El objetivo de este artículo es determinar si un hongo es comestible o venenoso. mediante la aplicación de aprendizaje automático y minería de datos usando 17 variables independientes y una base de datos online de hongos.

Keywords: Clasificación de hongos; aprendizaje automático; minería de datos; algoritmo J48; plataforma Weka

1. Introducción

Dado que los hongos crecen de forma natural en todos los ecosistemas y su mercado supera los 30 billones de dólares anuales por su gran potencial alimenticio, es importante poder identificar de una forma clara identificar los hongos comestibles de los que no lo son en especial cuando en los mismos cultivos pueden aparecer hongos venenosos por diversar razones ambientales. Considerando esto, el objetivo fundamental de este artículo es determinar por medio de la relación de 17 variables independientes tomadas de una base de datos existente en la web y referenciadas en (Dua y Graff, 2019; Wagner et al., 2021). Este analisis se realiza por medio de la plataforma libre de aprendizaje automático y minería de datps denominada Weka, bajo el algoritmo J48 de comportamiento similar a los algoritmos bayesianos.

2. Materiales y Métodos

La metodología se basa en el uso de la plataforma Weka, que permite aplicar diferentes algoritmos y técnicas de aprendizaje automático y minería de datos existente con 61.096 registros de hongos, a la que se le aplica un proceso de selección de variables, clasificación, optimización y comparación con otras técnicas inteligentes.

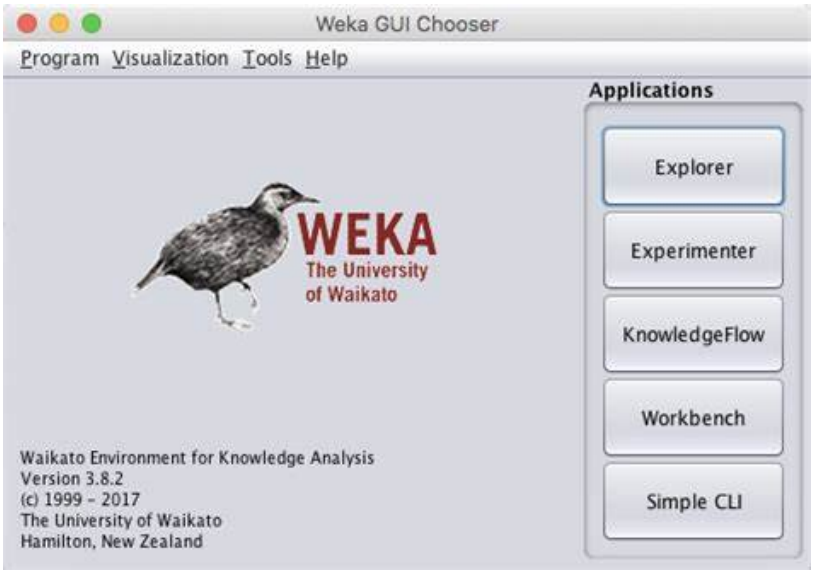

Figure 1: Software Weka

#	*Variable*	*Abrev*	*Unidades de medida / valores*	*Tipo Variable*
1	Diámetro sombrero	T_som	Cms	Independiente
2	Forma sombrero	F_som	b=campana, c=conica, x=convexa, f= plana, s=hundida, p= esférica, o= otras, NE = dato en blanco.	Independiente
3	Superficie sombrero	S_som	i=fibroso, g=surcos, y=escamosa, s= lisa, h=brillante, l=correosa, k=sedosa, t=pegajosa, w=arrugada, e=carnosa, d= otra, NE=dato en blanco	Independiente
4	Color sombrero	C_som	n=marrón, b= brillante, g=gris, r= verde, p=rosado, u=purpura, e=rojo, w=blanco, y=amarillo, l= azul, o=naranja, k=negro, NE=dato en blanco	Independiente
5	moretones o sangra	M_s	t=verdadero, f=falso	Independiente
6	Branquias	B	a=adnato, x=anexo, d=decurrente, e=libre, s=ondeado, p=poros, f=ninguno, ?=desconocido, NE = dato en blanco.	Independiente
7	Espacio branquias	B_e	c=estrecho, d=distante, f=ninguno, NE = dato en blanco.	Independiente
8	Color branquias	B_c	n=marrón, b= amarillo piel, g=gris, r= verde, p=rosado, u=purpura, e=rojo, w=blanco, y=amarillo, l= azul, o=naranja, k=negro, f=otro, NE = dato en blanco.	Independiente
9	Altura tallo	A_t	Cms	Independiente
10	Grosor tallo	G_t	Cms	Independiente
11	Raíz del tallo	R_t	b=bulbosa, s=hinchada, c= agrupada, u=sombrero, e=igual, z=rizomorfos, r= arraigada, f=otro, NE = dato en blanco	Independiente
12	Superficie del tallo	S_t	i=fibroso, g=surcos, y=escamosa, s= lisa, h=brillante, l=correosa, k=sedosa, t=pegajosa, f=otro, NE = dato en blanco	Independiente
13	Color del tallo	C_t	n=marrón, b= brillante, g=gris, r= verde, p=rosado, u=purpura, e=rojo, w=blanco, y=amarillo, l= azul, o=naranja, k=negro, f=otro, NE=dato en blanco	Independiente
14	Anillo	A	t=verdadero, f = falso	Independiente
15	Clase anillo	A_c	c=telarañas, e=evanescente, r=resplandeciente, g=estriado, l=largo, p=colgante, s=revestido, z=zona, y=escamoso, m=móvil, f=otro, ?=desconocido, NE = dato en blanco	Independiente
16	Habitat	H	g=pastos, m=prados, p=caminos, h=brezales, u=urbano, w=basuras, d=bosques, l=hojas, NE = dato en blanco	Independiente
17	Estación	E	s=primavera, u=verano, a=otoño, w=invierno, NE = dato en blanco	Independiente
18	Clase	C	e=comestible, p=venenoso	Dependiente

Figure 2: Estructura de la base de datos empleada

2.1 Selección de variables

Se hizo uso de la opción ConsistencySubSetEval de la plataforma Weka para seleccionar el meor de todos los subconjuntos de variables independientes, con un rendimiento igual al rendimiento completo de todas las variables independientes, esto para evitar la perdida de información al momento de realizar la predicción de la variable dependiente.

2.2 Clasificación por medio del algoritmo J48

Las variables seleccionadas son empleadas para realizar el proceso de clasificación en la plataforma Weka, por medio del algoritmo J48, el cual permite realizar sus clasificaciones mediante el manejo de la entropía de la información. Este proceso se realiza mediante tres parametrizaciones: 1) Entrenando y validando con el 100 por ciento de los datos, 2) Entrenando con el 50 y 75 por ciento de los datos y validando con el 50 y 25 por ciento de los datos respectivamente, 3) Por medio de una validación cruzada con los siguientes porcentajes de entrenamiento y validación.

2.3 Optimización por medio de variables calculadas

Se optimizan los porcentajes de clasificación usando una variable calculada, que es la relación entre dos variables numéricas. Se encuentra que la relación A-t/G-t es la mejor de ellas, con una efectividad del 99.76 por ciento.

2.4 Comparación con otras técnicas inteligentes de clasificación

Con el fin de establecer la efectividad de las técnicas de clasificación empleadas, se realiza una comparación frente a otras técnicas inteligentes de clasificación, bajo la misma misma plataform denominada Weka.

2.5 Clasificación por clústeres

Se emplea el algoritmo expectation maximisation de la plataforma Weka, con el fin de establecer el número de grupos de interés según sus probabilidades. Una vez establecido el número de grupos de interés, se emplea el algoritmo SimpleKMeans de la plataforma Weka para generar los diferentes grupos de interés con el fin de definir cuál o cuáles pueden ser clasificados en cada uno de los valores de la variable dependiente.

3. Resultados

3.1 Selección de variables y Clasificación mediante J48

Se presentan las reglas de clasificación obtenidas por el algoritmo J48, que permiten determinar si un hongo es comestible o venenoso según sus características. Se usan siete variables independientes: T-som, S-Som, C-som, A-t, G-t, S-t y C-t.

Variable	*Peso*
G_t	6906,5974***
C_t	4557,8906***
A_t	4552,9497***
S_t	4045,5581***
T_som	3581,2771***
C_som	3541,0798***
S_som	3188,13 ***
B	3094,3009
B_c	2245,0489
B_e	799,8579

Figure 3: Ranking de las variables

Instancias	#	%
Instancias clasificadas en forma correcta	60637	99.2926 %
Instancias clasificadas en forma incorrecta	432	0.7074 %
Estadística Kappa	0.9857	
Error absoluto medio	0.0095	
Error cuadrático medio	0.082	
Error absoluto relativo	1.915 %	
Error cuadrático relativo	16.5069 %	
Número de instancias totales	61069	

Figure 4: Matriz de clasificación

3.2 Optimización con variables calculadas

Se optimizan los porcentajes de clasificación usando una variable calculada, que es la relación entre dos variables numéricas. Se encuentra que la relación A-t/G-t es la mejor de ellas, con una efectividad del 99.76 por ciento.

3.3 Comparación con otras técnicas

Se comparan los resultados con otras técnicas de clasificación basadas en inteligencia artificial, usando una validación cruzada 90.10. Se obtiene una efectividad del 100 con las técnicas IB1, IBK, RandomForest y Random Tree.

3.4 Clasificación por clústeres

Se usan los algoritmos EM y SimpleKmeans para determinar 13 grupos de interés, según las características de los hongos. Se describen las propiedades de cada grupo y se identifican cinco grupos donde se clasifican los hongos venenosos.

4. Conclusiones

Se usó la plataforma Weka y el algoritmo J48 para seleccionar las variables más influyentes y generar un árbol de decisión que predice la comestibilidad de los hongos con una efectividad del 99.76 por ciento. En donde las variables más importantes fueron: diámetro sombrero, superficie sombrero, color sombrero, altura tallo, grosor tallo, superficie tallo y color tallo. Se comparó el metodo propuesto con otras técnicas de clasificación inteligente, como Bayes, regresión, Lazy, reglas y arboles, encontrando resultados similares. Después se identificó 13 grupos de interés mediante la clasificación por clústeres donde se encontró que cinco grupos correspondían a hongos venenosos con una efectividad del 85.33 por ciento.

References

Ardabili, S., Mahmoudi, A., y otros dos autores, Modeling and Comparison of Fuzzy and On/Off Controller in a Mushroom Growing Hall, http://dx.doi.org/10.1016/j.measurement.2016.04.050, Measurement, 90, 127-134 (2016)

Arunachalam, K., Puthanpura, S.S., y Yang, X., A Concise Review of Mushrooms Antiviral and Immunomodulatory Properties that May Combat Against COVID-19, https://doi.org/10.1016/j.foch Food Chemistry Advances, 1, 100023 (2022)

Braat, N., Koster, M., y Wösten, H., Beneficial Interactions Between Bacteria and Edible Mushrooms, https://doi.org/10.1016/j.fbr.2021.12.001, Fungal Biology Reviews, 39, 60 -72 (2022)

Cano-Estrada, A., y Romero-Bautista, L., Valor Económico, Nutricional y Medicinal de Hongos Comestibles Silvestres, http://dx.doi.org/10.4067/S0717-75182016000100011, Rev Chil Nutr., 43(1), 75-80 (2016)

Castrillón, O., Predicción del Divorcio por Técnicas de Minería de Datos, http://dx.doi.org/10.4067/S0718-07642021000500111, Información Tecnológica, 32(5), 111-120 (2021)

Dong, J., Zhang, J., y otros dos autores, Deep Learning for Species Identification of Bolete Mushrooms with two-Dimensional Correlation Spectral (2DCOS) Images, https://doi.org/10.1016/j.saa.202 Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 249, 119211 (2021)

Dua, D., y Graff, C., UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science (2019)

Lim, C., Chhabra, N., y otros cuatro autores, Atlas of Select Poisonous Plants and Mushrooms, http://dx.doi.org/10.1016/j.disamonth.2015.12.002, Disease-a-Month, 62, 41-66 (2016)

Lua, C., y Liaw, J., A Novel Image Measurement Algorithm for Common Mushroom Caps Based on Convolutional Neural Network, https://doi.org/10.1016/j.compag.2020.105336, Computers and Electronics in Agriculture, 171, 105336 (2020)

Moumita, S., y Das, B., Assessment of the Prebiotic Potential and Bioactive Components of Common Edible Mushrooms in India and Formulation of Synbiotic Microcapsules, https://doi.org/10.1016/j.lwt.2021.113050, LWT - Food Science and Technology, 156, 113050 (2022)

Oliveira, V., Almeida, A., y otros siete autores, A New Circular Economy Approach for Integrated Production of Tomatoes and Mushrooms, https://doi.org/10.1016/j.sjbs.2021.12.058, Saudi Journal of Biological Sciences, 29, 2756-2765 (2022)

Panda, S., y Luyten, W., Medicinal Mushrooms: Clinical Perspective and Challenges, https://doi.org/10.1016/j.drudis.2021.11.017, Drug Discovery Today, 27(2), 636-651 (2022)

Peter, M., Liu, Z., y otros cinco autores, Computational Intelligence and Mathematical Modelling in Chanterelle Mushrooms' Drying Process Under Heat Pump Dryer, https://doi.org/10.1016/j.biosy Biosystems Engineering, 212, 143 -159 (2021)

Rahman, H., Faruq, O., y otros ocho autores, IoT Enabled Mushroom Farm Automation With Machine Learning to Classify Toxic Mushrooms in Bangladesh, https://doi.org/10.1016/j.jafr.2021.100267 Journal of Agriculture and Food Research, 7, 100267 (2022)

Schunko, C., Li, X., y otros cinco autores, Local Communities' Perceptions of Wild Edible Plant and Mushroom Change: A Systematic Review, https://doi.org/10.1016/j.gfs.2021.100601, Global Food Security, 32,100601 (2022)

Saetang, N., Amornlerdpison, D., y otros tres autores, Processing of Split Gill Mushroom as a Biogenic Material for Functional Food Purpose, https://doi.org/10.1016/j.bcab.2022.102314, Biocatalysis and Agricultural Biotechnology, 41, 102314 (2022)

Tian, R., Liang, Z., y otros dos autores, Analysis of Aromatic Components of two Edible Mushrooms, Phlebopus portentosus and Cantharellus yunnanensis Using HS-SPME/GC-MS, https://doi.org/10.1016/j.rechem.2022.100282, Results in Chemistry, 4, 100282 (2022)

Valencia, M., Correa, J., y Díaz, F., Métodos Estadísticos Clásicos y Bayesianos para el Pronóstico de Demanda. Un Análisis Comparativo, https://doi.org/10.15446/rev.fac.cienc.v4n1.49775, Revista Facultad de Ciencias Universidad Nacional de Colombia, 4(1), 52 -67 (2015)

Wagner, D., Heider, D., y Hattab, G., Mushroom Data Creation, Curation, and Simulation to Support Classification Tasks, https://doi.org/10.1038/s41598-021-87602-3, Sci Rep, 11, 8134 (2021)

Wang, C., Wang, Y., y otros nueve autores, A Self-Floating and Integrated Bionic Mushroom for Highly Efficient Solar Steam Generation, https://doi.org/10.1016/j.jcis.2021.12.064, Journal of Colloid and Interface Science, 612, 88-96 (2022)

Wen, X., y Jing, P., Dietary Cerebrosides in Seven Edible Mushrooms: One Step Detection, Quantification, and Si-SPE Assisted Isolation, https://doi.org/10.1016/j.jfca.2022.104452, Journal of Food Composition and Analysis, 108, 104452 (2022)

Witten, I., Frank, E., y otros dos autores, Data Mining Practical Machine Learning Tools and Techniques, Morgan and Kaufman publication (Elsevier), ISBN-13: 978-0128042915, Cambrige, USA (2017)

Xu, M., Zhu, S., y otros cuatro autores, Effect of Selenium on Mushroom Growth and Metabolism: A Review, https://doi.org/10.1016/j.tifs.2021.10.018, Trends in Food Science Technology, 118, 328-340 (2021)

Zhao, Z., Fan, T., y otros ocho autores, A Simple Derivatization Method for Simultaneous Determination of Four Amino Group-Containing Mushroom Toxins in Mushroom and Urine by UPLC-MS/MS, https://doi.org/10.1016/j.foodcont.2021.108720, Food Control, 137, 108720 (2022)

Zotti, M., Zappatore, S., y Tosa, M., A Decision Support System for the Management of Accidental Mushroom and Plant Poisoning, https://doi:10.1016/s0014-827x(01)01103-x, II Farmaco, 56(5-7), 391-395 (2001)