Clasificación de hongos.

Rodrigo Daniel Hernández Barrera, Mateo Maya Martínez.

Introducción.

Los champiñones son el cuerpo frutal de algunos hongos que surgen de un grupo de micelios enterrados en sustrato. La mayoría de los champiñones pertenecen a la sub-división *Basidiomycotina* y algunos pertenecen a la subdivisión *Ascomycotina* del reino fungi (Mushroom culture, 2012).

Se ha reportado que hay alrededor de 50,000 especies conocidas de hongos y aproximadamente 10,000 de ellos son comestibles. De estos, alrededor de 180 pueden ser probadas para su cultivo artificial y 70 son ampliamente reconocidos como comida (Mushroom culture, 2012).

Aproximadamente de 1 a 2% de champiñones son venenosos para los humanos. y aunque solo unos cuantos de las 70-80 especies de champiñones venenosos son fatales cuando se ingieren, muchos de estos hongos mortales desafortunadamente tienen una aparienca muy similar a las especies comestibles, así que son especialmente peligrosos. Contrario a lo que se piensa, no hay una prueba casera que pueda distinguir entre variedades comestibles y venenosas. (Petruzzello, s.f.)

La meta de este análisis es clasificar entre hongos comestibles y venenosos de un set de datos obtenido de UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Mushroom), con el método de clasificación *K-Nearest-Neighbor (KNN)*. Los hongos se presentan en una gran variedad de formas, tamaños y colores, algunos son comestibles mientras que con otros se debería evitar contacto. La clasificación de hongos puede parecer crucial para la supervivencia para aquellas personas que viven en áreas rurales, por lo que es necesaria una buena clasificación.

Materiales y métodos.

El dataset de entrada contenía 8,416 muestras hipotéticas que corresponden a veintitrés especies de champiñones en las familias *Agaricus* y *Lepiota*. Cada muestra está identificada como definitivamente comestible, definitivamente venenosa, o de comestibilidad desconocida y no recomendado (también clasificado como venenoso).

Las clases se distribuyen de la siguiente manera:
-- Comestible: 4208 (51.8%)
-- Venenoso: 3916 (48.2%)

Cada muestra cuenta con 22 atributos (cap-shape, , cap-surface, cap-color, bruises, odor, gill-attachment, gill-spacing, gill-size, gill-color, stalk-shape, stalk-root, stalk-surface-above-ring, stalk-surface-below-ring, stalk-color-above-ring, stalk-color-below-ring, veil-type, veil-color, ring-number, ring-type, spore-print-color, population, habitat).

Se dividió el dataset original para las fases de entrenamiento, validación y predicción destinando un 95% para las dos primeras fases y y 5% para la última. Se utilizó el método *K-Nearest-Neighbor* que es un algoritmo basado en instancia de tipo supervisado de machine learning. Puede usarse para clasificar nuevas muestras (valores discretos) o para predecir (regresión, valores continuos). Sirve esencialmente para clasificar valores buscando los puntos de datos “más similares” (por cercanía) aprendidos en la etapa de entrenamiento y haciendo predicciones de nuevos puntos basado en esa clasificación.

Resultados y discusión.

Los datos de entrada fueron analizados encontrando 8416 entradas, cada una con 22 atributos. Para empezar con la manipulación de los datos, se comprobó que no se tuvieran valores nulos y se convirtieron los valores categóricos a ordinales utilizando *LabelEncoder*, esto último fue posible luego de modificar el tipo de dato de las características, ya que por default es considerado un "*object*" es necesaria la conversión a "*category*". Posteriormente se encontró que el atributo "*veil-type*" poseía solo un valor, lo cual la convertía en una variable no categórica y no contribuiría en los resultados, por lo que se removió del dataset.

Se dividió el dataset original para las distintas fases, destinando un 95% para las de entrenamiento y evaluación y 5% para la de predicción. Esto implicó que para las primeras dos fases contáramos con 7,995 muestras y para la última parte del análisis con 421.

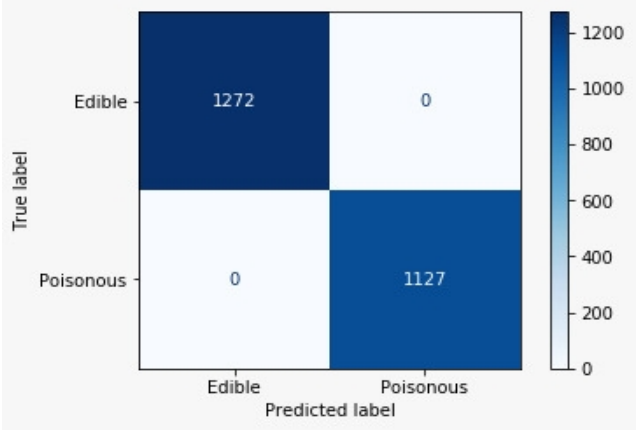
El método fue entrenado y evaluado con distintos valores de k, para ver la diferencia en su comportamiento. Se eligieron 1, 3, 5 y 7 como valores. Las medidas de rendimiento del clasificador fueron bastante similares con los diferentes valores de k que se usaron (Figura 1), pero únicamente alcanzó el valor de 1 cuando k = 1.

Valor de k.	Accuracy	Precision	Recall	F- score
k = 1.	1.0	1.0	1.0	1.0
k = 3.	0.9995831596498541	0.9996072270227808	0.9995563442768411	0.9995816100253143
k = 5.	0.9991663192997082	0.999215070643642	0.9991126885536823	0.9991631772892107
k = 7.	0.9983326385994165	0.9984326018808778	0.9982253771073647	0.9983261817547532

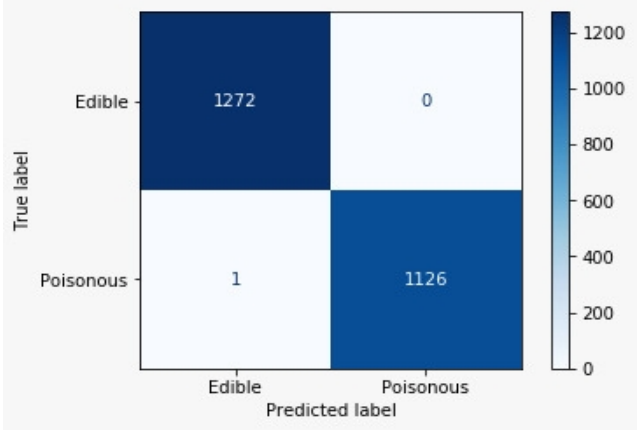
Tabla 1. Medidas de rendimiento del clasificador.

Posteriormente se graficó una matriz de confusión para observar los resultados de la fase de validación gráficamente. Se observó que conforme el valor de k aumentaba, el método cometía más equivocaciones. Con k = 1 no registró falsos positivos, ni falsos negativos, con k = 3 registró 1 error, para k = 5 registró 2 y con k = 7 registró 4 errores (Figura 1). Esto debido a que conforme se aumenta el valor de k aumente también el riesgo de *overfitting*.

A.



B.



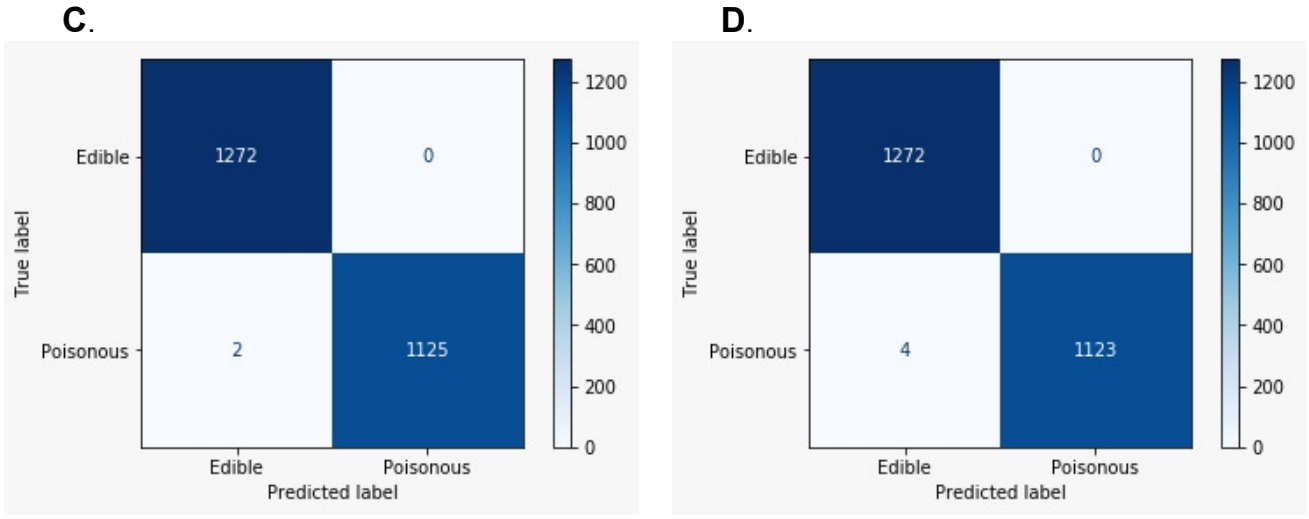


Figura 1. Matrices de confusión con los diferentes valores de k. A) k=1. B) k=3. C) k=5. D) k=7.

Para la parte final del análisis, se utilizaron el 5% de los datos anteriormente destinados y una k=1 porque fue el que entregó mejores resultados en las etapas previas, obteniendo los siguientes resultados.

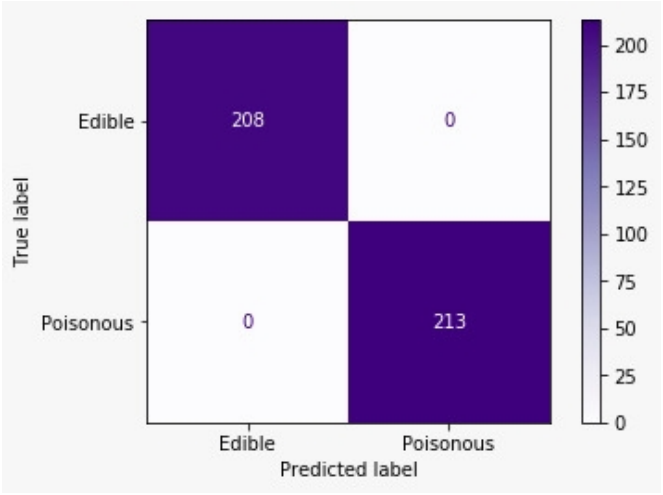


Figura 2. Matriz de confusión de la predicción.

Conclusiones.

El uso del método *K-Nearest Neighbor* en este análisis demostró ser útil para realizar la clasificación de nuevos datos correspondientes a los champiñones con base en sus características. Se comprobó que el aumento en el valor de k puede ocasionar un aumento en el sobreajuste del modelo. Hacer el mismo análisis con métodos diferentes de machine learning podría ser de utilidad en un análisis real para comparar resultados entre los diferentes algoritmos y elegir el más apropiado para cada situación a la que nos enfrentemos.

Referencias.

1. The Audubon Society Field Guide to North American Mushrooms. (1981). *Mushroom Data Set*. Recuperado de: https://archive.ics.uci.edu/ml/datasets/Mushroom

2. Mushroom Culture, (2012), *Classification of Mushrooms*. Recuperado de: http://ecoursesonline.iasri.res.in/mod/page/view.php?id=103103

3. Petruzzello, C. (s.f.). *How to Identify Poisonous Mushrooms*. Recuperado de: https://sciencing.com/identify-poisonous-mushrooms-2057768.html