Proyecto. clasificación de hongos
Reconocimiento de Patrones y Aprendizaje Automatizado

Adame Serrano Eduardo
Sánchez Rodríguez Ursula Vanessa

---

1. Introducción

Sabemos que en el reino Fungi los hongos son organismos representativos del mismo, y que están distribuidos ampliamente en la naturaleza pues desempeñan un papel importante en diversos ecosistemas. Sin embargo, la distinción entre hongos comestibles y venenosos puede resultar un verdadero reto de supervivencia para aquellos que no están familiarizados con sus características. Por ello, en este informe, abordaremos el problema de clasificación de hongos comestibles y venenosos mediante el uso de técnicas de aprendizaje automático. Nuestro objetivo es desarrollar un modelo preciso que prediga la comestibilidad de un hongo en función de sus características físicas. Lo cual genera un gran valor práctico por el hecho de evitar posibles intoxicaciones y promueve una recolección segura de hongos en entornos naturales.

2. Objetivo

El objetivo de este proyecto es desarrollar un modelo de aprendizaje automático que pueda predecir la comestibilidad de un hongo en función de sus características físicas.

3. Conjunto de datos

El conjunto de datos que utilizamos fue extraído de Kaggle y consta de un total de 8124 registros cada uno con 23 características, las cuales corresponden a diferentes aspectos físicos tales como colores, olores, texturas de ciertas partes o habitats en los que se encuentra cada hongo. Para el proyecto resulta de interés conocer si éstas características pueden ayudar a identificar la comestibilidad de los hongos.

Para comprender mejor a que se refieren las características del dataset se presenta el siguiente diagrama que muestra algunas de las partes más destacables de los hongos.

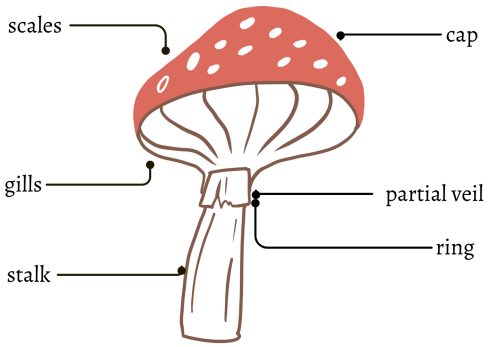

Figura 1: Anátomía de un hongo.

4. Método.

1. Preprocesamiento de datos: Se realizó un análisis exploratorio de los datos para identificar y tratar los valores faltantes. Dado que no se encontraron datos faltantes en el dataset, no fue necesario realizar un tratamiento adicional. A continuación, se aplicó la codificación one-hot a las características categóricas del dataset para convertirlas en variables numéricas con ceros y unos. A consecuencia de este procedimiento se pasó de tener 23 atributos a 119.

2. División de datos: El conjunto de datos se dividió en conjunto de entrenamiento, validación y prueba. El 70 % de los datos se utilizó como conjunto de entrenamiento, el 15 % como conjunto de validación y el 15 % como conjunto de prueba. Esta división se realizó utilizando la biblioteca scikit-learn.

3. Test chi-cuadrado: Se aplicó el test chi-cuadrado dado que es una técnica estadística utilizada para determinar si existe una relación significativa entre dos variables categóricas. En el contexto de selección de características, se aplica para identificar las características más relevantes que están asociadas de manera significativa con la variable objetivo.

4. Entrenamiento del modelo: Se utilizó el algoritmo de clasificación seleccionado para entrenar el modelo con el conjunto de entrenamiento. En este caso, el algoritmo elegido fue una red neuronal completamente conectada o densa.

   Se generaron dos tipos de modelos, ambos con 64 neuronas de entrada y 2 de salida, pero con distinto número de capas ocultas, el primero de ellos con una sola capa de 64 neuronas y el segundo con una capa de 64 y otra de 32. Este último fue el que obtuvo mejor precisión.

5. Validación del modelo: Se evaluó el rendimiento del modelo utilizando el conjunto de validación y prueba. Se calcula la matriz de confusión para evaluar la calidad de las predicciones de cada modelo.

6. Ajuste y evaluación final: Se realizaron ajustes en los parámetros del modelo para mejorar su rendimiento, en primera instancia el modelo 1, presentaba un muy buen ajuste y bastó agregar una capa oculta para que se obtuviera la mejor optimización. Una vez finalizado el

ajuste, se evaluó el modelo final utilizando el conjunto de prueba para obtener una estimación realista de su rendimiento en datos no vistos.

5. Resultados.

En ambos modelos se logró alcanzar una presición del 100 % en el conjunto de prueba. En las figuras 2 y 3 se muestran las matrices de confusión en las que es más sencillo apreciar la eficacia de los modelos al momento de clasificar cada muestra de hongo.

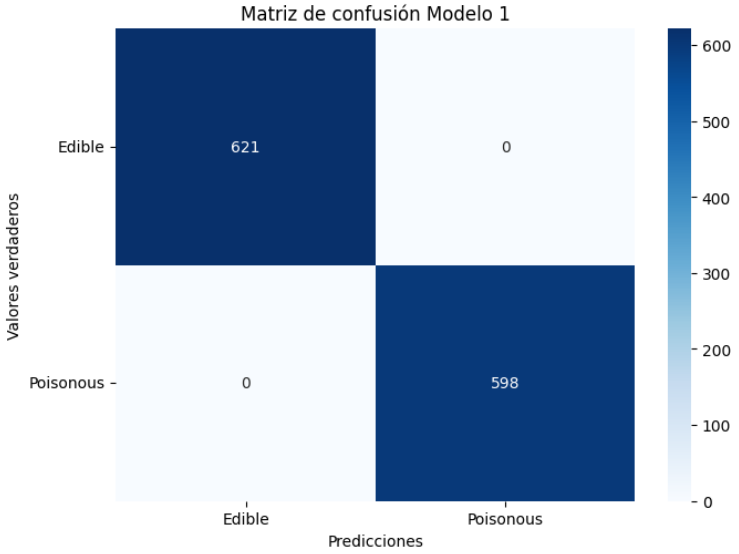

Figura 2: Matriz de confusión Modelo 1.

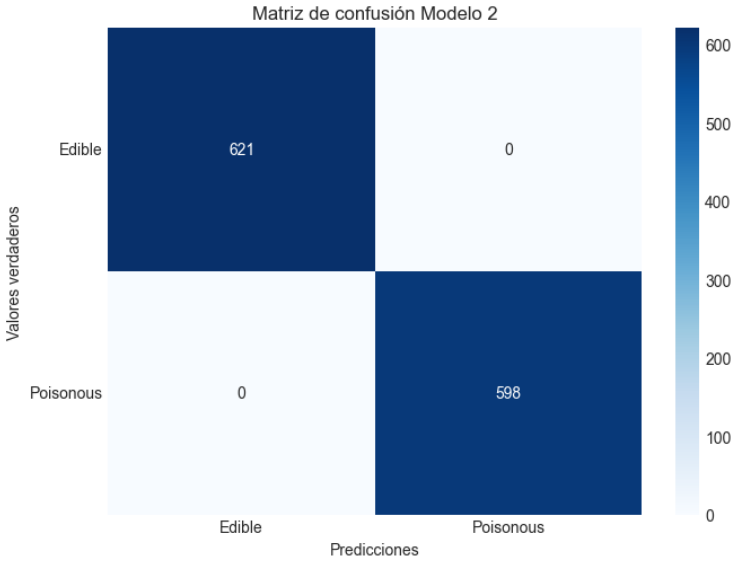

Figura 3: Matriz de confusión Modelo 2.