BIG DATA

Hecho por: Bautista Herrera Baker, Santiago Valle & Lorenzo Piqué.

Grupo: STRATOS

Dataset: Secondary Mushroom

https://archive.ics.uci.edu/dataset/848/secondary+mushroom+dataset

UNIVERSIDAD AUSTRAL

Definición de Problema

El problema que buscamos resolver en este trabajo práctico es la detección de hongos venenosos a partir de sus características morfológicas y ecológicas. Para ello, desarrollamos un modelo predictivo que clasifica los hongos en dos categorías: venenoso (p) y comestible (e).

La importancia de este problema radica en su impacto directo sobre la salud pública: la ingestión de hongos venenosos puede provocar intoxicaciones graves o incluso letales. Por lo tanto, contar con un sistema automatizado que permita anticipar el nivel de peligrosidad de un hongo en función de sus atributos resulta altamente relevante para prevenir riesgos y proteger la vida de las personas.

En términos prácticos, buscamos predecir si un hongo es venenoso (p), utilizando algoritmos de machine learning aplicados a un conjunto de datos con variables como el cap-surface, stem-surface, veil-color, stem-root, stem-color, entre otros. Esta clasificación nos permitirá evaluar diferentes modelos y seleccionar aquel que logre minimizar los falsos negativos, es decir, aquellos casos en los que un hongo venenoso es clasificado incorrectamente como comestible.

El dataset *Secondary Mushroom* está compuesto por dos bases: Primary Mushroom y Secondary Mushroom. La base Primary Mushroom contiene datos originales sobre especies de hongos, mientras que la Secondary Mushroom fue generada artificialmente a partir de ella mediante un proceso de simulación, utilizando un script en Python provisto por los autores del dataset. La Secondary Mushroom incluye más de 61.000 registros simulados, que representan diferentes características morfológicas de hongos (como color, forma, tamaño y estación de aparición). Cada hongo está etiquetado como comestible (e) o venenoso (p), combinando la clase “desconocido” junto a los venenosos por motivos de seguridad. El dataset final contiene 20 variables (17 categóricas y 3 métricas) que fueron aleatorizadas y ordenadas para imitar variabilidad real entre especies. La variable “class”, que indica si el hongo es venenoso (p) o comestible (e) será nuestra variable target.

Análisis Exploratorio de Datos

Con el objetivo de familiarizarnos con el dataset Secondary, realizamos un análisis exploratorio utilizando distintas herramientas de Orange, como Distributions, Scatter Plot, Feature Statistics y Correlations. El objetivo fue identificar patrones entre las características de los hongos y su clase (p, e), y formular hipótesis sobre qué variables podrían ser más relevantes para la clasificación.

Utilizando el widget Correlations, se calcularon las correlaciones entre variables numéricas. Se identificaron asociaciones entre:

- cap-diameter y stem-width (correlación: +0.695)
- stem-height y stem-width (correlación: +0.436)
- cap-diameter y stem-height (correlación: +0.423)

Esto indica que las variables que describen el tamaño del hongo están relacionadas entre sí, lo cual puede ser útil para modelar su morfología general.

Con Scatter Plot, visualizamos la relación entre atributos numéricos y la clase. Los gráficos muestran la clase en función de:

- stem-width: mayor densidad de venenosos para ciertos anchos específicos.
- stem-height: diferencia de distribución entre ambas clases.
- cap-diameter: los hongos venenosos tienden a ocupar un rango más estrecho de diámetros que los comestibles.

De estos gráficos también se pueden sacar varias conclusiones. Al analizar stem-width en relación a la clase, observamos que el ancho de los hongos puede llegar hasta 103 milímetros. Haciendo un análisis más detallado, detectamos que no existen casos de que el hongo sea venenoso cuando el mismo supera los 60 milímetros de ancho, es decir, que todos los hongos que superan esta medida son comestibles en el dataset.

Por otro lado, al comparar el stem-height respecto a la clase, pudimos observar que no hay hongos venenosos que superen los 21 centímetros. Mientras que si hay casos de hongos comestibles que superen esta medida. De esta forma, todos los hongos que superan esta medida son comestibles en el dataset.

Por último, analizando el cap-diameter en relación a la clase detectamos que existe un grupo separado de hongos en donde su cap-diameter se encuentra por encima de los 31 cm. De este grupo, pudimos observar que todos los hongos son comestibles. De esta forma, todos los hongos que superan esta medida son comestibles en el dataset.

A través de gráficos de distribución se evaluaron otras variables categóricas. Las observaciones que obtuvimos fueron:

- veil-color: valores como e, k, n, u (red=e, black=k, brown=n, purple=u) están 100% asociados a hongos venenosos, mientras que el color y (yellow=y) esta 100% asociado a hongos comestibles.
- stem-root: categorías como c, f y r (club=c, rooted=r, filamentous=f) también se asocian exclusivamente con los hongos venenosos.
- cap-color: si bien hay valores compartidos por ambas clases, algunos colores presentan mayor presencia en una clase que otra, lo que puede aportar información útil al modelo.

Preprocesamiento y Selección de Variables

Durante esta etapa realizamos tareas de limpieza y transformación para preparar los datos de forma adecuada para su análisis en Orange. En particular, se identificaron valores faltantes representados con el símbolo ?, los cuales fueron reemplazados por el valor más frecuente dentro de cada variable, con el objetivo de preservar la consistencia general del conjunto de datos sin perder registros valiosos.

En cuanto a la selección de variables, decidimos trabajar con un subconjunto de características que demostraron tener mayor relevancia predictiva respecto a la variable objetivo class, que distingue entre hongos comestibles (e) y venenosos (p). Las variables seleccionadas fueron:

- cap-surface
- stem-surface
- veil-color
- stem-root
- stem-color

El resto de las variables (por ejemplo, cap-color, season, gill-color, entre otras) fueron excluidas por baja correlación, alta cardinalidad o redundancia. Esta selección buscó reducir la complejidad del modelo, facilitar su interpretación y mejorar su capacidad de generalización, evitando el sobreajuste.

Modelado

Construimos un flujo de trabajo en Orange que nos permitió analizar el dataset *Secondary Mushroom* de forma visual y estructurada. A través de esta herramienta,

organizamos el análisis en distintas etapas: exploración inicial, selección de variables, preprocesamiento, entrenamiento de modelos y evaluación. A continuación, detallamos cada paso.

Comenzamos conectando el dataset al widget *Data Info* para obtener un panorama general: cantidad de registros, atributos y tipos de variables. Luego, utilizamos *Data Table* para visualizar los datos individualmente y confirmar que los atributos se habían cargado correctamente. Con *Feature Statistics* evaluamos la distribución estadística de cada variable, identificando valores faltantes, frecuencias y rangos de los atributos métricos.

Complementamos esta revisión haciendo un análisis exploratorio de los datos para familiarizarnos con la base de datos, con herramientas gráficas como *Distributions* y *Scatter Plot*, que nos permitieron visualizar la relación entre las variables y la clase objetivo (comestible o venenoso). Esta etapa fue fundamental para identificar patrones tempranos y formular hipótesis. Además, empleamos el widget *Correlations* para cuantificar relaciones lineales entre variables numéricas, lo cual confirmó asociaciones relevantes entre atributos como *cap-diameter*, *stem-height* y *stem-width*.

Para profundizar en la selección de variables, incorporamos el widget *Rank*, que nos permitió cuantificar la importancia relativa de cada atributo respecto a la variable objetivo. Esta evaluación nos ayudó a priorizar los atributos más informativos para el modelo, descartando aquellos con baja relevancia o alta redundancia.

A partir de los resultados obtenidos en *Rank* y del análisis exploratorio, utilizamos *Select Columns* para conservar únicamente las variables con mayor poder predictivo. Luego, aplicamos transformaciones en el widget *Preprocess*, reemplazando valores faltantes (representados con '?') por el valor más frecuente de cada columna. Esta estrategia nos permitió mantener la integridad del dataset sin eliminar registros valiosos.

Para evaluar el rendimiento real de los modelos, dividimos el dataset en dos subconjuntos: uno de entrenamiento y otro de prueba. Utilizamos el widget *Data Sampler* para seleccionar aleatoriamente el 80% de los registros como conjunto de entrenamiento, y el 20% restante como test. Esto garantizó que los modelos fueran evaluados con datos no vistos durante el entrenamiento.

Entrenamos tres modelos de clasificación distintos: Árbol de Decisión (Tree), Regresión Logística, Red Neuronal. Cada modelo fue entrenado con el conjunto de entrenamiento y conectado a los widgets de evaluación. Buscamos comparar modelos con

diferentes niveles de complejidad e interpretabilidad para seleccionar el más adecuado para nuestros objetivos del proyecto.

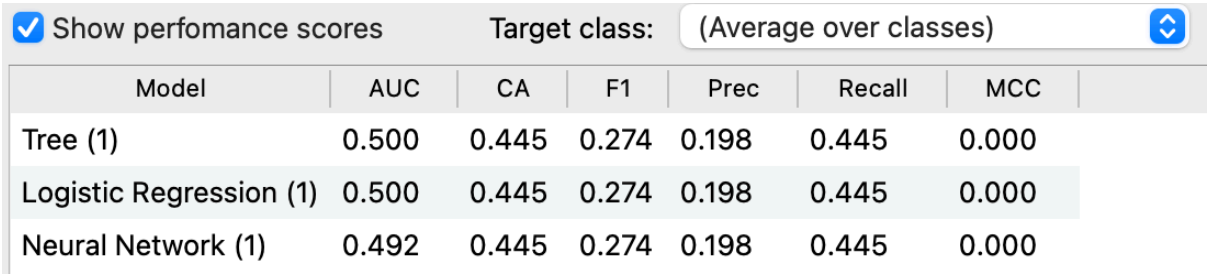
Utilizamos el widget *Test & Score* para obtener métricas clave como *Accuracy*, *Precision*, *Recall*, *F1 Score*, *MCC* y *AUC*. Este análisis comparativo nos permitió identificar fortalezas y debilidades de cada modelo, con especial foco en el recall de la clase 'p' (venenoso), ya que minimizar los falsos negativos era prioritario. Este análisis será explicado detalladamente en la sección de Interpretación de Predicciones.

Complementamos la evaluación con *Confusion Matrix* para entender en detalle cómo se distribuían los aciertos y errores en cada clase. También empleamos *ROC Analysis* para comparar visualmente la capacidad de discriminación de los modelos.

Segunda etapa de Modelado

Luego de entrenar nuestros modelos de clasificación sobre la base *Secondary*, decidimos probar su capacidad de generalización aplicándolos sobre el conjunto de datos *Primary*, que contiene observaciones reales del mismo dominio micológico. Esta instancia nos permitió evaluar si los patrones aprendidos por los modelos en un entorno simulado eran lo suficientemente robustos como para predecir correctamente casos del mundo real.

Para llevar a cabo esta evaluación, mantuvimos los mismos modelos que habíamos utilizado anteriormente (Árbol de Decisión, Regresión Logística y Red Neuronal), entrenados únicamente con la base *Secondary*. La base *Primary* fue pre-procesada de forma consistente, asegurando la aplicación de los mismos pasos de limpieza y transformación. Finalmente, se utilizó el widget *Predictions* para obtener los resultados y el widget *Test and Score* para calcular las métricas de desempeño sobre la base *Primary*.



Model	AUC	CA	F1	Prec	Recall	MCC
Tree (1)	0.500	0.445	0.274	0.198	0.445	0.000
Logistic Regression (1)	0.500	0.445	0.274	0.198	0.445	0.000
Neural Network (1)	0.492	0.445	0.274	0.198	0.445	0.000

Los resultados obtenidos al aplicar los modelos sobre este conjunto de datos fueron notablemente bajos y lejos de lo esperado. Tanto el Árbol de Decisión como la Regresión Logística arrojaron un AUC de 0.500, mientras que la Red Neuronal apenas alcanzó un AUC de 0.492. Estos valores indican que los modelos no fueron capaces de discriminar entre clases, operando prácticamente al nivel de una clasificación aleatoria.

El Accuracy fue idéntico en los tres modelos: 0.445, lo que implica que menos de la mitad de las instancias totales fueron correctamente clasificadas. A su vez, tanto el F1-score como el Recall se mantuvieron en 0.274 y 0.445 respectivamente, reflejando un equilibrio muy pobre entre precisión y sensibilidad. El Precision fue extremadamente bajo (0.198), lo que indica un alto número de falsos positivos.

Por último, la métrica MCC, que evalúa la correlación entre predicciones y clases reales considerando todos los cuadrantes de la matriz de confusión, fue igual a 0 para los tres modelos. Esto confirma que no hubo ninguna capacidad real de predicción, y que los resultados fueron equivalentes al azar.

Cuando analizamos específicamente la clase *p* (venenoso), observamos que todas las métricas relevantes (precisión, recall, F1) dieron como resultado un valor de 0. En otras palabras, ninguno de los modelos logró identificar correctamente un solo hongo venenoso en la base *Primary*. Este comportamiento representa una clara falla en la capacidad del modelo para generalizar el conocimiento adquirido en un conjunto a otro, lo que nos llevó a reflexionar sobre las causas de este bajo desempeño.

Entre los factores que pueden explicar esta situación, destacamos en primer lugar las diferencias entre los dominios de entrenamiento y prueba. Si bien la base *Secondary* fue construida artificialmente a partir de *Primary*, el proceso de simulación pudo haber introducido patrones o relaciones estadísticas que no existen en los datos reales. Esto puede haber llevado a que los modelos aprendieran “regularidades artificiales” que no son aplicables fuera del entorno simulado.

En segundo lugar, consideramos que los modelos podrían haber sufrido de overfitting a los datos del conjunto *Secondary*. Al ajustarse demasiado a las características específicas de ese dataset, los algoritmos perdieron la capacidad de adaptarse a nuevas distribuciones, como la de *Primary*, donde la variabilidad es mayor o los casos son más complejos.

Finalmente, esta experiencia nos deja una lección importante: un modelo que funciona bien en un entorno artificial no necesariamente tendrá buen desempeño en la

realidad. Esto refuerza la necesidad de contar con datos representativos del dominio real durante el entrenamiento y de validar los modelos en contextos distintos al que fueron entrenados antes de considerar su implementación práctica.

Evaluación del Modelo e Interpretación de Predicciones

Volviendo a nuestro objetivo principal, se entrenaron tres modelos (Árbol de Decisión, Regresión Logística y Red Neuronal) utilizando la base *Secondary* con una división del 80% para entrenamiento y 20% para testeo. El análisis de resultados se hace para tres niveles:

1. Promedio general (todas las clases).
2. Target = e (edible/comestible).
3. Target = p (poisonous/venenoso).

Queremos desarrollar un modelo que pueda predecir si un hongo es venenoso ('p') o comestible ('e'), con especial foco en minimizar los falsos negativos, ya que clasificar erróneamente un hongo venenoso como comestible podría tener consecuencias graves para la salud.

Haciendo un análisis general del modelo, el widget test and score nos dió los siguientes resultados:

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.913	0.818	0.818	0.818	0.818	0.633
Logistic Regression	0.781	0.708	0.708	0.708	0.708	0.410
Neural Network	0.915	0.820	0.820	0.820	0.820	0.635

En este nivel general, tanto el Árbol de Decisión como la Red Neuronal muestran muy buenos resultados en todas las métricas clave. El valor de AUC (área bajo la curva ROC) supera el 0.9 en ambos modelos, lo que indica una buena capacidad para distinguir entre hongos comestibles y venenosos. En contraste, la Regresión Logística presenta

valores más bajos en todas las métricas, evidenciando un menor poder predictivo. El Árbol presenta un buen equilibrio entre precisión y recall, con una MCC aceptable. La Red Neuronal se destaca con el MCC un poco más alto (0.635), lo que indica una clasificación general ligeramente más robusta.

Además, hicimos un análisis específicamente para cada variable dentro de la clase, para poder ver que modelo se ajustaba mejor a la hora de predecir si es venenoso o comestible:

Evaluation results for target e

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.913	0.818	0.797	0.796	0.798	0.633
Logistic Regression	0.781	0.708	0.674	0.671	0.678	0.410
Neural Network	0.915	0.820	0.795	0.806	0.785	0.635

Cuando analizamos exclusivamente la capacidad de los modelos para identificar hongos comestibles, el Arbol se destaca con un recall de 0.798, lo que significa que detecta correctamente el 79% de los hongos comestibles.A su vez, la precision de este modelo tiene un valor similar. La Red Neuronal, si bien tiene un recall más bajo (0.785), presenta un mejor equilibrio con la precisión, lo que lo convierte en una opción más conservadora pero segura.

Evaluation results for target p

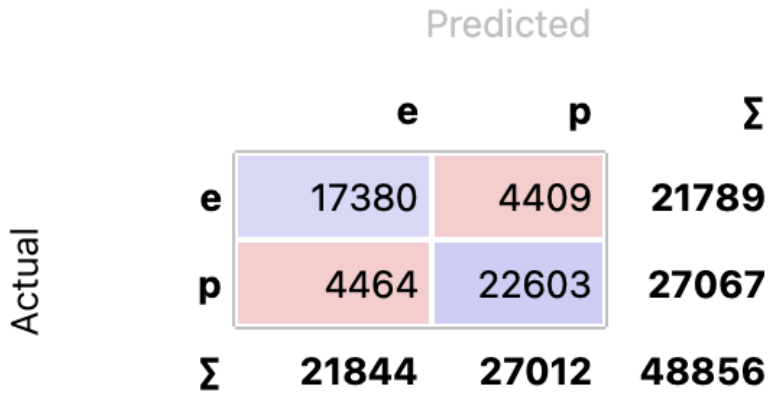
Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.913	0.818	0.836	0.837	0.835	0.633
Logistic Regression	0.781	0.708	0.735	0.739	0.732	0.410
Neural Network	0.915	0.820	0.839	0.830	0.848	0.635

Esta sección es clave para los objetivos del proyecto, ya que nos enfocamos en analizar el desempeño de los modelos al predecir correctamente los hongos venenosos (clase “p”), priorizando la minimización de falsos negativos, dado el riesgo que implicaría clasificar incorrectamente un hongo peligroso como comestible.

Según los resultados actualizados, el modelo de Red Neuronal fue el que obtuvo el mejor desempeño general. Alcanzó un recall del 84,8%, lo que significa que identificó correctamente casi 85 de cada 100 hongos venenosos. También presentó una precisión del 83,0%, lo cual indica que la gran mayoría de las veces que predijo un hongo como venenoso, acertó. Además, logró el AUC más alto (0.915) y el mejor F1 Score (0.839), consolidándose como el modelo más robusto y confiable del análisis.

El modelo de Árbol de Decisión también mostró un rendimiento muy sólido, con un recall de 83,5%, precisión de 83,7% y un AUC de 0.913. La diferencia respecto a la Red Neuronal fue mínima, y si bien es levemente inferior en métricas globales, su mayor interpretabilidad lo convierte en una opción válida, especialmente en contextos donde la transparencia del modelo es importante.

Por otro lado, la Regresión Logística tuvo un rendimiento claramente inferior. Obtuvo un recall de 73,2% y un AUC de 0.781, lo que refleja una menor capacidad para distinguir correctamente los casos de hongos venenosos. Esto la posiciona como la opción menos adecuada para los objetivos del proyecto.



La matriz de confusión del modelo refuerza estos resultados: de los 27.067 hongos realmente venenosos, el modelo logró identificar correctamente 22.603 casos (TP) y clasificó erróneamente como comestibles 4.464 casos (FN). A su vez, de los 21.789 hongos comestibles, clasificó correctamente 17.380 (TN) y se equivocó con 4.409 (FP). Estos valores muestran un buen equilibrio entre sensibilidad y seguridad, manteniendo bajo control tanto los falsos negativos como los falsos positivos.

En síntesis, el modelo de Red Neuronal se destaca como la mejor alternativa, combinando un alto nivel de recall y precisión con una sólida capacidad discriminativa (AUC), cumpliendo con el principal objetivo del proyecto: identificar con la mayor seguridad posible los hongos peligrosos. El modelo de Árbol de Decisión también se presenta como

una opción robusta, especialmente por su claridad interpretativa, mientras que la Regresión Logística quedó descartada por sus métricas significativamente más bajas.

Estas observaciones se alinean con nuestras hipótesis iniciales. Anticipamos que variables como el tamaño del hongo, el color del velo y el entorno en el que crecen tendrían fuerte poder discriminativo. El modelo confirmó esta intuición, basándose efectivamente en estas variables para separar ambas clases. En particular, en el modelo red neuronal se puede observar cómo estas variables aparecen en los nodos superiores, lo cual indica su relevancia en la toma de decisiones.

En conclusión, el análisis de la importancia de variables y su relación con nuestras hipótesis demuestra una buena alineación entre la exploración inicial y el aprendizaje del modelo. Esto refuerza la validez del enfoque metodológico adoptado y justifica la elección de las variables incluidas en la etapa de preprocesamiento.

Conclusiones y Lecciones Aprendidas

A modo de cierre, este trabajo nos permitió aplicar técnicas de análisis exploratorio y modelado predictivo para abordar un problema de alto impacto: la detección de hongos venenosos a partir de sus características morfológicas. Utilizando Orange, desarrollamos un flujo de trabajo estructurado que nos permitió evaluar diferentes modelos y métricas de desempeño.

Los resultados obtenidos sobre la base *Secondary* fueron muy positivos. El modelo de Red Neuronal, en particular, logró un buen equilibrio entre precisión y recall, siendo el más eficaz para identificar hongos venenosos y minimizar los falsos negativos. Además, confirmamos que muchas de las variables destacadas en el análisis exploratorio fueron efectivamente relevantes para la predicción.

No obstante, al aplicar estos modelos sobre la base *Primary*, los resultados fueron decepcionantes. Todos los modelos fallaron en generalizar el conocimiento aprendido, obteniendo métricas cercanas al azar y sin identificar correctamente ningún hongo venenoso. Esto evidenció las limitaciones de trabajar con datos simulados y la importancia de contar con datos representativos del dominio real.

Como lección principal, entendimos que un buen rendimiento sobre datos artificiales no garantiza una correcta aplicación práctica. En futuros trabajos, será clave incorporar más

datos reales y reforzar las estrategias de validación para mejorar la robustez de los modelos.