Breve introducción

Nuestro modelo de datos para realizar el entrenamiento fue la base de datos de secondary mushrooms, que contiene 2 bases de datos:

- Primary: Se trata de una base con 173 entradas, que contiene datos varios sobre hongos:
    - Si son comestibles o no
    - Su familia
    - Forma del sombrero
    - Tamaño del sombrero

  Entre otras variables que ayudarán a predecir si es venenoso o no el hongo
- Secondary: Es una base de datos hipotética con 61.069 entradas, con lo cual es mucho más amplia que la primera. De por si las variables para predecir si es comestible o no son las mismas, lo único que cambia es la presentación de los datos ya que secondary estaba un poco mejor presentada para se utilizada desde el primer momento

Problema: Tenemos que vender hongos a una cadena de supermercados en La Pampa, pero no sabemos si todos los hongos que nos llegan son venenosos o comestibles, con lo cual no podemos vender un hongo a esta cadena y que se acabe intoxicando alguien. Para ello entonces necesitamos un modelo que clasifique entre venenoso y comestible nuestros hongos.

Luego de haber explicado de qué se trata nuestros datasets, ahora explicaremos cómo realizamos el análisis de cada uno, sea pre procesamiento u otros aspectos como qué variables dejamos de lado.

Análisis exploratorio de datos

Las variables que más creemos que pueden llegar a impactar en la predicción de si un hongo es venenoso o son:

1. **Does-bruise-bleed (¿se magulla o sangra?):** Algunos hongos venenosos presentan reacciones visibles al ser manipulados, como cambio de color o secreción de líquidos. Hipótesis: Los hongos que sangran o se magullan al tocarlos tienen más probabilidad de ser venenosos.
2. **Cap-color (color del sombrero):** En la naturaleza, colores llamativos (rojo, verde, púrpura) pueden ser señales de advertencia. Hipótesis: Los hongos con colores como rojo, verde, púrpura o azul podrían tener mayor probabilidad de ser venenosos.
3. **Ring-type (tipo de anillo):** Algunos tipos de anillo son característicos de géneros venenosos como Amanita. Hipótesis: Tipos de anillo como pendant (p) o sheathing (s) podrían estar más asociados a hongos tóxicos.
4. **Hábitat (hábitat):** Determinadas especies venenosas prefieren ambientes específicos. Hipótesis: Los hongos que crecen en bosques (woods = d) o en hábitats perturbados (waste = w) podrían tener mayor frecuencia de toxicidad.

Descripción de primary

Inicialmente primary tiene 3 variables numéricas:

- Cap diameter
- Stem height
- Stem width

Estas 3 sin embargo, vienen en el dataset como intervalos y con lo cual tuvimos que tratarlos con código de python y los dividimos en 3 categorías: mín, máx y avg, aunque

solo usamos el promedio en el modelo para que no se confunda con tantas variables numéricas.

A su vez al analizar los datos en general tenía un 26,9% de missing data, lo cual es bastante y hay que tratarlo de alguna forma.

Luego lo hicimos a través del widget de select columns y decidimos que la variable target debía ser si eran o no venenosos. A su vez quitamos variables como nombre ya que al ser un metadato no servirá para el análisis y quitamos otras variables que estaban casi vacías para no entorpecer el análisis.

Luego lo que hicimos fue pasar el dataset por un preprocesamiento en el cual las variables que tenían datos faltantes fueron rellenadas por los valores promedios.

Finalmente partimos el modelo en dos partes un 80% y un 20% y lo entrenamos con 3 modelos (al 80%):

- Árbol
- Regresión logística
- Red neuronal

Lo conectamos a test and score y obtuvimos los siguientes datos considerando que el hongo sea venenoso como clase positiva

Modelo	Accuracy	F1	Precision	Recall
Tree	0,583	0,613	0,630	0,597
Logistic regression	0,612	0,649	0,649	0,649
Neural Network	0,633	0,679	0,659	0,701

Decidimos utilizar stem-color ya que de las variables usadas fue la que mejores métricas en general lograba. Como uno puede observar este modelo es bastante malo, con métricas cercanas al 0,5 es casi un juego de azar más que un modelo de adivinanza, con lo cual si tuviésemos que confiar en que este modelo se predice si un hongo es comestible o no, lo más probable es que el que consuma ese hongo acabe envenenado.

Tomando los valores de la red neuronal para hacer el análisis de métricas (dado a que es el que mejor valores tiene) hacemos el siguiente análisis:

- **Accuracy:** Con un valor de 0,633 el porcentaje total de predicciones correctas está en un 63,3% lo que indica que, si bien no es tan aleatorio como un 50:50 de tirar una moneda, sigue siendo bastante malo el porcentaje de predicciones. Mucho más teniendo en cuenta que la consumición de algún hongo venenoso podría matar a alguien.
- **F1:** Dado a que el valor es 0,679 estamos hablando de que hay un cierto balance entre precisión y recall, pero el modelo puede estar puede estar dejando pasar algunos casos importantes (bajo recall) o clasificando incorrectamente ejemplos (baja precisión).
- **Precisión:** Con una precisión del 0,659 esto indica que el modelo tiene un cierto margen de error con respecto a los falsos positivos y que a veces indica que un hongo comestible es venenoso. Como tal esto no es un problema ya que no daña

a nadie, el único impacto que tendría seria en que es un desperdicio tirar algo comestible pensando que era venenoso, pero más allá de eso no hay problema con que el modelo identifique negativos como positivos ya que no atenta contra la salud de nadie.

- **Recall:** Con recall del 0,659 este modelo identifica a los venenosos como tal un 65,9% de las veces y las otras veces lo predice cómo comestibles. Esto indica que el modelo no es muy confiable, ya que el 34,1% de las veces va a pedir un hongo venenoso como comestible y puede que llegue a dañar a alguien o incluso llegar a matarlo.

En general es un modelo decente, pero dado a que tienen un margen de error importante al clasificar hongos venenosos como comestibles no lo usaría ya que puede ser peligroso. Sin embargo, puede llegar a ocurrir que el modelo no se desempeñe bien dado a underfitting ya que tiene muy pocos datos (173) para la gran cantidad de variables y la complejidad del dataset en general, con lo cual quizás con más datos podría llegar a ser un modelo decente.

Descripción de secondary

Al igual que primary, secondary también tiene ciertas variables numéricas:

- Cap diameter
- Stem-height
- Stem-width

Sin embargo y para nuestro alivio esta vez no están separadas en intervalos, lo que simplifica el análisis ya que no tenemos que pasarlo por el código de python.

Por su parte secondary tiene un 19,6% de missing data, con lo cual si bien sigue siendo un número considerable no es tanto como primary, al menos porcentualmente ya que numéricamente al ser mucho más grande faltan muchos más datos que en primary.

Luego pasamos los datos por el módulo de select-columns para seleccionar la variable target quitar del análisis variables casi vacías de datos. Nuevamente el target fue si el hongo es comestible (negativo) o venenoso (positivo). Las variables que quitamos fueron gill-spacing y spore-sprint-color ya que eran columnas prácticamente vacías.

Seguido a esto aplicamos el pre-procesamiento para rellenar los otros datos faltantes con valores promedios.

Finalmente aplicamos nuevamente el data sampler y enviamos un 80% para el entrenado del modelo y un 20% para datos de prueba con los mismos tres modelos:

- Árbol
- Regresión logística
- Red neuronal

Y a raíz de esto obtuvimos los siguientes datos y métricas:

Modelo	Accuracy	F1	Precision	Recall
Tree	0,981	0,983	0,987	0,979
Logistic regression	0,835	0,849	0,862	0,836
Neural Network	0,998	0,999	0,999	0,998

Incluso a primera vista el peor modelo de secondary es más confiable que el mejor modelo de primary. Al igual que antes usaremos el modelo más confiable para hacer el análisis, en este caso Neural Network.

- **Accuracy:** Este modelo es totalmente preciso, muy cercano al 100% su tasa de acierto, con lo cual es muy bueno para realizar predicciones.
- **F1:** Es casi 1, lo que indica que hay un gran balance entre precisión y recall, lo que significa que es un análisis que no está sesgado por una métrica o la otra.
- **Precisión:** Al ser de 0,999 este modelo casi no identifica hongos comestibles como si fueran venenosos. Con lo cual incluso si era un problema menor tener que tirar estos hongos en el modelo de primary, con este modelo incluso ese problema menor desaparece.
- **Recall:** Al tener un recall del 0,998 esto indica que el 99,8% de las veces los hongos venenosos son clasificados como venenosos y solo el 0,02% de las veces un hongo venenoso es identificado como comestible. Esto significa que el modelo es extremadamente bueno para clasificar lo venenoso como tal y si hubiese que confiar en que si el hongo es comestible o no dado a que el modelo lo clasificó como tal, uno podría confiar sin problemas dado el pequeño margen de error. Con lo cual en este caso donde los falsos negativos pueden ser mortales es un excelente modelo.

En general este modelo es preciso casi al 100%, lo que indica que uno puede confiar en la clasificación que realiza. En todo aspecto supera al modelo de primary, y si tuviese que elegir este modelo para implementarlo no dudaría. Sin embargo si tuviese que decir una diferencia con el modelo de primary que puede llegar a ser negativo puede ser que el modelo sufra de overfitting, dado al gran volumen de datos y que se evidenciaría con las métricas tan precisas. Esto puede llegar a tener un impacto negativo a la hora de analizar un set de datos nuevo ya que puede que el modelo no realice las predicciones correspondientes.

Interpretación de las predicciones

Primary

Nuestras hipótesis iniciales señalaban que variables como does-bruise-or-bleed, cap-color, ring-type y hábitat serían relevantes para predecir si un hongo es venenoso.

Al aplicar un análisis de importancia de atributos mediante el widget Rank en Orange, observamos que ring-type y hábitat sí tienen un peso importante en el modelo, lo cual apoya parcialmente nuestras hipótesis.

Sin embargo, variables como does-bruise-or-bleed y cap-color, que esperábamos fueran significativas, resultaron tener baja o nula importancia en la predicción, lo cual sugiere que su relación con la toxicidad podría no ser tan clara en el conjunto de datos utilizado.

Con lo cual si bien puede llegar a ser variables usadas para predecir la toxicidad de un hongo, el modelo no lo consideró como relevante.

Secondary

Nuevamente creíamos que las variables que predecían si un hongo es venenoso o no serían does-bruise-or-bleed, cap-color, ring-type y hábitat.

En este segundo conjunto de datos, observamos que las variables ring-type y habitat, planteadas en nuestra hipótesis, vuelven a aparecer con una relevancia significativa, lo

cual refuerza su posible vínculo con la toxicidad en hongos. También aparece cap-color, aunque con una importancia baja, lo cual sugiere una relación débil con la variable objetivo.

Por otro lado, la variable does-bruise-or-bleed, que considerábamos clave, no aparece entre las más relevantes en ninguno de los modelos, indicando que en este conjunto de datos no aporta significativamente a la predicción.

Finalmente, variables no consideradas en nuestra hipótesis de el dataset de primary cómo stem-root, season o stem-surface mostraron un mayor peso predictivo, lo que nos invita a reconsiderar su relevancia biológica en el análisis.

Conclusión

A lo largo del proyecto, analizamos dos bases de datos diferentes (primary y secondary) con el objetivo de desarrollar un modelo confiable para clasificar hongos como comestibles o venenosos, lo cual es crucial ante el riesgo sanitario de vender hongos tóxicos.

Partimos de una hipótesis basada en el conocimiento biológico y visual de los hongos, identificando variables como does-bruise-or-bleed, cap-color, ring-type y hábitat como potencialmente clave para la predicción. Sin embargo, el análisis de importancia de atributos con el widget Rank en Orange y la evaluación de los modelos entrenados revelaron una realidad más compleja.

En el dataset primary, observamos un modelo con un rendimiento bajo (accuracy de apenas 63% en el mejor caso), y un importante porcentaje de datos faltantes. En este caso, sólo dos de nuestras variables hipótesis (ring-type y hábitat) mostraron relevancia predictiva. Otras como does-bruise-or-bleed o cap-color, que esperábamos fueran importantes, no aportaron significativamente al modelo.

En contraste, el dataset secondary, mucho más grande y limpio, permitió construir modelos con un rendimiento altísimo, especialmente la red neuronal (accuracy del 99,8%). Aquí, ring-type y hábitat volvieron a aparecer como variables relevantes, lo cual fortalece nuestra hipótesis inicial en parte. Sorprendentemente, variables como veil-color y stem-root, que no habíamos considerado, resultaron ser de alta importancia, mientras que does-bruise-or-bleed, nuevamente, fue insignificante.

Es probable que secondary sufra de un overfitting y primary de un underfitting o simplemente que las variables que no usamos en primary cómo veil-color realmente hayan impactado en la capacidad predictiva del modelo primary.