Informe Final - Proyecto de Programación

Nombres de los estudiantes: Nicolás Uribe Arboleda, Valeria Orozco Gómez y Verónica Restrepo García

Curso: Programación

Profesor: Andrés Quintero

Fecha: 31.05.2025

1. Introducción

Este proyecto lo hicimos con el objetivo de poner en práctica lo que aprendimos durante el curso de programación, usando Python y algunas librerías como pandas, sklearn, seaborn y matplotlib. Trabajamos con un conjunto de datos real sobre hongos, llamado "agaricus-lepiota", que tiene información sobre distintas características de los hongos y si son comestibles o venenosos. La idea fue predecir si un hongo es comestible o no, basándonos en esas características.

2. Exploración de datos

Primero cargamos el dataset y miramos sus primeras filas para entender cómo venían los datos. Vimos que había muchas columnas con palabras (variables categóricas), como la forma del sombrero, el olor, el color, entre otras. También hicimos algunas gráficas para ver mejor cómo estaban distribuidos los datos, especialmente la variable que indica si el hongo es venenoso o no.

Nos dimos cuenta de que esa variable está marcada como "e" para comestible y "p" para venenoso. Además, notamos que había algunas casillas con signos de pregunta (?) que luego tuvimos que manejar.

3. Preprocesamiento

En esta parte limpiamos los datos. Lo primero fue quitar las filas que tenían valores faltantes (los signos de pregunta). Después, como casi todo estaba en texto, usamos una herramienta llamada LabelEncoder para convertir esos textos en números. Finalmente, dividimos los datos en dos partes: una para entrenar el modelo y otra para probarlo, usando 80% para entrenar y 20% para probar.

4. Modelos implementados

Modelo 1: Árbol de Decisión

Este fue nuestro primer modelo. Usamos una clase de sklearn llamada DecisionTreeClassifier. Entrenamos el modelo con los datos de entrenamiento y lo evaluamos con los de prueba. El modelo funcionó bien, aunque puede que se ajuste demasiado a los datos. Calculamos la precisión y también sacamos la matriz de confusión.

Modelo 2: Random Forest

Este modelo es como una versión mejorada del anterior. En vez de un solo árbol, hace muchos árboles y decide por mayoría. Usamos RandomForestClassifier con 100 árboles. Nos dio un resultado mejor que el modelo anterior y también hicimos el reporte de clasificación.

5. Comparación de modelos

Al final comparamos los dos modelos. El Random Forest fue el que mejor funcionó. Mostramos la

precisión de cada uno en esta tabla:

Modelo	Precisión
Árbol de Decisión	0.95
Random Forest	0.99

Por eso, creemos que el Random Forest es más confiable para este tipo de datos.

6. Conclusiones

Con este proyecto aprendimos a manejar datos reales, a limpiarlos y prepararlos para entrenar modelos. También entendimos cómo funcionan dos algoritmos de clasificación y cómo comparar sus resultados. Aunque todavía no dominamos todo, sentimos que logramos aplicar lo que vimos en clase de una manera práctica y que nuestros códigos funcionaron bien. Fue interesante ver cómo con la programación se pueden hacer predicciones útiles a partir de los datos.