Implementación del Método de Árbol de Decisión

Nadir Madahi Apaza Escalante[1], Wilber Santiago Flores Vidal[2], Maria Belen Franco Ovando[3], and Ahmed Israel Ruiz Soliz[4]

Abstract. The abstract should briefly summarize the contents of the paper in 150–250 words.

Keywords: First keyword · Clothes · Datasets

1 Introducción

En el presente trabajo se analiza datos de precios y costos de tres categorías de una tienda dedicada a la venta de artículos de ciclismo. El análisis tiene el objetivo implementar un método predictivo para elegir la categoria que genere beneficios a la empresa y no así costos extras.

1.1 Estado del Arte

Realizando una busqueda en "Google Academy" con las siguientes palabras claves:

- CNN
- Clothes

Obtuvimos por resutados bastantes artículos. De todos ellos, seleccionamos por el título solamente los siguientes artículos:

- LEITHARDT, V. (2021). Classifying garments from fashion-MNIST dataset through CNNs. Advances in Science, Technology and Engineering Systems Journal, 6(1), 989-994.

Este artículo detalla cuatro modelos diferentes de redes neuronales convolucionales (CNN) para clasificar prendas a partir del conjunto de datos Fashion-MNIST. Los modelos se comparan con los resultados originales, y uno de ellos alcanza una precisión del 99,1%. El uso de CNN y técnicas de abandono ayuda a mitigar el sobreajuste y mejorar el rendimiento. Los investigadores concluyen que las CNN pueden ser clasificadores más precisos para productos de moda y que el uso de TensorFlow 2 y la GPU para el entrenamiento puede mejorar tanto el tiempo de entrenamiento como la precisión.

- Li, M., Ji, S., & Liu, G. (2018). Forecasting of Chinese E-commerce sales: an empirical comparison of ARIMA, nonlinear autoregressive neural network, and a combined ARIMA-NARNN model. Mathematical Problems in Engineering, 2018, 1-12.

Podemos apreciar la comparación de tres modelos de previsión de las ventas del comercio electrónico chino: ARIMA, red neuronal autorregresiva no lineal (NARNN) y un modelo combinado ARIMA-NARNN. El estudio concluye que el modelo ARIMA-NARNN es más eficaz para predecir las ventas del comercio electrónico que los otros dos modelos. El artículo concluye que una previsión precisa y eficaz de las ventas del comercio electrónico es crucial para las estrategias de inventario y compras del sector.

– Korobko, A. V., Korobko, A. A., & Markovin, A. V. (2021). Creating catalogues of clothes images using neural networks. In CEUR Workshop Proceedings (pp. 1-9).

Este artículo se centra en la aplicación práctica de redes neuronales para la catalogación automática de imágenes de ropa procedentes de cuentas comerciales de Instagram. Los autores describen los métodos y enfoques utilizados para construir modelos de redes neuronales convolucionales para determinar el color y la categoría de la ropa a partir de sus imágenes. También analizan la precisión de los modelos y los comparan con la clasificación aleatoria. El estudio pretende sentar las bases para futuras investigaciones y establecer valores de referencia para la precisión de la clasificación.

Luego de revisar todos los artículos, vemos conveniente añadir las siguientes palabras claves:

– Clothes
– Datasets
– Árbol de Decisión

1.2 Justificación

El presente trabajo es en beneficio a una tienda de productos exclusivamente de ciclismo, la misma que requiere un código en R que facilite la predicción y elección de una de sus tres categorias de productos (Accesorios, Bicicletas y Ropa), para que traiga a la tienda mayores ganancias y menos costos.

2 Materiales y Métodos

Los materiales se refiere al set de datos y los métodos a la forma de análisis de dicho dataset para conseguir un resultado y el propósito de este artículo científico.

2.1 Materiales

El dataset implementado en este trabajo fue proporcionado en el módulo de "Nuevos Topicos de Logística I" del Diplomado de Logística Integral Articulando la Cadena de Suministros con Analítica de Datos, a cargo del Ing. Maurcio García Soria Galvarro.

Se seleccionaron cuatro documentos Excel que corresponden a “Ventas 2018”, “Ventas 2019”, “Ventas 2020” y “Lista de Productos”. Estos documentos fueron transformados en la plataforma Power BI, donde se vio conveniente eliminar en los documentos de Ventas las columnas irrelevantes como “Territorio Venta ID” y “Distribuidor ID”, dado que no son necesarios para la resolución del trabajo. Se realizaron arreglos en la columna de “Fecha de Envio” por existencia de datos en blanco, y en la columna de “Cliente ID” por contener un cliente con ID negativo.

Por otro lado, para el documento de “Lista de Productos”, se eliminaron las columnas de “Color” y “Modelo” debido a que la columna Descripcion Producto contiene la misma información. También se eliminó “Subcategoria”. Una vez conseguida la Tabla de Data.xlsx, se procedió a eliminar las columnas que contengan texto; como ser “Orden de venta ID”, “Cliente ID”, “Año”, “Mes”, “Dia”, “SKU” y “Descripción”, para que solo existiera la columna de “Categoria” como variable de carácter.

De esta manera, resultó un sólo documento en formato xlsx para trabajar en RStudio con las columnas numéricas de “Costo” y “Precio”.

2.2 Métodos

Existen bastantes métodos que pueden aplicarse para el análisis de datos extensos, sólo depende de lo que se quiera lograr; como etiquetar, pronosticar, clasificar. etc. Considerando la estructura del dataset que obtuvimos, nuestra elección fue el método de Árbol para determinar...

El método de Árbol de Decisión es un algoritmo de aprendizaje supervisado no paramétrico que refleja de manera gráfica los posibles resultados, costos, probabilidades, beneficios o consecuencias comparando éstas entre si. Su estructura se basa en un nodo raíz del cual salen ramificaciones siendo los nodos de decisión y de ellos; a su vez, surgen mas ramificaciones. Los nodos terminales son las ultimas ramificaciones mostrando el resultado definitivo de una ruta de decisión. El Árbol de Decisiones emplea una estrategia que va buscando e identificando los puntos de división óptimos dentro del árbol recursivamente, hasta que todos o la mayoria de los datos se hayan clasificado homogénamente bajo etiquetas.

Algunos de sus fines son calcular valores esperados, comparar resultados para determinar el mejor plan de acción, resolver problemas, gestionar costos, pronosticar resultados y situaciones, optimizar estrategias o identificar oportunidades.

Esta técnica sigue el Teoréma de Bayes,dado que, es utilizado para calcular la probabilidad de un suceso condicionado a la información de otro suceso, afectando al primero.

La fórmula del Teoréma de Bayes se define matemáticamente como:

P[An/B] = P[B/An]*P[An] / Sumatoria*(P[B/Ai]*P[Ai])

Donde B es el suceso sobre el que tenemos información previa y A(n) son los distintos sucesos condicionados. La parte del numerador corresponde a la probabilidad condicionada, y en la parte del denominador la probabilidad total.

3 Caso de Estudio

A continuacion, se muestra los datos usados para la ejecución de este caso:

3.1 Datos

Muestra de los datos de Costos y Precios de la Tienda de Ciclismo:

Costo <dbl>	Precio <dbl>	Categoria <fctr>
13.09	34.99	Accesorios
1265.62	2319.99	Bicicletas
8.22	21.98	Accesorios
41.57	53.99	Ropa
1.87	4.99	Accesorios
10.84	28.99	Accesorios
1481.94	2384.07	Bicicletas
13.09	34.99	Accesorios
6.92	8.99	Ropa
1265.62	2319.99	Bicicletas

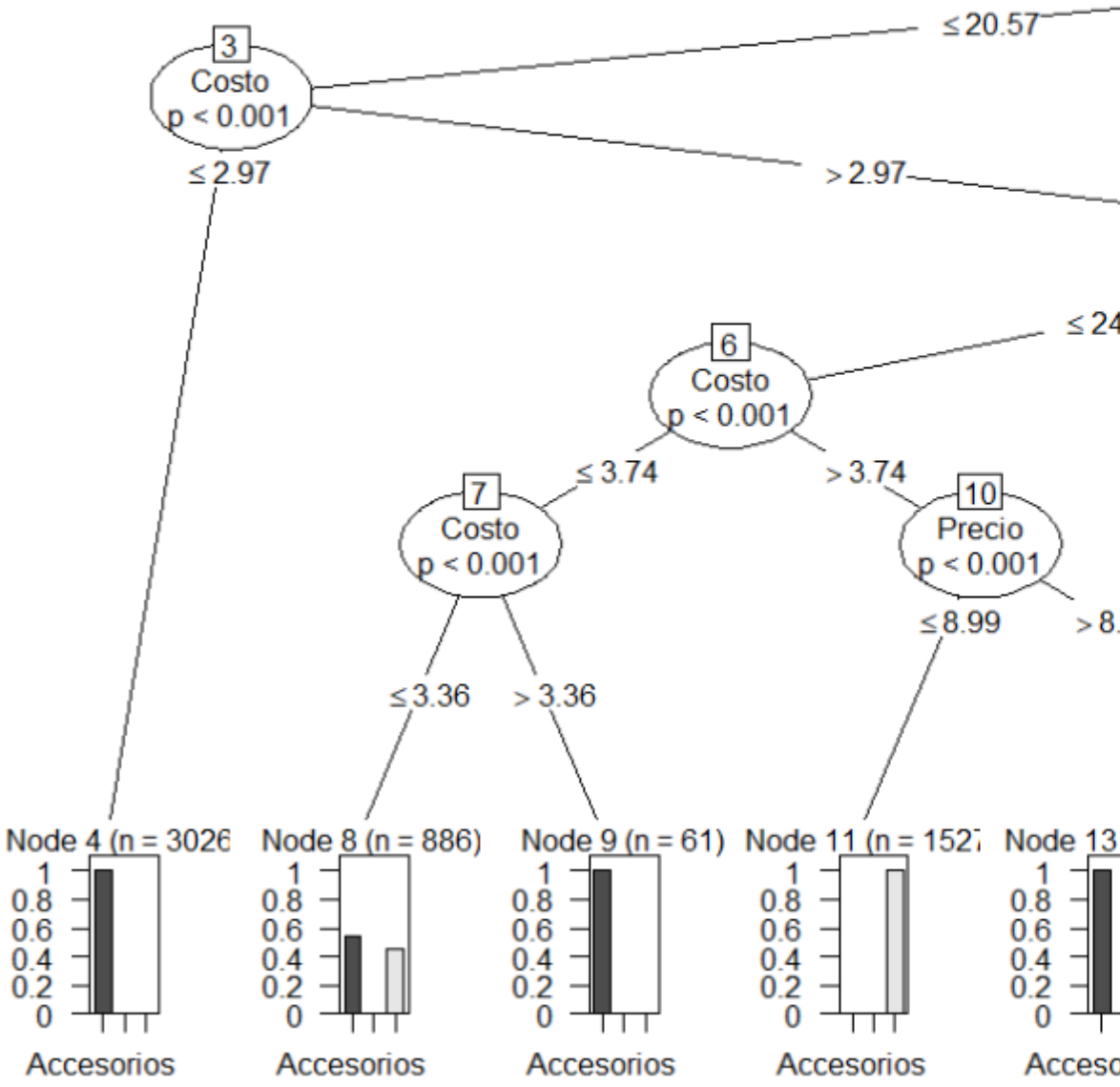
3.2 Matriz de Confusión

La matriz de confusion denota que la predicción fue bastante buena; ya que la diagonal principal presenta cantidades grandes. Sin embargo, hubo 404 predicciones erróneas al establecer Ropa y no Accesorios, estos se refiere a falsos positivos.

	Accesorios	Bicicletas	Ropa
Accesorios	6487	0	404
Bicicletas	0	6351	0
Ropa	0	0	5927

3.3 Árbol de Decisión

Con el codigo implementado (Arbol-Ventas), el árbol de decisión se muestra de la siguiente manera:


3
Costo
p < 0.001
≤ 20.57
≤ 2.97
> 2.97
6
Costo
p < 0.001
≤ 24
≤ 3.74
> 3.74
7
Costo
p < 0.001
10
Precio
p < 0.001
≤ 8.99
> 8.
≤ 3.36
> 3.36
Node 4 (n = 3026
Node 8 (n = 886)
Node 9 (n = 61)
Node 11 (n = 152
Node 13
1
0.8
0.6
0.4
0.2
0
Accesorios
Acceso

4 Interpretasión de Resultados

Las condiciones a las que somete el árbol de decisión son:

```
           Conditional inference tree with 10 terminal nodes

Response:  Categoria
Inputs:  Costo, Precio
Number of observations:  19169

1) Costo <= 59.47; criterion = 1, statistic = 13557.217
  2) Costo <= 20.57; criterion = 1, statistic = 3887.697
    3) Costo <= 2.97; criterion = 1, statistic = 94.186
      4)*  weights = 3026
    3) Costo > 2.97
      5) Precio <= 24.49; criterion = 1, statistic = 2250.232
        6) Costo <= 3.74; criterion = 1, statistic = 492.043
          7) Costo <= 3.36; criterion = 1, statistic = 48.458
            8)*  weights = 886
          7) Costo > 3.36
            9)*  weights = 61
        6) Costo > 3.74
          10) Precio <= 8.99; criterion = 1, statistic = 279.028
            11)*  weights = 1527
          10) Precio > 8.99
            12) Costo <= 8.22; criterion = 1, statistic = 1440.623
              13)*  weights = 437
            12) Costo > 8.22
              14)*  weights = 1008
      5) Precio > 24.49
        15)*  weights = 2405
  2) Costo > 20.57
    16) Precio <= 69.99; criterion = 1, statistic = 2377.933
      17)*  weights = 3392
    16) Precio > 69.99
      18)*  weights = 76
1) Costo > 59.47
  19)*  weights = 6351
```

Estas condiciones dependen de los valores de costos y precios.

5 Conclusiones

El árbol de decisiones determino que la categoría Accesorios tendria una cantidad mayor de predicciones correctas. Sin embargo, también algunos falsos positivos que podrian ser mejorados con un análisis de los datos o la implementación de otras herramientas para mitigar éstos. Por otro lado, las categorias de Bicicletas y Ropa, obtuvieron cantidades de predicciones buenas y ningún falso postivo.

6 Bibliografía:

José Francisco López. (2023). Teorema de Bayes - Definición, qué es y concepto | Economipedia. Economipedia. https://economipedia.com/definiciones/teorema-de-bayes.html

¿Qué es un árbol de decisión? | IBM. (2023). Ibm.com. https://www.ibm.com/es-es/topics/decision-trees#:~:text=Un%20%C3%A1rbol%20de%20decisi%C3%B3n%20es,nodos%20inte

Team Asana. (2023, February 27). El árbol de decisiones: un análisis de 5 pasos para tomar mejores decisiones [2023] • Asana. Asana; Asana. https://asana.com/es/resources/decision-tree-analysis

Qué es un diagrama de árbol de decisión. (2023). Lucidchart. https://www.lucidchart.com/pages/es/que-es-un-diagrama-de-arbol-de-decision