

Arboles de Decisión.

Instituto Politécnico

Nacional ESCOM

**Profesora. Dr. Fabiola Ocampo Botello.  
Data Mining  
  
Martínez López Sebastián  
Ramírez Reséndiz Luis Roque  
Rodriguez Yepez Jaasiel   
Fecha: 05 de Noviembre del 2022.**

Todos nuestros datos fueron obtenidos de la plataforma Kaggle, la cual contine una amplia variedad y gama de datasets para poder realizar los diversos arboles de decisión. A continuación, se agrega la dirección URL del dataset obtenido para la realización del árbol CART.

<https://www.kaggle.com/competitions/titanic>

Diccionario de Datos:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Definición** | **Dominio de Valores** | **Tipo de Dato** |
| survival | Sobreviviente | 0 = No  1 = Si | Nominal |
| pclass | Tipo de Boleto | 1 = 1st 2 = 2nd  3 = 3rd | Ordinal |
| sex | Sexo |  | Nominal |
| age | Edad |  | Cuantitativo |
| sibsp | # de hermanos y cónyuges abordo |  | Cuantitativo |
| parch | # de parientes y de hijos abordo |  | Cuantitativo |
| ticket | Numero de Boleto |  | Nominal |
| fare | Tarifa |  | Continua |
| cabin | Numero de cabina |  |  |
| embarked | Puerto donde abordo | C = Cherbourg Q = Queenstown S = Southampton | Nominal |

Árbol CART

Introducción:

CART es un algoritmo basado en un árbol que funciona examinando diversas maneras de particionar o dividir localmente los datos en segmentos más pequeños con base en diferentes valores y combinaciones de predictores. Es decir, se encarga de hacer subdivisiones. CART selecciona las divisiones de mejor rendimiento y luego repite este proceso de forma recursiva hasta encontrar el conjunto optimo. De otra manera el algoritmo repite el procedimiento hasta cumplir con una condición de paro o hasta que ya no puede continuar haciendo más divisiones. El resultado es un árbol de decisión representado por una serie de divisiones binarias que conducen a nodos terminales que pueden ser descritos por un conjunto de reglas específicas. El árbol y su diseño son visualmente estimulantes y se pueden interpretar de manera intuitiva, de, modo que no tienen que ser un científico de datos para entenderlo y obtener información de él. Estableciendo así que CART es bastante sencillo de comprender y muy intuitivo de interpretar.

Para la realización de este árbol de decisión se utilizo el conjunto de datos *“Titanic – Machine Learning from Disaster”*, el cual se obtuvo del siguiente enlace: <https://www.kaggle.com/competitions/titanic>

De este conjunto de datos nuestra clase objetivo fue la “survival”, es decir ver si sobrevive o no, por medio del árbol de decisión, en el cual se puede ver los porcentajes y las probabilidades de sobrevivir dependiendo del género, del tipo de boleto, incluso de la tarifa que se pagó.

Objetivo:

Descripción del estudio a realizar

Generar un modelo de clasificación por medio de un árbol de decisión el cual proporcione las posibilidades de sobrevivir dependiendo de los diferentes atributos de cada persona, ver la exactitud del algoritmo CART y así poder implementarlo en otro tipo de desastres verificando así si la tripulación o la muerte tiene cierta preferencia o no. Y si es así cuales son los atributos que podrían seleccionar como base para la toma de decisión o cuales son los que mas influyen en esto.

Como atributo objetivo se considera Si sobrevivió (survival) o no, el cual tiene dos clases: si y no. Los atributos independientes son: Passenerid, Pclass, Name, Sex, Age, SibSP, Parch, Ticket, Fare, Cabin, Embarked.

Descripción del estudio a realizar

Crear un modelo de generalización y las reglas de decisión que nos permita conocer, saber cuáles son las características de las diversas personas que sobrevivieron y las que no lo hicieron para utilizarlo en la identificación de muestras futuras.

Diccionario de Datos:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Definición** | **Dominio de Valores** | **Tipo de Dato** |
| survival | Sobreviviente | 0 = No  1 = Si | Nominal |
| pclass | Tipo de Boleto | 1 = 1st 2 = 2nd  3 = 3rd | Ordinal |
| sex | Sexo |  | Nominal |
| age | Edad |  | Cuantitativo |
| sibsp | # de hermanos y cónyuges abordo |  | Cuantitativo |
| parch | # de parientes y de hijos abordo |  | Cuantitativo |
| ticket | Numero de Boleto |  | Nominal |
| fare | Tarifa |  | Continua |
| cabin | Numero de cabina |  |  |
| embarked | Puerto donde abordo | C = Cherbourg Q = Queenstown S = Southampton | Nominal |

Árbol:

Captura de pantalla de computadora

Descripción generada automáticamente

Figura 1.1. Parte del árbol generado por el programa.

En la figura 1.1 se puede apreciar solo uan pequeña parde de todo el árbol generado, en esta primera parte observamos que el atributo que considera le da más ganancia de información es el género, es decir, el atributo Sex es el atributo el cual aporta mas respecto a la decisión de si sobrevivirá o no. La interpretación de esta primera parte del árbol quiere decir lo siguiente.

Los hombres tienes mas probabilidades de morir que las mujeres, para ser exactos la probabilidad que tiene un hombre de morir es de 81.1% y la probabilidad de una mujer de sobrevivir es de 74.2%, a partir de esto el árbol sigue tomando mas atributos los cuales considera le dan ms ganancia de información y va mostrando los diversos porcentajes obtenidos.

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

Figura 1.2. Parte del árbol generado.

En la figura 1.2 se muestra como a partir de la primera “rama” generada por el árbol del lado derecho al tomar el atributo de mujer, se observa que el siguiente atributo que le otorga más ganancia de información es el topo de boleto (Pclass). La interpretación que se le puede dar es que las mujeres que tenían boletos de primera clase tienen mas probabilidades de sobrevivir que las mujeres que no los tienen. Y que el porcentaje de probabilidad de sobrevivir es de 94.7%.

Medidas:

-Matriz de Confusión.

Captura de pantalla de computadora

Descripción generada automáticamente

Figura1.3 Matriz de Confusión.

La figura 1.3 es la matriz de confusión generada por el modelo, la cual indica las predicciones que realizo correctamente y las que no, por ejemplo, de los que no sobrevivieron tuvo 29 errores de clasificación y de los que si sobrevivieron tuvo 35 errores.

Una captura de pantalla de una red social

Descripción generada automáticamente

Figura 1.4. Distintos datos obtenidos del método de clasificación.

Interfaz de usuario gráfica, Aplicación, Tabla, Excel

Descripción generada automáticamente

Figura 1.5 Distintas medidas obtenidas.

- Sensibilidad para 0 es de 0.947

La sensibilidad se interpreta como la capacidad del predictor de clasificar correctamente, es decir que tiene una sensibilidad del 0.947 para decir que alguien realmente no sobrevivió, es decir que hizo una predicción correcta nos indica la capacidad de nuestro estimador para dar como casos positivos. Es decir, la capacidad de decir que la persona no sobrevivirá y esta no sobreviva.

- Precisión para 0 es de 0.937, para 1 es de 0.914

Se observa que la precisión es mas alta en el caso de las personas que no sobrevivieron, de igual manera ambas precisiones son muy buenas porque se acercan mucho a 1, el cual quiere decir que la precisión es del 100%

- Tasa de error 7.183%

La tasa de error fue de 7.183%, lo que quiere decir que de los 891 registros solo clasifico mal 64

- Exactitud 92.817%

La exactitud denota ser buena ya que es una exactitud superior al 90%, es decir que tiene una confiabilidad alta de predecir o clasificar correctamente, aunque hay que considerar que el margen de error es menor al 10%, es decir que tiene un sesgo pequeño.

- Especificidad para 0 es de 0.898

Esta nos indica la capacidad de nuestro estimador para dar como casos negativos, es decir se interpreta como la capacidad del predictor de clasicr correctamente, es decir que tiene una especificidad del 0.898 para decir que alguien realmente sobrevivió, es decir que hizo una predicción correcta. Es decir, la capacidad de decir que la persona sobrevivirá y esta sobreviva.

- Explicación de los positivos verdaderos y positivos falsos (520, 35)

Se aprecia que los positivos verdaderos fueron 520 de 555, es decir que solo clasifico mal 35, es decir que el total debería de ser 555 y no 520.

Conclusiones:

De esto se concluye que nuestro modelo es muy preciso y tiene un sesgo pequeño, de igual manera la interpretación que se puede dar es que las mujeres tienen mas probabilidades de sobrevivir que los hombres, esto se le puede atribuir a que en caso de emergencia siempre se dice “primero mujeres y niños”, ahora lo siguiente es que de las mujeres las que tienen mas probabilidades de sobrevivir son las que tienen una mejor posición económica en otras palabras son las que tienen boletos de primera clase, esto no es un misterio para nadie, ya que la sociedad siempre ha sido clasista y esto nos da a entender que los que tienen más tienes más y mejores opciones en la vida, es duro decirlo y admitirlo pero entre más poder económico tengas mejor te ira en la vida.

Anexos:

ARBOL CART

Captura de pantalla de computadora

Descripción generada automáticamente

Nodo Number To String

Una captura de pantalla de una red social

Descripción generada automáticamente

Nodo Decision Tree Learner

Captura de pantalla de computadora

Descripción generada automáticamente

Nodo Decision Tree Predictor

Una captura de pantalla de una red social

Descripción generada automáticamente

Nodo Scorer

Una captura de pantalla de una red social

Descripción generada automáticamente