# Resolución del Problema

En este apartado se detallara la metodología seguida para resolver el problema planteado.

Luego de plantear el problema y de analizar las variables que están involucradas en él, se determinó que un posible algoritmo que permita su correcta resolución seria el árbol de decisión ya que este nos permite examinar los datos e inducir las reglas para realizar predicciones, creando un aprendizaje que nos permita alcanzar decisiones lo más acertadas posibles. En el anexo 1 se puede encontrar un resumen con los puntos más importantes de este algoritmo.

Luego de decidir cuál sería el algoritmo a utilizar, ha surgido la posibilidad de plantearlo como un árbol de clasificación o un árbol de regresión. A partir de la definición de estos dos términos se ha optado por el primero ya que este se utiliza para predecir variables categóricas, que son las que nos atañen en este trabajo.

La implementación que se utilizó es la del algoritmo id3 detallado en el Anexo 1. Para esta implementación se decidió usar como lenguaje programático Java, el cual nos provee de una estructura de objetos que nos permite controlar el árbol de decisión y sus componentes fácilmente.

Para el diseño programático del algoritmo se tomó como premisas que para la etapa de aprendizaje es necesario la obtención de los datos de muestra en formato de tabla o matriz. Esta debería contener la cantidad de columnas necesarias para alojar los atributos que determinan una cierta categoría, la cual se encontrara en la última columna de la matriz. Para el ingreso de estos datos al sistema se optó por cargarlos en una tabla de Excel.

Otra de las premisas planteadas es que el árbol de decisión no se genere totalmente desde que se obtiene la muestra, sino que sus nodos se generen a medida que sea necesario profundizar en uno de ellos. Esto le provee al algoritmo mayor eficiencia ya que no desperdicia tiempo en crear nodos del árbol que quizás nunca sean requeridos para categorizar un conjunto de atributos.

Una vez planteadas estas premisas se pasó a separar al algoritmo en 4 capas, la primera para la obtención de los datos del documento Excel, la segunda para la realización de los cálculos de entropía e información de ganancia, la tercera para la creación del árbol a partir de los cálculos proveídos por la segunda capa y finalmente la cuarta para obtener la categoría de un conjunto de datos a partir del árbol generado en la capa anterior.

Para las primeras 2 capas se plantearon las siguientes clases:

* La clase **Matriz** que se compone de:
  + Un vector de columnas
    - Cada Columna representa una columna de atributos del Excel.
  + Una columna llamada “categorías”
    - Es la columna llamada categoría en el Excel.
  + El método Calcular Entropía
    - Calcula la entropía de una columna específica respecto a las categorías y un valor posible.
  + El método generar ganancias de información
    - Obtiene la ganancia de información de las columnas.
  + El método Obtener Resultado
    - Devuelve, según un valor ingresado, la categoría a la que pertenece según la columna de mayor ganancia de información o si hay más de una categoría para ese valor, genera las SubMatrices necesarias.
  + El método generar SubMatriz
    - Genera una matriz más pequeña, cuyas filas corresponden a aquellas donde un valor pasado como parámetro aparece en la columna de mayor ganancia.
* La clase **Columna** se compone de:
  + Un Nombre
    - Es el nombre de la columna en el Excel.
  + Un vector de valores
    - Son los valores de la columna en el Excel.

Para las capas 3 y 4 se plantearon las siguientes clases:

* La clase **Resultado**(clase abstracta), sería la categoría, se compone de:
  + Nombre de columna
    - Es el nombre de la columna en por la cual discriminamos para obtener la categoría.
  + Valor de columna
    - Es el valor en la columna discriminadora del cual queremos obtener la categoría.
  + Método Obtener Categoría
    - Método que va a ser implementado por las clases **ResultadoFinal** o **Árbol.**
* La clase **ResultadoFinal** (extiende de Resultado o sea que contiene todos los atributos de esta) se compone de:
  + La categoría
    - Es la categoría para la columna y el valor especificado.
  + Método Obtener Categoría
    - Devuelve el atributo categoría
* La clase **Árbol** (extiende de Resultado o sea que contiene todos los atributos de esta) se compone de:
  + Una matriz
    - Puede ser:
      * La matriz que se obtuvo del Excel, o
      * Una matriz obtenida luego de simplificar otra más grande a través del método **obtener SubMatriz** de la clase **Matriz**.
  + Una columna que es la de mayor ganancia
  + Un vector de resultados (serían los nodos del árbol)
    - Contiene los posibles resultados para cada valor de la columna de mayor ganancia
  + Método Obtener Categoría
    - Se le pasa como el conjunto de valores que se quieren categorizar
    - Sabiendo cual es la columna con mayor ganancia de ese árbol, se busca en la matriz de datos el valor que esta columna tiene y a partir de este valor se busca cuál es su categoría en el vector resultados (encontrando el Resultado con ese valor y aplicando el método obtenerCategoria a ese Resultado)

Para la obtención de la categoría de un conjunto de datos se creó el método obtener categoría de la clase Resultado, el cual tiene dos implementaciones distintas según sea una hoja o un nodo del árbol. Si el resultado para el valor que se encuentra en la columna de mayor información de ganancia de la matriz, es otro nodo del árbol y no una hoja se utilizó recursividad para iterar dentro de este nodo y obtener la categoría.

Para finalizar y poder demostrar la resolución del algoritmo mediante la programación, utilizamos el siguiente ejemplo de datos de entrada:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Día** | **Estación** | **Viento** | **Lluvia** | **Categoría** |
| lunes | primavera | no | no | PUNTUAL |
| miércoles | invierno | no | leve | PUNTUAL |
| jueves | invierno | no | leve | PUNTUAL |
| martes | invierno | fuerte | fuerte | ATRASADO |
| sábado | verano | normal | no | PUNTUAL |
| lunes | otoño | normal | no | MUY ATRASADO |
| viernes | verano | fuerte | leve | PUNTUAL |
| domingo | verano | normal | no | PUNTUAL |
| martes | invierno | fuerte | fuerte | MUY ATRASADO |
| jueves | verano | no | leve | PUNTUAL |
| sábado | primavera | fuerte | fuerte | CANCELADO |
| miércoles | verano | fuerte | leve | PUNTUAL |
| sábado | invierno | normal | no | ATRASADO |
| viernes | verano | fuerte | no | PUNTUAL |
| miércoles | invierno | normal | fuerte | MUY ATRASADO |
| sábado | otoño | fuerte | leve | PUNTUAL |
| lunes | otoño | fuerte | fuerte | PUNTUAL |
| viernes | primavera | no | leve | PUNTUAL |
| martes | primavera | normal | no | PUNTUAL |
| miércoles | primavera | normal | leve | PUNTUAL |

Tabla 1. Datos Obtenidos por una muestra

Y se desea conocer cuál sería la categoría de un conjunto de datos como el que sigue:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Día** | **Estación** | **Viento** | **Lluvia** | **Categoría** |
| sábado | otoño | fuerte | no | ¿? |

Tabla 1. Datos Obtenidos por una muestra

EL árbol generado por el sistema es el siguiente

Estación

Primavera

Invierno

Verano

Otoño

?

?

?

Viento

Fuerte

Normal

Muy Atrasado

Puntual

Para el conjunto de datos propuesto, su categoría, como indica el árbol de decisión generado, seria “Puntual”. Esto se debe a que primero busca a partir de la “Estación”, al ser esta “Otoño”, el algoritmo busca luego por “Viento” y al ser este “Fuerte” la categoría es “Puntual”.

Como se puede apreciar en el Árbol generado, no todos los nodos fueron desarrollados, ya que como se mencionó antes la implementación realizada profundiza los nodos según se precisen.