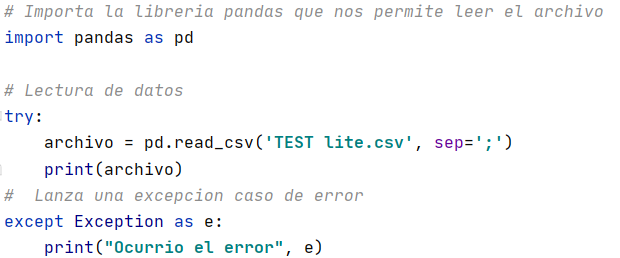
**Integrantes:**

**\***Daniel Andrés Hernández Oyola

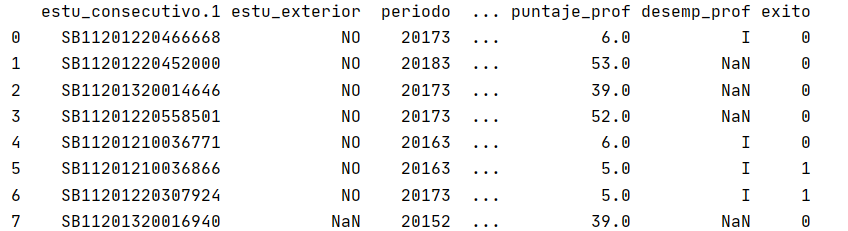
**\***Jose Miguel Blanco

**\***Dónovan Castrillón Franco

Decidimos cambiar la manera en la que estábamos realizando la entrada de datos ya que estaba generando distintos problemas. Ahora lo estamos haciendo con la librería Pandas:



Aparte, al usar esta librería la salida queda de una manera mas organizada y buena a la vista:



**Algoritmos de machine learning:**

**CART**

El termino CART hace referencia a los algoritmos de árboles de decisión que son usados en problemas de predicción por medio de la clasificación o la regresión.

CART es uno de los algoritmos más importantes en el machine learning porque establece la basa para otros tipos de algoritmos más sofisticados.

Básicamente la representación de un modelo CART es mediante un árbol binario, para crear un algoritmo basado en este modelo se deben seleccionar variables de entrada y se establecen unos puntos de división en las variables con el fin de lograr conseguir un árbol adecuado.

Un algoritmo CART se basa principalmente en los siguientes pasos:

* División codiciosa
* Criterio de detención
* Podado del árbol

División codiciosa: Para crear un árbol de decisiones se debe dividir el espacio de entrada, para esto se utiliza un enfoque codicioso para minimizar la función de costo. Consiguientemente, se selecciona la división con el costo más bajo, para esta clasificación se utiliza el índice Gini.

Criterio de detención: Para saber cuándo dejar de dividir se utiliza un procedimiento que consiste en asignar un mínimo en el entrenamiento, si el recuento es menor que el mínimo la división no se acepta y el nodo se toma como nodo hoja final. El criterio de detención es muy importante a la hora de implementar un árbol de decisión porque influye fuertemente en el rendimiento del algoritmo.

Podado del árbol: El podado del árbol permite mejorar el rendimiento del árbol, cabe aclarar que el rendimiento mejora dependiendo la implementación del algoritmo de podado.

**CHAID**

El algoritmo CHAID es uno de los más antiguos, sin embargo, es utilizado comúnmente.

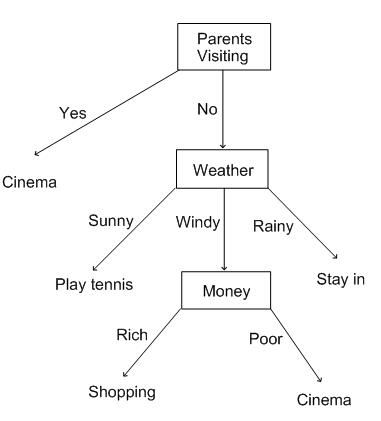
Cuando se va hacer un algoritmo CHAID se debe tener en cuenta las variables que recibe como entrada, es decir, las variables dependientes, que para este caso son cuantitativas y cualitativas. Igualmente, se deben considerar las variables de salida, es decir, variables independientes, que en este caso son las categóricas.

El algoritmo CHAID itera cíclicamente con el fin de buscar el par de categorías menos significativas con respecto a la variable dependiente, después dependiendo un valor denominado alfa se evalúa que, si el par es estadísticamente significativo, en caso verdadero sigue buscando otro par, en caso falso hace un ajuste con el método Bonferroni, luego busca la división más significativa para seguir dividiendo el árbol; El proceso anterior continua hasta que no se pueda hacer más divisiones

Cabe aclarar que la corrección Bonferroni es uno de los varios métodos que se utilizan para contrarrestar problemas de las comparaciones múltiples, Este método se basa en el error tipo I(Falso o verdadero).

**C4.5**

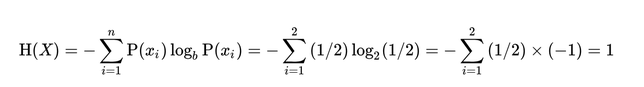
El algoritmo C4.5 es usado en Data Mining y Arboles de decisión clasificación para generar una decisión basada en datos, como el siguiente ejemplo:



El árbol consiste en una condición, donde en caso de ejecutarse se toma una decisión y esto abre nuevas condiciones, que a su vez pueden desembocar en nuevas condiciones o “estados” donde se termina de ejecutar el árbol.

De esto, tras varias ejecuciones del árbol podemos conseguir información valiosa, como las probabilidades en porcentaje de ejecutarse una acción en concreto.

Por ejemplo, si queremos ver una las posibilidades de que una moneda sea lanzada y su cara superior sea “sello”, sabemos que las posibilidades de que esto sea posible son de un 50% (0.5) y las posibilidades de obtener cara o sello son de un 100%(1.0) entonces el algoritmo seria ejecutado así.



Donde:

- E: Entropía del sistema del evento

-P: Probabilidad de que la cara superior de la moneda sea sello

-q: Probabilidad de que la cara superior de la moneda sea cara

Pruning:

Se refiere a pruning a la “eliminación” de ramas en un árbol de decisión clasificación, ya que se considera que no aportan nada significativo al proceso de la decisión final.

El algoritmo que vamos a escoger es el CART ya que nos parece que es el mas efectivo y adecuado para este proyecto.