**Proyecto 1**

**Caso Consumo de alcohol**

Tabla de contenido

1. Comprensión del negocio y enfoque analítico
   1. Tabla de oportunidad y requerimiento de minería de datos
   2. Criterios de éxito
   3. Resumen
2. Comprensión de los datos y preparación de los datos
   1. Perfilamiento
   2. Análisis de los datos
   3. Acciones concretas
3. Modelado y evaluación
   1. Tarea de clasificación
      1. Modelo utilizado en KNIME
      2. Descripción detallada del modelo
      3. Resultados: Matriz de confusión y otras estadísticas
      4. Vistazo al árbol de decisión
   2. Tarea de asociación
      1. Modelo utilizado en KNIME
      2. Descripción detallada del modelo
      3. Resultados: Reglas de asociación y respectivo soporte-confianza
      4. Resultados filtrados: Reglas de asociación para la población alcohólica
4. Análisis de resultados
   1. Tarea de Clasificación
   2. Tarea de Asociación
5. **Comprensión del negocio y enfoque analítico**

|  |  |  |
| --- | --- | --- |
| Oportunidad/Problema de Negocio | Realizar una atención oportuna y buscar control para ayudar a los problemas de adicción al alcohol | |
| Descripción del requerimiento desde el punto de vista de minería de datos | Determinar los factores que pueden influir en que un estudiante presente problemas de adicción al alcohol, y cuáles pueden ser estudiantes adictos al alcohol | |
| Detalles de la actividad de minería de datos | | |
| Tarea | Técnica | Algoritmo y parámetros utilizados |
| Clasificación | Árboles de decisión | C4.5 |
| Asociación | Patrones frecuentes | A priori |

* 1. Tabla de oportunidad y requerimiento de minería de datos
  2. Criterios de éxito

Como el estudio que quiere realizar UniAlpes está relacionado con hallar una predicción para los estudiantes propensos a sufrir de adicción al alcohol, se busca establecer un modelo que prediga acertadamente al menos el **80%** de los estudiantes con respecto a su condición de adicción al alcohol. Por otra parte, también se busca hallar reglas de inferencia que asocien atributos con la adicción, con un soporte de **70%** y una confianza de **70%**.

* 1. Resumen

Los modelos entregados se alinean con los objetivos del negocio, puesto que los resultados se pueden adaptar para que UniAlpes realice planes de acción para controlar el bienestar de los estudiantes que pueden presentar problemas de adicción al alcohol, y también, se pueden tomar medidas preventivas o campañas para impedir que los estudiantes tomen decisiones incorrectas en su vida causadas por el alcohol.

1. **Comprensión de los datos y preparación de los datos**
   1. Perfilamiento:

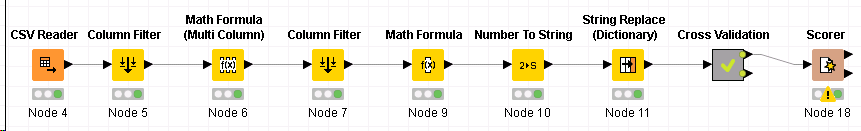


* 1. Análisis de los datos:

En general, los datos se encuentran bien definidos, y no existen respuestas que incumplan con las reglas de negocio para ningún atributo. Por ende, no es necesario eliminar registros específicos, y sobre todas las encuestas recogidas se puede obtener conocimiento. Sin embargo, para obtener un estudio satisfactorio, se pueden tomar ciertas acciones concretas de forma previa a las tareas de minería de datos, que se describen a continuación:

* 1. Acciones concretas:
* Para la tarea de asociación, sería necesario convertir todas las columnas numéricas a valores nominales, así, se pueden crear reglas entre los atributos al tener en cuenta los rangos en los que se encuentran. En este punto es necesario que el negocio determine los rangos que le interesan para cada atributo (según lo que ellos determinen, la edad se puede dividir en muchos rangos entre 15 y 22 años).
* Como se indica en el diccionario, la columna “fatherd” no tiene ningún significado relevante para el negocio, y por esto se procede a eliminarla.
* Igual que en el caso anterior, la columna “famrel” no tiene ningún significado relevante para el negocio, también se debe eliminar.
* Los expertos y el negocio deben llegar a un acuerdo para determinar los diferentes rangos en que se va a agrupar la variable “absences”, puesto que tiene un dominio numérico suficientemente amplio como para considerar no tomar cada valor como una etiqueta. Además, los datos no están centralizados, por lo que puede que en los valores iniciales haya más ocurrencias que en los finales, y esto haría que los rangos no sean simétricos. Esto hace falta en la tarea de asociación, pero no en la tarea de clasificación.
* Para la tarea de asociación, las columnas (como Medu, Fedu, traveltime, studytime, freetime, y demás) que tienen valores puntuales representados por un número, se debe reemplazar el número por el correspondiente valor nominal que se indica en el diccionario.
* Como las notas finales (G1, G2, G3) de los estudiantes están en un rango inusual para una calificación (0-19), es necesario definir grupos nominales para subrangos de notas en la población. Puede suceder que la columna de notas tenga valores nominales que no aparezcan en estas encuestas (en caso de que sea 19/100, nunca se verá un dato con “excelente”). Esto hace falta en la tarea de asociación, pero no en la tarea de clasificación.
* Existen algunas variables en las que se define un rango entre 1 y 5 con las etiquetas muy bajo hasta muy alto. Sin embargo, no se define nada para los valores intermedios, y sería necesario indicar cuáles de estos son importantes y cuál es la etiqueta indicada.

1. **Modelado y evaluación**
   1. Tarea de clasificación
      1. Modelo utilizado en KNIME:



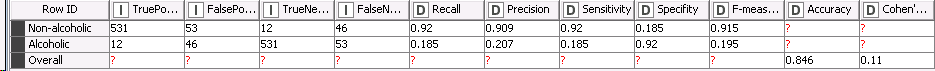
* + 1. Descripción detallada del modelo:

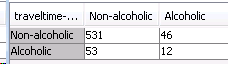
Todo el preprocesamiento fue realizado haciendo uso de Knime, pues cuenta con componentes que permiten realizar las operaciones indicadas en el numeral anterior. Primero se incluye un CSV reader que permita obtener los datos desde un archivo, luego se excluyen las columnas que no aportan conocimiento, como fatherd y famrel (Pues los expertos no saben decir qué representa) en el nodo 5. En el nodo 6 se crea la nueva columna que representa el valor Alc, calculado con respecto a Walc y Dalc, y se eliminan estas dos columnas inmediatamente después, en el nodo 7, para evitar que hagan parte del árbol de decisión (En caso de dejar alguna de las dos, el árbol solo las tendría en cuenta, pues la predicción de alcoholismo depende directamente del valor calculado con la fórmula Alc\*7 = 2\*Walc + 5\*Dalc).

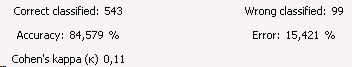
El nodo 9 se encarga de crear una nueva columna luego de que verifique si el valor de la nueva columna Alc es mayor que 3, para representar si el sujeto es alcohólico o no. Los nodos 10 y 11 se encargan de transformar la columna generada como double en una columna String que tome los valores “alcoholic” o “non-alcoholic”, de forma que la etiqueta de clase del árbol de decisión que se genera en los nodos de cross validation y scorer sea intuitiva para el negocio.

* + 1. Resultados: Matriz de confusión y otras estadísticas:

Luego de ejecutar todo el modelo, se obtuvieron los siguientes resultados:

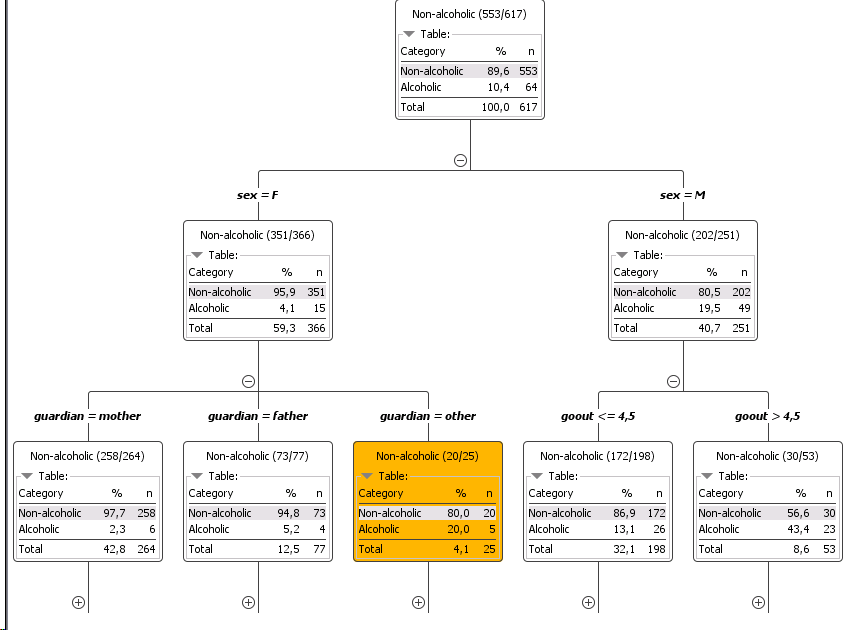




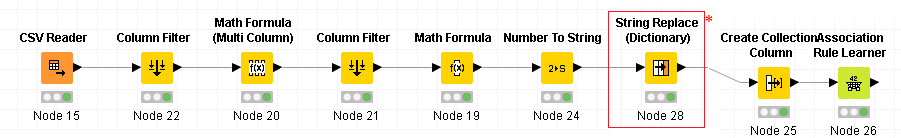


Se puede evidenciar que la precisión estuvo por encima del objetivo planteado para la minería, que era del 80% al menos. Esto quiere decir que el árbol generado es suficientemente bueno para realizar predicciones sobre los estudiantes propensos a sufrir de adicción al alcohol.

* + 1. Vistazo al árbol de decisión:



* 1. Tarea de asociación
     1. Modelo utilizado en KNIME:



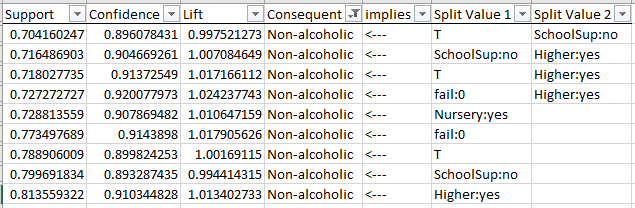
* + 1. Descripción detallada del modelo:

La primera etapa de preprocesamiento de datos en este modelo fue realizada de la misma forma que en el modelo de la tarea de clasificación, puesto que las acciones que se realizan son necesarias para entender el problema particular de este caso: Calcular la columna Alc, relacionarla con el alcoholismo y eliminar las columnas que no aportan conocimiento, que se realiza en los nodos 15, 22, 20, 21, 19 y 24.

Luego de tener los datos iguales que en el modelo anterior, se procede a nominalizar los datos. Para esto se utilizan múltiples diccionarios en componentes String Replace, los cuales van a asignar un valor intuitivo a cada columna numérica, de forma que las reglas de inferencia tengan valores claros para el negocio en antecedentes y consecuentes. Posteriormente, en el nodo 25, se deben juntar todos los datos en formato de conjunto de ítems, de tal forma que el algoritmo para hallar las reglas de asociación se pueda ejecutar sobre esta columna y descubrir relaciones interesantes entre varios atributos, incluyendo el alcoholismo.

* + 1. Resultados: Reglas de asociación y respectivo soporte-confianza:

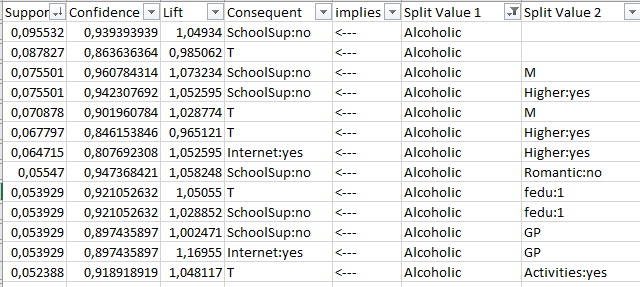
Luego de ejecutar el algoritmo, se obtienen los siguientes resultados:



Como se puede ver, entre las reglas de asociación que cuentan con el soporte (**70%**) y la confianza (**70%**) indicados en los objetivos iniciales, que no existe ninguna que determine el alcoholismo en una persona, pues el soporte significa que **455** encuestados cumplan con estas reglas, y solo se cuenta con **66** encuestados que efectivamente presentan adicción al alcohol (lo que constituye un 14,5%). Sin embargo, se encuentran reglas que incluyen la condición de no-alcohólico, lo cual igualmente puede servir al negocio para garantizar las condiciones que implican esto, y así controlar este problema en la población objetivo.

* + 1. Resultados filtrados: Reglas de asociación para la población alcohólica:

Adicionalmente, si se busca crear reglas de inferencia que incluyan la condición de alcohólico, se puede disminuir el soporte hasta alcanzar el valor de **10%**, con una confianza de **70%**, y filtrar únicamente las que tengan la condición alcohólica como su consecuente, puesto que el número de reglas generado con este soporte aumenta significativamente. Esto da los siguientes resultados:



1. **Análisis de resultados**
   1. Tarea de clasificación:

Como la precisión que da el árbol de decisión generado supera el 60%, puede ser considerado como un buen árbol. También hay que tener en cuenta que el árbol intenta estar bien balanceado, a pesar de que tiene una cantidad mucho mayor de personas en condición de no adicción al alcohol que las que sí presentan esta condición. Esto se logra al elegir un nodo padre adecuado, que en este caso pregunta por la condición del sexo de cada encuestado.

La precisión supera el 80%, que fue el criterio de evaluación elegido al inicio del proyecto, por lo que puede representar información relevante para el negocio y una herramienta adecuada para clasificar en uso futuro.

Con este árbol, se puede predecir si un estudiante puede llegar a tener problemas de adicción al alcohol, con un desempeño sobresaliente. Además, la forma en que el árbol realiza preguntas (en los nodos), llega a mostrar información descriptiva para el negocio, pues se ve la respectiva influencia en el alcoholismo para cada atributo. Por ejemplo, con el primer nodo, se puede ver que el porcentaje de hombres alcohólicos supera al porcentaje de mujeres alcohólicas.

* 1. Tarea de asociación:

Con un soporte del 70% y una confianza del 70%, se puede afirmar que el modelo de asociación realizado tiene relevancia para el negocio. El soporte, particularmente, es diciente debido a la cantidad de datos a evaluar en el proyecto.

Se filtraron solamente las reglas que tuvieran como consecuente a los valores “alcoholic” o “non-alcoholic” con los procesos descritos anteriormente. Con esto, se buscaba encontrar factores que se pudieran relacionar como causa de cada una de las condiciones. Los resultados arrojan que condiciones positivas como tener a los padres juntos, no reprobar notas, recibir acompañamiento de enfermería y poco soporte educativo, resultan en un estudiante sin problemas de adicción al alcohol. El negocio puede tomar estas reglas como fundamento para tomar ciertas decisiones que puedan garantizar estas condiciones a todo estudiante.

En cuanto a las reglas que causan el alcoholismo en un estudiante, no se pueden generar reglas con buen soporte y confianza que asocien algún antecedente con el consecuente de ser alcohólico, pues hay muy pocos estudiantes que presentan adicción al alcohol, y todos se comportan diferente. Sin embargo, se generan reglas, en las que la condición de alcoholismo se toma como un antecedente, que el negocio puede usar para tomar planes de acción frente a las condiciones que representan los otros atributos y buscar la forma de evitar la exposición de los estudiantes a estos factores, para proteger su salud.