## Explicación del Problema

|  |
| --- |
| Este proyecto está basado en una competencia de Kaggle que tiene como objetivo la búsqueda de un algoritmo que permita generar un modelo que prediga de manera eficiente el estado civil de una pareja dado un set de datos el cual contiene una serie de características de la pareja, tales como: la edad del hombre, edad de la mujer, ingreso anual de ambos, entre otros. Todos estos datos pasarán por un pre-procesamiento y finalmente para la generación de los modelos se utilizarán los algoritmos Árbol de Decisión, K-Vecinos y Reglas de Clasificación. |

## Instalación de paquetes necesarios

La instalación de los paquetes se muestra comentada porque ya se encuentran instalados en el ambiente de desarrollo

## Carga del set de datos

\* Este archivo contiene los datos relacionados al estado civil de una pareja, los cuales serán utilizados para implementar y evaluar los modelos anteriormente mencionados, y de esta forma poder elegir el mejor modelo para la predicción que se desea realizar.

## Preprocesamiento de la data

Para realizar el pre-procesamiento de la data se realizó un análisis exploratorio de los mismos en donde se observó que existían columnas irrelevantes en cuanto al objetivo del problema, dichas columnas fueron el ID y AINCOME los cuales representan el identificador de la fila y el sueldo anual de la pareja respectivamente. En cuanto a este último atributo lo consideramos irrelevante ya que no hallamos ningún tipo de patrón que nos encaminara a predecir el valor real del estatus de la pareja. Seguido de esto se decidió aplicar una numeración a las columnas que contenían variables nominales las cuales podían asumir n valores, dichas columnas son: GP, BP y STATUS. Cabe destacar que al realizar el análisis se pudo concluir que la diferencia entre ambas edades era un factor fuerte para determinar el estado civil debido a que se observaron diversas similitudes de diferencia de edad para un mismo valor de clase, por lo que se procedió a eliminar ambas columnas y crear una nueva almacenando la diferencia de edades.

##Muestreo

Dado que se tienen posibles valores (1-3) para la variable que se desea predecir "STATUS", se realiza un muestreo estratificado para asegurar que el mismo sea proporcional y exista un equilibrio en las muestras de entrenamiento y prueba.

##modelos

Los algoritmos utilizados para lograr predecir el estado civil de una pareja con los datos obtenidos y ya pre-procesados fueron Árbol de Decisión, K-Vecinos y Reglas de Clasificación. Se decidió utilizar estos métodos de clasificación porque fueron los estudiados en la materia de Minería de Datos y de los cuales se tenía mayor conocimiento a la hora de implementar la solución al problema planteado. Además es bueno destacar que son métodos con un funcionamiento fácil de entender, y dado que cada algoritmo de estos se basa en diferentes principios, permite que dependiendo del problema algunos modelos sean mejores que otros, ampliando así la posibilidad de lograr un buen resultado. Para la evaluar la eficiencia de los modelos se tomó en cuenta la matriz de confusión y la curva ROC

##Arbol de decisión

Se generó el modelo para el árbol de decisión usando la función rpart, cuyos parámetros fueron: (STATUS ~.) el cual denota que se quiere predecir la característica STATUS (estado civil) de la pareja basándose en las demás características provistas en el set de datos y el parámetro method: (class) indicando se quiere un modelo de clasificación. En cuanto a los parámetros de control: minsplit, cp, maxdepth (los cuales nos indican la cantidad mínima de individuos por nodo, el grado de brusquedad en los cambios y la profundidad de árbol respectivamente) no fueron utilizados debido a que en al realizar la permutación de los mismos, los resultados del árbol fueron muy similares. Para la evaluación de este modelo se tomó en cuenta la matriz de confusión y el área bajo la curva (función roc del paquete pROC) obteniendo como resultado un noventa (90) por ciento de aciertos y 0.97 de área bajo la curva respectivamente

##K nn

Se generó el modelo de K-Vecinos haciendo uso de la función knn, pasando los conjuntos de prueba y entrenamiento, y el parámetro k (número de vecinos considerados) asignándole el valor 28 debido a que se tomó como medida la raíz cuadrada del número total de individuos en el set de datos (750) que suele ser la medida que mejor se comporta teóricamente. Para la evaluación de este modelo se tomó en cuenta la matriz de confusión y el área bajo la curva (función roc del paquete pROC) obteniendo como resultado un cincuenta (50) por ciento de aciertos y 0.62 de área bajo la curva respectivamente

##Reglas de clasificación

Se generó el modelo de Reglas de Clasificación haciendo uso de la función JRip provista por la interfaz RWeka, para ello debimos realizar una transformación de los datos haciendo uso de \textbf{as.factor}, y luego aplicamos dicha función JRip, donde el primer parámetro es formula: (STATUS ~ .) el cual denota que se quiere predecir la característica STATUS de la pareja basándose en las demás características provistas en el set de datos, y el segundo parámetro que indica la data a utilizar que en este caso es la data de entrenamiento. Para la evaluación de este modelo se tomó en cuenta la matriz de confusión y el área bajo la curva (función roc del paquete pROC) obteniendo como resultado un cien (100) por ciento de aciertos y 1 de área bajo la curva respectivamente

##conclusion

Como mencionamos anteriormente para realizar la comparación de los resultados obtenidos entre los tres modelos implementados, se tomó en cuenta la matriz de confusión para obtener la tasa de aciertos y el área bajo la curva (función roc del paquete pROC) de cada uno de ellos. Se tiene que el mejor modelo obtenido es el basado en Reglas de Clasificación debido a que la tasa de aciertos es perfecta al igual que el área bajo la curva, es decir, para ambos casos el resultado fue de 1. Seguidamente el modelo basado en árbol de decisión con una tasa de aciertos de 0.91 y con un área bajo la curva de 0.97. Y por último se tiene el modelo de K-Vecinos el cual su tasa de aciertos fue de 0.5 y el área bajo la curva de 0.62 por lo que queda totalmente descartado como mejor modelo. Sin duda alguna el mejor modelo para el caso estudiado es el modelo basado en Reglas de Clasificación