Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación

Máquinas de aprendizaje

Reporte: Evaluación de un modelo de machine learning

Docente: Abraham Sánchez López

Alumno	Matrícula
Taisen Romero Bañuelos	202055209

Proyecto de evaluación de modelos

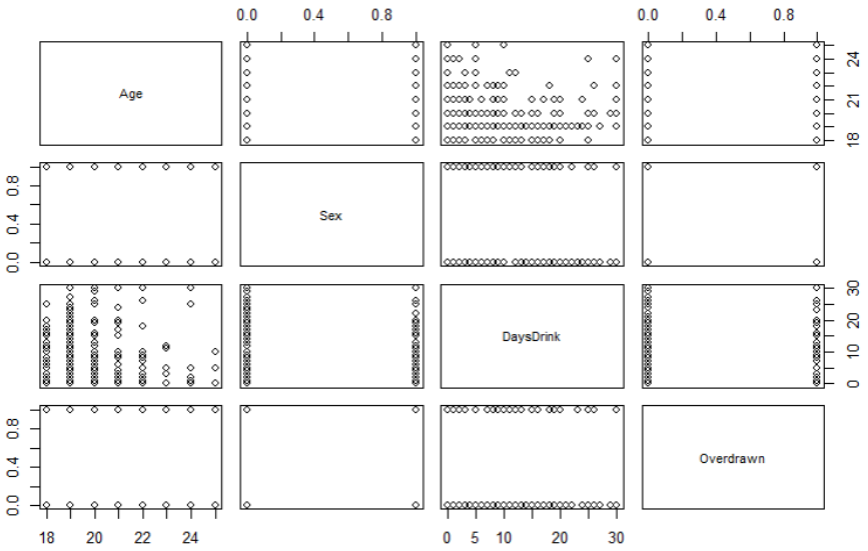
Análisis exploratorio de los datos (EDA)

Se dice que un modelo es tan bueno como lo son los datos que usa y analiza, pero también como lo es su análisis de datos. Para ello empezaremos limpiando datos. Tanto NA's como la columna X (X es una columna de conteo de número de estudiantes, por lo que no aportará nada útil a los modelos). Antes de hacer la conversión a factor de DaysDrink haré unos ploteos para el análisis de datos. También trataré de ser breve en aquellas cosas que no aporten mucho al análisis del EDA porque lo que hice va a llevar muchas páginas para explicarlo, así que prefiero detenerme en los detalles que aporten algo de valor y ser breve con esas cosas que son más como un trámite.

Desde el inicio ya podemos observar algo importante que consideraremos a lo largo de todo el reporte pero sobre todo al final. Esto es el desbalance de las clases.

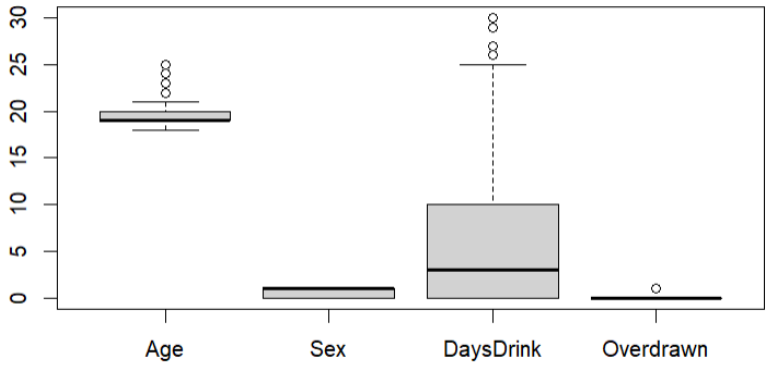
```
      Age             Sex            DaysDrink        Overdrawn      
 Min.   :18.00   Min.   :0.0000   Min.   : 0.000   Min.   :0.0000  
 1st Qu.:19.00   1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.:0.0000  
 Median :19.00   Median :1.0000   Median : 3.000   Median :0.0000  
 Mean   :19.62   Mean   :0.5584   Mean   : 6.497   Mean   :0.1281  
 3rd Qu.:20.00   3rd Qu.:1.0000   3rd Qu.:10.000   3rd Qu.:0.0000  
 Max.   :25.00   Max.   :1.0000   Max.   :30.000   Max.   :1.0000  
```

En algunas variables podemos observar que hay un desbalance en los cuartiles, pero hay un desbalance menos sencillo de observar si sólo usamos la función summary(). Estos desbalances surgen de la combinación de variables.



Empecé a notar esta clase de desbalances “ocultos” porque por ejemplo, en la celda Age - DaysDrink se ve que hay más puntos hacia la izquierda, y aunque sea más o menos difícil de notar, algo similar sucede con Sex - DaysDrink. Esto podría significar que hay edades y sexos más propensos al alcohol, y como veremos más adelante, también al sobregiro.

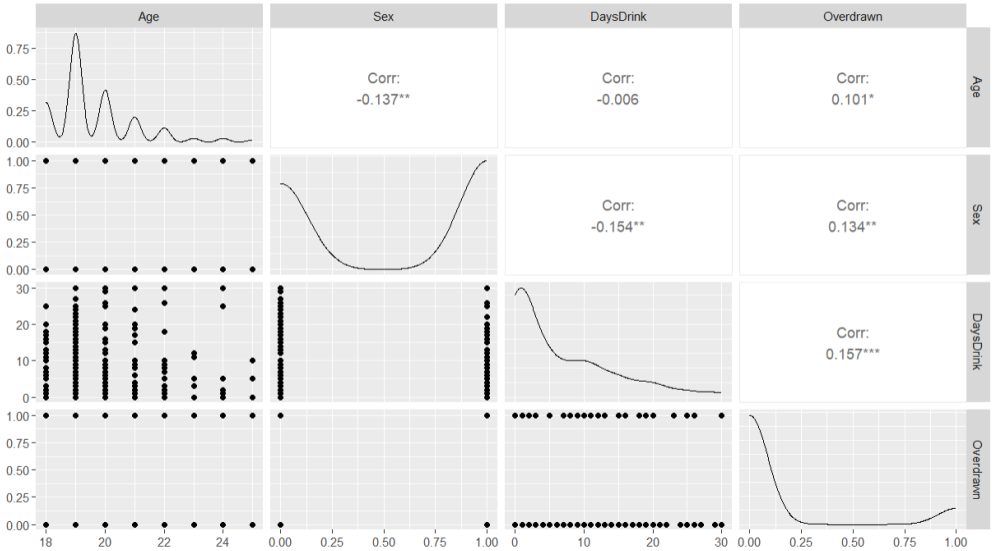
Aquí, además del desbalance de los datos también podemos observar que hay datos atípicos, lo que tiene sentido, pues más adelante veremos que Age y DaysDrink tienen datos muy variopintos.



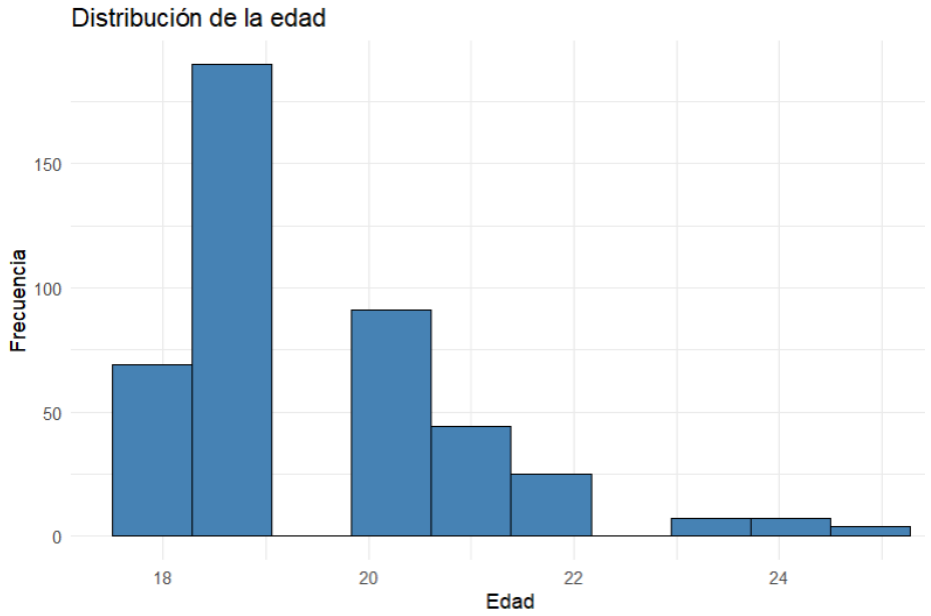
De momento veamos cómo está la correlación entre variables.

```
> cor_matrix
                Age        Sex    DaysDrink Overdrawn
Age        1.000000000 -0.1372684 -0.006016559 0.1005108
Sex       -0.137268433  1.0000000 -0.154262729 0.1341731
DaysDrink -0.006016559 -0.1542627  1.000000000 0.1571219
Overdrawn  0.100510787  0.1341731  0.157121948 1.0000000
```

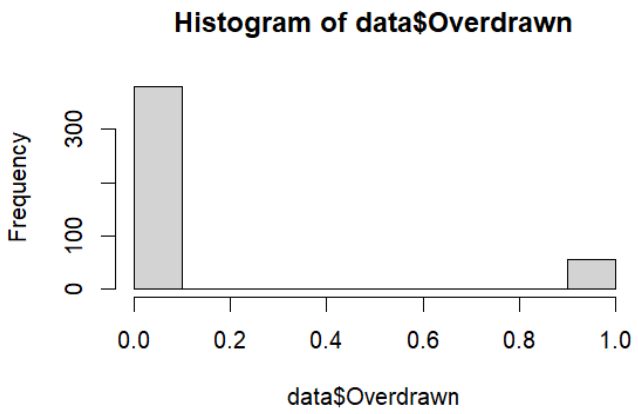
Aquí podemos empezar a notar que DaysDrink tiene mucha influencia en el sobregiro, y a su vez, que DaysDrink se ve muy influenciado por el sexo.



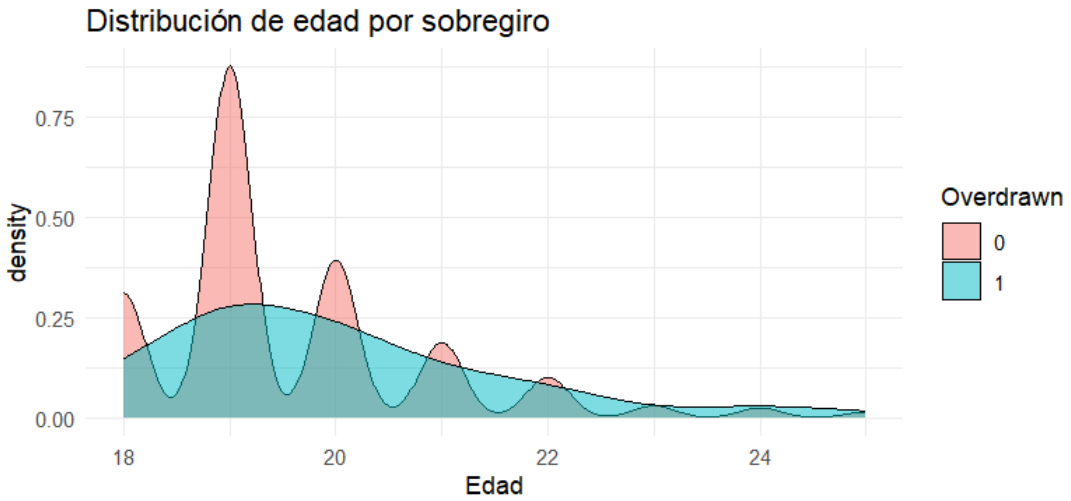
En la matriz anterior vemos unas curvas en forma de “U”, lo que podría ser preocupante de no ser porque el tipo de dato de esas variables son binarios, entonces resulta natural que exista esa tendencia hacia los extremos. Además, también las celdas entre variables refuerzan cada vez más la idea de que hay un desbalance entre variables y no sólo con las variables de forma aislada. Por ejemplo, ya se puede empezar a notar que las edades más jóvenes tienden a pasar más días bebiendo, aunque esto podría explicarse por un desbalance en la distribución de la edad. Como vemos en el siguiente histograma, la mayoría de los encuestados fueron personas jóvenes, lo que podría sesgar un poco el análisis de datos.



Podemos ver este sesgo en otras variables. Aunque en Sex no es demasiado el desbalance, si lo es en el sobregiro y en DaysDrink (buenas noticias, al parecer el alcoholismo no está tan presente).



Entonces, si la mayoría de personas son jóvenes y la mayoría bebe ocasionalmente, esto cómo se verá reflejado en un diagrama de dispersión de la edad por sobregiro?



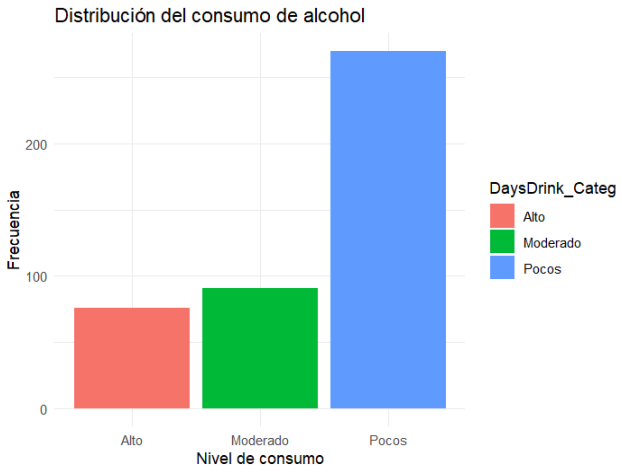
Como vemos, el sobregiro también se ve influenciado por la edad. Bien podría ser por el desbalance de clases pero supongamos que así es la tendencia natural que capturaron en los datos. Propongo este supuesto porque pese a que la mayoría de los sobregiros positivos están sesgados a la izquierda (al igual que Age), realmente la mayoría de las observaciones tienen un Overdrawn negativo, lo que me hace pensar que más bien esa es la tendencia natural de los datos.

Para corroborar la proporción de Overdrawn calculé la proporción de Overdrawn (y de paso muestro también la distribución de DaysDrink después de volverla a factor como se sugirió).

```
> prop.table(table(data$Overdrawn))
        0         1
0.8718535 0.1281465
```

```
> table(data$DaysDrink_Categ)

  Alto Moderado    Pocos
    76       91      270
```

Y para que sea más evidente la diferencia en la proporción de DaysDrink_Categ veamos el siguiente plot



Como podemos ver, la mayoría de gente no bebe mucho, sin embargo, la diferencia entre Alto y Moderado es poca, cosa que resulta interesante, pues quiere decir que es más fácil pasar de beber moderadamente a beber mucho que pasar de beber moderadamente a beber poco. Quizá sin quererlo encontramos una tendencia al alcoholismo.

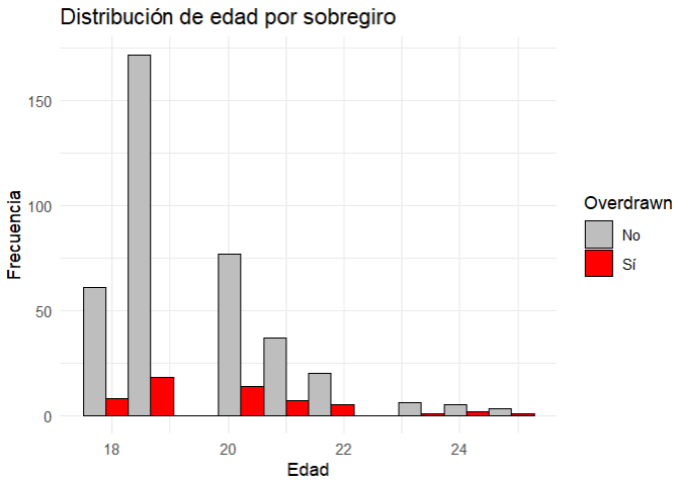
En fin, pasemos a hacer la prueba de Chi-cuadrado para buscar asociaciones significativas de las variables con Overdrawn (sobregiro).

```
> chisq.test(chi_table_age)

        Pearson's Chi-squared test

data:  chi_table_age
X-squared = 6.1613, df = 7, p-value = 0.5211
```

```
> chisq.test(chi_table_sex)

        Pearson's Chi-squared test with Yates' continuity correction

data:  chi_table_sex
X-squared = 7.0795, df = 1, p-value = 0.007797
```

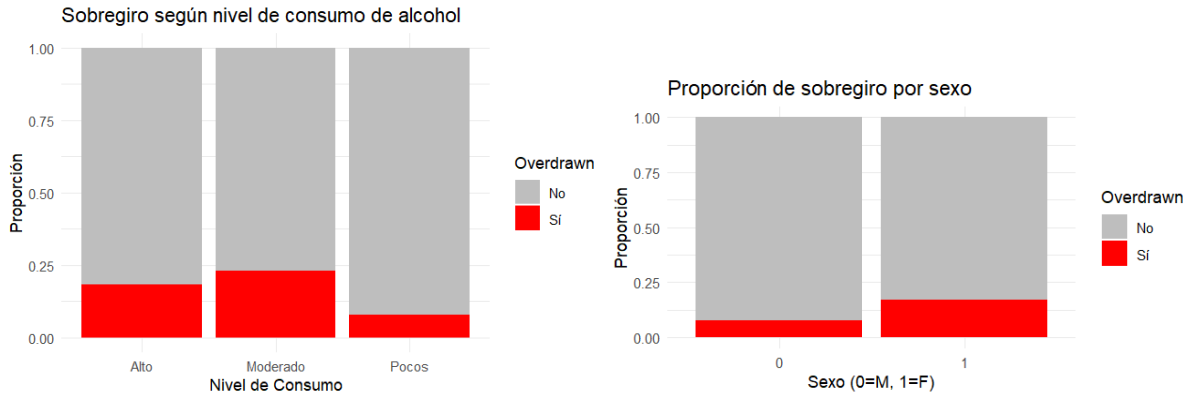
```
> chisq.test(chi_table_drink)

        Pearson's Chi-squared test

data:  chi_table_drink
X-squared = 16.847, df = 2, p-value = 0.0002196
```

Los resultados corroboran que hay una **asociación significativa** entre DaysDrink y el sobregiro, seguido de Sex. En cuanto a la edad, pareciera que no hay una asociación significativa, lo que es bueno para hacer más sólido el supuesto que hice anteriormente. De todas formas, más vale corroborar los datos y no confiarse porque como se discutió el los PDF's sobre las medidas de evaluación, muchas veces uno puede irse con la finta de un resultado conveniente cuando en realidad hace falta ver las cosas desde otra perspectiva.

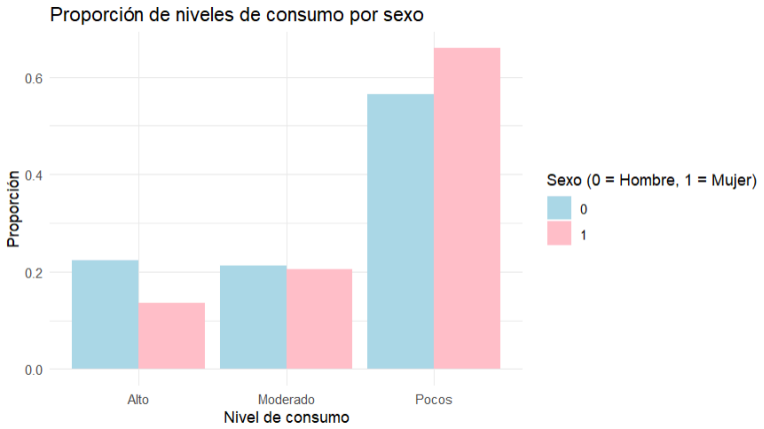


Bueno, este plot nos permite observar de mejor manera la relación entre el sobregiro y la edad. Con esto podemos notar que la edad en sí no parece tener una relación lineal clara con el sobregiro, si acaso podemos decir que visualmente, la proporción de sobregiros parece más alta en estudiantes de 21 años en adelante, pero no en un patrón fuerte. Esto coincide con la prueba de chi-cuadrado (no hay una asociación significativa entre edad y sobregiro) aunque cierto es que podría haber algo interesante en un grupo específico, la mayoría de los estudiantes tienen entre 18 y 21 años, lo que coincide con esos grupos de alumnos con más sobregiro. En fin, podemos decir que nuestro supuesto sobre la naturalidad de los datos es suficientemente sólido como para continuar. No sólo por lo expuesto hasta ahora, sino que también porque encaja con el contexto de un entorno universitario. Por sí solo el argumento del entorno podría darle solidez suficiente al supuesto, sin embargo, consideré necesario hacer todo esto, no solo para reforzar el supuesto, sino que también para ayudarnos a hacer un buen EDA de forma orgánica.

Bueno, continuemos con el análisis entre dos variables (aquello de lo que hablé al inicio del documento).

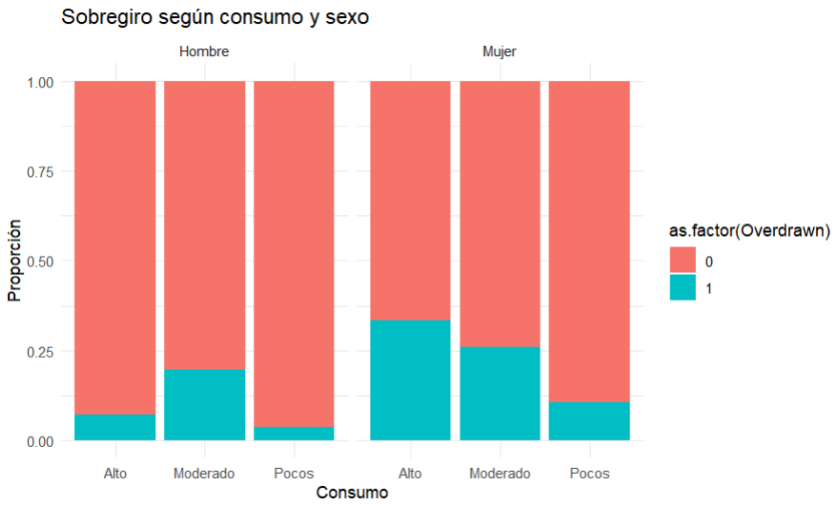


El plot de Sobregiro según el consumo de alcohol (Overdrawn vs DaysDrink) muestra que el grupo de alumnos con menor consumo de alcohol tiene también la menor proporción de sobregiros. Como tal, hay una relación positiva entre el consumo de alcohol y la probabilidad de sobregiro, aunque el patrón no es perfectamente lineal (Moderado tiene más peso, luego Alto y finalmente Pocos). Por otro lado, el plot de Overdrawn vs Sex muestra que las mujeres son más propensas al sobregiro, incluso pese a que los hombres beban más que las mujeres. Véase el siguiente plot para observar que los hombres beben un poco más que las mujeres.



Algunas de estas relaciones ya se veían venir desde los primeros ploteos del EDA, como por ejemplo la matriz generada por la función ggpairs(), pero hay que revisarlas manualmente para corroborar las cosas, sobre todo porque en la matriz mencionada no se observaba con precisión la diferencia entre la proporción de una categoría y otra porque aparecían como puntos sobrepuestos.

El siguiente plot es una combinación de la relación entre tres variables, Consumo & Sexo vs Overdrawn.



Este plot es muy importante porque muestra la fuerte tendencia de las mujeres al sobregiro. Como ya lo habíamos mencionado, el consumo de alcohol está asociado al sobregiro en ambos sexos, pero este plot revela ciertas diferencias en intensidad. Para las mujeres, hay una relación más clara y lineal: a mayor consumo, mayor sobregiro. Mientras que en hombres, el patrón es menos lineal; el consumo moderado muestra mayor proporción de sobregiro que el alto. Esto refuerza que hay una interacción/relación entre consumo y sexo.

Construcción de modelos

Bueno, con la información presentada hasta ahora creo que es suficiente como para que pasemos a lo que nos concierne, la evaluación de los modelos. Pero antes de eso, hay que construirlos. Para poder hacer un buen análisis de las medidas de evaluación decidí construir 4 modelos para su comparación. Para los tres primeros modelos se usó el paquete Rpart y se diferencian en el esquema de validación utilizado.

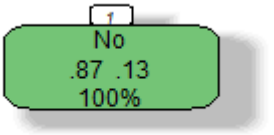
- **Modelo 1: Árbol de decisión con partición 70:30**. El conjunto de datos se dividió en 70% para entrenamiento y 30% para prueba. Este es el enfoque clásico que hemos venido usando para evaluar el modelo con datos no vistos.
- **Modelo 2: Árbol de decisión con partición 50:25:25**. Aquí se utilizó un 50% de los datos para entrenamiento, un 25% para prueba y otro 25% para evaluación. Este es el enfoque que me dio mejores resultados en la tarea anterior por lo que le tengo cierta fe.
- **Modelo 3: Árbol de decisión con validación por bootstrap** (100 iteraciones). Se entrena el modelo con el 70% de los datos (misma división del Modelo 1) y se utiliza el método bootstrap para validación interna, permitiendo evaluar la estabilidad del modelo a través de múltiples muestras con reemplazo. No usé un entrenamiento con el 100% de los datos porque si el objetivo de este análisis es usar las medidas de evaluación creo que es más ideal usar una partición de los datos para que al comparar con los otros modelos los resultados sean más representativos.
- **Modelo 4: Regresión logística** (partición de 70:30 del Modelo 1). Este modelo busca predecir el sobregiro a partir de las mismas variables, usando un enfoque probabilístico. Debido al desbalance en la variable objetivo, se ajustó el umbral de corte para mejorar su capacidad de clasificación. Usé la partición del modelo 1 porque como veremos, resultó ser la más adecuada.

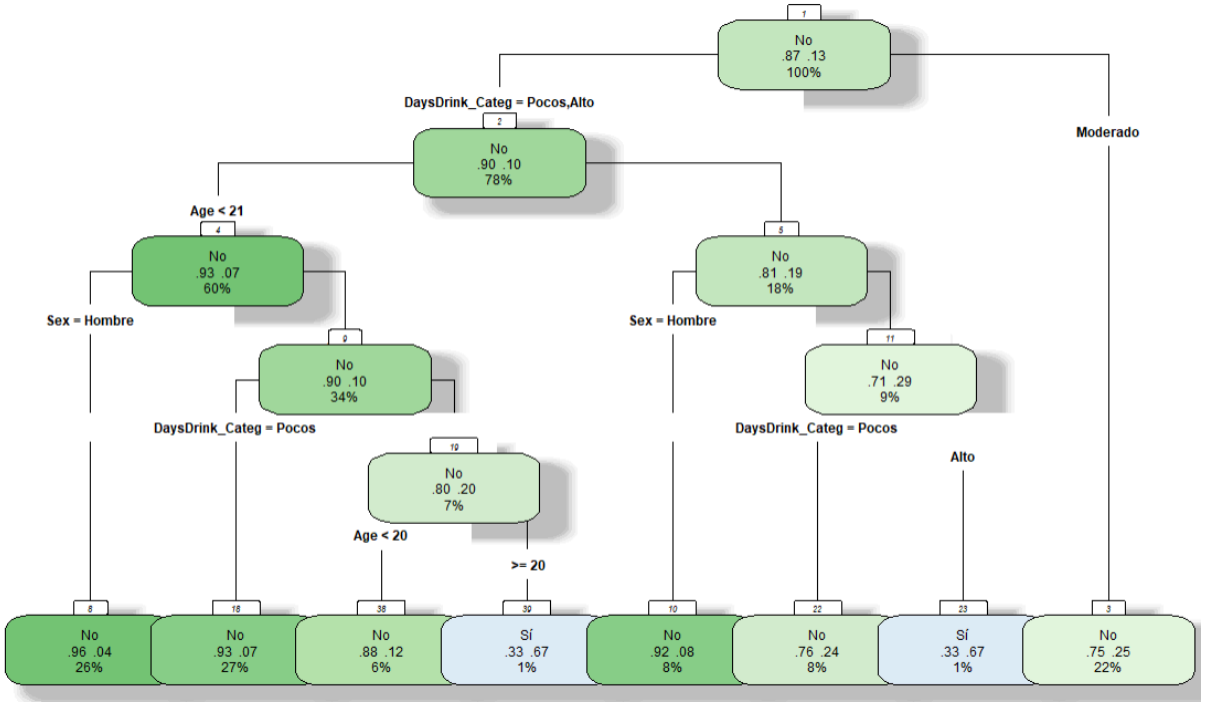
Tras la construcción de estos modelos, el siguiente paso será comparar su desempeño utilizando métricas como Kappa y AUC (ROC). No incluí el MCC porque que yo

recuerde no hemos aplicado el MCC en las prácticas, así que siguiendo mi enfoque de centrarme en lo que vemos en clase, usaré sólo Kappa y AUC/ROC.

La construcción y entrenamiento de los modelos no tiene mucho misterio pero sí hay algunos puntos importantes a considerar. El primero es que el CP fue un argumento importante en la construcción de los árboles, pues, si dejaba el CP por defecto de 0.01 el árbol sólo construía un único nodo. Esto significa que el CP es muy estricto y no permite crecer el árbol, por lo que probé con un CP que permitía al árbol expandirse lo suficiente como para considerar los caminos principales. Después, quise podar el árbol calculando el CP óptimo, sin embargo, resulta que el CP óptimo me genera otra vez un árbol de un nodo, por lo que opté por conservar el CP que expandía el árbol de manera adecuada bajo mi criterio.

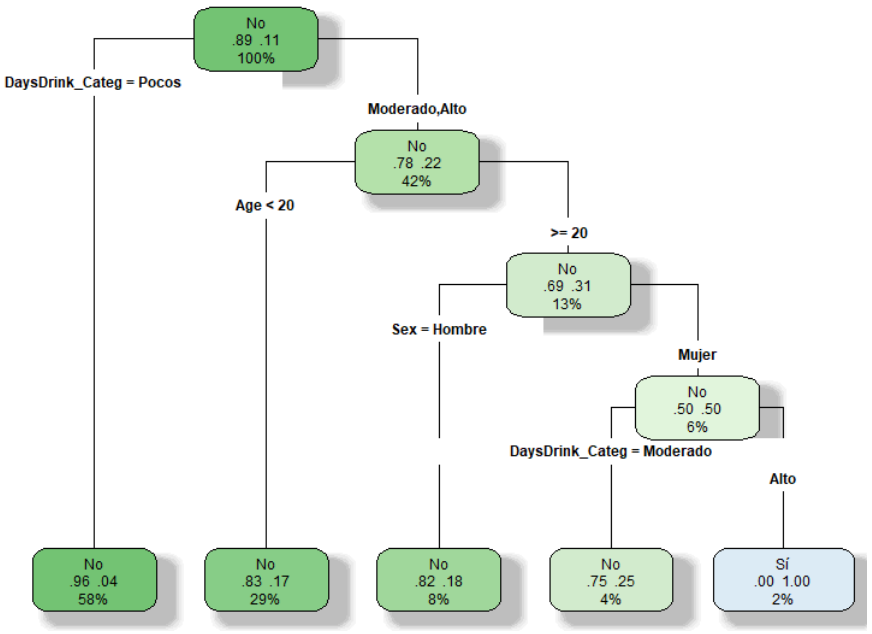
Adjunto una captura del árbol con un CP óptimo vs el árbol (modelo 1) con el CP que consideré mejor.





Como podemos ver, la diferencia es notable. Usé el mismo CP para el resto de modelos, por lo que será algo a considerar.

Este es el árbol generado por el modelo 2:



Si observamos las métricas generadas vemos de nuevo que el consumo de alcohol es una variable que influye mucho en el sobregiro.

```
> tree_70$variable.importance
DaysDrink_Categ             Age             Sex
       4.137972        2.799274        1.486538
```

```
> tree_50$variable.importance
DaysDrink_Categ             Sex             Age
      6.4812507       1.4726166       0.7318246
```

En cuanto a las métricas de resumen, el **árbol 70:30** generó un árbol con 7 divisiones, donde la variable más importante fue el nivel de consumo de alcohol. Aunque se observó una buena capacidad de separación inicial (error base de 13%), el aumento en el error de validación cruzada con más divisiones sugiere un posible sobreajuste. Ya veremos si esto es cierto en la evaluación.

```
> summary(tree_70)
Call:
rpart(formula = Overdrawn ~ ., data = train_1, method = "class",
    control = rpart.control(cp = 1e-04, minsplit = 5))
  n= 307

           CP nsplit rel error xerror      xstd
1 0.007142857      0      1.00  1.000 0.1474540
2 0.000100000      7      0.95  1.175 0.1577271

Variable importance
DaysDrink_Categ             Age             Sex
             49              33              18
```

El **árbol 50:25:25** fue más conservador porque solo hizo 4 divisiones. DaysDrink también fue la variable más relevante, aunque el sexo ganó más peso en comparación con el modelo anterior. Esto podría indicar que, en menor tamaño de muestra, ciertas variables ganan relevancia.

```
> summary(tree_50)
Call:
rpart(formula = Overdrawn ~ ., data = train_2, method = "class",
    control = rpart.control(cp = 1e-04, minsplit = 5))
  n= 218

     CP nsplit rel error xerror     xstd
1 4e-02      0      1.00   1.00 0.188183
2 1e-04      4      0.84   0.88 0.177898

Variable importance
DaysDrink_Categ             Sex             Age
             75              17               8
```

El árbol con bootstrap no logró un buen desempeño. Esto puede deberse a que las variables predictoras no son suficientemente fuertes para captar patrones robustos dadas las múltiples particiones. Pero ya hablaremos de esto en la sección de evaluación.

```
  CP nsplit rel error
1  0      0         1

Node number 1: 307 observations
  predicted class=No  expected loss=0.1302932  P(node) =1
    class counts:   267    40
   probabilities: 0.870 0.130
```

Evaluación de los modelos

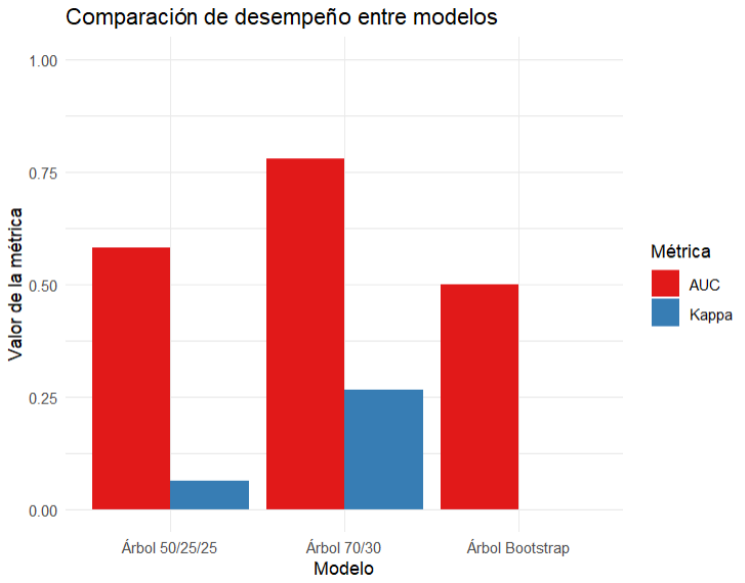
Bueno, ahora si vamos con el mero mole. Para evaluar los modelos usé Kappa y ROC/AUC. El proceso es similar a como lo hicimos en las prácticas anteriores, con la diferencia de que para agilizar los procesos y explorar la programación funcional de R hice una función de evaluación de Kappa y el AUC para los primeros dos árboles. La función (o pipeline) no tiene mucho misterio. Le pasamos el modelo, los datos de prueba y el nombre del modelo para el etiquetado. Internamente calcula las predicciones y todo lo necesario para sacar el Kappa y el AUC. Esto lo menciono en parte porque fue el único pipeline que me dio buenos resultados. Quise hacer una especie de funciones modulares para hacer el entrenamiento y evaluación de un modelo “x” en unas pocas líneas (y así reutilizar el código para futuras actividades), pero a veces las soluciones más simples son las más efectivas.

En fin, pasemos a la comparación del rendimiento de los tres modelos.

```
> resultados_modelos
           Modelo Kappa   AUC
1     Árbol 70:30 0.264 0.780
2  Árbol 50:25:25 0.062 0.582
3 Árbol Bootstrap 0.000 0.500
```

Lo primero es que el mejor modelo fue el árbol de 70:30. Lo segundo más importante es que el árbol de bootstrap tiene un Kappa de cero. Esto lo explicaré más a detalle más adelante porque con el modelo de regresión logística pasó algo aún más

interesante. Obtuvo un Kappa igual a cero pese a tener un buen AUC. De mientras, observemos la comparación del rendimiento de los modelos en un gráfico de barras.



Realmente el árbol 70:30 fue dominante en estas pruebas, por lo que usaremos esa misma partición para crear un modelo de regresión logística, pero antes, ¿Por qué regresión logística y no lineal o lineal múltiple?, básicamente porque de entrada la variable dependiente es categórica (Overdrawn = Si/No), y la regresión lineal asume que la variable dependiente es continua, lo cual no es nuestro caso. Pero además de esa clase de tecnicismos, si usamos un modelo de regresión lineal podría obtener probabilidades <0 o >1 (lo cual no tiene sentido para la respuesta que buscamos). Pero si incluso usamos un umbral numérico para la categorización de la respuesta, no tendría una función de pérdida adecuada para clasificación y el modelo no estaría optimizado para maximizar la discriminación entre clases. Lo cual dificultará no solo el manejo de las respuestas sino también su fiabilidad para compararlo con los otros modelos.

Bien, dicho esto veamos que nos arroja el modelo de regresión logística.

```
> cat("— Regresión Logística (70/30) —\n")
— Regresión Logística (70/30) —
> cat("Kappa:", round(kappa_logit, 3), "\n")
Kappa: 0
> cat("AUC:", round(auc_logit, 3), "\n")
AUC: 0.789
```

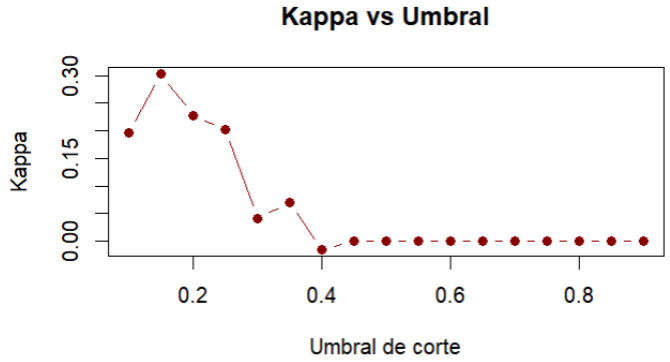
Aquí observamos que pese a que nuestro Kappa es malo tenemos un buen AUC. ¿Por qué pasa esto?, sucede que AUC tiene cierta independencia con respeto a Kappa, pese a que calculan lo mismo (el rendimiento de un modelo) el cálculo de uno no debería afectar al del otro puesto que se basan en preceptos diferentes para medir el

rendimiento. En términos sencillos, Kappa hace algo parecido a una matriz cruzada (compara predicciones categóricas con los valores reales, tomando en cuenta la clase mayoritaria y el azar). Y el AUC mide la capacidad del modelo para ordenar correctamente las clases, independientemente del umbral de clasificación. Solo importa que los casos positivos tengan mayor probabilidad que los negativos. Entonces, ¿Qué rayos significa que el Kappa sea igual a cero?, significa que todas las predicciones salen como "No", por lo que no hay verdaderos positivos ni falsos positivos. Entonces, Kappa = 0, porque no se logró clasificar ningún "Sí" correctamente. Por eso el modelo tiene un AUC alto, pero el umbral de corte de 0.5 no logra detectar bien los casos "Sí", porque hay pocos. Entonces la solución natural es cambiar el umbral de corte de las predicciones.

La siguiente imagen confirma las sospechas, todas las predicciones se categorizaron como “No”.

```
> pred_logit
  1   3   6   7   9  10  15  18  20  21  22  23  24  27  34  36  37  44  45  48  54  57  60  61  68  69  72  74  76  79  82  84  85  87  89
 No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No
 93  96 106 111 112 115 118 122 131 133 139 140 143 144 149 152 153 161 162 169 171 172 177 181 187 193 195 196 197 208 215 217 218 219 231
 No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No
235 236 245 250 251 252 253 254 263 265 266 270 277 279 280 283 285 288 291 295 296 302 304 309 310 313 317 330 332 334 335 336 345 351 363
 No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No
366 372 375 377 378 380 384 389 392 396 397 399 401 403 407 409 411 412 413 419 421 428 432 444 450
 No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No
Levels: No Sí
```

Tan solo con cambiar el punto de corte a 0.3 noté cambios, pero para no andar adivinando el mejor umbral de corte hice una comparación entre umbrales que van desde 0.1 hasta 0.9 con incrementos de 0.05. Estos fueron los resultados.



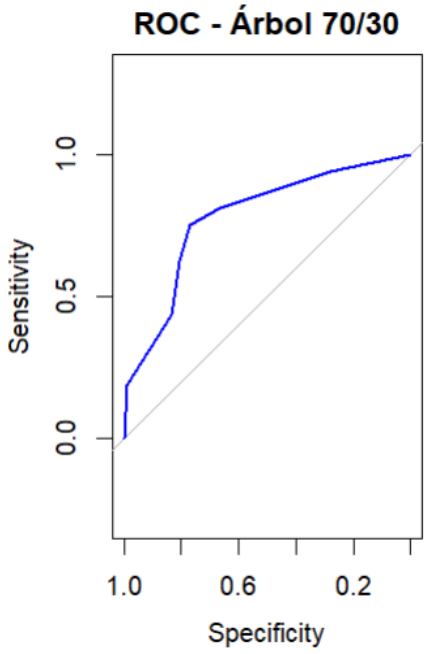
El mejor punto de corte fue con diferencia de 0.18 aproximadamente, por lo que habrá que refinar el modelo con este nuevo valor.

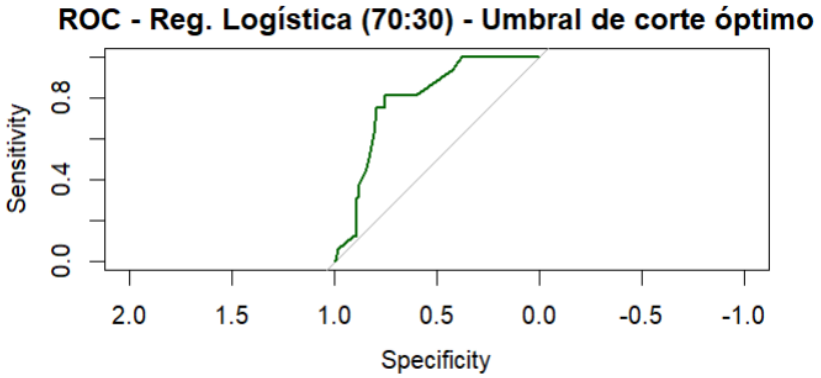
```
> cat("— Regresión Logística (umbral óptimo ≈ 0.18) —\n")
— Regresión Logística (umbral óptimo ≈ 0.18) —
> cat("Kappa:", round(kappa_opt, 3), "\n")
Kappa: 0.228
> cat("AUC:", round(auc_opt, 3), "\n")
AUC: 0.789
```

Estos nuevos resultados no decepcionan, compiten a la par con el árbol 70:30.

```
> resultados_modelos
                              Modelo Kappa   AUC
1                        Árbol 70:30 0.264 0.780
2                     Árbol 50:25:25 0.062 0.582
3                    Árbol Bootstrap 0.000 0.500
4 Reg. Logística (umbral óptimo) 0.228 0.789
```

Y sus curvas AUC son muy similares, de hecho.

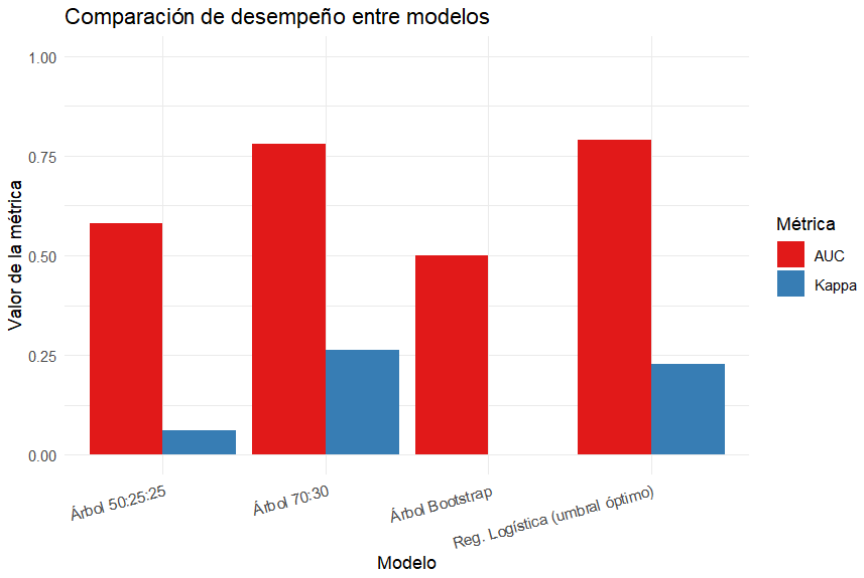




Bueno, dije que eran similares porque presentaban la misma tendencia, pero sin duda el modelo de regresión muestra un AUC más peculiar. Los cambios abruptos en la curva del modelo de regresión se pueden deber a que como hay pocas observaciones positivas cada vez que el umbral incluye o excluye uno de esos pocos casos, el cambio en sensibilidad o especificidad es más pronunciado. De hecho, estos cambios abruptos resultaron favorables, pues lograron un acercamiento mayor a la esquina superior

izquierda. Y no solo eso, de hecho el modelo de regresión si superó al modelo del árbol por poco.

```
> resultados_modelos
                                Modelo Kappa   AUC
1                    Árbol 70:30 0.264 0.780
2                 Árbol 50:25:25 0.062 0.582
3                 Árbol Bootstrap 0.000 0.500
4 Reg. Logística (umbral óptimo) 0.228 0.789
```



Aunque dije que el modelo de regresión es mejor que el del árbol, su Kappa de hecho es ligeramente peor. Aquí salta la pregunta de ¿Qué medida usar como medida principal para determinar qué modelo fue mejor?, creo que la respuesta es un depende. Depende del objetivo de nuestro trabajo. En este caso el objetivo era comparar las métricas de medición del rendimiento entre varios modelos en una tarea de clasificación, por lo que técnicamente el Kappa sería una medida más representativa del rendimiento, ya que como vimos, Kappa ayuda más a decidir bien a quién clasificar como “Sí” o “No” (recordar el caso en que el AUC fue bueno pese a que Kappa=0). Aunque siendo objetivos, creo que lo mejor es no guiarse por una única métrica, lo ideal en este caso es basarse en la combinación de los resultados de ambas, ya que una complementa lo que la otra no considera. En este sentido, la **regresión logística ajustada** ofrece la mejor combinación de sensibilidad, especificidad y discriminación de los datos.

Retomando el contexto de los universitarios, podemos decir que basándonos en el mejor modelo evaluado, ser mujer incrementa significativamente el log-odds de sobregiro ($p < 0.01$). Esto significa que combinado con las otras dos variables de alta significancia, las mujeres que beben moderadamente o beben mucho son el sector más propenso al sobregiro. Esto porque comparado con quienes beben poco, quienes

consumen alcohol de forma moderada tienen mayor riesgo de sobregiro, al igual que quienes beben mucho.

```
> summary(model_logit)

Call:
glm(formula = Overdrawn ~ ., family = "binomial", data = train_1)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               -7.5233     2.3419  -3.212 0.001316 **
Age                        0.2180     0.1124   1.940 0.052374 .
SexMujer                   1.0735     0.3923   2.737 0.006204 **
DaysDrink_CategModerado    1.4968     0.4054   3.692 0.000222 ***
DaysDrink_CategAlto        0.9985     0.4885   2.044 0.040940 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Aquí la pregunta que me hago es, ¿Si esto fuera un trabajo “real”, cómo debería manejar estos resultados en cuanto a la cuestión ética?, quizá, como sucedió con casos similares de la vida real, el señalar a un sector de la población tan específico puede provocar una especie de efecto discriminador en los sistemas contra dicho sector.