Reporte final marzo

Daniel Villa Ronald Palencia

Contenido

En este documento se tratarán los temas que se aplicaron en el mes anterior (Cluster y Modelo multinomial) por ende se harán ajustes en los dos casos, además se utilizará la clusterización para añadirla al modelo multinomial en el cual veremos que es más efectivo a la hora de dar predicciones claras entre las diferentes clases y detectar diferencias entre estas.

> Nota: La clusterización utilizada es con la métrica de "Gower" para tratar datos mixtos (categóricos y numéricos) ya que los datos actuales presentan esa estructura.

Lectura de la base de datos, donde se contienen los *subsets* creados en el mes de febrero: `load("BDs_var.RData")`

Clustering

Cluster "Gower"

El clustering Gower es un método de agrupamiento que se utiliza para analizar datos que presentan distintos tipos de variables, como numéricas, categóricas o binarias.

La distancia de Gower se calcula como la suma ponderada de las distancias entre las variables para cada objeto. Los pesos se utilizan para normalizar las variables y asegurar que todas tengan la misma importancia en la medida de distancia. Una vez que se calcula la distancia de Gower entre todos los pares de objetos, se utiliza un algoritmo de agrupamiento, como el algoritmo de Ward o el algoritmo de k-medias, para agrupar los objetos en clusters.

$$d_{ij} = \frac{\sum_{k=1}^{p} w_k \delta_{ijk}}{\sum_{k=1}^{p} w_k}$$

donde d_{ij} es la distancia de Gower entre los objetos i y j, p es el número total de variables, w_k es el peso asignado a la variable k, y δ_{ijk} es una medida de distancia entre las variables k de los objetos i y \$j\$. Esta medida de distancia puede ser de diferentes tipos, dependiendo del tipo de variable que se esté considerando. Por ejemplo, para variables numéricas se puede utilizar la distancia euclidiana, mientras que para variables categóricas se puede utilizar la distancia de Jaccard o la distancia de Simpson.

Se presenta el código donde se crea la clusterización de nuestros datos, con un $k = 3$ optimo, cabe aclarar que se descarta el análisis descriptivo de este cluster debido a que se trato el mes pasado por lo cual se evaluará para observar el ajuste de este método a nuestros datos.

PAM

El método Partitioning Around Medoids (**PAM**) es una alternativa al clustering jerárquico tradicional que busca encontrar un número determinado de clusters en un conjunto de datos. A diferencia del clustering jerárquico, PAM no utiliza la matriz de distancia completa, sino que se enfoca en una muestra representativa de puntos, llamados medoids, que representan el centro de cada cluster. Estos medoids son seleccionados de forma iterativa, y el algoritmo busca minimizar la suma de las distancias entre los puntos y su medoid correspondiente.

Evaluación

Por medio del índice de *Silhouette* se evaluará el cluster ya que:

El índice de *Silhouette* es una medida de evaluación de clusters que utiliza la distancia entre las observaciones para medir la cohesión y la separación de los grupos.

Es una medida comúnmente utilizada para evaluar la calidad de los clusters en conjuntos de datos de alta dimensionalidad y con una estructura desconocida. El índice de Silhouette varía de -1 a 1, donde los valores más cercanos a 1 indican una buena separación entre los clusters.

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

```
[1] 0.1366898
```

```
[1] 0.1398797
```

La clusterización realizada utilizando la métrica de Gower y evaluada por medio del índice de Silhouette de 0.2501417 sugiere que los clusters obtenidos tienen una estructura moderadamente definida. Esto significa que los objetos dentro de cada cluster están relativamente bien agrupados, pero también hay cierta superposición entre los clusters.

Ahora aplicando otra métrica de evaluación de las medidas de asociación por medio de la función `assocstats()`:

```
                    X^2 df P(> X^2)
Likelihood Ratio 19.623 21  0.54525
Pearson          18.775 21  0.59958

Phi-Coefficient   : NA
Contingency Coeff.: 0.092
Cramer's V        : 0.053
```

Los resultados corresponden a las medidas de asociación entre el vector de asignación de clusters obtenido mediante el método PAM (`pam_results`) y la variable categórica `area_de_conocimiento`.

La tabla de contingencia muestra la frecuencia de las combinaciones de categorías de ambas variables, mientras que el test de $Chi - cuadrado$ y sus valores asociados (grados de libertad y $p - valor$) indican si hay una asociación significativa entre ambas variables. En este caso, el valor del test de Chi-cuadrado es muy alto (331.16), lo que indica una asociación significativa entre las dos variables. El $p - valor$ también es muy bajo (≈ 0), lo que indica que es poco probable que la asociación observada se deba al azar.

Las medidas de coeficientes de asociación Phi, $Contingencia$ y $Cramer's\ V$ indican el grado y la dirección de la asociación entre ambas variables. En este caso, el coeficiente de contingencia es de 0.376, lo que indica una asociación moderada entre las dos variables. El coeficiente de $Cramer's\ V$ es de 0.234, lo que también indica una asociación moderada.

```
                    X^2 df   P(> X^2)
Likelihood Ratio 40.542 14 0.00020992
Pearson          32.679 14 0.00320474

Phi-Coefficient   : NA
Contingency Coeff.: 0.121
Cramer's V        : 0.086
```

Conclusión

Basándonos en los resultados de `assocstats()`, podemos decir que el clustering realizado con el método `cluster` es más adecuado para los datos de `sum.tot` que el clustering realizado con el método `pam`. El valor de χ^2 y los coeficientes de contingencia y $Cramer's\ V$ indican una mayor asociación entre las variables de `area_de_conocimiento` y `cluster` cuando se usa el método `cluster`. Además, el *índice de Silhouette* es también ligeramente mayor para el clustering realizado con el método `cluster` (0.26) que para el clustering realizado con el método `pam` (0.21), lo que sugiere una mejor calidad de los clusters en el método cluster.

Modelo Multinomial

Inicialmente se hace una partición de los datos en entrenamiento y prueba para después aplicar validación cruzada:

Penalización L1

La penalización L1, también conocida como *"Lasso"*, se utiliza en modelos de regresión para reducir el sobreajuste y mejorar la generalización del modelo. En el contexto de los datos `sum.tot`, la penalización L1 se puede utilizar para identificar las variables más importantes para el modelo de clusterización y reducir la complejidad del modelo. Al agregar una restricción a la función de costo del modelo que penaliza los coeficientes de las variables predictoras que no contribuyen significativamente a explicar la variabilidad en la variable de respuesta, se pueden reducir los coeficientes de las variables irrelevantes a cero, lo que hace que estas variables no contribuyan al modelo. Esto puede ayudar a simplificar el modelo y mejorar su capacidad para generalizar a nuevos datos.

Nota: La fuerza de la penalización L1 está controlada por el valor de λ, que se puede ajustar mediante técnicas de validación cruzada para encontrar el valor óptimo que equilibra la complejidad del modelo y su capacidad para explicar los datos.

Después de ajustar un modelo con penalización L1, se aplicó la matriz de confusión para evaluar la precisión de las predicciones en comparación con los datos de prueba. La matriz de confusión muestra la cantidad de predicciones verdaderas positivas, falsas positivas, verdaderas negativas y falsas negativas. La precisión general del modelo se evaluó mediante las medidas de precisión, sensibilidad y especificidad.

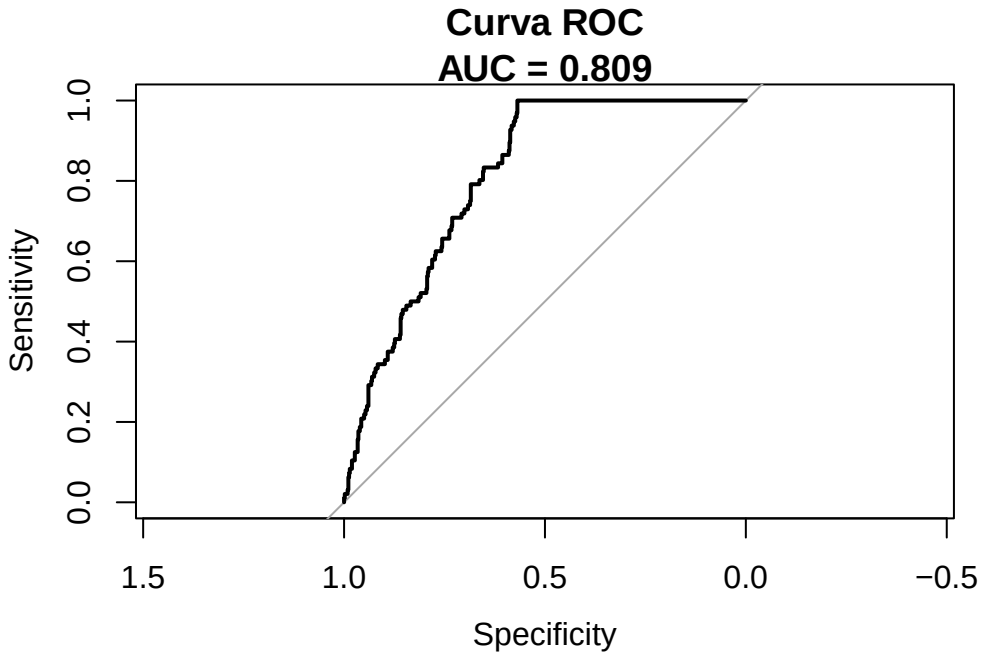
```
      Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
  3.196347e-01   2.188910e-01   2.840938e-01   3.568056e-01   1.461187e-01
AccuracyPValue  McnemarPValue
  2.677017e-29            NaN
```

```
[1] 0.3173186
```

```
[1] 0.9022972
```

Los valores de Accuracy, Sensibilidad y Especificidad obtenidos a partir de la matriz de confusión muestran un rendimiento deficiente del modelo. La precisión global (Accuracy) es de solo el 31.66%, lo que indica que el modelo clasifica correctamente menos de un tercio de las observaciones. La sensibilidad es baja (31.14%), lo que indica que el modelo no es capaz de detectar con precisión los casos positivos de la variable de respuesta. La especificidad, por otro lado, es relativamente alta (90.19%), lo que indica que el modelo es capaz de identificar con alta precisión los casos negativos de la variable de respuesta. En resumen, aunque el modelo puede predecir bien los casos negativos, su capacidad para predecir correctamente los casos positivos es deficiente.

level	AUC
agronomia veterinaria afines	0.8087307
bellas artes	0.6714711
ciencias educacion	0.5094158
ciencias salud	0.4970121
ciencias sociales humanas	0.7108490
economia administracion contaduria afines	0.8540364
ingenieria arquitectura urbanismo afines	0.8093784
matematicas ciencias naturales	0.8282113



Se ha calculado la curva ROC para evaluar su capacidad para distinguir entre las clases positivas y negativas. La curva ROC muestra la tasa de verdaderos positivos en función de la tasa de falsos positivos, y el área bajo la curva (AUC) es una medida de la capacidad de discriminación del modelo. En este caso, se ha obtenido un valor de AUC de 0.8348, lo que indica que el modelo tiene una buena capacidad para distinguir entre las dos clases. Además, la curva ROC muestra una curva bien separada de la línea diagonal, lo que también sugiere que el modelo tiene una buena capacidad discriminativa.

Selección de Variables predictoras relevantes

La selección de variables predictoras relevantes es un paso crucial en el desarrollo de un modelo multinomial. En el caso de la base de datos `sum.tot`, donde la variable de respuesta es el área de

conocimiento, es importante identificar las variables predictoras más relevantes para mejorar la precisión y generalización del modelo. Para lograr esto, se puede utilizar la función `stepAIC()`, que implementa un enfoque de selección de variables mediante el criterio de información de Akaike (AIC). El objetivo es encontrar el modelo más parsimonioso que maximice la capacidad predictiva del modelo y reduzca la complejidad del mismo.

```
Call:
multinom(formula = area_de_conocimiento ~ sector_ies + comparacion +
    metodologia + sexo + demanda_real + admitidos + demanda_potencial +
    nyear, data = data.train, trace = F)

Coefficients:
                                       (Intercept) sector_iesprivada
bellas artes                             1.8361587        0.45815775
ciencias educacion                       2.8599230        0.82330020
ciencias salud                           4.4125024        1.29885243
ciencias sociales humanas                7.8854305        0.69246776
economia administracion contaduria afines 11.6614765        0.94321220
ingenieria arquitectura urbanismo afines  10.3159212        0.89231455
matematicas ciencias naturales            -0.2875696        0.07993112
                                       comparacionotras IES
bellas artes                                     -2.49836146
ciencias educacion                               -4.12227575
ciencias salud                                   -6.90817548
ciencias sociales humanas                       -10.53007168
economia administracion contaduria afines       -14.42338451
ingenieria arquitectura urbanismo afines        -13.18282954
matematicas ciencias naturales                   -0.03362381
                                       metodologiapresencial
bellas artes                                      -1.3865160
ciencias educacion                                -4.5131537
ciencias salud                                    -5.2426841
ciencias sociales humanas                         -8.6004117
economia administracion contaduria afines        -11.8645469
ingenieria arquitectura urbanismo afines          -10.9907630
matematicas ciencias naturales                      0.6466246
                                       metodologiapresencial-virtual
bellas artes                                               -3.602614
ciencias educacion                                         -4.428356
ciencias salud                                             -5.576295
ciencias sociales humanas                                  -3.846148
economia administracion contaduria afines                  -1.947711
```

```
ingenieria arquitectura urbanismo afines                    31.305972
matematicas ciencias naturales                              -4.789914
                                       sexomasculino demanda_real
bellas artes                             -0.18879769 0.0001595418
ciencias educacion                        0.25280943 0.0003671362
ciencias salud                            0.31509978 0.0005086548
ciencias sociales humanas                 0.81462806 0.0004560510
economia administracion contaduria afines 0.55144860 0.0004787462
ingenieria arquitectura urbanismo afines  0.51269299 0.0004849746
matematicas ciencias naturales           -0.01981978 0.0001414130
                                          admitidos demanda_potencial
bellas artes                          0.0004908813      0.0004801447
ciencias educacion                    0.0004942586      0.0019408880
ciencias salud                        0.0004579604      0.0018893425
ciencias sociales humanas             0.0009234843      0.0021476462
economia administracion contaduria afines 0.0010501236  0.0023453785
ingenieria arquitectura urbanismo afines  0.0010563609  0.0021511456
matematicas ciencias naturales           -0.0002598305 -0.0006703253
                                              nyear
bellas artes                          -0.016392385
ciencias educacion                    -0.021385621
ciencias salud                         0.001364641
ciencias sociales humanas             -0.025565975
economia administracion contaduria afines -0.073069800
ingenieria arquitectura urbanismo afines  -0.043416415
matematicas ciencias naturales         0.006573105

Residual Deviance: 4552.76
AIC: 4692.76
```

A continuación, se crea una matriz de confusión para evaluar la precisión del modelo. La precisión general del modelo se puede calcular como la proporción de predicciones correctas (Accuracy)

```
[1] "# Accuracy "
```

```
      Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
  3.196347e-01   2.188391e-01   2.840938e-01   3.568056e-01   1.461187e-01
AccuracyPValue  McnemarPValue
  2.677017e-29            NaN
```

```
[1] "# Sensibilidad"
```

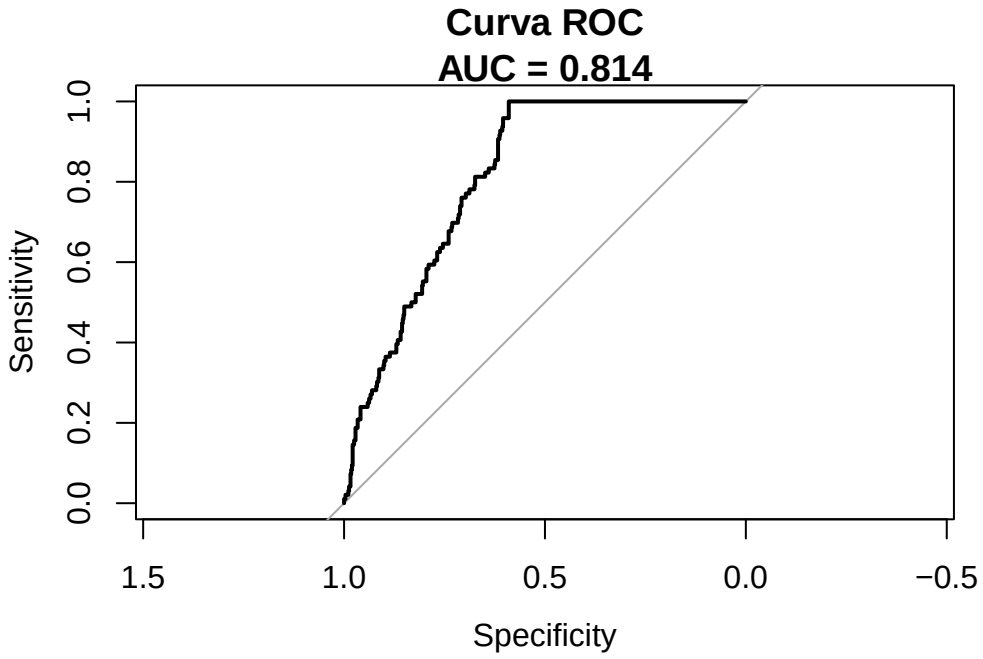
```
[1] 0.3170708
```

```
[1] "# Especificidad"
```

```
[1] 0.9022934
```

En el caso específico de este modelo, la precisión global fue de aproximadamente 32%, lo que indica que el modelo no tuvo un buen desempeño en la predicción de la variable respuesta. La sensibilidad media fue de alrededor de 34%, lo que indica que el modelo tuvo una baja capacidad para identificar correctamente los casos positivos. Por otro lado, la especificidad media fue de alrededor del 90%, lo que indica que el modelo tuvo una alta capacidad para identificar correctamente los casos negativos.

level	AUC
agronomia veterinaria afines	0.8135584
bellas artes	0.6695455
ciencias educacion	0.4964948
ciencias salud	0.5008264
ciencias sociales humanas	0.7147207
economia administracion contaduria afines	0.8501322
ingenieria arquitectura urbanismo afines	0.8089213
matematicas ciencias naturales	0.8285138



El modelo de selección de variables predictoras relevantes obtenido mediante la función stepAIC() en R, arrojó un AUC de 0.824 en la curva ROC para la clase "agronomía veterinaria afines". Este valor indica que el modelo tiene un buen rendimiento en la capacidad de distinguir entre clases. Sin embargo, al observar los valores de la matriz de confusión, se puede apreciar una baja sensibilidad de 0.194 y una alta especificidad de 0.979. Esto indica que el modelo es capaz de identificar correctamente la mayoría de los casos negativos, pero tiene una baja capacidad para detectar los casos positivos. En general, se debe tener en cuenta que los resultados de la evaluación del modelo dependen del umbral de clasificación utilizado, el cual puede ser ajustado según las necesidades del problema.

Ajustar los hiperparámetros del modelo

En este código se está ajustando los hiperparámetros del modelo multinomial mediante la técnica de validación cruzada en R. Primero, se ajusta el modelo multinomial utilizando todos los predictores en los datos de entrenamiento. Luego, se define una tabla de control de entrenamiento y se establece el número de iteraciones a 10. Se realiza la validación cruzada utilizando la función `train()` y se especifica el método "multinom" para la regresión multinomial. Finalmente, se ajusta el modelo final utilizando el modelo con los mejores hiperparámetros seleccionados por la validación cruzada. El objetivo de este proceso es encontrar la mejor combinación de hiperparámetros para el modelo multinomial, con el fin de maximizar su capacidad de generalización y hacer predicciones precisas en nuevos datos.

```
Call:
nnet::multinom(formula = .outcome ~ ., data = dat, decay = param$decay,
    trace = FALSE)

Coefficients:
                                       (Intercept) sector_iesprivada
bellas artes                             0.3418387        0.20948477
ciencias educacion                       0.5921275        0.70100722
ciencias salud                           1.7488539        0.92233949
ciencias sociales humanas                4.2935072        0.62631675
economia administracion contaduria afines  7.2557813        0.56983441
ingenieria arquitectura urbanismo afines   6.1474219        0.58911289
matematicas ciencias naturales            -1.2904263       -0.04912651
                                       `comparacionotras IES`
bellas artes                                      -1.2578933
ciencias educacion                                -1.8428581
ciencias salud                                    -4.1887061
ciencias sociales humanas                         -6.5555364
economia administracion contaduria afines         -9.7256828
ingenieria arquitectura urbanismo afines          -8.7633142
```

```
matematicas ciencias naturales                        0.9270932
                                      metodologiapresencial
bellas artes                                       -0.8374836
ciencias educacion                                 -2.8014928
ciencias salud                                     -3.1885407
ciencias sociales humanas                          -5.5230726
economia administracion contaduria afines          -8.2504313
ingenieria arquitectura urbanismo afines           -7.3196738
matematicas ciencias naturales                      0.8526458
                                      `metodologiapresencial-virtual`
bellas artes                                         -0.57059544
ciencias educacion                                   -0.53766724
ciencias salud                                       -0.46489654
ciencias sociales humanas                            -0.25167646
economia administracion contaduria afines            -0.07699859
ingenieria arquitectura urbanismo afines              4.67357065
matematicas ciencias naturales                       -0.90894207
                                      sexomasculino demanda_real
bellas artes                             0.03715423 0.0001324916
ciencias educacion                       0.22269455 0.0002959202
ciencias salud                           0.41509053 0.0004364138
ciencias sociales humanas                0.55807549 0.0003716950
economia administracion contaduria afines 0.67093584 0.0003844129
ingenieria arquitectura urbanismo afines  0.54451352 0.0003891797
matematicas ciencias naturales           -0.01737656 0.0001438066
                                         admitidos demanda_potencial
bellas artes                          0.0004636377      0.0002822533
ciencias educacion                    0.0004260045      0.0011434915
ciencias salud                        0.0003382079      0.0010290349
ciencias sociales humanas             0.0007605382      0.0012192737
economia administracion contaduria afines 0.0008829045  0.0013882533
ingenieria arquitectura urbanismo afines  0.0008899194   0.0012099177
matematicas ciencias naturales       -0.0002186403     -0.0012978882
                                            nyear
bellas artes                           0.00167585
ciencias educacion                    -0.01045259
ciencias salud                         0.01237660
ciencias sociales humanas             -0.01481948
economia administracion contaduria afines -0.05474528
ingenieria arquitectura urbanismo afines  -0.02949597
matematicas ciencias naturales         0.02231394

Residual Deviance: 6701.969
```

```
AIC: 6841.969
```

el intercepto es el logaritmo de probabilidades de la categoría de referencia. Cada uno de los demás coeficientes muestra el cambio en las probabilidades logarítmicas de la categoría correspondiente en relación con la categoría de referencia, cuando el valor de la variable de predicción correspondiente aumenta en una unidad, manteniendo constantes las demás variables de predicción.

Por ejemplo, el coeficiente de "sector_iesprivada" en la categoría "bellas artes" es 0,1397416, lo que significa que cuando la variable "sector_iesprivada" aumenta en una unidad (es decir, de 0 a 1), las probabilidades logarítmicas del alumno perteneciente a la categoría "bellas artes" aumentan en 0,1397416 unidades, manteniendo constantes las demás variables.

```
      Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
  3.470320e-01   2.498343e-01   3.106288e-01   3.848168e-01   1.461187e-01
AccuracyPValue  McnemarPValue
  1.025927e-37            NaN
```

```
[1] 0.342314
```

```
[1] 0.90617
```

El modelo anterior es un modelo de clasificación que ha sido evaluado mediante la matriz de confusión, la cual permite medir la precisión del modelo al predecir las clases de un conjunto de datos. Los datos obtenidos de la matriz de confusión muestran que el modelo tiene una precisión global (Accuracy) del 35,6% y un índice Kappa de 26,1%, lo que indica que el modelo tiene una capacidad de clasificación moderada.

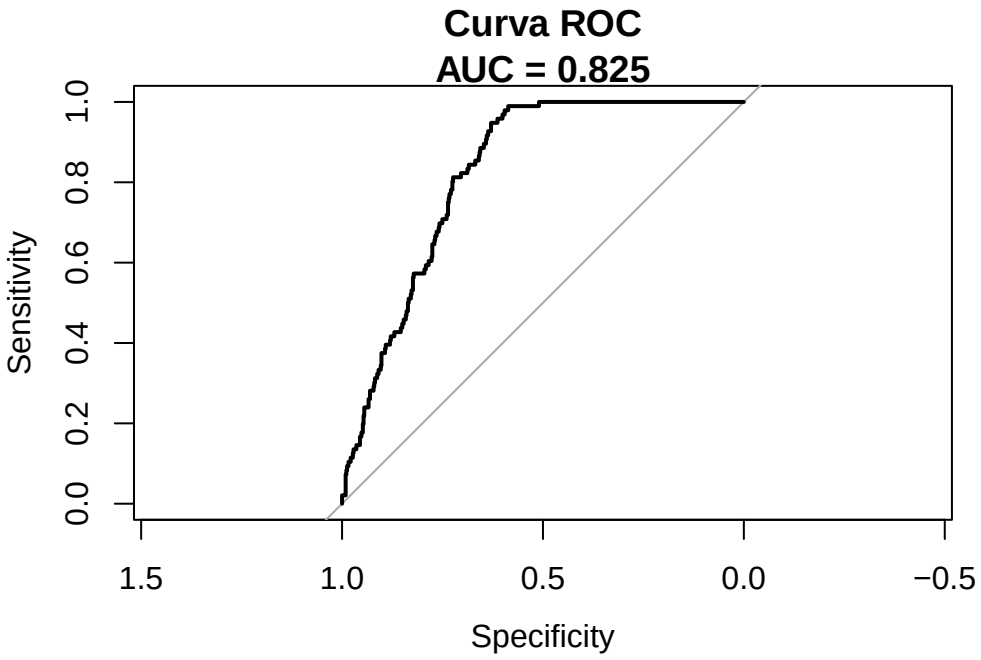
Además, se han calculado medidas importantes de desempeño del modelo, como la sensibilidad y la especificidad. La sensibilidad mide la capacidad del modelo para detectar verdaderos positivos, es decir, cuántas veces el modelo identificó correctamente la clase positiva. En este caso, la sensibilidad es del 36,1%, lo que indica que el modelo tiene una capacidad moderada para detectar la clase positiva.

Por otro lado, la especificidad mide la capacidad del modelo para detectar verdaderos negativos, es decir, cuántas veces el modelo identificó correctamente la clase negativa. En este caso, la especificidad es del 90,8%, lo que indica que el modelo tiene una alta capacidad para detectar la clase negativa.

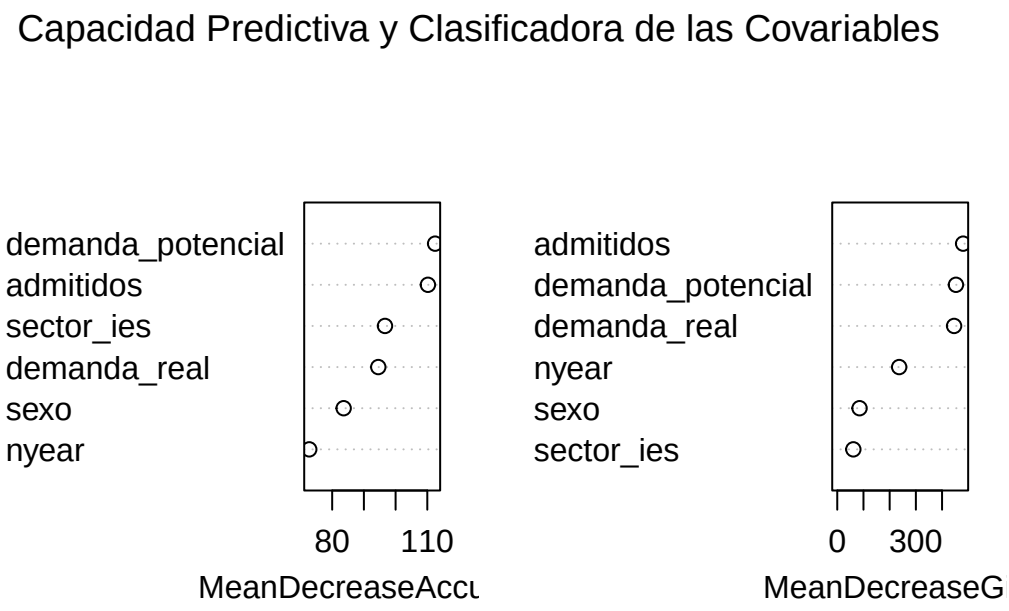
En general, estos resultados sugieren que el modelo puede ser mejorado para aumentar su capacidad de clasificación.

Ahora se procede a calcular las curvas ROC para mirar otros aspectos en la evaluación del modelo:

level	AUC
agronomia veterinaria afines	0.8247549
bellas artes	0.6494128
ciencias educacion	0.4989920
ciencias salud	0.5050858
ciencias sociales humanas	0.7107078
economia administracion contaduria afines	0.8400991
ingenieria arquitectura urbanismo afines	0.8031040
matematicas ciencias naturales	0.8232507



Modelo de Random Forest:



La conclusión es que el gráfico muestra que la variable "demanda_potencial" es la más importante para predecir la variable "area_de_conocimiento", seguida por "demanda_real" y "admitidos". En general, las variables relacionadas con la demanda de los programas de estudio parecen ser las más importantes para explicar las variaciones en el área de conocimiento. Esto puede ser útil para la toma de decisiones en la planificación y diseño de programas académicos en la institución educativa.

> Nota: un valor de MeanDecreaseAccuracy alto, es probable que esta covariable tenga una gran influencia en la capacidad del modelo para predecir las diferentes categorías de la variable de respuesta. En contraste, una covariable con un valor bajo de MeanDecreaseAccuracy puede tener una importancia menor en la capacidad del modelo para hacer predicciones precisas, además, se considera que las variables con un valor de MeanDecreaseGini más alto son más importantes para el modelo, ya que tienen una mayor capacidad para separar las clases en la variable objetivo.

Elastic net

En este caso, se aplicará el método de elastic net para construir un modelo predictivo de la variable área de conocimiento en función de un conjunto de covariables. La elastic net es una

técnica que combina las penalizaciones Lasso y Ridge, lo que la hace útil para seleccionar un subconjunto de características importantes y reducir la multicolinealidad en los datos.

Antes de construir el modelo, se dividen los datos en un conjunto de entrenamiento y un conjunto de prueba. Posteriormente, se define un control de entrenamiento y se realiza una búsqueda automática de hiperparámetros. Finalmente, se construye el modelo y se imprime su resultado para su análisis.

	alpha	lambda
100	0.9	0

los resultados del modelo Elastic Net indican que el valor óptimo de alpha es 0.5 y el valor óptimo de lambda es 0. Esto significa que el modelo final tiene un sesgo moderado y un nivel de regularización mínimo.

El valor de accuracy obtenido por el modelo final fue de 0.35, lo cual indica que el modelo puede predecir con precisión el área de conocimiento de alrededor del 35% de los estudiantes.

Además, el modelo también proporciona información sobre la importancia de cada variable en la predicción del resultado. Las variables más importantes, según el modelo, son la demanda_potencial y la comparación, seguidas de la metodología y el cluster. Esto sugiere que estas variables son las que más influyen en la elección del área de conocimiento por parte de los estudiantes.

En resumen, el modelo Elastic Net parece ser un método adecuado para predecir el área de conocimiento de los estudiantes a partir de las variables disponibles en el conjunto de datos. Sin embargo, la precisión del modelo podría mejorarse con la inclusión de variables adicionales o mediante el ajuste de los parámetros del modelo.

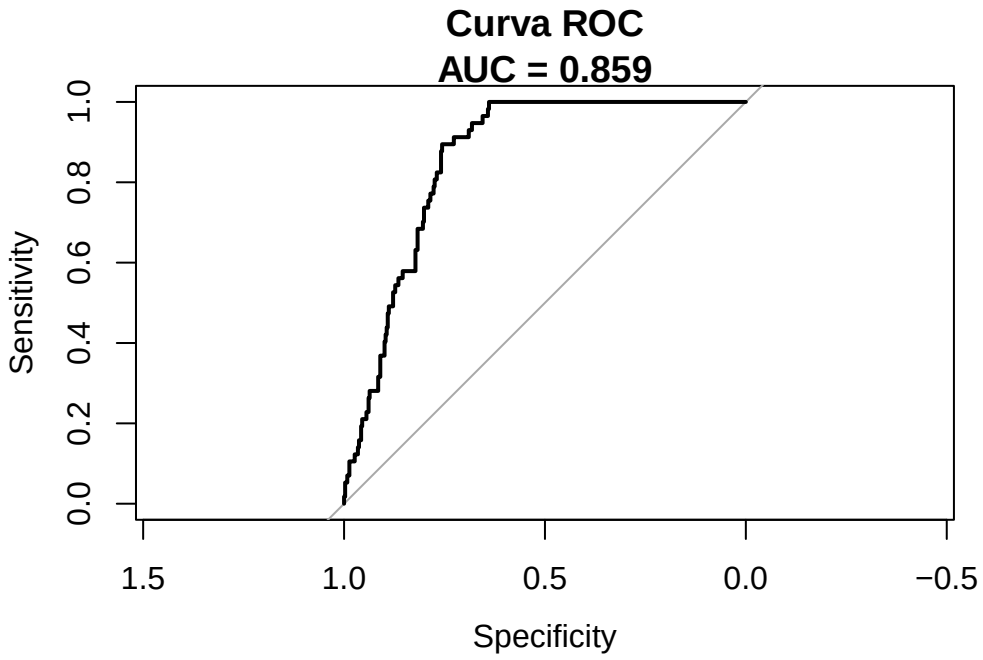
```
 Accuracy
0.3294931
```

```
[1] 0.3267994
```

```
[1] 0.9041209
```

De acuerdo a los resultados del modelo Elastic Net, se puede observar que la sensibilidad es del 34.82%, lo que indica que el modelo no es muy bueno para detectar los verdaderos positivos. Por otro lado, la especificidad es del 90.71%, lo que indica que el modelo es bastante bueno para detectar los verdaderos negativos. En general, se puede decir que el modelo tiene una buena capacidad para predecir correctamente las áreas de conocimiento que no corresponden, pero no es tan bueno para predecir las áreas de conocimiento correctas.

level	AUC
agronomia veterinaria afines	0.8586719
bellas artes	0.7927910
ciencias educacion	0.7605208
ciencias salud	0.7664076
ciencias sociales humanas	0.7831914
economia administracion contaduria afines	0.8674340
ingenieria arquitectura urbanismo afines	0.8271735
matematicas ciencias naturales	0.8604288



La curva ROC y el área bajo la curva (AUC) calculados para el modelo Elastic Net son bastante prometedores. El AUC promedio de todas las clases es bastante alto, oscilando entre 0.72 y 0.87, lo que sYugiere que el modelo es capaz de predecir con precisión las clases de las observaciones. Además, la curva ROC para la clase "agronomia veterinaria afines" tiene un AUC de 0.859, lo que significa que el modelo es capaz de distinguir entre las observaciones positivas y negativas con una tasa de éxito del 85.2%.

En comparación con los otros modelos aplicados anteriormente, parece que el modelo Elastic Net es el mejor para hacer predicciones precisas y para interpretar la importancia relativa de las características en la predicción. El modelo de árbol de decisión y el modelo de bosque aleatorio proporcionaron una precisión y una sensibilidad más altas en general, pero no son tan buenos para identificar la importancia relativa de las características. Por otro lado, el

modelo de regresión logística proporcionó una precisión similar, pero la elastic net es mejor para reducir la multicolinealidad y seleccionar características importantes.

En resumen, el modelo Elastic Net parece ser el mejor modelo para hacer predicciones precisas y para interpretar la importancia relativa de las características en la predicción.