Universidad de
SanAndrés

Big Data

Profesora: Romero, María Noelia

Tutora: Oubiña, Victoria

Trabajo Práctico 3

Cucher Maximiliano, Saucedo Federico & Soares Gache Manuel

Fecha de entrega: 26/5/2024

Parte I: Analizando la base

1. Para identificar si una persona es pobre, a partir de los ingresos de los hogares (calculados mediante la EPH) se establece si éstos tienen capacidad de satisfacer -por medio de la compra de bienes y servicios- un conjunto de necesidades alimentarias y no alimentarias consideradas esenciales. El procedimiento parte de utilizar una canasta básica de alimentos (CBA) y ampliarla con la inclusión de bienes y servicios no alimentarios (vestimenta, transporte, educación, salud, etc.) con el fin de obtener el valor de la Canasta Básica Total (CBT). Mientras que la Canasta Alimentaria es una canasta normativa, la Canasta Básica Total se construye en base a la evidencia empírica que refleja los hábitos de consumo alimentario y no alimentario de la población de referencia. Por lo tanto, para expandir la CBA se utiliza el coeficiente de Engel ($Coeficiente\ de\ Engel\ =\ \frac{Gasto\ Alimentario}{Gasto\ Total}$). Una vez obtenido este coeficiente, se calcula la inversa del coeficiente de Engel (ICE) y calculamos la CBT de la siguiente forma: $CBT = CBA * ICE$. Si los ingresos de las personas/hogar no superan el valor de la Canasta Básica Total, la persona será considerada pobre.

2. b) Lo que hicimos en este inciso fue eliminar todas las variables que contuviesen algún valor negativo, ya que analizando el diccionario de variables, nos dimos cuenta que no podían tomar valores negativos en ninguna de ellas. Esto se debe a que la mayoría de variables suelen ser variables dicotómicas, es decir toman valor 1 o 2, cuantiles, o variables numéricas, como el ingreso que no puede ser negativo, o días/meses/años que tampoco tiene sentido que sean negativos.

c)

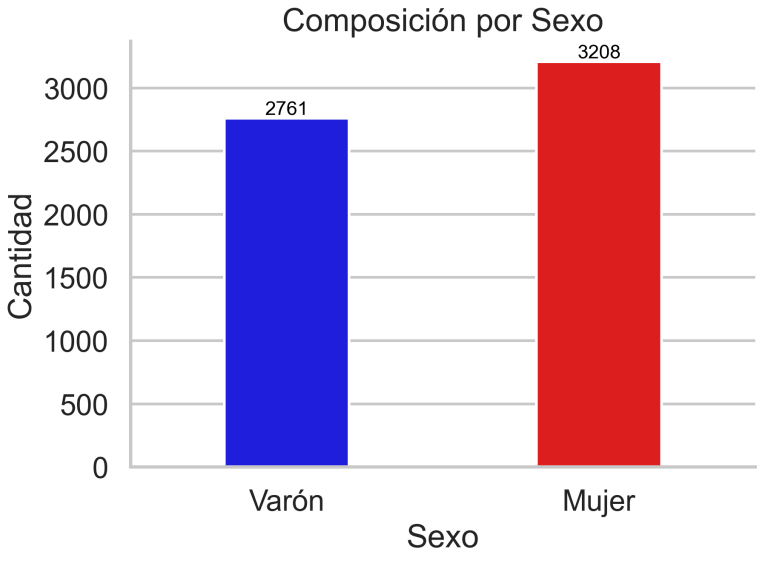

Figura 1: Gráfico de barras de la composición por sexo.

El gráfico que presentamos permite analizar la composición por sexo de la Encuesta Permanente de Hogares (EPH), identificando un total de 2761 hombres y 3208 mujeres en la muestra. La preponderancia de las mujeres puede atribuirse a varios factores. Primero, la esperanza de vida de las mujeres suele ser mayor en comparación con la de los hombres, lo

que podría influir en la composición de la muestra. Además, es plausible suponer que las mujeres tienen una mayor disposición que los hombres a participar en encuestas que abordan temas relacionados con el hogar, el cuidado de la familia o el trabajo doméstico.

d)

Figura 2: Matriz de correlación.

En este gráfico podemos analizar las distintas correlaciones entre las variables CH04 (sexo), CH07 (estado civil), CH08 (cobertura medica), NIVEL_ED (nivel educativo), ESTADO (condición de actividad), CAT_INAC (categoría de inactividad), IPCF (ingreso per cápita familiar). Notamos una correlación positiva de 0.44 entre el estado ocupacional de la persona y su estado civil, lo cual no resulta tan intuitivo. Hay una fuerte correlación positiva de 0.82 entre el estado civil de la persona y su categoría de actividad. Esto tiene sentido, ya que, por ejemplo, es más probable que los estudiantes no estén casados, y los trabajadores casados, debido a factores de edad.

La correlación entre el nivel de educación de la persona y el ingreso per cápita de la familia es de 0.2, bastante menor de lo que podríamos esperar, resultaría plausible esperar que la correlación sea más alta en el sentido de que a mayor nivel de educación, se esperaría que en promedio el individuo consiga trabajos más calificados, y, trabajos más calificados en promedio se observan salarios mayores.

En esta línea, un argumento posible es que el hecho de medir el ingreso per cápita por familia en lugar de por persona atenúa esta correlación. Sin embargo, estudios para argentina como

Gabrielli et al (2017)[1] observa *positive assortative mating*[2] para las características de las parejas, lo que debería reflejarse en un mayor ingreso per cápita en esos hogares. Esto podría sugerir que las personas no están declarando sus ingresos de manera precisa.

Resulta relevante observar que este correlograma no es completamente interpretable. Esto se debe a que deberíamos crear variables dummy para cada una de las categorías de las variables, lo cual permitiría comprar de manera más granular las categorías y obtendremos análisis más ricos en términos de información y posibles argumentos.

e) Cantidad de desocupados en la muestra: 226.

Cantidad de inactivos en la muestra: 2507.

Media de IPCF para ocupados: 190809.67.

Media de IPCF para desocupados: 61605.87.

Media de IPCF para inactivos: 93740.53.

Resulta esperable que la media del Ingreso Per Cápita Familiar (IPCF) sea superior para los individuos inactivos en comparación con los desocupados. Esta diferencia podría explicarse por el hecho de que los inactivos podrían no sentir la necesidad de buscar empleo, dado que cuentan con un ingreso familiar relativamente alto (podría ser que su pareja tenga ingresos altos, por ejemplo, lo que trae como consecuencia que pueda permitirse no trabajar). Por el contrario, los desempleados se encuentran en búsqueda activa de trabajo, posiblemente motivados por un IPCF más bajo, aunque enfrentan dificultades para encontrarlo.

Un ejemplo ilustrativo de esta situación podría ser el de una persona, ya sea hombre o mujer, que opta por no trabajar porque su pareja percibe un salario suficiente para sostener el hogar. En contraste, un desocupado podría ser alguien que, a pesar de necesitar y buscar activamente empleo, no logra encontrarlo. Este escenario explica por qué el ingreso per cápita familiar de un desocupado suele ser menor en comparación con el de alguien que se mantiene inactivo.

En esta línea de razonamiento, también es esperable que el IPCF para los ocupados sea mayor que los desocupados e inactivos, porque en promedio, esperaríamos que un salario más al ingreso familiar haga que sea más grande en comparación de un salario menos porque el individuo ya sea que está desocupado o inactivo.

3. Cantidad de ceros en la variable 'ITF': 1618. Esta manera de medir la cantidad de personas que no dieron su ingreso es un tanto ineficiente, ya que toma a las personas que realmente no tienen ingreso (aunque deberían ser pocas o hasta nulas) como missing values.

5. Cantidad de personas identificadas como pobres: 1789. Porcentaje de personas identificadas como pobres: 41.12%

---

[1] Gabrielli, M. F. y M. Serio. (2017). "Testing Assortative Mating: Evidence from Argentina". Revista de Análisis Económico, v. 32, n. 2, p. 109-129.

[2] Aumento en la probabilidad de que un individuo esté en pareja con otra persona con características socioeconómicas y educativas similares.

Parte II: Clasificación

3. Antes de proceder al análisis comparativo de los tres modelos, es crucial mencionar la importancia de trabajar con un conjunto de datos escalado, dado que nuestro objetivo es realizar una comparación efectiva entre ellos. Por esta razón, hemos empleado el comando Standard Scaler para estandarizar el dataset, asegurando que las variables tengan una escala común y no distorsionen los resultados del análisis.

Modelo 1: Regresión Logística

AUC (área bajo la curva): 0.983.

Accuracy: 0.95.

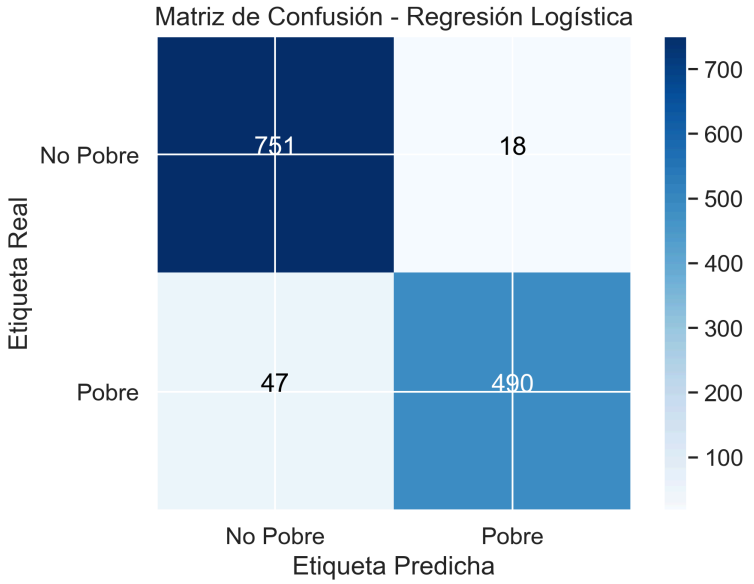

Figura 3: Matriz de Confusión Regresión Logística.

Antes que nada es bueno recordar cómo se interpreta una matriz de confusión. En el eje vertical tenemos los valores reales de predicción, mientras que en el eje horizontal tenemos los valores predichos. En el primer cuadrante (No pobre, No pobre) marca los verdaderos negativos. El segundo cuadrante (Pobre, No Pobre) marca los falsos positivos dado que la regresión logística predice que es pobre cuando realmente el individuo era no pobre. El tercer cuadrante (No Pobre, Pobre) marca los falsos negativos con un análisis similar al anterior. Finalmente, el cuarto cuadrante (Pobre, Pobre) marca los verdaderos positivos.

El objetivo es tener un modelo que tenga la mayor cantidad de verdaderos positivos y verdaderos negativos posible, con la menor cantidad de falso positivo y de falso negativo, para que nuestra predicción sea lo más certera posible (el escenario *first best* es aquel sin falsos positivos ni falsos negativos). Obviamente, ese objetivo es casi inalcanzable. Al tener en cuenta que la presencia de errores en la predicción es casi un hecho, los investigadores deben decidir cuánto priorizar la minimización de falsos negativos o falsos positivos en

conjunto (dado que existe un *trade-off*). La decisión óptima de estas variables estará relacionada a la naturaleza y el contexto del experimento.

En este caso, vemos que los falsos positivos son únicamente 18, mientras que los falsos negativos son 47. Podemos ver que en ambos casos, la muestra es muy pequeña. Esto se ve reflejado en un nivel de accuracy de 0.95. El modelo ha categorizado correctamente a las observaciones en un 95% de los casos.

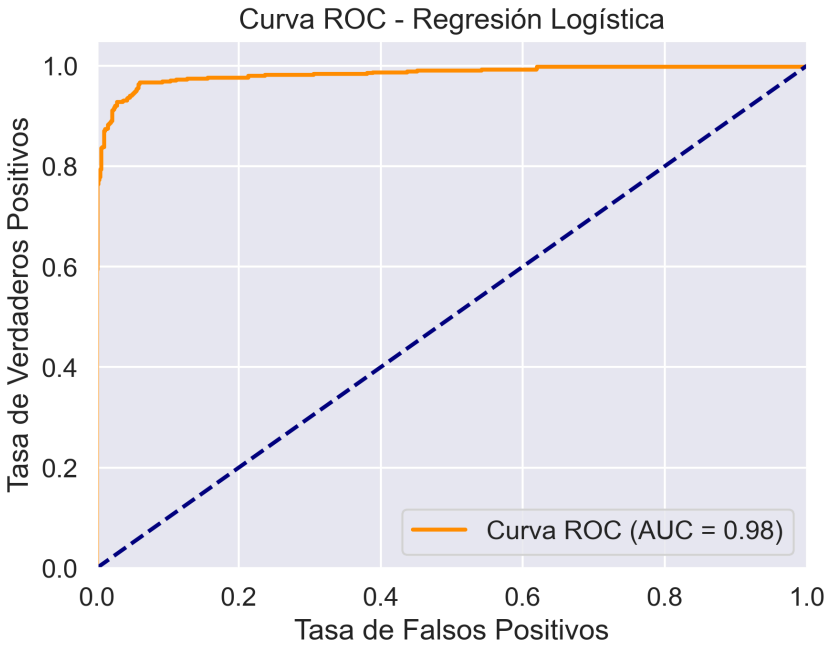

Figura 4: Curva ROC - Regresión Logística.

En este caso, tenemos una curva de ROC. En el eje vertical tenemos la tasa de verdaderos positivos y en el eje horizontal la tasa de falsos positivos. La línea azul es un modelo predictivo sin precisión que a medida de que aumenta la tasa de falsos positivos en una unidad aumenta la tasa de verdaderos positivos en una unidad. La curva naranja traza las distintas tasas de verdaderos positivos y falsos positivos para cada umbral.

Lo ideal sería que quede la mayor área posible entre la recta azul y la curva naranja, es decir que la curva naranja esté lo más cercano a la parte superior izquierda. En nuestro gráfico tenemos que la curva está muy pegada al eje vertical. Pareciera partir del 0.8, es decir que yo puedo casi eliminar la tasa de falsos positivos, con una tasa de verdaderos positivos de 0.8. A la vez, puedo eliminar los falsos negativos, con una tasa de falsos positivos bastante baja, como del 0.2/0.4.

El modelo cuenta con un AUC muy elevado, de 0.983. Indica que el modelo es extremadamente efectivo en la asignación de probabilidades más altas a los verdaderos positivos que a los falsos positivos, a través de la mayoría de los umbrales de decisión.

Modelo 2: Análisis de Discriminante Lineal

AUC: 0.60.

Accuracy: 0.59.

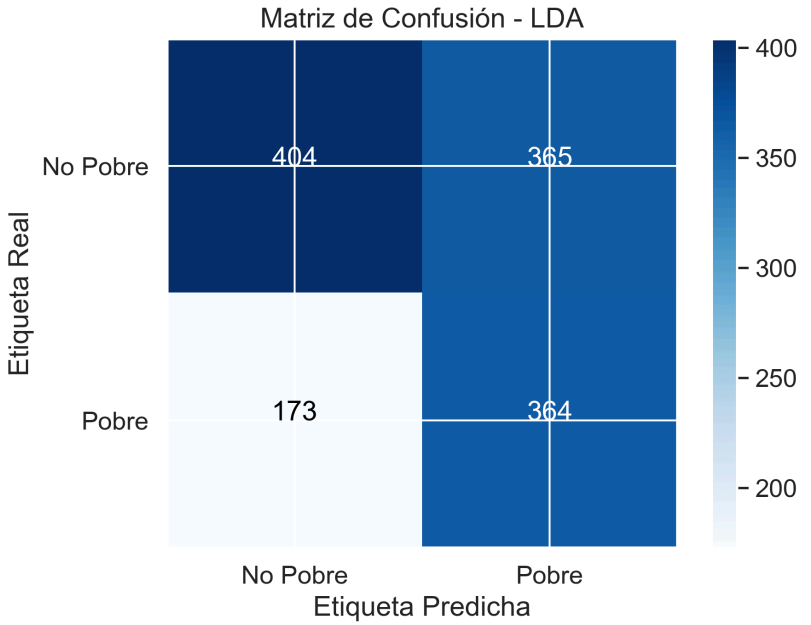

Figura 5: Matriz de Confusión - LDA

En este caso, notamos una alta frecuencia de falsos negativos (365). Podríamos decir que este modelo tiene un sesgo a la hora de identificar a las personas como pobres, ya que muchas veces comete un error al categorizarlos de esta manera. El modelo comete más errores del tipo falsos negativos que falsos positivos. Categoriza 365 veces como pobre a personas que no lo son, y 173 veces como no pobres a personas que sí son pobres.

Esta falla de precisión se ve reflejado en el valor de accuracy de este modelo, 0.59, un número bastante más bajo que el del modelo de regresión logística.

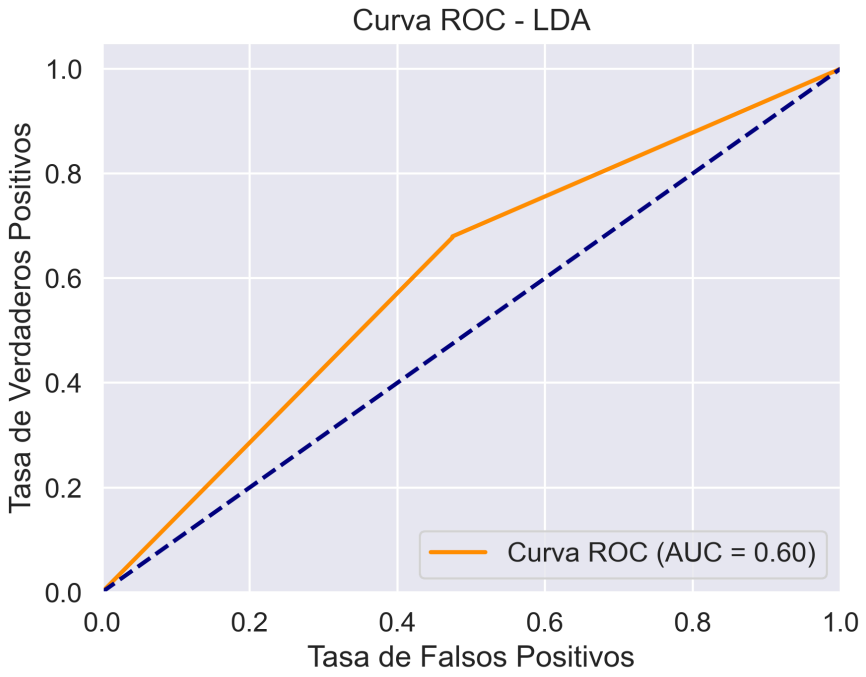

Figura 6: Curva ROC - LDA

La interpretación de la curva ROC y el AUC de 0.60 sugiere que el modelo presenta una capacidad de discriminación moderada, pero claramamente inferior comparado con un modelo de regresión logística que exhibe un rendimiento significativamente mejor. La curva ROC revela que para alcanzar una tasa de verdaderos positivos de 0.8, el modelo incurre en una tasa de falsos positivos de 0.6. Esto contrasta notablemente con el modelo de regresión logística, que logra la misma tasa de verdaderos positivos pero sin falsos positivos, indicando un rendimiento ideal.

Un AUC de 0.62 implica que, en términos de poder distinguir entre clases positivas y negativas, el modelo es ligeramente mejor que una clasificación aleatoria, pero aún deja mucho que desear en términos de eficacia. Este nivel de AUC muestra que, aunque el modelo puede identificar correctamente las clases en un 60% de las ocasiones, su capacidad para hacerlo de manera consistente y sin errores significativos es limitada.

Modelo 3: KNN con k = 3

AUC: 0.71.

Accuracy: 0.448.

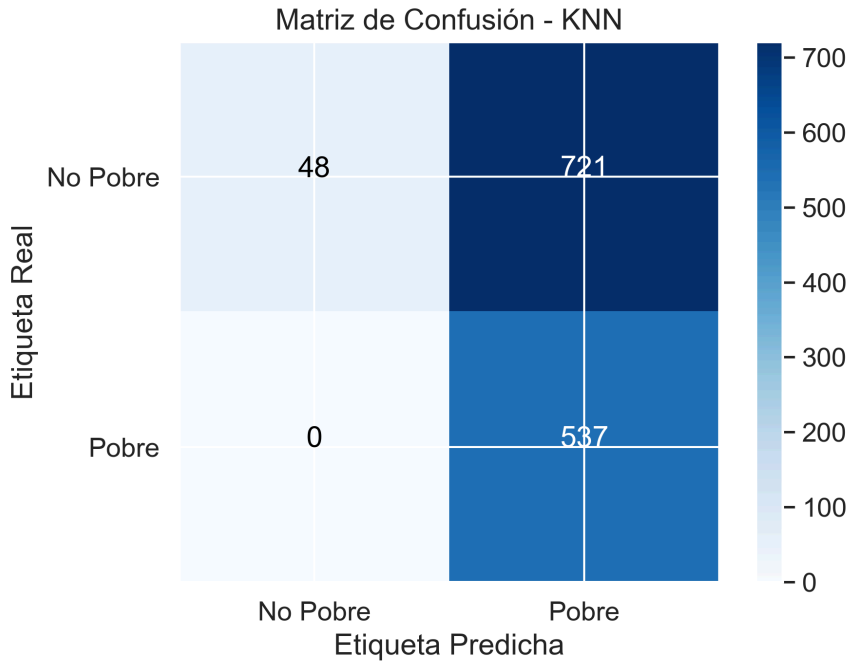

Figura 7: Matriz de Confusión KNN

La matriz de confusión para este modelo revela un número especialmente bajo de verdaderos negativos, lo que indica que el modelo tiende a clasificar erróneamente a los individuos que no son pobres como si lo fueran. Esto sugiere que el modelo tiene un sesgo considerable hacia la clasificación de casi todos los individuos en la muestra como pobres, fallando así en identificar correctamente a aquellos que no lo son.

Un nivel de precisión (accuracy) de 0.448, el más bajo entre los tres modelos evaluados, refleja esta tendencia del modelo a cometer errores de clasificación significativos. Esta baja

precisión sugiere que, en la mayoría de los casos, el modelo no es eficaz para predecir correctamente la condición de pobreza o no pobreza de los individuos.

Sin embargo, este modelo podría tener un valor práctico en un escenario hipotético donde sea crítico evitar clasificar a cualquier persona pobre como no pobre, es decir, maximizando la sensibilidad a costa de la especificidad. En un contexto donde el costo de no identificar a un pobre como tal es mucho mayor que el costo de clasificar erróneamente a un no pobre como pobre, este modelo podría ser útil, aunque con importantes limitaciones en términos de la cantidad de falsos positivos y su impacto.

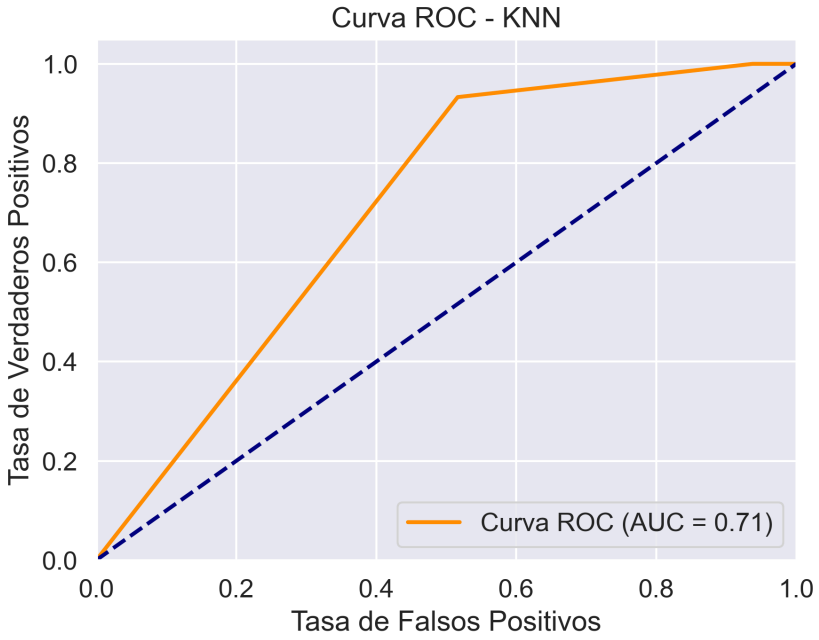

Figura 8: Curva ROC - KNN

En cuanto a la curva de ROC, notamos que el área por debajo de la línea naranja es más grande que en la regresión lineal, pero más baja que el caso de la regresión logística. Para una tasa de verdaderos positivos de 0.8, tenemos una tasa de falsos positivos de 0.45 aproximadamente.

El AUC es de 0.7104. Notamos como un modelo que etiqueta a casi todas las personas de la muestra como pobres, tiene un AUC mayor que el de la regresión lineal. Esto indica que ambos modelos son bastante malos a la hora de predecir correctamente, y que hay más pobres que no pobres en la muestra.

4. En conclusión, notamos que el modelo de regresión logística es el que mejor predice, ya que tiene un AUC y una precisión (accuracy) mayores. Esto puede deberse a que la variable dependiente tiene una relación lineal con las variables independientes. Además, el LDA puede no ser efectivo debido a que las variables no tienen una distribución normal, las clases no comparten la misma matriz de covarianza, además este estimador resulta ser poco flexible. Por otro lado, nuestras variables tienen una alta correlación entre ellas, lo que hace que KNN no sea eficaz. En esta línea, otra característica por la cual vecinos cercanos puede no ser bueno prediciendo es por la maldición de dimensionalidad; el problema que puede tener esta

variante es que cuanto tenemos muchos predictores no resulta relevante pensar cómo definir la distancia entre múltiples dimensiones y es necesario pensar si hay que dar diferentes ponderaciones a los predictores según su importancia para predecir si es pobre o no.

5. Proporción de personas que no respondieron identificadas como pobres: 0.423. Como podemos observar, según nuestra estimación con el modelo Logit el 42% de las personas que no respondieron son identificadas como pobres. Esto resulta convincente ya que como mencionamos anteriormente, el porcentaje de personas pobres en la muestra respondieron, fue de 41,12%. Por lo tanto, tendría sentido que el porcentaje de pobres sea más o menos el mismo.

6. Hay algunas variables de la muestra que pueden no ser tan relevantes a la hora de predecir si una persona es o no pobre. Si incluimos estas variables en la muestra, como hicimos en los ejercicios anteriores, estaríamos agrandando la varianza de nuestros estimadores sin bajar el sesgo, lo que aumentaría el MSE.

Adicionalmente, no resultaría correcto usar todas las variables para KNN por problemas de correlación entre predictores y problemas de dimensionalidad como resaltados anteriormente.

Es por eso que vamos a eliminar estas variables que nosotros consideramos que no son relevantes, correremos el modelo Logit de nuevo y analizaremos cómo es que cambian las medidas de precisión.

Notar que una manera más precisa y eficaz de hacer esto es utilizando Lasso o Ridge, donde estas metodologías eliminaran (en Ridge no se elimina ninguna variable), o le quitan peso a las variables que menos útiles sean.

Las variables que vamos a sacar son: TRIMESTRE, CH05, CH13 y CH14 (notar que si son variables relevantes, pero están muy correlacionadas entre sí estas variables con CH12, entonces al quedarnos solo con CH12 ya estamos captando el componente de educación. De todos modos seria aun mejor hacer un índice de educación en donde tengamos en cuenta todas esas variables). PP03C por la misma explicación que antes. PP11R.

Precisión del modelo logit con todas las variables: 0.95.

AUC del modelo logit con todas las variables: 0.983.

Precisión del modelo logit con variables seleccionadas: 0.96.

AUC del modelo logit con variables seleccionadas: 0.992.

Al eliminar estas variables irrelevantes, o altamente correlacionadas, hemos logrado mejorar la precisión y el AUC del modelo logit. En cuanto a la precisión, pasamos de 95% a 96% y en cuanto al AUC pasamos de 0.983 a 0.992. Si bien no son cambios de gran magnitud, podemos observar mejoras. Nos queda pendiente ver cómo podríamos mejorar esto utilizando Lasso o Ridge.