Trabajo Práctico N°2

Big Data

Universidad de
SanAndrés

Alumnos:

Juan Diego Barnes

Fransisco Legaspe

Rodrigo Martin

Profesora: Maria Noelia Romero

Tutora: Victoria Oubiña

1. Parte I: Analizando la base

Ejercicio 1

Las nociones de pobreza e indigencia empleadas por el INDEC para el cálculo de la pobreza se corresponden con el método de medición indirecta, denominado también "método de la línea". Se definen dos linea, la linea de "línea de indigenciaz "linea de pobreza", donde un hogar es pobre o indigente si su ingreso ,recolectado por la Encuesta Permanente de Hogares (EPH), logra superar o no estos umbrales o lineas.

La "línea de indigencia"(LI) es un concepto que se utiliza para determinar si los hogares tienen ingresos suficientes para adquirir una canasta de alimentos que satisfaga las necesidades mínimas de energía y proteínas. Los hogares que no pueden cubrir este umbral mínimo se consideran indigentes. Para calcular la LI, se utiliza una canasta básica de alimentos de costo mínimo (CBA) basada en los hábitos de consumo de la población de referencia, según los datos de la Encuesta de Gastos e Ingresos de los Hogares (ENGHo). D

El enfoque de "Línea de Pobreza" (LP) busca establecer si los hogares cuentan con ingresos suficientes para cubrir una canasta de alimentos capaz de satisfacer un umbral mínimo de necesidades energéticas y proteicas, lo que se mide con la canasta básica alimentaria (CBA) y otros consumos básicos no alimentarios. Si a los consumos básicos que se encuentran en la canasta básica alimentaria, le sumamos los otros consumos básicos no alimentarios, obtenemos la Canasta Básica Total (CBT), la cual es contrastada con los ingresos de los hogares relevados por la Encuesta Permanente de Hogares(EPH).

Para la línea de pobreza se aplia la CBA, incluyendo bienes y servicios no alimentarios (vestimenta, transporte, etc.) con el objetivo de construir el valor de la Canasta Básica Total (CBT). Dado que los requerimientos nutricionales varían según la edad, el sexo y la actividad de las personas, se realiza una adaptación que considera las características individuales en relación con estas variables. Se toma como referencia al varón adulto de 30 a 60 años con actividad moderada, al que se denomina .ªdulto equivalentez se calcula el valor de la canasta para este. Luego se repondrá los participantes del hogar según su edad y sexo en base a esta equivalencia para determinar el nivel de ingresos necesario del hogar para superar la linea de la pobreza o indigencia.

En tanto que las líneas se construyen por hogar, el valor de las canastas que estas suponen debe ser contrastado con el ingreso total familiar del hogar, lo que permite clasificarlos en hogares indigentes (ingresos totales debajo de los necesarios para cubrir la CBA), pobres no indigentes (ingreso totales pueden cubrir la canasta básica alimenticia pero no la total), pobres (incluye las dos anteriores) y no pobres (ingresos totales pueden cubrir la canasta básica total), extendiéndose esa caracterización a cada una de las personas que los integran.

Ejercicio 2

En este trabajo utilizaremos solo los datos de EPH del Gran Buenos Aires (CABA o Gran Buenos Aires). Es decir, nos quedamos solamente con la información donde la variable AGLOMERADO es igual a 32 e igual a 33 respectivamente.

Por otro lado, para tener una muestra bien defininda y un análisis robusto, eliminamos los valores que no tienen sentido como valores negativos de edades e ingreso.

Filtramos por sexo y reemplazamos los nombres de la encuesta por Varón y Mujer respectivamente para obtener una mejor exposición de la composición por sexo de la muestra. La misma se presenta a continuación:

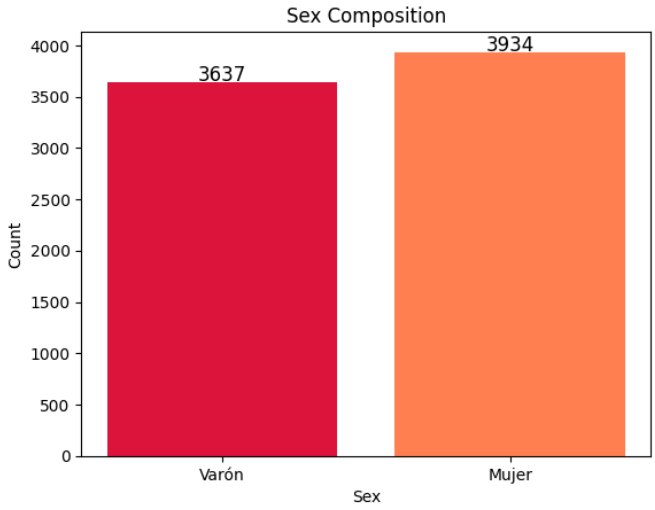

Figura 1: Composición por sexo de la muestra tomada.

Como se puede observar, en nuestra muestra hay una cantidad mayor de mujeres que varones. Siguiendo con la descripción de la muestra, nos interesa realizar una matriz de correlación para las variables `CH04`, `CH07`,`CH08`,`NIVEL_ED`,`ESTADO`,`CAT_` y `IPCF`.

Pero como muchas de estas variables son categóricas no ordinales, graficar estas como están en la muestra no brinda ninguna posible interpretación. Por esto, aunque bajo cierta subjetividad realizamos cierta agrupación o redefinimos algunas categorías para lograr una interpretación de los coeficientes de la matriz. Para la variable que indica el estado civil, `CH07`, la redefinimos como una dummy que toma valor 1 si la persona esta en pareja (Casado/Unido) y 0 en caso contrario. la variable `CH08`, se redefinió como 1 si la persona tiene algún tipo de cobertura media y cero de caso contrario. La variable estado se redefinió como una variable indicadora de si la persona se encuentra ocupada. Para la variable de categoría de inactividad (`CAT_INAC`) generamos dummies para las tres categoría que consideramos mas interesantes de interpretar Rentista Jubilado/Pensionado y Estudiante.

La matriz de correlaciones es la siguiente:

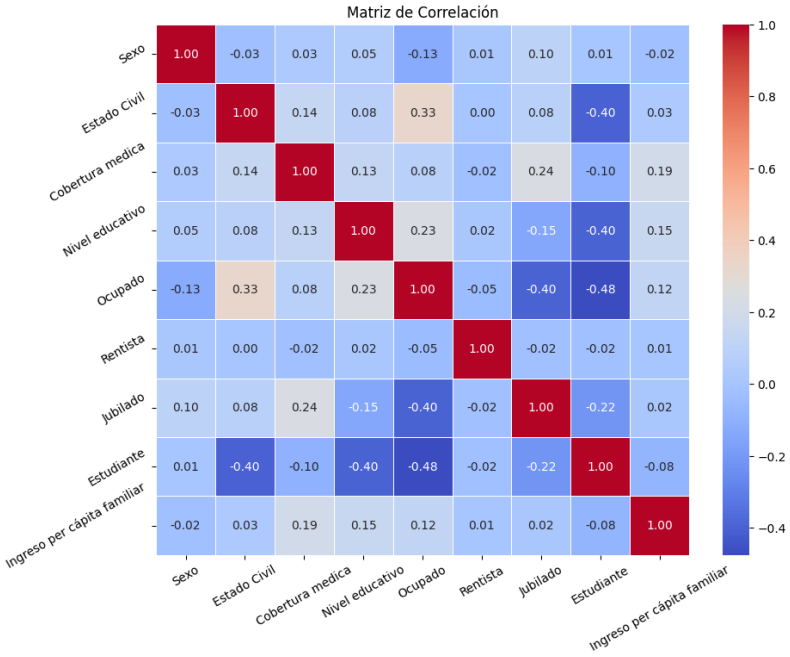

Figura 2: Composición por sexo de la muestra tomada.

En la diagonal principal de la matriz se encuentra la relación de una variable con sigo mismo por eso el valor es 1.00. Por otro lado, podemos observar que las variables que poseen una mayor correlación en valor absolutos son Estudiante y Ocupado (-48 %), Estudiante y Estado Civil (-40 %), y Estudiante y Nivel educativo (-40 %). Por otro lado, todas las variables con respecto a Rentista, tiene correlaciones muy bajas al igual que la variable Sexo las cuales reflejan las variables con menores correlaciones con las demás.

Siguiendo con el análisis nos gustaría saber cuantos desocupados, inactivos y ocupados hay en la muestra. Para eso generamos variables asociadas a los valores posibles de la variable `ESTADO`, la cuál toma valor 1 si la persona es ocupada, 3 si es una persona inactiva y 2 si es desocupado. Por otro lado, calculamos la media de ingreso per capita familiar según el estado laboral. Los resultados obtenidos para la cantidad de persona en cada estado laboral son los siguiente:

- Ocupados: 3523 personas
- Desocupados: 286 personas
- Inactivos: 2837 personas

Con respecto a las medias del ingreso por estado laboral, los resultados son los siguientes:

- Media ingreso per cápita familiar (Total): $ 47720.43
- Media ingreso per cápita familiar (Ocupados): $ 59579.4429

- Media ingreso per cápita familiar (Desocupados): $ 25536.02
- Media ingreso per cápita familiar (Inactivos): $ 40067.99

Ejercicio 3

Una vez vista la cantidad de personas en cada estado laboral, nos interesa entender cuantas personas no reportan sus ingresos. Nos interesa saber cuantas personas respondieron acerca de sus ingresos:

- Cantidad de personas: 7571
- Cantidad de personas que no respondieron ingresos: 3390
- Cantidad de personas que respondieron ingresos: 4181

Como podemos observar, uno de los grandes problema de la EPH es la cantidad de personas que no reportan sus ingresos la cual es del 44 % en nuestro periodo de análisis. A partir de esto separamos la muestra en dos, una primera con los que `respomdieron` y otra con los que `no respomdieron` (consideramos los ingresos iguales a cero como no respuesta).

Ejercicio 4 y 5

Una vez realizado esto y sabiendo que la Canasta Básica Total para un adulto equivalente en el Gran Buenos Aires en el primer trimestre de 2023 es aproximadamente $57.371,05, y tomando la tablas de `adulto equivalente`, calculamos el `ingreso_necesario` que es el producto entre este valor por ad_equiv_hogar para ver el valor mínimo que necesita ese hogar para no ser pobre. Gracias a este paso, pudimos generar una columna llamada `pobre` que es 1 si el ITF es menor al ingreso necesario que necesita esa familia y 0 sino. Los resultados fueron los siguientes:

- El porcentaje de pobres en la muestra fue del 34.7 %.

2. Parte II - Clasificación

2.1. Ejercicio 1

Nuestro objetivo en esta parte del trabajo es intentar predecir si una persona es o no pobre utilizando datos distintos al ingreso, dado que muchos hogares son reacios a responder cuánto ganan. Entrenaremos y validamos diferentes modelos utilizando los datos de los que `respondieron`, y luego realizamos la predicción sobre los que `no respondieron`

Para esto realizamos unos pasos preliminares antes de realizar el análisis:

1. Cargamos las bases generadas bajo en nombre respondieron y no respondieron explicadas anteriormente.
2. Eliminamos las variables relacionadas a los ingresos como : ingresos de la ocupación principal de los asalariados, ingresos de la ocupación principal, ingresos de otras ocupaciones, ingreso total individual, ingresos no laborales, ingreso total familiar, ingreso per cápita familiar.
3. Eliminamos las columnas relacionadas a las variables que generamos antes, estas fueron `adulto`$*equiv*, `ad__equiv_hogar` y `ingreso_necesario`.

2.2. Ejercicio 2

Una vez realizados los pasos preliminares, partimos la base `respondieron` en una base de *train* y otra base de *test*.[1] Para esto utilizamos el criterio de que la base de entrenamiento debe comprender el 70 % de los datos[2].

2.3. Ejercicio 3

Pasaremos a implementar el método logit, Análisis de determinante lineal y KNN con k=3 para la predicción y reportaremos la matriz de confusión, la curva ROC y los valores de AUC y de Accuracy para cada uno.

2.3.1. Logit

La matriz de confusión para la estimación Logit (con un máximo de iteraciones de 1000) es la siguiente:

[1] Como los algoritmos no funcionan si hay missing values, eliminamos las variables que tengan mas del 50 % de missing values y luego eliminamos las observaciones restantes con missing values del resto de las variables

[2] Utilizamos la semilla (random state instance) 201

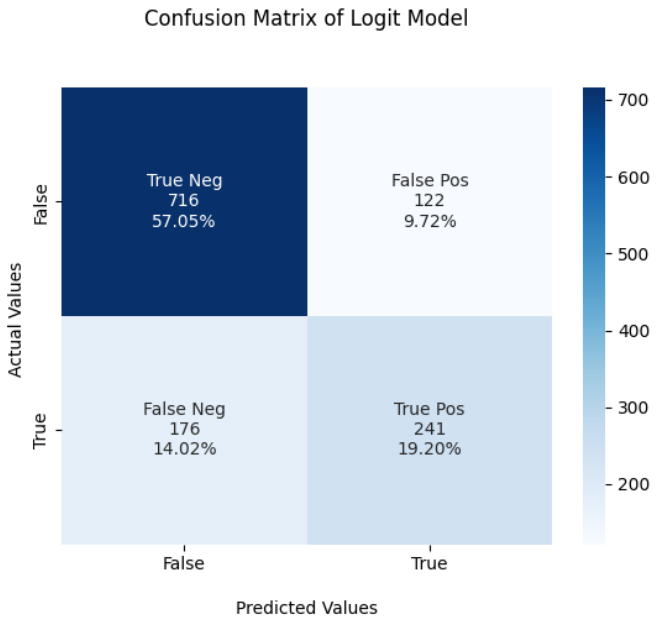

Figura 3: Matriz de Confusión para Logit.

Como podemos observar, los verdaderos negativos contienen el mayor porcentaje mientras que los falsos positivos el menor. A continuación exponemos la curva ROC para complementar el análisis:

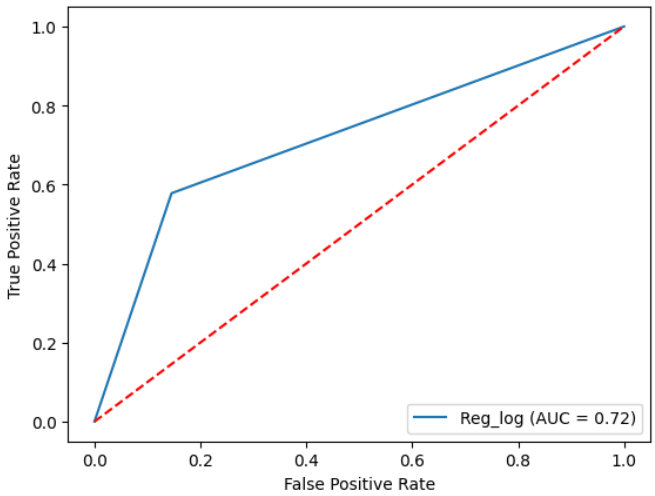

Figura 4: Curva ROC para Logit

Como podemos ver la curva ROC se encuentra por encima de la curva de 45°, reportando una valor de AUC de 0.72.

2.3.2. Análisis de discriminante lineal

Ahora utilizaremos el modelo de análisis de discriminante lineal (LDA por sus siglas en ingles), a continuación presentamos las diferentes medidas predicción:

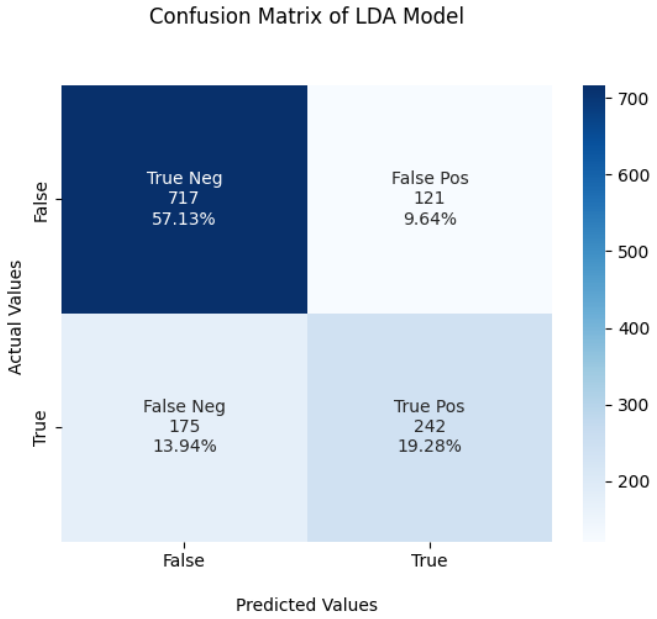

Figura 5: Matriz de Confusión para LDA.

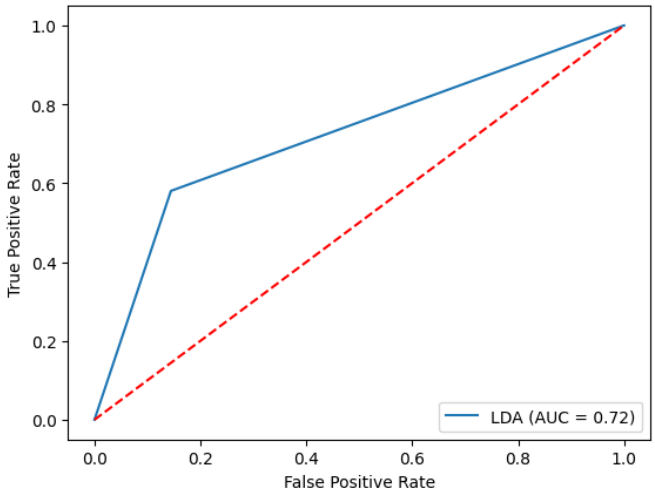

Figura 6: Curva ROC para LDA

Podemos observar que no hay grandes ganancias del Análisis discriminaste de datos por sobre el modelo logit, su AUC y Accuracy Score son prácticamente iguales. Sin embargo, podemos observar que en la matriz de confusión del modelo LDA, el porcentaje de verdaderos positivos y verdaderos negativos aumento mientras que el de falso negativos y falso positivos bajo.

2.4. K Nearest Neighbor con k=3

A continuación presentamos las estimaciones mediante el alogitmo de KNN,

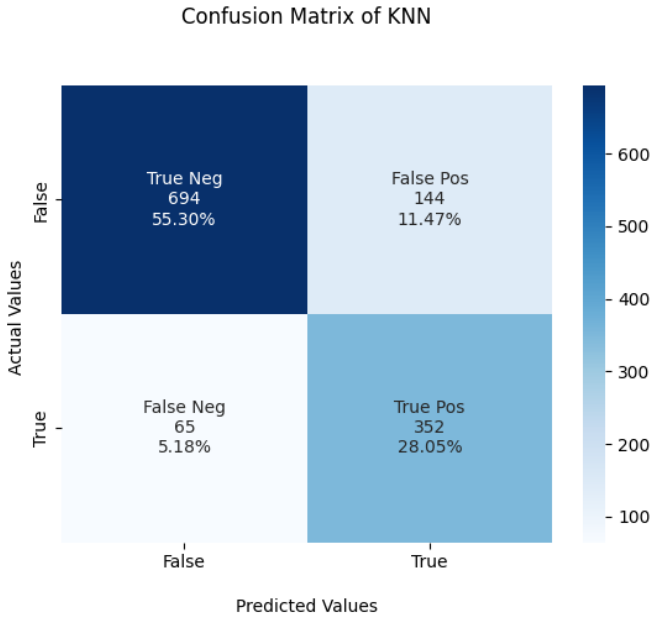

Figura 7: Matriz de Confusión para KNN con k=3.

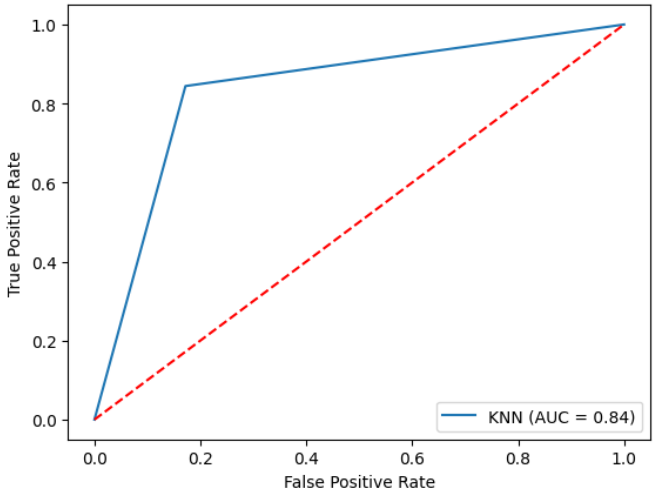

Figura 8: Curva ROC para KNN con k=3

Podemos observar que el método de KNN es el que presenta mejores resultados en la predicción, con un 0.83 de Accuracy Score y un AUC de 0.82. Por otro lado, su matriz de confusión presenta un alto porcentaje de verdaderos positivos, el mayor entre los métodos expuestos, pero al mismo tiempo el mayor porcentaje de falsos positivos entre los métodos expuestos.

2.5. Ejercicio 4

En términos de los modelos, los Accuracy Score de los mismos son los siguientes:

- Logit Accuracy Score: 0.763
- LDA Accuracy Score: 0.764
- KNN (con k=3) Accuracy Score: 0.833

Por lo tanto, utilizando la medida de Accuracy Score como medida de precisión, el modelo que mejor predice es KNN con K=3. Otra medida alternativa es el AUC-ROC (Area Under The Curve) el cual es de 0.71, 0.72 y 0.84 para cada modelo respectivamente, por lo que con ambas medidas llegamos la misma conclusión. Por lo tanto, en el siguiente punto haremos la predicción en base a ese modelo.

3. Ejercicio 5

En este ejercicio intentaremos predecir las personas que son pobres dentro de la base `norespondieron`. Utilizando el modelo de KNN (con k=3) entrenado y validado con la base de encontramos los siguientes resultados:

- La cantidad de pobres predicha en la muestra que no respondió es de: 1076 personas del total que es 3390 personas
- La proporción de personas predicha en la muestra de los que no respondieron y que es pobre es del: 31.74 %

4. Ejercicio 6

En este ejercicio seleccionamos las variables que pensamos relevantes e implementamos nuevamente el método logit. Las variables seleccionadas fueron:

Sexo, edad, cobertura médica, si sabe leer o escribir, ,educación y establecimiento, y variables sobre búsqueda de trabajo.

Ante partir de esta reportamos la matriz de confusión, la curva ROC y los valores de AUC y Accuracy Score, a continuación:

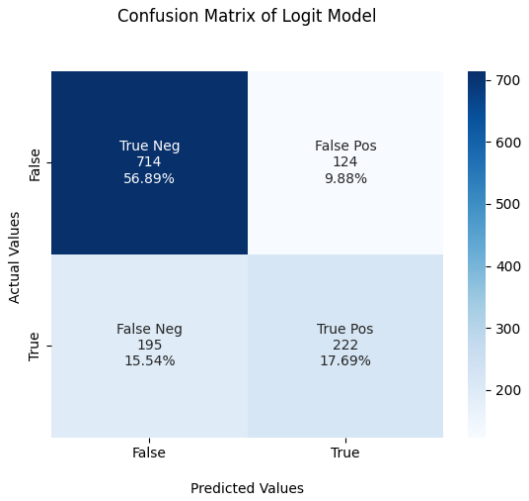

Figura 9: Matriz de Confusión para Logit con variables seleccionadas.

Como se puede observar, el mayor porcentaje lo tienen los verdaderos negativos mientras que los falsos positivos presentan el menor porcentaje entre la clasificación. A continuación exponemos la curva ROC y el valor de AUC para complementar el análisis:

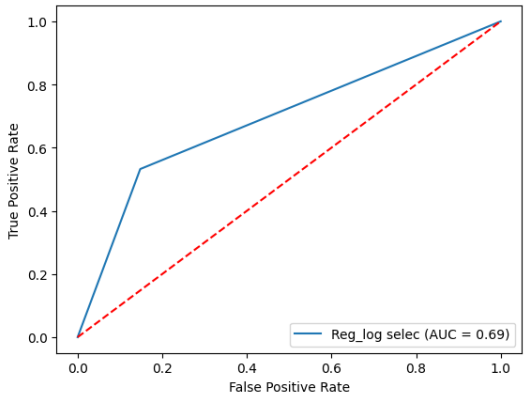

Figura 10: Matriz de Confusión para Logit con selección de variables.

Como podemos ver, el valor de AUC presenta un valor de 0.691 y la curva ROC se encuentra por encima de la curva de 45° grados dando respaldo a lo expuesto en la matriz de confusión.

Con respecto a nuestra medida de precisión (Accuracy Score) para este Logit con variables seleccionadas es de 0.746. Por lo tanto, el modelo Logit con la selección de variables tiene una su precisión menor relativo al modelo logit original. Sin embargo, los cambios no son demasiado grandes.