Universidad de San Andrés

Universidad de
SanAndrés

Big Data

Trabajo Práctico N° 4

Profesor: María Noelia Romero

Tutor: Victoria Oubiña

Alumnos: María Solana Cucher, María Victoria Rosino, María Florencia Ruiz

Parte 1

Inciso 3

Para "limpiar" nuestra base de datos prestaremos especial atención a las columnas y datos duplicados, a las observaciones atípicas, a los valores faltantes, a los *outliers* y al "tipo" de datos que contienen las variables. Para eso, de las columnas que se encuentran duplicadas, comenzaremos por conservar solamente una. El mismo procedimiento lo realizaremos con las observaciones duplicadas.

Por otro lado, para tratar los *missing values*, realizamos dos ejercicios. El primero de ellos, consiste en eliminar de la base de datos todas aquellas variables que tengan más de un 80% de valores faltantes. En segundo lugar, a aquellas variables no categóricas con más de la mitad de *missing values*, le imputamos el promedio a los valores nulos. Las variables que quedan con muchos *missing* (más de 50%) luego de estos dos procesos son directamente eliminadas. Al hacer estos pasos, la base de datos solo queda variables con menos de 1% de valores faltantes (estas observaciones son eliminadas al momento de predecir). Con respecto a los valores atípicos, eliminamos los valores negativos de las edades (CH06) y de las variables de ingresos (P21 y P47T). En esta misma línea, descartamos las observaciones que toman valores "9" o "99" en columnas de tipo categóricas. El fundamento de esta decisión es que, según el Diseño de Registro de la EPH, este valor implica un "no sabe/no responde". Por tanto, lo consideramos como un valor faltante.

En cuanto a los *outliers*, como las variables de ingresos son continuas, pueden tomar cualquier valor y, a simple vista, sería difícil determinar si algunos de estos son *outliers*. Por lo tanto, optamos por realizar un *boxplot* de las variables que miden los ingresos, tanto familiares como individuales (ITF, IPCF, P21, P47T). Esto se presenta en la Figura 1 a continuación. Notamos que en todas las variables graficadas existen valores extremos, pero, en particular en las columnas ITF y P47T, hay dos montos de ingresos demasiado altos que generan una escala exagerada en la figura. En particular, para eliminar los valores atípicos usamos la siguiente regla: aquellos que se encuentren más allá de 1.5 veces el rango intercuartílico, ya sea por debajo del primer cuartil o por encima del tercero, se consideran *outliers* y, por lo tanto, deben ser eliminados. En la Figura 2 podemos observar cómo queda el mismo *boxplot* de la Figura 1 sin estos

valores extremos. Si bien continúan existiendo *outliers* (que se representan como puntos por fuera de las "cajas") en los datos, dichas observaciones no se encuentran tan alejadas de la media del ingreso como sí sucedía en la Figura 1.

Figura 1 – Distribución de las variables de ingresos

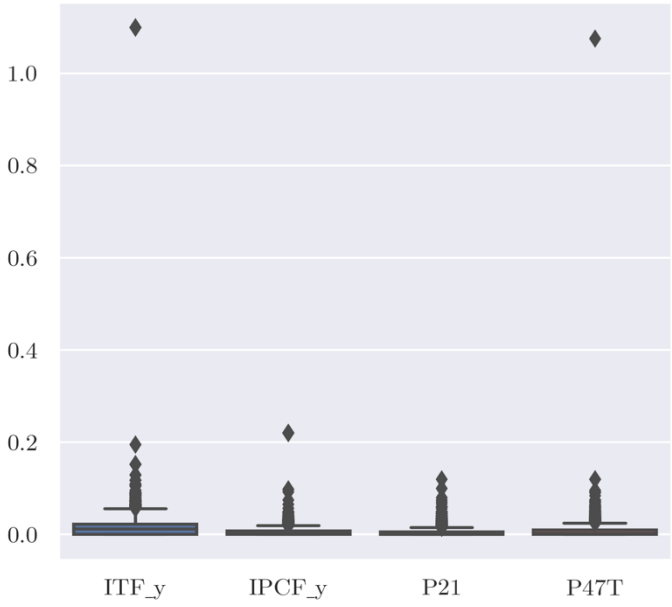
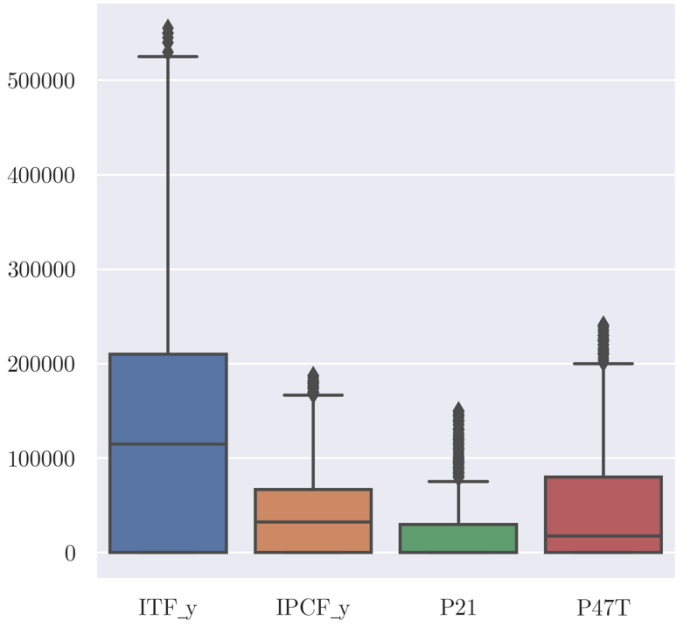


Figura 2 - Distribución de las variables de ingresos (luego de eliminar *outliers*)



En último lugar, trabajamos con el tipo de cada variable. En específico, transformamos aquellas columnas que poseen una estructura de variables categóricas

a este "tipo". También, descartamos de la base de datos dos variables (una que indica la fecha de la entrevista y otra que indica el aglomerado según su tamaño) dado que se encuentran en formato de texto y no resultan útiles para nuestra predicción de la pobreza.

Inciso 4

En esta instancia, generamos a partir de la EPH tres nuevas variables que, intuimos, ayudarán a predecir la pobreza en la muestra. Estas se denominan: vivienda marginal, hacinamiento grave y proporción de niños en el hogar. La nueva columna "vivienda marginal" es una variable binaria que toma el valor de 1 cuando clasificamos al hogar de esta manera y 0 en caso contrario. La razón para introducir esta variable radica en nuestra creencia de que las carencias físicas o de infraestructura del lugar de residencia se relacionan con la pobreza de los hogares. En particular, son los hogares pobres los que suelen caracterizarse por tener una vivienda marginal o deficitaria.

Para definir "vivienda marginal" en un principio nos basamos en el indicador de vivienda deficitaria del INDEC. Este último se construye a partir de las categorías Casa tipo B (si cumple alguna de estas condiciones: tienen piso de tierra, no tienen provisión de agua por cañería dentro de la vivienda, o no disponen de baño con descarga de agua) y vivienda precaria (incluye casillas, piezas en inquilinato, locales no construidos para habitación y viviendas móviles). Existen tres variables disponibles en nuestra base de datos que nos permiten replicar dicho indicador del INDEC: IV1 (tipo de vivienda), IV3 (material de los pisos) e IV10 (características del baño). Además de la definición de vivienda deficitaria, consideramos necesario agregar características sobre la ubicación del hogar. En particular, un entorno inseguro, insalubre o potencialmente riesgoso podría ser algo relevante a la hora de definir la marginalidad de una vivienda. Por ello, sumamos las variables IV12_1 (si la vivienda se encuentra cerca de un basural), IV12_2 (si la vivienda está ubicada en zona inundable) e IV12_3 (si la vivienda está ubicada en villa de emergencia) a la definición de vivienda deficitaria del INDEC y, de esta manera, construimos nuestra variable de "vivienda marginal".

Por otro lado, definimos a "hacinamiento" como la relación entre la cantidad total de miembros del hogar (variable IX_TOT) y la cantidad de habitaciones de uso

exclusivo del hogar (variable II1). A priori, pensamos que un valor más alto de la variable "hacinamiento" se correlacionaría con mayor pobreza. Según el INDEC, vivir en condiciones de hacinamiento genera menor privacidad, higiene y carencia de entornos aptos para el estudio y la socialización. A partir de esta variable, generamos una *dummy* denominada "hacinamiento grave" que vale 1 cuando en un hogar hay más de 1.5 personas por cuarto.

Por último, generamos la variable "proporción de niños en el hogar" la cual consideramos que puede ser un predictor importante de la pobreza porque los hogares de bajos ingresos suelen tener, en promedio, más hijos. En particular, según el estudio de la Fundación Observatorio de la Maternidad, las madres de menores recursos tienen en promedio 3.3 hijos en contraposición a 1.6 hijos que tienen las de mayores ingresos. Podemos explicar la tendencia de los hogares pobres a tener más descendencia por múltiples motivos. Por ejemplo, los hijos pueden verse como un seguro de vida más "económico" durante la vejez. Por otro lado, la mayor cantidad de niños en el hogar puede explicarse por falta de conocimientos o acceso al cuidado sexual, así como también por la falta de planeamiento familiar. A la vez, la proporción de niños en un hogar también puede retroalimentar la situación de pobreza: más personas en un hogar disminuye los ingresos per cápita de este porque el salario de los adultos se tiene que distribuir en más individuos. A partir de este sustento teórico, creamos la variable "proporción de niños en el hogar", que se construye como el total de menores de 10 años que residen en el hogar (IX_MEN10) sobre el total de individuos en él (IX_TOT).

Inciso 5

La Figura 3 muestra la correlación entre un subconjunto de variables de nuestra base de datos que consideramos relevantes para predecir la pobreza. Estas son: sexo, estado civil, *dummy* de cobertura médica, nivel educativo, ocupación, categoría de inactividad, ingreso per cápita familiar (IPCF), cantidad de miembros del hogar, *dummy* de vivienda marginal, *dummy* de hacinamiento y proporción de niños en el hogar.

Dentro de las correlaciones positivas, la que más se destaca es aquella entre la variable de categoría de inactividad y la de ocupación, que poseen una correlación de 0.81. Esto quiere decir que, por ejemplo, los individuos menores de 6 años o

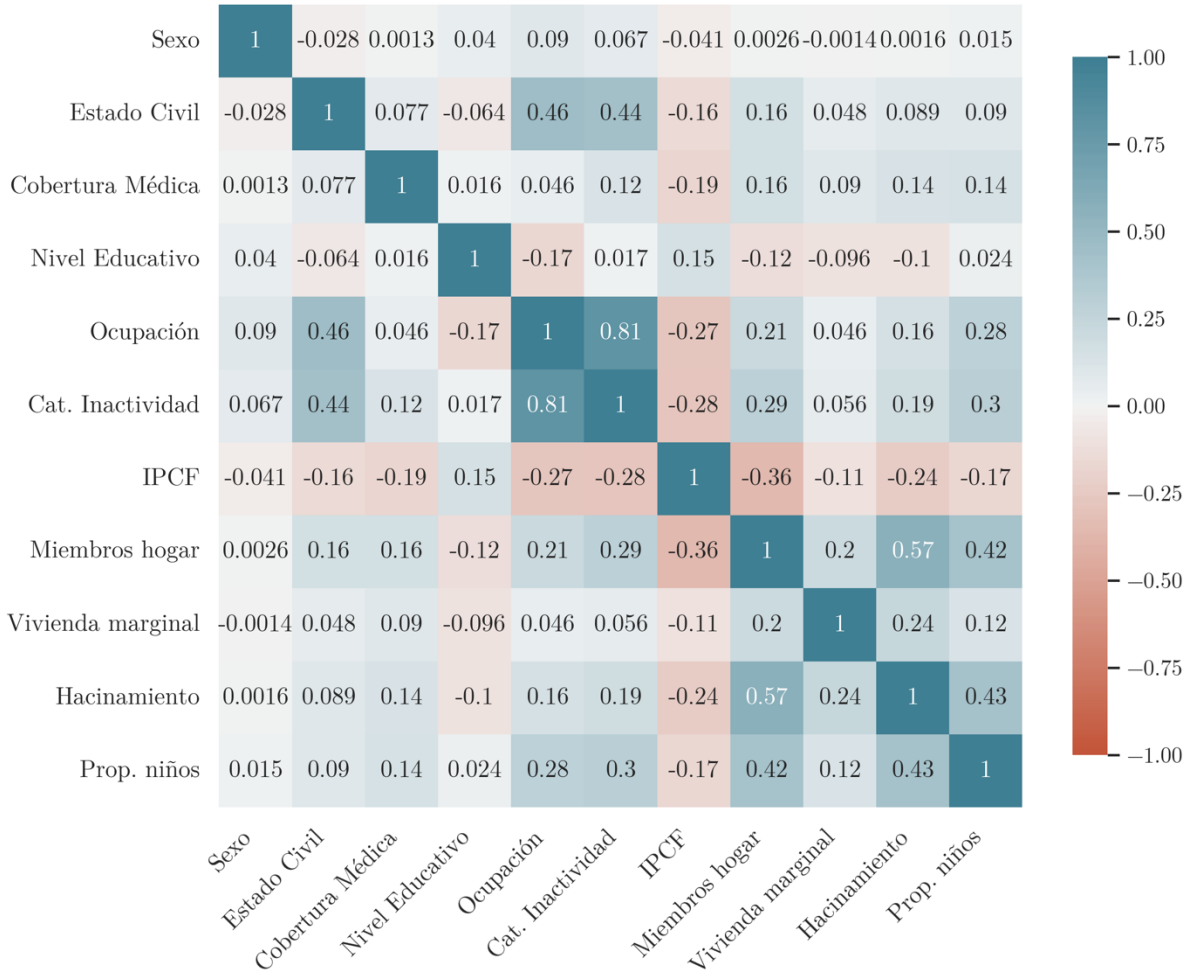
discapacitados suelen estar desocupados, inactivos o, tautológicamente, menores de 10 años. En segundo lugar, se destaca la correlación positiva de 0.57 entre la variable de hacinamiento, creada por nosotros, y la cantidad de miembros del hogar. Así como también entre el hacinamiento y la proporción de niños viviendo en el hogar, otra variable creada por nosotros. Es decir, mientras mayor sea la cantidad de personas viviendo en un mismo hogar y, dentro de estas personas, más alta sea la cantidad de niños, más habitaciones compartirán. La cantidad de miembros del hogar también correlaciona positivamente (coeficiente de 0.42) con la proporción de niños viviendo en él. En este sentido, no es sorprendente que mientras más personas habiten en una misma vivienda, una mayor proporción de ellas sean niños menores de 10 años.

También se destaca las altas correlaciones entre las variables de ocupación y categoría de inactividad con la variable que indica el estado civil. En el primer caso, el coeficiente resulta de 0.46; mientras que en el segundo es 0.44. Intuitivamente, la correlación entre estas variables refleja que, por ejemplo, aquellas personas inactivas o desocupadas (categorías más altas de la variable de ocupación), así como aquellas menores de 6 años o discapacitadas (categorías más altas de inactividad), suelen estar separadas, divorciadas, viudas o solteras (categorías más altas del estado civil).

Luego, pasamos a analizar las cuatro correlaciones negativas más relevantes. En primer lugar, el coeficiente entre el ingreso per cápita familiar (IPCF) y la cantidad de miembros del hogar es de -0.36. Es decir, a mayor cantidad de personas en una misma menor es su ingreso per cápita, lo cual tiene sentido dado que, a mayor cantidad de miembros, el ingreso de quienes trabajen se reparte entre más individuos.

También se observa una correlación negativa relativamente alta entre el IPCF y la categoría de inactividad (coeficiente de -0.28), así como entre el el IPCF y la ocupación (coeficiente de -0.27). Esto indica que las personas con menor ingreso per cápita dentro del hogar son aquellas menores de 6 años o discapacitadas (categorías más altas de inactividad) y aquellas personas inactivas o desocupadas (categorías más altas de la variable de ocupación). Por último, el IPC también se relaciona negativamente con la variable de hacinamiento (coeficiente de -0.24), lo cual indica que mientras más personas por habitación haya en un hogar, estas tendrán un menor ingreso per cápita.

Figura 3 - Matriz de correlaciones



Inciso 7

Según el Cuadro 4.3 (Pobreza en hogares y personas - Regiones estadísticas y 31 aglomerados urbanos) del informe del INDEC "Incidencia de la pobreza y la indigencia en 31 aglomerados urbanos", el porcentaje de hogares pobres en el Gran Buenos Aires en el primer semestre de 2023 fue 30.3%. Si calculamos la incidencia de la pobreza en hogares en base a la EPH (junto con la variable de ponderación para expandir la muestra al total de la población), encontramos que el porcentaje de hogares debajo de la línea de pobreza es 31.09%. Es decir, solo 0.79 puntos porcentuales mayor de lo que expone el INDEC.

Parte 3

Inciso 2

Evaluamos los siguientes métodos: Regresión Logística, Análisis de discriminante lineal, Vecinos cercanos, Árbol de decisión, Bagging, Random Forest y Boosting. En algunos casos, optimizamos el hiperparámetro que consideramos más relevante para la predicción. En particular, al hacer Regresión Logística elegimos por validación cruzada el hiperparámetro de regularización, ya sea de LASSO o de Ridge; para esto generamos una serie de 10 valores entre 10^{-5} y 10^{5}, entre los cuales será seleccionado el hiperparámetro de regularización. Al implementar Árbol de decisión, optimizamos la profundidad del árbol, seleccionándola de una lista de valores que va desde 1 hasta el total de variables de la base de datos (99). En Vecinos cercanos, elegimos la cantidad óptima de vecinos, seleccionando por validación cruzada un número entero entre 1 y 10. Finalmente, para realizar Random Forest, elegimos de forma óptima el tamaño del *subset* de características que se tendrán en cuenta en cada nodo del árbol. En este caso, los parámetros posibles van desde 1 hasta el total de variables de la base de datos, al igual que en el Árbol de decisión.

Inciso 3

La Tabla 1 nos permite evaluar las métricas de predicción de los distintos modelos. En primer lugar, buscamos un área debajo de la curva ROC lo más grande posible, en tanto esto nos indica que estamos más cerca del ideal de la curva ROC (esquina izquierda superior, sin FP o FN). El método de Árbol de decisión obtiene un AUC de 0.827496, seguido por el AUC de 0.777314 de Bagging. Luego, la métrica de *accuracy* nos indica la proporción de clasificaciones correctas, ya sean verdaderos positivos o verdaderos negativos, por lo que esperamos un valor alto en el modelo que mejor predice: Árbol de decisión obtiene el mayor valor igual a 0.839111, seguido por Bagging con una *accuracy* de 0.8080.

La tasa de falsos negativos es la probabilidad de que no se detecte un verdadero positivo, por lo que bajo una buena predicción buscaríamos el menor valor posible. En este caso, CART tiene un valor de 0.221445, mientras que el resto de los modelos supera el 0.3 en esta métrica. En relación con la tasa de verdaderos negativos (o especificidad), que es la probabilidad de que un resultado negativo resulte efectivamente negativo, esperamos un valor lo más alto posible, y Regresión Logística obtiene el máximo valor (1.0).

Tabla 1 – Evaluación de modelos

Modelo	Hiperparámetro	AUC	Accuracy	False N. Rate	True N. Rate	False P. Rate	True P. Rate	ECM	Precisión
Regresión logística	10 (Ridge)	0.500000	0.618667	1.000000	1.000000	0.000000	0.000000	0.381333	0.000000
Análisis de discriminate lineal	-	0.771090	0.792000	0.317016	0.859195	0.140805	0.682984	0.208000	0.749361
KNN	1	0.734452	0.746667	0.317016	0.785920	0.214080	0.682984	0.253333	0.662896
Árbol de decisión	21	0.827496	0.839111	0.221445	0.876437	0.123563	0.778555	0.160889	0.795238
Bagging	-	0.777314	0.808000	0.351981	0.906609	0.093391	0.648019	0.192000	0.810496
Random Forest	39	0.770210	0.796444	0.340326	0.880747	0.119253	0.659674	0.203556	0.773224
Boosting	-	0.767146	0.789333	0.326340	0.860632	0.139368	0.673660	0.210667	0.748705

Por otro lado, la tasa de falsos positivos nos indica la probabilidad de que se dé un resultado positivo cuando el valor verdadero es negativo, por lo que buscamos el menor valor entre los métodos evaluados y Regresión Logística, con un hiperparámetro λ igual a 10, bajo el método de regularización de Ridge, obtiene un valor nulo de esta probabilidad. La tasa de verdaderos positivos o sensibilidad es la probabilidad de que un positivo real de positivo, por lo que buscaríamos el mayor valor posible. La mejor *performance* en esta medida la tiene Árbol de decisión, con una tasa igual a 0.778555.

En cuarto lugar, el ECM, entendido como la diferencia cuadrática media entre los valores estimados y los valores reales de los datos, debería ser menor a medida que las predicciones sean mejores: el método de CART tiene el menor ECM entre los métodos, igual a 0.160889, y es seguido por el ECM de 0.1920 de Bagging. Por último, el método de Bagging es el elegido si consideramos la métrica de precisión. La precisión (o*positive predicted value*) es la proporción de positivos predichos que son positivos reales; es decir, a diferencia de *accuracy*, solo consideramos haber predicho bien los valores positivos. Bagging obtiene el mayor valor de precisión entre los métodos, igual

a 0.810416, y se encuentra seguido por el método de Árbol de decisión, con una precisión de 0.795238.

De esta manera, si consideramos todas las métricas descritas, podemos concluir que el método de Árbol de decisión (CART) es el que mejor predice ya que tiene el mejor desempeño en la mayoría de las métricas (*accuracy*, tasa de falsos negativos, tasa de verdaderos positivos, ECM y AUC). En particular, tenemos en cuenta un árbol con una profundidad de 21, según la optimización llevada a cabo sobre este parámetro. Si bien el método de Random Forest no resulta el elegido por ninguna de las métricas obtenidas, es importante mencionar que, de emplearse este método, el número óptimo de características (variables) que se deben considerar al dividir un nodo durante la construcción del árbol es de 39. De manera similar, el hiperparámetro óptimo para el método de vecinos cercanos (la cantidad de vecinos) resultó ser 1 y el hiperparámetro óptimo de penalización para Regresión logística es de 10 (Ridge).

Inciso 4

En el trabajo práctico 3 el modelo que mejor predecía la pobreza, en función de las métricas consideradas, resultó ser Análisis de Discriminante Lineal. Entonces, para saber si nuestras predicciones mejoraron, podemos comparar las métricas obtenidas en el trabajo anterior para Análisis de Discriminante Lineal con las obtenidas en este trabajo para el modelo seleccionado (Árbol de decisión). Al hacer esto, encontramos que la mayoría de los indicadores mejoraron en esta nueva predicción (ver en Tabla 2). En particular, el área debajo de la curva ROC (AUC) aumentó de 0.780625 en la primera estimación a 0.827496 en esta nueva especificación. En cuanto a la *accuracy,* también podemos señalar una mejoría: su valor subió de 0.803556 a 083911. Un cambio positivo se puede destacar también en el valor de la tasa de falsos negativos, la cual bajó de 0.367946 a 0.221445. Otra medida que mejoró fue la tasa de verdaderos negativos, que pasó de valer 0.794721 a 0.876437. También se puede notar un gran cambio en la tasa de verdaderos positivos (pasó de 0.67 a 0.78 aproximadamente) y, en menor medida, pero no menos relevante, en el ECM (pasó de 0.20 a 0.16 aproximadamente). La única métrica que empeoró fue la tasa de falsos positivos. Esta medida era 0.111437 cuando estimamos mediante Análisis discriminante lineal en el trabajo anterior y ahora, toma un valor de 0.123563.

De esta manera, como cinco de las seis métricas comparadas entre los trabajos han mejorado sustancialmente en esta nueva estimación, podríamos concluir que nuestras predicciones actuales tendrán un mejor resultado que las anteriores.

Tabla 2 – Comparación métricas TP3 y TP4

Modelo	Hiperparámetro	AUC	Accuracy	False N. Rate	True N. Rate	False P. Rate	True P. Rate	ECM	Precisión
Árbol de decisión (TP4)	74	**0.827496**	**0.839111**	**0.221445**	**0.876437**	0.123563	**0.778555**	**0.160889**	0.795238
Análisis de discriminante lineal (TP3)	-	0.780625	0.803556	0.367946	0.794721	**0.111437**	0.672686	0.196444	-

Inciso 5

Según lo descripto en el inciso 2 y 3, el método elegido para predecir la pobreza es Árbol de decisión. En particular, utilizamos un modelo que toma una profundidad del árbol igual a 21. La elección de este hiperparámetro se debe a que 21 es la profundidad del árbol óptima según los resultados de *cross validation*. Ahora bien, usando Árbol de decisión, la tasa de pobreza predicha dentro de la muestra "no respondieron" es del 46.58% aproximadamente.