# Reporte final

164889 - Estadística aplicada 2

Diciembre 2020

## 1 Introducción

Actualmente las computadoras son parte esencial en la vida de las personas, sus componentes son cada vez más avanzados y el rendimiento de las mismas depende de múltiples factores por lo que muchas veces es difícil saber cual máquina presenta un rendimiento superior al promedio.

En el siguiente reporte se presentan 3 modelos de regresión lineal que explican el Rendimiento relativo de la CPU a partir de 8 variables disponibles que se describen en el siguiente apartado (en la mayoría de los casos no se usaron las 8 variables disponibles). También se muestra y valida el modelo que consideramos más apropiado y pronosticamos el rendimiento relativo de 20 CPU´s como muestra de la eficacia del modelo seleccionado. Toda esta información está basada en el artículo "Un modelo de predicción del desempeño relativo de CPUs".

## 2 Descripción de los datos

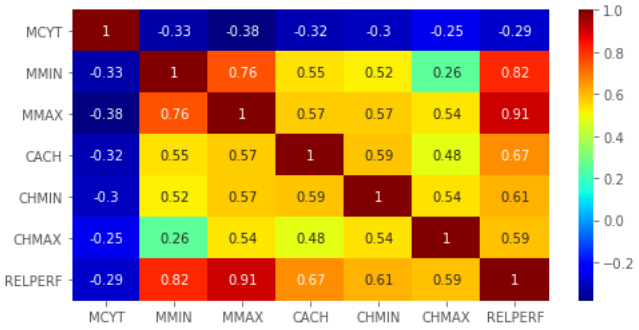
Para la creación de los modelos que se presentan en las siguientes hojas se utilizó una muestra de 189 modelos de CPU. Cada una de estas unidades de procesamiento central cuentan con la información de las variables MCYT, MMIN,MMAX,CACH, CHMIN, CHMAX, RELPERF, VENDMOD. Estas variables se describen a continuación:

- RELPERF: desempeño relativo, es una variable cuantitativa continua además de ser la variable respuesta. Las variables restantes se consideran regresores.
- MCYT: tiempo de ciclo, es una variable cuantitativa continua y se mide en nanosegundos.

- MMIN: mínimo memoria principal, es una variable cuantitativa continua y se mide en kilobytes.
- MMAX: máximo memoria principal, es una variable cuantitativa continua y se mide en kilobytes.
- CACH: tamaño de caché, es una variable cuantitativa continua y su unidad son 10 kilobytes.
- CHMIN: mínimo número de canales E/S, es una variable cuantitativa discreta y su unidad son canales.
- CHMAX: máximo número de canales E/S, es una variable cuantitativa discreta y su unidad son canales.
- VENDMOD: marca y modelo de CPU, es una variable cualitativa.

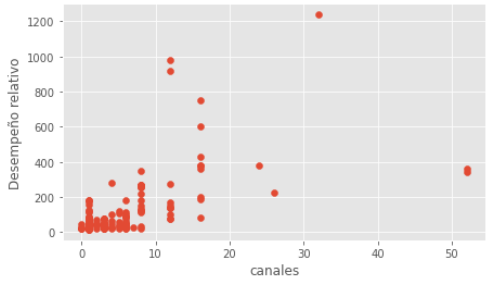
La variable *VENDMOD* fue descartada al presentar 189 categorías diferentes, no aporta a nuestra investigación. Ahora presentamos la matriz de correlación, recordemos que la correlación es un valor entre -1 y 1 donde mientras más cercano se esté de los extremos mayor será la relación entre ambas variables.

Figura 1: Matríz de correlación

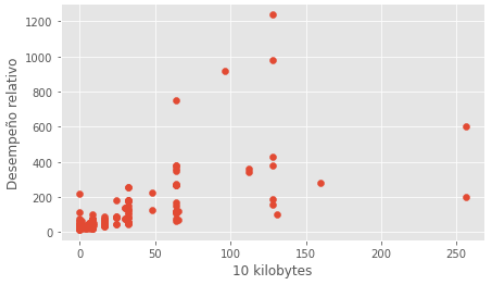



Como podemos observar, las variables que presentan una correlación positiva más fuerte frente a *RELPERF* son *MMIN, MMAX* así como *CACH*. La única variable que parece presentar una correlación débil es *tiempo de ciclo*. Si además graficamos los diagramas de dispersión (Figura 2) de estas variables podemos notar que la mayoría de los datos se encuentran agrupados enlos valores de rendimiento relativo más pequeños, esto nos podría

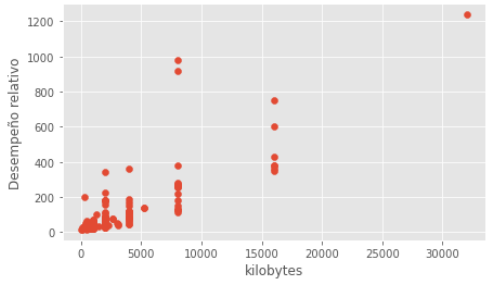
indicar que pocos CPU´s presentan un rendimiento relativo alto. Por otro lado también confirmamos lo que nos indicaba la matriz de correlación, es decir, existe una relación positiva entre el rendimiento relativo y las variables mínimo y máximo de memoria principal así como caché. Notemos que a partir de estos diagramas de dispersión no podemos encontrar relaciónes lineales, sólo podemos saber que existen tendencias positivas o negativas como es el caso de *MCYT* y *RELPERF*



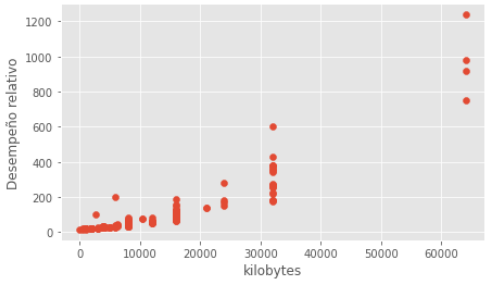

(a) Mínimo de canales vs rendimiento relativo



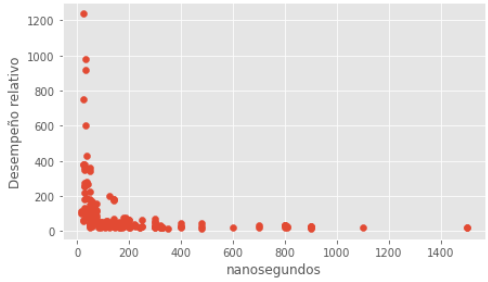

(b) Caché vs rendimiento relativo



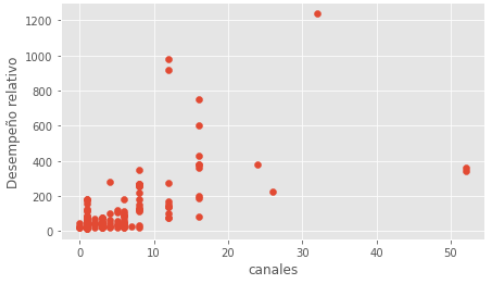

(c) Mínimo de memoria vs rendimiento relativo




(d) máximo de memoria vs rendimiento relativo



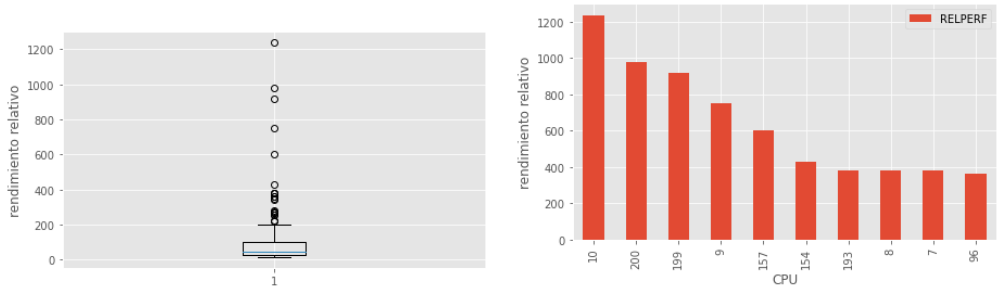

(e) Tiempo de ciclo vs rendimiento relativo




(f) mínimo de canales vs rendimiento relativo

Figura 2: Diagramas de dispersión

Si ahora analizamos la variable RELPERF, podemos notar que presenta una gran agrupación de sus datos para valores menores a 200 puntos de rendimiento relativo mientras que existen CPU´s particulares que presentan un rendimiento relativo muy superior al promedio(Figura 3).




(a) Gráfico de caja

(b) Caché vs rendimiento relativo

Figura 3

Notemos que son 5 las unidades de procesamiento central que presentan rendimientos muy alejados de la mayoría, estas CPU´s son la número secuencial 10, 200, 199, 9 y 157 y podrían generar problemas al ajustar los modelos. En particular, podemos saber que el rendimiento superior que presenta la número 10 se debe a que este CPU presenta el nivel de memoria mínima más grande de toda la muestra de 189 CPU´s.

# 3 Modelo

Para encontrar el mejor modelo que explicara la variable respuesta se inició ajustando un modelo de regresión lineal con todos los regresores disponibles, se analizó la tabla de la regresión y se descartaron aquellos regresores que tuvieran un valor p alto. En este primer paso se eliminó el regresor CHMIN al presentar un valor p de 0.616. Una vez hecho esto, el modelo resultante presenta un valor

$$R^2 = 0.914 \tag{1}$$

En un siguiente proceso verificamos si existen posibles valores atípicos en el gráfico de residuales estandarizados vs apalancamiento y encontramos que las CPU´s con el número secuencial 200,199 y 10 presentan una Distancia de Cook mayor a 1. Al retirar estos valores encontramos que el modelo anterior tiene un coeficiente de determinación mayor, es decir, el ajuste mejoró al

retirar estos 3 valores atípicos. Entonces, el primer modelo propuesto es:

$$\widehat{RELPERF} = -31.94 + .0054(MMAX) + 0.739(CACH) + 0.2465(CHMAX)$$
$$+0.1(MMIN) + 0.03(MCYT)$$
$$R^2 = 0.943, PRESS = 153975, ee(intercepto) = 4, ee(MMAX) = 0,$$
$$ee(CACH) = .06 ee(CHMAX) = 0.1, ee(MMIN) = .01, ee(MCYT) = .03$$

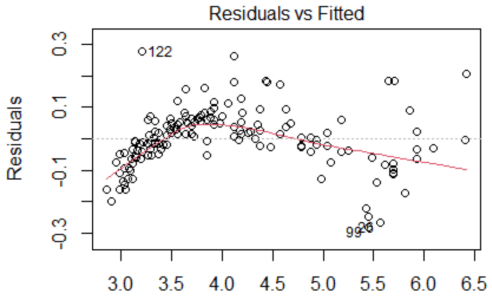
Lo siguiente que hacemos para encontrar un mejor modelo es verificar si se presentan interacciones, probamos una a una las posibles interacciones y encontramos que el modelo más completo con el mayor número de regresores e interacciones es el siguiente

$$\widehat{RELPERF} = 12.81 + .0023(MMAX) + .368(CACH) +$$
$$(2.7e-5)(MMAX : CHMAX) + (2.45e-5)(MMAX : CACH)$$
$$+3.7e-7(MMAX : MMIN), R^2 = 0.995, PRESS = 14696, ee(intercepto) = .1,$$
$$ee(MMAX) = .002, ee(CACH) = .03 ee(interacciones) = 0$$

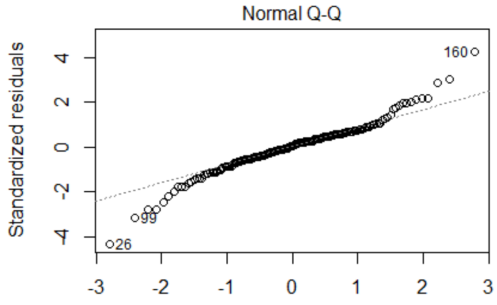
Este segundo modelo propuesto ocupa gran cantidad de regresores, busquemos un modelo más parsimonioso y con un PRESS más bajo. El modelo final que proponemos para explicar el rendimiento relativo de las CPU es:

$$\widehat{LN(RELPERF)} = 2.86 + (7e-5)(MMAX) + .01(CACH) + .0001(MMIN)$$
$$(-2.35e-7)(MMAX : CACH) - (2.9e-9)(MMAX : MMIN)$$
$$R^2 = .9893, PRESS = 2.47, ee(intercepto) = .1, ee(regresores) = 0$$

Con la transformación Ln(y) a la respuesta se intenta corregir problemas de heteroscedasticidad y falta de normalidad en los errores. Las gráficas de residuales y cuantil - cuantil normal se presentan a continuación:




(a) Residuales vs respuesta ajustada




(b) Cuantil-Cantil normal

# 4 Pronósticos

En la siguiente sección pronosticaremos la respuesta media y construiremos los correspondientes intervalos al 95% de confianza para un nivel de regresores

$$x_q = (x_{q1}; x_{qn}; ...; x_{qp})$$

donde q corresponde al primer cuartil, la mediana y el tercer cuartil de cada uno de los regresores y p es el número de regresores en el modelo final. En nuestro caso obtenemos lo siguiente:

$$x_1 = (4000, 0, 1); x_2 = (8000, 8, 2); x_3 = (16000, 32, 6)$$
$$x_q = (MMAX, CACH, MMIN)$$

Los correspondientes intervalos de confianza al 95% son:

$$Parax_1 : (25.28, 26.18), RespuestaMedia = 25.8$$
$$Parax_2 : (40.85, 42.09), RespuestaMedia = 41.26$$
$$Parax_3 : (88.82, 91.23), RespuestaMedia = 90.01$$

Lo siguiente que realizamos fue la predicción de los 20 datos de prueba a partir del modelo elegido. En la siguiente tabla se muestran los valores exactos, los valores predichos y los errores. El error relativo más bajo fue de .003 mientras que el más alto fue de 0.22. Por otro lado, para conocer si nuestra predicción fue acertada calculamos dos estadísticos, recordemos que el estadístico PRESS es una medida de qué tan bueno es el modelo de regresión para predecir datos, buscamos un valor PRESS pequeño. Por otro lado también existe el R2 de la predicción que se basa en el PRESS y este valor da una idea de la capacidad predictiva del modelo. Estos estadísticos están definidos como sigue:

$$\text{PRESS} = \sum_{i=1}^{n} \left[ y_i - \hat{y}_{(i)} \right]^2$$
$$= \sum_{i=1}^{n} \left( \frac{e_i}{1 - h_{ii}} \right)^2 \qquad (c)$$

$$R^2_{\text{prediction}} = 1 - \frac{\text{PRESS}}{SS_{\text{T}}} \qquad (d)$$

Al calcular el valor de estos estadísticos para nuestro modelo final obtenemos PRESS=2.47304 Y R2predicción=0.9827 por lo que esperamos que

este modelo explique 98% de la variabilidad en predecir nuevas observaciones, la forma en que predice este modelo parece satisfactoria. En general este modelo fue capaz de pronosticar de forma certera los datos verdaderos, el error relativo sólo en una ocasión superó el .2 y en ningún momento parece que se vea afectado por observaciones extremas.

Figura 4: Tabla de pronósticos

| | RELPERF | PREDICCIÓN | ERROR ABSOLUTO | ERROR RELATIVO |
|---|---|---|---|---|
| **0** | 253 | 274.327211 | -247.385678 | 0.084297 |
| **1** | 290 | 297.384287 | -284.304975 | 0.025463 |
| **2** | 124 | 153.306131 | -118.967563 | 0.236340 |
| **3** | 102 | 85.082132 | -97.556383 | 0.165861 |
| **4** | 47 | 41.664393 | -43.270353 | 0.113524 |
| **5** | 25 | 26.362348 | -21.728063 | 0.054494 |
| **6** | 32 | 31.622609 | -28.546128 | 0.011793 |
| **7** | 82 | 80.302740 | -77.614196 | 0.020698 |
| **8** | 18 | 20.717804 | -14.969007 | 0.150989 |
| **9** | 31 | 31.622609 | -27.546128 | 0.020084 |
| **10** | 80 | 74.171735 | -75.693617 | 0.072853 |
| **11** | 46 | 44.483599 | -42.204879 | 0.032965 |
| **12** | 95 | 97.612409 | -90.418995 | 0.027499 |
| **13** | 119 | 119.402759 | -114.217498 | 0.003385 |
| **14** | 19 | 21.132270 | -15.949199 | 0.112225 |
| **15** | 41 | 39.023858 | -37.335827 | 0.048199 |
| **16** | 80 | 61.611349 | -75.879154 | 0.229858 |
| **17** | 142 | 160.753664 | -136.920127 | 0.132068 |
| **18** | 25 | 25.458341 | -21.762957 | 0.018334 |
| **19** | 24 | 24.326335 | -20.808440 | 0.013597 |

# 5 Conclusiones

Aunque los datos de CPU con los que trabajamos no sean recientes nos permiten darnos una idea de la gran disparidad que existe en el mundo de la computación. En este caso, la memoria con la que cuenta una CPU es determinante para poder explicar el rendimiento que mostrará. Algunas

variables que parecían influír a simple vista el desempeño relativo terminaron siendo rechazadas por los diferentes modelos que planteamos y sólo unas pocas, en especial las relacionadas con memoria, fueron las más influyentes para los diferentes modelos posibles. Esto nos ayuda a entender el camino que siguienton las unidades de procesamiento central en su desarrollo a lo largo de los años pues el requerimiento de cada vez mayor capacidad de memoria impulsó la creación de sistemas más avanzados, más veloces y con un rendimiento superior. Por el lado de la creación de estos modelos propuestos podemos comentar que no fue una proceso de elección sencillo, el ajuste de la mayoría de los modelos disponibles es superior a R2=.9 y aunque desde el inicio hay indicios de variables que no contribuyen para explicar la variable respuesta, todos estos modelos presentan problemas de falta de normalidad donde eliminar ciertos valores atípicos ayuda parcialmente pero no dan solución al problema. También intentamos realizar las transformaciones usadas para estos casos y no aportan, encontramos que la transformación que más ayuda es aplicar Ln a la respuesta. Podemos concluir que este fue el grupo de datos que presentó un mayor reto para su manipulación de todos los trabajados a lo largo del curso pero al final pudimos encontrar un modelo que hiciera pronósticos muy certeros y eficaces a la hora de ingresar cualquier tipo de datos sin importar los valores de los regresores.