Análisis de Modelo de Regresión para estimación de frecuencia en CPU y GPU

Stalin Pillajo[1], Erick Angulo[2], Michael Mendoza[3], and Jesús Yaranga[4]

[1] Universidad Politécnica Salesiana, Carrera Ciencias de la Computación, Quito, Ecuador
spillajom@est.ups.edu.ec
[2] Universidad Politécnica Salesiana, Carrera Ciencias de la Computación, Quito, Ecuador
eangulom@est.ups.edu.ec
[3] Universidad Politécnica Salesiana, Carrera Ciencias de la Computación, Quito, Ecuador
mmendozar1@est.ups.edu.ec
[4] Universidad Politécnica Salesiana, Carrera Ciencias de la Computación, Quito, Ecuador
jyarangac@est.ups.edu.ec

Abstract
This report explains how a regression model was developed to determine which variables are capable of explaining a certain phenomenon, therefore, it is desired to identify which variables contribute substantially to the creation of a statistical model based on data. For this analysis, a dataset from the public Kaggle repository was used, which has 4845 instances. This gives us information about the performance of various CPUs and GPUs based on five variables that were process size, TDP, die size, transistors, and freq. Previously, to obtain the statistical analysis, a data cleaning was carried out since there were some empty and incomplete data. In addition, to observe the relationship of these, graphical representations, correlation analysis, significance tests, goodness of fit of the model, identification of outliers, influential values, the respective verifications of the assumptions of the model were made and with all this it was possible to obtain the conclusions relevant to the project. More advanced graphs were made to determine which model best suits our case study and to choose the best regression model we combined the independent variables in different ways until we obtained the best result analytically.

Keywords: k-Nearest Neighbor, Knowledge Data Discovery, Data curation, Imputation, Hot-Deck, TDP, MNAR, MICE, Histogram

1 Introducción

La Estadística computacional es una disciplina del conocimiento científico tecnológico, que se ocupa de estudiar el impacto de la computación en la metodología estadística [3]. La Estadística busca como acumular y optimizar la información extraída de los datos, como recolectar los datos para maximizar la información y como hacer inferencias de los mismos para extender nuestro conocimiento. Por otro lado, la Ciencia de Computación, nos dice cómo calcular y procesar de manera óptima los datos, cómo

medir el costo asociado al procesamiento de la información, cómo la información y el conocimiento pueden ser útilmente representados y como comprender los límites de lo que puede ser calculado etc. [4]. Cada modelo de un problema del mundo real contiene un conjunto de suposiciones que se agregan para simplificar el modelo, lo que a menudo implica descartar o ignorar algunos efectos o parámetros del problema que se estudia. Es importante comprender estas limitaciones del modelo para evitar el mal uso. Por lo general, un modelo matemático se define mediante un sistema de ecuaciones lineales o no lineales, ecuaciones integrales o diferenciales u otros tipos de ecuaciones. [2]. A veces es posible estudiar estas ecuaciones Matemáticamente, ni siquiera a resolución completa. Sin embargo, a menudo es difícil o imposible encontrar exactamente estas soluciones y deben resolverse con la ayuda de una computadora.

2 Metodología

2.1 Limpieza de Datos (Data curation)

La limpieza y/o curación de datos es un punto clave dentro del proceso KDD (Knowledge Data Discovery) ya que nos permite organizar y mantener los conjuntos de datos accesibles y utilizables para el uso correcto de los mismos. Esto con el fin de obtener información basado en un análisis estadístico con el cual podamos ya sea conocer la tendencia de los datos o realizar tareas de regresión basado en una suposición de correlación entre los datos. Teniendo esto en mente podemos decir que sin una limpieza correcta de los datos un análisis estadístico seria difícil por no decir imposible.

Tomando en cuenta lo anteriormente mencionado, nuestro conjunto de datos como mucho otros poseerá una serie de valores vacíos los cuales nos impiden realizar una correcta implementación de un modelo de regresión lineal, por el simple hecho de que dichos datos vacíos podrían poseer información de importancia de la realidad que buscamos determinar la cual es la tendencia del incremento de la frecuencia (velocidad) basado en la cantidad de transistores y el tamaño de placa del procesador o la cantidad de temperatura que puede soportar basado en la capacidad de procesamiento de la matriz del mismo. Pero para conocer con certeza esta realidad que buscamos interpretar debemos también con considerar ciertos criterios para la detección y tratamiento de datos.

2.1.1 Métodos de Detección

Un método de detección posee diferentes criterios para determinar la naturaleza de un valor vacío o perdido en un conjunto de datos, estos criterios son tres: **Valores Perdidos Completamente Aleatorios *(MCAR)***; En donde se determina que la razón de la perdida de los datos es ajena a los mismos, dando a entender que sin importar la categoría a la que pertenezca un dato, este puede presentar un valor vacío por una determinada razón. **Valores Perdidos no Aleatorios *(MNAR)***; Donde la falta de los aleatorios posee un razón de no estar ya sea basado en la naturaleza de los datos o el conjunto, es decir que hay una razón de que un dato no deba estar en dicha sección o categoría. **Valores Perdidos Aleatorios *(MAR)***; Aquí la razón de falta de los datos es ajena a los mismos datos, pero puede tener una razón de fondo, es decir que la perdida de los datos se deba a que pertenezca a una categoría especifica.

2.1.2 Métodos de tratamiento

Obtenida una idea de la manera de conocer la naturaleza de la perdida de datos debemos poseer mecanismos de acción para solventar dicha perdida, para esto tenemos dos caminos a seguir: el **descarte** y la **imputación**.

El descarte de datos vacíos consiste en la eliminación parcial o total de los datos ya sea por *Eliminación de filas*; Elimina la fila a la que pertenece el dato, o *Eliminación de dato*; Elimina a la posición especifica que pertenece el dato mas no a la fila a la que pertenece. En ambos criterios se debe tener sumo cuidado ya que la eliminación de datos podría alterar la distribución original de los datos algo que nos daría un diseño de modelo totalmente diferente al de los datos originales, pero para poder saber cuando aplicar o escoger este camino usamos un **porcentaje de 10 en los datos vacíos** ya que esto no influiría gravemente en la distribución original de los datos.

La imputación de los datos consiste en agregar los datos vacíos con valores que se aproximen al valor original esperado pero con métodos que no afecten la distribución original entre estos métodos tenemos **Mediana, Media, Modelos de Regresión, Hot-Deck, MICE**. Aunque los métodos mas usados por su naturaleza de evitar alteraciones en la distribución de los datos son el *Modelos de Regresion, Hot-Deck: kNN y MICE*. Debemos tener en cuenta que para elegir uno de estos métodos en los modelos de regresión y MICE debemos asegurar que exista un correlación entre los datos, osea debe a existir una relación entre la variable que vamos a predecir y la que va ayudar a realizar dicha predicción, ya que si no existe la distribución de los datos puede llegar a cambiar. En contra parte el método de Hot-Deck: k Nearest Neighbour (knn) nos ayudara a predecir o estimar valores cuando estos no posean una correlación entre ellos.

2.1.3 Implementación de algoritmo de imputación Hot-Deck: K-Nearest-Neighbor

Particularmente en nuestro set de datos podemos establecer que siguen una naturaleza de datos vacíos completamente aleatorios, sabiendo esto debemos elegir el camino para el tratamiento de los datos, analizando cada una de las variables y verificando la relación entre cada una de ellas se elegí una método de descarte para la variable de Tamaño del Procesador debido a que posee menos del 10% (9 datos perdidos) se procede a la eliminación de las filas por el hecho de que no afectaría a la distribución original de los datos. En contra parte el resto debido al hecho que las demás variables poseen mas del 10 % (mas de 600 datos perdidos) y se conoce que no existe una relación entre cada una de las variables para aplicar un modelo de regresión se propone el uso de un método de imputación Hot-Deck: k Nearest Neighbor por el hecho de que la distribución de los datos se mantendría sin cambio significantes muy fiel al original.

K-Nearest-Heighbor

$$x_i = \frac{\sqrt{(x_2-x_1)^2*(y_2-y_1)^2}}{2}$$

Gower Distance

$$D_{Gower}(x_1, x_2) = 1 - (\frac{1}{p}\sum_{p}^{j=1} s_j(x_1, x_2) = 1 - \frac{|y_{1j}-y_{2j}|}{R_j})$$

Algoritmo de k nearest neighbor es un algoritmo para estimar y sustituir valores perdidos[1] y en conjunto con el criterio de *distancia de gower* podemos aplicar una estimación con mayor con fiabilidad entre los datos que se van a sacar la estimación, ya que la distancia de gower nos permite conocer la similitud entre los datos elegidos y con dicha certeza de similitud el calculo para un valor perdido va a ser mucho mas cercano al que se esperaría fuera real cuando en el conjunto de datos no existe una relación entre las variables con la cual aplicar un modelo de regresión lineal.

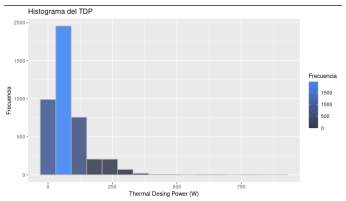

Figure 1: Histograma sin Imputación

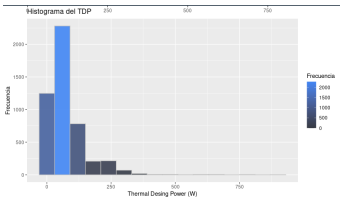

Figure 2: Histograma con Imputación

2.2 Modelo y variables

El modelo que se usó en este proyecto fue el de regresión lineal múltiple que nos permite generar un modelo lineal en el que el valor de la variable dependiente o respuesta (Y) se determina a partir de un conjunto de variables independientes (X1, X2, X3...). donde la ecuación es la siguiente:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + u \tag{1}$$

2.2.1 Regresion lineal multiple estimada (muestral)

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p \tag{2}$$

Donde b0, b1, b2 son estadisticas muestrales

2.2.2 Relación entre las variables dependientes e independientes

En este caso debido a que la influencia de la cantidad de los transistores aumenta la frecuencia o velocidad de procesamiento vemos menester usar la cantidad de transistores para predecir la tendencia de la frecuencia que tendría si la cantidad de transistores aumentara.Así mismo podemos relacionar las otras variables, en el caso de la temperatura es bien conocido que cuando esta aumenta,la potencia o frecuencia de una GPU O CPU disminuyen por ende el rendimiento es afectado, sucede lo contraria cuando las temperaturas están controladas permitiendo que funcione correctamente el dispositivo. El tamaño de la matriz esta ampliamente relacionada con la variable de la frecuencia puesto que mientras más grande sea el tamaño de esta permitirá que los datos se transmitan con mayor velocidad y frecuencia.Por ultimo, el tamaño del procesador tiene una relación directamente proporcional con la frecuencia lo que nos permite poder realizar los respectivos análisis estadísticos.

Variable Dependiente.

- Frecuencia: Velocidad máxima que puede llegar a alcanzar el CPU y GPU en base a las siguientes variables independientes.

Variable Independiente.

- Temperatura: Cantidad de calor que puede llegar a soportar el procesador
- Tamaño del procesador: Capacidad del procesador a la hora de realizar varias tareas dentro de la CPU y GPU.
- Tamaño de la matriz: Procesamiento de datos, de manera que mientras mas grande sea con mayor velocidad o frecuencia se transmitirán los datos.
- Transistores: Elemento que permite el paso de potencia o energía al procesador de manera que este es el encargado de identificar si el procesador es semi-conductor.

3 Experimentos

En esta sección incluimos algunos experimentos para evaluar el desempeño de los métodos propuestos. Llevamos a cabo dos escenarios diferentes de experimentos. Primero, probamos la validez de los métodos sobre pequeños conjuntos de datos. En segundo lugar, analizamos el rendimiento en un escenario de agrupación de documentos, es decir, empleamos conjuntos de datos reales.

3.1 Análisis de Modelos

Para realizar este análisis usaremos como variable dependiente (y) a la variable frecuencia (freq) medida en (MHz) y como variables dependientes usaremos:

- Tamaño del procesador (prcsz, medido en nanómetros (nm)
- Temperatura (tdp), medido en vatios o watts (w)
- Tamaño de la matriz (diesz), medido en milímetros cuadrados
- Transistores (transs), medido en millones

3.1.1 Análisis de Residuales con dos variables independientes

Tabla de comparaciones

Regresión Lineal Múltiple	Coeficiente de Determinación	Prueba de Significación
Frecuencia en función del Procesador y la Temperatura	0.08080846	Valor P: 0.01 P-Value-t: -19.80648 | 3.287078 P-Value-f: 2.780541e-89
Frecuencia en función del Procesador y la Matriz	0.08014625	P-Value: 0.01 P-Value-t: -20.48161 | -2.70423 P-Value-f: 1.58853e-88
Frecuencia en función del Procesador y los Transistores	0.08836463	P-Value: 0.01 P-Value-t: -21.65161 | -7.142096 P-Value-f: 5.876554e-98
Frecuencia en función de la Temperatura y de la Matriz	0.006885451	P-Value: 0.01 P-Value-t: 5.61459 | -1.685556 P-Value-f: 5.47548e-08
Frecuencia en función de la Temperatura y los transistores	0.006910981	P-Value: 0.01 P-Value-t: 5.777506 | -1.722091 P-Value-f: 5.145222e-08
Frecuencia en función de la Matriz y los transistores	0.0004489156	P-Value: 0.01 P-Value-t: 1.368729 | -0.3922921 P-Value-f: 0.3373549

3.1.2 Análisis de Residuales con tres variables independientes

Tabla de comparaciones				
Regresión Lineal Múltiple	Coeficiente de Determinación	Prueba de Significación		
Frecuencia en función del Procesador, Temperatura y Matriz	0.08571058	Valor P: 0.01		
		P-Value-t:		
		-20.42241	5.417986	-5.084282
		P-Value-f: 1.191033e-93		
Frecuencia en función del Procesador, temperatura y transistores	0.09606638	P-Value: 0.01		
		P-Value-t:		
		-21.84343	6.412536	-9.031509
		P-Value-f: 1.58853e-88		
Frecuencia en función del Procesador y los Transistores	0.007091193	P-Value: 0.01		
		P-Value-t:		
		5.68896	-0.9333459	-0.9977754
		P-Value-f: 1.605171e-07		
Frecuencia en función de los transistores, procesador y matriz	0.08897774	P-Value: 0.01		
		P-Value-t:		
		-6.841714	-21.68191	1.777304
		P-Value-f: 2.105703e-97		

3.1.3 Análisis de Residuales con cuatro variables independientes

Tabla de comparaciones				
Regresión Lineal Múltiple	Coeficiente de Determinación	Prueba de Significación		
Frecuencia en función del Procesador, Temperatura, Matriz y Transistores	0.09617224	Valor P: 0.01		
		P-Value-t:		
		-21.83351	6.204849	-0.7526587
		-7.482233		
		P-Value-f: 1.46237e-104		

3.1.4 Coeficiente de Correlación

Tabla de Coeficiente de Correlacion					
	y	x1	x2	x3	x4
y	1.000000000	-0.2805777	0.07937182	0.02041964	0.007873595
x1	-0.280577722	1.000000000	-0.12268732	-0.20286000	-0.354643686
x2	0.079371817	-0.1226873	1.000000000	0.51750656	0.385857985
x3	0.020419636	-0.2028600	0.51750656	1.000000000	0.605305158
x4	0.007873595	-0.3546437	0.38585799	0.60530516	1.000000000

3.2 Análisis de residuales gráficamente

A continuación se mostrara cuales fueron las regresiones que mas trataron de ajustar al modelo en función de dos, tres y cuatro variables independientes.

- Dos variables independientes: Frecuencia en función de la temperatura y el tamaño de la matriz
- Tres variables independientes: Frecuencia en función de la temperatura, tamaño de la matriz y los transistores

- Cuatro variables independientes: Frecuencia en función de todas las variables independientes

3.2.1 Dos Variables Independientes

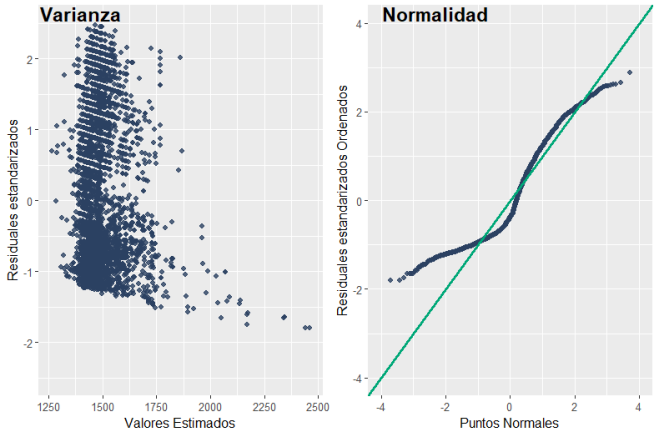

Figure 3: Frecuencia - TDP + T.Matriz

3.2.2 Tres Variables Independientes

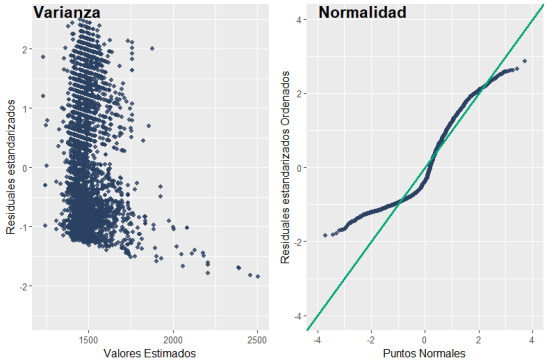

Figure 4: Frecuencia - TDP + T.Matriz + Transistores

3.2.3 Cuatro Variables Independientes

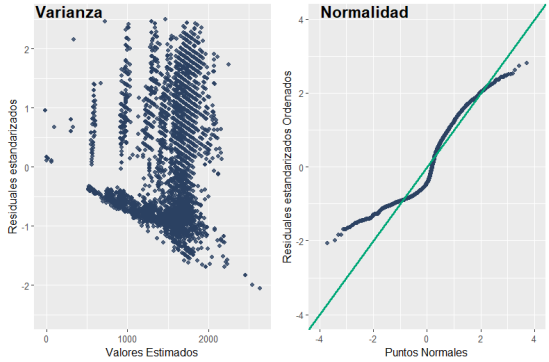

Figure 5: Frecuencia - T.Procesador + TDP + T.Matriz + Transistores

4 Resultado de la Experimentación (Modelo a escoger)

El resultado del análisis residual para la predicción de la Frecuencia en función de las demás variables (temperatura, Tamaño de Matriz, Procesador y transistores) a través de las tablas de comparación, dio como resultado que no hay un modelo que se ajuste correctamente a los datos. De manera que el modelo que mas se trata de ajustar es la regresión de la predicción de la Frecuencia en función de la temperatura, Tamaño de Matriz, Procesador y transistores. Este modelo es el que mas trata de ajustarse ya que como podemos ver en la tabla de comparaciones, su bondad de ajuste llega a 0.09617224, donde es la que mas se acerca al umbral de (0.7 o -0.7) y tiene las siguientes propiedades:

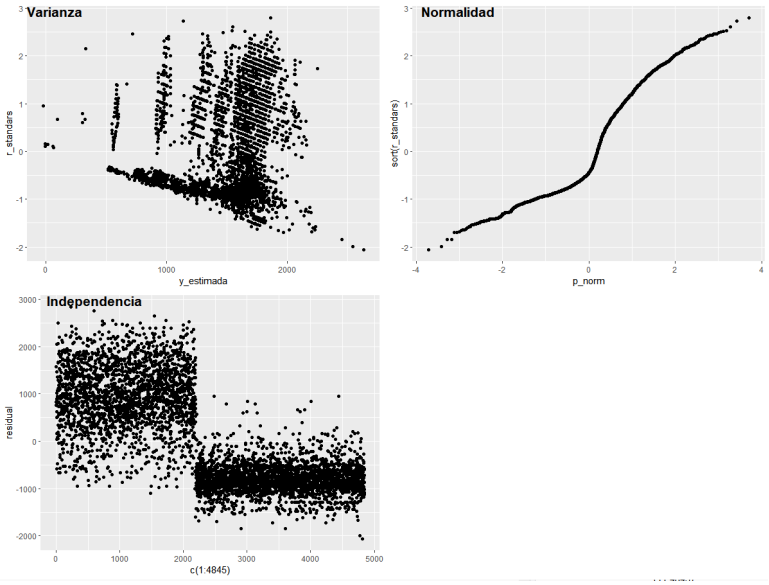

Figure 6: Frecuencia: Todas las variables

No cumple con la propiedad de normalidad ya que los datos no están cerca de la linea de los 45 grados, No cumple con la propiedad de homocedasticidad ya que los datos no están cerca del cero, y sus datos son independientes.

4.1 Prueba de Significancia

Prueba de Significancia	
P - value: 0.01	
P - value - F (Global): 1.46237e-104	
Variable	**P - value - T**
Frecuencia en función del procesador	-21.83351
Frecuencia en función de la temperatura	6.204849
Frecuencia en función del tamaño de la matriz	-0.7526587
Frecuencia en función de los transistores	-7.482233

Se determina que los datos no son útiles para realizar una regresión lineal simple ya que ninguno supera el p - value de 0.01 para ciencias exactas.

4.2 Observaciones atípicas e influyentes

Datos Atípicos e influyentes		
Frecuencia en función de la Temperatura, Tamaño de la matriz Procesador, y Transistores	Datos Atípicos	Datos influyentes Treshold: 0.002683179
	120	220

5 Conclusiones

- Se ha realizado la presentación del análisis profundo de datos utilizando el modelo de regresión lineal múltiple, permitiendo llegar a la conclusión de que el modelo no es el más adecuado ya que la bondad de ajuste siempre otorga un resultado no favorable al umbral. Entonces posiblemente en este caso en particular se podría utilizar otro tipo de modelo estadístico.
- Se ha realizado el análisis de residuales de forma gráfica, empleando dos variables independientes, llegando a la conclusión de que la frecuencia en función de la temperatura y el tamaño de la matriz otorgan alrededor de 174 valores atípicos, sin embargo otorgando al mismo tiempo un total de 355 valores influyentes.
- Haciendo uso de cuatro variables, de modo que se analice la frecuencia en función de la temperatura, tamaño de la matriz, procesador y transistores, se determina la existencia de 120 datos atípicos, sin embargo es posible analizar alrededor de 220 datos influyentes, además de la existencia de un coeficiente de determinación de 0.1194541.
- Tras a ver realizado la experimentación, a través de la formula de regresión lineal múltiple, s determino que no existe un modelo que se ajuste de manera correcta a los ya que no existía una buena bondad de ajuste entre el modelo y los datos y su correlación no superaba el umbral.

References

[1] Gustavo EAPA Batista, Maria Carolina Monard, et al. A study of k-nearest neighbour as an imputation method. *His*, 87(251-260):48, 2002.

[2] Pablo Negrón Marrero. *Fundamentos del Análisis Computacional*. Universidad de Puerto Rico, 2021.

[3] Toby Segaran. *Programming collective intelligence: building smart web 2.0 applications*. O'Reilly Media, Inc., 2007.

[4] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.