

Uso de correo electrónico para analizar la comunicación bilateral aplicando Big Data y Regresión Lineal Simple

Use of Email to analyze bilateral communication with Big Data and Simple Linear Regression

HERNÁNDEZ-CRUZ, Luz María†*, MEX-ALVAREZ, Diana Concepción, ESTRADA-SEGOVIA, Guadalupe Manuel y CASTILLO-TELLEZ, Margarita

Universidad Autónoma de Campeche

ID 1^{er} Autor: *Luz María, Hernández-Cruz* / ORC ID: 0000-0002-0469-5298, Researcher ID Thomson: H-3153-2018, CVU CONACYT ID: 662220

ID 1^{er} Coautor: *Diana Concepción, Mex-Alvarez* / ORC ID: 0000-0001-9419-7868, Researcher ID Thomson: I-4164-2018, CVU CONACYT ID: 842039

ID 2^{do} Coautor: *Guadalupe Manuel, Estrada-Segovia* / ORC ID: 0000-0002-5700-258X, Researcher ID Thomson: G-3542-2019, CVU CONACYT ID: 95199

ID 3^{er} Coautor: *Margarita, Castillo-Tellez* / ORC ID: 0000-0001-9639-1736, Researcher ID Thomson: S-2283-2018, CVU CONACYT ID: 210428

DOI: 10.35429/JITC.2019.10.3.21.28

Recibido 06 de Junio, 2019; Aceptado 20 de Diciembre, 2019

Resumen

En la actualidad el correo electrónico es el servicio de red más usado como medio de comunicación para envío/recepción de mensajes y archivos. El objetivo de este estudio es realizar un análisis de correos electrónicos institucionales aplicando una estrategia que asegure la existencia de una comunicación bilateral entre el personal. La investigación es de tipo aplicada, la cual permitirá predecir grupos de trabajo asertivos con relaciones laborales prosperas y productivas. El estudio integra la aplicación de una herramienta Tecnológica de Big Data llamada Immersion y el análisis de un modelo de Regresión Lineal Simple (PLS) utilizando Microsoft Office Excel. La metodología adaptada se compone de tres fases: primeramente, la "Fase de campo" donde se recoge un gran volumen de datos (personal data) desde una cuenta de correo electrónico institucional para el caso de estudio, enseguida tenemos la "Fase de Análisis" donde se construye un modelo de regresión lineal simple para analizar la relación entre los datos recogidos y finalmente, la "Fase de Interpretación" donde se explican los resultados obtenidos. Teniendo aplicaciones importantes como pueden ser la integración de cuerpos académicos, redes temáticas, comités disciplinarios y/o miembros de colaboración en proyectos de investigación atendiendo al contexto de estudio.

Big data, Correo electrónico, Comunicación

Abstract

Currently, the email is the most used network service as a means of communication for sending and receiving messages and files. The objective of this study is to perform an analysis of institutional emails by applying a strategic that ensures the existence of a bilateral communication between the employees. The research is of applied type, which will allow to predict assertive working groups with prosperous and productive labor relations. The study integrates the application of a Technological Big Data tool called Immersion and the analysis of a Simple Linear Regression (PLS) model using Microsoft Office Excel. The adapted methodology is composed of three phases: first, the "Data Collection" where a large volume of data is collected (personal data) from an institutional email account for the case study, then we have the "Analysis" where a simple linear regression model is constructed to analyze the relationship between the collected data and finally, the "Interpretation" where the obtained results are explained. Having important applications such as the integration of academic group, thematic networks, disciplinary committees or collaborative members in projects.

Big data, Email, Communication

Citación: HERNÁNDEZ-CRUZ, Luz María, MEX-ALVAREZ, Diana Concepción, ESTRADA-SEGOVIA, Guadalupe Manuel y CASTILLO-TELLEZ, Margarita. Uso de correo electrónico para analizar la comunicación bilateral aplicando Big Data y Regresión Lineal Simple. Revista de Tecnologías de la Información y Comunicaciones. 2019. 3-10: 21-28

* Correspondencia del Autor (lmhernan@uacam.mx)

† Investigador contribuyendo como primer autor.

Introducción

La comunicación en el ámbito laboral es un elemento fundamental en cualquier tipo de empresa independientemente de las características que presente. La comunicación es imprescindible para poder mantener las relaciones tanto de forma interna como de forma externa. (Sánchez, 2014)

La comunicación adquiere un rol fundamental como eje de transmisión de la organización para su funcionamiento. Lo que la empresa comunica tanto hacia el exterior como de forma interior, la imagen que trasmite es de vital importancia para diferenciarse de la competencia. (Sánchez, 2014)

En la presente investigación se aborda la problemática de evaluar la comunicación bilateral entre colaboradores para optimizar el trabajo. Hoy día, en diversos contextos es necesario poder discernir grupos de trabajo y colaboración para aprovechar de manera óptima y eficiente las relaciones interpersonales.

La Hipótesis central está enfocada en validar la existencia de una relación recíproca o bilateral entre colaboradores. El estudio se realiza utilizando como dato de entrada la carga de correo electrónico institucional o corporativo extrayendo la información con una herramienta de Big Data y evaluando las conexiones existentes mediante un Modelo de Regresión Lineal.

Antecedentes

Al Imaginar la carga de correo electrónico institucional o corporativo pensamos en un gran cúmulo de datos en los cuales sería muy difícil encontrar alguna relación o conexión de datos a partir de un mecanismo manual. Según Castilla el modo usual de recolección de datos implica el uso de un equipo para cada variable de interés, lo cual dificulta y encarece la integración y el procesamiento conjunto.

Según Martínez y Lara (2014), la sociedad de la información, donde el volumen de datos crece de forma exponencial, la eclosión del Big data ha impactado en ámbitos diversos. La popularidad del término ha desdibujado las fronteras de un concepto que no sólo incide en la dimensión sino también en el valor de los datos recopilados y procesados.

En el sector de las Tecnologías de la Información y la Comunicación el término Big Data es una referencia a grandes conjuntos de datos. Existen una gran variedad de las técnicas de Big Data. Siempre que sea necesario extraer el conocimiento inmerso en grandes volúmenes de datos podemos emplear aplicaciones de Big Data. Además, existen herramientas de Big Data propias para la recogida, transformación, puesta a disposición de los datos y, evidentemente, para el análisis de datos.

“Las Tecnologías de Big Data se clasifican aquellas que dan soporte a la captura, transformación, procesamiento y análisis de los datos, ya sean estructurados, semi-estructurados o no estructurados.” (Hernández, et al.,2017)

El uso de Big Data ha ayudado en el área de investigación a estudiar cosas que podrían llevar años en descubrirse por sí mismos sin el uso de estas herramientas, debido a la velocidad del análisis, es posible que el analista da datos pueda cambiar sus ideas basándose en el resultado obtenido. En este sentido, Immersion nos permite recoger metadatos de Gmail para una cuenta de correo electrónico y visualizar conexiones de dicha cuenta con otras cuentas.

Big Data es un método de *machine learning* para crear modelos predictivos, tanto para clasificación como regresión, a partir de datos numéricos o cualitativos. Con la regresión lineal, modelamos la relación entre la variable dependiente, y , y una variable explicativa o variable independiente, x . (Anand, 2018)

En esta investigación se usa una herramienta de Big Data para la extracción de datos que sirva de entrada a un modelo de regresión lineal con el objetivo de evaluar la comunicación bilateral entre colaboradores. La Figura 1 muestra un esquema del caso de estudio.

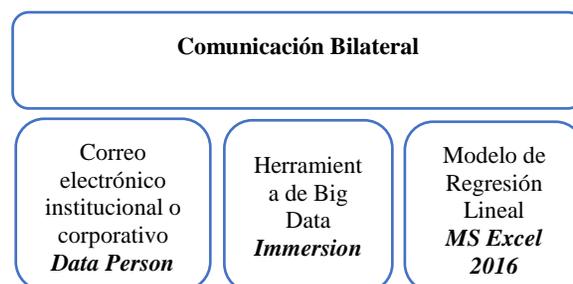


Figura 1 Planteamiento

Fuente: Fuente Propia

El caso de estudio toma para el análisis una cuenta de correo institucional de Gmail, y a partir de ella, se extrae y analiza la información obtenida de los colaboradores principales (top) para evaluar, mediante un modelo de regresión lineal, si existe una comunicación bilateral entre ellos, contribuyendo de esta forma a ofrecer un mecanismo capaz de réplica para diferentes aplicaciones de interés que acredite un entorno colaborativo.

Metodología

La investigación, se lleva a cabo en tres fases: recogida de datos, análisis de datos e interpretación de los resultados.

Recogida de datos

Para iniciar, es importante reunir los datos significantes que permitan realizar el estudio. Optamos, en esta investigación, por utilizar una herramienta de Big Data denominada Immersion. Esta genera un diagrama de red web que permite mostrar las conexiones de relación de la cuenta de correo en uso con otras cuentas. La Figura 2 muestra la interfaz presentada por Immersion.

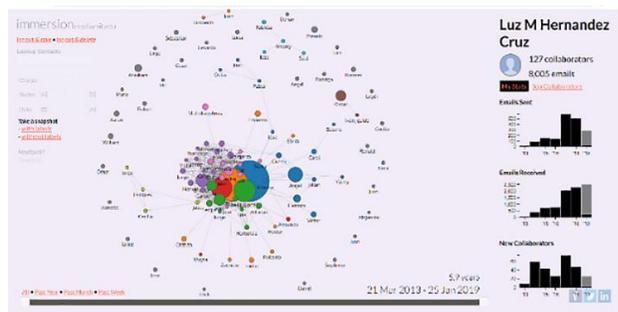


Figura 2 Interfaz Immersion
Fuente: Fuente Propia

La herramienta recoge todos los correos electrónicos (incluyendo los enviados y recibidos) a partir de la fecha en la cual fue creado. Es decir, la cantidad de datos recogidos varía principalmente en función a la carga de envío/recepción de correos electrónicos y el tiempo transcurrido a partir de la creación de la cuenta, esto de igual forma considerando el espacio máximo asignado por el administrador. Immersion por su parte ofrece una serie de datos significativos como el total de correos enviados, recibidos y nuevos colaboradores. Por otro lado, nuestro interés radica en extraer los principales colaboradores encontrados en el año 2018 para la cuenta en estudio.

La Figura 3 muestra la lista obtenida del Immersion de los 15 colaboradores con mayor conexión con la cuenta de estudio.



Figura 3 Top de Colaboradores, Immersion
Fuente: Elaboración Propia

Específicamente por cada colaborador se generan los detalles de la colaboración a partir de la relación con la cuenta de estudio. En la Figura 4 se muestra el detalle de conexión de la cuenta de estudio con su principal colaborador.

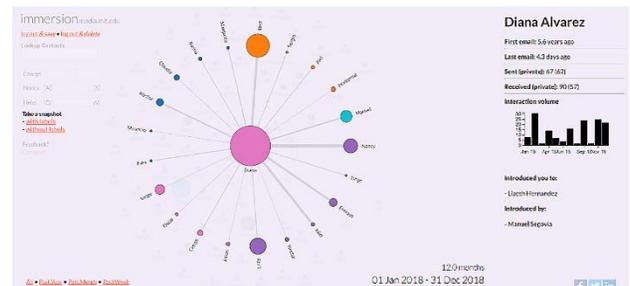


Figura 4 Detalles de la conexión con Colaboradores, Immersion
Fuente: Fuente Propia

En el caso de estudio se tiene como objetivo analizar la correlación entre personal docente con la finalidad de discernir líneas de generación de conocimiento o cuerpos académicos (CA) y el trabajo colaborativo. Con este enfoque se analizan los cinco principales colaboradores identificados por Immersion. Otro factor que pudiera afectar el análisis radica en los correos masivos, que no sugieren una relación directa con el propietario de la cuenta de estudio, por lo tanto, se ha decidido considerar como datos relevantes únicamente los correos enviados y recibidos de forma *privada*.

De esta forma se obtiene los datos significantes para el análisis de la correlación bilateral. La Tabla 1 muestra los datos de estudio.

Colaborador	Enviados	Recibidos
1	62	57
2	31	31
3	29	28
4	21	14
5	18	19

Tabla 1 Datos de estudio

Fuente: Fuente Propia

Análisis de datos

Inicialmente, se crea un gráfico de dispersión con los datos de estudio mostrados en la Tabla 1 para analizar si son admisibles para un estudio de regresión lineal simple. La Gráfica 1 muestra el gráfico de dispersión obtenido.

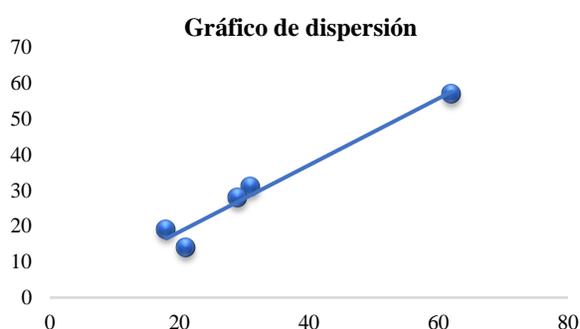


Gráfico 1 Gráfico de Dispersión de Datos de Estudio

Fuente: Fuente Propia

En el gráfico de dispersión se observa que, si existe una tendencia lineal, por lo tanto, los valores de estudio si son admisibles para un estudio de regresión lineal simple. Por medio de un modelo de regresión lineal simple se puede explicar la relación que existe entre la variable respuesta Y y una única variable explicativa X.

El modelo de regresión lineal simple tiene la siguiente expresión:

$$Y = \alpha + \beta X + \varepsilon \quad (1)$$

Donde:

Y: Variable dependiente,

α : Intersección, término constante,

β : Parámetro respectivo a cada variable independiente,

X: Variable independiente,

ε : Perturbación, error aleatorio.

En este momento, se decide utilizar una herramienta informática capaz de procesar los datos de estudio y aplicar el análisis de regresión lineal simple. Microsoft Office Excel 2016 permite añadir un complemento de análisis de datos idóneo para cumplir con este propósito. La herramienta de análisis de Regresión de MS Excel efectúa el análisis de Regresión Lineal utilizando el método de “Mínimos cuadrados” para ajustar una línea a un conjunto de observaciones. En nuestro caso de estudio, con ello, se analiza la forma en que la variable independiente X afecta a la variable dependiente Y en un estudio de regresión lineal simple.

En palabras más exactas, se analiza de qué modo incide los correos electrónicos enviados (variable independiente o explicativa X) con respecto a los correos electrónicos recibidos (variable dependiente o de respuesta Y) correspondiente a los colaboradores top de la cuenta de correo electrónico institucional de interés para evaluar una relación bilateral. El análisis se realiza proporcionando los datos de estudio a la herramienta de análisis de datos de MS Excel, siendo procesados por funciones de macros estadísticas para realizar los cálculos de forma automática y finalmente, mostrar los resultados en una tabla de resumen.

La Figura 5 muestra las opciones configuradas en la herramienta de análisis de datos de MS Excel para generar el Modelo de Regresión del caso de estudio.

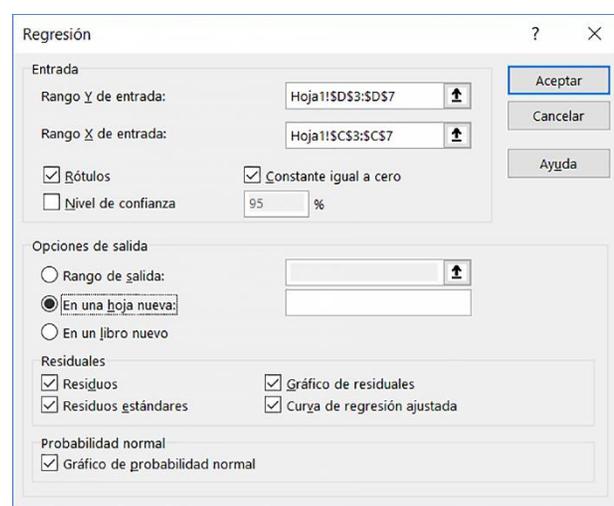


Figura 5 Regresión en MS Excel 2016 para el caso de estudio

Fuente: Fuente Propia

Interpretación de Resultados

MS Excel ofrece los resultados del análisis de regresión lineal simple de nuestro caso de estudio en forma automática. A continuación, se interpretan los valores obtenidos del modelo: El primer cuadro de resultados proporciona los coeficientes de ajuste del modelo. La Tabla 2 muestra las estadísticas de la regresión lineal simple aplicada al caso de estudio.

Estadísticas de la regresión	
Coefficiente de correlación múltiple	0.99622452
Coefficiente de determinación R ²	0.9924633
R ² ajustado	0.7424633
Error típico	3.23404619
Observaciones	5

Tabla 2 Resultados. Estadísticas de Regresión Lineal Simple

Fuente: Fuente Propia

Resume el ajuste del modelo propuesto para los datos observados. Un dato importante de mencionar es el coeficiente de correlación que permite medir la relación lineal existente entre las variables X y Y. El valor del estudio para el coeficiente de correlación es de 0.98101697 lo que indica que la asociación lineal entre las dos variables es fuerte, asumiendo que existe una tendencia lineal entre las dos variables del estudio.

El R² (coeficiente de determinación) proporciona un % de variabilidad de la variable a modelizar (Y), explicado por la variable explicativa (X). Mientras más cerca está de 1 este coeficiente, mejor es el modelo. Se observa que, el coeficiente de determinación R² resultado de caso de estudio es de 96%, lo cual indica que existe una relación evidente entre la variable independiente X y la variable dependiente Y. También obtener una tabla de análisis de varianza. La Tabla 3 muestra la tabla de Análisis de Varianza para nuestro Caso de estudio.

Análisis de Varianza					
	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F
Regresión	1	1069.028	1069.02758	76.7751	0.0031
Residuos	3	41.77242	13.9241387		
Total	4	1110.8			

Tabla 3 Resultados. Tabla de Análisis de Varianza

Fuente: Fuente Propia

En el análisis de varianza tenemos las fuentes de variación. Donde la Regresión tiene un Valor crítico de F igual a 0.0031307, por lo cual, se rechaza la Hipótesis nula debido a ser un valor menor a 0.05 establecido como valor de significación. Dado lo anterior, se puede concluir que se rechaza la Hipótesis nula del análisis de varianza, asumiendo que por lo menos uno de los coeficientes de regresión es diferente de 1.

Otro dato importante para considerar dentro del análisis de varianza es el intervalo de confianza, dado por las columnas “Inferior 95%” y “Superior 95%” que permite determinar si un coeficiente debe ser o no incluido en el modelo de regresión ajustado definitivo; si el valor 0 se encuentra en el intervalo de confianza para su estimación entonces el coeficiente no deberá ser incluido en el modelo. La Tabla 4 muestra la tabla de resultados generados para los coeficientes del modelo de regresión lineal simple del caso de estudio. Los valores obtenidos para el intervalo de confianza del coeficiente interceptación son *Inferior* de -12.39872397 y *Superior* de 11.88225837, que incluye en su rango al valor 0, en consecuencia, se hace notar que dicho coeficiente no debería estar considerado en el modelo de regresión definitivo. Para la variable X, el intervalo es de 0.59443969 a 1.272531293, siendo aceptada en el modelo regresivo definitivo.

	Coefficientes	Error típico	Estadístico t	Probabilidad
Intercepción	-0.258232801	3.814829854	-0.067691827	0.950289987
Variable X1	0.933485491	0.106536221	8.762141971	0.003130701
	Inferior 95%	Superior 95%	Inferior 95.0%	Superior 95.0%
Intercepción	-12.39872397	11.88225837	-12.39872397	11.88225837
Variable X1	0.594439689	1.272531293	0.594439689	1.272531293

Tabla 4 Resultados. Coeficientes de Regresión Lineal

Fuente: Fuente Propia

Por otro lado, la herramienta de análisis de datos de MS Excel permite crear un Gráfico de probabilidad normal, por medio del cual se puede determinar si la variable dependiente Y sigue una distribución normal o aproximadamente normal. La Gráfica 2 muestra el Gráfico de probabilidad normal.

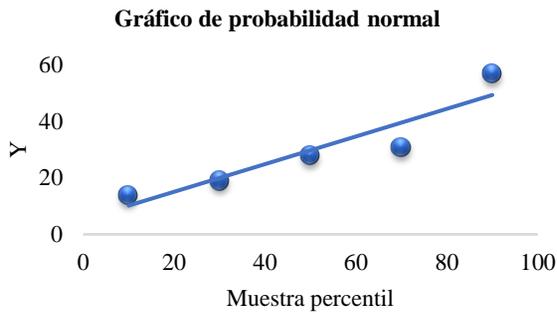


Gráfico 2 Gráfico de Probabilidad Normal
Fuente: Fuente Propia

Visualmente los puntos del Gráfico de probabilidad normal siguen una tendencia aproximadamente normal con lo que se puede inferir que si existe una distribución normal para el modelo de regresión lineal analizado. De igual forma, es generada la Gráfica de curva de regresión ajustada que evalúa hasta qué punto el modelo se ajusta a los datos y si el modelo cumple sus objetivos.

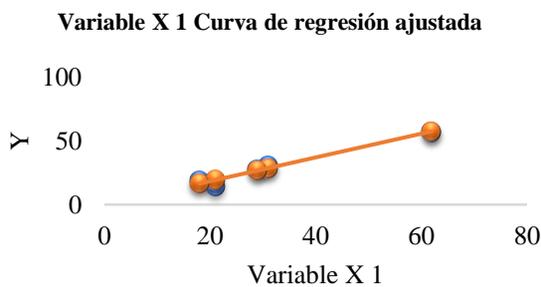


Gráfico 3 Resultados. Gráfico de Regresión Ajustada
Fuente: Fuente Propia

La Gráfica 3 muestra la Curva de regresión ajustada que presenta una comparación de los valores de Y en comparación con los valores pronóstico de Y.

En el caso de los valores pronóstico es notable que se ajustan a los valores observados, reafirmando un modelo de regresión lineal aceptable. La última gráfica generada corresponde al Gráfico de los residuos, muestra los valores residuos con respecto a los valores que toma la variable X. Lo que permite observar cómo se comportan los valores de X por encima o por debajo de la recta de regresión, donde se espera que los valores no posean una tendencia y puedan considerarse con un comportamiento aleatorio con respecto al eje horizontal. La Gráfica 4 muestra el Gráfico de residuales generado para el caso de estudio.

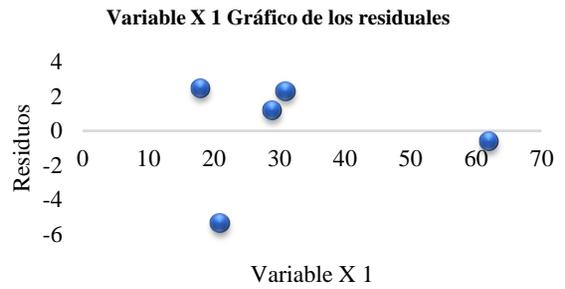


Gráfico 3 Resultados. Gráfico de los residuos
Fuente: Fuente Propia

Distinguimos en la Gráfica de los residuos que se cumple con el residuo de varianza homogénea o de homocedasticidad.

Para terminar, considerando el análisis de coeficientes, donde se concluye que el intercepto no debe considerarse en el modelo de regresión estudiado, se aplica nuevamente la herramienta de análisis de datos de MS Excel, omitiendo el valor de constante igual a cero. La Figura 6 muestra la configuración de la herramienta de análisis de datos de MS Excel con las opciones del nuevo modelo de regresión.

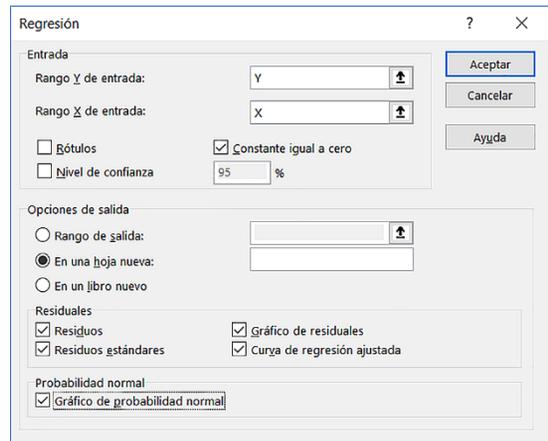


Figura 6 Nuevo análisis de Regresión en MS Excel 2016
Fuente: Fuente Propia

La Tabla 5 muestra la nueva tabla de resumen del modelo de regresión lineal generado para el caso de estudio.

Estadísticas de la regresión	
Coefficiente de correlación múltiple	0.99622452
Coefficiente de determinación R ²	0.9924633
R ² ajustado	0.7424633
Error típico	3.23404619
Observaciones	5

Tabla 5 Resultados. Análisis de Regresión Definitivo
Fuente: Fuente Propia

El nuevo valor del coeficiente de determinación R^2 es de 0.9924633 para el Modelo de Regresión Lineal definitivo (véase Anexo 1, 2 y 3), lo que implica la aceptación de este.

Conclusiones

La nueva era tecnológica está inmersa en grandes volúmenes de datos, y pueden diseñarse mecanismos para extraer información significativa en pro del beneficio de las organizaciones. En particular, en el presente artículo se ha mostrado una estrategia para integrar una herramienta de Big data con otra herramienta para generar un estudio de Regresión Lineal que permitieron evaluar la existencia de una comunicación bilateral.

Puntualmente, en el caso de estudio se comprueba, sin lugar a duda, la comunicación bilateral entre un grupo de tres docentes del área disciplinaria de ingeniería en sistemas computacionales. Incluso, dicho grupo han formado un comité de colaboración y trabajan en una línea de investigación común, con la visión de formar en un futuro próximo un cuerpo académico de investigación.

Además, la participación integral del comité es relevante en diversos proyectos de investigación como “Diseño e Implementación de Educación a Distancia en la Facultad de Ingeniería” con clave de registro 066/UAC/2015, “Diseño e implementación del Sistema de Convenios de la Facultad de Ingeniería” con clave de registro 087/UAC/2017, “Aplicación Web de Administración de Servicios Odontológicos de la Facultad de Odontología” con clave de registro 088/UAC/2017, “Plataforma Virtual de Gestión de Asesorías Académicas” con clave de registro 096/UAC/2018, entre otros. Claramente, la comunicación bilateral evidencia un trabajo colaborativo para el caso de estudio.

El campo de aplicación de herramientas de Big Data es muy amplio, y van emergiendo debido a la utilidad que brindan para el análisis de información con volúmenes considerables de datos.

Referencias

- Anand Deshpande, Manish Kumar. (2018). Artificial Intelligence for Big Data: Complete guide to automating Big Data solutions using Artificial Intelligence techniques. Packt Publishing: Packt Publishing Ltd.
- Badii, M. H., Guillen, A., Cerna, E., Valenzuela, J., y Landeros, J. (2012). Análisis de Regresión Lineal Simple para Predicción. Revista Daena (International Journal of Good Conscience).
- Camargo Vega, J. J., Camargo Ortega, J. F., y Joyanes Aguilar, L. (2015). Conociendo Big Data. Facultad de Ingeniería.
- Carrasquilla Batista, A., Chacón Rodríguez, A., Núñez Montero, K., Gómez Espinoza, O., Valverde-Cerdas, J., y Guerrero Barrantes, M. (2016). Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal. Revista Tecnología en Marcha.
- Garriga Domínguez, A. (2016). Nuevos retos para la protección de datos personales. En la Era del Big Data y de la computación ubicua. Dykinson.
- Gutiérrez González Eduardo y Vladimirovna Panteleeva Olga (2016). Estadística inferencial 1 para ingeniería y ciencias. Grupo Editorial Patria.
- Hernández Leal Emily J., Duque Méndez Néstor D. y Moreno Cadavid Julián. (2017). Big Data: una exploración de investigaciones, tecnologías y casos de aplicación. TecnoLógicas, Vol 20, S/N.
- Immersion. Immersion.media.mit.edu. Recuperado de: <https://immersion.media.mit.edu>
- Lavalle Andrea L., Micheli Elda B., Rubio Natalia. (2016). Análisis didáctico de regresión y correlación para la enseñanza media. Revista latinoamericana de investigación en matemática educativa.
- Levin Richard y Rubin David (2004). Estadística para Administración y Economía. Editorial Pearson Prentice Hall. 7ª Edición.

Martínez Martínez, S., y Lara Navarra, P. (2014). El Big data transforma la interpretación de los medios sociales. El profesional de la información.

Moreno, R. S. G., y Pereira, R. (2017) Aplicaciones de la Big Data a la Computación Urbana.

Rigollet Pierre (2016). Cuadro resumen y cuadros de mando. Tratamiento y análisis de grandes volúmenes de datos con Excel 2016. Ediciones ENI. 2º Edición.

Sánchez Fernández María D. (2014). Comunicación efectiva y trabajo en equipo. Editorialcep.

Vegas, A. M. E., y Reverter, F. (2017). Big Data. Hacia la cuarta revolución industrial. Edicions Universitat Barcelona.

Agradecimientos

Un agradecimiento al apoyo otorgado por la Universidad Autónoma de Campeche, al director de la Facultad de Ingeniería, M.C.C. Guadalupe Manuel Estrada Segovia y a la Coordinadora de Sistemas Computacionales, MTE Nancy Georgina Ortiz Cuevas. De igual forma, extendemos el agradecimiento a todas las personas que directa o indirectamente colaboraron para el logro de esta investigación.

Anexos

Análisis de Varianza					
	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F
Regresión	1	5509.16378	5509.16378	526.736298	0.00018118
Residuos	4	41.836219	10.4590547		
Total	5	5551			

Tabla 6 Análisis de Varianza del Modelo de Regresión definitivo

Fuente: Fuente Propia

Variable X 1 Curva de regresión ajustada

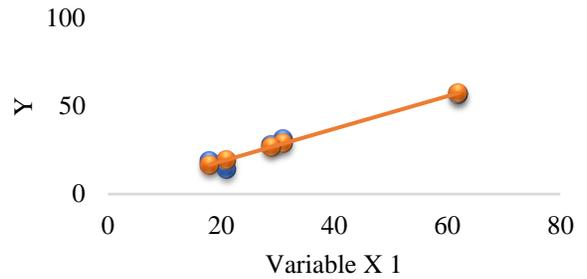


Grafico 6 Gráfico de los residuales del Modelo de Regresión definitivo

Fuente: Fuente Propia

Variable X 1 Gráfico de los residuales

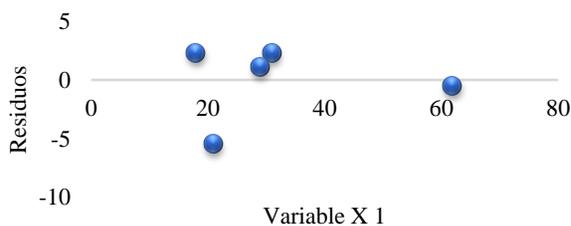


Grafico 5 Gráfico de Curva de regresión ajustada de la variable X del Modelo de Regresión definitivo

Fuente: Fuente Propia