TRABAJO DE GRADO
Opción Seminario-Diplomado.

Análisis de Datos sobre el éxito académico de estudiantes con respecto a la profesión de sus padres

Corporación Universitaria Remington
Facultad de Ingeniería
Seminario de Ciencia de Datos

Juan Esteban Carrasquilla Barrientos[1]

Ivonne Castaño Osorio
Jhon Edison Amórtegui Granada

Opción de Trabajo de grado Seminario-Diplomado
2024

[1] Estudiante de 9.° Semestre de Ingeniería de Sistemas de UniRemington Pereira. E-mail: juan.carrasquilla.9347@miremington.edu.co

Tabla de contenidos

Resumen

Este documento está basado en la aplicación de la metodología CRISP del Análisis de Datos para el entendimiento y la interpretación de un data set libre. Siendo esta metodología esencial para la extracción de información relevante de un conjunto de datos. Se usan métodos y técnicas como lo es el análisis descriptivo haciendo enfoque en distintas áreas como en este caso que va dirigido a el consumo de tabaco en jóvenes, con este ejercicio se logra dar un mayor valor a los datos que nos brindan, encontrando patrones que no se podían ver a simple vista y a partir de ahí tomar decisiones

Palabras clave: Metodología - Patrones – Análisis de Datos – Consumo – Conjunto de Datos.

Pregunta orientadora de la búsqueda

Se inicia la investigación del éxito académico de los estudiantes según la profesión de los padres debido a que hay una aparente relevancia cuando los padres de los estudiantes alcanzan un grado de escolaridad mayor, así mismo se evidencian en algunos casos cuando la situación de lo contrario. “El término éxito está sujeto a la realidad y aspiración de cada individuo tomando en cuenta las potencialidades que le permitan desarrollarse de manera integral. En el ámbito educativo.” (Lira & Aular, 2018)

Según los anterior se plantea la siguiente pregunta: ¿Desde la metodología del Análisis de Datos qué tanto influye el éxito académico en los estudiantes si los padres tienen un nivel de educación alto?

Metodología de búsqueda de la información

Entendimiento del negocio:

Este paso inicia dedicando tiempo a la recopilación de información para saber cuál es y cómo funciona el Core del negocio. Es fundamental para entender la problemática que tiene la empresa o lo que están buscando para que así se pueda generar una buena pregunta.

Enfoque analítico:

Después de identificar el problema o la necesidad de la empresa, se escoge la estrategia analítica más acorde a las necesidades del cliente junto con una persona que será la guía para definir el método más eficiente.

Cuando ya se tiene la pregunta se escoge el método analítico según los patrones que se hayan encontrado y en base a esto se decide si se utiliza el modelo descriptivo o predictivo.

Requisitos de datos:

Para esto es necesario analizar nuestro dataset, con la finalidad de poder darle un enfoque a la búsqueda y poner generar un análisis basado en datos reales, así se evita reprocesos en la investigación.

Recopilación de datos:

Se realiza una recopilación inicial de datos para evaluar si cumplen con los requisitos necesarios; cuando ya se tiene el dataset con la información concreta, se puede tener una visual más amplia de lo que se quiere analizar.

De acuerdo con nuestra pregunta planteada se debe identificar que datos hacen falta para el análisis y así poder reemplazarlos o de ser el caso, dar otro enfoque a el interrogante.

Comprensión de datos:

En esta etapa se utiliza tácticas de análisis descriptivo para hacer una comparativa de las variables relevantes y así poder identificar qué relación tienen entre sí del dataset. En este paso se puede usar el apoyo de histogramas con la finalidad de compactar los datos para que se pueda realizar un análisis y lectura más precisa.

"La fase de comprensión de datos comienza con la recopilación inicial de datos y continúa con actividades que le permiten familiarizarse con los datos, identificar problemas de calidad de los datos, descubrir los primeros conocimientos sobre los datos y/o detectar". (Chapman et al., 1999)

Preparación de datos:

Este paso se cataloga como el más extenso ya que se deben de preparar los datos y pulirlos con la finalidad de automatizar el proceso de búsqueda y así reducir el tiempo considerablemente para brindar más a el modelado. En la preparación de los datos se validan que no estén duplicados, faltantes o errores.

Modelado:

En esta etapa, se busca abordar dos preguntas fundamentales: en primer lugar, cuál es la finalidad del modelado de datos y, en segundo lugar, cuáles son los atributos de este procedimiento. El modelado de datos se centra en la creación de modelos descriptivos o predictivos, los cuales se fundamentan en un enfoque analítico, ya sea estadístico o por aprendizaje automático.

Sustentación teórica de la pregunta

El éxito académico si bien es un factor determinante para la vida de una persona, se puede decir que este término es subjetivo ya que dependerá de las metas que se plantee una persona, sin embargo, si se puede decir que las variables como habilidades cognitivas, motivación e interés, hábitos de estudio, autoestima entre otros son las que marcan más para que el estudiante triunfe en su vida académica. En esta ocasión se resaltará el factor de los padres, específicamente su nivel de educación, ya que a partir de allí se desprenden infinidad de posibilidades como lo es el factor económico y el apoyo con conocimiento adquirido según si es educación superior o no.

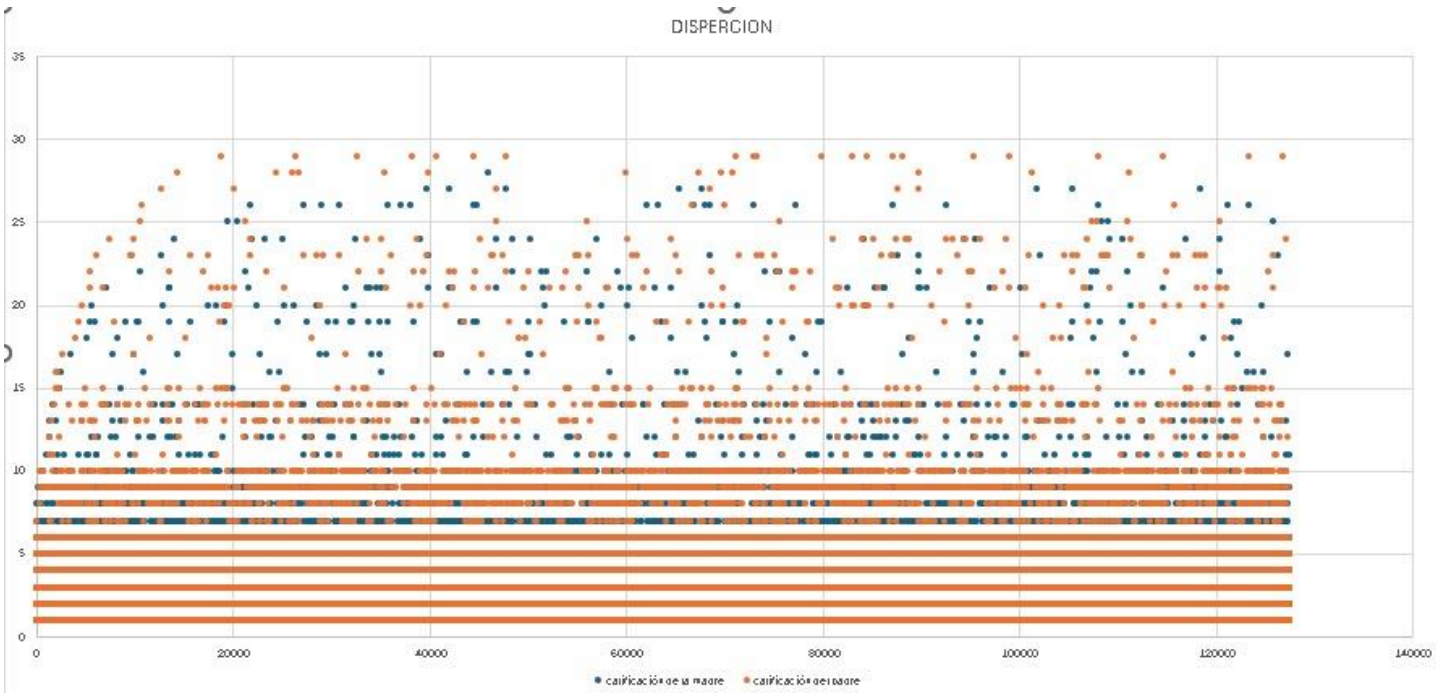
A partir de este contexto se decide hacer un análisis sobre casos de éxito académico en distintos ambientes, abarcando desde el bachiller hasta estudios superiores para determinar con hechos propios la veracidad de esta inclinación algo notable en diferentes sectores del mundo, tomando como punto base el dataset expuesto en la plataforma kaggle.

Gráfica 1. **Histograma**

(Carrasquilla Barrientos, Histograma)

La gráfica muestra la relación entre la profesión de los padres y el curso del hijo. Los datos se presentan en forma de barras agrupadas, donde cada barra representa el número de hijos que siguen un determinado curso en función de la profesión de sus padres

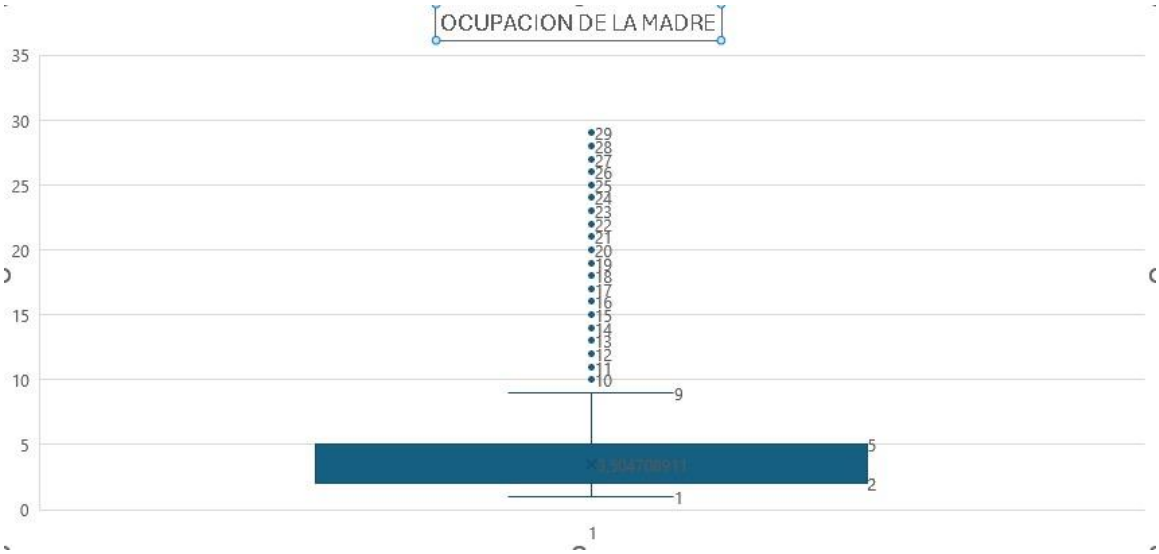
Gráfica 2. **Dispersión**



(Carrasquilla Barrientos, Gráfica de Dispersión)

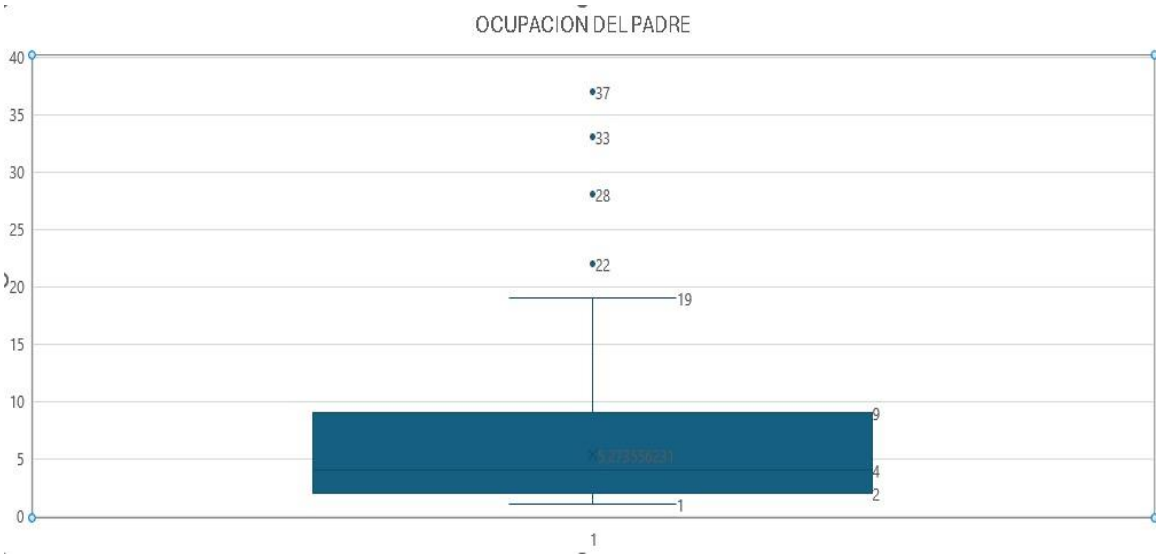
El diagrama de dispersión muestra la relación entre la calificación de matemática y la calificación de español de un grupo de estudiantes. Cada punto en el gráfico representa a un estudiante, y la ubicación del punto indica las calificaciones del estudiante en ambas materias.

Gráfica 3. **bigotes de la ocupación de la madre**



El gráfico de barras muestra que la distribución de la ocupación de la madre en este grupo de personas no es uniforme. La mayoría de las madres son amas de casa o profesionales de la salud.

Gráfica 4. **bigotes de la ocupación del padre**



(Carrasquilla Barrientos)

El gráfico de barras muestra que la distribución de la ocupación del padre en este grupo de personas no es uniforme. La mayoría de los padres son obreros o trabajadores manuales.

Tabla 1**. Análisis de Varianza**

Grados	Cuadro	Cuadro	Cuadro	F	Crítico de F
Regresión	12	96769.053	8064.0877	1496.0236	0
Residuos	127155	685409.68	5.3903478		
Total	127167	782178.73			

Coeficiente	0.3517347
Coeficiente	0.1237173
R2 ajustado	0.1236346
Error típico	2.3217123
Observación	127168

(Carrasquilla Barrientos, Gráfica de Regresión)

Las dos tablas de regresión proporcionan información sobre la relación entre las variables independientes y la variable dependiente. El modelo de regresión es estadísticamente significativo, lo que significa que hay evidencia suficiente para rechazar la hipótesis nula de que no hay relación entre las variables. El R2 ajustado indica que el modelo explica el 12.36% de la variación de la variable dependiente

Tabla 2**. Regresión**

Coeficientes	*Error Típico*	*Estadístico t*	*Probabilidad*	*Inferior a 95%*	*Superior a 95%*	*Inferior a 95%*	*Superior a 95%*
Intercepción	0,8650102	0,0953573	9,0712528	1,193819	0,6781116	1,0519089	0,6781116
1	-0,1331991	0,0054965	-24,2333	1,95E-129	-0,1439722	-0,122426	-0,1439722
1	0,024804	0,0013603	18,23385	3,46E-74	0,0221378	0,0274702	0,0221378
1	0,2086022	0,0279238	7,4704074	8,05E-14	0,1538721	0,2633324	0,1538721
1	0,3730689	0,0054449	68,51654	0	0,3623969	0,3837409	0,3623969
1260	-0,0004172	7,22E-05	-5,7761609	7,66E-09	-0,0005588	-0,0002756	-0,0005588
1	0,0011981	0,003616	0,3313326	0,7403938	-0,0058893	0,0082855	-0,0058893
1	0,0154827	0,0034009	4,5524961	5,31E-06	0,0088169	0,0221484	0,0088169
1	0,0070432	0,002929	2,404653	0,0161892	0,0013024	0,012784	0,0013024
1	0,0081095	0,0019052	4,2564891	2,08E-05	0,0043753	0,0118437	0,0043753
122	-0,0027723	0,0006322	-4,3852043	1,17E-05	0		

Existe una diferencia significativa en la media de la variable dependiente entre los tres niveles del factor A y existe una diferencia significativa en la media de la variable dependiente entre los tres niveles del factor B, sin embargo, no hay interacción significativa entre el factor A y el factor B.

Tabla 3. **Descriptiva.**

do de Aplicación		Orden de Aplicación		Curso		Asistencia diurna/nocturna		Titulación previa		Titulación previa (grado)		calificación de la madre		calificación del padre		ocup
Media	2,4568271	Media	1,6461803	Media	7,3375347	Media	1,0834262	Media	1,362076	Media	1323,5933	Media	2,9287149	Media	3,2591	
Error típico	0,0069527	Error típico	0,0034522	Error típico	0,0143409	Error típico	0,0007747	Error típico	0,0035738	Error típico	0,3068172	Error típico	0,0052263	Error típico	0,0054805	
Mediana	2	Mediana	1	Mediana	6	Mediana	1	Mediana	1	Mediana	1331	Mediana	2	Mediana	3	
Moda	1	Moda	1	Moda	3	Moda	1	Moda	1	Moda	1331	Moda	1	Moda	3	
Desviación e	2,4818119	Desviación e	1,2323031	Desviación e	5,1190688	Desviación e	0,2765264	Desviación e	1,2756903	Desviación e	109,52043	Desviación e	1,8655565	Desviación e	1,9562945	
Varianza de l	6,1593901	Varianza de l	1,5185709	Varianza de l	26,204865	Varianza de l	0,0764669	Varianza de l	1,6273859	Varianza de l	11994,725	Varianza de l	3,480301	Varianza de l	3,827088	
Curtosis	12,545362	Curtosis	3,4199376	Curtosis	-0,3466182	Curtosis	7,077986	Curtosis	29,395478	Curtosis	1,1786296	Curtosis	17,667243	Curtosis	23,775871	
Coeficiente	3,1967742	Coeficiente	2,0470187	Coeficiente	0,7338109	Coeficiente	3,0129512	Coeficiente	4,892086	Coeficiente	0,2275363	Coeficiente	2,3542759	Coeficiente	2,7734354	
Rango	19	Rango	9	Rango	19	Rango	1	Rango	27	Rango	950	Rango	27	Rango	28	
Mínimo	1	Mínimo	0	Mínimo	1	Mínimo	1	Mínimo	1	Mínimo	950	Mínimo	1	Mínimo	1	
Máximo	20	Máximo	9	Máximo	20	Máximo	2	Máximo	28	Máximo	1900	Máximo	28	Máximo	29	
Suma	313044	Suma	209753	Suma	934934	Suma	138048	Suma	173553	Suma	168649610	Suma	373171	Suma	415268	
Cuenta	127418	Cuenta	127418	Cuenta	127418	Cuenta	127418	Cuenta	127418	Cuenta	127418	Cuenta	127418	Cuenta	127418	
Nivel de conf	0,0150882	Nivel de conf	0,0074918	Nivel de conf	0,0311213	Nivel de conf	0,0016811	Nivel de conf	0,0077555	Nivel de conf	0,6658286	Nivel de conf	0,0113416	Nivel de conf	0,0118933	

ocupación de la madre		ocupación del padre		grado de admisión			Género		Edad de Inscripción	
,2591	Media	3,5047089	Media	5,2619724	Media	125,04397	Media	1,3145709	Media	22,247822
4805	Error típico	0,00701	Error típico	0,0107355	Error típico	0,0352056	Error típico	0,0013008	Error típico	0,0191538
3	Mediana	2	Mediana	4	Mediana	124	Mediana	1	Mediana	19
3	Moda	2	Moda	2	Moda	120	Moda	1	Moda	18
2945	Desviación e	2,5022584	Desviación e	3,8321245	Desviación e	12,566864	Desviación e	0,4643466	Desviación e	6,8370739
7088	Varianza de l	6,261297	Varianza de l	14,685178	Varianza de l	157,92607	Varianza de l	0,2156178	Varianza de l	46,74558
5871	Curtosis	10,405696	Curtosis	5,2011681	Curtosis	0,842695	Curtosis	-1,362133	Curtosis	5,883421
4354	Coeficiente	2,2504427	Coeficiente	1,1570857	Coeficiente	0,4277213	Coeficiente	0,7986792	Coeficiente	2,3781607
28	Rango	28	Rango	42	Rango	95	Rango	1	Rango	53
1	Mínimo	1	Mínimo	1	Mínimo	95	Mínimo	1	Mínimo	17
29	Máximo	29	Máximo	43	Máximo	190	Máximo	2	Máximo	70
5268	Suma	446563	Suma	670470	Suma	15932853	Suma	167500	Suma	2834773
7418	Cuenta	127418	Cuenta	127418	Cuenta	127418	Cuenta	127418	Cuenta	127418
8933	Nivel de conf	0,0152125	Nivel de conf	0,0232974	Nivel de conf	0,0764002	Nivel de conf	0,002823	Nivel de conf	0,0415659

(Carrasquilla Barrientos, Gráfica descriptiva)

Los datos indican que la mayoría de los estudiantes son mujeres, asisten a clases diurnas y no tienen titulación previa. Las calificaciones de los padres y el grado de admisión parecen estar relacionados con la edad de inscripción. Sin embargo, se necesita más información para comprender mejor las relaciones entre estas variables.

Tabla 4. **Covarianza.**

	Modo de Aplicac	en de Aplicac	Curso	cia diurna/nc	tulación previ	ción previa (g	ación de la m	icación del p	ación de la m	pación del p	ado de admisi	Género	id de Inscripción
Modo de Apli	6,1593417												
Orden de Apl	-0,447251	1,518559											
Curso	1,6757113	-0,838994	26,20466										
Asistencia di	0,1080925	-0,049356	0,43497	0,0764663									
Titulación pr	0,8310465	-0,207306	0,6580925	0,0445002	1,6273731								
Titulación pr	-11,37468	-1,856501	-30,88368	-2,053962	5,9909	11994,631							
calificación c	0,3047008	-0,111058	0,6473326	0,0704591	0,2235535	-0,347809	3,4802737						
calificación c	0,1818213	0,0039634	0,3464474	0,0090707	0,1810531	4,67952	0,5580484	3,827058					
ocupación d	0,2781154	-0,02386	0,5198204	0,0074553	0,3246108	2,0758215	0,0345184	0,5176156	6,2612478				
ocupación d	0,3591416	-0,035085	0,1379796	-0,012304	0,3369689	2,0271542	0,1428095	0,7249544	4,2959772	14,685063			
grado de adn	-0,765038	-0,61863	-0,997035	-0,037125	1,3805529	773,11227	-0,249506	0,2819638	-0,746848	-1,153957	157,92483		
Género	0,1432643	-0,069654	0,3961287	0,0100544	0,0573723	-2,992429	0,0206741	0,0507444	0,036912	0,0332576	-0,229116	0,2156161	
Edad de Insc	4,6332133	-2,021531	7,4912937	0,9229673	2,6134512	-82,30058	2,6525999	0,4036039	0,698275	0,9504206	-8,517953	0,6416895	46,745213

(Carrasquilla Barrientos, Gráfico de covarianza)

La mayoría de los estudiantes se matriculan en el curso de forma presencial. De estos que están presencial tienen un mejor rendimiento académico que los que se matriculan de forma no presencial. No hay una relación significativa entre la titulación previa del estudiante y su rendimiento académico. pero si el rendimiento académico de los estudiantes está relacionado con la ocupación de sus padres. Los estudiantes cuyos padres tienen una ocupación de mayor nivel educativo tienen un mejor rendimiento académico, también está relacionado con el género y la edad ya que las mujeres tienen un mejor rendimiento académico que los hombres y los estudiantes más jóvenes tienen un mejor rendimiento académico que los estudiantes más mayores.

Conclusiones.

Se define que hay una cierta correlación entre la profesión de los padres y el curso que eligen los hijos. Los hijos de padres con profesiones de alto nivel educativo o ingresos profesionales tienden a elegir cursos más exigentes y prestigiosos, por ejemplo

Las profesiones más nombradas entre los hijos de padres con profesiones de buen rango según su alcance educativo o ingresos profesionales son medicina, derecho, ingeniería y administración de empresas, mientras que las profesiones más nombradas entre los hijos de padres con profesiones de bajo rango según su alcance educativo o ingresos no profesionales son técnicos, operarios y artesanos. Así mismo se evidencia que existe una correlación positiva entre las calificaciones de los padres y las calificaciones de sus hijos**,** significa que los hijos de padres con calificaciones altas tienden a tener calificaciones altas mientras que los hijos de padres con calificaciones bajas tienden a tener calificaciones bajas.

Se ha señalado que el nivel académico de los jefes de familia influye en el nivel de involucramiento en la vida académica de los adolescentes, es decir a mayor nivel de escolaridad toman mayor participación crítica que aquellos que demuestran menos niveles educativos. (Hernández & Hernández, s.f.)

Se tiene evidencia de que las expectativas de los padres son un elemento que influye en la toma de decisiones en las elecciones académicas a futuro de los jóvenes. (Hernández & Hernández, s.f.)

Finalmente se puede concluir que el análisis de datos en el área de la ingeniería en sistemas se ha vuelto indispensable para generar programas más confiables, eficientes y

centrados en los usuarios, a su vez es una herramienta eficaz para dar solución a ciertas problemáticas que se puedan estar generando y no encuentren la causa raíz.

Lista referencias

Carrasquilla Barrientos, J. E. (s.f.). *Gráfica de bigotes.*

Carrasquilla Barrientos, J. E. (s.f.). *Gráfica de Dispersión.*

Carrasquilla Barrientos, J. E. (s.f.). *Gráfica de Regresión.*

Carrasquilla Barrientos, J. E. (s.f.). *Gráfica descriptiva.*

Carrasquilla Barrientos, J. E. (s.f.). *Gráfico de covarianza.*

Carrasquilla Barrientos, J. E. (s.f.). *Histograma.*

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (s.f.). *CRISP-DM 1.0.*

Hernández, C. A., A.Cárdenas, C., O.Romero, P., & Hernández, M. (s.f.). *Los Padres de Familia y el Logro Académico de losAdolescentes de una Secundaria en Milpa Alta, Ciudad deMéxicoo*. Obtenido de research.ebsco.com: https://research.ebsco.com/c/4sprsq/viewer/pdf/qymhqcgs6j

Hernández, C. A., A.Cárdenas, C., O.Romero, P., & Hernández, M. (s.f.). *Los Padres de Familia y el Logro Académico de losAdolescentes de una Secundaria en Milpa Alta, Ciudad deMéxicooh*. Obtenido de research.ebsco.com: https://research.ebsco.com/c/4sprsq/viewer/pdf/qymhqcgs6j

Lira, A., & Aular, E. (s.f.). *Factores asociados al éxito académico en la UNEFA BEJUMA: caso: Carrera de Ingeniería*. Obtenido de elibro.net: https://elibro.net/es/ereader/remington/119369

OPENDATA. (s.f.). Obtenido de https://opendata.agriculture.gov.ie/dataset/h5n1-wild-bird-species-identification

OPS. (s.f.). Obtenido de https://www.paho.org/es/temas/influenza-aviar