TRABAJO DE GRADO
Opción Seminario-Diplomado.

Análisis de Datos sobre la Deserción Escolar en Dosquebradas, Risaralda según la ubicación de las Instituciones Educativas por Comuna año lectivo de 2023

Corporación Universitaria Remington
Facultad de Ingenierías
Ingeniería de Sistemas

Oscar Alberto Campiño Medina[1]
Julio César Sicachá Amaya[2]
Ivonne Castaño Osorio[3]
John Edisson Amortegui Granada[4]

Seminario de Grado
2024

[1] Estudiante de noveno semestre de Ingeniería de Sistemas Uniremington sede Pereira. E-mail: oscar.campino.7513@miremington.edu.co
[2] Estudiante de noveno semestre de Ingeniería de Sistemas Uniremington sede Pereira. E-mail: julio.sicacha.8580@miremington.edu.co
[3] Asesor Temático Uniremington sede Pereira. E-mail: Ivonne.castano@uniremington.edu.co
[4] Asesor Metodológico Uniremington sede Pereira. E-mail: jhon.amortegui@uniremington.edu.co

Dedicatoria

A la familia por el apoyo recibido constantemente durante toda la carrera.

Agradecimientos

A los compañeros por el apoyo y el esfuerzo.

Tabla de contenidos

Tabla de Gráficos

Tablas

Resumen

Para el desarrollo del presente trabajo, se tuvo en cuenta una base de datos proporcionada por el MEN (Ministerio de Educación Nacional) que contiene información detallada del total de estudiantes registrados en el sistema educativo oficial para el año lectivo de 2023; así mismo, los datos de aquellos estudiantes que una vez matriculados cambiaron de estado a retirados dentro de la misma vigencia. Con esta información y aplicando la analítica de datos se obtuvo la clasificación por comuna de las instituciones educativas del municipio de Dosquebradas – Risaralda, en donde se analizó en cuál de ellas se encuentra la mayor cantidad de retiros, los cuales a largo plazo fueron caracterizados como deserción escolar de estudiantes. Igualmente, se realizó un análisis de la georreferenciación de las instituciones educativas según su estrato socio-económico para determinar si la ubicación de las mismas tiene una relación con los retiros de los alumnos. De igual manera, se tuvo en cuenta la tasa de deserción con las variables de nivel educativo respecto a la edad; con base a lo anterior, se pudo determinar el factor de extraedad en la continuidad de los alumnos en el sistema educativo, teniendo en cuenta las estrategias desarrolladas por el MEN en metodologías flexibles para el acceso y permanencia de los NNJA en extraedad. Por último, se pudo evidenciar que las instituciones educativas ubicadas en los estratos 1 y 2, contienen un alto porcentaje de alumnos retirados

Palabras clave: Índice de deserción, información, clasificación, análisis y base de datos

Pregunta orientadora de la búsqueda

Los datos siempre han estado en nuestro entorno, antes se procesaban de diferentes maneras y cada día evoluciona la forma de analizarlos; tal como lo dice Maldonado (2022):

> La minería de datos, como concepto, se encuentra algo <<pasada de moda>>, siendo reemplazada por <<ciencia de datos>> (data science, o DS). La ciencia de los datos engloba esencialmente las mismas técnicas, pero populariza la inclusión de nuevas tecnologías y tendencias, tales como herramientas de Big Data, manejos de datos no estructurados y aprendizaje profundo o Deep Learning. (P. 24)

En la era actual es muy importante utilizar una metodología de análisis de datos, la cual se basa en la forma de convertir información en toma de decisiones fundamentadas. Gran cantidad de datos son generados diariamente lo que lleva a contar con análisis y un enfoque metodológico que se vuelve necesario para generar un significado y valor a esta información.

El uso de una metodología estructurada de análisis de datos asegura la coherencia y la precisión de los resultados, teniendo en cuenta que estos deben ser organizados y estructurados por consiguiente se pueden evitar errores que pueden llevar a conclusiones erróneas.

En el sector educativo, una metodología de análisis de datos puede ser instrumento para el comparar la evolución de matrícula, la identificación de los motivos de los retiros y los posibles alumnos desertores. Este último es un problema muy

complejo que afecta la cobertura educativa y la asignación de recursos para las instituciones educativas, lo cual permite generar acciones para buscar soluciones efectivas y basadas en datos reales.

Con base en lo anterior se puede plantear el siguiente cuestionamiento ¿Cuál es el impacto de la ubicación de las instituciones educativas en las diferentes comunas de Dosquebradas en relación con la tasa de deserción escolar entre estudiantes de educación regular durante la vigencia 2023 utilizando metodología de Análisis de Datos?

Metodología de búsqueda de la información

La metodología que se empleará para el desarrollo del trabajo será la de CRISP-DM por ser una metodología que se enfoca en el conocimiento del negocio, recopilando datos que existen para hallar la información oculta.

Las fases de la metodología de la búsqueda de información son:

1. **Entendimiento del negocio o proyecto**

Es fundamental para el desarrollo de cualquier proyecto, se enfoca en comprender los requisitos y los objetivos del mismo, para tener clara una definición del problema, sin un marco de referencia adecuado para su interpretación, los datos perderán toda relevancia práctica, estadística y numérica, impidiendo la toma de decisiones significativas.

2. Enfoque Analítico

Es una fase muy importante de la metodología, ya que se pasa del entendimiento del negocio en definir cuál enfoque se llevará a cabo ya sea descriptivo, predictivo, análisis estadístico, aprendizaje automático y enfoques de asociación *Clustering*, con el fin de realizar un análisis concreto para dar respuesta los problemas identificados.

3. Requisitos de Datos

Son las condiciones que deben reunir los datos que se van a utilizar en la implementación y solución de un problema. Dentro de los requisitos de los datos se pueden encontrar su estructura, formato y tipos de variables.

4. Recopilación de Datos

Es la fase en la que se inicia la recolección de los datos de acuerdo con los requisitos aportados inicialmente, la cual se puede realizar de diferentes formas como entrevistas, encuestas, cuestionarios, estudios estadísticos.

Con la recopilación de los datos se puede realizar un análisis de estos y de acuerdo con los resultados arrojados evaluar si se hace necesario aumentar la base de datos o no. En esta fase se puede trabajar con varias bases de datos, teniendo en cuenta que puede llevar al científico de datos a unificarlas y a tomar mejores decisiones.

5. Comprensión de Datos

Es la fase en la que se identifica la calidad de los datos que hacen parte de la base de datos, refinándolos y optimizando la calidad de estos, con el fin de llegar a un mejor análisis de los datos, por medio de gráficas tales como cajas y bigotes, dispersión e histogramas; por ejemplo, se puede revisar si se encuentran sesgos y/o valores atípicos.

6. Preparación de Datos

Es la fase en la que ya se toma decisiones en cuanto a determinar que los datos tengan calidad, integridad y relevancia y con los que se realizará el modelado. En esta fase de la metodología de la ciencia de datos se realiza la normalización y la estandarización de los datos, que consiste en darles un formato consistente en cuanto a medidas consistentes, tipos de datos y formatos de fecha.

De igual manera, se realiza la codificación de los datos, en donde se convierten datos cualitativos en cuantitativos, asignando códigos numéricos a datos de tipo texto. En la transformación de los datos, eliminación de información que se considera que no es necesaria para hacer el análisis.

7. Modelado

Es la fase en la cual se aplican diferentes algoritmos y el modelo a seguir con el fin de conseguir los valores óptimos para llegar al mejor resultado dentro del problema planteado. Adicional a lo anterior, se debe tener en cuenta que en muchas ocasiones (así se considere que se tienen los datos requeridos) se debe realizar una nueva preparación de datos y de esta manera cumplir con los requerimientos de los datos de entrada y conseguir los resultados esperados.

Dentro de la fase de modelado aparece una herramienta importante como lo es el *machine learning* que la convierte en un factor invaluable dentro de la analítica descriptiva y predictiva.

8. Evaluación del Modelo

Se cumple con los objetivos del proyecto. Si la respuesta es negativa o el grupo de trabajo considere que hay lugar a realizar mejoras, se deben aunar esfuerzos con el fin de realizar los cambios necesarios. La eliminación de atributos que estadísticamente no son importantes, la corrección de la entrada de datos y el tratamiento de atributos son algunas de las muchas formas en que pueden ocurrir estos cambios.

9. Implementación

Si todo ha sido hecho correctamente, este será el paso final. El modelo debe implementarse en la producción para agregar valor al negocio. La forma en que se realiza depende del tipo de modelo y del proyecto. Es necesario exponer ese modelo para el acceso, que suele estar almacenado en servidores locales de la propia empresa o en la nube.

10. Retroalimentación

Una vez obtenidos los resultados del modelo implementado, la organización recibe retroalimentación sobre el desempeño del modelo y su impacto en el entorno de implementación. Por ejemplo, esta retroalimentación puede manifestarse en forma de porcentajes de respuesta a una campaña promocional dirigida a un grupo de clientes identificados por el modelo como de alto potencial. Los científicos de datos pueden analizar esta retroalimentación para ajustar el modelo con el fin de mejorar su precisión y utilidad. Algunos o todos los pasos de la evaluación del modelo, la recolección de retroalimentación, el ajuste y la reimplementación del modelo pueden automatizarse para agilizar el proceso de actualización del modelo y lograr resultados más favorables.

Aplicando las fases de la metodología CRISP- DM en este proyecto se detalla lo siguiente:

1. Entendimiento del Negocio

Se requiere determinar si los factores de ubicación de las instituciones educativas tienen alguna relación con la deserción escolar interna del municipio de Dosquebradas.

Se entiende por deserción escolar, tal como lo define el Ministerio de Educación Nacional, al abandono del sistema escolar por parte de los estudiantes, provocado por la combinación de factores que se generan tanto al interior del sistema como en contextos de tipo social, familiar, individual y del entorno. (Ministerio de Educación Nacional, 2017)

Teniendo en cuenta lo anterior, dentro de la problemática de tipo social se encuentra que las instituciones educativas suelen encontrarse distantes de los lugares de residencia de los estudiantes, conllevando, que ante la falta de recursos económicos se opte por parte de los padres de familia, no volver a enviar a sus hijos a las aulas de clase.

De igual manera, en los municipios industriales como el caso de Dosquebradas – Risaralda, las familias pernoctan por espacios cortos de tiempo mientras existe la posibilidad laboral de los padres de familia y una vez se termina la dicha vinculación, nuevamente se inicia con la búsqueda de oportunidades de trabajo, por lo cual muchos de ellos deben emigrar a otras regiones, trayendo como consecuencia el retiro de los estudiantes y la posible deserción escolar de los mismos.

Partiendo de lo descrito con antelación, se pretende realizar un comparativo de la deserción escolar en el municipio de Dosquebradas – Risaralda, en cuanto a las comunas que lo conforman y las instituciones educativas que hacen parte de ellas y con ello establecer en donde es más notorio este flagelo que impacta la educación de ese territorio.

2. Enfoque Analítico

Se define hacer un análisis descriptivo Correlacional teniendo en cuenta por la comprensión integral y detallada para dar respuesta a la pregunta orientadora, por cuanto se va a realizar un comparativo de los retiros anuales internos de los estudiantes en el municipio de Dosquebradas – Risaralda, en lo referente a la georreferenciación de las instituciones educativas sus diferentes comunas.

Además, se puede identificar patrones en el análisis para determinar que los resultados sean efectivos y así poder establecer estrategias de permanencia de los alumnos en el sistema educativo del municipio de Dosquebradas.

3. Requisitos de datos

La data set deberá tener las siguientes características.

- Longitud: Mayor a 1000 registros de retiros.
- Formatos: csv o Excel.
- Base de Datos estructurada.
- Origen de los datos: MEN SIMAT.
- Tipo de contenido: Información de matrículas y retiros de las I.E. oficiales.
- Tipo de datos: Texto y numéricos.
- Tipo de fuente: Información oficial.
- Características: Población registrada académicamente.

4. Recopilación de datos

Se define que la Data Set cumple con las siguientes características.

- Cantidad de registros: 3651
- Completos: Sí
- Coherentes: Sí
- Columnas vacías: No
- Tipo de datos: Numérico y texto
- Formato: Excel
- Número de columnas: 35
- Número de filas: 3651
- Dueño de los datos: MEN SIMAT (Contiene información sensible)

5. Comprensión de los Datos

En esta fase se revisó la información de la base de datos inicial, identificando información que no era relevante para la obtención de los resultados. Por medio de las

gráficas de cajas y bigotes, y dispersión, se pudo establecer que no se encontraban datos erróneos o valores por fuera de los rangos normales.

Las variables que se consideraron que no servían son:

ETC, JERARQUÍA, CALENDARIO, SECTOR, FECHA FIN, NUI, DER_ID, APELLIDO1, APELLIDO2, NOMBRE1, NOMBRE2, BARRIO, APOYO ACADÉMICO ESPECIAL, SRPA, GRUPO, CORREO; por cuanto, se consideró que no llegarían a tener injerencia en el desarrollo del ejercicio y se encontraba información sensible como es la información de nombres y apellidos.

Por otro lado, las variables que se identificaron que servían para la toma de decisiones son AÑO, ESTADO, DANE (identifica el nombre de la institución educativa), CODIGO_DANE (identifica el nombre de la sede la institución educativa), ZONA_SEDE, JORNADA, GRADO, MODELO, MOTIVO_RETIRO, FECHAINI, ESTRATO, GENERO, FECHA_NACIMIENTO, DISCAPACIDAD y PAIS_ORIGEN.

6. Preparación de los Datos

Dentro de esta fase se eliminaron las variables ETC, JERARQUÍA, CALENDARIO, SECTOR, FECHA FIN, NUI, DER_ID, APELLIDO1, APELLIDO2, NOMBRE1, NOMBRE2, BARRIO, APOYO ACADÉMICO ESPECIAL, SRPA, GRUPO, CORREO, por las razones expuestas en el ítem anterior.

En cuanto a las variables ESTADO, INSTITUCIÓN, SEDE, ZONA_SEDE, JORNADA, MODELO, MOTIVO, GÉNERO, DISCAPACIDAD y PAIS_ORIGEN, se generaron nuevas columnas con el fin de codificarlas y de esta manera llevar a cabo el análisis de los datos. Así mismo, creó la columna EDAD para determinar la edad con la

cual se presenta la deserción escolar y la columna COMUNA para establecer en cuál de ellas se encuentra ubicada la institución educativa.

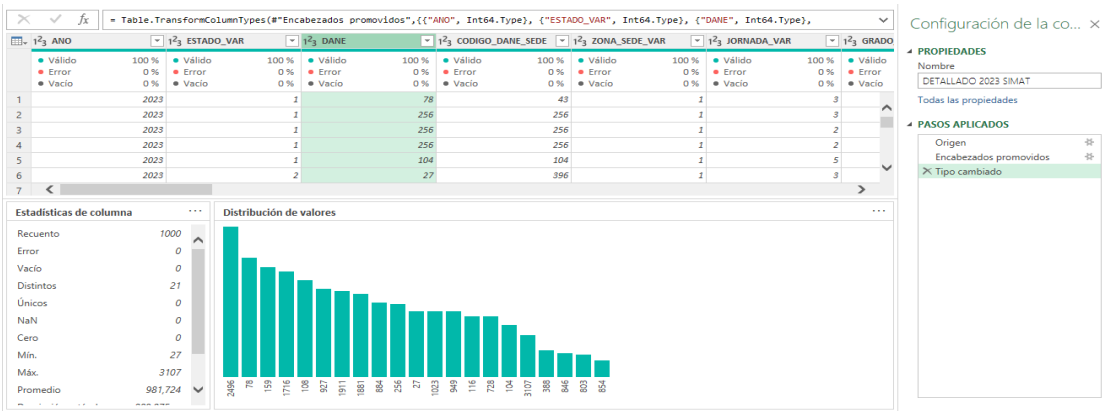


Gráfico 1. limpieza de datos.

En el gráfico anterior (Ver gráfico 1. Limpieza de datos) se puede evidenciar que los datos se encuentran sin errores y sin celdas vacías lo que permite hacer un análisis con información coherente y que el resultado de esta sea satisfactorio.

7. Modelado: Se explica en Sustentación teórica de la pregunta.

8. Evaluación del modelo: Se explica en Sustentación teórica de la pregunta.

9. Implementación: No se aplicará en este proyecto.

10. Retroalimentación: No se aplicará en este proyecto.

Sustentación teórica de la pregunta

Modelado y evaluación del modelo

Para el desarrollo del análisis se debe tener en cuenta las fuentes de información y qué tipo de base de datos se utilizará según como lo define Pulido et al. (2019) “Las bases de datos son las más adecuadas para almacenar datos en un sistema de información debido

a sus diversas características como seguridad, capacidad de recuperación ante fallos y gestión centralizada.” (p. 12)

Con base a lo anterior, se inicia un proceso de tratamiento de los datos con el cual se pretende encontrar bases de datos de gran tamaño y reciben el nombre de big data, la cual se puede definir según Caballero (2015), así:

> La traducción literal de big data serían “datos masivos” o “datos a gran escala” un término muy general. Sin embargo, en el ámbito de la informática, aunque sin perder del todo la ambigüedad, acostumbra a referirse a datos masivos que cumplen tres características las conocidas 3V Volumen, velocidad y variedad. (p. 25)

Una vez se tiene la base de datos y se hace una primera revisión de la información contenida en ella se puede identificar aspectos relevantes con el fin de que sirvan como referentes para un análisis detallado y además poder descubrir lo que narran los datos como lo determina Alcalde (2015) “Se trata de descubrir qué nos cuentan los datos y saber narrarlo. Una vez descubierto lo que nos dicen, es muy relevante saber dimensionar nuestra visualización en función de la complejidad de lo que queremos explicar.” (p. 52)

Otro aspecto fundamental para el ejercicio del análisis de los datos y con la finalidad de que el resultado de este análisis sea el más preciso se define en la limpieza de los datos como lo explica López (2009) “La limpieza de datos es el proceso mediante el cual se detectan y corrigen los errores, y su aplicación es válida tanto en bases de datos de uso operacional, como en aquellos conjuntos que sirven de fuentes de almacenes de datos.” (p. 4)

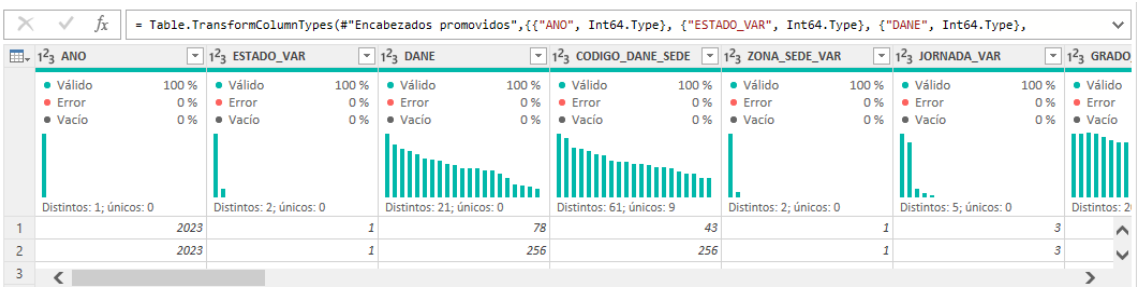


Gráfico 2. Limpieza y coherencia de datos

En el grafico anterior (Ver gráfico 2. Limpieza y coherencia de datos) se puede verificar que los datos son coherentes entre los registros. Con lo anterior, una vez finalizada la limpieza de datos se determina cual será el modelo a seguir para el análisis, teniendo en cuenta los datos encontrados para este trabajo se define la metodología descriptiva correlacional, al respecto Menoyo (2021) afirma que:

> La estadística descriptiva se encarga de extraer información relevante de los datos, de manera que resuman y reflejen sus propiedades más importantes. No pretende alcanzar conclusiones ni demostrar hipótesis, simplemente se centra en describir los datos de una forma que nos sea de utilidad. (p. 65)

Esto permite tener claridad de cuál es el enfoque que se dará al resultado del análisis y con el fin de dar respuesta a la pregunta orientadora, por eso al momento de modelar es importante tener claro lo de dice García et al. (2017). “El objetivo de tal modelación es encontrar los conceptos esenciales presentes en la fase de análisis de requerimientos." (p. 8)

En las siguientes gráficas se puede evidenciar que no se encuentran datos erróneos o valores por fuera de los rangos normales **gráfico 3** (Ver gráfico 3. Cajas y bigotes) **y gráfico 4** (Ver gráfico 4. Dispersión)

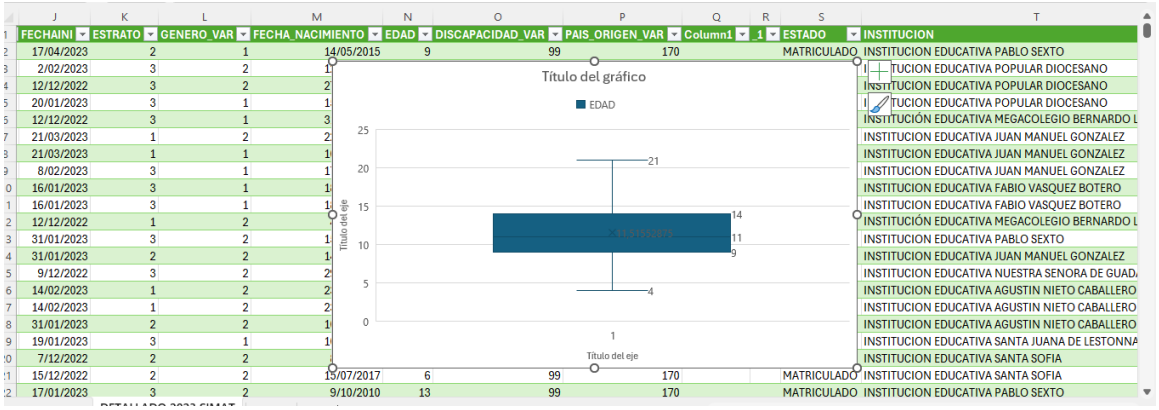


Gráfico 3. Cajas y bigotes

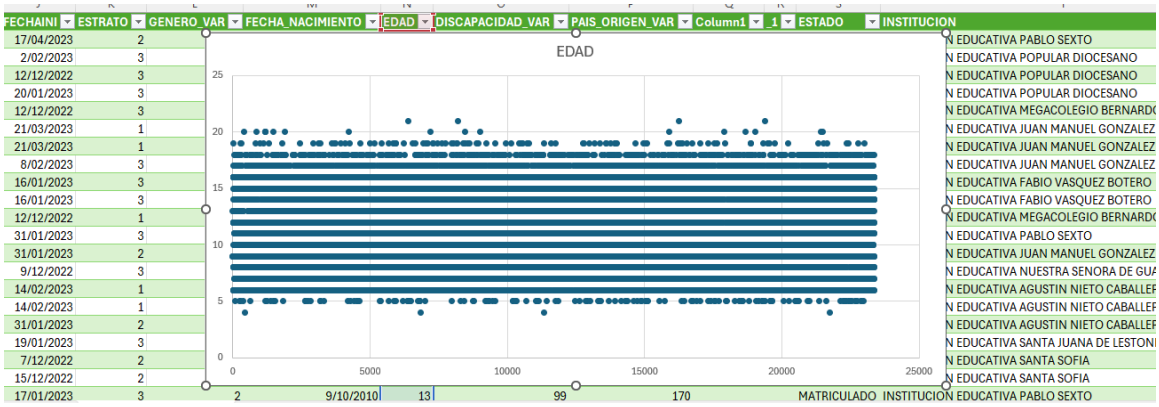


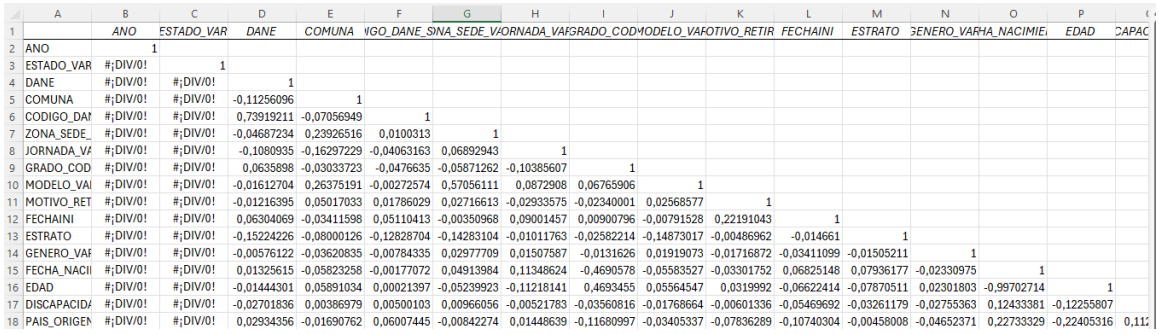
Gráfico 4. Dispersión

	ANO		ESTADO_VAR		DANE		COMUNA		CODIGO_DANE_SEDE		ZONA_SEDE_VAR		JORNADA_VAR		GRADO_COD		MODEL
Media	2023	Media	2	Media	1031,93375	Media	6,33552886	Media	966,133669	Media	1,07826424	Media	2,62998838	Media	6,06586594	Media
Error típico	0	Error típico	0	Error típico	17,1533895	Error típico	0,07777267	Error típico	16,9981048	Error típico	0,0052878	Error típico	0,01719846	Error típico	0,12494487	Error t
Mediana	2023	Mediana	2	Mediana	884	Mediana	7	Mediana	884	Mediana	1	Mediana	2	Mediana	6	Media
Moda	2023	Moda	2	Moda	78	Moda	2	Moda	78	Moda	1	Moda	2	Moda	6	Moda
Desviación e	0	Desviación e	0	Desviación e	871,452962	Desviación e	3,9511274	Desviación e	863,56395	Desviación e	0,26863899	Desviación e	0,87374249	Desviación e	6,34764202	Desvia
Varianza de	0	Varianza de	0	Varianza de	759430,265	Varianza de	15,6114078	Varianza de	745742,695	Varianza de	0,07216691	Varianza de	0,76342593	Varianza de	40,2925592	Varia
Curtosis	#¡DIV/0!	Curtosis	#¡DIV/0!	Curtosis	-0,6704839	Curtosis	-1,37823618	Curtosis	0,46464254	Curtosis	7,87971771	Curtosis	1,95148321	Curtosis	157,782208	Curto
Coeficiente c	#¡DIV/0!	Coeficiente c	#¡DIV/0!	Coeficiente c	0,59545801	Coeficiente c	0,13949232	Coeficiente c	0,84086967	Coeficiente c	3,14223048	Coeficiente c	1,5876239	Coeficiente c	10,8817696	Coefic
Rango	0	Rango	0	Rango	3080	Rango	13	Rango	6461	Rango	1	Rango	3	Rango	99	Rango
Mínimo	2023	Mínimo	2	Mínimo	27	Mínimo	1	Mínimo	19	Mínimo	1	Mínimo	2	Mínimo	0	Mínim
Máximo	2023	Máximo	2	Máximo	3107	Máximo	14	Máximo	6480	Máximo	2	Máximo	5	Máximo	99	Máxin
Suma	5221363	Suma	5162	Suma	2663421	Suma	16352	Suma	2493591	Suma	2783	Suma	6788	Suma	15656	Suma
Cuenta	2581	Cuenta	2581	Cuenta	2581	Cuenta	2581	Cuenta	2581	Cuenta	2581	Cuenta	2581	Cuenta	2581	Cuent
Mayor (1)	2023	Mayor (1)	2	Mayor (1)	3107	Mayor (1)	14	Mayor (1)	6480	Mayor (1)	2	Mayor (1)	5	Mayor (1)	99	Mayor
Menor(1)	2023	Menor(1)	2	Menor(1)	27	Menor(1)	1	Menor(1)	19	Menor(1)	1	Menor(1)	2	Menor(1)	0	Menor
Nivel de conf	0	Nivel de conf	0	Nivel de conf	37,2450099	Nivel de conf	0,16886715	Nivel de conf	36,9078416	Nivel de conf	0,01148136	Nivel de conf	0,03734286	Nivel de conf	0,27129174	Nivel

Tabla 1. Análisis Descriptivos

En este análisis podemos evidenciar la tendencia que tiene la mediana según lo explica Garriga (2009). “La mediana es un valor más apropiado para representar la tendencia central de la distribución.” (p. 58)

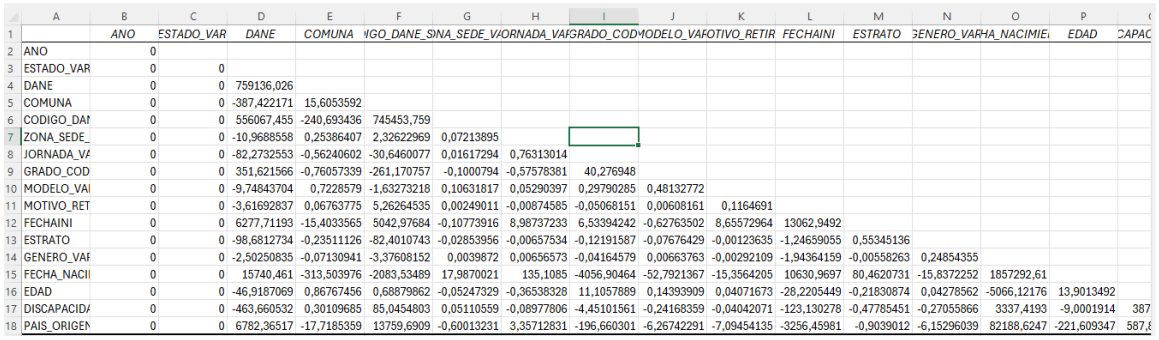
Esta tabla es el resultado del Análisis de Pearson



	ANO	ESTADO_VAR	DANE	COMUNA	IGO_DANE_	SINA_SEDE_V	ORNADA_VAR	GRADO_COD	MODELO_VA	OTIVO_RETIR	FECHAINI	ESTRATO	GENERO_VAR	HA_NACIMIE	EDAD	CAPAC
ANO	1															
ESTADO_VAR	#¡DIV/0!	1														
DANE	#¡DIV/0!	#¡DIV/0!	1													
COMUNA	#¡DIV/0!	#¡DIV/0!	-0,11256096	1												
CODIGO_DAI	#¡DIV/0!	#¡DIV/0!	0,73919211	-0,07056949	1											
ZONA_SEDE_	#¡DIV/0!	#¡DIV/0!	-0,04687234	0,23926516	0,0100313	1										
JORNADA_VA	#¡DIV/0!	#¡DIV/0!	-0,1080935	-0,16297229	-0,04063163	0,06892943	1									
GRADO_COD	#¡DIV/0!	#¡DIV/0!	0,0635898	-0,03033723	-0,0476635	-0,05871262	-0,10385607	1								
MODELO_VAI	#¡DIV/0!	#¡DIV/0!	-0,01612704	0,26375191	-0,00272574	0,57056111	0,0872908	0,06765906	1							
MOTIVO_RET	#¡DIV/0!	#¡DIV/0!	-0,01216395	0,05017033	0,01786029	0,02716613	-0,02933575	-0,02340001	0,02568577	1						
FECHAINI	#¡DIV/0!	#¡DIV/0!	0,06304069	-0,03411598	0,05110413	-0,00350968	0,09001457	0,00900796	-0,00791528	0,22191043	1					
ESTRATO	#¡DIV/0!	#¡DIV/0!	-0,15224226	-0,08000126	-0,12828704	-0,14283104	-0,01011763	-0,02582214	-0,14873017	-0,00486962	-0,014661	1				
GENERO_VAF	#¡DIV/0!	#¡DIV/0!	-0,00576122	-0,03620835	-0,00784335	0,02977709	0,01507587	-0,0131626	0,01919073	-0,01716872	-0,03411099	-0,01505211	1			
FECHA_NACII	#¡DIV/0!	#¡DIV/0!	0,01325615	-0,05823258	-0,00177072	0,04913984	0,11348624	-0,4690578	-0,05583527	-0,03301752	0,06825148	0,07936177	-0,02330975	1		
EDAD	#¡DIV/0!	#¡DIV/0!	-0,01444301	0,05891034	0,00021397	-0,05239923	-0,11218141	0,4693455	0,05564547	0,0319992	-0,06622414	-0,07870511	0,02301803	-0,99702714	1	
DISCAPACIDA	#¡DIV/0!	#¡DIV/0!	-0,02701836	0,00386979	0,00500103	0,00966056	-0,00521783	-0,03560816	-0,01768664	-0,00601336	-0,05469692	-0,03261179	-0,02755363	0,12433381	-0,12255807	
PAIS_ORIGEN	#¡DIV/0!	#¡DIV/0!	0,02934356	-0,01690762	0,06007445	-0,00842274	0,01448639	-0,11680997	-0,03405337	-0,07836289	-0,10740304	-0,00458008	-0,04652371	0,22733329	-0,22405316	0,11

Tabla 2. Análisis de Pearson

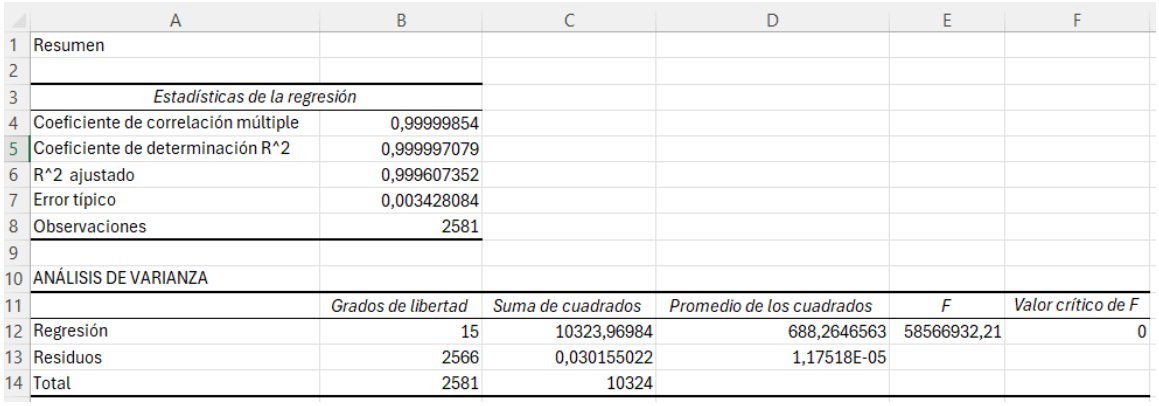
Esta tabla es el resultado del Análisis de Covarianza.



	ANO	ESTADO_VAR	DANE	COMUNA	IGO_DANE_	SINA_SEDE_V	ORNADA_VAR	GRADO_COD	MODELO_VA	OTIVO_RETIR	FECHAINI	ESTRATO	GENERO_VAR	HA_NACIMIE	EDAD	CAPAC
ANO	0															
ESTADO_VAR	0	0														
DANE	0	0	759136,026													
COMUNA	0	0	-387,422171	15,6053592												
CODIGO_DAI	0	0	556067,455	-240,693436	745453,759											
ZONA_SEDE_	0	0	-10,9688558	0,25386407	2,32622969	0,07213895										
JORNADA_VA	0	0	-82,2732553	-0,56240602	-30,6460077	0,01617294	0,76313014									
GRADO_COD	0	0	351,621566	-0,76057339	-261,170757	-0,1000794	-0,57578381	40,276948								
MODELO_VAI	0	0	-9,74843704	0,7228579	-1,63273218	0,10631817	0,05290397	0,29790285	0,48132772							
MOTIVO_RET	0	0	-3,61692837	0,06763775	5,26264535	0,00249011	-0,00874585	-0,05068151	0,00608161	0,1164691						
FECHAINI	0	0	6277,71193	-15,4033565	5042,97684	-0,10773916	8,98737233	6,53394242	-0,62763502	8,65572964	13062,9492					
ESTRATO	0	0	-98,6812734	-0,23511126	-82,4010743	-0,02853956	-0,00657534	-0,12191587	-0,07676429	-0,00123635	-1,24659055	0,55345136				
GENERO_VAF	0	0	-2,50250835	-0,07130941	-3,37608152	0,0039872	0,00656573	-0,04164579	0,00663763	-0,00292109	-1,94364159	-0,00558263	0,24854355			
FECHA_NACII	0	0	15740,461	-313,503976	-2083,53489	17,9870021	135,1085	-4056,90464	-52,7921367	-15,3564205	10630,9697	80,4620731	-15,8372252	1857292,61		
EDAD	0	0	-46,9187069	0,86767456	0,68879862	-0,05247329	-0,36538328	11,1057889	0,14393909	0,04071673	-28,2205449	-0,21830874	0,04278562	-5066,12176	13,9013492	
DISCAPACIDA	0	0	-463,660532	0,30109685	85,0454803	0,05110559	-0,08977806	-4,45101561	-0,24168359	-0,04042071	-123,130278	-0,47785451	-0,27055866	3337,4193	-9,0001914	387
PAIS_ORIGEN	0	0	6782,36517	-17,7185359	13759,6909	-0,60013231	3,35712831	-196,660301	-6,26742291	-7,09454135	-3256,45981	-0,9039012	-6,15296039	82188,6247	-221,609347	587,8

Tabla 3. Análisis de Covarianza

Las tablas 4 y 5 son es el resultado del análisis de Regresión



	A	B	C	D	E	F
1	Resumen					
2						
3	*Estadísticas de la regresión*					
4	Coeficiente de correlación múltiple	0,99999854				
5	Coeficiente de determinación R^2	0,999997079				
6	R^2 ajustado	0,999607352				
7	Error típico	0,003428084				
8	Observaciones	2581				
9						
10	ANÁLISIS DE VARIANZA					
11		*Grados de libertad*	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
12	Regresión	15	10323,96984	688,2646563	58566932,21	0
13	Residuos	2566	0,030155022	1,17518E-05		
14	Total	2581	10324			

Tabla 4. Análisis de Regresión 1

	A	B	C	D	E	F	G	H	I
16		*Coeficientes*	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*uperior 95,0%*
17	Intercepción	0	#N/D	#N/D	#N/D	#N/D	#N/D	#N/D	#N/D
18	DANE	-1,5358E-07	1,19463E-07	-1,285587899	0,19870293	-3,87833E-07	8,0673E-08	-3,878E-07	8,0673E-08
19	COMUNA	-1,59789E-06	1,84651E-05	-0,086535508	0,931047482	-3,7806E-05	3,461E-05	-3,781E-05	3,461E-05
20	CODIGO_DANE_SEDE	2,10226E-08	1,18587E-07	0,177274857	0,85930649	-2,11514E-07	2,5356E-07	-2,115E-07	2,5356E-07
21	ZONA_SEDE_VAR	0,000680141	0,00031249	2,17652221	0,029607137	6,73832E-05	0,0012929	6,7383E-05	0,0012929
22	JORNADA_VAR	-0,000311999	8,08514E-05	-3,85891483	0,000116699	-0,00047054	-0,0001535	-0,0004705	-0,0001535
23	GRADO_COD	-1,56541E-05	1,23894E-05	-1,263502898	0,206523287	-3,99484E-05	8,6402E-06	-3,995E-05	8,6402E-06
24	MODELO_VAR	-3,70785E-05	0,000122205	-0,303411523	0,761600862	-0,000276709	0,00020255	-0,0002767	0,00020255
25	MOTIVO_RETIRO	-0,001297555	0,000202696	-6,401491839	1,82329E-10	-0,001695019	-0,0009001	-0,001695	-0,0009001
26	FECHAINI	2,10869E-05	4,57702E-07	46,07113806	0	2,01894E-05	2,1984E-05	2,0189E-05	2,1984E-05
27	ESTRATO	6,09608E-05	9,38262E-05	0,649720561	0,515930896	-0,000123022	0,00024494	-0,000123	0,00024494
28	GENERO_VAR	0,000211613	0,000135883	1,557321679	0,11951745	-5,48378E-05	0,00047806	-5,484E-05	0,00047806
29	FECHA_NACIMIENTO	2,32505E-05	4,53848E-07	51,22962324	0	2,23605E-05	2,414E-05	2,2361E-05	2,414E-05
30	EDAD	0,008544963	0,000164605	51,91184842	0	0,008222191	0,00886774	0,00822219	0,00886774
31	DISCAPACIDAD_VAR	4,97946E-06	3,48261E-06	1,42980768	0,152893964	-1,84955E-06	1,1808E-05	-1,85E-06	1,1808E-05
32	PAIS_ORIGEN_VAR	5,85457E-07	2,65485E-07	2,205236277	0,027525905	6,48705E-08	1,106E-06	6,487E-08	1,106E-06

Tabla 5. Análisis de Regresión 2

De acuerdo con la matriz de correlaciones, las variables asociadas con la deserción son la edad, la zona de la sede del colegio y la comuna, y en menor medida el estrato y el modelo. De las demás variables pueden decirse que no tienen ninguna injerencia sobre el fenómeno de la deserción. También se ve por los resultados de la regresión que estas variables, al tratarse de manera conjunta, son decisivas sobre este fenómeno para esta población en específico.

Según el **gráfico 5** (Ver gráfico 5. Retiros por Comuna) se puede observar que la mayor cantidad de retiros se refleja en la comuna 2, en la cual están ubicadas las instituciones educativas Santa Sofia y Agustín Nieto Caballero.

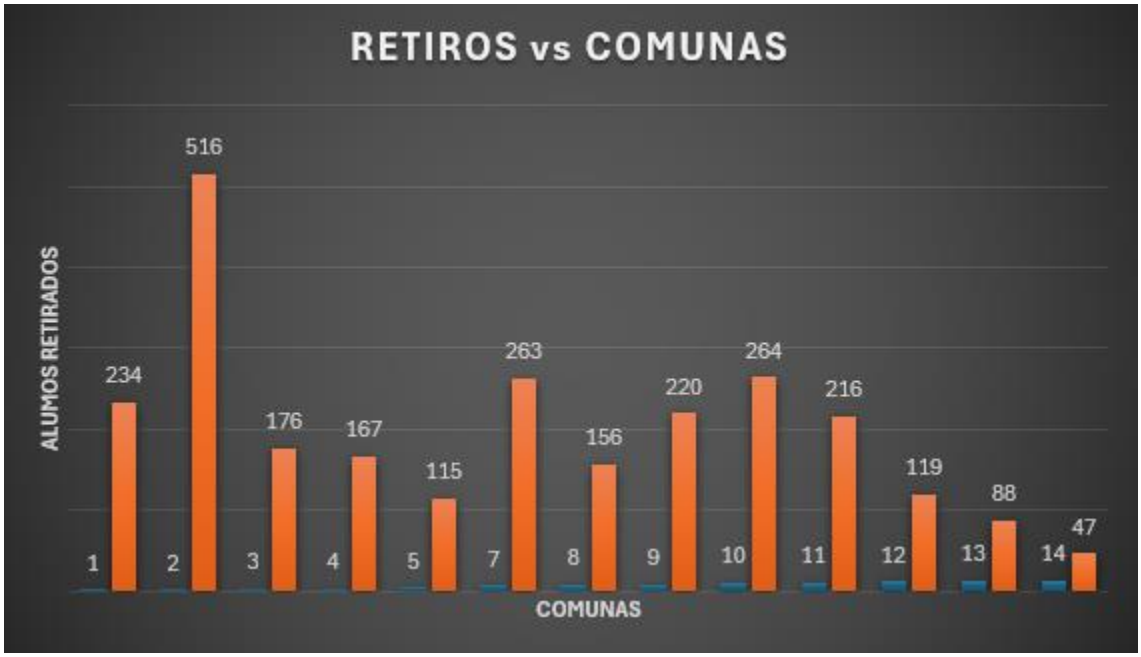


Gráfico 5. Retiros por Comuna

En cuanto al **gráfico 6** (Ver gráfico 6. Retiros por institución educativa) se puede establecer que la institución educativa con mayor cantidad de retiros es Pablo Sexto la cual está ubicada en la comuna 7.

Gráfico 6. Retiros por institución educativa

Conclusiones.

Con la evolución de los sistemas informáticos y la gran cantidad de información que manejan son de vital importancia el big data y la ciencia de datos; teniendo en cuenta, que los datos son parte vital del mundo moderno. De igual manera, con el análisis de datos se obtienen tendencias, preferencias, estadísticas, entre otras que ayudan a las compañías en la toma de decisiones y el mejoramiento continuo de los procesos.

Con base en lo anterior y teniendo en cuenta el caso tratado en el presente informe, podemos evidenciar que no es relevante la ubicación geográfica de las instituciones educativas como motivo para el retiro y posterior deserción de los alumnos del sistema educativo en el municipio de Dosquebradas.

Lista de referencias

Alcalde, I. (2015). *Visualización de la información: de los datos al conocimiento:* (ed.). Editorial UOC. https://elibro.net/es/lc/remington/titulos/57832

Caballero, R. & Martín, E. (2015). *Las bases de Big Data:* (1 ed.). Los libros de la Catarata. https://elibro.net/es/lc/remington/titulos/234185

García González, C. E. González González, L. M. & Rodríguez Morffi, A. (2017). *Modelación de datos: un enfoque sistémico:* (ed.). Editorial Universo Sur. https://elibro.net/es/lc/remington/titulos/120860

Garriga Trillo, A. J. (2009). *Introducción al análisis de datos:* (ed.). UNED - Universidad Nacional de Educación a Distancia. https://elibro.net/es/lc/remington/titulos/48460

López Porrero, B. E. (2009). *Limpieza de datos:* (ed.). Editorial Feijóo. https://elibro.net/es/lc/remington/titulos/71744

Maldonado, S. (2022). Analytics y Big Data: ciencia de los Datos aplicada al mundo de los negocios: (1 ed.). RIL editores. https://elibro.net/es/lc/remington/titulos/225562

Menoyo Ros, D. García López, E. & García Cabot, A. (2021). *Fundamentos de la ciencia de datos:* (ed.). Editorial Universidad de Alcalá. https://elibro.net/es/lc/remington/titulos/177631

Ministerio de Educación Nacional, 07 de febrero de 2017 Deserción Escolar, https://www.mineducacion.gov.co/portal/secciones/Glosario/82745:DESERCION-ESCOLAR

Pulido Romero, E. Escobar Domínguez, Ó. & Núñez Pérez, J. Á. (2019). *Base de datos:* (ed.). Grupo Editorial Patria. https://elibro.net/es/lc/remington/titulos/121283