Modelo de detección de anomalías en información financiera de seguros

**Logotipo

Descripción generada automáticamente**

**Título del trabajo**

María Camila Duarte Foronda

Cristian Joel Lozano Duran

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor  
Nombres completos, Título académico más alto

Universidad de Antioquia  
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2024

|  |  |
| --- | --- |
| **Cita** | ( Duarte Foronda& Lozano Duran, 2023) |
| **Referencia**  **Estilo APA 7 (2020)** | Duarte Foronda, M. C., & Lozano Duran, C. J. (2023). *Modelo de detección de anomalías en información financiera de seguros*  ]. Universidad de Antioquia, Medellín, Colombia. |

**** 

Especialización en Analítica y Ciencia de Datos, CohorteIV.

Centro de Investigación Ambientales y de Ingeniería (CIA).

|  |  |
| --- | --- |
|  | Diagrama  Descripción generada automáticamente con confianza media |

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

Modelo de detección de anomalías en información financiera de seguros

**Dedicatoria**

Este trabajo de grado es el resultado de un viaje de aprendizaje, esfuerzo y crecimiento personal que no hubiera sido posible sin el apoyo incondicional de muchas personas a lo largo del camino.

Alas familias, ya que con su amor, paciencia y constante aliento han sido el motor que nos ha impulsado a superar cada obstáculo. Gracias por creer en nosotros.

A los profesores, quienes no solo han compartido su vasto conocimiento y experiencia, sino que también han inculcado la pasión por el aprendizaje en analítica. Su dedicación y compromiso con la educación han sido fundamentales en la realización de este trabajo.

A los compañeros de clase, con quienes hemos compartido innumerables horas de estudio, discusiones y risas.

A todos ustedes, gracias por ser parte de esta travesía y por contribuir de manera significativa a la culminación de este importante logro.

**Agradecimientos**

En primer lugar, queremos expresar nuestro más profundo agradecimiento a las familias. A los padres, por su amor incondicional, su apoyo constante y por ser la mayor fuente de inspiración.

A mis profesores y mentores, quienes han sido guías esenciales en mi formación académica

A los amigos, por su comprensión y apoyo durante los momentos más difíciles, y por brindar momentos de esparcimiento necesarios para mantener el equilibrio.

**Tabla de contenido**

[Resumen 9](#_heading=h.3rdcrjn)

[Abstract 10](#_heading=h.26in1rg)

[1.](#_heading=h.35nkun2) Descripción del problema 11

[1.1.](#_heading=h.1ksv4uv) Problema de negocio 11

[1.2.](#_heading=h.44sinio) Aproximación desde la analítica de datos 11

[1.3.](#_heading=h.z337ya) Origen de los datos 11

[1.4.](#_heading=h.3j2qqm3) Métricas de desempeño 11

[2.](#_heading=h.4i7ojhp) Objetivos 12

[2.1.](#_heading=h.2xcytpi) Objetivo general 12

[2.2.](#_heading=h.3whwml4) Objetivos específicos 12

[3.](#_heading=h.2bn6wsx) Datos 13

[3.1.](#_heading=h.qsh70q) Datos originales 13

[3.2.](#_heading=h.3as4poj) Datasets 13

[3.3.](#_heading=h.1pxezwc) Analítica descriptiva 13

[Referencias 22](#_heading=h.1jlao46)

[Anexos 23](#_heading=h.43ky6rz)

[Anexo 1. Autoarchivo en Repositorio y documentos de interés 24](#_heading=h.xvir7l)

[Anexo 2. Gestor de citas y referencias de Microsoft Word Microsoft Word - Wikipedia 25](#_heading=h.1baon6m)

[Anexo 3. Citas y referencias de material legal (leyes, decretos, sentencias, etc.) 27](#_heading=h.3vac5uf)

[Anexo 4. Ortografía y gramática 30](#_heading=h.2afmg28)

[Anexo 5. Buscar, reemplazar y eliminar espacios (o palabras) 32](#_heading=h.pkwqa1)

[Anexo 6. Atajos de teclado útiles en Microsoft Word 33](#_heading=h.39kk8xu)

[Anexo 7. Sinónimos y antónimos 34](#_heading=h.1opuj5n)

[Anexo 8. Copiar y pegar sin formato 35](#_heading=h.48pi1tg)

[Anexo 9. Comparar dos documentos 36](#_heading=h.2nusc19)

[Anexo 10. Control de cambios 37](#_heading=h.1302m92)

[Anexo 11. Insertar salto de página 39](#_heading=h.3mzq4wv)

[Anexo 12. Recortar y abreviar direcciones web largas 40](#_heading=h.2250f4o)

**Lista de tablas**

[**Tabla 1**  Resultados del test PBQ-SF (Personality Belief Questionnaire Short Form) 19](#_heading=h.2u6wntf)

[**Tabla 2** Características demográficas y tipo de tratamiento de hemodiálisis y diálisis peritoneal con la adherencia (SMAQ) 20](#_heading=h.19c6y18)

[**Tabla 3** Categorías de la investigación 21](#_heading=h.3tbugp1)

**Lista de figuras**

[**Figura 1** Portada Normas APA séptima edición 2020 en inglés 22](#_heading=h.2r0uhxc)

[**Figura 2** Logo Universidad de Antioquia 22](#_heading=h.1664s55)

**Siglas, acrónimos y abreviaturas**

**APA** American Psychological Association

**Cms.** Centímetros

**ERIC** Education Resources Information Center

**Esp.** Especialista

**MP** Magistrado Ponente

**MSc** Magister Scientiae

**Párr.** Párrafo

**PhD** Philosophiae Doctor

**PBQ-SF** Personality Belief Questionnaire Short Form

**PostDoc** PostDoctor

**UdeA** Universidad de Antioquia

# Resumen

Una compañía de seguros necesita desarrollar un modelo para identificar transacciones inusuales en la generación de recibos 'manuales' de producción, cumpliendo con la norma de auditoría, circular externa 008 de 2023. Se analizarán datos de recibos generados, identificando patrones de comportamiento. Los datos provienen de registros internos del último año. Las métricas de desempeño incluirán homogeneidad, completitud, precisión y la proporción de transacciones inusuales identificadas.

El problema es que la compañía no puede monitorear todos los recibos manuales debido a su volumen anual. Se solicitó controlar solo los recibos más inusuales, pero no se puede identificar cuáles son para investigarlos.

La solución propuesta es un modelo no supervisado de clustering, específicamente DBSCAN, para identificar ruido y agrupar datos, permitiendo encontrar etiquetas para un modelo supervisado que clasifique recibos usuales e inusuales.

Incluye al final de dicha página la dirección de los repositorios GitHub.

1. *Palabras clave*: Transacciones Inusuales, Recibos Manuales, Analítica de Datos, Clustering DBSCAN, Cumplimiento.

# Abstract

An insurance company needs to develop a model to identify unusual transactions in the generation of 'manual' production receipts, complying with the audit standard, External Circular 008 of 2023. Data from generated receipts will be analyzed to identify behavior patterns. The data comes from internal records from the past year. Performance metrics will include homogeneity, completeness, accuracy, and the proportion of identified unusual transactions.

The problem is that the company cannot monitor all manual receipts due to their annual volume. They have been asked to control only the most unusual receipts, but they cannot identify which ones to investigate.

The proposed solution is an unsupervised clustering model, specifically DBSCAN, to identify noise and cluster data, allowing for labels that can be used in a supervised model to classify usual and unusual receipts.

1. *Keywords***:** Unusual Transactions, Manual Receipts, Data Analytics, DBSCAN Clustering, Compliance.

# Descripción del problema

## Problema de negocio

En una compañía de seguros, hay una opción a nivel de aplicación que permite a los usuarios que tienen acceso, realizar recibos de cobro cuyo valor se digita de forma manual, el área de control financiero debe velar por la integridad y exactitud de las cuentas de ingreso. Por tal razón, debe realizar un monitoreo y diseñar un control para la ejecución de recibos manuales. Pero el área de negocio responsable afirma, no poder monitorear todos los recibos manuales, debido a la cantidad de recibos generados en el año; Se solicitó al área responsable solo realizar control sobre los recibos más inusuales. Pero tampoco identifican de la cantidad de recibos manuales generados, cuáles podrían ser inusuales, para investigar la justificación de estas.

## Aproximación desde la analítica de datos

La idea es desarrollar un modelo no supervisado, de clustering, específicamente DBSCAN, para identificar el ruido en los datos y una agrupación en datos que posibilite encontrar las etiquetas que posteriormente, se puedan ingresar en un modelo supervisado de clasificación en usuales e inusuales.

## Origen de los datos

Contamos con el historial de recibos manuales realizados en 2023, de acuerdo con la regional, tipo de operación, fecha de ejecución, entre otros. Estos datos son extraídos del ERP SAP de la organización.

## Métricas de desempeño

**Métricas de Machine Learning:**

* Coeficiente de Silueta: Mide qué tan bien se agrupan los puntos en sus clusters. Mayor que 0.5 indica una buena agrupación.
* Índice de Davies-Bouldin: Mide la separación y compactación de los clusters. menor que 1 indica clusters bien formados.
* Ruido Identificado: Proporción de puntos identificados como ruido. La métrica depende de los datos y la evolución del modelo. Por ahora 10%.

**Métricas de Negocio:**

1. Tasa de Falsos Positivos y Negativos: Proporción de recibos usuales clasificados incorrectamente como inusuales y viceversa. menor o igual al 5%
2. Consistencia de anomalías: Validación de los recibos inusuales identificados mediante auditorías. Mayor o igual al ≥ 80%
3. Reducción en la calificación del riesgo en la cuenta inherente y residual. Se espera una reducción de al menos una categoría en la escala de evaluación de riesgos.

# Objetivos

## Objetivo general

Desarrollar un modelo analítico, que utilice técnicas de aprendizaje automático no supervisado en la identificación de transacciones inusuales de ‘recibos manuales’ en una compañía de seguros. Para mejorar el cumplimiento de la norma de auditoría y garantizar la integridad y exactitud de las cuentas de ingreso.

## Objetivos específicos

* Exploratorio los registros de recibos manuales, Identificando tendencias y patrones e indicios de información inusual en los recibos generados.
* Seleccionar y aplicar un modelo no supervisado adecuado para la identificación de etiquetas de transacciones inusuales.
* Entrenar el modelo seleccionado utilizando los datos preprocesados.
* Evaluar el rendimiento del modelo mediante el uso de métricas pertinentes como homogeneidad, completitud, precisión y tasa de identificación de transacciones inusuales.
* Comparar las transacciones inusuales identificadas por el modelo con los resultados de auditoría para validar su efectividad y precisión.
* Definir y Documentar las etiquetas de clasificación de información inusual
* Realizar un análisis de costos y beneficios para evaluar el impacto del proyecto.
* Desplegar el modelo.

# Datos

## Datos originales

Los datos se encuentran en el sistema interno de información de la compañía el cual se maneja por medio de SAP, la información es cargada de forma manual por los operadores y se pueden descargar si se tienen las credenciales de acceso al sistema. Este Dataset contiene los manuales tanto notas débito como crédito que se realizaron en 2023 y las columnas se refieren a cada una de las siguientes variables.

**Numero\_Poliza**: Es el núemro de poliza al cual corresponde el recibo manual.

**Codigo\_Ramo\_Op**: Es le código de ramo al cual pertenece la poliza (producto)

**Num\_Recibo**: número asignado al recibo manual.

**Fecha\_Emision**: fecha de emisión del recibo manual.

**Fecha\_Movimiento**: Fecha en la cual se realizó el movimiento.

**Fecha de cobro**: fecha en la cual se hará efectivo el cobro.

**Cod\_Sucursal**: Código de la sucursual donde se realizó el recibo.

**Cod\_Agente**: Código del agente que solicitó el recibo.

**Estado\_Contrato\_Cd**: Número de asignación al estado del contrato.

**Estado\_Contrato\_Desc**: Estado en el cual se encuentra la poliza.

**Nombre\_Regional**: Nombre de la regional.

**Nombre\_Canal\_Comercial:** Canal comercial que solicitó el recibo.

**estado\_rec**: Estado del recibo.

**Tipo\_Oper\_Pago\_Cd**: Número asigando por Sap a la operación

**Desc\_Causa\_Estado** Justificación del recibo manual.

**Valor\_Cobro**: Valor del recibo del cobro.

**Desc\_Operacion\_Pago**: Descripción de la operación.

El Dataframe cuenta con 36098 filas y 17 columnas para un total de 613666 datos y se puede descargar de SAP en formato CSV, dado que cuenta con información sensible para la compañía no se encuentra en ningún repositorio público y su acceso es limitado.

## Datsets

La preparación de los datasets se inicia eliminando algunas columnas que no aportan información relevante para el modelo que se desea construir, dichas columnas son:

Fecha\_Movimiento: Dado que es un proceso automático realizado por un tercero, no tiene correlación alguna con el proceso de las transacciones.

Num\_Recibo: Este es un código secuencial que se asigna al entrar al sistema, el mismo no tiene ningúna influencia en el proceso de las transacciones.

Cod\_Sucursal: este campo puede ser de importancia para el caso de estudio, pero dada la gran cantidad de sucursales se hace necesario agrupar las categorías de algún modo, dicha agrupación puede ser por región y dado que el campo “Nombre\_Regional” ya existe, sería redundante trabajar la sucursal.

Cod\_Agente: al igual que el caso anterior, las categorías de esta columna se pueden agrupar en “Nombre\_Regional”

Estado\_Contrato\_Cd: dado que se trabará con el estado del contrato, no hace ncesario trabajar con el codigo del mismo

Nombre\_Canal\_Comercial: el nombre del canal comercial está relacionado con la forma en la que se vendió la póliza y dado que esta actividad no se realiza por el mismo equipo de transacciones, no es relevante para el caso de estudio.

estado\_rec: todos los recibos en esta base de datos deben estar legalizados o proximos a legalizar, por lo que esta columna no aporta información al modelo.

Tipo\_Oper\_Pago\_Cd: esté es un número asignado por SAP a la operación y no tiene relevancia para el modelo.

Numero\_poliza: dado que este es un número secuencial asignado al momento de vender la póliza este no debería tener relación con la transacción en el modelo que se plantea, sin embargo, esta información puede ser utilizada en un futuro con un modelo paralelo que revise la frecuencia con la que una póliza genera cobros.

Fecha\_cobro: está fecha depende de un tercero y no tiene que ver con la actividad de la compañía, no se tendrá en cuenta en el modelo de manera inicial.

seleccionadas las columnas con las que se va a trabajar, se procede a convertir en string las columnas categóricas y en float la columna numérica, posteriormente se procede a agrupar las categorías de la columna ‘Codigo\_Ramo\_Op’ así:

Bienes muebles e inmuebles:

007, aeronaves casco

009, equipo electrónico

015, sustracción establecimientos

019, rotura maquinaria

028, hogar

030, incendio

032, maquinaria y equipo

040, automoviles

039, obras civiles

Gestión empresarial o bancaria

011, manejo

012, cumplimiento

013, responsabilidad civil

017, manejo bancario

020, transporte de mercancias

021, navegación

024, transporte de valores

033, montaje todo riesgo

034, construcción

BAN, bancarios

181, vida

Persona natural individual/colectivo

081, vida individual

083, vida de grupo

084, accidentes personales

085, juvenil

086, exequial

090, salud familiar

091, salud colectiva

092, seguros previsionales

095, seguro colectivo de pensiones.

113, plan complementario familiar

114, plan complementario colectivo

Así mismo en la columna ‘DESC\_CAUSA\_ESTADO’ se hacen las siguientes agrupaciones.

renovación: renovado + prorroga

nuevo/rehabilitación: nuevo + rehabilitación

error: por error en expedición + convertido + por error de trámite + sin información

cancelación: todas las demás

Por último de la columna ‘Fecha\_Emisión’ se tendrá únicamente en cuenta el mes como categoría, para extraer dicha información se usaron los comandos de la librería pandas to\_datetime y month\_name.

Cabe recalcar que dado que se usará un modelo no supervisado no dividieron los datos en entrenamiento y prueba.

## Analítica descriptiva

Con el fin de analizar la distribución de los datos se utilizan las librerías Matplotlib y pandas para graficar la frecuencia de cada una de las variables categóricas con las que trabará el modelo.

# Gráfico, Gráfico en cascada Descripción generada automáticamente

Ilustración 1. frecuencia de instancias para variables categóricas

En las gráficas se puede observar que la distribución de las categorías del ‘mes de emisión’, el ‘estado del contrato’ es bastante uniforme, para el ‘nombre regional’ la distribución es uniforme también con la clara excepción de la categoría ‘Antioquia” lo que demuestra que la compañía tiene gran parte de la cuota de mercado del departamento de Antioquia, por otra parte las variables ‘cod\_ramo\_operación’, ‘Desc\_causa\_estado’ y ‘Desc\_operación\_pago’ están claramente desbalanceadas, para la primera variable mencionada se puede ver que los seguros dirigidos a personas o grupos de personas tienen muchas más pólizas que los dirigidos a empresas, algo que se puede considerar normal el mercado de personas naturales es mucho más grande (en cantidad de potenciales clientes) que el de personas jurídicas. En la descripción del estado, podemos ver como la mayoría de los recibos manuales se hacen por renovación y muy pocos se deben a errores o ingreso de nuevos clientes, esto se puede deber a la naturaleza misma del proceso del recibo manual dentro de la compañía.

Análisis de Datos faltantes:

Se reviso la porción datos faltantes por columna. Se puede observar en la siguiente tabla.Interfaz de usuario gráfica, Tabla

Descripción generada automáticamente

Ilustración 2 Razón de datos faltantes por columna

Se realizó un análisis de los datos faltantes pro cada columna y en las más relevantes se evidenciaron las columnas, Fecha\_ cobro, Fecha\_ movimiento, Desc\_Causa\_ Estado y Desc\_Operación\_ Pago. La mayor porción de faltantes la tiene la columna de fecha de movimiento con el 94 % de faltantes seguido por fecha de cobro con el 13%. Para estas dos columnas se decidió, eliminarlas del modelo ya que la columna vital para la identificación de transacciones inusuales en la generación del recibo es la fecha de movimiento, que es la que contiene la fecha y hora de realización de estos. Pero las columnas de descripción de la causa y de la operación son relevantes en el modelo.

Una de las hipótesis que teníamos era que de pronto los datos faltantes se debieran a una fecha en específico, para esto creamos una tabla en la que pudiéramos visualizar si los faltantes se debían a algún periodo en específico. En este caso, los periodos se definieron como agrupaciones mensuales de la columna Fecha\_ Movimiento, respecto a la variable que consideramos una de las más relevantes ‘Desc\_Operación\_Pago’:

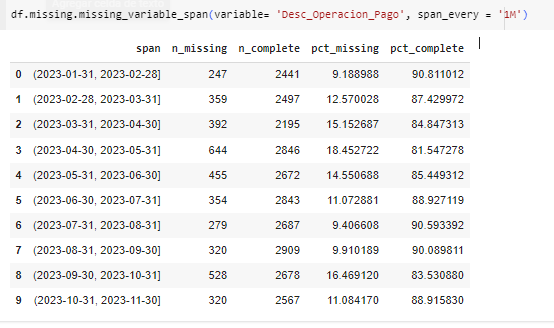


Ilustración 3:Datos faltantes por periodo de la columna 'Desc\_Operación\_Pago'

Como se Observa en la tabla anterior los datos faltantes no se deben a ningún periodo en especial, manejan cierta frecuencia promedio en todos los periodos.

Otra de las premisas que tenemos es que la información de la columna Desc\_Operación\_Pago' sé podría complementar o incluso reemplazar con la de la columna ‘Desc\_causa\_estado’Por esta razón queríamos identificar si hay información faltante en ambas columnas.

Gráfico

Descripción generada automáticamente

Ilustración 4 Datos faltantes comunes en las columnas de descripción de causa y operación

Como se Observa en la tabla anterior cero es el numero de filas en las cuales faltan datos tanto en la columna descripción de la causa como en la columna descripción de la operación.

Por último, queríamos observar cómo era la distribución e faltantes en la tabla inicial:

Imagen que contiene objeto, peine, cerca, cruce de peatones

Descripción generada automáticamente

Ilustración: Distribución de faltantes en el Dataset

Como se Observa en el esquema anterior, en la columna ‘Span’ podemos visualizar que en total hay 30233 filas que contienen la información completa y representan más del 80% d ellos datos.

Reducción de dimensionalidad y análisis previo de columnas.

Para la columna "Codigo\_Ramo\_Op" Agrupamos las categorías de la siguiente manera Nota: Nota: El texto al que refiere cada código se tomó de SAP.

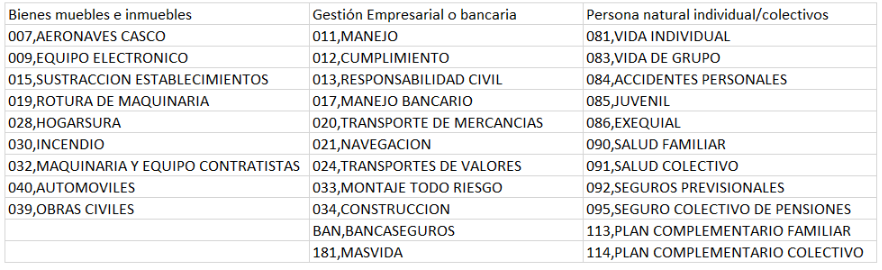


Ilustración 5: Reducción dimensionalidad de producto

Limpieza columna Des\_Causa\_estado:

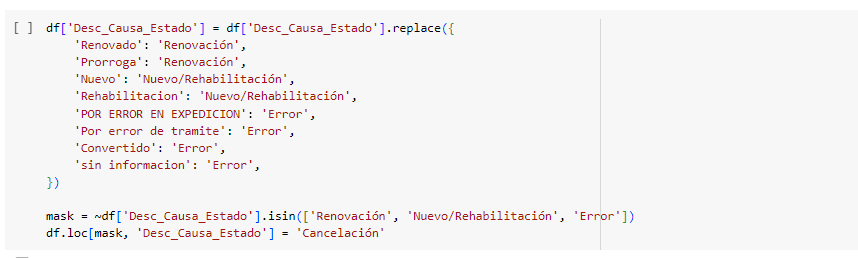


Ilustración 6: Limpieza columna descripción de la causa

.

Definición y Análisis de las variables categóricas para el modelo:

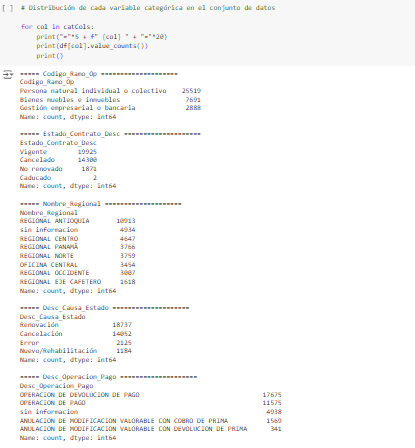


Ilustración 7: distribuciones variables categóricas

Aplicación de One Hot variables categóricas:

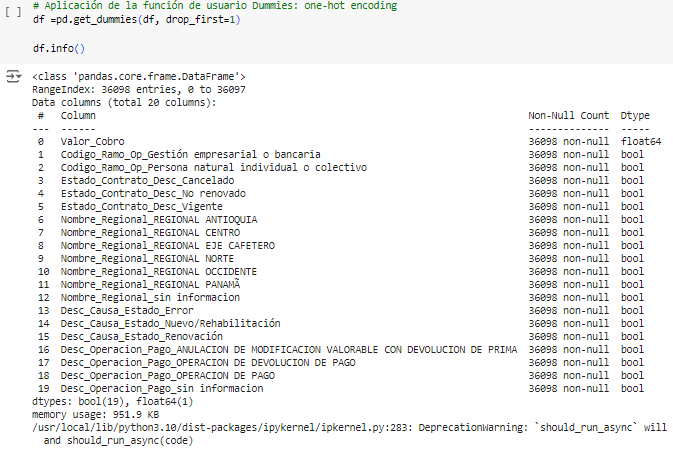


Ilustración 8: Aplicación One hot variables categóricas

**Figura 2** *Logo Universidad de Antioquia*

****

*Nota.* Fuente http:/www.udea.edu.co

# Conclusiones

Son las interpretaciones finales que recopilan los datos de la investigación, describe lo que se obtuvo, qué se logró y cuáles son los resultados. Guardan relación directa con lo que se mencionó en el planteamiento del problema y en los objetivos. Haz un breve juicio crítico de tu proyecto.

# Recomendaciones

Las recomendaciones son las futuras y posibles líneas de investigación que llevarán a resolver problemas relacionados con la presente investigación.

# Referencias