Estadística III para Ingenieros de Sistemas

Jose Daniel Ramirez Soto 2023
jdr2162@columbia.edu

Acerca de mí, y de ustedes…

EGRESADO 2005 ING. SISTEMAS

Entré a la universidad muy joven :)

MAESTRÍA EN ING.SISTEMAS Y DATA SCIENCE

12 AÑOS TRABAJANDO EN SOFTWARE Y MODELOS DE INTELIGENCIA ARTIFICIAL.

AHORA EMPRENDEDOR EN AI.

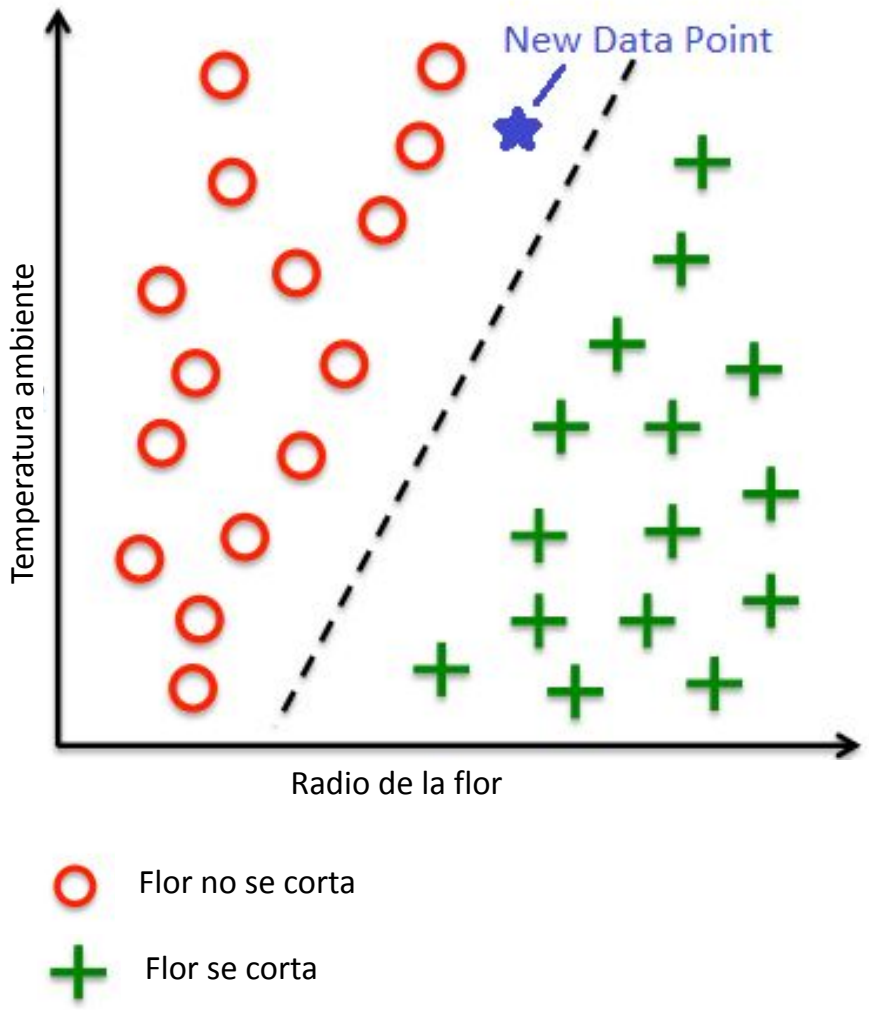
Metodología

- Clase teórica primer bloque, parte práctica segundo bloque.
  - Lenguaje (Python)
- Datasets reales:
  - 2 Proyectos prácticos
  - 1 competencia de kaggle https://www.kaggle.com/
- Comunicación:
  - Slack: chat office hours. estadistica3.slack.com
  - correo: 24 horas
  - github: https://github.com/jdramirez/UCO_ML_AI
  - office hours: google meet

Agenda

- Conceptos básicos de machine learning y big data
- “Big picture” flujo de trabajo para crear un modelo
- Exploración de los datos
- Transformación de datos
- Split and Sampling

QUÉ ES MACHINE LEARNING ?

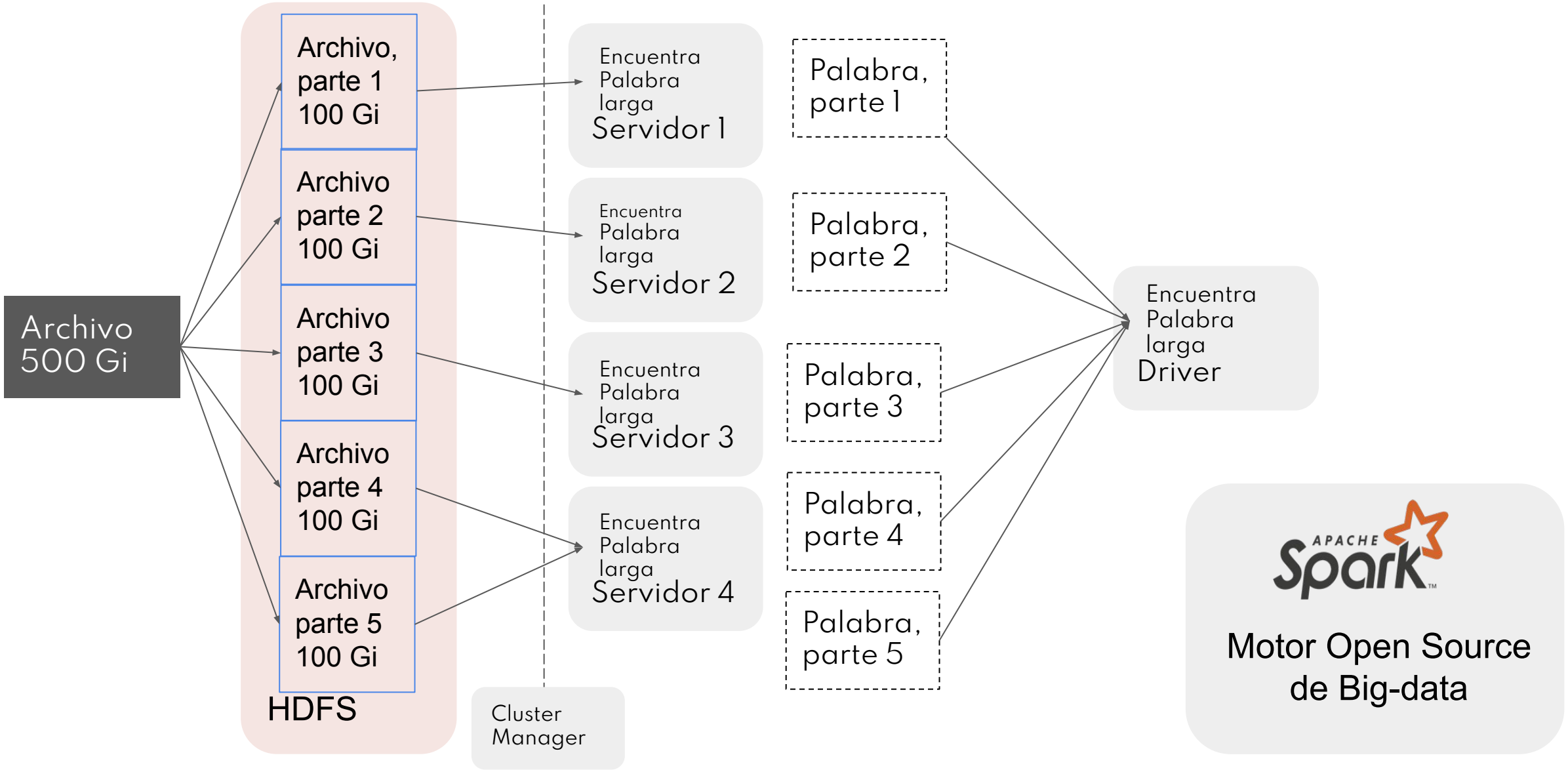


¿QUÉ ES MACHINE LEARNING?

Aprende de los datos sin ser explícitamente programado.

Machine learning (ML) es el estudio de algoritmos que mejoran automáticamente a través de los datos observados. Los modelos de machine learning aprenden de unos datos observados llamados "Training data", lo que le permite al modelo clasificar datos que no ha observado.

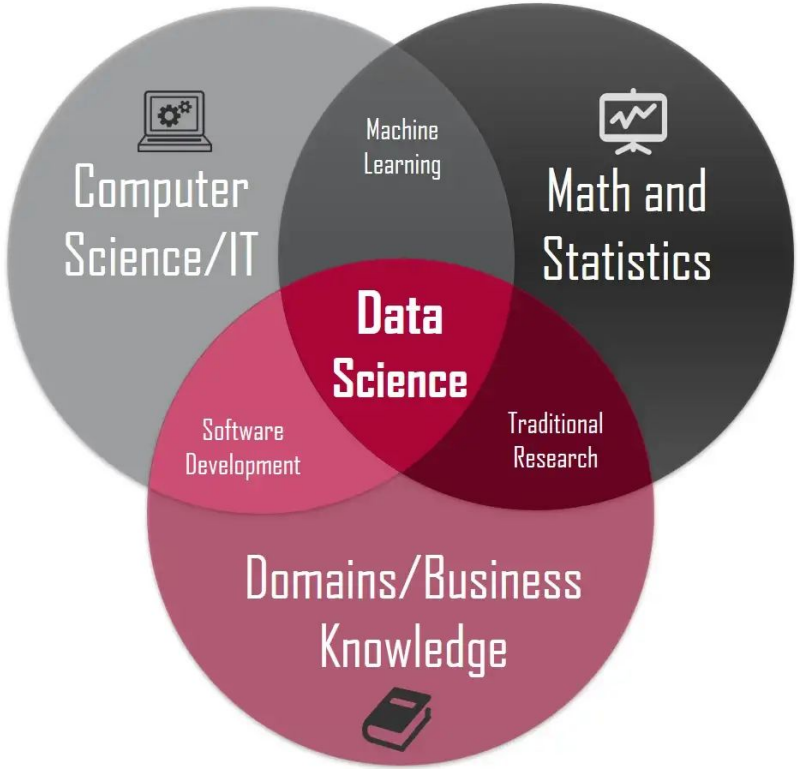
¿Cómo funciona una plataforma de Big Data?

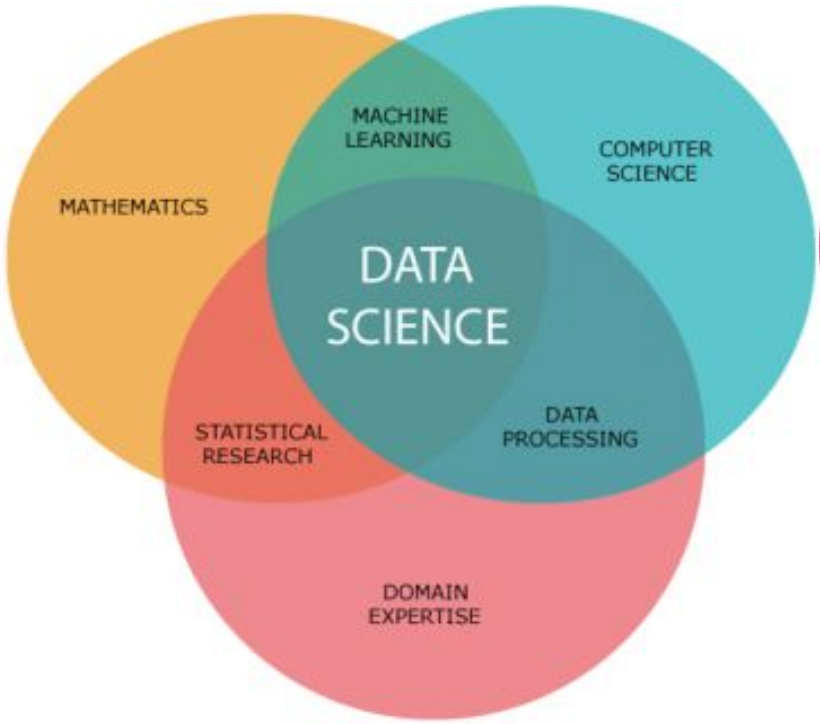


¿Qué significa ser un científico de datos?

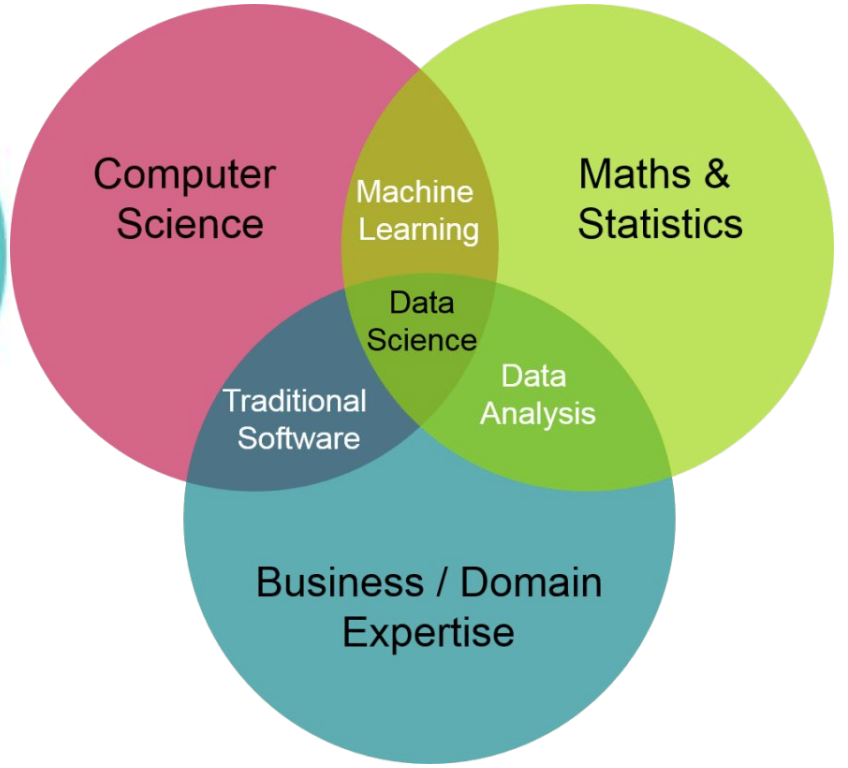
UCO
Universidad Católica de Oriente

CONOCIMIENTO





ROLES



Matemática: Algebra lineal, probabilidad y estadística

https://medium.com/data-science-in-2019/what-is-data-science-87e9dc225cf9
https://dev.to/amananandrai/most-popular-tools-for-data-scientists-in-2020-5eki

¿Preguntas?

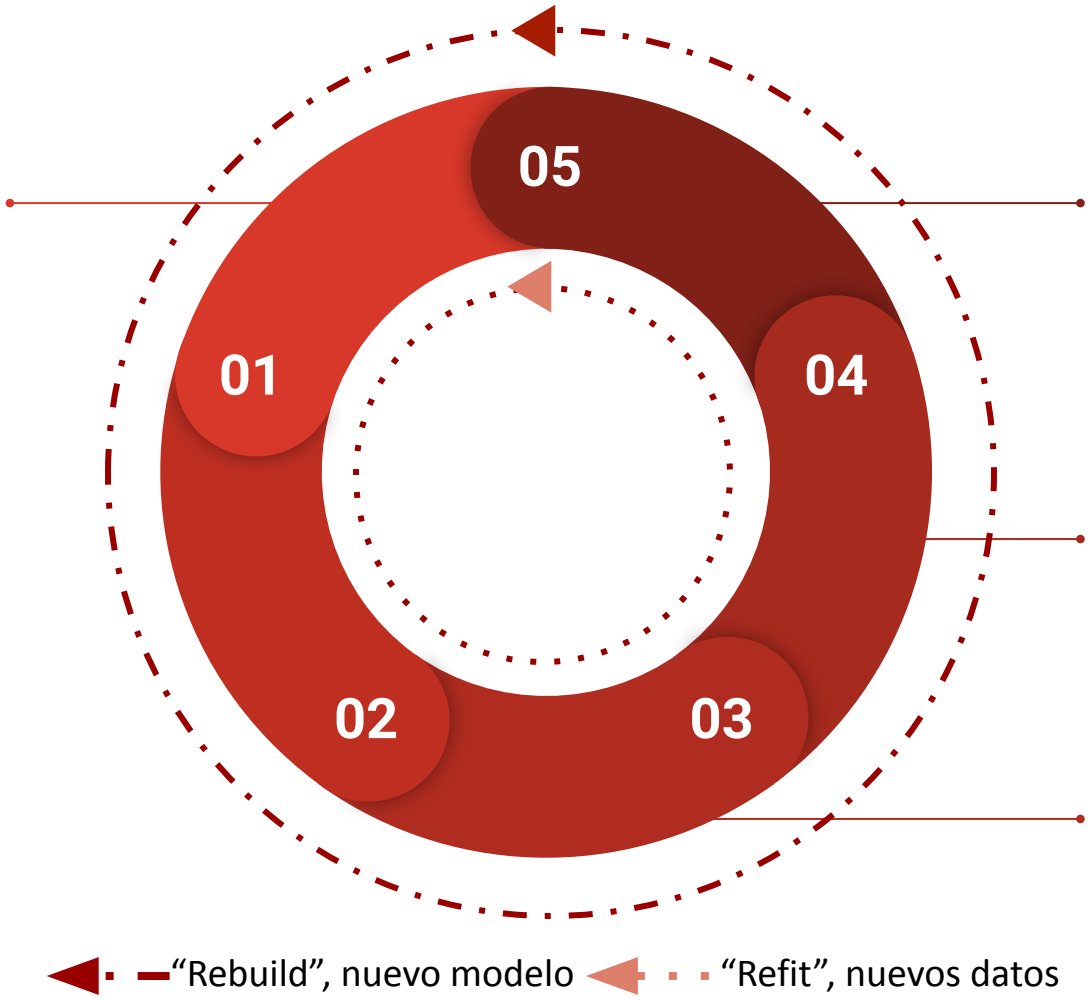
Numero de meses de experiencia del curso Estadística 3?

2,4,5,6,8,10,20,28
¿Cuál es el promedio? 11.3
¿Cuál es la media? 8
¿Cuál es la desviación estándar?
¿Cuál es el valor del primer cuartil?4

Flujos de trabajos para crear un modelo desde el negocio

Proceso de negocio y captura de datos

¿Cómo se utilizará el modelo? ¿"humans in the loop"? ¿Tengo etiquetas? ¿Tengo nuevos datos? ¿El modelo es válido y estable?



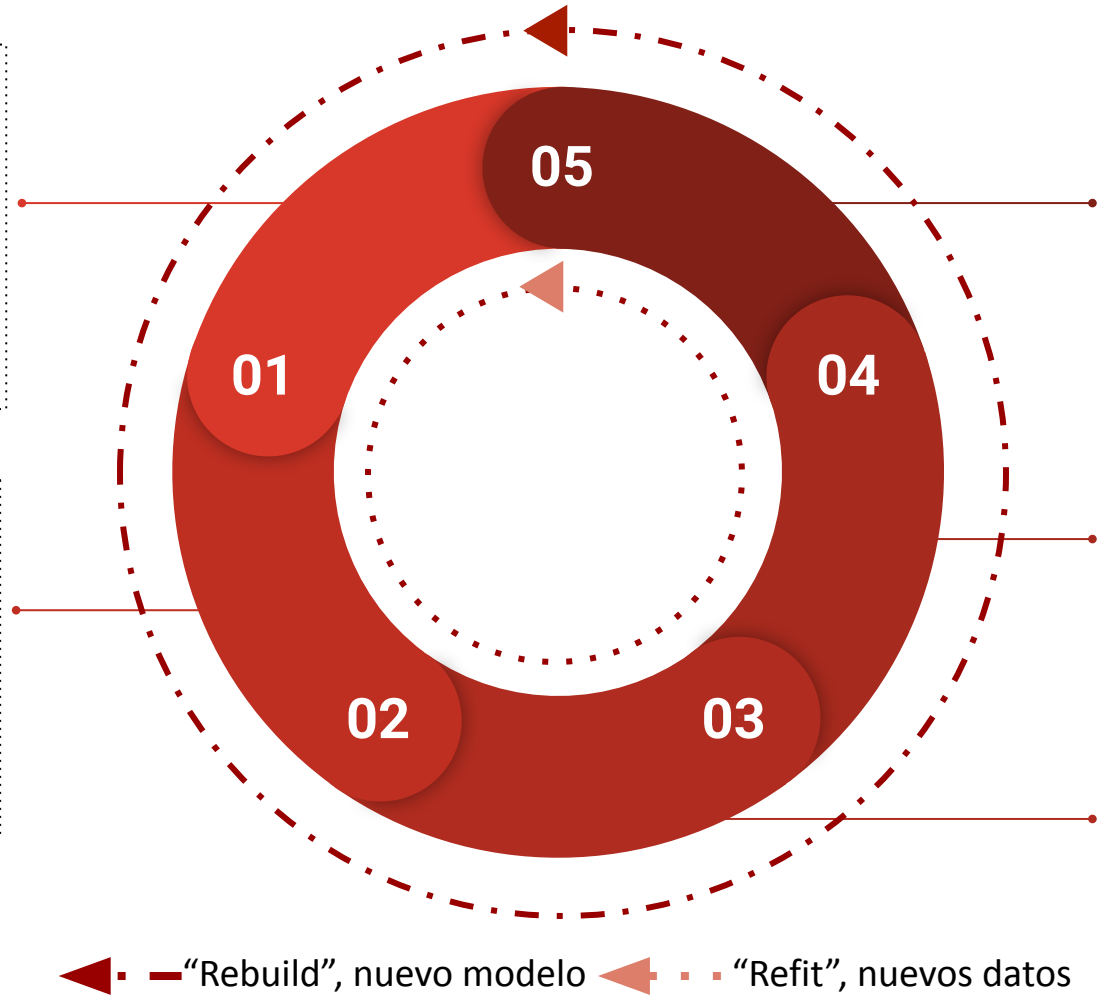
Flujos de trabajos para crear un modelo desde el negocio

Proceso de negocio y captura de datos

¿Cómo se utilizará el modelo? ¿"humans in the loop"? ¿Tengo etiquetas? ¿Tengo nuevos datos? ¿El modelo es válido y estable?

Crear/entrenar modelo y evaluar

Data scientist: crea el modelo, mide el impacto desde el error.
Business: ayuda a evaluar el beneficio desde "$ dollars"



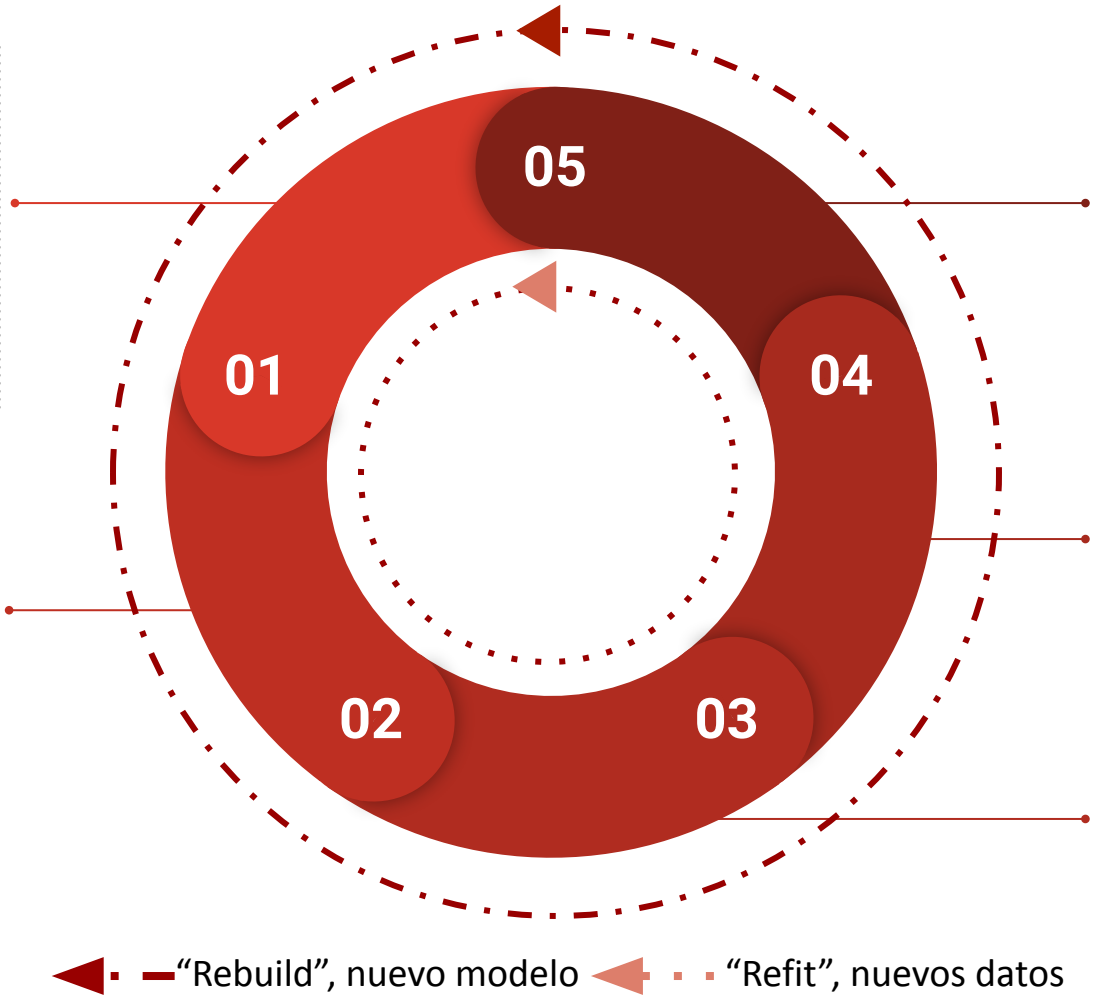
Flujos de trabajos para crear un modelo desde el negocio

Proceso de negocio y captura de datos

¿Cómo se utilizará el modelo? ¿"humans in the loop"? ¿Tengo etiquetas? ¿Tengo nuevos datos? ¿El modelo es válido y estable?

Crear/entrenar modelo y evaluar

Data scientist: crea el modelo, mide el impacto desde el error.
Business: ayuda a evaluar el beneficio desde "$ dollars"



Despliegue en producción

Ingeniero de Datos: incluye el modelo en las aplicaciones utilizadas por el cliente.
Data scientist: válida resultado "esperado" vs "obtenido", "research data" vs "producción"

Flujos de trabajos para crear un modelo desde el negocio

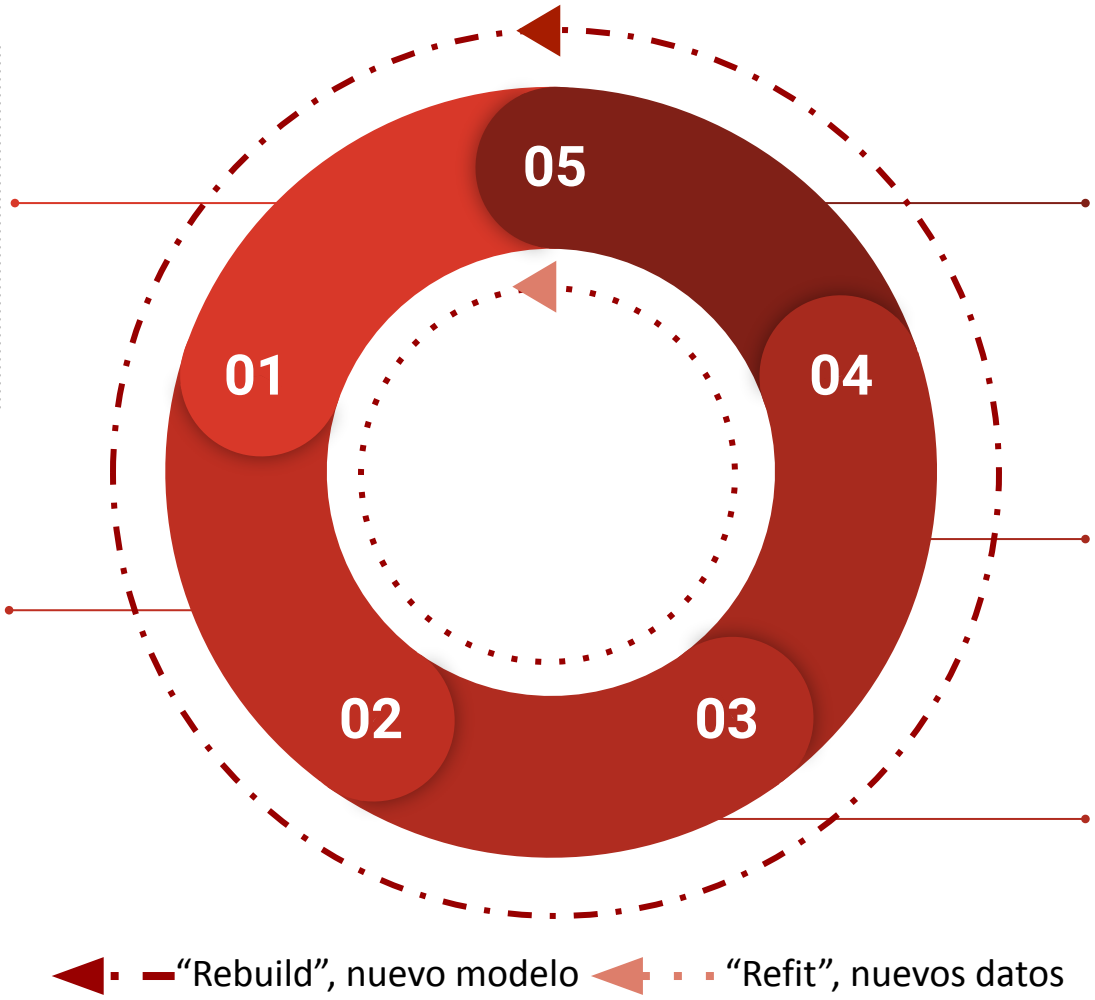
Universidad Católica de Oriente

Proceso de negocio y captura de datos

¿Cómo se utilizará el modelo? ¿"humans in the loop"? ¿Tengo etiquetas? ¿Tengo nuevos datos? ¿El modelo es válido y estable?

Crear/entrenar modelo y evaluar

Data scientist: crea el modelo, mide el impacto desde el error.
Business: ayuda a evaluar el beneficio desde "$ dollars"



Expansión/Comprensión

Business: incluir un nuevo producto en el modelo. O dejar de utilizar el modelo para un producto existente.
Data scientist: predecir el comportamiento del modelo

Despliegue en producción

Ingeniero de Datos: incluye el modelo en las aplicaciones utilizadas por el cliente.
Data scientist: válida resultado "esperado" vs "obtenido", "research data" vs "producción"

Flujos de trabajos para crear un modelo desde el negocio

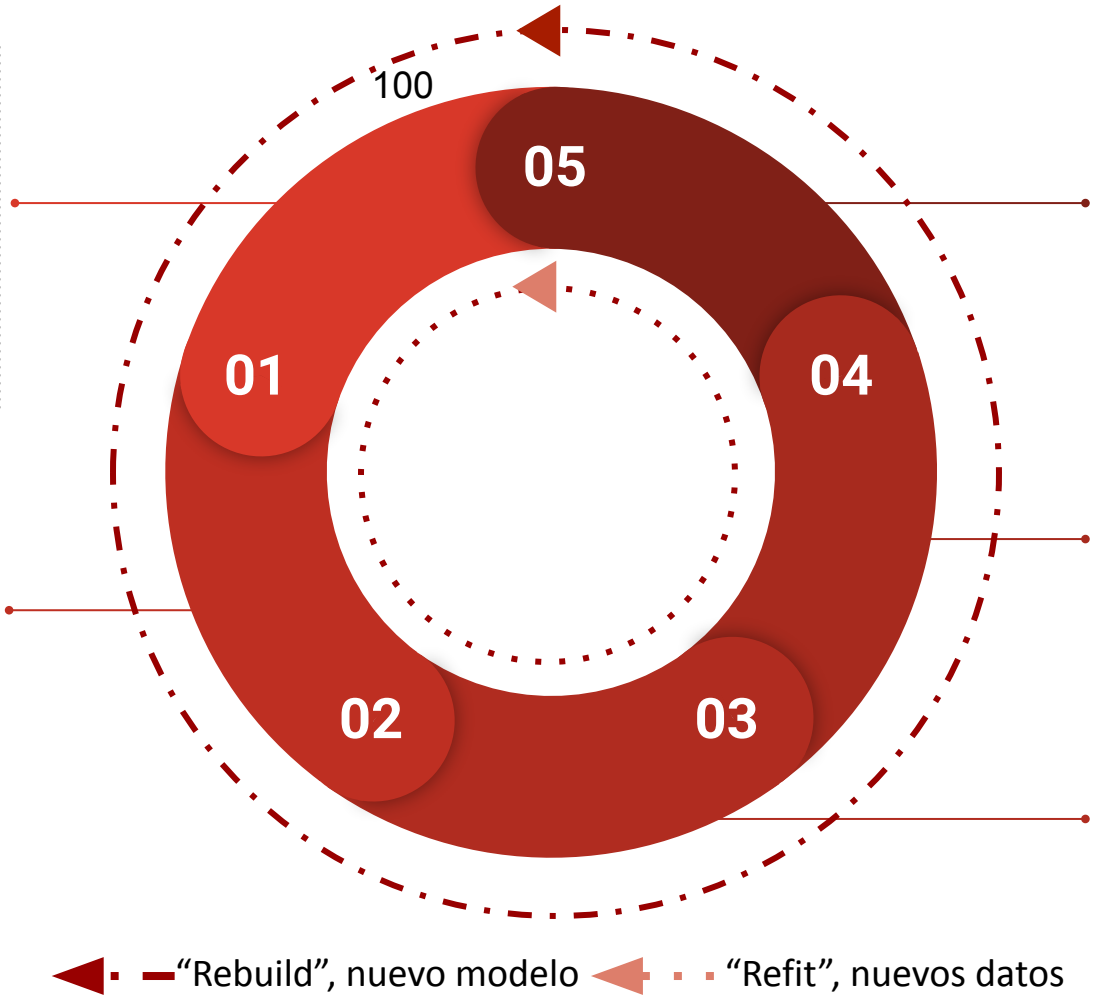
Universidad Católica de Oriente

Proceso de negocio y captura de datos

¿Cómo se utilizará el modelo? ¿"humans in the loop"? ¿Tengo etiquetas? ¿Tengo nuevos datos? ¿El modelo es válido y estable?

Crear/entrenar modelo y evaluar

Data scientist: crea el modelo, mide el impacto desde el error.
Business: ayuda a evaluar el beneficio desde "$ dollars"



Monitorear

Business: desempeño del modelo e impacto ($$).
Data scientist: modelo sea estable y válido

Expansión/Comprensión

Business: incluir un nuevo producto en el modelo. O dejar de utilizar el modelo para un producto existente.
Data scientist: predecir el comportamiento del modelo

Despliegue en producción

Ingeniero de Datos: incluye el modelo en las aplicaciones utilizadas por el cliente.
Data scientist: válida resultado "esperado" vs "obtenido", "research data" vs "producción"

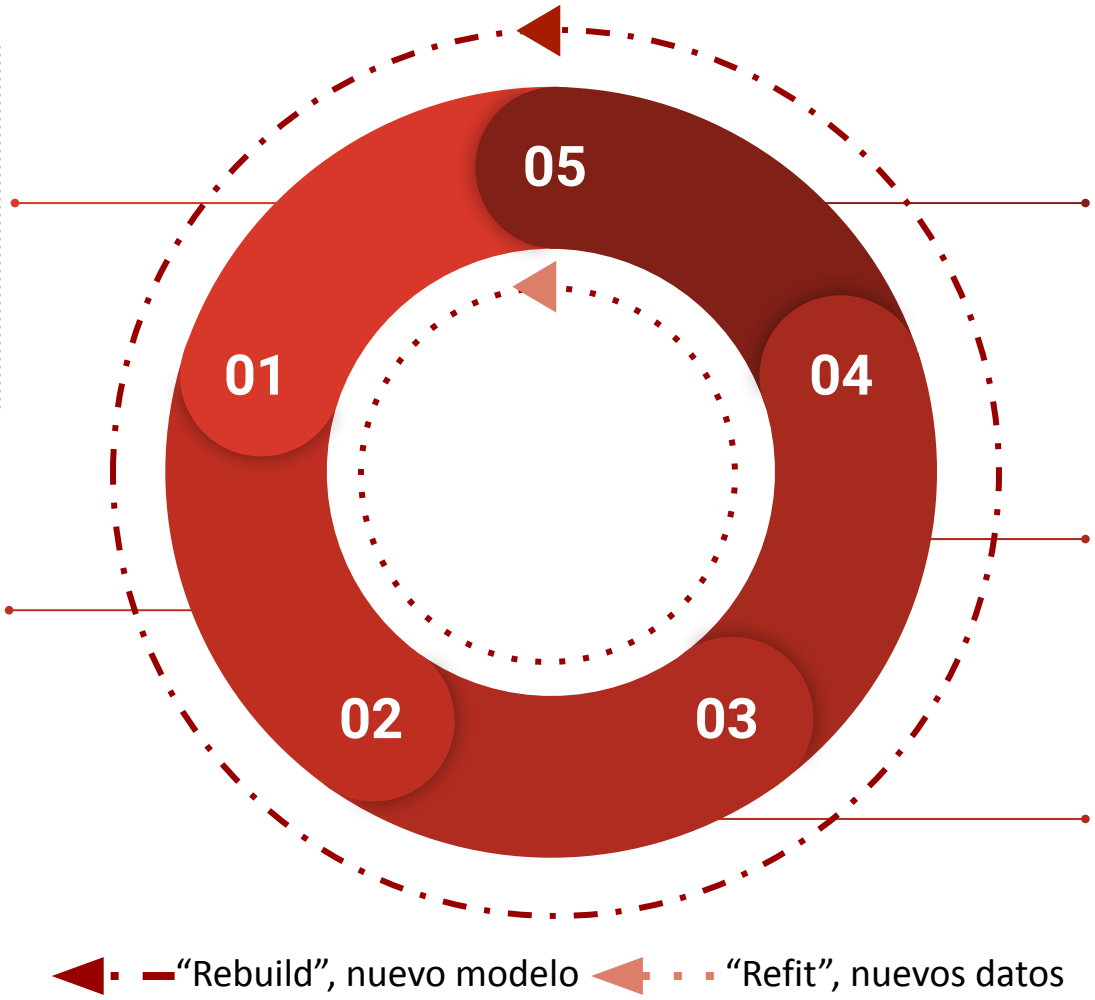
Flujos de trabajos para crear un modelo desde el negocio

Proceso de negocio y captura de datos

¿Cómo se utilizará el modelo? ¿"humans in the loop"? ¿Tengo etiquetas? ¿Tengo nuevos datos? ¿El modelo es válido y estable?

Crear/entrenar modelo y evaluar

Data scientist: crea el modelo, mide el impacto desde el error.
Business: ayuda a evaluar el beneficio desde "$ dollars"



Monitorear

Business: desempeño del modelo e impacto ($$).
Data scientist: modelo sea estable y válido

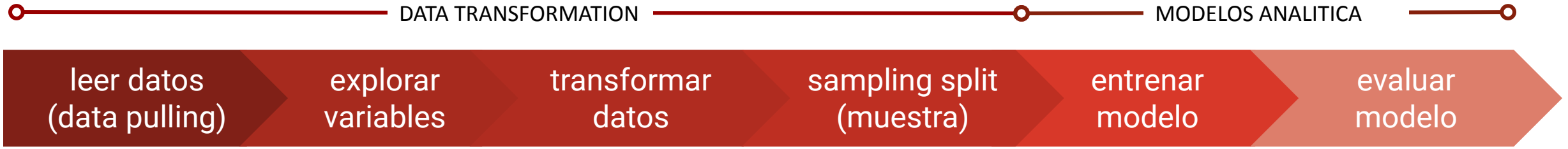
Expansión/Comprensión

Business: incluir un nuevo producto en el modelo. O dejar de utilizar el modelo para un producto existente.
Data scientist: predecir el comportamiento del modelo

Despliegue en producción

Ingeniero de Datos: incluye el modelo en las aplicaciones utilizadas por el cliente.
Data scientist: válida resultado "esperado" vs "obtenido", "research data" vs "producción"

“workflow” Flujos de trabajo para crear un modelo



Caso de uso, detección de fraude en el Sisben

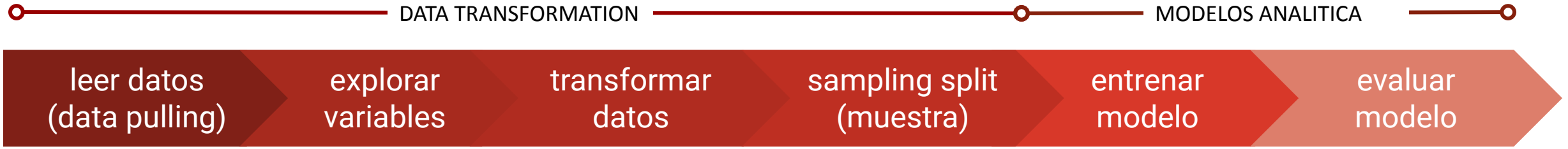
Que es el SISBEN, Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales:

- Sistema de Puntaje entre **0 y 100** para clasificar a la población colombiana tomando como base su condición socio-económica. De los **38’057,617** personas existen **122,401** marcadas como anomalias(**0.3%** de etiquetas).
- Mide la pobreza basado en el índice multidimensional de pobreza. Cuatro categorías principales: salud, educación, vivienda y vulnerabilidad. Contiene datos a nivel de persona, hogar y vivienda.
- Puntajes menores a 60 pueden ser beneficiarios de algunos de los programas del Gobierno Colombiano.
- 3 áreas geográficas: ciudades, zonas urbanas y zonas rurales.

OBJETIVO : Identificar los factores que motivan el comportamiento de las personas registradas en la búsqueda de beneficios mediante la manipulación de la información y predecir si un nuevo formulario tiene probabilidad de fraude.

* https://www.sisben.gov.co/Paginas/landing.aspx

“workflow” Flujos de trabajo para crear un modelo



- leer las fuentes de datos y explorar

leer datos o data pulling

Identificar las fuentes de datos que se incluirán en el proyecto, la granularidad (nivel de persona, municipio, barrio). Considerar:

- porcentaje de equivalencias entre las fuentes
- numero de registros repetidos y manejo.
- formatos de cada variable (numero, texto, fecha).

Ejemplo Sisben

Validar la misma información desde varias fuentes funciona muy bien para la detección de fraude. Por ejemplo, el 88.65% de las personas que tenían vehículo no lo habían reportado en el Sisben. También, utilizando nombre, lugar y fecha se detectaron 300k duplicados

“workflow” Flujos de trabajo para crear un modelo

DATA TRANSFORMATION — MODELOS ANALITICA

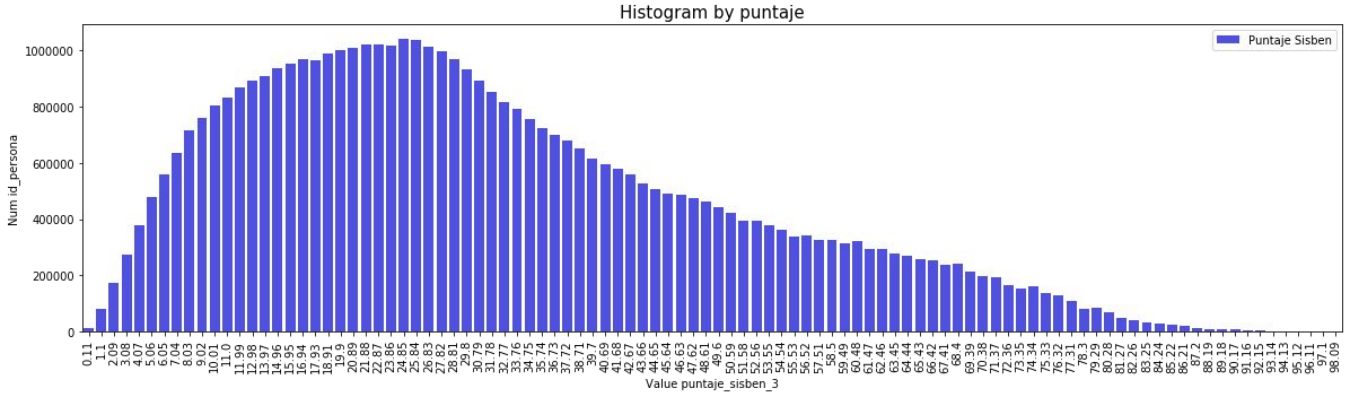
- leer las fuentes de datos y explorar

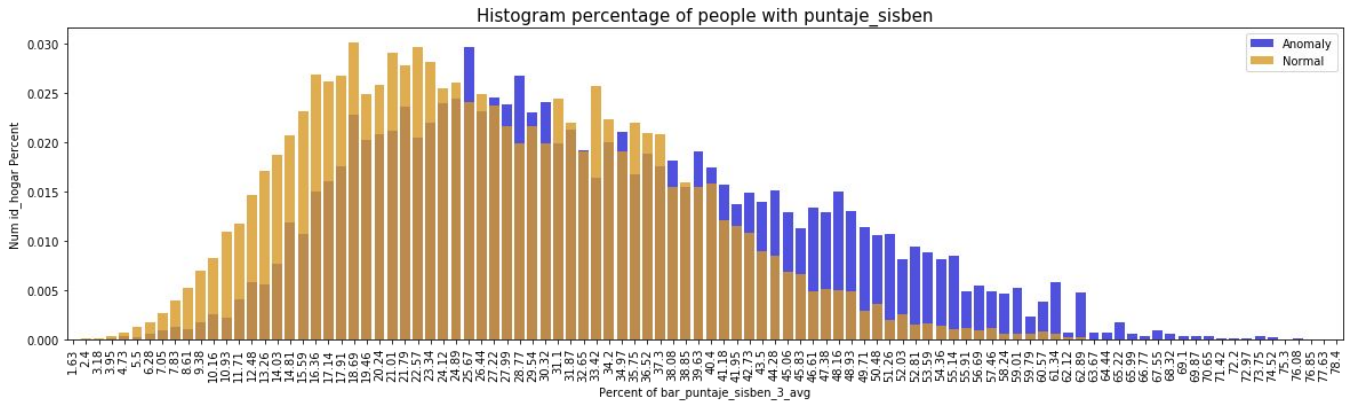
- calidad en los datos, nulos
- distribuciones
- encontrar outliers

Explorar variables

Cada variable debe tener una exploración:

- Numero de nulos.
- histograma
- Dominio de la variable por ej: [0 a 100] otro valor está fuera del dominio
- relación entre la variable objetivo y la variable actual.





Histograma es mucho mejor que los scatter plots en grandes volúmenes de datos. Se debe calcular el porcentaje de valores en cada canasta. Por ejemplo el 86.25% estan por debajo de 54.86 puntos

Explorar variables

Universidad Católica de Oriente

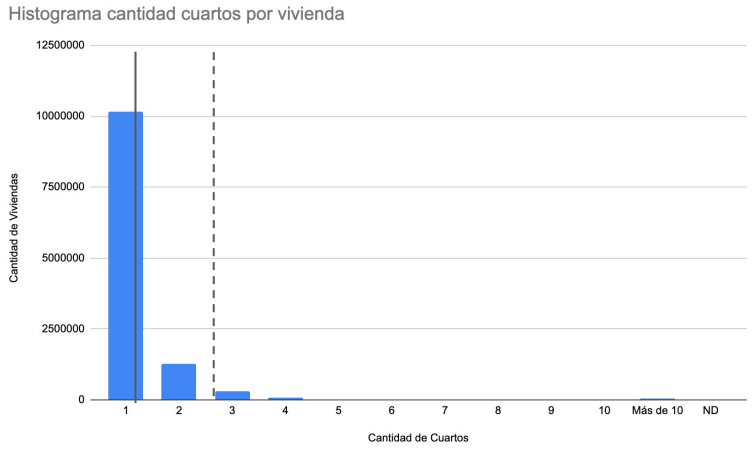
Variables numéricas:

numero de cuartos en el hogar

utilizar: describe, hist

Cantidad de Cuartos	Cantidad de Viviendas
Total de Datos	11’801.875
Promedio o media	2,70
Desviación Estándar	1,59
Mediana	1
Valor Mínimo	0
Valor Máximo	99

grandes diferencias entre el promedio y la mediana. Significa que la distribución tiene muchos valores a la derecha, cola larga



Una forma de detectar outliers (valores extremos) es utilizar la media y sumar 3 veces la desviación estándar.
Por ej todas las viviendas con más de 2,70 + 3* (1,59) = 6.36 cuartos son consideradas valores extremos.

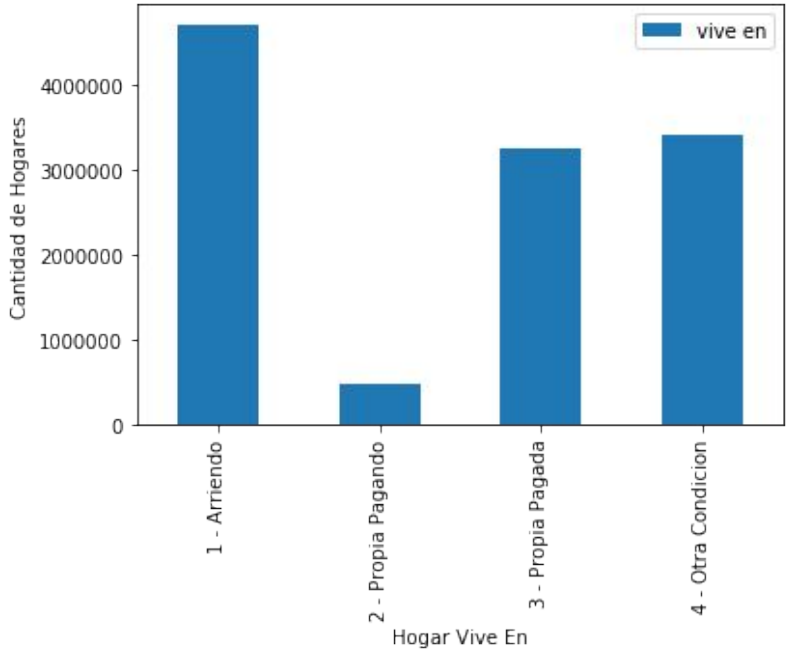
Explorar variables

Variables categóricas:

tipo de vivienda

utilizar: unique_values, or group by, catplot

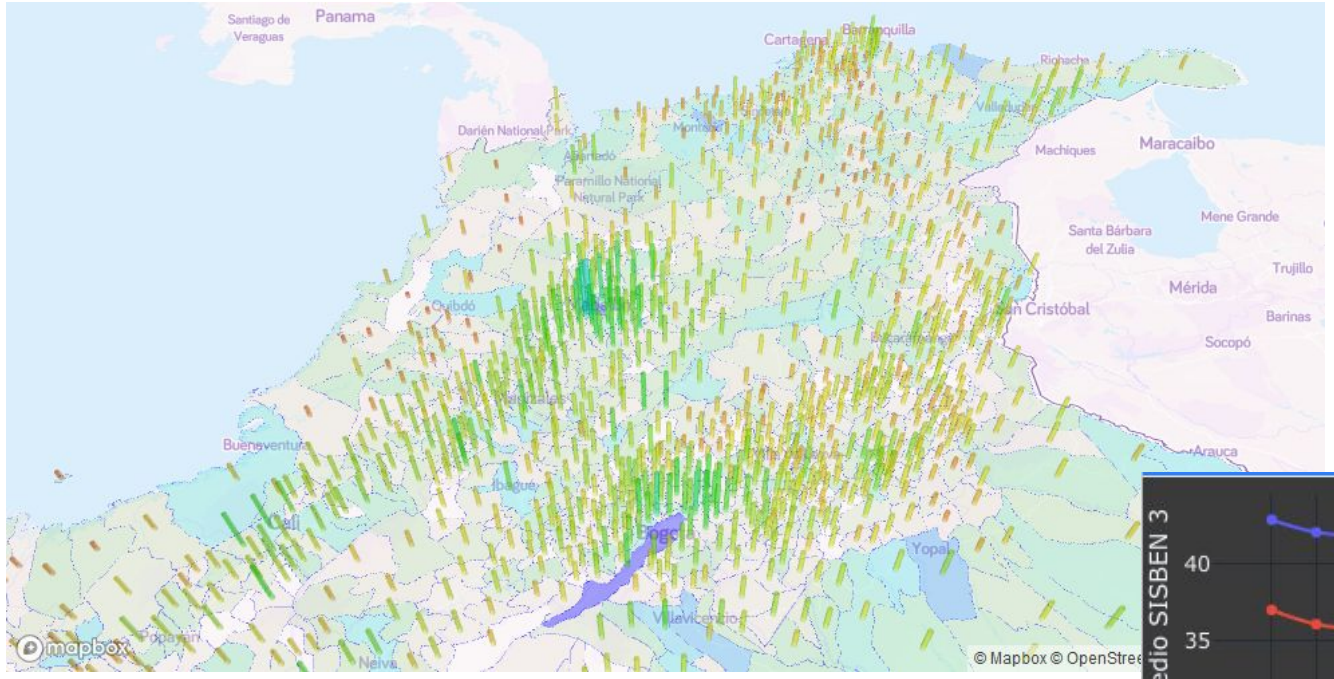
Cantidad de Cuartos	Cantidad de Viviendas
Total de Datos	11’801.875
Num valores unicos	4
Moda	Arriendo
Numero de veces que se repite	4’705.424



* https://pandas.pydata.org/docs/reference/api/pandas.Series.value_counts.html

Explorar variables, en tiempo y espacio

Porcentaje de anomalias por municipio y promedio de score



Puntaje promedio por área en el tiempo

area 1 = Ciudades, area 2 = centros poblados, area 3 = Rural

Explorar variables, en tiempo

“workflow” Flujos de trabajo para crear un modelo

DATA TRANSFORMATION — MODELOS ANALITICA

leer datos (data pulling)	explorar variables	transformar datos	sampling split (muestra)	entrenar modelo	evaluar modelo
- leer las fuentes de datos y explorar	- calidad en los datos, nulos - distribuciones - encontrar outliers	- imputar - centrar, escalar - agrupar variables con muchas categorías			

transformar datos - imputar

Asignar un valor a los registros que estan nulos o valores que están por fuera del dominio de la variable.

Cantidad de Cuartos	Cantidad de cuartos	SISBEN Score
Persona 1	1	40
Persona 2	ND(1)	38
Persona 3	2	60
Persona 4	1	31

Ej, qué valor le deberíamos poner a ND?

- Un nuevo valor numérico defecto como -1 o 0.
- Utilizar la media (2.73)
- Utilizar la mediana (1)

transformar datos - imputar

Asignar un valor a los registros que estan nulos o valores que están por fuera del dominio de la variable.

Cantidad de Cuartos	Cantidad de cuartos	SISBEN Score
Persona 1	1	40
Persona 2	ND	38
Persona 3	2	60
Persona 4	1	31

Ej, qué valor le deberíamos poner a ND?

- Un nuevo valor numérico defecto como -1 o 0.
- Utilizar la media (2.73)
- Utilizar la mediana (1)

Si el numero de ND es grande, dibuja la distribución de las personas que tienen 1 cuarto, 2, 3 y más de 4. Compara esas distribuciones con las personas ND y selecciona alguna similar.

* Utilizar SimpleImputer o KNNImputer de https://scikit-learn.org/stable/modules/impute.html

transformar datos - escalar y centrar

Universidad Católica de Oriente

Escalar los valores:

Scaler MinMax = (x - min(x))/(max(x) - min(x))
Todos los valores entre 0 y 1, positivos

Cantidad de Cuartos	Cantidad de cuartos	SISBEN Score
Persona 1	1 (0)	40 (0.31)
Persona 2	1 (0)	38 (0.24)
Persona 3	2 (0.33)	60 (1)
Persona 4	4 (1)	31 (0)

Scaler Standard = (x - mean(x))/std(x)
La suma de cada columna es 0 y la std es 1, funciona bien con outliers

Cantidad de Cuartos	Cantidad de cuartos	SISBEN Score
Persona 1	1 (-0.816)	40(-0.20)
Persona 2	1(-0.816)	38(-0.39)
Persona 3	2(0)	60(1.64)
Persona 4	4(1.63)	31(-1.04)

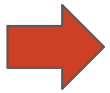
Al escalar los datos solo se debe incluir los datos del training. * Escalar los datos puede mejorar los resultados del modelo

* https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.scale.html

transformar datos - agrupar variables con alta cardinalidad

Convertir las variables categóricas en números creando “dummy variables”

Cantidad de Cuartos	Vive en
Persona 1	Propia Pagando
Persona 2	Propia Pagada
Persona 3	Propia Pagando
Persona 4	Arriendo



Cantidad de Cuartos	Propia Pagada	Propia Pagada	Arriendo
Persona 1	1	0	0
Persona 2	0	1	0
Persona 3	1	0	0
Persona 4	0	0	1

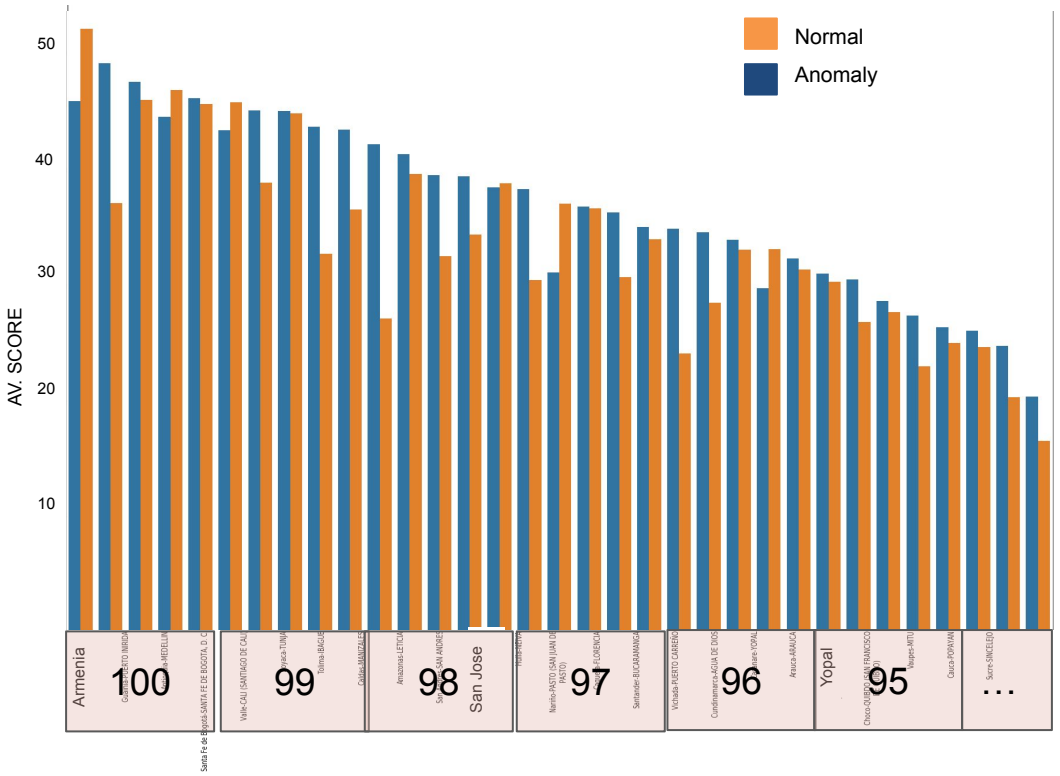
La mayoría de modelos transforman los datos iniciales con multiplicaciones para minimizar el error en la predicción o regresión, crear una dummy variable permite tener trazabilidad de la importancia de cada categoría.

* https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html

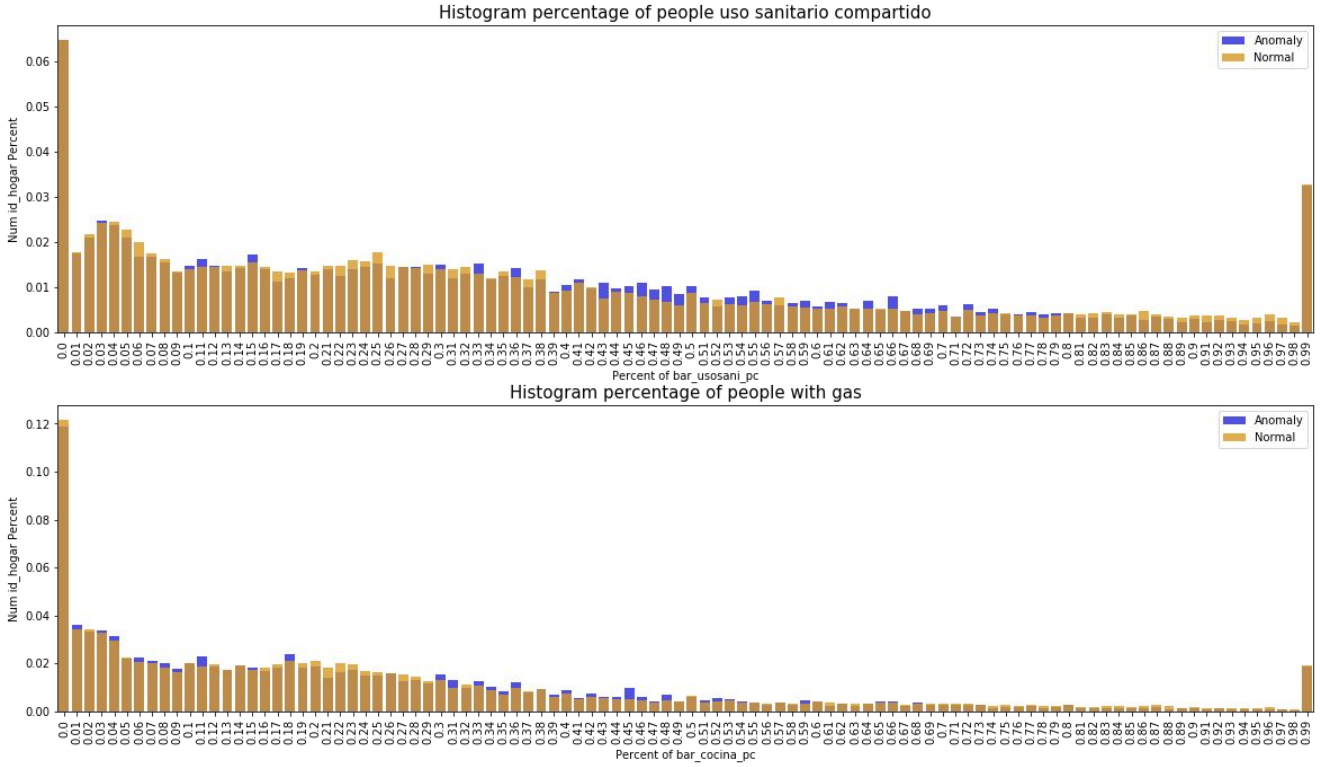
transformar datos - agrupar variables con alta cardinalidad

Existen variables con muchas categorías: ej, La variable municipio en el Sisben +1000 Municipios:

Opción 1. Ordenar por una variable y crear nuevas categorías del mismo tamaño (100)



Opción 2. Reagrupar (solo 14 ciudades y no ciudad) o nuevas variables numéricas. Utilizar barrio para describir el contexto de la encuesta en infraestructura y servicios públicos



Utilizar muchas dummy variables requiere muchos datos, por cada variable se debe tener exponencial veces más datos

* https://en.wikipedia.org/wiki/Curse_of_dimensionality
* https://pandas.pydata.org/docs/reference/api/pandas.qcut.html

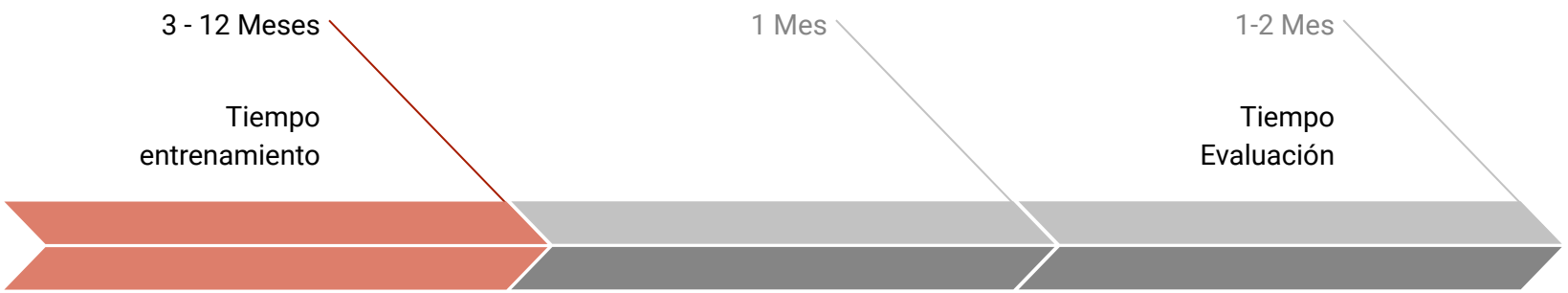
“workflow” Flujos de trabajo para crear un modelo

DATA TRANSFORMATION — MODELOS ANALITICA

leer datos (data pulling)	explorar variables	transformar datos	sampling split (muestra)	entrenar modelo	evaluar modelo
- leer las fuentes de datos y explorar	- calidad en los datos, nulos - distribuciones - encontrar outliers	- imputar - centrar, escalar - agrupar variables con muchas categorías	- split por tiempo, % de datos. - subsampling, oversampling		

sampling split - split

Para modelos que tienen una variable objetivo que estan cambiando en el tiempo. Solo se pueden hacer sampling o muestreo con los datos de training y validación



Training period

- **10% Validación, data utilizada para ver el progreso del modelo en el training.**
- **20% Testing: Data utilizada para medir el modelo después de entrenarlo**
- **70% Training: Datos utilizados para entrenar el modelo**

Período seguridad
Tiempo excluido de la evaluación, para garantizar que ninguna información filtrada del periodo de evaluación a entrenamiento

sampling split - sampling

Para modelos que tienen una variable objetivo que estan cambiando en el tiempo. Solo se pueden hacer sampling o muestreo con los datos de training y validación.
Ej En el Sisben tenemos 0.3% de formularios con fraude.

Opción 1 - Subsampling:
Hacer un muestreo por departamento e incluir todo el 0.3% de fraude y un 3% de no fraude (relación 1 a 10). Los efectos:

- La salida del modelo es un ratio y no una probabilidad
- Reduce la cantidad de datos, y hace que sea más propenso a equivocarse.

Opción 2 - No hacer sampling.

Dejar todos los datos. Los efectos:

- La salida del modelo es una probabilidad.
- Se debe utilizar una métrica que permita evaluar el modelo con datos desbalanceados.

Opción 3 - Oversampling:
Aumentar sintéticamente los formularios fraudulentos. Synthetic Minority Oversampling Technique, or SMOTE.

- Ayuda a reducir los falsos positivos. Formularios que son clasificados como Fraude pero no lo son.
- La salida del modelo es un ratio y no una probabilidad

Este tipo de decisiones depende del resultado del modelo. Es importante evaluar múltiples opciones en la construcción del modelo

“workflow” Flujos de trabajo para crear un modelo

DATA TRANSFORMATION — MODELOS ANALITICA

leer datos (data pulling)	explorar variables	transformar datos	sampling split (muestra)	entrenar modelo	evaluar modelo
- leer las fuentes de datos y explorar	- calidad en los datos, nulos - distribuciones - encontrar outliers	- imputar - centrar, escalar - agrupar variables con muchas categorías	- split por tiempo, % de datos. - subsampling, oversampling	- utilizar datos de training y evaluar en test	

“workflow” Flujos de trabajo para crear un modelo

DATA TRANSFORMATION — MODELOS ANALITICA

leer datos (data pulling)	explorar variables	transformar datos	sampling split (muestra)	entrenar modelo	evaluar modelo
- leer las fuentes de datos y explorar	- calidad en los datos, nulos - distribuciones - encontrar outliers	- imputar - centrar, escalar - agrupar variables con muchas categorías	- split por tiempo, % de datos. - subsampling, oversampling	- utilizar datos de training y evaluar en test	- evaluar los resultados del modelo en múltiples segmentos.

“workflow” Flujos de trabajo para crear un modelo

DATA TRANSFORMATION — MODELOS ANALITICA

leer datos (data pulling)	explorar variables	transformar datos	sampling split (muestra)	entrenar modelo	evaluar modelo
- leer las fuentes de datos y explorar	- calidad en los datos, nulos - distribuciones - encontrar outliers	- imputar - centrar, escalar - agrupar variables con muchas categorías	- split por tiempo, % de datos. - subsampling, oversampling	- utilizar datos de training y evaluar en test	- evaluar los resultados del modelo en múltiples segmentos.

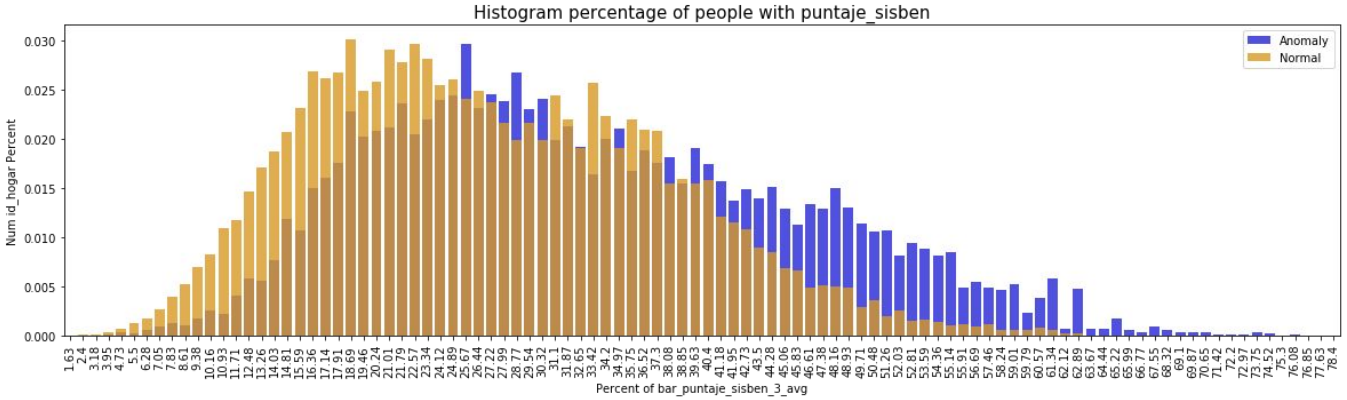
“workflow” Flujos de trabajo para crear un modelo

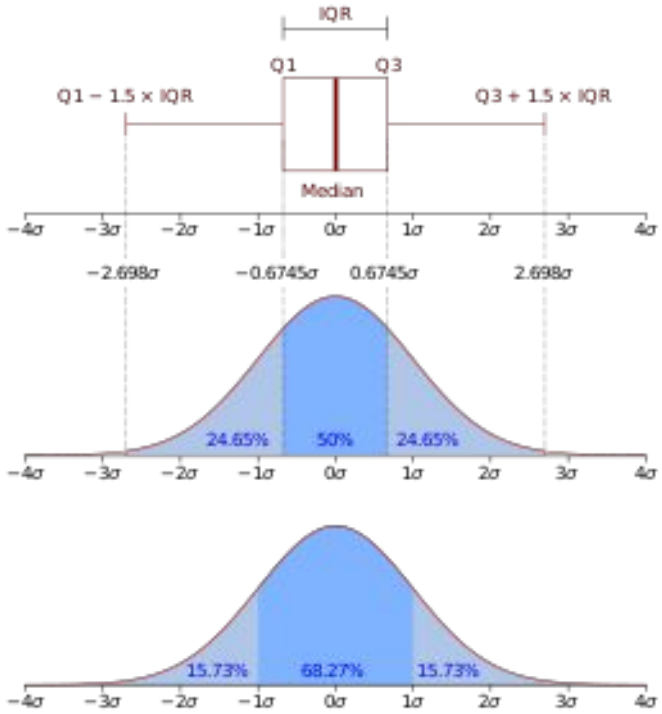
DATA TRANSFORMATION — MODELOS ANALITICA

leer datos (data pulling)	explorar variables	transformar datos	sampling split (muestra)	entrenar modelo	evaluar modelo
- leer las fuentes de datos y explorar	- calidad en los datos, nulos - distribuciones - encontrar outliers	- imputar - centrar, escalar - agrupar variables con muchas categorías	- split por tiempo, % de datos. - subsampling, oversampling	- utilizar datos de training y evaluar en test	- evaluar los resultados del modelo en múltiples segmentos.

APÉNDICE

Ejemplo histograma y equivalencia con boxplots





https://en.wikipedia.org/wiki/Box_plot