UNIVERSIDAD
SERGIO ARBOLEDA

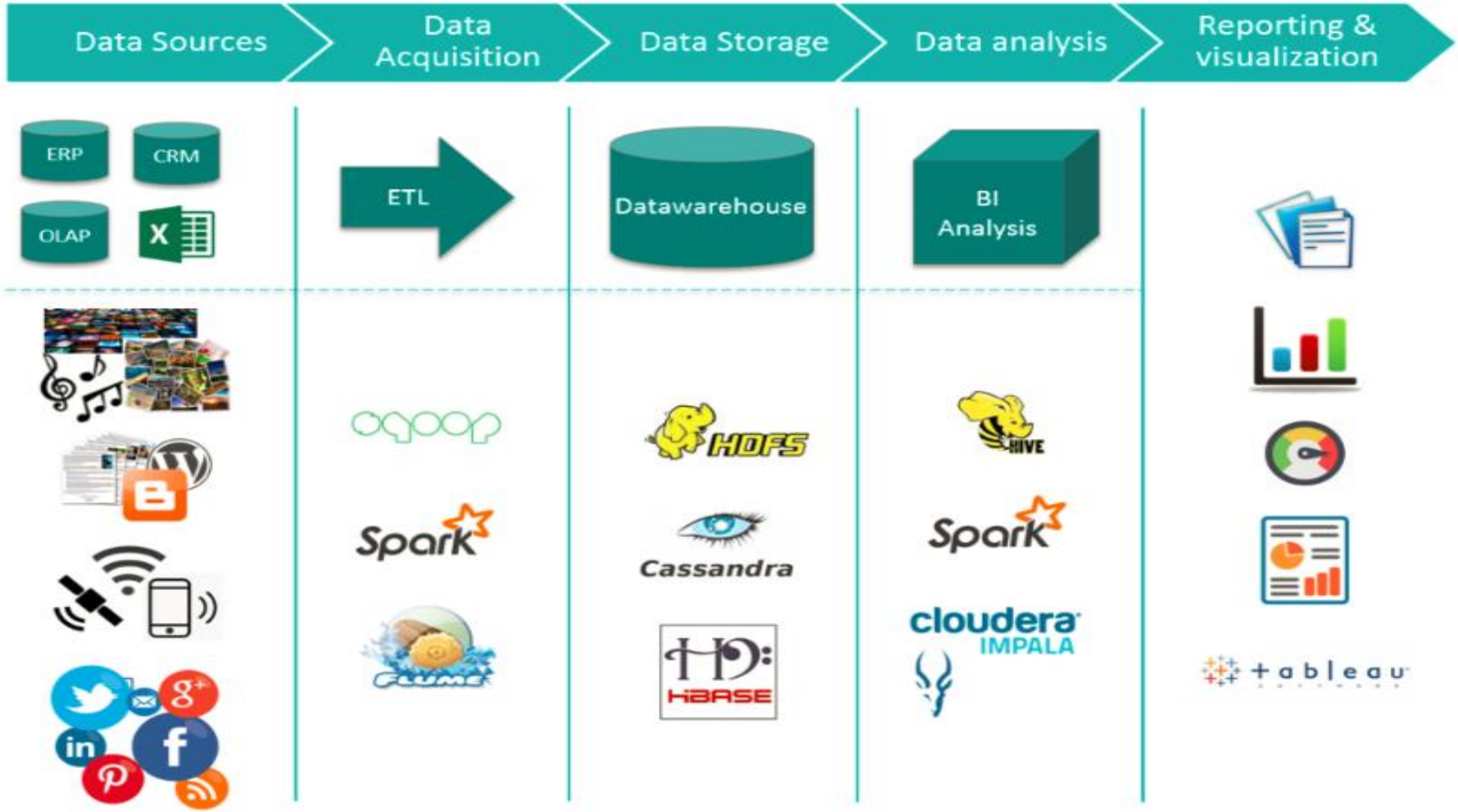
Big Data

Introducción, presentación y motivación

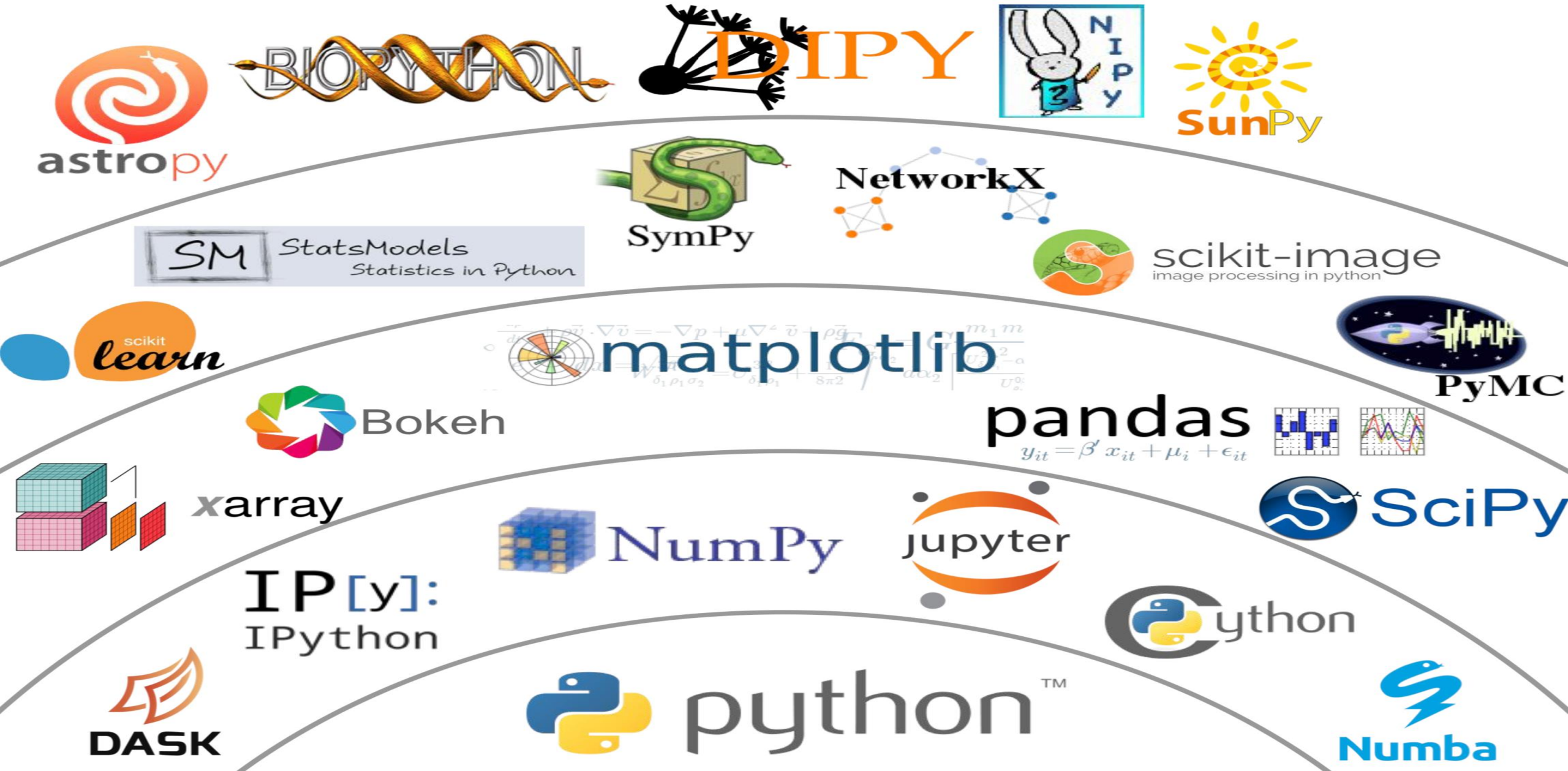
Camilo Yate Támara

Introducción

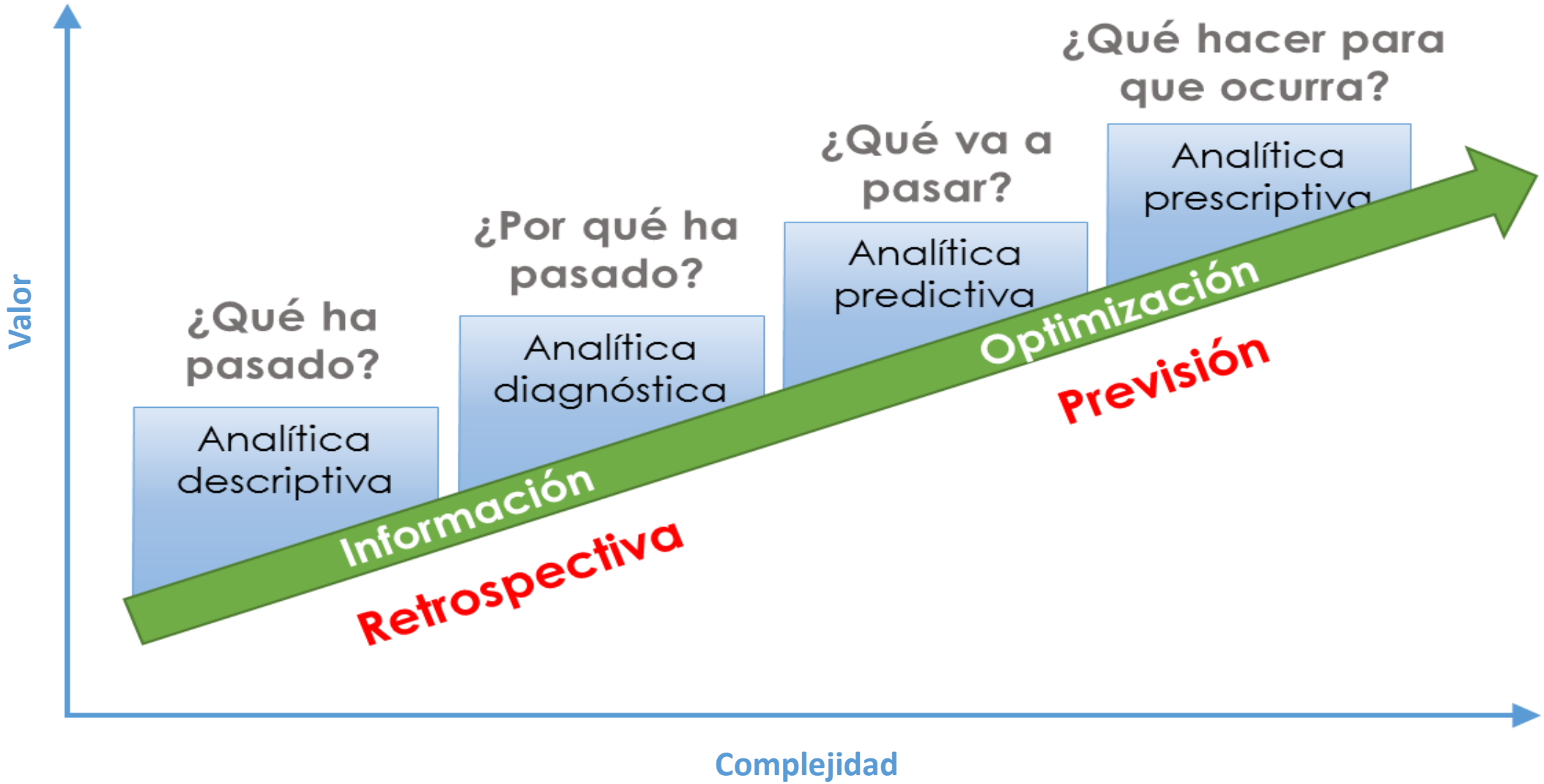
Introducción



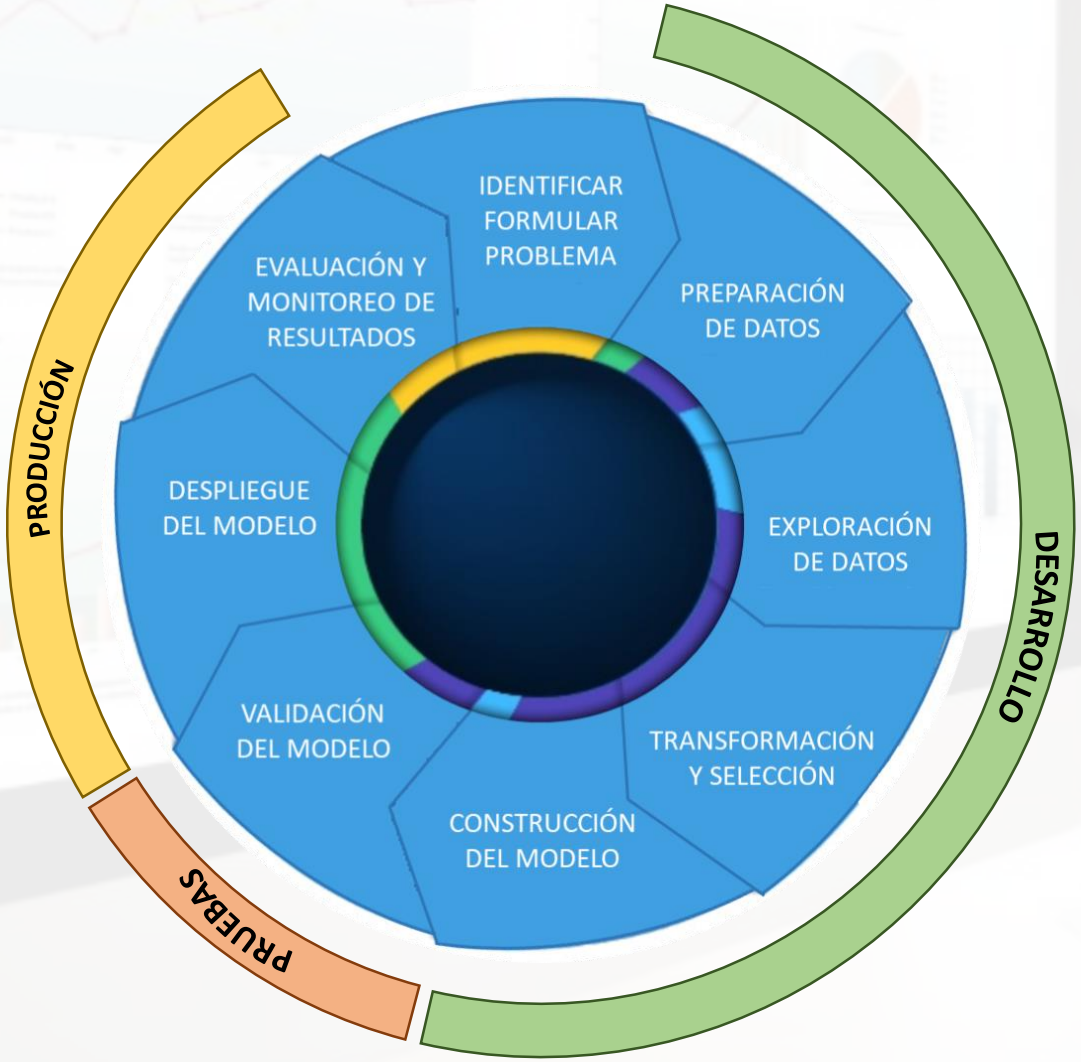
Introducción



Introducción



0 Metodología Analítica



- Modelo Cupo Sin Uso
- Recomendación Medicamentos
- Propensión Uso Recreacion
- Propensión a Servicios - Dados

Cupo sin Uso - Crédito

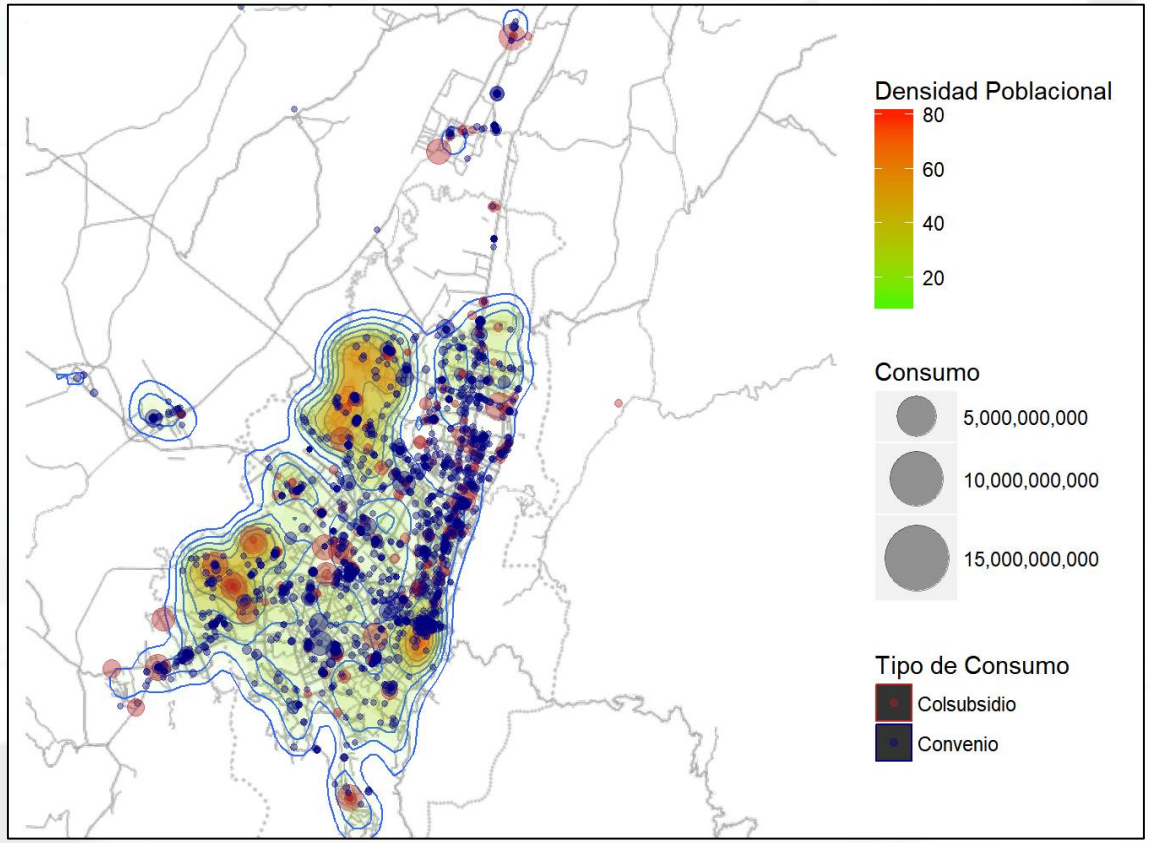
1 Cupo Sin Uso – Pregunta Negocio

Pregunta de Negocio

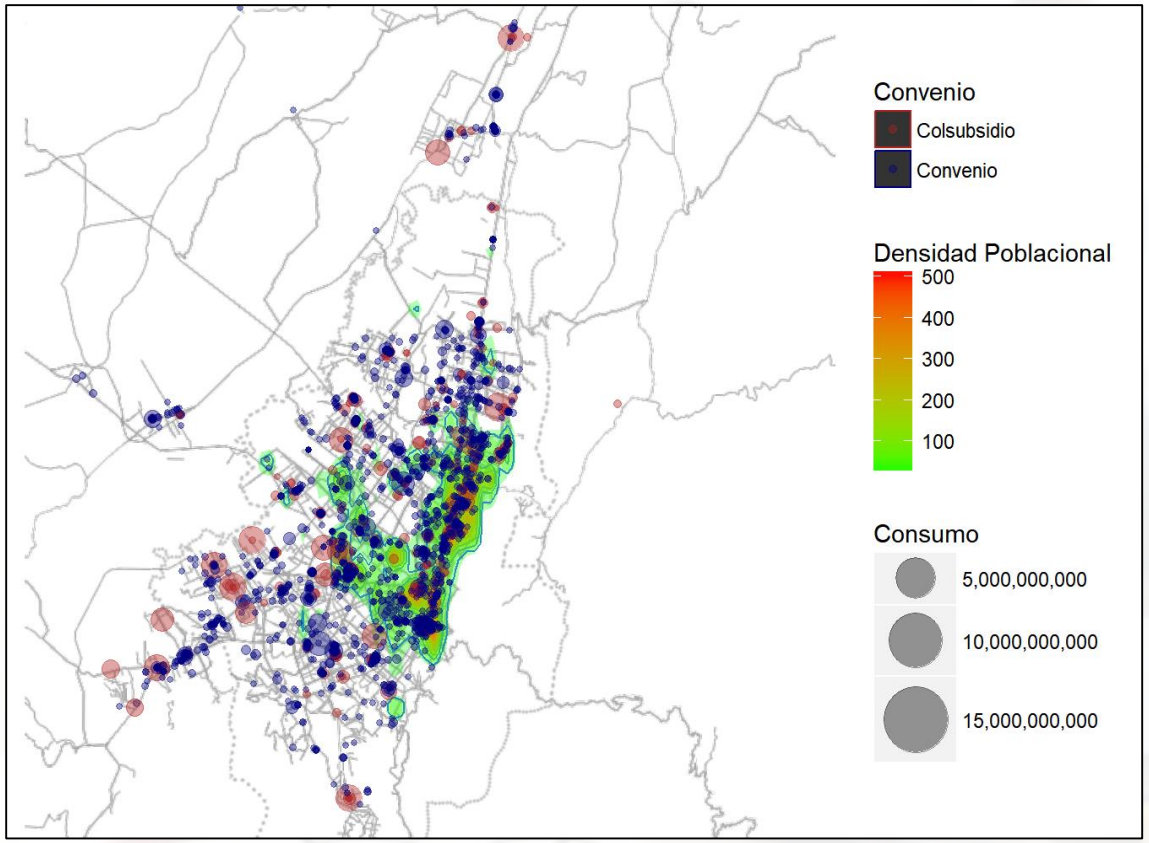
¿Qué estrategias desarrollar para incentivar el uso de TMS en aquellos clientes que nunca la han usado?

1 Cupo Sin Uso – Analisis Exploratorio

Consumo según lugar de residencia



Consumo según lugar de trabajo



1 Cupo Sin Uso – Construcción Modelo

Fase 1

Cálculo de la probabilidad de Compra en instalaciones de Colsubsidio

Partición

Entrenamiento (70%)	Prueba (30%)

Modelo de Clasificación

(conjunto de entrenamiento)

- **Regresión logística binomial**
- **Análisis Discriminante lineal**
- **Random Forest**

Comparación de Modelos (Conjunto de Prueba)

- **Comparación de AUROC & F1 Score.**

Promedio armónico entre la sensibilidad y especificad

Selección de Modelo

- **Regresión Logística**

Fase 2

Calculo de la UES de compra más probable

Partición

Entrenamiento (70%)	Prueba (30%)

Modelo de Clasificación

(conjunto de entrenamiento)

- **Regresión logística multinomial**
- **KNN**
- **Elastic Net**

Comparación de Modelos (Conjunto de Prueba)

- **Comparación de AUROC & F1 Score.**

Selección de Modelo

- **Regresión logística multinomial**

Fase 3

Calculo del Convenio de compra más probable

Partición

Entrenamiento (70%)	Prueba (30%)

Modelo de Clasificación

(conjunto de entrenamiento)

- **Regresión logística multinomial**
- **KNN**
- **Elastic Net**

Comparación de Modelos (Conjunto de Prueba)

- **Comparación de AUROC & F1 Score.**

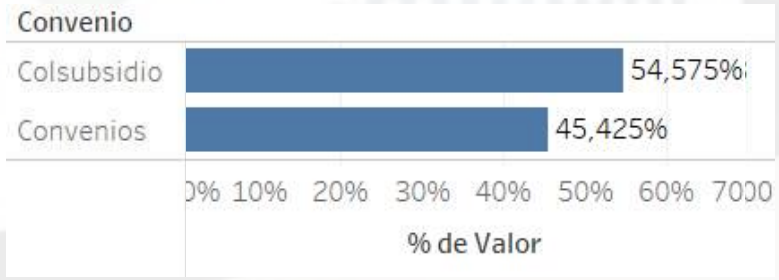
Selección de Modelo

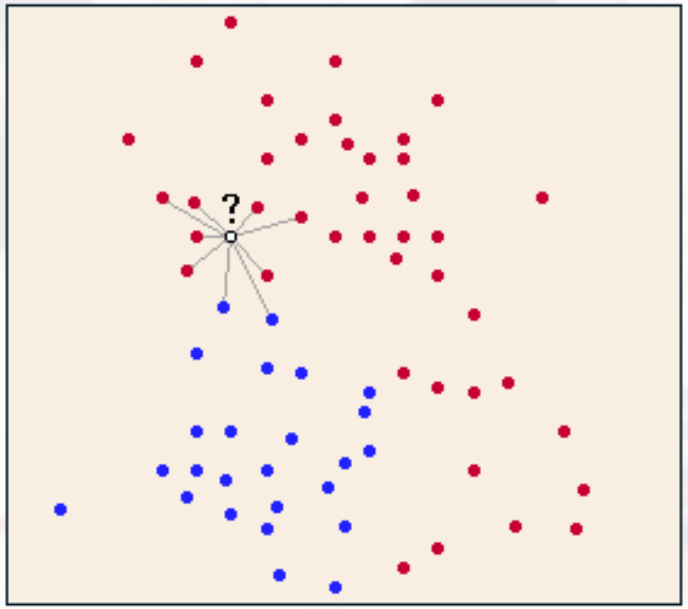
- **KNN**

Modelo Cupo Sin Uso – Resultados

Crédito
(Modelo Cupo sin Uso)

23,490	$4,847 mll	$360 mill
Activaciones Cupo	Consumo 2018	**Promedio mes 2019**





Sistema de Recomendación - Medicamentos

Analitica Avanzada - Colsubsidio

2 Sistema Recomendación - Pregunta

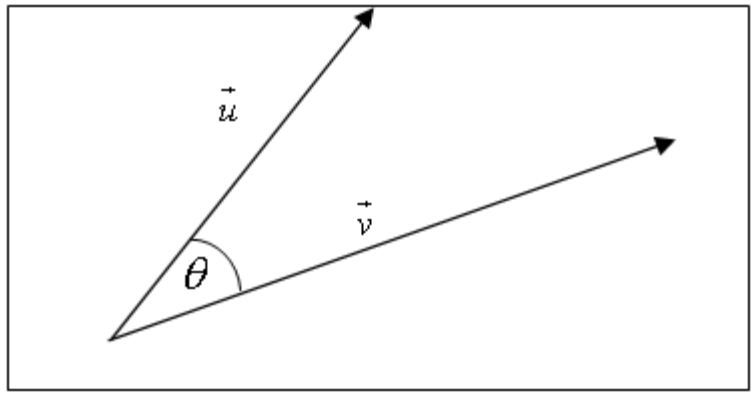
Pregunta de Negocio

¿Cuáles son los clientes que tendrían mejor respuesta a campañas de comunicación y promoción para los diferentes artículos ofrecidos por las droguerías Colsubsidio?

2 Sistema Recomendación – Metodologia

Los algoritmos de recomendación se basan en el producto punto entre dos vectores y en las fórmulas de correlación .

Producto Punto $\vec{u} \cdot \vec{v} = \left|\vec{u}\right|\left|\vec{v}\right| \cos\theta$



Correlación $sim(u,v) = \dfrac{\sum_{i=1}^{m}\left(r_{u,i} - r_{\vec{u}}\right)\left(r_{v,i} - r_{\vec{v}}\right)}{\sigma_u \sigma_v}$

- Algoritmos de Recomendación Basado en Usuario
- Algoritmos de Recomendación Basado en Articulos

	Artículos o Productos				
Usuario	A	B	...	...	n
1	1	1	0	0	0
2	1	1	0	1	0
3	1	0	1	0	1
4	0	0	1	0	0

En general, para un dataset con **n usuarios** y **m ítems**, para cada usuario se deben realizar **n-1 comparaciones**, en total **n(n-1).** En el peor de los casos cada comparación implica **m operaciones**

2 Sistema Recomendación – Metodologia

- **Coseno del ángulo de dos vectores** (invarianza, salvo signo, frente a homotecias)
- **Coeficiente de correlación** (invarianza frente a traslaciones y salvo signo frente a homotecias)
- **Medidas para datos dicotómicos**

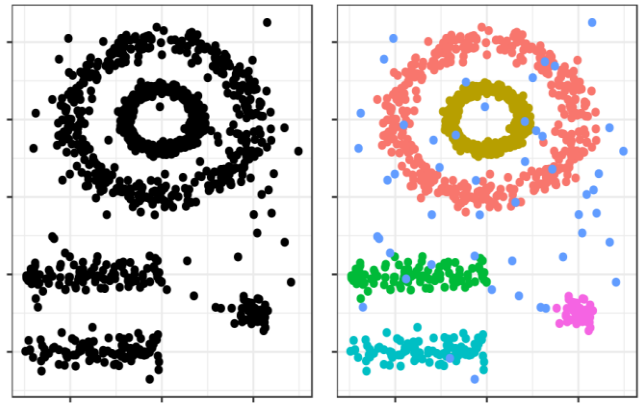
$X_i \setminus X_j$	1	0	Totales
1	a	b	a + b
0	c	d	c + d
Totales	a + c	b + d	m = a + b + c + d

- **Medida de Ochiai** → $\frac{a}{\sqrt{(a+b)(a+c)}}$
- **Medida Φ** → $\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$
- **Medida de Russell y Rao** → $\frac{a}{a+b+c+d}=\frac{a}{m}$
- **Medida de Parejas simples** → $\frac{a+d}{a+b+c+d}=\frac{a+d}{m}$
- **Medida de Jaccard** → $\frac{a}{a+b+c}$
- **Medida de Dice** → $\frac{2a}{2a+b+c}$
- **Medida de Rogers-Tanimoto** → $\frac{a+d}{a+d+2(b+c)}$
- **Distancia Euclídea:** $d(x_i,x_j)=\sqrt{\sum_{c=1}^{p}(x_{ic}-x_{jc})^2}$
- **Distancia de Minkowski:** $d_q(x_i,x_j)=\left(\sum_{c=1}^{p}\left|x_{ic}-x_{jc}\right|^q\right)^{\frac{1}{q}}$ donde $q\geq 1$
- **Distancia d_1 o ciudad (City Block):** $d(x_i,x_j)=\sum_{c=1}^{p}\left|x_{ic}-x_{jc}\right|$
- **Distancia de Tchebychev o del máximo** $(q=\infty)$: $d_{\infty}e(x_i,x_j)=\max(c=1,\cdots,p)\left|x_{ic}-x_{jc}\right|$
- **Distancia de Mahalanobis:** $D_S(x_i,x_j)=\sqrt{(x_i-x_j)'S^{-1}(x_i-x_j)}$
- **Distancia χ^2:** $\chi^2=m\left[\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{n_{ij}^2}{m_{\bullet}m_{\bullet j}}-1\right]$

2 Sistema Recomendación – Metodologia

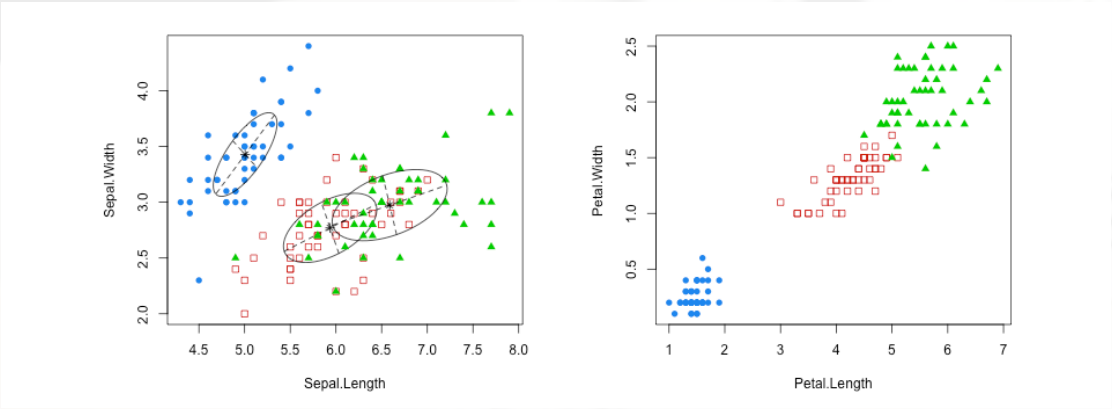
Métodos Basados en Densidades:

Buscan eliminar el supuesto de esfericidad de los datos. Sigue una forma de identificar clúster siguiendo el modo intuitivo en el que lo hace el cerebro humano, identificando regiones con alta densidad de observaciones separadas por regiones de baja densidad.



Métodos basados en distribuciones:

Considera que las observaciones proceden de una distribución (normal multivariante). En principio, cada clúster puede estar descrito por cualquier función de densidad, pero normalmente se asume que siguen una distribución multivariante normal.



2 Sistema Recomendación – Metodologia

Sistemas de Recomendación Planteados

Para los sistemas de recomendación se excluyen artículos los artículos de las categorías dispositivos médicos y *retail* y clientes esporádicos.

MÉTODO HÍBRIDO: FACTORIZACIÓN MATRICIAL CON CONTENIDO

Permite lidiar con el problema de cold-start incluyendo informacion adicional del usuario y del item.

MÉTODOS BASADOS EN REDUCCIÓN DE DIMENSIONALIDAD

En este método se supone que las filas y las columnas de la matriz de interacciones están altamente correlacionadas y puede ser aproximada por una matriz de menor rango usando los factores latentes de la matriz.

MÉTODOS BASADOS EN POPULARIDAD

Este método no personalizado calcula la proporción de cada ítem como estimación de la probabilidad de uso de cada ítem

MÉTODOS BASADOS EN MEMORIA

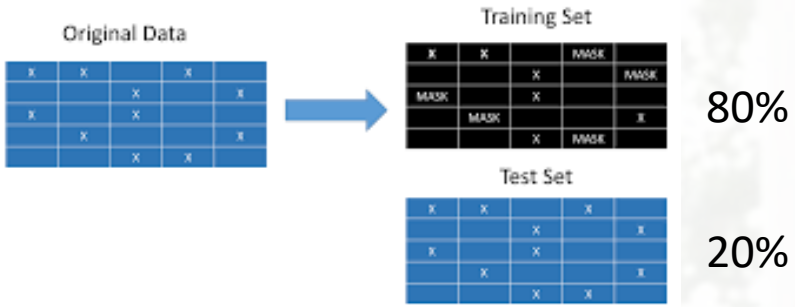
Las predicciones de cada interacción entre usuario y artículo están basada en vecindades entre usuarios o artículos. Pueden ser basados en Ítems o en Usuarios

RECOMENDACIÓN ALEATORIO

Recomienda a cada usuario un ítem aleatorio con el que el usuario no haya interactuado.

2 Sistema Recomendación – Metodologia

Con el fin de evaluar la calidad de un sistema de recomendación, es necesario particionar correctamente el conjunto de datos entre conjunto de entrenamiento y conjunto de prueba. En los sistemas de recomendación es usual el método *Holdout.*



Con el fin de determinar el modelo que otorga mejores rankings se utilizaran las siguientes métricas de evaluación

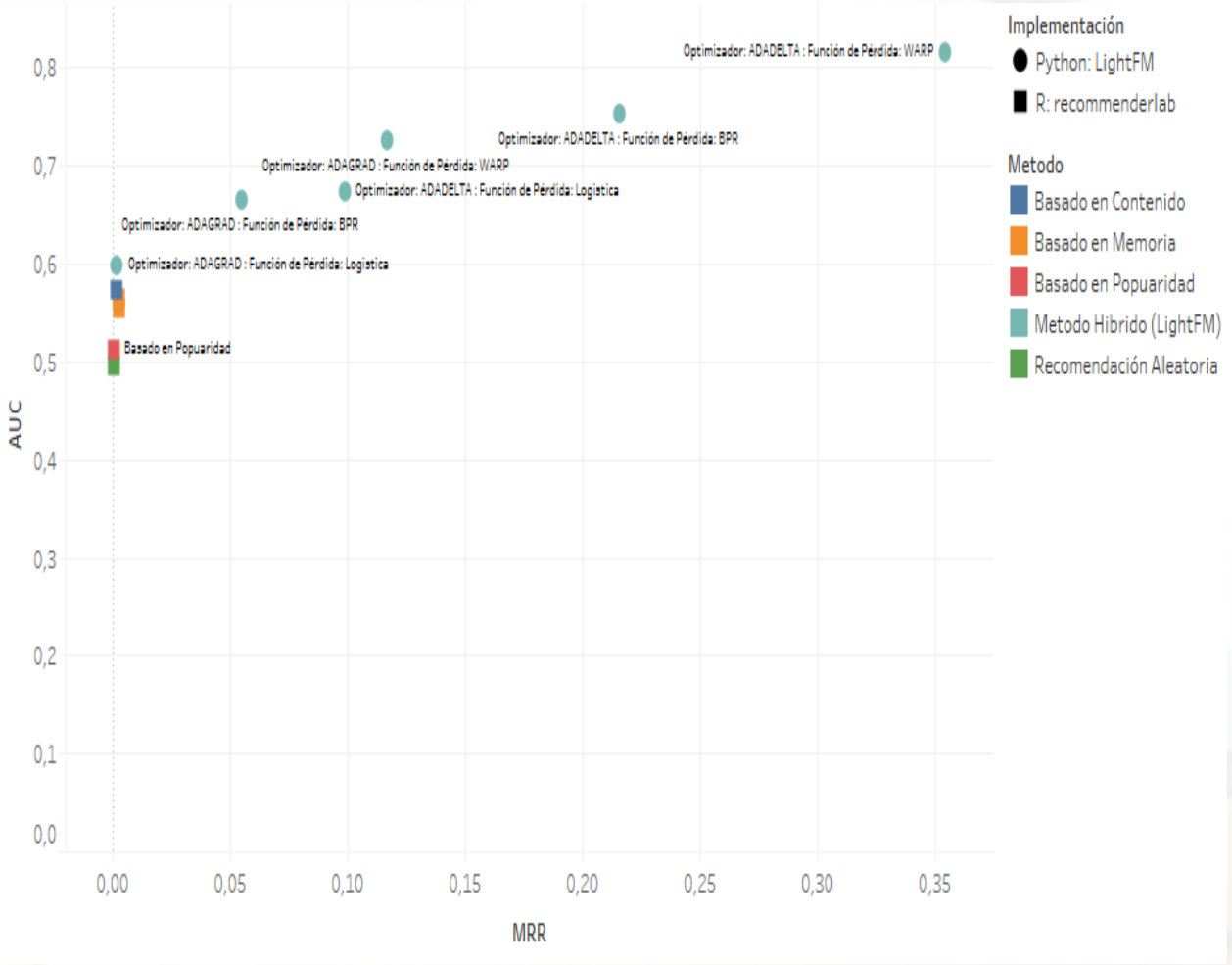
AUC: **la probabilidad que para un usuario en particular, un ítem con interacción positiva esté en un ranking de recomendación superior a un artículo sin interacción**

$$MRR = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{rango_i}$$

Los hiper parámetros de cada modelo son calibrados con el fin de maximizar ambas medidas

2 Sistema Recomendación – Metodologia

Modelo		Parámetros		Entrenamiento AUC	Entrenamiento MRR	Prueba AUC	Prueba MRR	Implementación
Recomendación Aleatoria				0,49578	0,00005	0,44620	0,00003	R: recommenderlab
Basado en Popularidad				0,51257	0,00008	0,34854	0,00007	R: recommenderlab
Basado en Memoria	Basado en usuarios	Numero de Vecinos: 100		0,56485	0,00265	0,56485	0,00218	R: recommenderlab
	Basado en Ítems	Numero de Vecinos: 5		0,55419	0,00254	0,36022	0,00160	R: recommenderlab
Basado en Contenido				0,57249	0,00150	0,48089	0,00085	R: recommenderlab
Método Hibrido (LightFM)		Optimizador: ADAGRAD	Función de Pérdida: Logística	0,59846	0,00154	0,49074	0,00131	Python: LightFM
		Optimizador: ADAGRAD	Función de Pérdida: BPR	0,66476	0,05482	0,45869	0,05372	Python: LightFM
		Optimizador: ADAGRAD	Función de Pérdida: WARP	0,72466	0,11655	0,68118	0,10606	Python: LightFM
		Optimizador: ADADELTA	Función de Pérdida: Logística	0,67246	0,09845	0,59849	0,07876	Python: LightFM
		Optimizador: ADADELTA	Función de Pérdida: BPR	0,75250	0,21548	0,73745	0,13360	Python: LightFM
		Optimizador: ADADELTA	Función de Pérdida: WARP	0,81462	0,35448	0,79945	0,31978	Python: LightFM



Uso de Servicios Clubes Horas Valle - RyT

3 Uso Clubes RyT– Metodologia

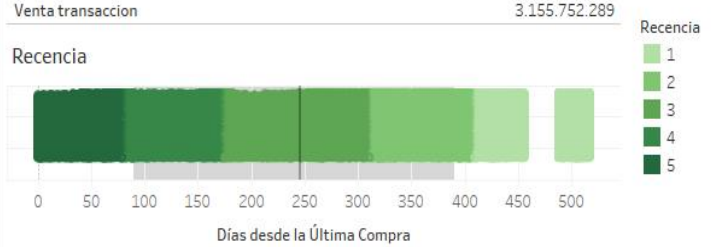
Pregunta de Negocio

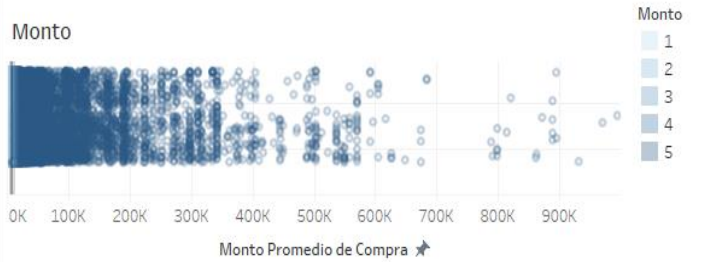
¿Cuáles son los clientes mas propensos a utilizar los servicios de los Clubes Colsubsidio, en servicios y horarios de baja demanda?

3 Uso Clubes RyT – Exploración RFM

Análisis Total

Venta Promedio Mes	$ 262.979.357
Promedio de Venta	$ 33.655
Venta Promedio Cliente por Mes	$ 2.805
Número de Clientes	93.767
Venta transaccion	3.155.752.289

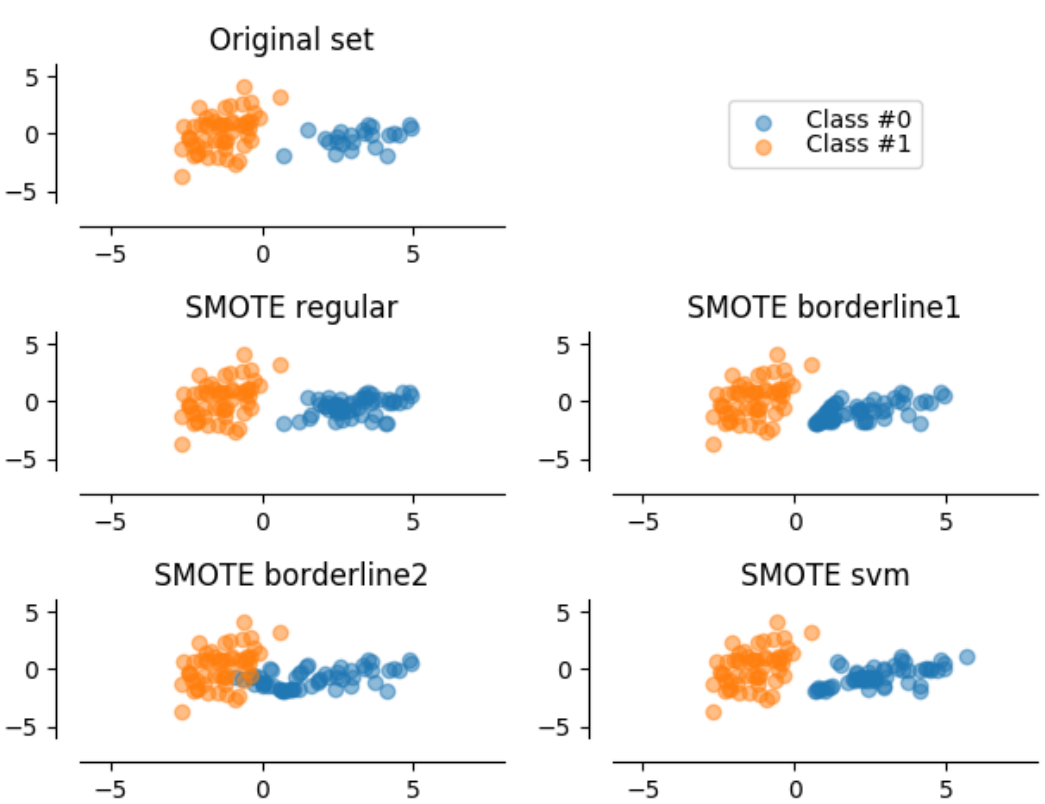




3 Uso Clubes RyT– Balanceo

La base de datos esta evidentemente desbalanceada, es decir, existe una gran proporción de productos que han sido muy poco usados. Esto suele representar problemas a la hora de estimar un modelo. Por lo cual se propone la siguiente metodología:

Synthetic Minority Over-sampling Technique (SMOTE)



3 Uso Clubes RyT – Modelo Analítica

Modelo de Clasificación:

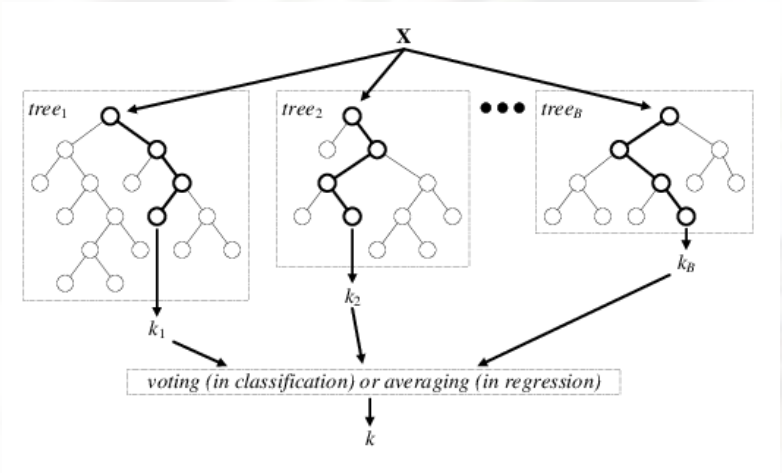
Se propone modelar la probabilidad la intención de compra de cada producto en cada club por para todos los clientes afiliados a la caja a partir de la información de los clientes que han consumido

Partición

Entrenamiento (70%)	Prueba (30%)

Se particiona la base de datos en conjunto de entrenamiento y validación con el fin de minimizar los errores de predicción del modelo

A partir de estadísticas de bondad de ajuste, se decide utilizar el Modelo Random Forest para estimar la probabilidad de deserción de los tarjetahabientes



AUROC: 0,9056

Modelo propensión a Servicios - Dados

4 Propensión Uso – Pregunta

Pregunta de Negocio

¿Cuáles serían los clientes con mayor propensión a consumir los de productos o servicios en el dado y asimismo adquirir una membresía?

4 Propensión Uso – Datos

Consumos

Información de Afiliado

Bases de Datos UES

Base de Datos Analítica

Modelo de Propensión Dados

- Modelos de Clasificación Multiclase
- Calificación individual de la probabilidad de consumir algún servicio.
- Agregación para calcular la probabilidad de adquirir membresía.

Construcción del Modelo

4 Propensión Uso – Modelo

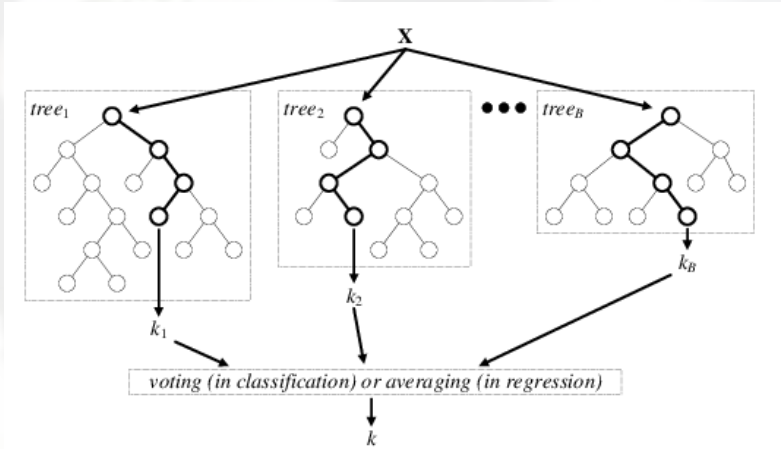
Modelo de Clasificación:

Se propone modelar la probabilidad la intención de uso de alguno de los servicios propuestos para el dado, basados en los consumos individuales (para afiliados o beneficiarios) en alguno de los servicios similares ofrecidos actualmente.

Partición

Entrenamiento (70%)	Prueba (30%)

Se particiona la base de datos en conjunto de entrenamiento y validación con el fin de minimizar los errores de predicción del modelo



Se utiliza una metodología OnevsAll con un Random Forest para estimar la probabilidad individual de cada servicio dados los consumos pasados y las demás covariables de la base

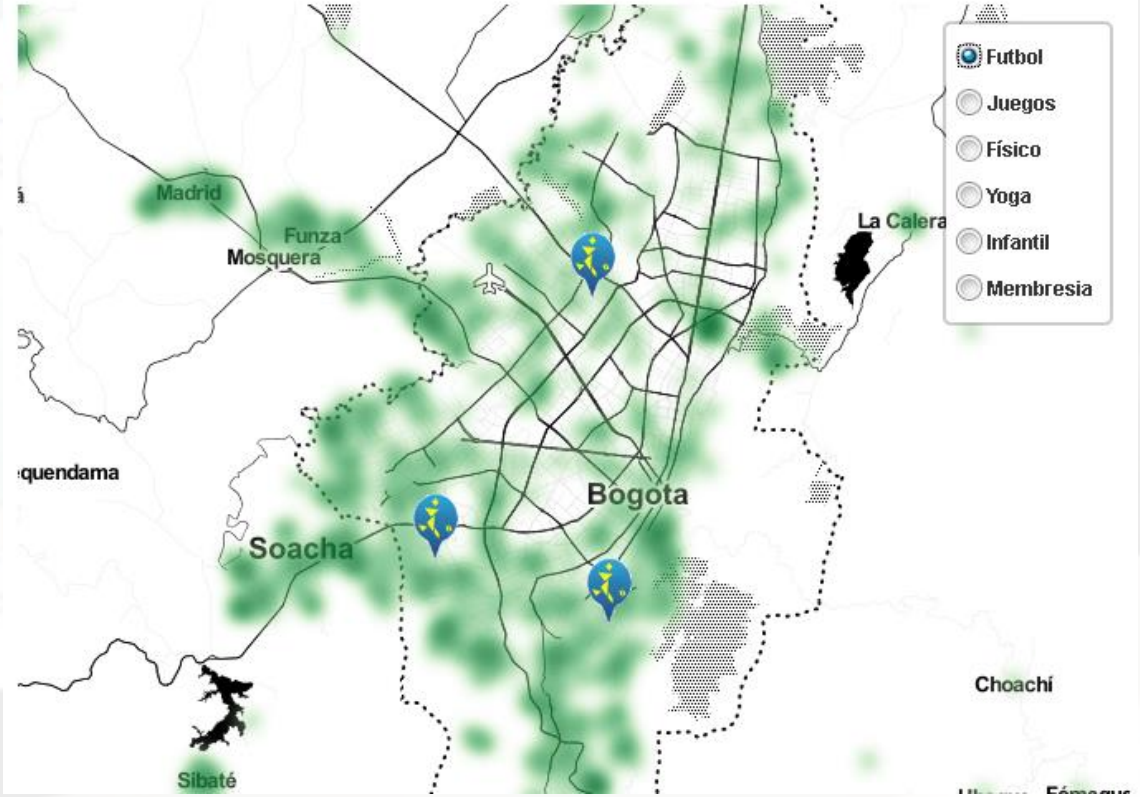
```
OneVsRestClassifier(estimator=RandomForestClassifier(bootstrap=True, class_weight='balanced',
            criterion='gini', max_depth=20, max_features='auto',
            max_leaf_nodes=None, min_impurity_decrease=0.0,
            min_impurity_split=None, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            n_estimators=1000, n_jobs=-1, oob_score=False,
            random_state=31415, verbose=0, warm_start=False),
          n_jobs=1)
```

https://colsubsidio.shinyapps.io/Dados/

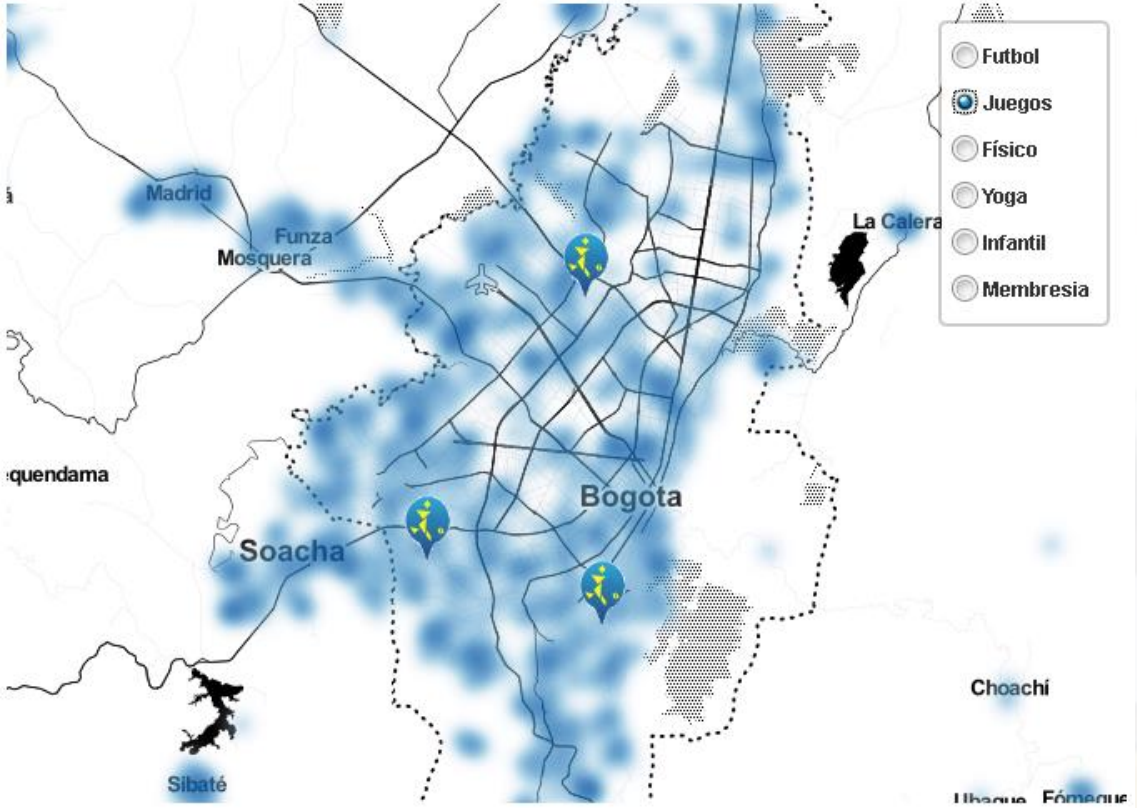
AUROC: 0,9127

4 Propensión Uso – Modelo

Mapa de Propensión de Servicios



Mapa de Propensión de Servicios



Contenidos

Contenidos

El curso se dividirá en tres capítulos principales:

1. Introducción a Python.
    I. Numpy
    II. SicPy
    III. Pandas
    IV. Matplotlib
    V. Plotly
2. Conceptos de Big Data
    I. Arquitectura Big Data
    II. Data Wrangling
    III. Herramientas Distribuidas
3. Machine Learning
    I. Descripción de Datos
    II. Aprendizaje no supervisado
    III. Aprendizaje Supervisado
    IV. Sistemas de Recomendacion

semana	TEMA	ACTIVIDADES DE APRENDIZAJE		
		ACOMPAÑAMIENTO DEL DOCENTE		TRABAJO INDEPENDIENTE
		TEORÍA	PRÁCTICA	
1	Presentación del curso	• Lectura de contenido programático • Acuerdos • Introducción		
2,3,4,5	Capítulo 1 Programación en Python	• Tipos de variables, datos, estructuras de control y funciones. • Data wrangling y Exploración de datos. • Modelos de regresión.		TALLER
6	PRIMER PARCIAL			
7,8	Capítulo 2 Ficheros, limpieza y descripción	• Arquitectura Big data, ¿cómo funciona? • RDD's • Estructuras de control y funciones. • Data wrangling		TALLER
9,10	Capítulo 3 Segmentación	• Exploración de datos • Clustering, K-means y Knn	SEGUNDA ENTREGA	
11	SEGUNDO PARCIAL			
12,13	Capítulo 4 Modeling	• Conceptos: modelos de regresión • Regresión Lineal • Regresión logística	LABORATORIO	ADELANTO PROYECTO
14, 15, 16	Capítulo 5 Predicción	• Matriz de confusión • Métricas de resultados • Análisis de resultados y selección de modelos	ADELANTO PROYECTO	ADELANTO PROYECTO
PRESENTACIÓN FINAL				

9 de Septiembre

14 de Octubre → PreProyecto

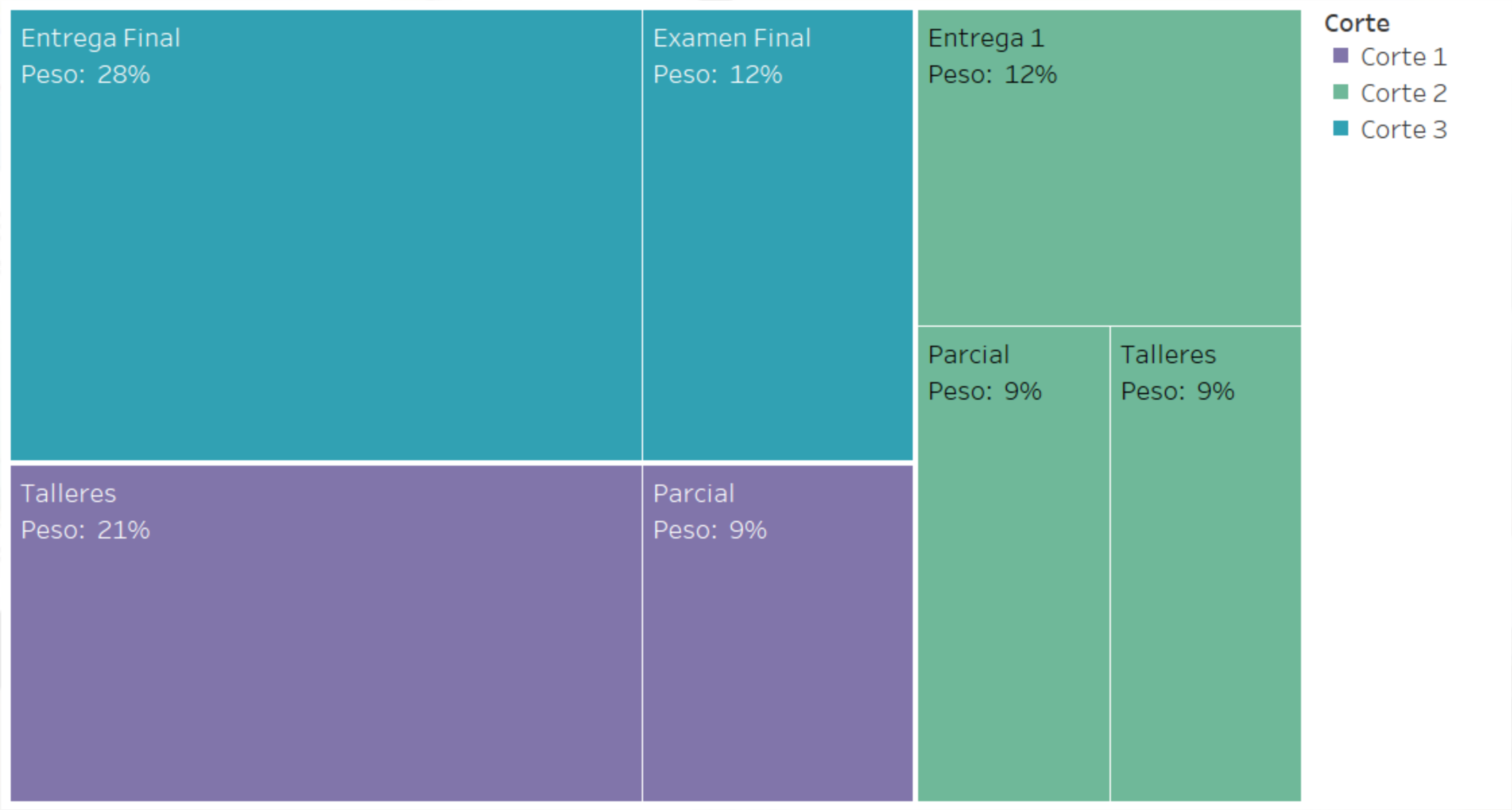
21 de Octubre → Parcial 2

18 de Noviembre→ Proyecto

25 de Noviembre→ Examen Final

Evaluación

Evaluación


Entrega Final
Peso: 28%
Examen Final
Peso: 12%
Entrega 1
Peso: 12%
Talleres
Peso: 21%
Parcial
Peso: 9%
Parcial
Peso: 9%
Talleres
Peso: 9%
Corte
Corte 1
Corte 2
Corte 3