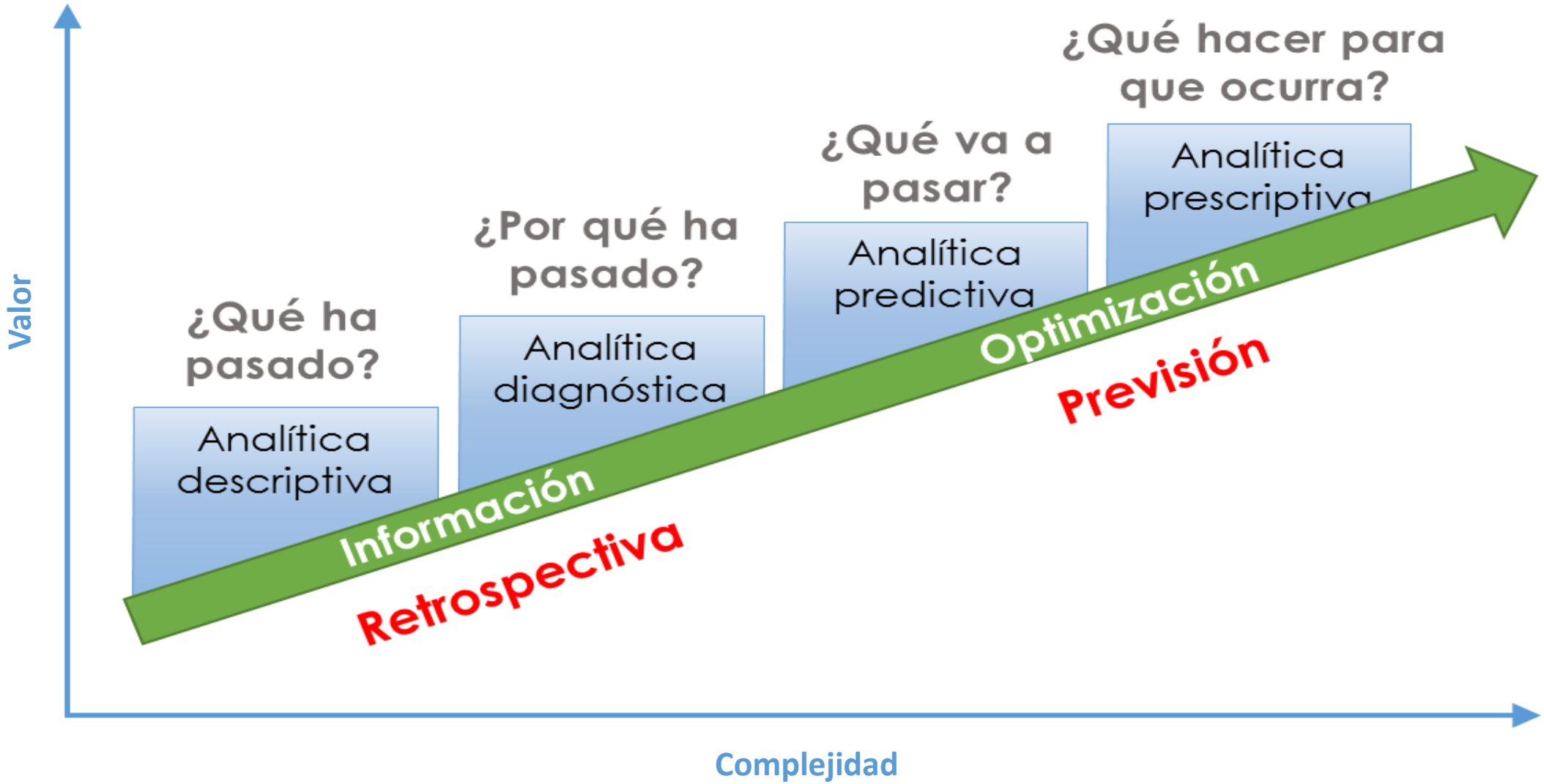
UNIVERSIDAD
SERGIO ARBOLEDA

Programación en R

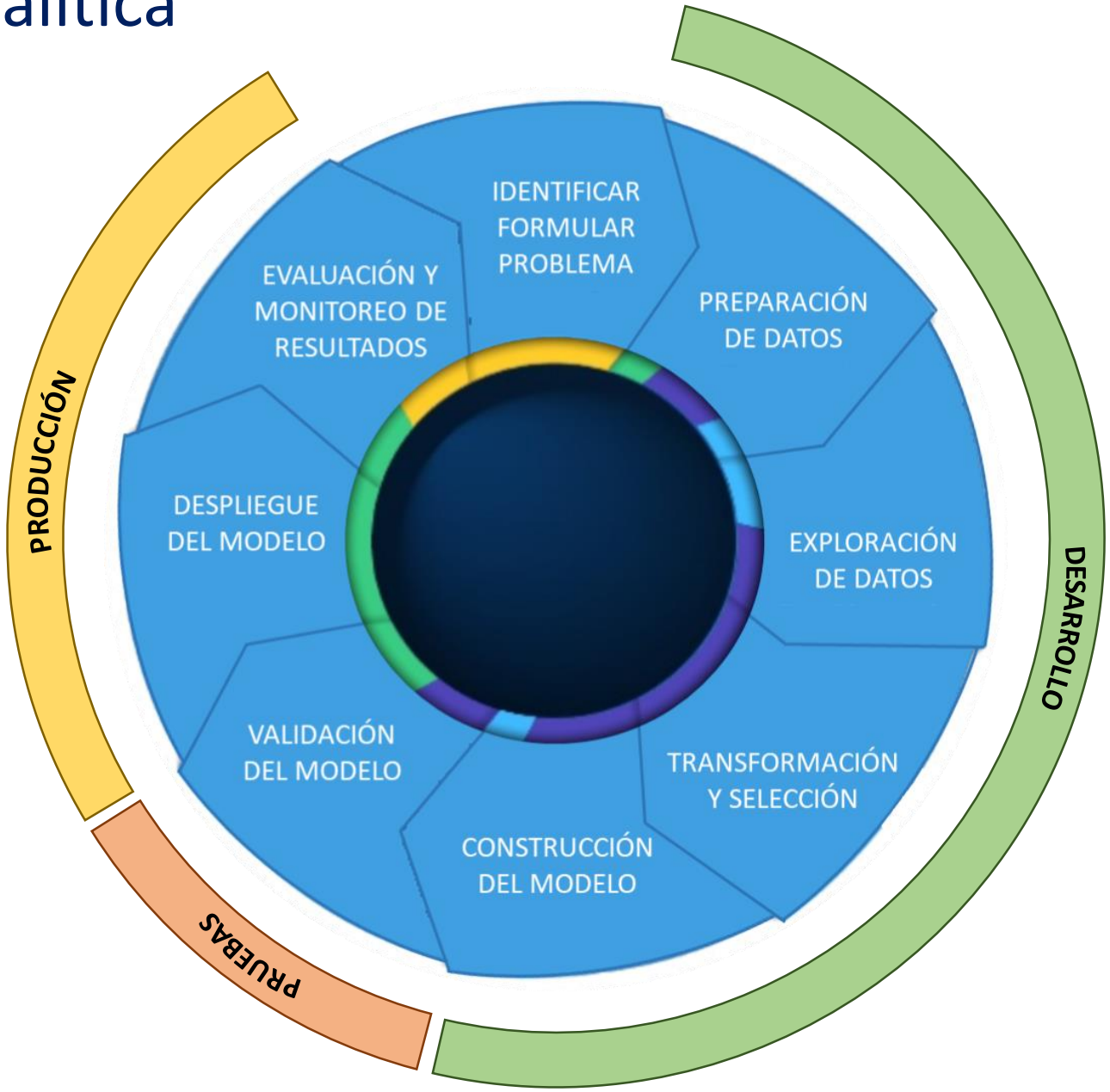
Introducción, presentación y motivación

Camilo Yate Támara

Introducción



Metodología Analítica



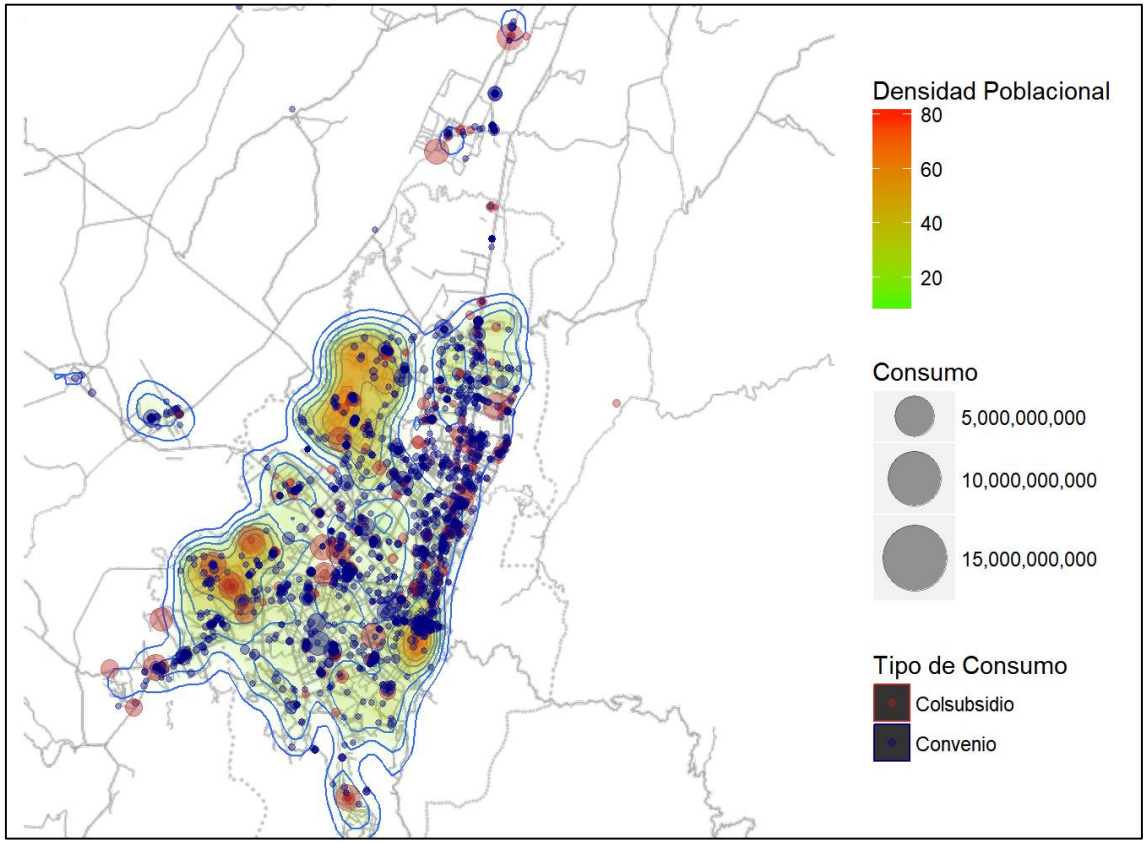
Casos de Uso – Fomento de Uso de Líneas de Crédito

Pregunta de Negocio

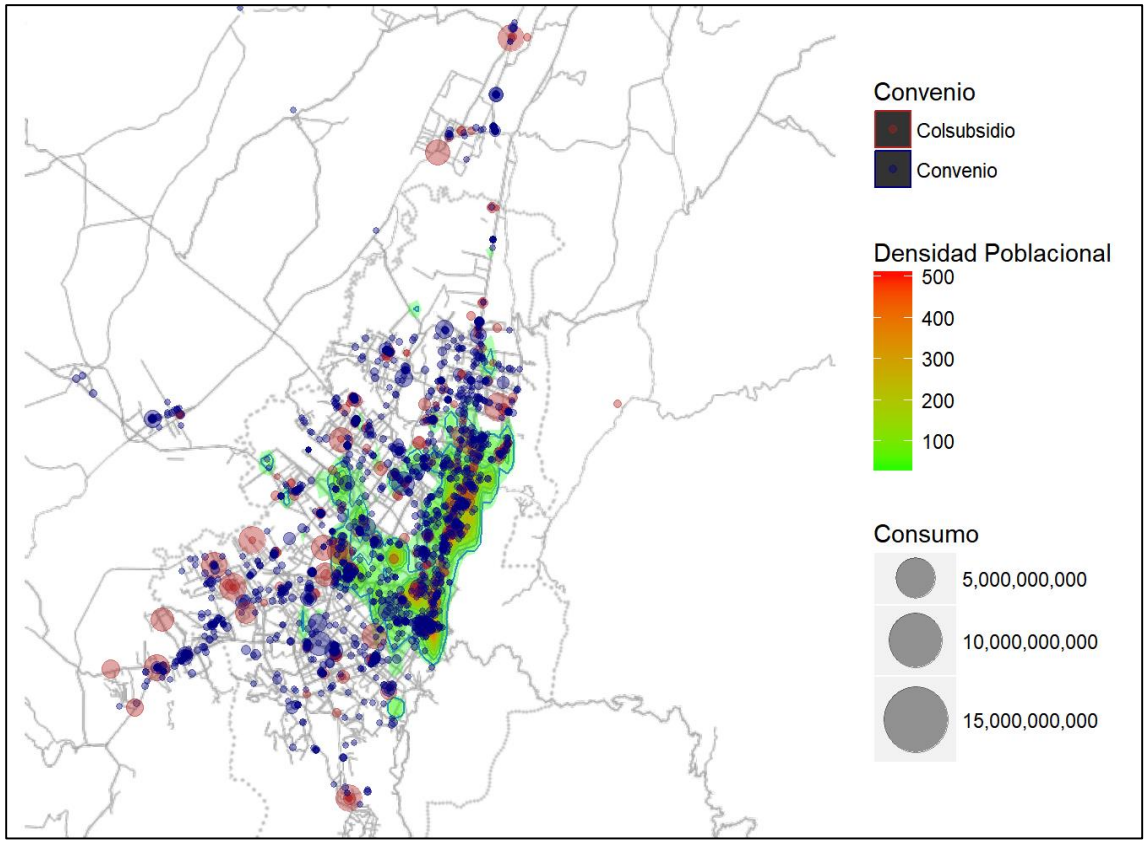
¿Qué estrategias desarrollar para incentivar el uso de tarjeta de crédito en aquellos clientes que nunca la han usado?

Casos de Uso – Fomento de Uso de Líneas de Crédito

Consumo según lugar de residencia



Consumo según lugar de trabajo



Casos de Uso – Fomento de Uso de Líneas de Crédito

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**

Casos de Uso – Sistemas de Recomendación

Pregunta de Negocio

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

Casos de Uso – Sistemas de Recomendación

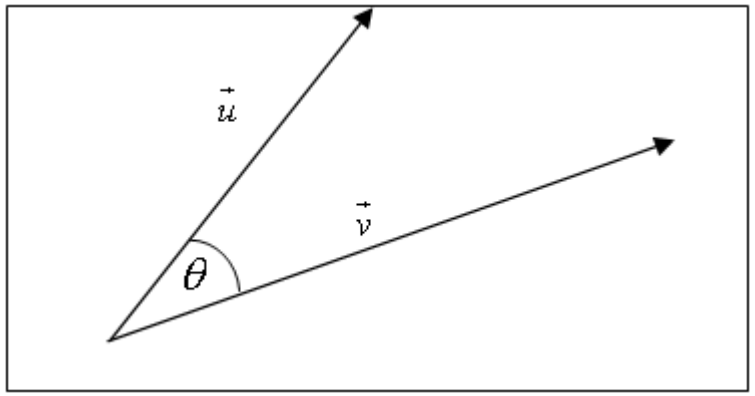
Pregunta de Negocio

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

Casos de Uso – Sistemas de Recomendación

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

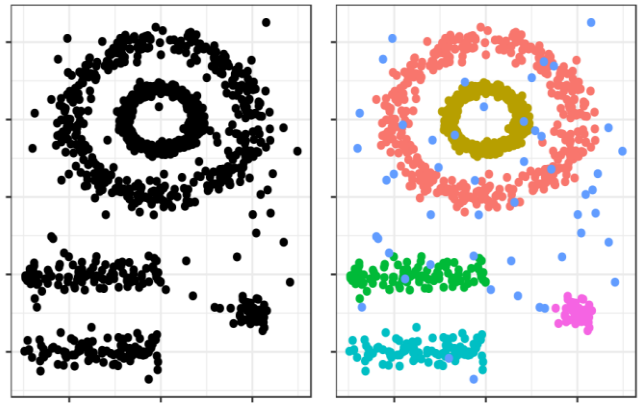
- Algoritmos de Recomendación Basado en Usuario
- Algoritmos de Recomendación Basado en Articulos

	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

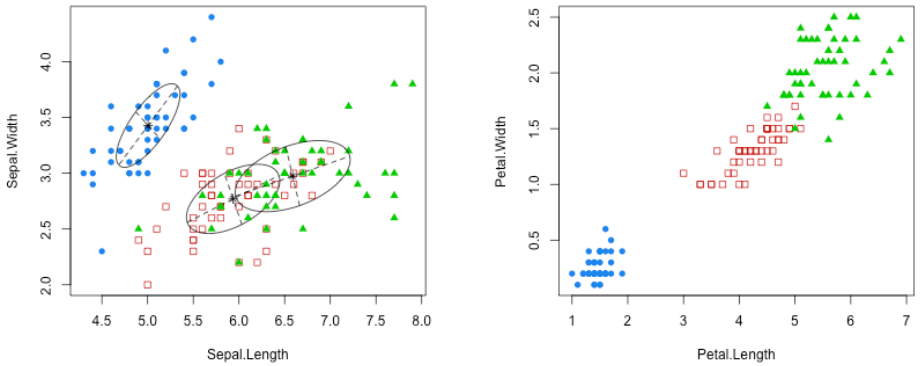
En general, para un dataset con **n usuarios** y **m ítems**, para cada usuario se deben realizar **n-1 comparaciones**, en total **n(n-1).** En el peor de los casos cada comparación implica **m operaciones**

Casos de Uso – Sistemas de Recomendación

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.

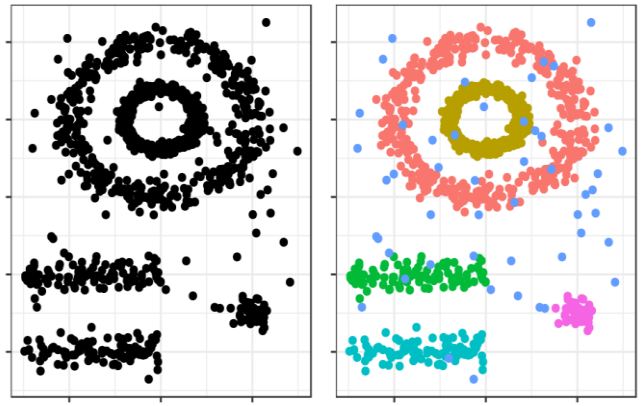


Métodos basados en distribuciones:
Considera que las observaciones proceden de una distribución (normal multivariante). En principio, cada clúster puede estar descrito por cualquier función de densidad, pero normalmente se asume que siguen una distribución multivariante normal.

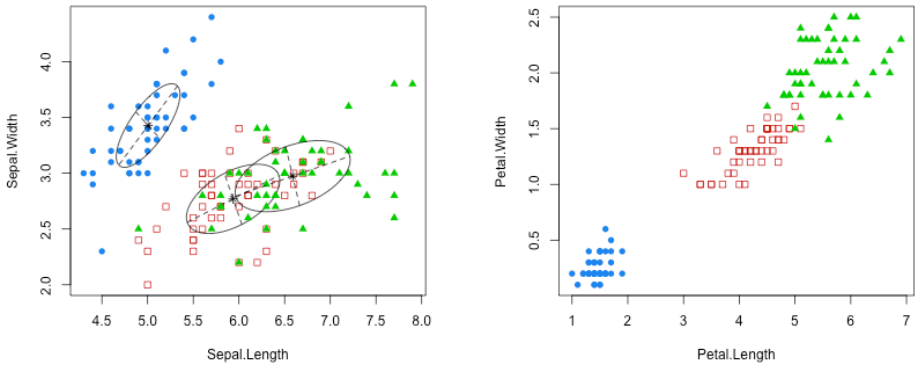


Casos de Uso – Sistemas de Recomendación

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.

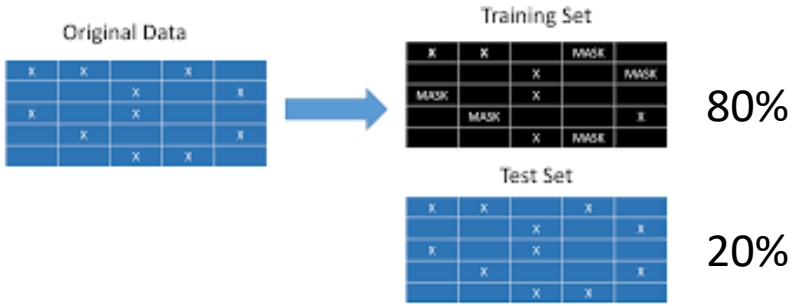


Métodos basados en distribuciones:
Considera que las observaciones proceden de una distribución (normal multivariante). En principio, cada clúster puede estar descrito por cualquier función de densidad, pero normalmente se asume que siguen una distribución multivariante normal.



Casos de Uso – Sistemas de Recomendación

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

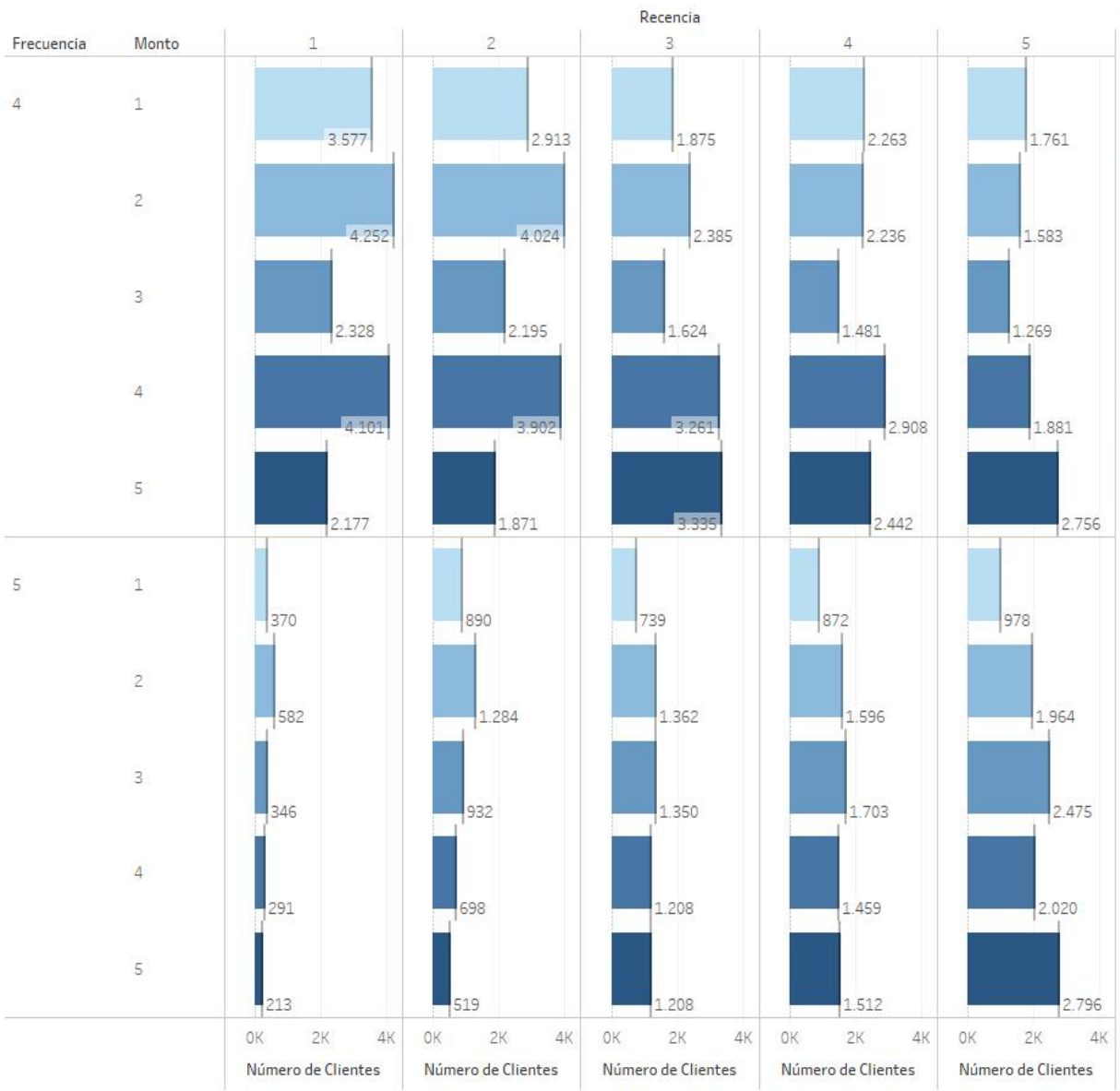
Casos de Uso – Propensión de Uso

Pregunta de Negocio

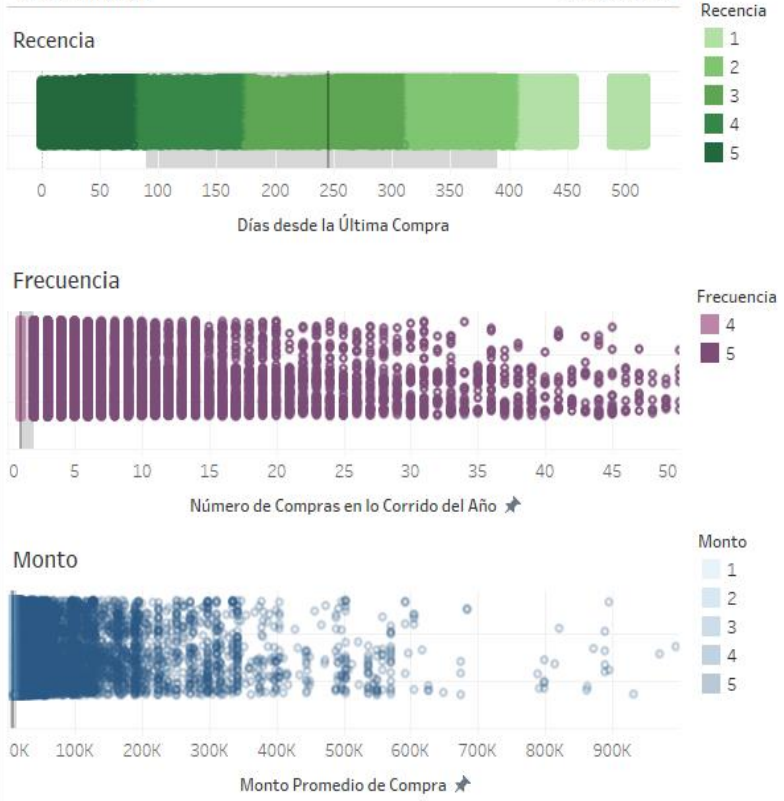
¿Cuáles serían los clientes con mayor propensión a consumir los de productos o servicios en el dado y asimismo adquirir una membresía?

Casos de Uso – Propensión de Uso

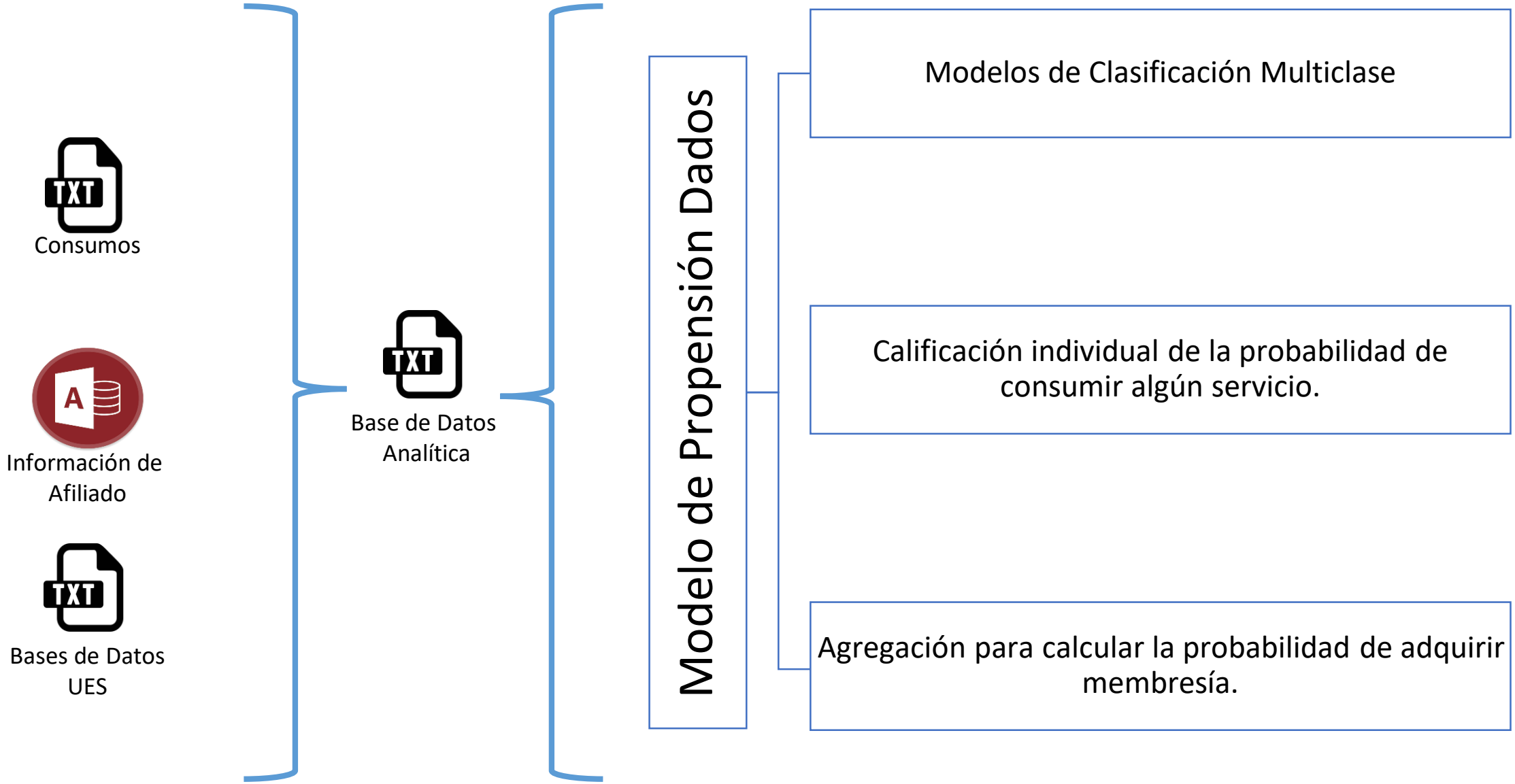
RFM - Análisis Total



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



Casos de Uso – Propensión de Uso



Casos de Uso – Propensión de Uso

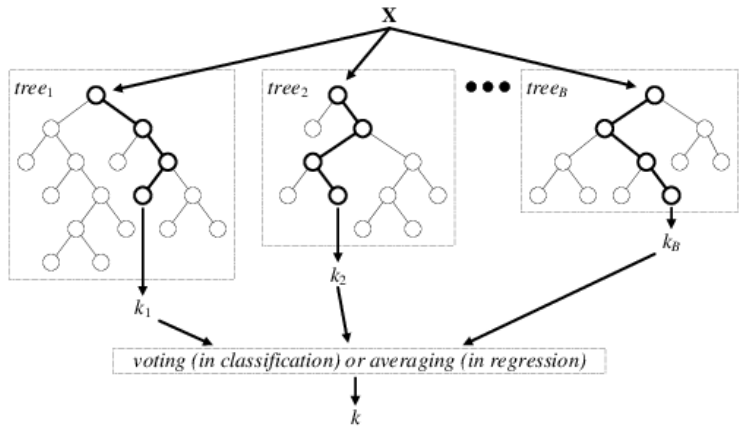
Modelo de Clasificación:

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



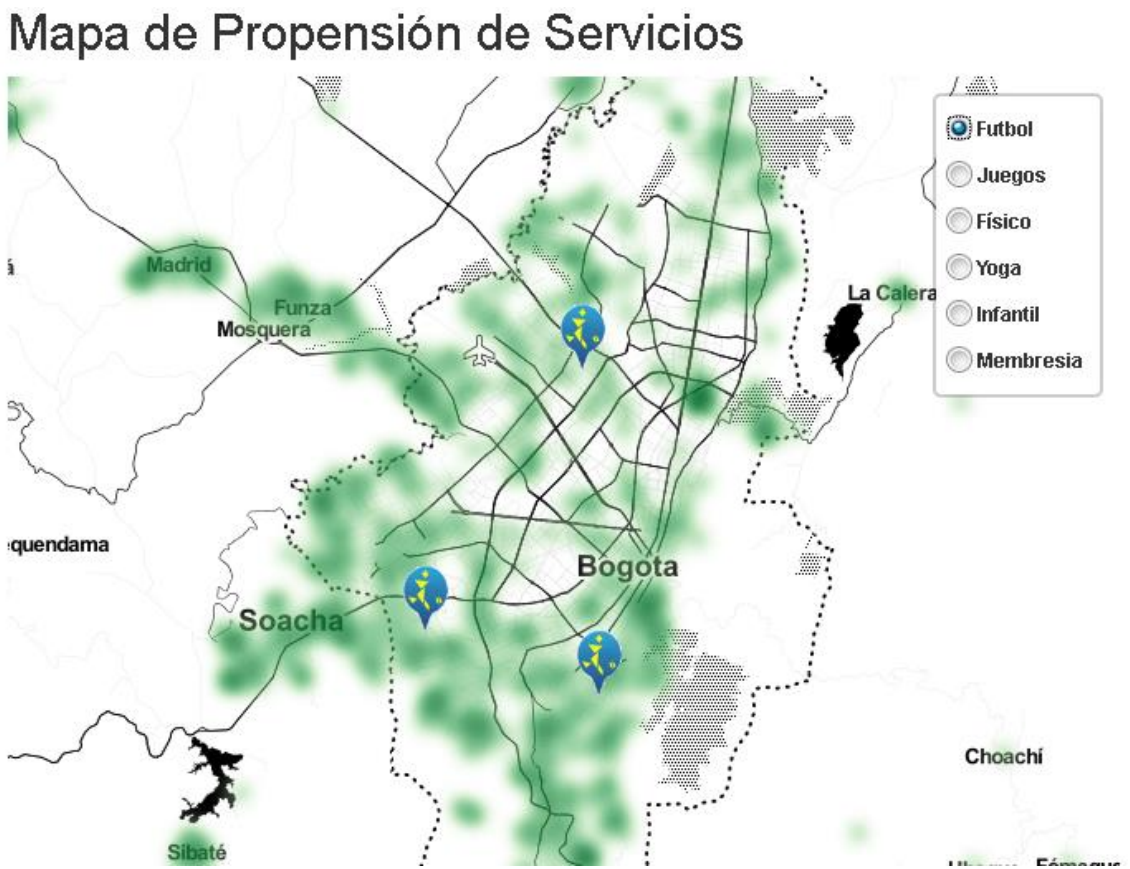
Se utiliza una metodología OnevsAll con un Random Forest para estimar la probabilidad individual de cada servicio dados los consumos pasados y las demás covariables de la base

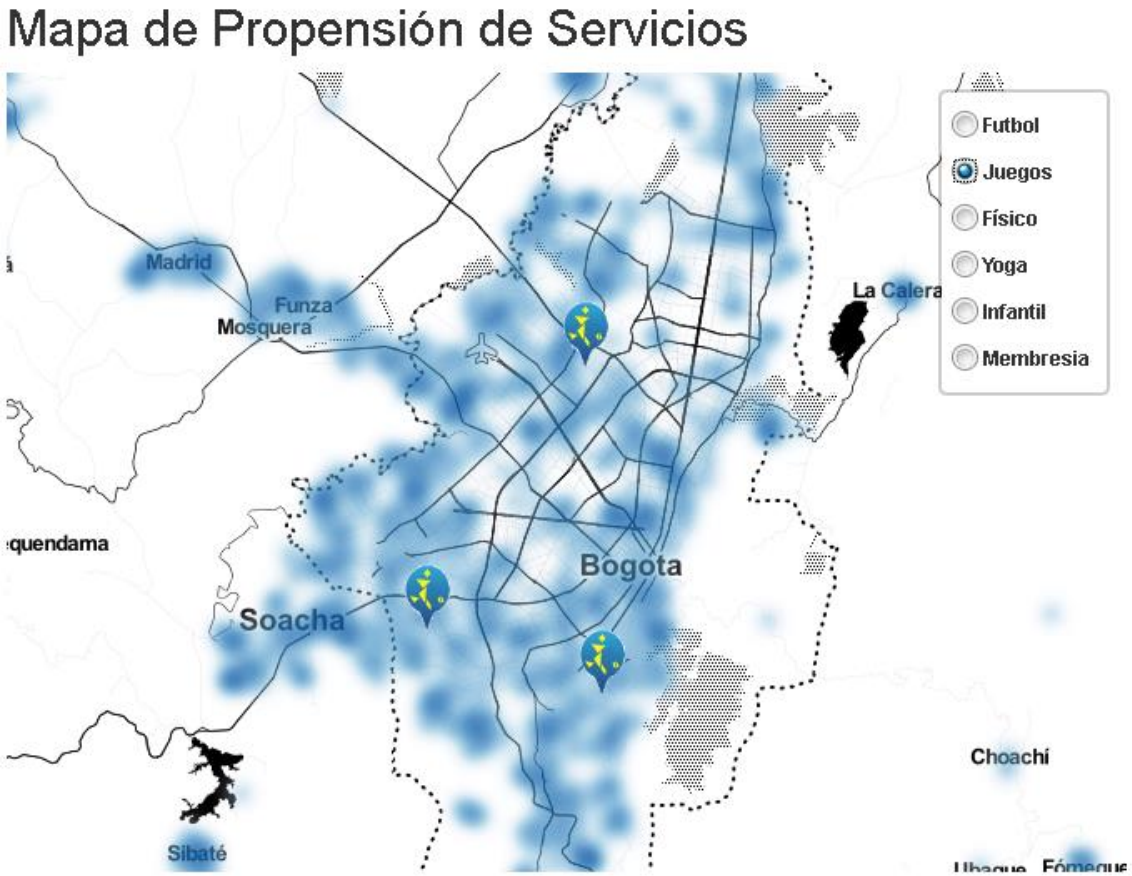
```
OneVsRestClassifier(estimator=RandomForestClassifier(bootstrap=True, class_weight='balanced',
            criterion='gini', max_depth=20, max_features='auto',
            max_leaf_nodes=None, min_impurity_decrease=0.0,
            min_impurity_split=None, min_samples_leaf=1,
            min_samples_split=2, 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

Casos de Uso – Propensión de Uso





semana	TEMA	ACTIVIDADES DE APRENDIZAJE		
		ACOMPAÑAMIENTO DEL DOCENTE		TRABAJO INDEPENDIENTE
		TEORÍA	PRÁCTICA	
1	Presentación del curso	• Lectura de contenido programático • Acuerdos		NA
2,3,4,5	Capítulo 1 Conceptos básicos	• Tipos de datos • Loops • Estructura de datos • Funciones. • Estadística descriptiva, tablas. • Gráficos • Introducción a paquetes		TALLER
6	PRIMER PARCIAL			
7,8	Capítulo 2 Ficheros, limpieza y descripción	• Lectura de ficheros • Análisis de ficheros (Estructura) • Data Wrangling • Análisis descriptivo (Descripción) • Visualización	SEGUNDA ENTREGA	TALLER
9,10	Capítulo 3 Segmentación	• K-Means • KNN		
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				

7 de Septiembre

12 de Octubre → PreProyecto

19 de Octubre → Parcial 2

16 de Noviembre→ Proyecto

23 de Noviembre→ Examen Final

Evaluación


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