ESTUDIO DEL CONJUNTO DE DATOS NHANES MEDIANTE EL EMPLEO DE TÉCNICAS DE APRENDIZAJE NO SUPERVISADO.

Raúl Sánchez Temporal
Máster universitario en Ciencia de Datos
Área: Aprendizaje automático en medicina

Universitat Oberta de Catalunya

CONTENIDO

1. Contexto, Motivación y objetivos
2. Metodología y planificación
3. Diseño e implementación de los modelos de agrupamiento
4. Interfaz web
5. Conclusiones y trabajo futuro

1. Contexto, motivación y objetivos

Universitat Oberta de Catalunya

Programa de estudios diseñado para evaluar el estado de salud de niños y adultos en Estados Unidos

nhanes National Health and Nutrition Examination Survey

1960
Desde

+5000
Entrevistas anuales

Datos publicados
Factores de riegos
Hábitos de salud
Constitución
Factores hereditarios
Hábitos dietéticos
...

Oportunidades de investigación
Múltiples trabajos publicados se basan en los datos NHANES

Beneficios para la sociedad
Son muchos e importantes los resultados obtenidos en base a estos datos para la mejora de la saludos de la gente.

*** Imágenes, logos e información propiedad de CDC**

1. Contexto, motivación y objetivos

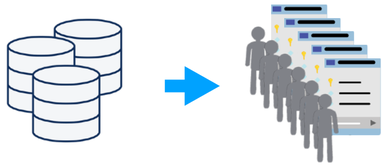
Universitat Oberta de Catalunya

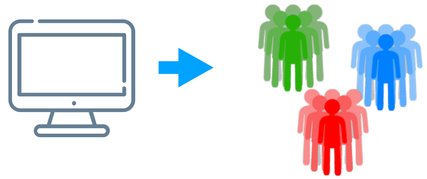
Objetivos

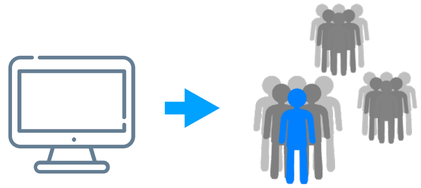
Conjunto de datos a partir de NHANES mediante el paquete estadístico SAS

Preparar los datos y los **modelos de agrupamiento** para perfilar los individuos

Aplicación WEB para la clasificación de individuos en base a sus características







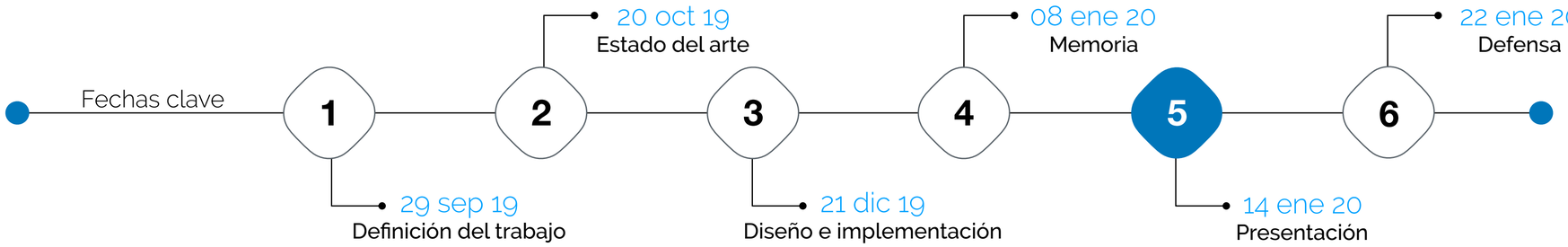
Productos

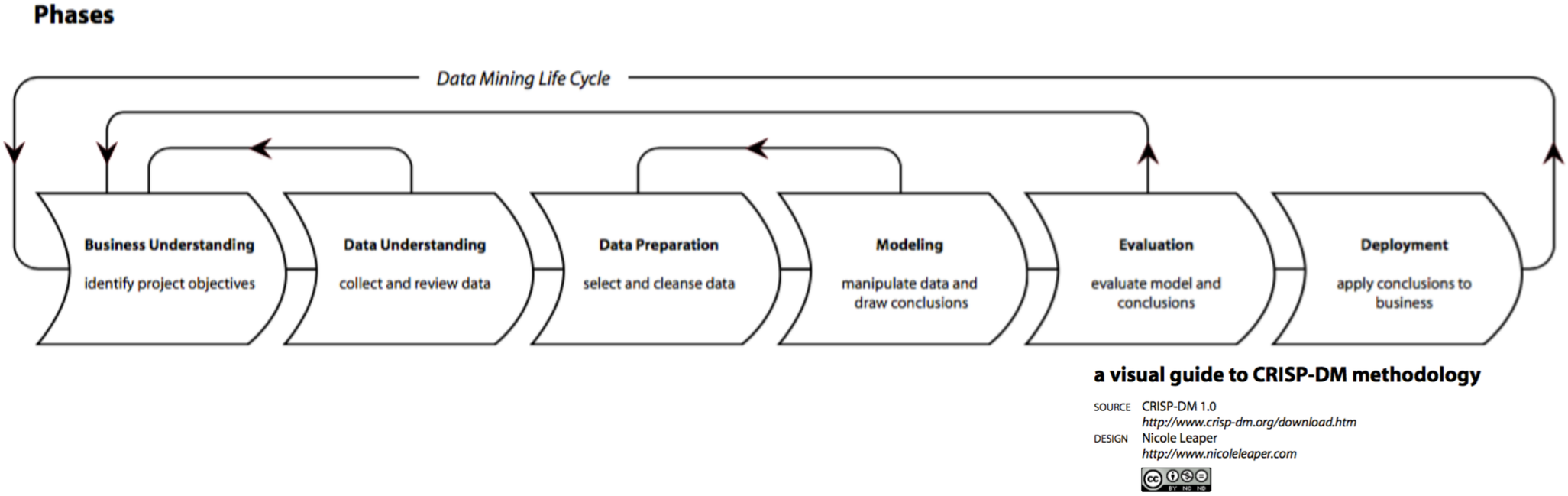
Data set
Selección de observaciones y sus características.

Clustering
Definición de grupos de individuos similares entre ellos y diferentes al resto de grupos

Clasificación
Según las características del individuo asignar a grupo con características más similares

2. Metodología y planificación





3. Diseño e implementación de los modelos de agrupamiento

Recopilación de datos

SEQN	Id
BMXBMI	Índice de masa corporal
BMXWAIST	Waist Circumference (cm)
BPXDI1	Presión en sangre
DIQ010	Diagnóstico de diabetes
LBXGLU	Nivel de glucosa
LBXIN	Nivel de insulina
HIQ011	Seguro de salud
HOD050	N habitaciones en casa
HOQ065	Propiedad de la vivienda
HSD010	Estado de salud
INQ020	¿Tiene ingresos?
MCQ010	¿Tiene asma?
MCQ300C	¿Pariente con diabetes?
PAQ650	Actividad física intensa
PAQ665	Actividad física moderada
PAQ710	Horas uso Televisión
PAQ715	Horas uso ordenador
RHQ160	N embarazos
SLD012	Horas de sueño
SMQ040	Do you now smoke cigarettes
RIAGENDR	Género
RIDAGEYR	Edad
RIDRETH3	Raza
DMDEDUC2	Nivel educación - Adults 20+

Ciclos seleccionados

2011-2012	2013-2014	2015-2016
9756	10175	9971

29902

Observaciones

Componentes

Datos demográficos

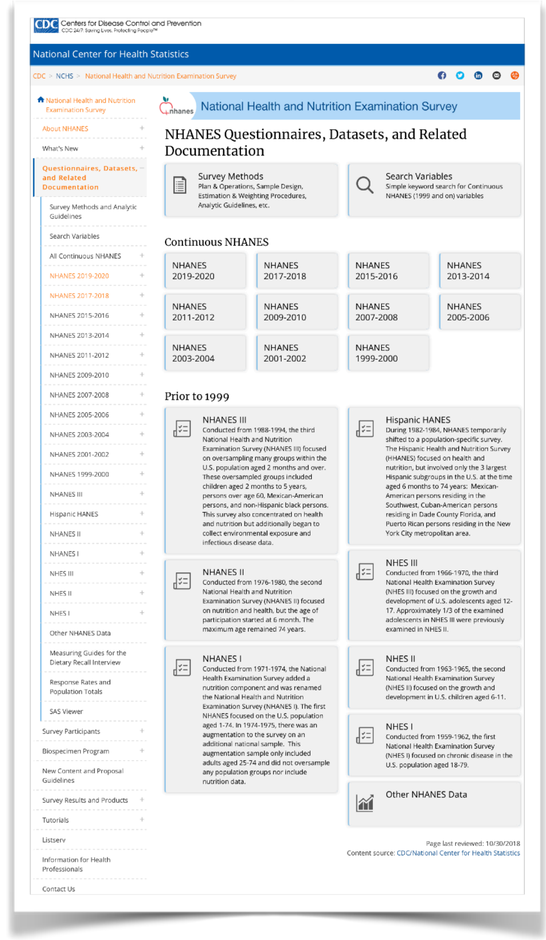
Datos Dietéticos

Datos de Exámen físico

Datos de cuestionario

25

Variables



web NHANES: https://wwwn.cdc.gov/nchs/nhanes/Default.aspx

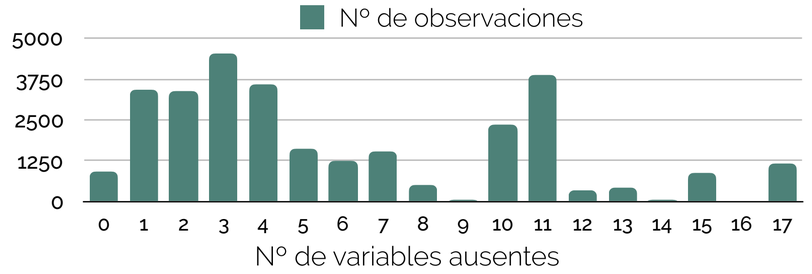
3. Diseño e implementación de los modelos de agrupamiento

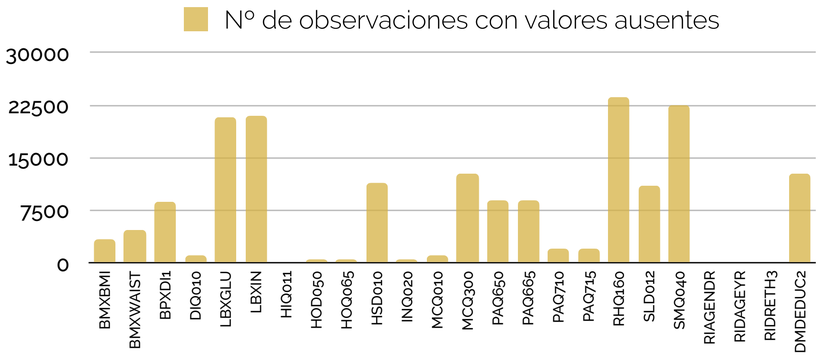
Preparación de los datos recopilados

Universitat Oberta de Catalunya

Características del conjunto recopilado:

- Existe un gran número de valores ausentes.
- Existen valores comodín del tipo 7, 9, 77, 99 ó 777 y 999
- Se detecta presencia de outliers en los datos.





Tareas de preparación y transformación aplicadas:

- Eliminación de observaciones con gran nº variables ausentes.
- Imputaciones por la media, por valor o por valor más frecuente.
- Detección de outliers en los datos.
- Eliminación de outliers (Método Z-score)
- Recodificación de variables.
- Escalado

Variables imputadas

RHQ160, SLD012, INQ020, HOD050, HOQ065

Variables afectadas con presencia outliers

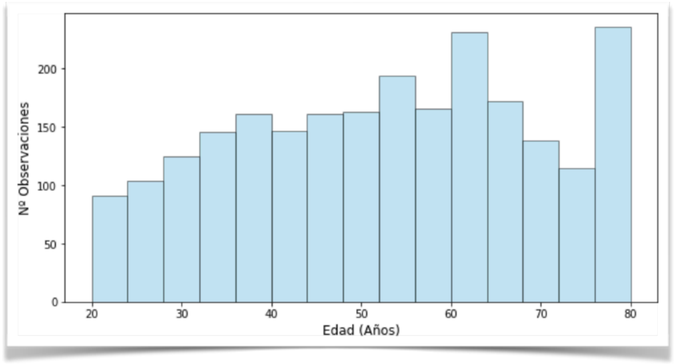
BPXDI1, LBXIN, LBXGLU, BMI

Resultado:

- Conjunto final de datos con 2350 observaciones y 25 variables.

3. Diseño e implementación de los modelos de agrupamiento

Entendimiento de los datos



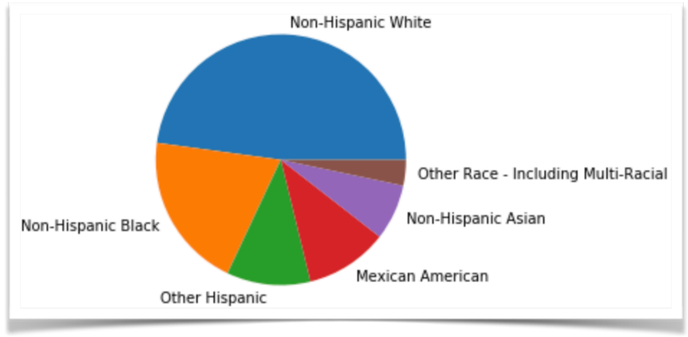
Mujeres
35.3%

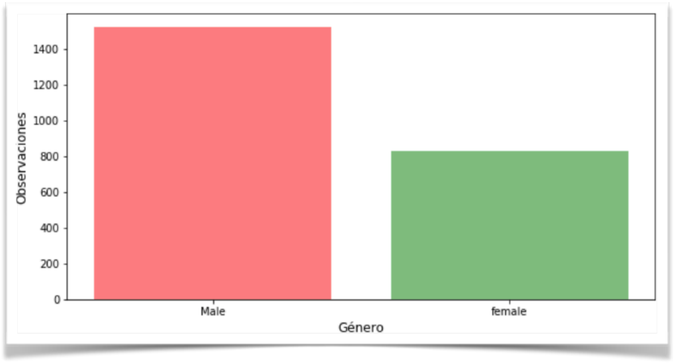
Hombres
64.7%

Edad media
52.28 años

Raza blanca
47.8%

No fumadores
55%





3. Diseño e implementación de los modelos de agrupamiento

Entendimiento de los datos

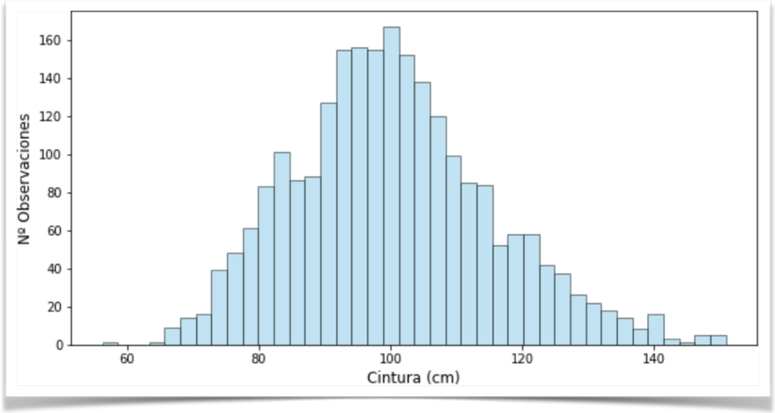
Media BMI

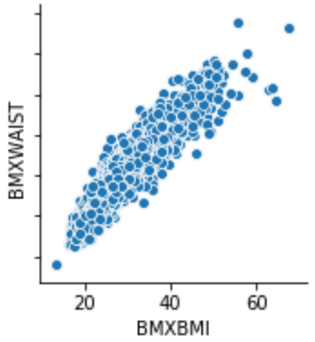
28.62

Grupo > 40 Años

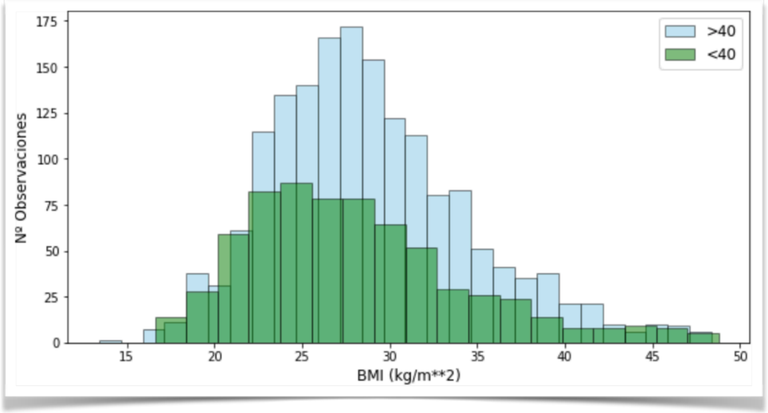
Mayores índices de BMI

Diámetro cintura



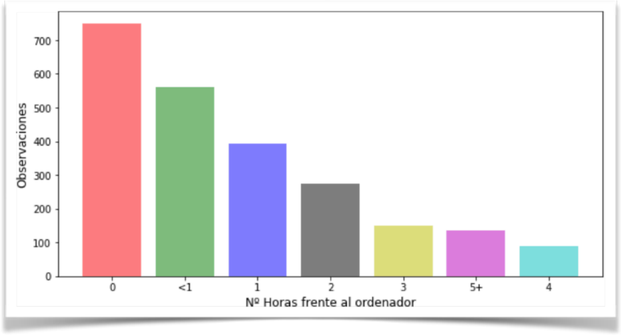


Índice de masa corporal

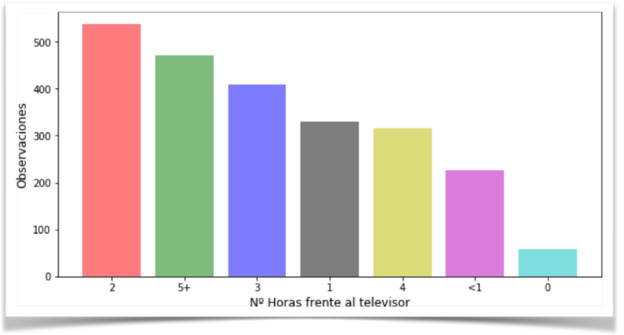


3. Diseño e implementación de los modelos de agrupamiento

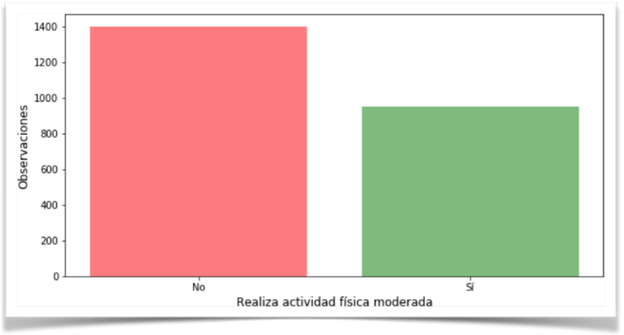
Entendimiento de los datos



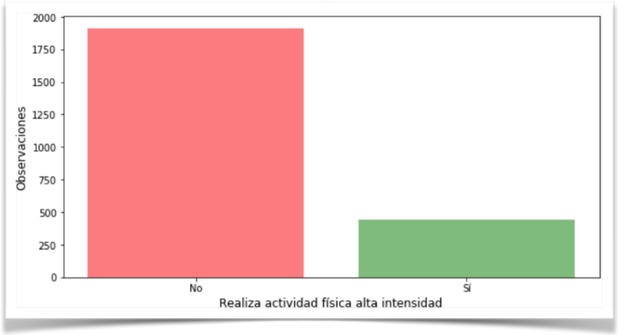
Media uso TV
2.71 horas/día



Media uso ordenador
1.14 horas/día



Actividad física moderada
40.5%

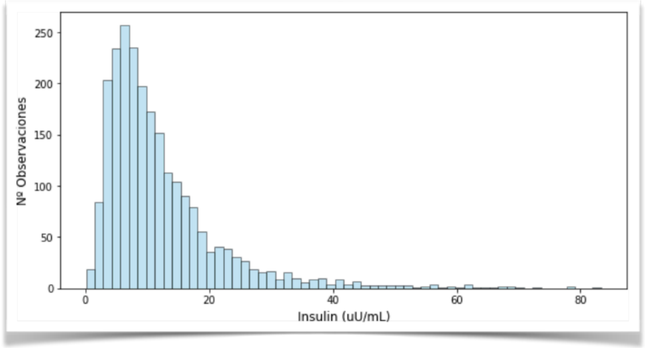


Actividad física intensa
18.7%

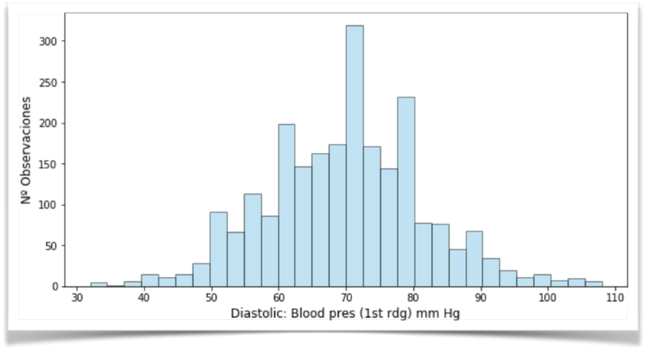
3. Diseño e implementación de los modelos de agrupamiento

Entendimiento de los datos

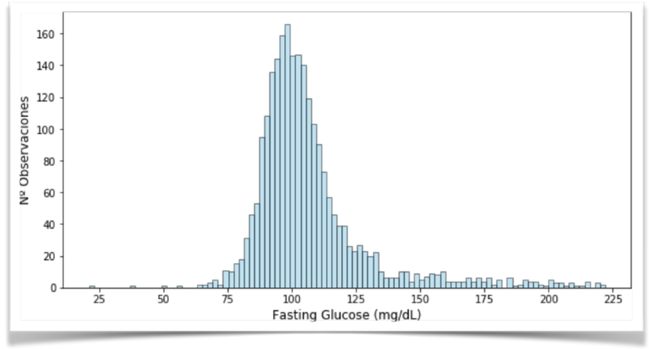
Media Insulina
12.36uU/ml



Media Glucosa (Ayunas)
106.89 mg/dL

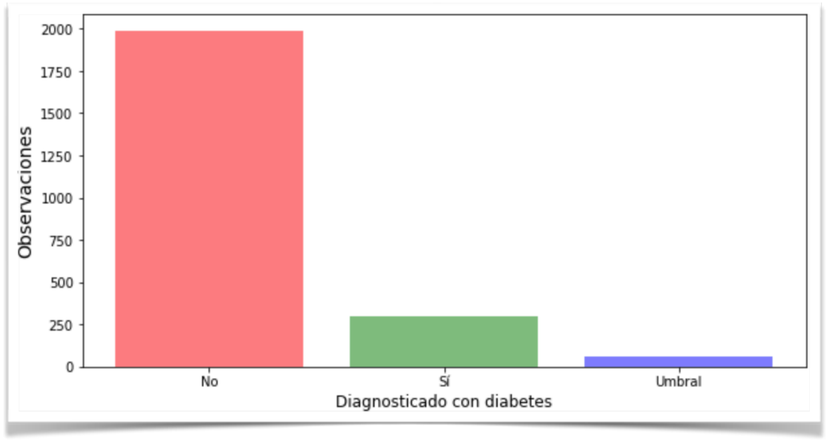


Media presión en sangre
69.78 mmHg

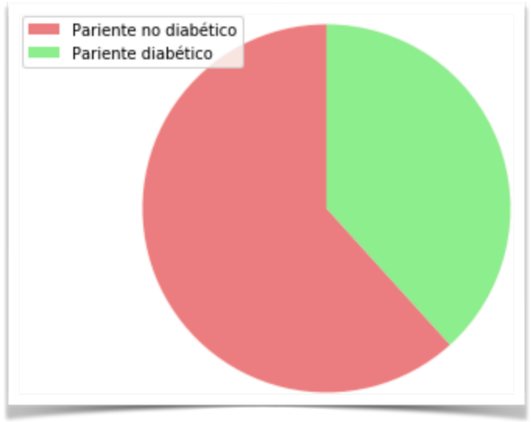


3. Diseño e implementación de los modelos de agrupamiento

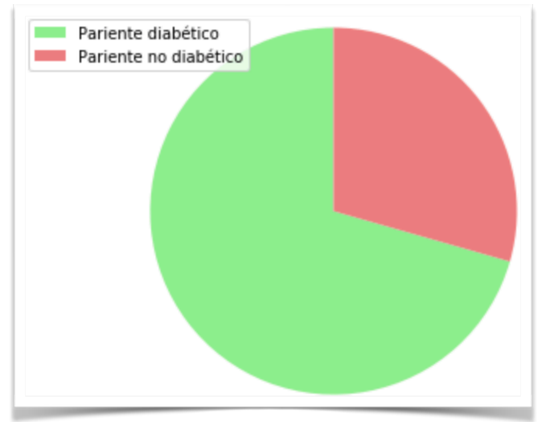
Entendimiento de los datos



Individuos no diabéticos



Individuos diabéticos



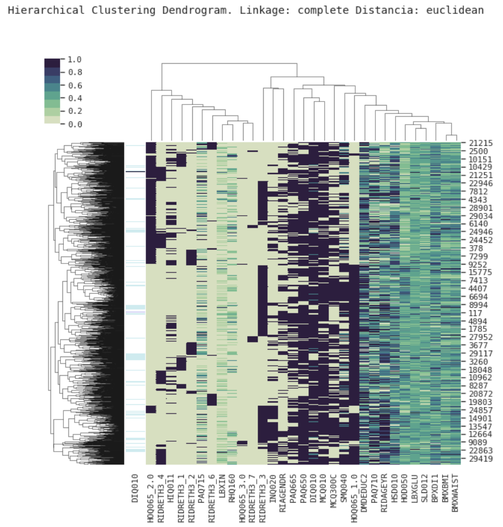
Individuos diagnosticados con diabetes 12.7%

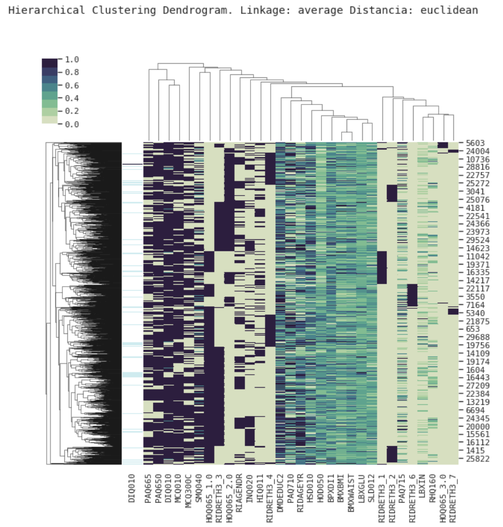
Un 70.5% de los diabéticos tienen un pariente cercano diabético

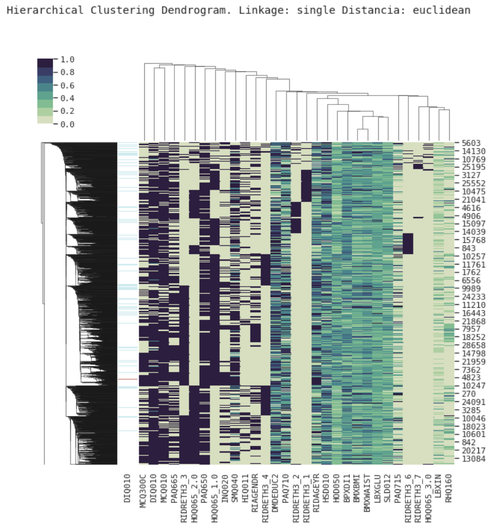
Un 38.3% de los no diabéticos tienen un pariente cercano diabético

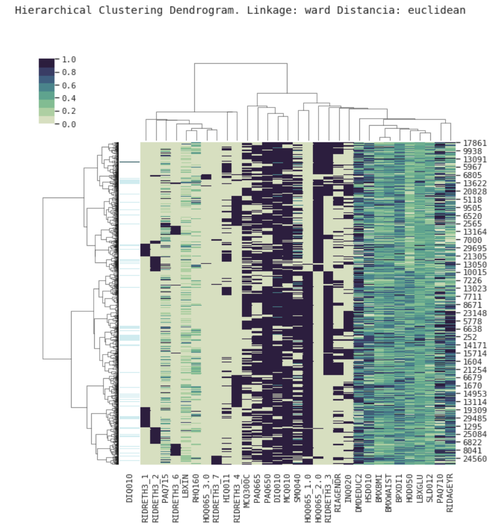
3. Diseño e implementación de los modelos de agrupamiento

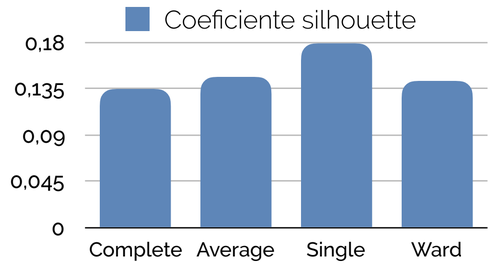
Modelo HAC

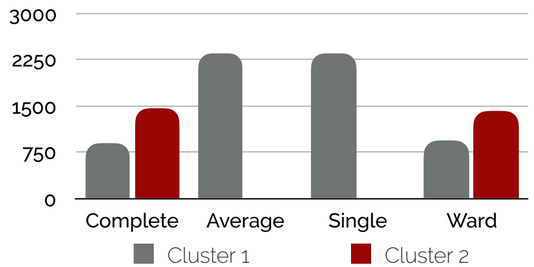


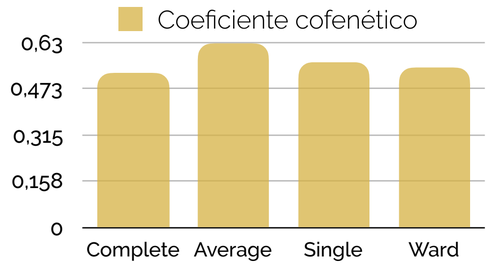












Nº óptimo de clústeres **2** en todos los métodos

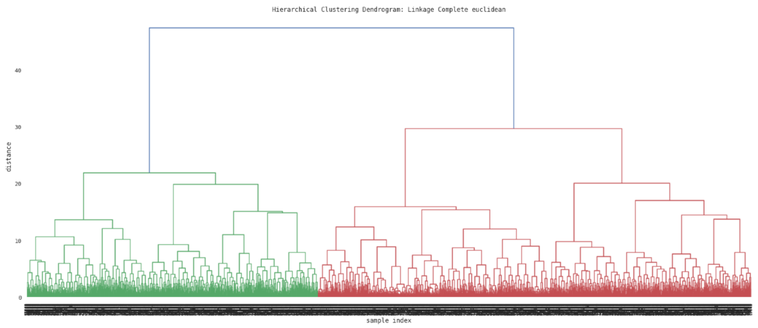
3. Diseño e implementación de los modelos de agrupamiento

Modelo HAC

Universitat Oberta de Catalunya

Método: Ward
Distancia: Euclidean

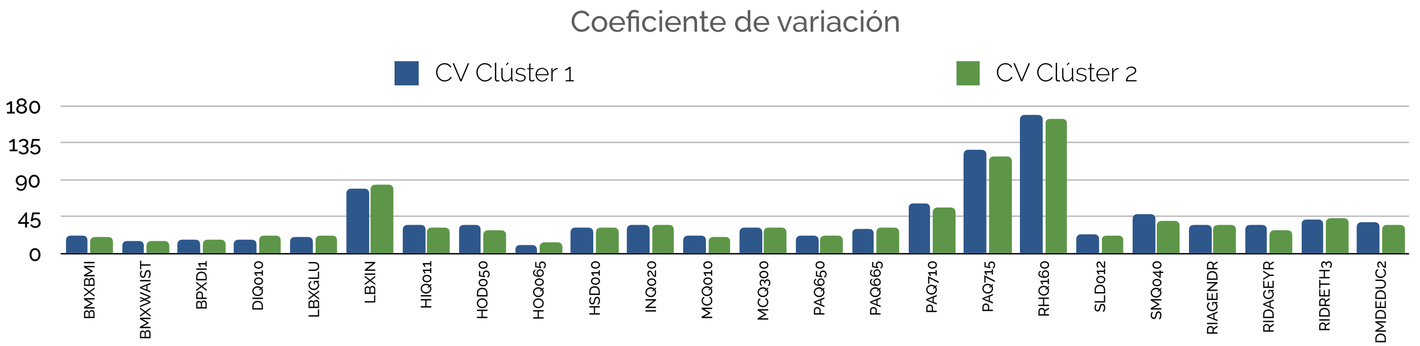
Nº de clústeres: 2
Clúster 1: 946
Clúster 2: 1404



Descripción de los clústeres:

Clúster 1	Clúster 2
Vivienda en alquiler (94.5%)	Vivienda en propiedad (98%)
Media de habitaciones (4.5)	Media de habitaciones (6.5)
Media de edad (46)	Media de edad (56)
Mayor uso de ordenador	Mayor uso de televisión
Más actividad deportiva intensa	Más actividad deportiva moderada

Características de los 2 grupos muy similares

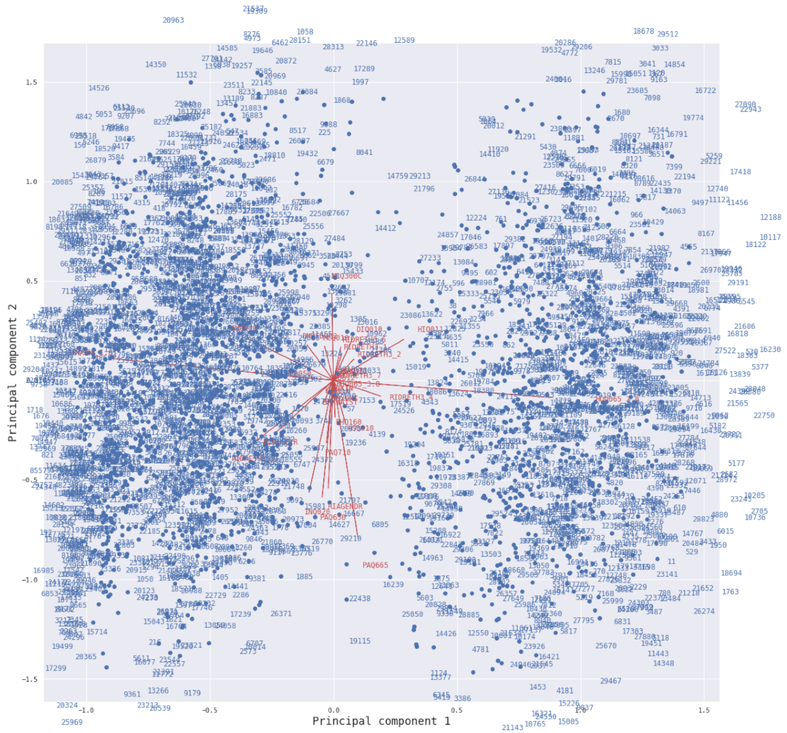


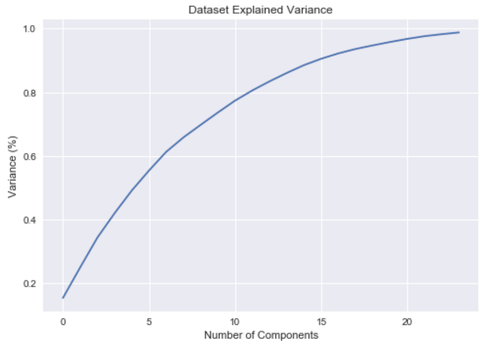
3. Diseño e implementación de los modelos de agrupamiento

Modelo HAC

Universitat Oberta de Catalunya

Análisis de componentes principales (PCA):






```
pca.explained_variance_ratio_
```

```
array([0.15449265, 0.09571529, 0.0935799 , 0.07666424, 0.07129137,
       0.06267205, 0.05882476, 0.0451747 , 0.03938747, 0.03882761,
       0.03733706, 0.03188885, 0.02853565, 0.0263318 , 0.02456721,
       0.0200701 , 0.01679431, 0.01387804, 0.01123066, 0.01067149,
       0.00975786, 0.00849294, 0.00622987, 0.00548537])
```

Varianza explicada por los dos principales componentes : 25%

Variable más influyente

HOQ065 (Tipo de propiedad vivienda)

3. Diseño e implementación de los modelos de agrupamiento

Modelo DBSCAN

Parametrización

radio épsilon (ε) = 7

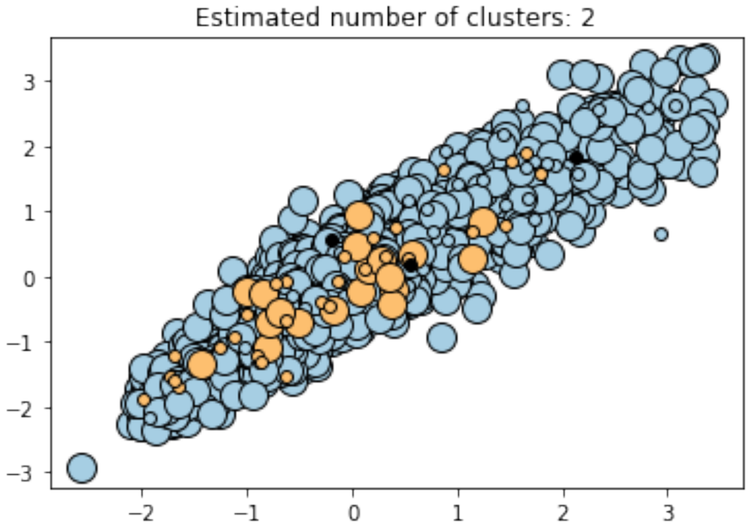
minPts = 30

Nº de clústeres: 2

Clúster 1: 2229

Clúster 2: 48

Outliers: 3



Coeficiente Silhouette : 0.254

No se han encontrado estructuras basadas en densidad

3. Diseño e implementación de los modelos de agrupamiento

Evaluación

No existe conocimiento previo

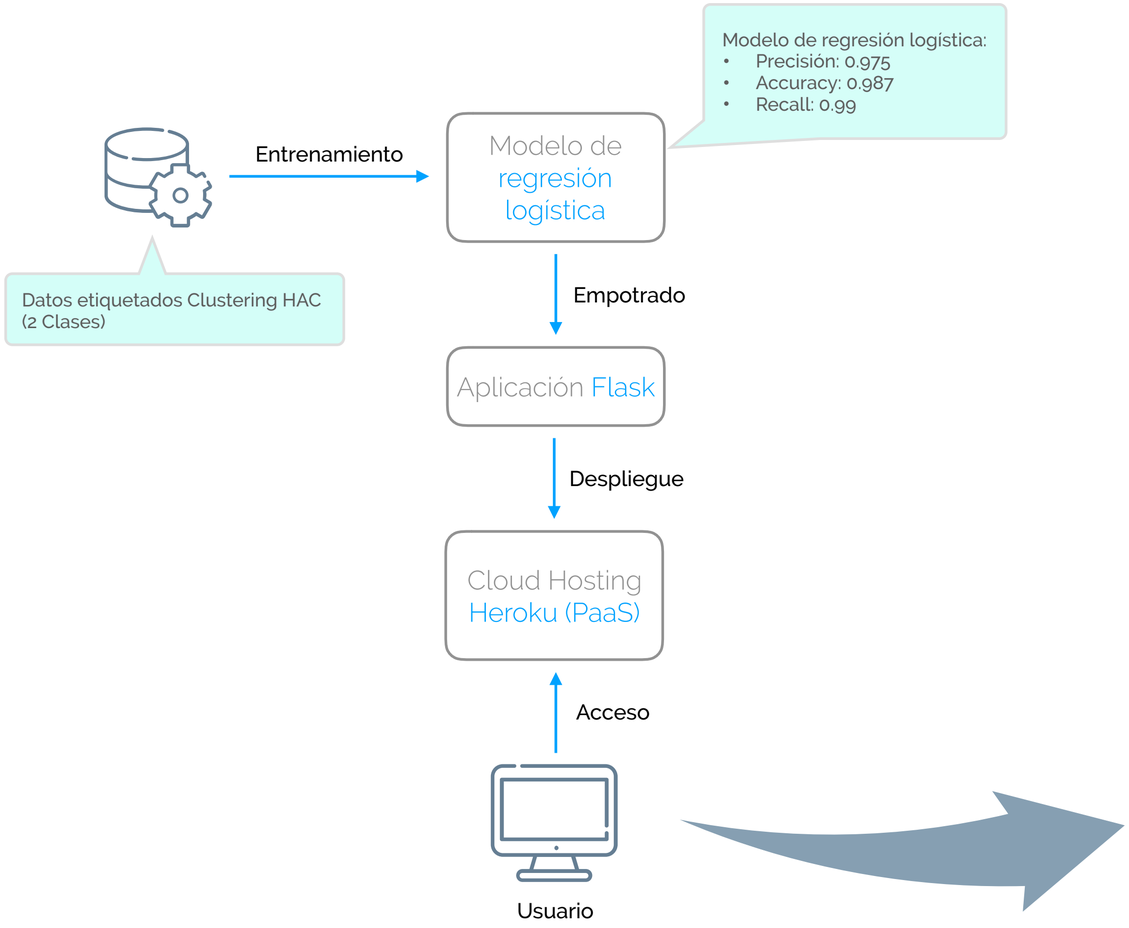
Los índices de calidad indican agrupaciones pobres

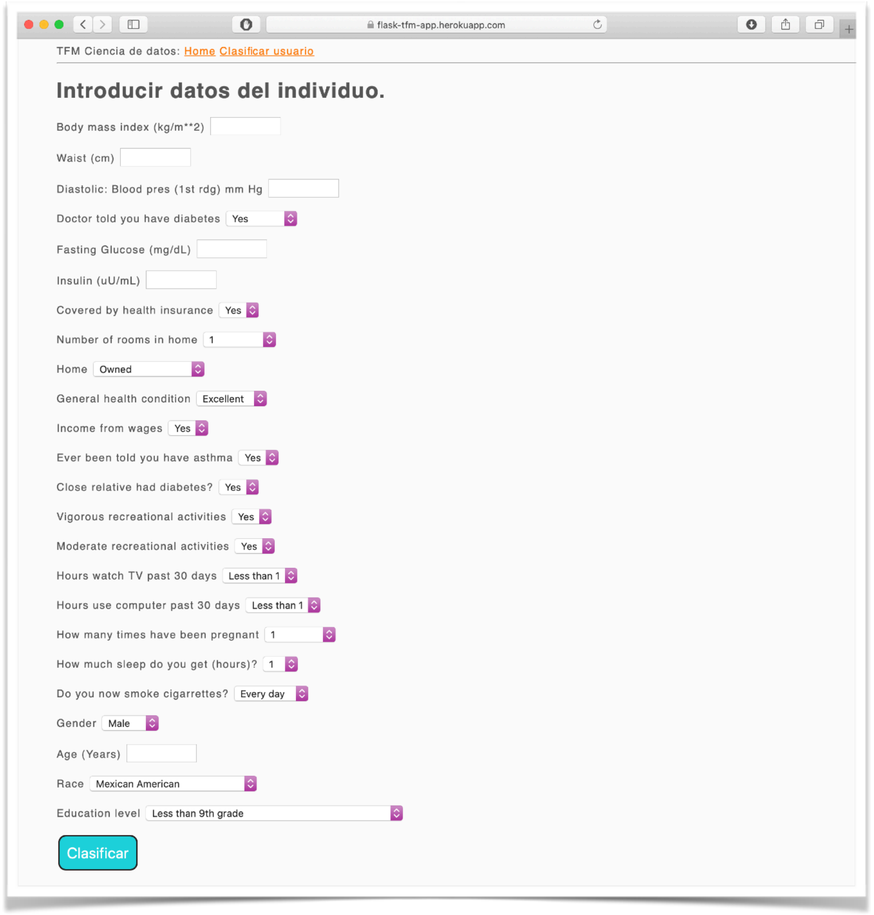
¿Es significante la diferencia entre los grupos creados?

¿Hay básicamente un grupo?

Davies-Bouldin criteria (maximum interclass-intraclass distance ratio)		
Modelo	Nº grupos	Valor
HAC Ward	2	2.45
Hac Ward	3	2.70
DBSCAN	2	1.79

4. Aplicación Web



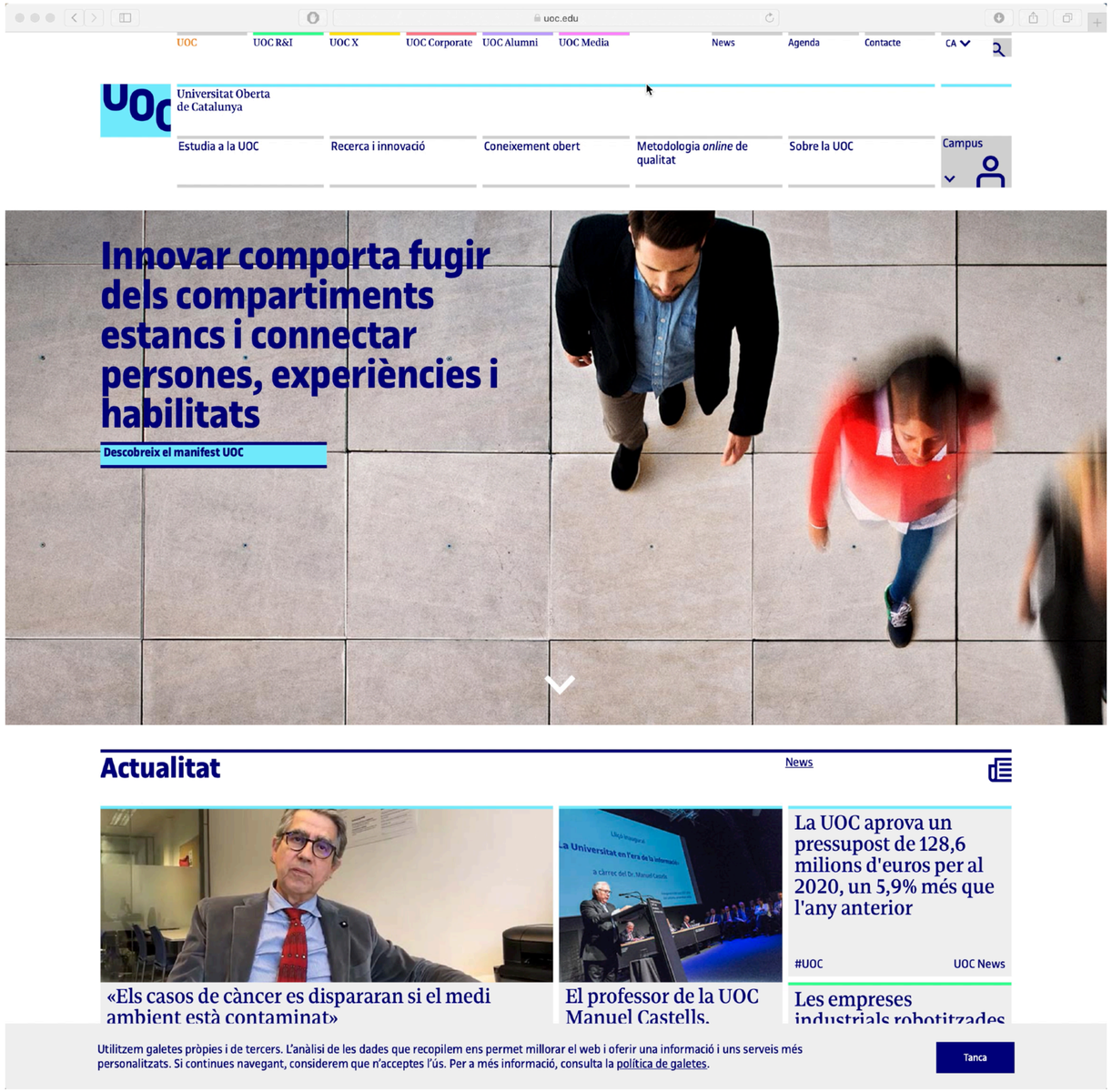


Aplicación: https://flask-tfm-app.herokuapp.com

4. Aplicación Web

Demo

Universitat Oberta de Catalunya



Aplicación: https://flask-tfm-app.herokuapp.com

5. Conclusiones y trabajo futuro

Universitat Oberta de Catalunya

Conclusiones

- Importancia del conocimiento del dominio.
- Escasa estructura subyacente en los datos.
- Perfiles de individuos similares.

Trabajo futuro

- Una muestra mayor de observaciones.
- Mejora en la selección de variables inicial que explique mejor nuestro problema.
- Incluir almacenamiento de individuos que se clasifican en aplicación.

Muchas gracias por su atención.

Raúl Sánchez
rsancheztem@uoc.edu