Universidad de Valladolid

ESCUELA DE INGENIERÍA INFORMÁTICA DE VALLADOLID

Master en Ingeniería Informática

Uso de técnicas de Data Mining sobre series temporales obtenidas por simulación y aplicación de resultados en videojuego Crossroads

---

Alumno: Adrián Manzano Santos

Tutor: David Escudero Mancebo

Resumen

Aunque el cambio climático es un desafío actual de una magnitud impredecible, existen modelos de simulación que permiten obtener previsiones a largo plazo de la evolución del clima y la economía en función de las decisiones que adoptemos. Estos simuladores permiten obtener abundante información sobre la que se pueden aplicar técnicas de *data mining* para conocer la relación entre las políticas y sus efectos. Crossroads es un videojuego que permite explotar estas evidencias. Este trabajo se centra en la clasificación de series temporales, resultado de un proceso de simulación, para el reconocimiento de patrones característicos implícitos. Esto se desarrolla bajo un enfoque de doble clasificación: en primer lugar, se agrupan las series que son similares entre sí para reducir el cuerpo de datos, y sobre el cuerpo reducido se extraen los patrones implícitos. El trabajo se desarrolla en el contexto del videojuego educativo Crossroads, cuya salida es el resultado de simulación, con el fin último de desarrollar un proceso autónomo de recomendaciones. Para ello, se presenta un estudio basado en la información mutua que trata de evaluar la influencia que tienen las entradas, determinadas por el jugador, sobre el patrón característico asignado. Esto permite realizar una descripción experta del patrón. Por otro lado, tras clasificar los patrones como "buenos" o "malos", es posible guiar al jugador aconsejando cambios en la entrada para obtener un resultado mejor.

Palabras clave: simulación, clustering de series temporales multivariables, feedback en videojuego educativo, cambio climático.

Abstract

Although climate change is a current challenge of unpredictable magnitude, there are simulation models that allow us to obtain long-term forecasts of the evolution of the climate and the economy based on the decisions we make. These simulators provide a wealth of information on which to apply data mining techniques to understand the relationship between policies and their effects. Crossroads is a video game that allows us to exploit this evidence. This work focuses on the classification of time series, resulting from a simulation process, for the recognition of implicit characteristic patterns. This is developed under a double classification approach: first of all, series that are similar to each other are grouped to reduce the data corpus, and, over the reduced data, the implicit patterns are extracted. The work is developed in the context of the educational video game Crossroads, whose output is the simulation result, with the ultimate goal of developing an autonomous process of recommendations. For this purpose, a study based on mutual information is presented, which tries to evaluate the influence that the inputs, determined by the player, have on the assigned characteristic pattern. This allows for an expert description of the pattern.On the other hand, after classifying the patterns as "good" or "bad", it is possible to guide the player by advising changes in his input to obtain a better result.

Key words: simulation, multivariate time series clustering, feedback in educational video game, climate change.

Índice general

Capítulo 1

Introducción

Actualmente, la especie humana y el mundo se encuentra en un punto de inflexión provocado por el calentamiento global y el colapso de los ecosistemas. Mientras que la sociedad civil y los representantes políticos son los responsables de buscar y liderar un cambio necesario y una transición a un futuro sostenible con cero emisiones netas, LOCOMOTION[1] contribuye a este empeño desarrollando sofisticados modelos para evaluar el impacto socioeconómico y medioambiental de las distintas opciones políticas con el fin de ayudar a la sociedad a tomar decisiones informadas sobre la transición a un futuro sostenible bajas emisiones.

El presente trabajo se engloba dentro del proyecto LOCOMOTION y se enfoca en el reconocimiento de patrones característicos sobre un conjunto de series temporales, resultados del proceso de simulación basado en los modelos de proyecto. En este ámbito, y con este objetivo, las técnicas de clústering son predominantes, basadas en la agrupación de series similares.

En este primer capítulo se introduce y contextualiza el problema a abordar así como la motivación para llevar a cabo este trabajo. También se presentan los objetivos que se desean alcanzar, y el plan de trabajo desarrollado para conseguirlos.

1.1. Contexto: Medeas y Crossroads

Se dispone de un simulador, basado en el modelo **MEDEAS**[2] (acrónimo del inglés de "modelizando la transición energética renovable en Europa"), que predice el comportamiento de determinados indicadores socioeconómicos (como el PIB *per Cápita* global o el Ïndice de Desarrollo Humano) y medioambientales (incremento de la temperatura media global y CO_2 respecto al valor preindustrial) relacionados con el cambio climático y el consumo de recursos. Este modelo está dominado por más de 5000 variables de entrada, que pueden ser continuas y no acotadas.

[1] Sitio web LOCOMOTION: https://www.locomotion-h2020.eu/
[2] Sitio web MEDEAS: https://www.medeas.eu/

El simulador basado en MEDEAS tiene la finalidad de alimentar un un juego educativo de concienciación medioambiental, denominado ***Crossroads***. El objetivo del videojuego es mostrar las consecuencias que diferentes decisiones políticas tienen sobre el medioambiente y el cambio climático, así como buscar la transición de una sociedad dependiente de energías fósiles a una sociedad basada en las energías renovables de cero emisiones netas. Cabe destacar que debido al tiempo que requiere cada simulación (de unos diez segundos), se utilizan datos pregrabados, evitando que el simulador forme parte del motor del juego.

La mecánica básica del juego consiste en que cada jugador se fija unos objetivos en términos medioambientales (incremento de la temperatura en el año 2100 con respecto a los valores preindustriales) y económicos (PIB persona en dólares de 1995). Posteriormente determina una serie de hipótesis sobre la disponibilidad de recursos, impacto del cambio climático y evolución de la población. Finalmente especifica un conjunto de medidas políticas de dominio amplio con efecto en la economía y en el medioambiente.

A priori, las hipótesis definidas son fijas, pues describen la situación del entorno, y mediante las medidas políticas se busca conseguir una estabilización de los incrementos de temperatura en valores que no supongan un riesgo de extinción. La **Figura 1.1** muestra un boceto de la interfaz utilizada para seleccionar una medida política.

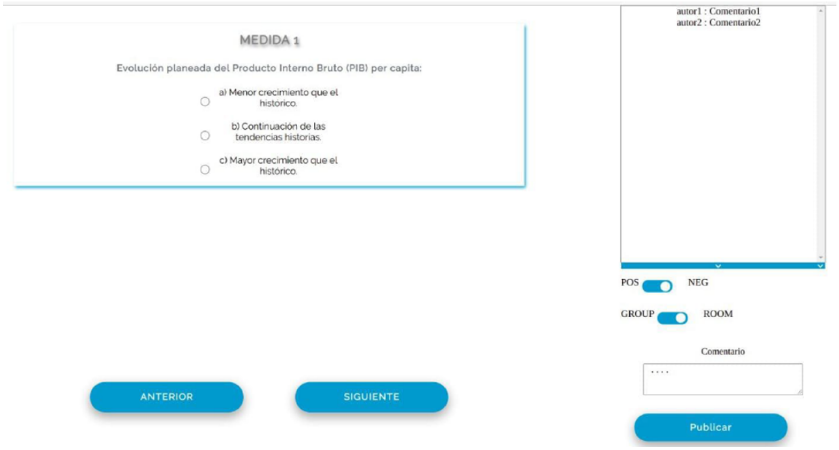


Figura 1.1: Boceto de Crossroads: selección de hipótesis y medidas. Fuente: [2]

La elección de las opciones para las hipótesis y medidas políticas no es arbitraria. En un trabajo previo[3], la complejidad de entradas que admite el modelo de simulación se ha simplificado a un conjunto de doce características discretas (hipótesis y medidas políticas). En consecuencia, se puede aplicar el modelo MEDEAS sobre las decisiones del jugador para generar un escenario futuro simulado.

[3]MEDEAS, Grupo de Economía, Energía y Dinámica de Sistema Universidad de Valladolid: https://geeds.es, http://www.eis.uva.es

El proceso de simulación produce la evolución temporal de numerosos indicadores socio-económicos y medioambientales (temperatura global, PIB mundial, CO_2, índice de desarrollo humano, etc.), comenzando en el año 1995 y finalizando en el año 2100. En lo que respecta al juego, los escenarios futuros quedan determinados por dos indicadores: incrementos de temperatura (medioambiental) y PIB *per Cápita* (económico).

En la **Figura 1.2** se presenta un boceto de la interfaz de visualización de resultados dados por el juego. Además de representar los indicadores simulados frente al objetivo, el juego da una valoración numérica de ciertos aspectos, tanto propios como derivados del escenario (**Figura 1.2**).

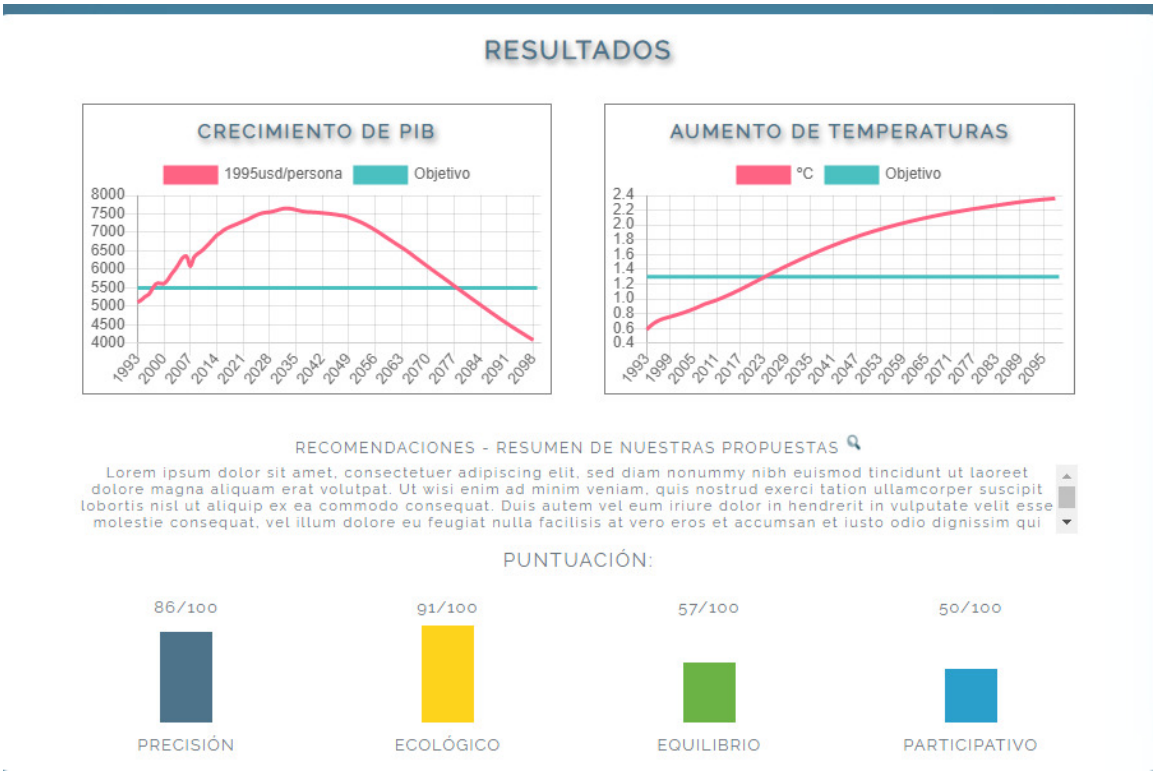


Figura 1.2: Captura de Crossroads: visualización de resultados

Por otro lado, el objetivo del juego es que los participantes comprendan que ha causado el obtener su escenario concreto y realicen cambios sobre las medidas (e hipótesis si fuese necesario) para alcanzar un escenario medioambiental y económicamente equilibrado. Se desea que el juego de una descripción del escenario y recomendará cambios a los usuarios. Esta información se presentará en la sección "RECOMENDACIONES - RESUMEN DE NUESTRAS PROPUESTAS" de la **Figura 1.2**, y será generada por el sistema autónomo de recomendaciones propuesto y desarrollado en el **Capítulo 4**.

En la **Tabla 1.1** se recogen el número de opciones que admite cada pregunta del cuestionario, lo que genera 437400 combinaciones diferentes de posibles respuesta completas al cuestionario. El cuestionario en detalle se recoge en el **Apéndice C**. El elevado número de escenarios y su imposibilidad de un analisis directo es lo que motiva este trabajo.

Cuestión	Núm. resp.	Cuestión	Núm. resp.
H1	3	H2	3
H3	5	M1	5
M2	2	M3	3
M4	3	M5	3
M6	2	M7	3
M8	2	M9	4

Tabla 1.1: Preguntas y número de respuestas al cuestionario

1.2. Motivación

El objetivo principal de este trabajo es construir el módulo de diagnosis y recomendación autónomo para el videojuego Crossroads. Aunque el cuestionario supone una importante reducción de las entradas admitidas en el juego, el estar compuesto de doce preguntas que admiten dos, tres, cuatro o cinco opciones posibles (**Apéndice A1**), supone que existen 437.400 posibles respuestas diferentes.

En consecuencia, el gran número de escenarios simulados imposibilita su estudio directo por expertos en la materia, así como realizar una descripción detallada de todos ellos. Además, el hecho de trabajar con datos de simulación presenta una peculiaridad: existen evoluciones de temperatura y PIB que son muy similares entre sí (el "ojo humano" las consideraría como iguales), pero que numéricamente difieren ligeramente. . Es más, analíticamente estas series representan un mismo escenario, pues valores finales muy similares tienen un mismo significado (obtener un incremento de temperatura valores en el año cercano al 2,25ºC es un buen resultado, pero incrementos de 5ºC suponen un riesgo de extinción). Esta idea se ilustra formalmente en la **Sección 3.4**.

Este proyecto se motiva en la incapacidad para trabajar directamente con el conjunto total de las simulaciones, así como en la necesidad de identificar como una unidad aquellos escenarios que presenten evoluciones de temperatura y PIB similares. Por tanto, se desea reducir el número de escenarios posibles a un conjunto pequeño y sencillo (denominado conjunto de **escenarios característicos**), para su posterior análisis por expertos en la materia. Además, se considera necesario conocer la contribución de las respuestas sobre el escenario característico resultante.

Por último, dado que en el juego el usuario se fija unos objetivos, se desea evaluar si el escenario resultante del cuestionario cumple con dicho objetivo. Además, en el conjunto de simulaciones se han observado que existen escenarios más beneficiosos y adecuados que otros:

1. Respecto a la temperatura, existen escenarios con valores de temperatura en el último año superiores a 3ºC, o en los que la temperatura no se llega a estabilizar en algún valor. En esto casos, existe un riesgo de extinción por lo que son poco deseables.

2. Respecto a la economía, existen escenarios en los que el PIB se hunde completamente, finalizando en valores inferiores al año inicial e incluso llegando a colapsar la economía mundial. También existen escenarios irreales en los que la economía crece a placer, dados por unas hipótesis nada restrictivas.

Por tanto, una vez que se dispone de un escenario desfavorable, se quiere dar una recomendación al jugador, en forma de cambios sobre sus medidas política e hipótesis, que le permitan alcanzar un escenario más favorable[4].

1.3. Objetivos

Este proyecto busca aportar un análisis de las simulaciones que complementen el videojuego de concienciación sobre el cambio climático, dando una base para la justificación de cada escenario. Con este fin, se identifican los siguientes objetivos (**Tabla 1.2**).

Id	Objetivo
Obj-01	Aplicar técnicas de análisis automático para ayudar a interpretar los escenarios simulados y relacionarlos con las decisiones tomadas.
Obj-01.1	Extracción y construcción, de forma autónoma, de los escenarios característicos, que representen el abanico general de posibles escenarios futuros.
Obj-01.2	Valorar la influencia de las respuestas dadas al cuestionario en la asignación de un escenario característico.
Obj-02	Construcción de un sistema de valoración y recomendaciones autónomo
Obj-02.1	Recopilar un análisis experto sobre los escenarios característicos, que incluya una descripción de este escenario, sus posibles causas y su idoneidad.
Obj-02.2	Construcción del sistema de valoración del cumplimiento de los objetivos del usuario en función del escenario obtenido como respuesta al cuestionario.
Obj-02.3	Construcción de un módulo de recomendación autónoma, que proponga diferentes cambios en sus hipótesis y medidas políticas para alcanzar un escenario considerado aceptable, a partir de las hipótesis y medidas que determinaron un escenario de simulación inaceptable.

Tabla 1.2: Objetivos

[4]En este contexto, se entiende por escenario favorable aquel en el que la temperatura se estabiliza en valores que no suponen riesgo de extinción, y la economía evita su colapso.

1.4. Propuesta de solución

Se propone un estudio de tres etapas: agrupación de escenarios para identificar los escenarios característicos, análisis de las hipótesis y medidas políticas y construcción de un recomendador sobre los escenarios característicos.

El proceso de clasificación y agrupación propuesto consta de tres fases. En la primera fase, se busca agrupar aquellos escenarios que son "casi iguales". En la segunda fase, a partir de los escenarios agrupados, se extraen patrones característicos para la temperatura y el PIB, de forma independiente. Por último, en la tercera fase, a partir de los patrones extraídos se construyen los clústeres finales o escenarios característicos, y se procede a asignar cada escenario a su clúster.

Una vez asignado cada escenario a un clúster final, lo que determina el escenario característico asociado, se procede a analizar que respuestas al cuestionario determinan la pertenencia o no a un clúster fijo. Este estudio se apoya en la Información Mutua sobre el problema de pertenencia.

Por último, dada una respuesta concreta del cuestionario y fijado un escenario característico como objetivo, se presenta un recomendador que indica los cambios mínimos necesarios sobe las respuestas dadas para alcanzar dicho escenario.

Cabe destacar que, los resultados obtenidos de este trabajo se van a englobar dentro de un videojuego educativo, lo que obliga a que presenten cierta consistencia y rigurosidad.

1.4.1. Plan de Trabajo

Para el cumplimiento de los objetivos propuestos, se identifican siete tareas fundamentales, que se ejecutan de forma consecutiva. Cinco de ellas están enfocadas a cumplir el primer objetivo **Obj-01**, mientras que las dos últimas se enfoca al segundo objetivo **Obj-02**.

1. Obtención y análisis preliminar de los resultados de simulación del modelo MEDEAS.
2. Identificación de los escenarios característicos (patrones típicos) a partir de los resultados de simulación.
3. Asociación de los escenarios característicos con las doce hipótesis y medidas políticas que alimentan el proceso de simulación.
4. Identificación de las respuestas más informativas con respecto a cada escenario característico.
5. Validación de los resultados obtenidos por el modelo:
   - Objetiva, basada en los errores cometidos al aproximar cada escenario por un escenario característico.
   - Basada en juicios de expertos, para determinar si el escenario tiene o no sentido.

6. Obtención de una narrativa de cada escenario característico, realizada por un experto, que describa y justifique el escenario. Para poder obtener esta narrativa, se recurre a los resultados del análisis realizado sobre las respuestas más informativas.

7. Diseño del recomendador y módulo de diagnosis autónomo, basado en los escenarios finales y en los objetivos dados por el usuario como un servicio REST implementado en *Python* y desplegado como una imagen *docker*.

En principio las fases definidas siguen un orden consecutivo, donde los resultados obtenidos en una fase alimentan las fases posteriores. Esto supone una verificación de los resultados, y puede retroalimentar la fase de origen o las fases anteriores, obligando a comenzar el proceso de nuevo. Es decir, los resultados de una fase son valorados en las fases siguientes, y si la calidad no es aceptable se comienza o retrocede a una fase ya superada.

1.5. Estructura del documento

Este documento se estructura en seis capítulos. El **Capítulo 1** ha presentado el caso de estudio y el problema a solucionar, incluyendo una primera aproximación a la solución deseada. En el **Capítulo 2** se abordan los conceptos fundamentales sobre series temporales, clústering y la medida de Información Mutua.

El **Capítulo 3** describe la metodología aplicada: estructurar la salida del proceso de simulación como series temporales, una metodología de agrupación para estas y valoración de las entradas al proceso según el grupo asignado mediante la Información Mutua. Esta metodología es aplicada en el **Capítulo 3** sobre el cuerpo de datos del videojuego *Crossroads*, justificando cada decisión y evaluando tanto la agrupación obtenida como la metodología utilizada.

Con dicho modelo de agrupación, en el **Capítulo 4** se construye un sistema autónomo de valoración de escenario y recomendación al jugador, con el objetivo de integrarse y completar Crossroads como un módulo generador de *feedback*. Finalmente, en el **Capítulo 5** se incluyen las conclusiones del trabajo, así como las mejoras futuras.

Capítulo 2

Series temporales multivariantes

En este capítulo se desarrollan los conceptos, fundamentos y técnicas básicas que servirán de soporte a la metodología propuesta posteriormente. En particular, se aborda el concepto de serie temporal, utilizado para dar una estructura homogénea a la información y algunas técnicas de clústering adaptadas a series de tiempo.

2.1. Series temporales multivariantes

Una **serie temporal** es una colección de observaciones realizadas cronológicamente. Se dice que una serie temporal es **discreta** o de tiempo discreto cuando las observaciones que la componen han sido tomadas en instantes de tiempo espaciados (generalmente equiespaciados). En caso contrario, se dice que la serie es de **tiempo continuo**.

De forma general, una serie temporal univariante $\mathbf{x} = \{x_t\}_{t\in T}$ es una realización de un proceso estocástico $\{X_t : 1 \leq t \in T\}, T \subset \mathbb{R}$, denominado modelo de la serie [12, *Capítulo 8*]. Al conjunto T se lo denomina **orden cronológico o temporal**. La serie es discreta o continua cuando así lo sea T. Así, una serie temporal discreta univariante (UTS) $\mathbf{x} = (x_1, \ldots, x_n)$ es una realización del proceso estocástico $\{X_t : 1 \leq t \leq n\}$, donde se supone que cada observación x_t es un valor realizado de la variable aleatoria X_t. En este documento se supondrán que todas las series son de tiempo discreto [5, *Capítulo 1*].

Sin embargo, en la práctica, para cada instante de tiempo los datos se presentan como una colección de varias observaciones. Es decir, el modelo de la serie es un conjunto de vectores aleatorios $\{\mathbf{X}_t = (X_{1,t}, \ldots X_{m,t}) : 1 \leq t \leq n\}$, y su realización se denomina serie temporal (discreta) multivariante (MTS). En este tipo de series se pueden dar dos relaciones de dependencia: en la sucesión temporal $\{\mathbf{X}_t\}_t$ y entre las componentes de cada vector $\{X_{i,t}\}_i,\ 1 \leq t \leq n$ [5, *Capítulo 8*].

Extendiendo la notación utilizada para las series univariantes, una serie multivariante se escribe como

$$\bar{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \quad \text{donde } \mathbf{x}_t = (x_{1,t}, \ldots, x_{m,t}) \text{ es una realización del vector aleatorio } \mathbf{X}_t$$

Sin embargo, fijado $i \in \{1, \ldots, m\}$, $(x_{i,1}, \ldots, x_{i,n})$ es una serie temporal univariante. Por ello, las series temporales multivariantes también se pueden denotar como

$$\bar{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$$

donde $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,n})$ es una serie temporal univariante con modelo $\{X_{i,t}\}_t$. Esta segunda notación será la que se utilice en este documento.

2.1.1. Comparación de series temporales

En esta sección se presentan diferentes métricas que permiten la comparación de dos series temporales con intervalos de tiempo iguales y uniformes, $\mathbf{x} = (x_1, \ldots, x_n)$ e $\mathbf{y} = (y_1, \ldots, y_n)$. En primer lugar, se presentan medidas de comparación derivadas de las distancias vectoriales tradicionales, como la distancia euclídea (2.1) y la distancia del máximo (2.2)

$$d_E(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.1}$$

$$d_{max}(\mathbf{x}, \mathbf{y}) = \max_{1 \leq i \leq n} |x_i - y_i| \tag{2.2}$$

Cuya extensión es inmediata para series temporales multivariables, $\bar{\mathbf{x}} = (\mathbf{x}_1, \ldots \mathbf{x}_m)$ e $\bar{\mathbf{y}} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)$

$$d_E(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = \sqrt{\sum_{j=1}^{m} \sum_{i=1}^{n} (x_{j,i} - y_{j,i})^2} = \sum_{j=1}^{m} \sqrt{d_E(\mathbf{x}_j, \mathbf{y}_j)^2} \tag{2.3}$$

$$d_{max}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = \max_{1 \leq j \leq m, 1 \leq j \leq n} |x_{j,i} - y_{j,i}| = \max_{1 \leq j \leq m} d_{max}(\mathbf{x}_j, \mathbf{y}_j) \tag{2.4}$$

Por otro lado, cuando se desea conocer el error cometido al sustituir la serie $\mathbf{x} = (x_1, \ldots, x_n)$ por la serie $\mathbf{y} = (y_1, \ldots, y_n)$ existen dos métricas específicas derivadas de las anteriores: la raíz del error medio cuadrático (RMSE) y el error máximo relativo (MRE)

$$RMSE(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{n}} = \frac{d_E(\mathbf{x}, \mathbf{y})}{\sqrt{n}} \tag{2.5}$$

$$MRE(\mathbf{x}, \mathbf{y}) = \max_{1 \leq i \leq n} \frac{|x_i - y_i|}{|x_i|}, \quad x_i \neq 0 \ \forall i \tag{2.6}$$

También existen otras medidas de comparación más avanzadas, que se basan en la idea de que un mismo evento puede suceder o extenderse en diferentes instantes de tiempo. Este tipo de medidas, buscan realizar deformaciones sobre el orden cronológico para hacer coincidir eventos idénticos que suceden en distintos instantes, como la *Dynamic Time Warping* (DTW) [23].

2.2. Clustering

El objetivo del clustering es identificar la estructura en un conjunto de datos no etiquetados (solo se dispone de la información que representa el propio dato), organizando sus elementos en grupos (denominados **clúseres**) lo más homogéneos posibles, donde la disimilitud se minimiza para los elementos de un mismo clúster y se maximiza entre los diferentes clústeres. Los métodos de clustering para datos se pueden dividir en estáticos en cinco categorías [9] [26]: métodos de partición, jerárquicos, basados en densidad, basados en cuadrículas y basados en modelos.

Dado un conjunto D, se dice que $C_1, \dots C_k$ es una **partición** de D si $C_i \neq \emptyset$ y se verifica que

$$D = \bigsqcup_{i=1}^{k} C_i (\text{unión disjunta})$$

Dado un conjunto de M elementos sin etiquetar, un ***método de partición*** construye k particiones (clústeres) con al menos un elemento en cada partición. Si cada elemento solo puede pertenecer a un único clúster se dice que la partición es nítida, y en caso contrario, se dice que es difusa. La mayoría de los métodos de partición se basan en una medida de comparación, generalmente una distancia. Comienza con la creación de una partición inicial, e iterativamente se reubican los elementos entre las particiones.

Desde el punto de vista computacional, la optimización global de un método basado en particiones suele ser prohibitivo, pues, para un k fijo, requiere construir todas las posibles particiones. Por ello se utilizan métodos heurísticos, como los algoritmo de ***k-means*** **(k-medias)**, *k-medians* (k-medianas) y *k-medoids*. En todos ellos, cada clúster posee un representante en función de sus elementos, y este representante es quien determina la reubicación iterativa.

Los métodos de partición realizan una división de un único nivel. Sin embargo, es posible que cada clúster contenga a su vez subclústeres, y estos también admitan sub-subclústeres, y así sucesivamente hasta llegar a clústeres formados por un único elemento. Los métodos jerárquicos se basan en esta idea. Un ***método de clústering jerárquico*** establece un "árbol" de contención de clústeres.

En función de cómo se realiza la descomposición, se diferencian dos tipos de métodos jerárquicos: **divisivo** y **aglomerativo**. El primero de ellos emplea una estrategia descendente: comienza considerando que todos los elementos pertenecen al mismo clúster, y en cada iteración un clúster se divide en otros más pequeños, repitiéndose hasta obtener n clústeres con un único elemento en cada uno de ellos. El clústering jerárquico aglomerativo emplea una estrategia ascendente, que comienza considerando que cada elemento pertenece a un clúster diferente y en cada iteración fusiona los dos clústeres más "cercanos" (dado por una medida de similitud).

Al contrario que en los métodos de partición, donde un cambio de clúster para un elemento puede deshacerse en futuras iteraciones, en los métodos jerárquicos, una vez que se realiza la fusión o división, esta no se puede deshacer. Por otro lado, mientras que los métodos de par-

tición suelen requerir una inicialización aleatoria, esta no es necesaria en los métodos jerárquicos.

Los ***métodos basados en cuadrículas*** (*grid*) discretizan el espacio de los elementos en una cantidad finita de celdas, que forman una estructura de red. Para formar la red, cada atributo de los objetos se divide en intervalos de igual longitud, dividiendo un espacio $n-$dimensional en un conjunto finito de celdas. Un algoritmo típico de este enfoque es STING [25].

Los ***métodos basados en densidad*** construyen los clústeres como áreas "densas" de elementos separadas por regiones dispersas. La idea general es que un clúster se expande (adquiere nuevos elementos) mientras su densidad (número de elementos) sea inferior a un cierto umbral. El algoritmo típico que se basa en esta idea es *DBSCAN* [10]. Estos métodos son capaces de identificar tanto formas variadas en los clústeres como *outliers*.

Finalmente, los ***métodos basados en modelos*** supone que cada grupo posee sigue cierto modelo en su generación, y trata de ajustar los elementos del conjunto a algún modelo. Se diferencian dos enfoques: el enfoque estadístico, como *AutoClass* [7] basado en el análisis bayesiano, y el enfoque de red neuronal, como *ART* [6] y los mapas de características autoorganizados [15].

2.2.1. K-means

Sea D un conjunto de n objetos en un Espacio Euclídeo. Un método de partición divide el conjunto en k clústeres $C_1, \ldots, C_k$. Los métodos de partición basados en representantes (denominados **centroides**), identifican cada clúster C_i, de forma unívoca, con un elemento del espacio μ_i. Conceptualmente, el centroide se puede ver como el punto central del clúster. En el algoritmo de **k-medias** el centroide se define como la media de los elementos pertenecientes a cada clúster.

La calidad de un clúster en los métodos de partición por centroides se puede medir por la variación intra-clúster, también denominada suma de los errores al cuadrado, definida como

$$\sum_{i=1}^{k} \sum_{x \in C_i} d_E(x, \mu_i)^2 \tag{2.7}$$

donde d_E denota la distancia euclídea. Por tanto, el problema de clasificación se traduce en encontrar la partición de k subconjuntos de D que minimice el valor de la distancia intra-clúster. Esto implica que se tendría que evaluar todas las posibles particiones de D en k subconjuntos, cuyo número guarda una relación exponencial con el número de elementos en D, y hace que sea computacionalmente inabordable. Se ha demostrado que el problema es *NP-hard* incluso para dos clústeres [18].

El algoritmo de k-medias, algoritmo de clústering por excelencia, es una aproximación escalable al problema anterior.

1. Se comienza con un conjunto inicial de centroides, que puede seleccionarse de forma aleatoria o mediante técnicas heurísticas.

2. En cada iteración, para cada muestra de D se calcula su distancia con respecto cada uno de los centroides, y se asigna al clúster del centroide de mínima distancia. Posteriormente, los centroides se recalculan como la media de los elementos que componen su clúster.

3. El paso anterior se repite hasta que se cumple cierto criterio de parada. El criterio general implica que los clústeres permanezcan invariantes (y por tanto los centroides) en una iteración. Criterios más avanzados implican limites en la reducción de la variación intra-clúster de una iteración con respecto a la anterior, de forma que, cuando la diferencia es inferior a este límite se finaliza.

Algorithm 1 K-medias

Input: D un conjunto de n muestras de un Espacio Euclídeo y k el número de clústeres.
Output: Un conjunto de k clústeres, representados 'por $mu_1, \ldots, \mu_k$.

```
1: begin Inicialización de μ_1, ..., μ_k
2:     repeat
3:         Clasificar las n muestras de D con respecto al μ_i más cercano.
4:         Recalcular μ_1, ..., μ_k
5:     until μ_1, ..., μ_k no cambien
6: end
```

Este método, propuesto por Stuart Lloyd en 1957 [17], es una forma aproximada de obtener estimaciones de máxima verosimilitud para las medias. En general, cuando el solapamiento entre las densidades de los componentes es pequeño, el enfoque del mínimo real para la variación intra-clúster y el procedimiento de k-medias producen resultados similares [9, *Sección 10.4.3*]. La complejidad computacional de este algoritmo es $\mathcal{O}(nTk)$, siendo n el número de muestras y T el número de iteraciones. En general, k y T son de varios ordenes de magnitud inferior a n, y fijado un límite máximo de iteraciones siempre es escalable.

2.2.2. Determinando el número de clústeres: el método del codo

Los métodos de clustering por particiones, como el algoritmo de k-medias, requieren que se determine previamente el número de clústeres a construir. Para determinar este valor, se recurre al **método del codo** [4]. Este método se basa en la idea de que se debe elegir un número de clústeres tal que la adición de un nuevo clúster no ofrezca una modelización mucho mejor de los datos.

Para cada valor de k (número de clústeres), este se enfrenta con la variación intra-clúster. Para los primeros valores de k se obtendrá un valor elevado de variación intra-clúster, que se reduce drásticamente a medida que se aumenta k. Sin embargo, a partir de un determinado valor,

k_0, la variación intra-clúster pasa de decrementarse drásticamente a hacerlo de forma moderada, manteniéndose este comportamiento para cada $k > k_0$. La causa de esto es que los clústeres quedan próximos entre sí (el nuevo clúster que se construye en k está muy cerca de algún clúster de los construidos en $k-1$). Dicho k_0 es el valor que se elige como óptimo para el número de clústeres. La justificación sobre el método se presenta en la **Sección 2.4.1**.

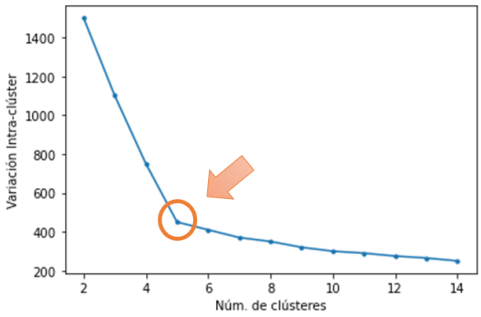


Figura 2.1: Ejemplo de aplicación del método del codo, donde la elección es $k = 5$

2.2.3. Clústering jerárquico aglomerativo

Se considera un conjunto D de n objetos a particionar. En primer lugar, se supone que cada objeto es un clúster, de forma que se tendrían n clústeres diferentes. Posteriormente, se unen los dos clústeres que estén más próximos entre sí, lo que produce una segunda partición de $n-1$ grupos. Esto se repite, obteniendo particiones una partición de $n-2$ clústeres. Así, en cada iteración se unen los dos clústeres más próximos, hasta la $n-$ésima partición, en la que todos los elementos pertenecen a un único clúster. Como notación general, a la iteración se le denomina nivel, de forma que el número de clústeres en el nivel i es de $n-i+1$.

En general, un clústering jerárquico se define como una sucesión de particiones de D

$$\{\mathcal{C}_i = \{C_1^i, \ldots, C_{n-i+1}^i\}\}_{i=1}^n$$

donde se verifica que para cada par de elementos $x, y \in D$ pertenecientes a un mismo clúster en la iteración i_0, entonces permanecen juntan en cualquier nivel superior (para todo $i > i_0$) [13, *Sección 10.3*].

La forma natural de representación de los agrupamientos jerárquicos es en forma de árbol, denominado dendograma **Figura 2.2**. Este muestra cómo se agrupan los objetos, y dada una medida de disimilitud (por ejemplo, la distancia euclídea), también muestra cómo se incrementa en cada clúster. El valor de disimilitud permite determinar si los agrupamientos son "naturales".

El problema de clasificación se resuelve eligiendo una partición de la sucesión de particiones. Para ello, se recurre al dendograma y se elige aquel nivel que presenta una diferencia notable

entre los valores de diferentes niveles de disimilitud. En el ejemplo de la **Figura 2.2**, se elegiría el nivel 8.

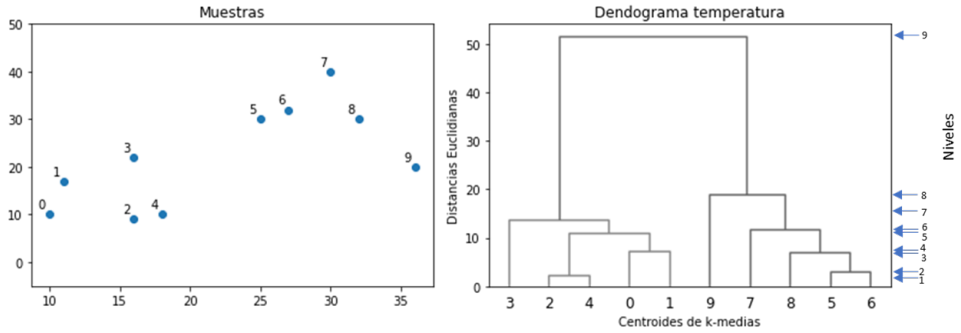


Figura 2.2: Ejemplo de agrupación jerárquico. A la izquierda se representan las muestras, y a derecha el dendograma con los diferentes niveles.

El siguiente algoritmo (**Algoritmo 2**) realiza las agrupaciones hasta un cierto nivel l predefinido.

Algorithm 2 Agrupación jerárquica

Input: $D = \{x_1, \ldots, x_n\}$ un conjunto de n muestras de un Espacio Euclídeo, k el número de clústeres deseados.
Output: Un conjunto de k clústeres.

begin Inicializar $C_i \leftarrow \{x_i\},\ i = 1, \ldots, n$
  $l \leftarrow n - k + 1$
  $i \leftarrow 1$
  repeat
    Buscar los dos clústeres más cercanos, denotados por $C_i,\ C_j$.
    $C_i \leftarrow C_i \cup C_j$
    Eliminar C_j
    $i \leftarrow i + 1$
  until $i == l$
end

El problema que presenta el clústering jerárquico es su escalabilidad, pues su complejidad computacional es de $\mathcal{O}(n^2)$, siendo n el número de elementos a clasificar [9, *Sección 10.9.2*]. Esto se debe a que en cada iteración se hace necesario calcular la distancia entre todos los clústeres para hallar su mínimo, y para un número fijo de clústeres k, se requerirán $n-k+1$ iteraciones.

2.3. Clustering de series temporales

La agrupación (clústering) de series temporales ha demostrado ser una técnica bastante eficaz para proporcionar información útil en determinados dominios [26]. En su aplicación sobre series temporales, se diferencian tres enfoques [26]: basado en datos en bruto, basado en características y basado en modelos.

El **enfoque basado en datos en bruto** utiliza directamente los datos que componen la propia serie, sin procesarlos. Este enfoque requiere definir una medida de distancia o similitud adecuada a las series temporales, siendo esta la modificación principal a realizar sobre los algoritmos tradicionales.

Los otros enfoques buscan extraer información inherente a la propia serie, que es utilizada en la clasificación. El **enfoque basado en características** representan la serie como un vector de características de dimensión fija, sobre el que aplicar los algoritmos tradicionales. El **enfoque basado en modelos** representa la serie temporal como su proceso estocástico, aplicando ciertos modelos estadísticos (modelo oculto de Markov [19] o los modelos ARMA y ARIMA [8] [14]).

El uso principal que se da al clústering de series temporales con los datos en bruto es el reconocimiento de patrones implícitos en las propias series. Así, se supone que las series agrupadas en un mismo conjunto comparten un patrón común, y utilizándolas se trata de hallar dicho patrón. Por ejemplo, en el algoritmo de k-medias se recurre la media componente a componente, que consideran las series como vectores, o técnicas más avanzadas basadas en buscar el mejor ajuste realizando "deformaciones temporales" (la media de Fréchet para la DTW [20]). El objetivo común es definir una nueva serie temporal que represente a todos los elementos del clúster. Otro enfoque supone que todas las series de un clúster provienen del mismo modelo (proceso estocástico), y tratan de deducirlo a partir de la muestra.

2.3.1. Clústering para reducir el número de elementos

Además del reconocimiento de patrones, los algoritmos de clustering pueden ser utilizados para reducir la cantidad de elementos que componen el cuerpo de los datos (*Numerosity Reduction* [13, *Sección 3.4*]). Las técnicas de reducción de la cantidad buscan sustituir el conjunto original por un conjunto alternativo, de menor tamaño y que mantenga y que sea representativo del original.

En el caso de las series temporales, esta aplicación es de gran relevancia, pues la igualdad de dos series es un requisito muy estricto. Dos series discretas con mismo orden temporal T son iguales si así lo son cada una de sus componentes. El uso de determinadas medidas de comparación, como la DTW, da cierta flexibilidad a la definición de igualdad.

Utilizando el clustering, y asumiendo un cierto error, es posible acotar el dominio de información, por ejemplo, considerando como iguales todos los elementos asociados a un clúster, y

sustituirlos por su representante. Cuando hay una gran similitud entre los elementos del clúster y su representante (error pequeño) esto produce buenos resultados.

Como ejemplo sobre datos estáticos, en [1] se aplica el algoritmo de k-medias sobre imágenes de resonancias en escala de grises para reducir el espectro posible de tonos de gris en las imágenes. Posteriormente se utiliza un segundo algoritmo para la segmentación de imágenes, obteniéndose mejores resultados si se aplicaba la reducción de tonalidad de grises.

2.4. Valoración de un clústering

Una vez construido un clústering interesa evaluar como de bueno es, en el sentido de que los agrupamientos sean naturales, maximizando la disimilitud entre los clústeres y minimizándola entre los elementos de un mismo clúster. Es decir, se desea responder a la pregunta *"¿Qué calidad tiene el clustering generado por un método, y cómo podemos comparar los clústerings generados por diferentes métodos?"* [13].

En general, los métodos de evaluación de clústering se pueden clasificar en dos grandes familias, teniendo en cuenta la disponibilidad de la *verdad sobre el terreno* (*ground truth*, agrupación construida por expertos humanos): métodos **intrínsecos** y **extrínsecos**. Los métodos extrínsecos exigen que la verdad esté disponible, y comparan las agrupaciones obtenidas con respecto a esta. Son una especie de métodos supervisados, pues se tiene cierta etiqueta además de la información dada por los propios datos.

Los métodos intrínsecos no utilizan la verdad sobre el terreno, sino que evalúan la bondad del clúster en función de la separación existente entre los datos. Intuitivamente, si el clúster minimiza la distancia dentro de los elementos de un clúster y la maximiza con respecto al resto, en una clasificación la distancia mínima entre los centros será inferior a la mitad de la máxima distancia entre cada centro y cada elemento de su clúster.

2.4.1. Variación intra -clúster

El primer problema por analizar es si el número de clústeres es adecuado (natural). Si $J(k),\ k = 1, \ldots, n$ denota la variación intra-clúster para la partición óptima de k elementos del conjunto D, entonces se verifica que $J(k) < J(k+1)$ [9, *Sección 10.10*]. Intuitivamente, si en la partición óptima de k clústeres se selecciona un elemento arbitrario y se crea el clúster formado por este único elemento, la nueva partición de $k+1$ clústeres tiene una variación intra-clúster menor que la partición de k clústeres. Por tanto, la partición óptima de $k+1$ clústeres tendrá una variación intra-clúster menor.

La nueva agrupación que se ha formado puede no ser natural. Intuitivamente, se puede elegir un elemento del clúster próximo a su centro, luego el nuevo clúster estaría rodeado de elementos

de otro. El método del codo se basa en esta idea: mientras se separan clústeres con sus elementos dispersos (por ejemplo, alejados de su centro), se tendrá una disminución importante de la variación intra-clúster, mientras que, si se divide un grupo de elementos cercanos entre sí, la disminución será pequeña. Es decir, se determina el punto en el que se pasa de clústeres naturales, separados por zonas dispersas, a clústeres forzados, con grupos próximos entre sí.

Cabe destacar que la variación intra-clúster es una medida en particiones con representante. Para otro tipo de particiones, existen técnicas similares como la distancia intra-clúster. Con la notación utilizada en (2.7), se describe como

$$\sum_{i=1}^{k} \frac{1}{2} \sum_{x \in C_i} \sum_{y \in C_i} d_E(x, y) \tag{2.8}$$

En particular, los dendogramas en el clústering jerárquico utilizan esta medida de disimilitud, sobre cada clúster, para determinar si los agrupamientos son naturales, y que partición elegir.

2.4.2. Coeficientes de la silueta

Los coeficientes de la silueta (*silhouette coefficient*) [13, *Sección 10.6.3*] es una medida de la bondad en la separación de los elementos de un clústering. Sea $D = \{x_1, \ldots, x_n\}$ un conjunto de n elementos, y sea $C_1, \ldots, C_k$ una partición de D (con $k < n$), para cada $x \in D$ con $x \in C_i$ se definen los coeficientes

$$\begin{aligned} a(x) &= \frac{1}{|C_i| - 1} \sum_{y \in C_i} d(x, y) \\ b(x) &= \min_{1 \leq j \leq k,\ i \neq j} \frac{1}{|C_j|} \sum_{y \in C_j} d(x, y) \end{aligned} \tag{2.9}$$

donde $|C_i|$ denota el cardinal (número de elementos) del conjunto C_i. El **coeficiente de la silueta** para x se define como

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}} \tag{2.10}$$

Es claro que $a(x), b(x) \geq 0$ para cualquier x siempre que d verifique el axioma de no-negatividad. En consecuencia $-1 \leq s(x) \leq 1$. El valor $a(x)$ es la distancia media a los elementos pertenecientes al mismo clúster que x, y evalúa lo compacto que es el clúster al que pertenece x. Por otro lado, $b(x)$ es la distancia media a los elementos que no están en el mismo clúster que x, y mide lo separado que está x del resto de clústeres.

La situación deseable es que $a(x)$ sea pequeño (clúster compacto) y $b(x)$ sea grande (x está alejado del resto de clústeres). En este caso, el valor $s(x)$ será próximo a 1. Por otro lado, si

$s(x) < 0$ entonces $a(x) > b(x)$, por lo que x está más cerca de los objetos de otros clústeres que de los miembros de su propio clúster. Generalmente, esta es una situación bastante indeseable que se desea evitar.

La evaluación de un clústering mediante los coeficientes de la silueta se realiza con el promedio de los coeficientes para todos los objetos del clúster.

2.4.3. Valoración del método de clústering

Los coeficientes de la silueta y la variación intra-clúster son dos medidas de la bondad de una agrupación. En esta sección se valora la metodología de clasificación empleada. En particular, se desea evaluar como de sensible es un método de clústering a la desaparición de algunos elementos sobre el conjunto de datos.

Sea $D = \{x_1, \ldots, x_n\}$ un conjunto de datos, y sean $Y = \{Y_1, \ldots, Y_{k_1}\}$ y $Z = \{X_1, \ldots, Z_{k_2}\}$ dos particiones de D, se define la medida de similitud c entre Y y Z como [21]

$$c(Y, Z) = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} \sum_{j>i} \gamma_{i,j} \tag{2.11}$$

donde

$$\gamma_{i,j} = \begin{cases} 1 & \text{Si existen } p, q \in \mathbb{N} \text{ tales que } x_i, x_j \in Y_p \text{ y } x_i, x_j \in Z_q \\ 1 & \text{Si existen } p, q \in \mathbb{N} \text{ tales que } x_i \in Y_p,\ x_i \in Z_q,\ X_j \notin Y_p \text{ y } x_j \notin Z_q. \\ 0 & \text{Resto de casos} \end{cases} \tag{2.12}$$

Es decir, dados dos puntos $x_i, x_j \in D$, $\gamma_{i,j} = 1$ si están juntos en un clúster para ambas clasificaciones o están en clústeres separados en ambas clasificaciones. Cuando en una clasificación los dos están en el mismo clúster y en otra no, entonces $\gamma_{i,j} = 0$. En particular, si $n_{i,j}$ es el número de elementos que están simultáneamente en las particiones Y_i y Z_j, entonces [21]

$$c(Y, Z) = \frac{1}{\binom{n}{2}} \left[\binom{n}{2} - \left[\frac{1}{2} \sum_{i=1}^{k_1} \left(\sum_{j=1}^{k_2} n_{i,j} \right)^2 + \frac{1}{2} \sum_{j=1}^{k_2} \left(\sum_{i=1}^{k_1} n_{i,j} \right)^2 - \sum_{i=1}^{k_2} \sum_{j=1}^{k_1} n_{i,j}^2 \right] \right] \tag{2.13}$$

En general, dado un conjunto de datos muestra, se puede dar el caso en que sea incompleto o la población no esté bien representada en la muestra. Para valorar si la metodología aplicada es sensible a la desaparición de datos, se considera $E \subset D$. Sea $Y = \{Y_1, \ldots, Y_k\}$ una partición de D, y sea $Z = \{Z_1, \ldots, Z_k\}$ una partición de E, se considera $Y' = Y'_1, \ldots, Y'_k$ donde $Y'_i = Y_i \cap E$.

En consecuencia, se tienen dos particiones de E, dadas por Z y por Y'[1] de las que interesa evaluar como son de diferentes, se recurre a la medida c anterior.

Cabe destacar que si dos particiones, $Y = \{Y_1, \ldots, Y_k\}$ y $Z = \{Z_1, \ldots, Z_k\}$, tienen conjuntos idénticos, salvo permutaciones, entonces el coeficiente $c(Y, Z) = 1$, y cuanto más cercano sea el valor c a uno mayor sera la similitud. Por otro lado, de la ecuación (2.11) se obtiene que $c(Y, Z) \geq 0$.

2.5. Información Mutua

Generalmente, se dispone de un cuerpo de datos etiquetados donde cada muestra está representada por conjunto numeroso de características, vistas como un punto en el espacio $n-$dimensional. El objetivo es elegir el mínimo número de características que permita discriminar entre las clases, sin redundancia.

La medida de la información mutua, originaria de la teoría de la información, ha sido utilizada de forma recurrente para la selección de atributos en múltiples dominios [3], destacando especialmente en la construcción de árboles de clasificación. Es una generalización de la Correlación de Pearson que no realiza suposiciones sobre el modelo que subyace[2].

Sea X una variable aleatoria discreta con función de probabilidad p_X. Con abuso de notación, denotaremos también por $X = \{x_1, \ldots, x_n\}$ el espacio muestral de la variable aleatoria. Dada una segunda variable discreta Y, $p_{X,Y}$ denota la distribución de probabilidad conjunta de (X, Y), y $p_Y(y) = \sum_{x \in X} p_{x,y}(x, y)$ es la distribución marginal de Y.

La información mutua entre dos variables aleatorias cuantifica la cantidad de información que comparten ambas variables. Es decir, estima la reducción de incertidumbre en una variable aleatoria (por ejemplo Y) que supone el conocimiento u observación de otra (X). En el caso discreto, se define como [22]:

$$I(X, Y) = \sum_{(x,y) \in X \times Y,\ p(x,y) \neq 0} p_{(X,Y)}(x, y) \log \left(\frac{p_{(X,Y)}(x, y)}{p_X(x) p_Y(y)} \right) \tag{2.14}$$

Por tanto, un valor de información mutua alto indica una gran reducción de la incertidumbre (mayor dependencia entre las variables), y un valor bajo indica una pequeña reducción de la incertidumbre.

[1] En sentido amplio de partición, pues es posible que $Y'_l = \emptyset$. Sin embargo, $E = \bigcup_{i=1}^{k} Y'_i$, con lo que valdría con "ignorar" los Y'_i vacíos para tener una partición (aunque no en k conjuntos). Para mantener la misma notación, trataremos Y'_i como una partición, aunque pueda haber conjuntos vacíos, pues estos no tienen ninguna influencia sobre la medida $c(Y', Z)$, si suponemos que $Y'_l = \emptyset$ entonces los coeficientes $n_{l,j} = 0$ para cualquier j.

[2] La correlación de Pearson supone la linealidad de dos variables, limitándose a este tipo de relaciones.

2.5.1. Normalización de la información mutua

La información mutua está muy relacionada con otro concepto de la teoría de la información, la Entropía de Shannon, que es una estimación de la cantidad de información que una variable aleatoria representa. En el caso discreto, se define como [24]

$$H(X) = -\sum_{x \in X} p_X(x) \log p_X(x) \tag{2.15}$$

Si X es una variable aleatoria con espacio muestral $\{x_1, \ldots, x_n\}$ elementos, y $p_X(x_i)$ es la probabilidad de x_i, entonces H alcanza su máximo cuando se verifica

$$p_X(x_1) = p_X(x_2) = \cdots p_X(x_n)$$

Además, en este caso $H(X) = \log(n)$ [22, *Sección 2*].

Además, el concepto de Entropía se puede extender a la información del vector aleatorio discreto dado por las variables X e Y, denominado entropía conjunta:

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p_{X,Y}(x,y) \log p_{X,Y}(x,y) \tag{2.16}$$

Cuando se supone independencia entre las variables, $p_{X,Y}(x,y) = p_X(x)p_Y(y)$, se verifica que $I(X,Y) = 0$. Es más, $I(X,Y) \geq 0$. Por otro lado, cuando ambas variables están perfectamente correlacionadas (conocer una determina unívocamente la otra) , $I(X,Y)$ alcanza su máximo en $H(X,Y) = H(X) = H(Y)$. Es más, en general se verifica la siguiente desigualdad[3]:

$$0 \leq I(X,Y) \leq H(X), H(Y) \leq H(X,Y)$$

Por tanto, no es posible definir una cota superior genérica para la información mutual. Sin embargo, se puede normalizar al intervalo $[0,1]$ utilizando la entropía o la entropía conjunta.

La metodología que utiliza la Información Mutua normalizada para la selección de un subconjunto de características sobre $X_1, \ldots X_n$ en función de una etiqueta Y se denomina **NMIFS** (*Normalized Mutual Information Feature Selection*) [11].

2.5.2. Extensión a variables aleatorias continuas

La medida de la Información Mutua, y la Entropía de Shannon se pueden extender sobre variables aleatorias continuas, X e Y con funciones de probabilidad conjunta $p_{X,Y}$ y marginales p_X y p_Y respectivamente:

$$\begin{aligned} IM(X,Y) &= \int_X \int_Y p_{X,Y}(x,y) \log \left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \right) dxdy \\ h(X) &= -\int_X p_X(x) \log(p_X(x)) dx \end{aligned} \tag{2.17}$$

[3]Véase [22, *Proposición 3.3*]

donde, con abuso de notación, X denota la variable aleatoria y el conjunto que lo soporta. A la extensión del concepto de entropía sobre variables aleatorias continuas, $h(X)$ se le denomina **entropía diferencial**.

En general, calcular la Información Mutua y la Entropía para variables aleatorias continuas es complejo, por lo que se suelen aplicar algoritmos para obtener estimaciones de máxima verosimilitud [16].

Capítulo 3

Identificación y caracterización de escenarios característicos

En este capítulo se presenta la metodología a seguir, enfocada a conjuntos de datos no etiquetados resultado de un proceso de simulación y estructurados como series temporales multivariantes. Por tanto, se diferencian tres componentes claves: el propio proceso de simulación, la entrada al proceso y la salida producida.

La metodología propuesta tiene dos etapas diferentes:

Etapa 1. En primer lugar, se desean agrupar los resultados simulados, clasificándolos en k clústeres. Así, al finalizar esta etapa, se podría sustituir cada resultado de simulación por el clúster al que fue asignado.

Etapa 2. Considerando el conjunto de las entradas al proceso de simulación (*inputs*) y el clúster al que pertenece su salida, se busca determinar si existe relación entre cada *input* y el clúster.

3.1. Formulación general del problema

Se dispone de un proceso de simulación, denotado por $\mathbb{S}$, que, dada una entrada, $\mathbf{x} = (x_1, \ldots, x_p)$, produce una salida $\bar{\mathbf{y}} = \mathbb{S}(\mathbf{x}) = (\mathbf{y}_1, \ldots, \mathbf{y}_m)$ representada como una serie temporal multivariante. Por tanto, para cada $i = 1, \ldots, m$, $\mathbf{y}_i$ es una serie temporal univariante de $n \in \mathbb{N}$ elementos

$$\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,n}) \in \mathbb{R}^n \quad i = 1, \ldots, m$$

En consecuencia, $\bar{\mathbf{y}} \in \mathcal{T} = \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_m}$. Por otro lado, supondremos que $\mathbf{x}$ pertenece a un espacio $\mathcal{H}$ de cardinal finito (es decir, el número de entradas posibles es acotado). En estas condiciones, toda la información de cada simulación se puede describir como

$$\mathcal{A} = \{(\mathbf{x}, \bar{\mathbf{y}}) \ : \ \bar{\mathbf{y}} \in \mathcal{T}, \ \mathbf{x} \in \mathcal{H}, \ \bar{\mathbf{y}} = \mathbb{S}(\mathbf{x}))\}$$

Además, se consideran los siguientes conjuntos derivados de $\mathcal{A}$:

$$\mathcal{Y} = \{\bar{\mathbf{y}} : \exists\ \mathbf{x} \in \mathcal{H} \text{ con } (\mathbf{x}, \bar{\mathbf{y}}) \in \mathcal{A}\}$$
$$\mathcal{X} = \{\mathbf{x} : \exists\ \bar{\mathbf{y}} \in \mathcal{T} \text{ con } (\mathbf{x}, \bar{\mathbf{y}}) \in \mathcal{A}\}$$
$$\mathcal{Y}^{(i)} = \{\mathbf{y}_i : \exists\ \bar{\mathbf{y}} \in \mathcal{Y} \text{ con } \mathbf{y}_i \text{ término } i\text{-ésimo de } \bar{\mathbf{y}}\}$$

Es decir, $\mathcal{X}$ es el conjunto de entradas para los que se ha realizado una simulación, e $\mathcal{Y}$ es el conjunto de las simulaciones realizada para para alguna entrada. Si se hubiese realizado una simulación en $\mathcal{A}$ para cada elemento de $\mathcal{H}$, entonces $\mathcal{X} = \mathcal{H}$.

El caso de estudio, las MTS están formadas por dos componentes: la simulación del incremento de la temperatura media global y la simulación de la evolución del PIB global. El *input* de este proceso son las respuestas al cuestionario descrito en la **Sección 1.1**.

En las **Secciones 3.2** y **3.3** se desarrolla la metodología aplicada para un caso general. En la **Sección 3.4** se presenta el caso de estudio concreto, y los resultados obtenidos se desarrollan en las **Secciones 3.6** y **3.7**.

3.2. Metodología de clasificación: extracción de los escenarios característicos

En esta sección se describe el procedimiento aplicado para identificar patrones característicos en series temporales multivariante. Como notación general, el conjunto $\{1, \ldots, P\}$ se representa como $[P]$.

1. **Reducción del cuerpo de datos**

   Sea $\bar{\mathbf{y}} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)$ una serie temporal multivariante contenida en $\mathcal{Y}$. En primer lugar, se propone aplicar el algoritmo de k-medias sobre las series multivariante, considerándolas como una unidad indivisible. De esta forma, el conjunto de las series multivariante, $\mathcal{Y}$ se divide en L particiones disjuntas $\mathcal{B}_1, \ldots, \mathcal{B}_L$ aplicando el algoritmo de k-medias. Así,

   $$\mathcal{Y} = \bigsqcup_{i=1}^{L} \mathcal{B}_i$$

   Además, para cada $i = 1, \ldots, L$, se considera el centroide de $\mathcal{B}_i$, definido como la media de sus miembros, que a su vez es una serie temporal multivariante denotada por $\bar{\mathbf{c}}_i$. Se define $\mathcal{C} = \{\bar{\mathbf{c}}_i : i = 1, \ldots, L\}$ como el conjunto de los centroides. Esta primera fase se puede ver como una aplicación $\alpha : \mathcal{Y} \to \mathcal{C}$ tal que, para cada elemento $\bar{\mathbf{y}} \in \mathcal{Y}$, $\alpha(\bar{\mathbf{y}})$ es el centroide del clúster al que pertenece $\bar{\mathbf{y}}$.

Es posible que el orden de magnitud de las series que componen una serie temporal multivariante sean diferentes entre sí. En este caso, teniendo en cuenta que se utiliza la distancia euclídea, estas series de mayor orden pueden condicionar los resultados de esta primera fase. Por ello, previo al uso del algoritmo de k-medias, se procede a la estandarización de los elementos de las series temporales multivariables.

La estandarización de las MTS se aplica sobre cada UTS que la compone. Así, para cada $i = 1, \ldots, m$ se consideran

$$\begin{aligned} \mu_i &= \frac{1}{N_s}\frac{1}{n}\sum_{\mathbf{y}_i \in \mathcal{Y}^{(i)}}\sum_{j=1}^{n} y_{i,j} \\ \sigma_i^2 &= \frac{1}{N_s}\frac{1}{n}\sum_{\mathbf{y}_i \in \mathcal{Y}^{(i)}}\sum_{j=1}^{n} (y_{i,j} - \mu_i)^2 \end{aligned} \tag{3.1}$$

donde N_s representa el número de series en $\mathcal{Y}^{(i)}$. Posteriormente, para cada observación de cada serie temporal en $\mathcal{Y}^{(i)}$ se le resta μ_i y se divide por σ_i.

$$y'_{i,j} = \frac{y_{i,j} - \mu_i}{\sigma_i}, \qquad i = 1, 2;\ j = 1, \ldots, n, \quad \mathbf{y}_i \in \mathcal{Y}^{(i)} \tag{3.2}$$

El procedimiento de estandarización es la metodología tradicional que se aplica sobre el conjunto de las observaciones $\{y_{i,j} : \mathbf{y}_i \in \mathcal{Y}^{(i)},\ \ j = 1, \ldots, n\}$ sin tener en cuenta el caracter temporal.

2. **Extracción de patrones de cada serie temporal**

Con la notación expuesta anteriormente, se consideran de forma independiente los subconjuntos $\mathcal{C}^{(i)}$, es decir, se separa cada serie que compone la serie temporal multivariante y que representan la evolución de la misma información. Para cada $\mathcal{C}^{(i)}$ se aplica un algoritmo de clústering jerárquico aglomerativo, dividiendo el conjunto en N_i particiones, $i = 1, \ldots, m$. De esta forma,

$$\mathcal{C}^{(i)} = \bigsqcup_{j=1}^{N_i} \mathcal{D}_{i,j}, \qquad i = 1, \ldots, m$$

Por otro lado, para cada partición

$$\mathcal{D}_{i,j}, \quad i = 1, \ldots, m,\ j = 1, \ldots, N_i$$

se considera el representante o patrón de dicho conjunto, $\mathbf{d}_{i,j}$, como la media componente a componente de sus elementos. Esta segunda fase se puede describir como una familia de funciones

$$\{\beta_i : \mathcal{C}^i \to \{1, \ldots N_i\} \text{ con } 1 \leq i \leq m\}$$

tal que, para cada $\mathbf{c}_i \in \mathcal{C}^{(i)}$ se verifica que $\beta_i(\mathbf{c}_i) = j$ si y solo si $\mathbf{c}_i \in \mathcal{D}_{i,j}$. Así, se define $\beta : \mathcal{C} \to [N_1] \times \cdots \times [N_m]$ como $\beta = (\beta_1, \ldots, \beta_m)$.

3. **Construcción de los clústeres y patrones característicos**

   Dado un multi-índice $\mathbf{i} = (i_1, \ldots, i_m) \in [N_1] \times \cdots \times [N_m]$, se define el clúster final

$$\begin{gathered}\mathcal{E}_{(i_1,\ldots,i_m)} = \{\bar{\mathbf{y}} \in \mathcal{Y} : \ \bar{\mathbf{y}} \in \mathcal{B}_j, \ \mathbf{c}_k \in \mathcal{D}_{1,i_k}, \ \text{para todo } k = 1, \ldots, m\} \\ \text{donde } \bar{\mathbf{c}} = (\mathbf{c}_1, \ldots, \mathbf{c}_m) \text{ es el centroide del clúster } \mathcal{B}_j\end{gathered} \tag{3.3}$$

   Es decir, dada una serie $\bar{\mathbf{y}} \in \mathcal{Y}$, el proceso de asignación a un clúster es como sigue:

   a) Se determina, mediante la clasificación de k-medias, la partición B_i a la que pertenece el clúster y se toma el centroide $\bar{\mathbf{c}} = (\mathbf{c}_1, \ldots, \mathbf{c}_m)$ correspondiente a la partición asignada. Es decir, $\bar{\mathbf{c}} = \alpha(\bar{\mathbf{y}})$.

   b) Para cada componente del centroide, $\mathbf{c}_j$ $j = 1, \ldots, m$, se aplica su respectiva clasificación jerárquica, obteniendo la partición a la que pertenece, $\mathcal{D}_{j,k_j}$. Esto es, $k_j = \beta_j(\mathbf{c}_j)$.

   c) Se determina el clúster final al que pertenece, $\mathcal{E}_{\mathbf{k}}$, uniendo las particiones a las que pertenecen cada una de sus componentes, $\mathbf{k} = (k_1, \ldots, k_m)$. Por tanto, $\mathbf{k} = \beta(\alpha(\bar{\mathbf{y}}))$.

   Por otro lado, cada clúster final $\mathcal{E}_{(i_1,\ldots,i_m)}$ tiene asociado un único escenario característico, definido como:

$$\bar{\mathbf{d}}_{(i_1,\ldots,i_m)} = (\mathbf{d}_{1,i_1}, \mathbf{d}_{2,i_2}, \ldots, \mathbf{d}_{m,i_m})$$

En resumen, el método aplicado para clasificar un conjunto de MTS comienza reduciendo el conjunto de datos $\mathcal{Y}$ en otro conjunto $\mathcal{C}$, que representa a los elementos del conjunto original y tiene un menor número de elementos. Posteriormente, de las MTS del conjunto $\mathcal{C}$ se separa cada una de sus componentes, que son UTS, y cada conjunto para cada componente $\mathcal{Y}^{(i)}$ se agrupa con un algoritmo jerárquico, obteniendo la partición $\{D_{i,1}, \ldots, D_{i,N_i}\}$. Finalmente, se procede a la construcción de los clústeres finales y sus MTS representativas, como la combinación de cada clúster para cada conjunto UTS de componentes.

Donde $NumReduction(\mathcal{Y}, L)$ es el resultado de aplicar el algoritmo de k-medias (**Algoritmo 1**) sobre el conjunto $\mathcal{Y}$ para L clústeres y retornar los centroides de cada clúster. Por otro lado, $ClassReduction(\mathcal{C}^{(i)}, N_i)$ es la aplicación del algoritmo de clasificación jerárquico aglomerativo (**Algoritmo 2**) sobre el conjunto $\mathcal{C}^{(i)}$ cortando en el nivel que produce N_i clústeres, para posteriormente, retornar la media[1] de los elementos asignados a cada clúster $\mathbf{d}_{i,1}, \ldots, \mathbf{d}_{i,N_i}$.

[1] En el caso de series de tiempo, se considera la media para cada instante de tiempo

Algorithm 3 Procedimiento de Asignación a clústeres finales

Input: $\mathcal{Y}$ el conjunto de MTS, L el número de elementos tras la reducción de datos, $N_1, \ldots, N_m$ el número de clústeres para cada los conjuntos $\mathcal{Y}^{(1)}, \ldots, \mathcal{Y}^{(m)}$.

Output: La partición $\{E_{\mathbf{k}} : \mathbf{k} \in [N_1] \times \cdots \times [N_m]\}$

$C = \bar{\mathbf{c}}_1, \ldots, \bar{\mathbf{c}}_L \leftarrow NumReduction(\mathcal{Y}, L)$
for $i \leftarrow 1 \ldots m$ **do**
  $\mathbf{d}_{i,1}, \ldots, \mathbf{d}_{i,N_i} \leftarrow ClassReduction(C^{(i)}, N_i)$
end for
for each $\bar{\mathbf{y}} \in \mathcal{Y}$ **do**
  $\bar{\mathbf{c}} := (\mathbf{c}_1, \ldots, \mathbf{c}_m) \leftarrow \arg\min_{\bar{\mathbf{c}} \in C} dist(\bar{\mathbf{c}}, \bar{\mathbf{y}})$
  for $i \leftarrow 1 \ldots m$ **do**
    $k_i \leftarrow \arg\min_j dist(\mathbf{c}_i, \mathbf{d}_{i,j})$
  end for
  $E_{(k_1,\ldots,k_m)} \leftarrow E_{(k_1,\ldots,k_m)} + \bar{\mathbf{y}}$
end for

3.3. Metodología de valoración del *input* del proceso de simulación

Dado un par $(\mathbf{x}, \bar{\mathbf{y}}) \in \mathcal{A}$, como resultado de la metodología de clasificación se puede construir la aplicación $\Psi : \mathcal{A} = \mathcal{X} \times \mathcal{Y} \longrightarrow \mathcal{X} \times \mathcal{K}$ (con $\mathcal{K} = [N_1] \times \cdots \times [N_m]$) tal que $\Psi(\mathbf{x}, \bar{\mathbf{y}}) = (\mathbf{x}, \mathbf{k})$ si y solo si $\bar{\mathbf{y}} \in \mathcal{E}_{\mathbf{k}}$. Con las notaciones de la sección anterior, $\Psi = (I_{\mathcal{X}}, \beta \circ \alpha)$, donde $I_{\mathcal{X}}$ representa la función identidad en $\mathcal{X}$. Se supone que $\mathcal{X}$ es discreto de dimensión n (es decir, la entrada está determinada por p características discretas).

sean $X_1, \ldots, X_p$ las variables aleatorias que representan cada característica del **input**, como $\mathcal{K}$ es discreto, se puede considerar la variable aleatoria K con soporte en $\mathcal{K}$, y determinar la "importancia" de cada característica X_i del *input* sobre el clúster final asignado K como $IM(X_i, K)$.

Por otro lado, fijado en clúster $\mathcal{E}_{\mathbf{k}}$, se puede considerar el problema binario de pertenencia de la serie temporal al clúster. Así, se considera la función

$$\Phi_{\mathbf{k}}(\bar{\mathbf{y}}) = \begin{cases} 0 & \text{si } \bar{\mathbf{y}} \in \mathcal{E}_{\mathbf{k}} \\ 1 & \text{si } \bar{\mathbf{y}} \notin \mathcal{E}_{\mathbf{k}} \end{cases}$$

De forma que, para cada elemento de $\mathcal{A}$ y para cada clúster final, se puede valorar la correlación existente entre cada característica del *input* y el propio clúster. Para ello, sea Z la variable aleatoria en $\{0, 1\}$ con función de probabilidad inducida por $\Phi(\mathcal{Y}) = \{\Psi(\bar{\mathbf{y}} : \ \bar{\mathbf{y}} \in \mathcal{Y})\}$, se considera $IM(X_i, Z)$.

3.4. El conjunto de datos. Análisis preliminar

En la **Sección 1.1** se describió el origen de los datos, lo que representan y su aplicación. Sea $\mathbf{x} = (x_1, \ldots, x_{12})$ una respuesta dada al cuestionario Crossroads y sea $\mathbb{S}$ el proceso de simulación, un escenario se representa como la MTS $\bar{\mathbf{y}} = (\mathbf{y}_1, \mathbf{y}_2) = \mathbb{S}(\mathbf{x})$. Donde, $\mathbf{y}_1$ es la simulación de la evolución de la temperatura e $\mathbf{y}_2$ de la economía global (en términos del PIB).

Con las notaciones introducidas en la **Sección 3.1**, $\mathcal{X}$ representa las respuestas al cuestionario para las cuales es posible realizar una simulación, $\mathcal{Y}$ representa el conjunto de escenarios resultantes de cada simulación, e $\mathcal{Y}^{(1)}$ e $\mathcal{Y}^{(2)}$ son el conjunto de series univariantes de simulaciones de temperatura y de PIB respectivamente. Se dice que dos UTS, una de temperatura y otra de PIB, están asociadas si son componentes de una misma MTS.

El principal problema que presenta este enfoque es que dos series temporales son consideradas iguales si lo son cada una de sus componentes. Sin embargo, en la **Figura 3.1** se muestran diferentes series de simulaciones que representan una misma situación final pero que no son idénticas. Además, existe la posibilidad de que series de temperaturas "casi idénticas" estén asociadas con series de PIB muy dispares, y viceversa. En la **Figura 3.2** se ejemplifica este caso.

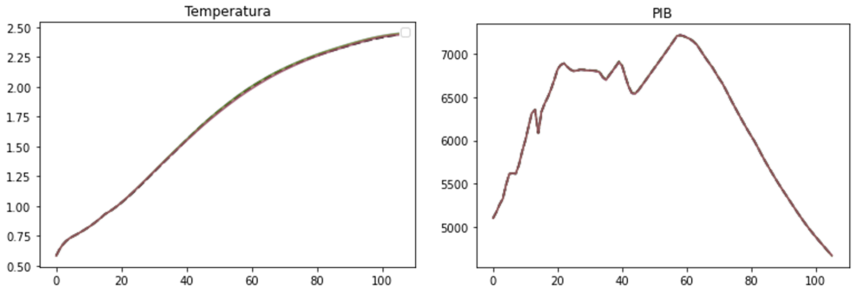


Figura 3.1: Representación de seis MTS de temperatura y PIB DE distintas simulaciones tales que, fijado un escenario $\bar{\mathbf{y}} = (\mathbf{y}_1, \mathbf{y}_2)$ de los representados se verifica que, para cualquiera de las series consideradas $\bar{\mathbf{z}} = (\mathbf{z}_1, \mathbf{z}_2)$ y para cualquier instante de tiempo $i \in \{1, \ldots, 106\}$, $|y_{1,i} - z_{1,i}| \leq 0,025\ |y_{1,i}|$ e $|y_{2,i} - z_{2,i}| \leq 0,025\ |y_{2,i}|$.

Por otro lado, hay algunas combinaciones de respuestas al cuestionario de Crossroads para las cuales no existe un resultado de simulación realista (la temperatura decrece al orden de -10^{33}). Este comportamiento particular se ha asociado al funcionamiento interno del propio modelo de generación. En particular, hay 19678 respuestas con simulaciones incoherentes.

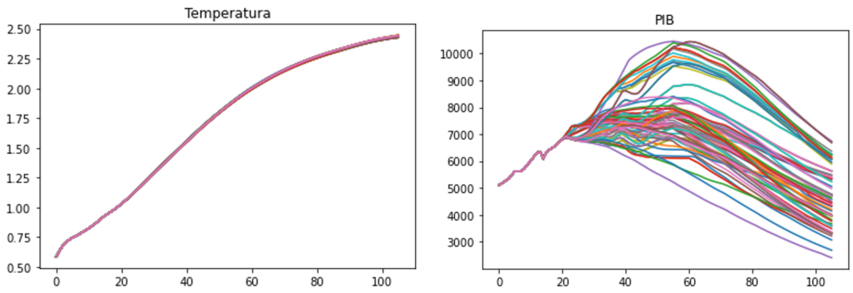


Figura 3.2: Representación de escenarios con similar evolución de temperatura y una amplia variedad en la evolución del PIB

3.5. Implementación de la metodología

Para la implementación de la metodología se utiliza el lenguaje de programación *Python*. En particular, se recurre al módulo `sklearn`:

- `sklearn.cluster.KMeans`: algoritmo de **k-medias**.
- `sklearn.cluster.AgglomerativeClustering`: algoritmo de **clasificación jerárquica aglomerativo**
- `sklearn.feature_selection.mutual_info_classif`: estimación de la información mutua

Además, se ha utilizado los módulos `pandas` y `numpy` para estructurar la información de las simulaciones, que se encuentra almacenada en diferentes documentos de texto plano estructurado en forma de tabla.

Dado un conjunto de datos $D = \{d_1, \ldots, d_n\}$, una partición $B_1, \ldots, B_k$ se representa como una lista $(\delta_1, \ldots, \delta_n)$ con $1 \leq \delta_i \leq k$, tal que $d_i \in B_{\delta_i}$. A esta lista se le denomina ***labels***. Además, se pueden considerar los representantes o centros de cada clúster, $\mu_1, \ldots, \mu_k$

Para la implementación de la metodología de clasificación, así como la evaluación por la información mutua se han implementado tres clases:

- `Clasificador`: representa un modelo de particionado básico. Está compuesto de los *labels* (atributo `labels_`) los representante de cada clúster (atributo `clusters_centers_`).
- `ModeloDoble`: representa un modelo de clasificación para MTS de dos componentes. Es decir, contiene las clasificaciones independientes para la temperatura y el PIB, cada una determinada por sus *labels* y sus centros, definidos como instancias de la clase `Clasificador`.
- `modelo2Fases`: clase principal que representa todo el proceso de clasificación según la metodología (**Sección 3.2**) como de evaluación de la relación de las entradas del proceso

de simulación con el clúster asignado a su salida (**Sección 3.3**). Además, dispone de un conjunto de funcionalidades adicionales que permiten la evaluación del clúster y el análisis de la distribución de las diferentes respuestas en cada clúster.

La implementación de las clases anteriores se recoge en el fichero `clasesAgrupacion.py`, mientras que la generación del modelo, basado en estas clases, esta implementada en el fichero `modelo2FasesMain.py`. También esta presente un fichero de configuración `conf.py` donde se determinan ciertos parámetros de ejecución.

3.6. Resultados

En esta sección se presenta el primer modelo construido, que identifica un número reducido de escenarios característicos diferentes entre sí para su posterior análisis. Para ello se sigue la metodología propuesta en la **Sección 3.2**.

Esta metodología ha sido diseñada para satisfacer las peculiaridades descritas en la sección anterior. En la primera fase, el algoritmo de k-medias con un número de clústeres lo suficientemente grande busca agrupar aquellas MTS que son muy similares entre sí. Posteriormente, con el algoritmo jerárquico se identifican los patrones para los resultados de temperatura y PIB, que se combinan para formar los escenarios característicos.

Previo a la clasificación, se han estandarizado de forma independiente los conjuntos de las series de temperatura y de PIB, pues presentan valores con diferentes magnitudes (inferior a 10 en el caso de la temperatura y superior a 1000 en el caso del PIB).

3.6.1. Construcción del modelo: determinando los hiperparámetros

Para construir el modelo es necesario determinar los hiperparámetros que lo gobiernan: número de clústeres para el algoritmo de k-medias y para las clasificaciones jerárquicas de temperatura y PIB.

Número de clústeres en la primera fase: k-medias

En primer lugar, se determina experimentalmente, mediante el método del codo (**Sección 2.2.2**) aplicado a la distancia intra-clúster, el número de particiones a realizar en la fase de *reducción*, representado en la **Figura 3.3**.

En la **Tabla 3.1** se representa la evolución de la variación intra-clúster con respecto a diferentes configuraciones. Se ha elegido 200 particiones, pues para este número se aprecia un cambio en el decremento de la variación intra-clúster, y en esta primera etapa es preferible tener clústeres similares a un clúster con elementos muy heterogéneos.

Cabe destacar que, las particiones obtenidas al aumentar el números de clústeres tienen un efecto reductor similar sobre la variación intra-clúster de la temperatura y del PIB, lo que indica que ninguno de los indicadores considerados esta dominando y condicionando el particionado.

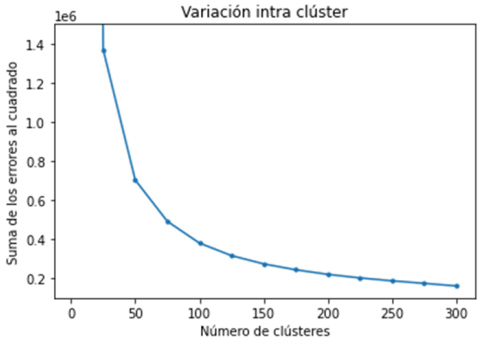


Figura 3.3: Distancia intra-clúster para el PIB y la temperatura en conjunto

Número de clústeres	**Variación intra-clúster Temperatura**	**Variación intra-clúster PIB**	**Variación intra-clúster total**
50	368654	334380	703034
100	198670	181808	380478
150	139909	134163	274072
200	112291	108177	220468
250	970578	90309	187366
300	84396	79877	161273

Tabla 3.1: Variación total intra-clúster para el PIB y la temperatura con diferentes configuraciones. Las medidas de error que se presentan se aplican sobre los datos estandarizados.

Número de clústeres en la segunda fase: algoritmo jerárquico

Para determinar el número de divisiones a realizar mediante el clústering jerárquico aglomerativo, considerando de forma independiente los conjuntos de simulaciones de temperatura y PIB, se aplican técnicas heurísticas y experimentales sobre el dendograma (**Figura 3.4**) de cada indicador. En el caso de la temperatura se construyen 5 particiones, y seis para el PIB.

3.6.2. Resultados de la clasificación

Patrones identificados

En la **Figura 3.5** se pueden observar los representantes de cada partición para el clúster jerárquico de la temperatura y del PIB. En gris se han representado las series resultantes de

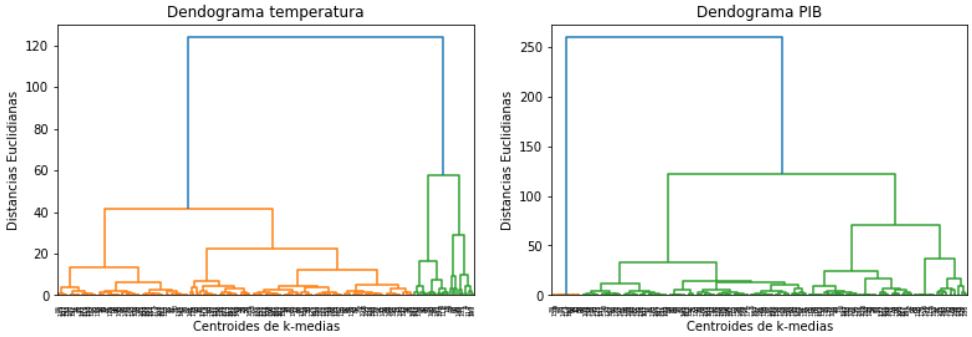


Figura 3.4: Dendograma para la agrupación jerárquica de la temperatura (izquierda) y PIB (derecha)

aplicar la *reducción del cuerpo de datos*.

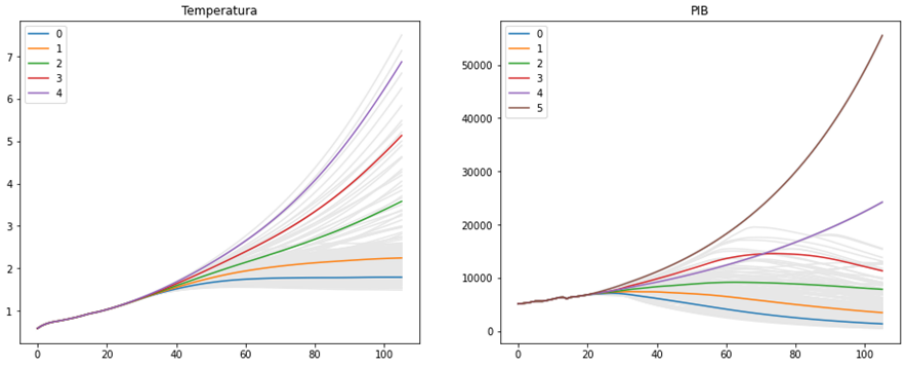


Figura 3.5: Representación de las evoluciones de temperatura y del PIB extraídas mediante k-medias del conjunto original (gris) y sus respectivos patrones característicos

En esta figura se puede apreciar la principal ventaja de aplicar la reducción de datos, dado que, al agrupar series muy similares en un único elemento, se evita que las series con comportamientos más extremos y menos comunes se diluyan en los casos más usuales al aplicar la media para obtener los representantes.

Construcción de los escenarios característicos

Como resultado de combinar los patrones de temperatura y PIB obtenidos se obtendrían 30 posibles escenarios característicos. Sin embargo, el procedimiento de clasificación de los resultados de simulación muestra que no todos los escenarios característicos son posible, lo que es

coherente con el conocimiento del tema[2].

Así, se obtiene que solo son posibles 17 escenarios característicos. Además, cerca del 80 % de las simulaciones son asignadas a cinco escenarios concretos, lo que indica que algunos escenarios característicos requieren respuestas muy concretas a determinadas preguntas.

Clúster Temperatura \ Clúster PIB	0	1	2	3	4	5
0	**30**	**15**	**5**	0*	0	0
1	**9**	**22**	**11**	1	1	0*
2	0	0*	2	0	1	1
3	0	0	0	0	0*	1
4	0	0	0	0	0	1

Figura 3.6: Porcentaje de escenarios asignados a cada clúster. Los patrones de cada clúster están definidos en la Figura 3.5. Los valores marcados con un $*$ indican un porcentaje inferior a 0,5 % no nulo.

Errores cometidos

Además de obtener un número reducido de escenarios característicos que permita su análisis, la clasificación propuesta permite sustituir cada MTS de simulación por el patrón característico del correspondiente escenario. En la **Tabla 3.2** se presenta el RMSE promedio del error cometido al sustituir cada MTS por el representante de cada clúster. Teniendo en cuenta que los datos están estandarizados, y que el RMSE se puede interpretar como la desviación estándar de los residuos, el error obtenido es aceptable e indican un buen ajuste (pues la dispersión de los residuos es inferior a la dispersión natural de los datos, con valor de 1 por estar estandarizados).

[2]Los escenarios con un PIB que crece exponencialmente se logran consumiendo gran cantidad de combustibles fósiles para producir la energía necesaria que permita este crecimiento. Por tanto, las emisiones de gases de efecto invernadero no se reducen y la temperatura global crece. Por otro lado, escenarios más conservadores en los que se logra la estabilización de la temperatura en valores próximos a los 2ºC requieren la reducción de emisiones, lo que conlleva el abandono progresivo de los combustibles fósiles en energías renovables y ciertos sacrificios sobre el modelo económico actual.

Clúster Tem/PIB	0	1	2	3	4	5
0	0,139	0,145	0,217	0,131		
1	0,03	0,159	0,154	0,132	0,204	0,164
2		0,076	0,113		0,088	0,118
3					0,166	0,303
4						0,069

Tabla 3.2: Media del error cuadrático medio cometido en cada clúster final al sustituir la serie multivariable por el centroide. La temperatura y el PIB están estandarizados.

Por otro lado, para interpretar el error cometido en la sustitución de cada los escenario simulado por su correspondiente escenario característico, se utiliza el error máximo. Los resultados se recogen en la **Tabla 3.3**.

		Temperatura (ºC)	PIB *per Capita* ($)
RMSE	max	0,41	3344,73
	min	$9{,}7 \cdot 10^{-4}$	18,32
	μ	0,082	1355,61
	σ	0,051	445,24
Error Máximo	max	1,08	7526,86
	min	$2{,}4 \cdot 10^{-3}$	24,43
	μ	0,17	1355,61
	σ	0,12	971,63
Error Máx. Relativo	max	0,30	8,6
	min	$1{,}4 \cdot 10^{-3}$	$2{,}3 \cdot 10^{-3}$
	μ	0,081	0,35
	σ	0,047	0,34
Rango dinámico	max	7,151	55787,8
	min	0,585	275,8
	μ	1,58	6117,45
	σ	0,53	3878,43

Tabla 3.3: Media del error cometido en cada clúster final al sustituir la serie multivariable simulada por el escenario característico correspondiente.

En general, el error promedio cometido es pequeño, teniendo en cuenta el dominio de cada indicador. Así, para la temperatura, cuyo valor promedio es de 1,583 y su valor máximo es de 7,715, se comente un error maximo medio de 0,17 (aproximadamente un 10 % de la temperatura promedio). En el caso del PIB, con valores promedio de 6193 y valor máximo de 55787, el error máximo medio que se comete es de 1355 (aproximadamente del 20 % del PIB promedio). En términos relativos, de media se comente un error del 8 % en la estimación de la temperatura y del 34 % en la del PIB.

El PIB justifica que a este modelo se le haya denominado "*modelo básico*": la variabilidad

de escenarios en el PIB es bastante grande, con errores no despreciables. Esto se debe a que series con un comportamiento similar (crecen hasta cierto punto y luego decrecen) toman valores muy diferentes en instantes concretos (por ejemplo, en su máximo o en el momento final) y se agrupan juntas.

Este modelo presenta otro inconveniente que se detalla en el **Capítulo 4**. Sin embargo, teniendo en cuenta que los datos son simulados y que los errores en el máximo tienen un carácter puntual no generalizado (pues el RMSE es pequeño y aceptable), este modelo es una buena reducción del número inicial de escenarios, en más del $99,9\,\%$, y en subconjuntos cuyos representantes (escenarios característicos) tienen comportamientos claramente diferenciados.

3.6.3. Evaluación del clústering: Coeficiente de la silueta

Para evaluar el particionado realizado se recurre al método de la silueta (**Sección 2.4.2**). Dado el elevado número de muestras, para disponer de un cálculo que se pueda ejecutar con los recursos disponibles, se utiliza la implementación de la librería `sklearn.metrics` denominada `silhouette_score`.

El coeficiente de la silueta promedio obtenido para este modelo es de 0,23. Dado que es mayor que cero, esto indica que, de forma global, los clústeres están separados y son "naturales". Sin embargo, dado que el valor obtenido no es próximo a 1, esto implica que los clústeres no son tan compactos como deberían, y que para ciertos elementos específicos estos están mal clasificados, provocando alguna sobreposición puntual de clústeres.

El procedimiento de asignación es el responsable de la existencia de estos elementos "mal clasificados" (son mas próximos a otro clúster que al suyo). Generalmente, este error se produce cuando para cierto concreto, su representante en la fase de reducción se obtiene con un error que sobrestima o subestima la MTS, y el escenario característico se vuelve a obtener sobrestimando o subestimando al representante de reducción respectivamente.

En la **Figura 3.7** se muestra, para cada clúster final construido por el procedimiento de asignación definido en **Algoritmo 3**, el porcentaje de miembros asignados a otros clústeres segun el principio de mínima distancia. La notación de los clústeres se realiza con formato XY donde X representa el clúster asociado a los incrementos de temperatura e Y al PIB.

Teniendo en cuenta que los clústeres no están distribuidos de forma uniforme en lo que a su población se refiere, por lo que el número de MTS que se reasignarían con el principio de mínima distancia es cercano al $10\,\%$. Cabe destacar que las diferencias de asignaciones se producen entre clústeres próximos, donde los patrones de temperatura y PIB presentan un orden en sentido creciente. Por ejemplo, el clúster 2 (02) y el clúster 3 (03) solo se produce un cambio en el patrón de PIB por el menor de los mayores. Análogamente, entre el clúster 3 (03) y el clúster 13 se produce un cambio en el patrón de temperatura por el inmediatamente superior.

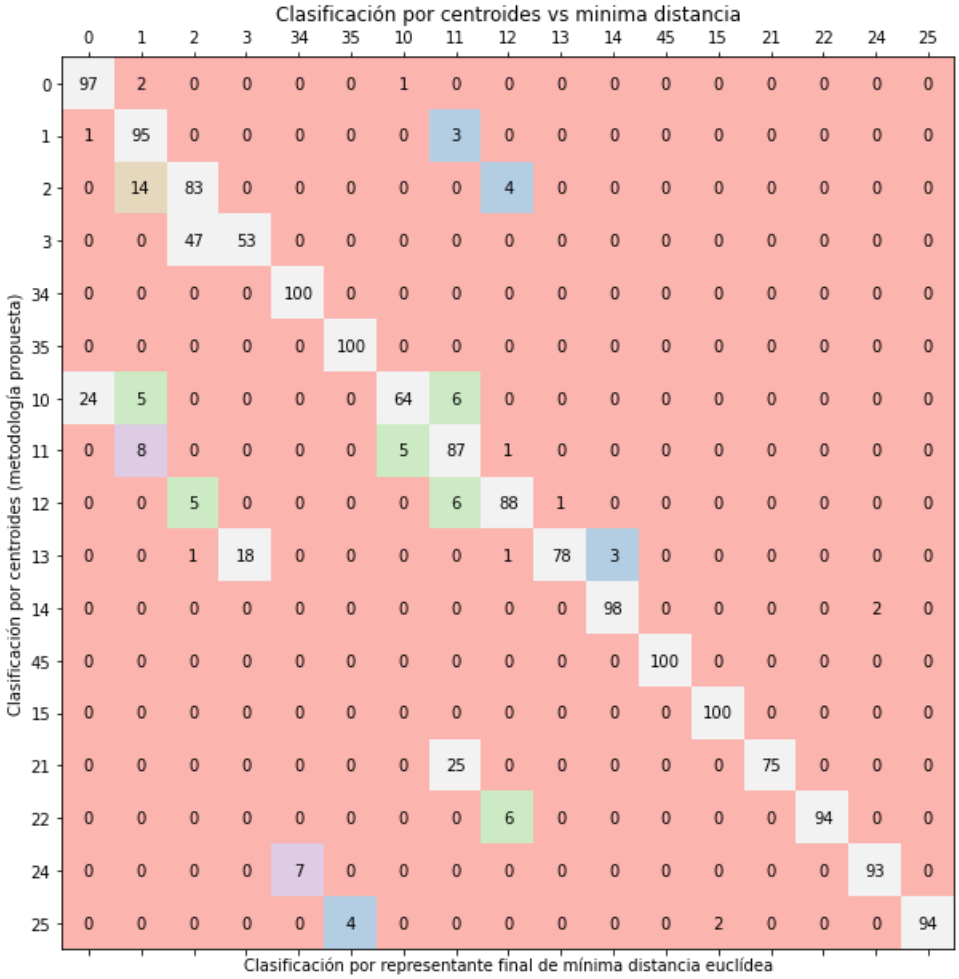


Figura 3.7: Asignaciones de las MTS a clústeres según la metodología de clasificación frente al principio de mínima distancia. Se indica el porcentaje de los miembros del clúster por la asignación propuesta que cambian al aplicar el principio de mínima distancia.

3.6.4. Valoración de la metodología: consistencia

En esta sección se pretende evaluar la sensibilidad del modelo construido a la desaparición de datos y/o existencia de *outliers*, aplicando la técnica descrita en la **Sección 2.4.3**. Para ello se eliminan de forma aleatoria un 20 % de las muestras, y se procede a calcular el coeficiente c (2.13) entre la agrupación original y la nueva agrupación obtenida. Los resultados obtenidos se presentan en

Semilla aleatoria	**Coeficiente** c
0	0,934
5	0,942
10	0,787
15	0,834
20	0,891
25	0,984
30	0,975
35	0,894
40	0,93
45	0,891

Tabla 3.4: Coeficientes c para diferentes eliminaciones aleatorias de datos

En consecuencia, con una probabilidad del 95 % y bajo hipótesis de normalidad, el verdadero valor del coeficiente c se encuentra en el intervalo $(0,87; 0,95)$. Es decir, dado que se obtiene un coeficiente muy próximo a 1, la metodología anterior es poco sensible a la supresión de pequeños subconjuntos de datos. Además, los clústeres definidos son consistentes a la pérdida de miembros.

3.6.5. Completando los datos

De forma externa a este trabajo, utilizando las agrupaciones realizadas por el modelo se ha construido un predictor que, dadas las respuestas al cuestionario determine el clúster que se le asocia. La tasa de acierto de este predictor es superior al 95 %.

Para las respuestas al cuestionario que no admitían un resultado de simulación, mediante el predictor anterior han sido asignadas a un clúster concreto. Además, como escenario de temperatura y PIB para estas combinaciones de respuestas, se toma el represente del clúster predicho. A este conjunto de respuestas y simulaciones se le denomina conjunto completo, pues con las notaciones anteriormente descritas se verifica que $\mathcal{X} = \mathcal{H}$.

3.7. Valoración del *input* en función del clúster asignado.

En esta sección se recogen los resultados del análisis de información mutua y se ilustra con el estudio de dos casos en concreto. Los resultados obtenidos para el total de los clústeres finales se presentan en el **Apéndice B**.

Para aplicar la Información Mutua, se considera cada hipótesis y medida política como una variable aleatoria X_i con soporte en las respuestas que admite con probabilidad inducida por la proyección sobre la componente $i-$ésima de $\mathcal{X}$. Denotaremos por Y la variable aleatoria con probabilidad inducida por el problema de pertenencia a un clúster concreto y cuyo soporte es $\{0, 1\}$.

En general, se puede observar que para cualquier clúster los valores de información mutua de hipótesis y medidas son cercanos a cero, pero no nulos. Es decir, el escenario característico depende de cada una de las hipótesis o medidas, y ninguna de ellas, por si sola, es determinante en un escenario característico concreto. En consecuencia, no existen preguntas redundantes en el cuestionario.

Sin embargo, sí que se puede observar cómo algunas respuestas a ciertas preguntas del cuestionario tienen más influencia que otras en el clúster asignado. Para este caso destacan dos situaciones diferentes:

- En la **Figura 3.8** se puede observar como en el valor de información mutua para el problema de la pertenencia destaca en dos hipótesis (**h1** y **h2**) y una medida política (**m1**). Los tres casos destacan porque admiten una única opción, es decir, permite establecer condiciones necesarias de pertenencia (si la respuesta a **h1** no es "c", entonces el resultado de simulación no pertenece al clúster).
- En la **Figura 3.9** se presenta la situación opuesta. En este caso, una hipótesis (**h2**) y una medida (**m1**) destacan sobre las demás, entre otras cosas, porque permiten establecer un valor que automáticamente descarta la pertenencia (opciones "a" y "c" respectivamente).

Por otro lado, se puede observar que existen algunas medidas políticas que, pese a ser necesarias, nunca destacan sobre las demás. Este es el caso de las medidas **m6** y **m8**. Cabe destacar que estas medidas solo admiten dos opciones posibles, por lo que las relaciones de dependencia son menores.

Nota sobre la normalización de la Información Mutua

Con las notaciones anteriores, si $\mathcal{X} = \mathcal{H}$ entonces cada combinación de respuestas al cuestionario (posible *input*) se utilizaría una y solo una vez en el proceso de simulación. En consecuencia, la variable aleatoria X_i sería equiprobable y $H(X_i) = \log(q)$, donde q denota el número de respuestas diferentes que admite la pregunta $i-$ésima.

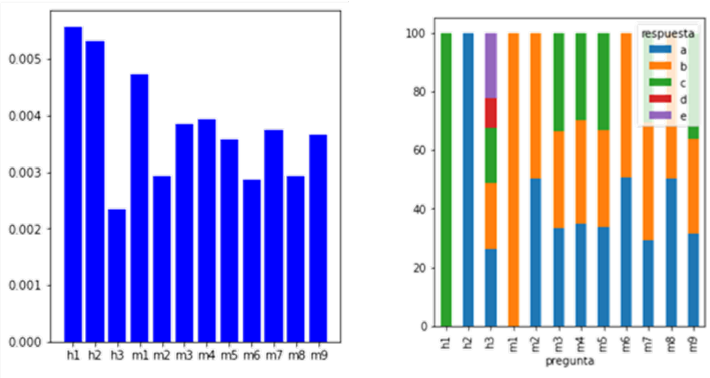


Figura 3.8: Distribución de las preguntas e información mutua estimada, ejemplo primero.

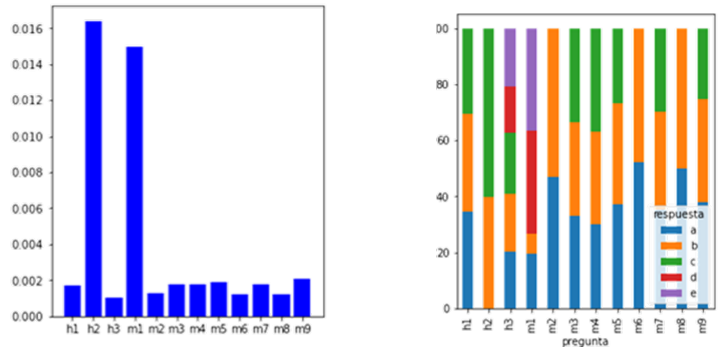


Figura 3.9: Distribución de las preguntas e información mutua estimada, ejemplo segundo.

Además, Y es una variable aleatoria binaria, luego $H(Y) \leq \log(2)$. Por tanto, en este caso se verifica que $H(Y) \leq H(X_i)$ para cualquier $i = 1, \ldots, 12$, y $IM(X_i, Y) \leq H(Y) \leq H(X_i)$, y se normaliza con la entropía de Y (**Sección 2.5.1**).

Aunque en el conjunto de datos considerado $\mathcal{X} \neq \mathcal{H}$, dado que existen algunas combinaciones de respuestas que no admiten resultado de simulación, este error no se ha asociado con ninguna respuesta concreta a una pregunta determinada, manteniéndose casi equidistribuidas las respuestas dadas a cada pregunta del cuestionario. En consecuencia, el razonamiento anterior es válido (más teniendo en cuenta que en Y la diferencia entre las probabilidades de los dos sucesos es como mínimo de $0,4$).

3.8. Conclusiones

En este capítulo se ha propuesto una metodología para la clasificación de datos de simulación estructurados como MTS. Esta metodología realiza una doble clasificación. Primero, mediante el algoritmo de k-medias se reduce el conjunto original de datos, pasando de más de 400000 elementos a un segundo conjunto de 200.

En segundo lugar, se considera el conjunto reducido para proceder a la extracción de los patrones característicos de cada una de las componentes de las MTS (temperatura y PIB), que a su vez son UTS. Esta identificación de patrones se realiza de forma independiente, aplicando un algoritmo de clasificación jerárquica aglomerativo sobre el conjunto formado por una componente concreta de todas las MTS del cuerpo de datos reducido.

Finalmente, se procede a la construcción de los escenarios característicos, como combinaciones de los patrones de las componentes extraídas anteriormente. Cada escenario es el representante de un clúster, al que se asignan parte de los datos iniciales. Los parámetros que dominan el modelo (número de elementos tras la fase de reducción, número de clústeres de temperatura y PIB) se han determinado utilizando técnicas heurísticas, basadas en el método del codo y en el uso de dendogramas.

Sin embargo, la distribución en clústeres no es uniforme, perteneciendo cerca de la mitad de los datos a solo dos clústeres, y conteniéndose el 90 % de los datos en cinco. En las **Figuras 3.6** y **3.10** se presenta la distribución de las MTS de simulación en los diferentes clústeres finales.

En consecuencia, este modelo nos permite identificar ciertos datos atípicos o poco comunes. Sin embargo, el hecho de obtener clústeres muy grandes implica el riesgo de haber unificado dos conjuntos que de forma "natural" puedan estar por separados. Por otro lado, el utilizar un modelo con reducción de datos ha supuesto que los datos extremos o poco comunes no se "diluyan" en los clústeres más usuales, ni que se forme un clúster de *outliers* donde el error de aproximación por el centroide final fuera grande.

En cuanto al modelo construido, la valoración mediante el coeficiente de la silueta muestra que se obtienen "agrupaciones naturales", aunque admite margen de mejora. Como alternativa, se sugiere modificar la técnica de asignación a clústeres, asignando por el representante (escenario característico) de mínima distancia con la MTS simulada. Sin embargo, el modelo es bastante bueno en términos de obtención de patrones diferentes entre sí, siendo cada escenario característico único. Esto ha llevado a obtener un análisis diferente para cada escenario.

Además, se ha justificado que la metodología utilizada para el conjunto de datos en cuestión es poco sensible a la pérdida de elementos, y que los clústeres construidos son consistentes pese a perder algunos de sus miembros.

Para realizar este análisis, se ha tenido en cuenta que cada MTS simulada es el resultado de aplicar un proceso de simulación sobre unas variables de entrada. Así, relacionando las posibles variables de entrada con su escenario característico ha sido posible construir un *feedback* de cada escenario, validado por un experto en la materia.

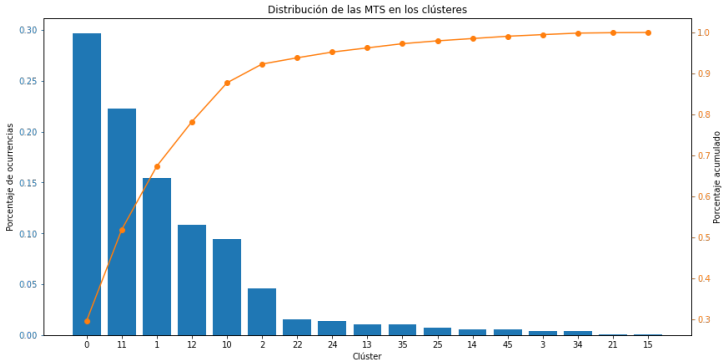


Figura 3.10: Distribución de los elementos en cada clúster

Discusión sobre el procedimiento de asignación a clústeres

En la **Figura 3.7** se comparan dos procedimientos de asignación a clústeres, obteniendo resultados con ciertas variaciones en función del procedimiento que se aplique. Aun así, cabe destacar que los clústeres a los que se asigna cada elementos, cuando son diferentes en función del procedimiento, sus representantes son próximos entre sí.

La elección del método de asignación según el representante de la fase de reducción presenta la ventaja de tener complejidad lineal en el tiempo, dependiente únicamente del número de elementos que formen el conjunto inicial. Sin embargo, la asignación por centroide final de mínima distancia tiene complejidad cuadrática, dependiente del número de elementos a clasificar y el número de clústeres existentes.

Por otro lado, el primer método de asignación trata de igual manera a todas las series que tengan un mismo representante en la fase de reducción, lo que se traduce en un incremento del RMSE. Así, para clasificar un nuevo elemento por el modelo, en la primera técnica de asignación sería necesario obtener su representante en la fase de reducción y el clúster al que se asigna dicho representante. Es decir, esta asignación mantiene la coherencia de trabajar con el conjunto de datos reducido. En el segundo método de asignación, un nuevo dato se clasificaría utilizando los representantes finales.

Capítulo 4

Recomendador

El objetivo final de este trabajo es construir un sistema de diagnosis autónomo, que se integre en el videojuego Crossroads, y que una vez que el jugador responde al cuestionario y obtiene su escenario simulado:

- Presente al jugador una breve descripción y diagnosis de su escenario simulado.
- Valore el cumplimiento de los objetivos que el jugador se propuso.
- Si procede, presente una serie de recomendaciones sobre las medidas políticas del jugador para llegar a un escenario medioambientalmente adecuado con una economiza equilibrada.

A dicho sistema autónomo se lo denomina **Recomendador**.

4.1. Diseño y funcionamiento

4.1.1. Requisitos

En esta sección se abordan los requisitos funcionales y no funcionales del sistema de recomendación propuesto. El Recomendador es un sistema con un único requisito de usuario: obtener una recomendación y evaluación para una respuesta a todas las cuestiones del videojuego *Crossroads* (incluyendo la determinación de objetivos y las respuestas a las preguntas del cuestionario que determinan los escenarios de simulación). Este único requisito, denominado **Generar Feedback** (**RU-01**) se describe en la **Tabla 4.2**.

Los requisitos de información recogidos en el caso de uso **RU-01** se recogen en la **Tabla 4.3**.

Id	Actor	Descripción
A-01	Crossroads	El videojuego *Crossroads* interactúa con el Recomendador de forma que este se integre como servicio REST externo es su funcionamiento.

Tabla 4.1: Actores del Recomendador

RU-01	**Generar Feedback**
Actor(es)	Crossroads (**Act-01**)
Descripción	El Recomendador deberá comportarse como indica el siguiente caso de uso cuando Crossroads solicite un *feedback*
Precondición	Existen unos parámetros predefinidos requeridos en el paso segundo de la secuencia normal
Flujo normal	1. Crossroads solicita un *feedback*, para lo que da las respuestas del jugador al cuestionario (**RI-03**), los objetivos de temperatura (**RI-01**) y de PIB (**RI-02**) que el jugador asumió. 2. El sistema, utilizando el modelo de clasificación construido (**RI-04**) elabora el *feedback*, y se lo entrega a Crossroads.
Flujos alternos	2. Si se produce un error durante la ejecución, el sistema notificará el error en lugar del *feedback*.
Postcondición	-

Tabla 4.2: Descripción del requisito de usuario del Recomendador **RU-01**

Id	**Req. Información**	**Descripción**
RI-01	Objetivo de temperatura	Objetivo de temperatura que el jugador determina en el videojuego Crossroads.
RI-02	Objetivo de PIB	Objetivo de PIB que el jugador determina en el videojuego Crossroads.
RI-03	Respuestas	Respuestas al cuestionario de doce preguntas dadas por el jugador en el videojuego Crossroads, y que determinan los escenarios de simulación.
RI-04	Modelo	Modelo de agrupación, en sentido amplió, que establece la relación entre las respuestas dadas al cuestionario y el escenario característico asociado.

Tabla 4.3: Requisitos de Información del Recomendador

En cuanto a los requisitos funcionales, se recogen en la **Tabla 4.4**. Además, debido a la naturaleza del juego se han identificado dos requisitos no funcionales, relacionado con cómo se interacciona con el Recomendador y con el tiempo de respuesta esperado (**Tabla 4.5**).

Id	Req. Funcional
RF-01	Recepción y procesamiento de peticiones HTTP, con la información en un formato definido para extraer los objetivos de temperatura y PIB y las respuestas al cuestionario.
RF-02	Generación autónoma del *feedback* de cumplimiento de objetivos.
RF-03	Generación autónoma del *feedback* de adecuación del escenario obtenido a unos objetivos globales de temperatura y PIB.
RI-04	Generación autónoma de un conjunto de cambios mínimos en las respuestas del jugador que le permitan alcanzar un escenario con valores económicos y medioambientales equilibrados.

Tabla 4.4: Requisitos Funcionales del Recomendador

Id	Req. no Funcional
RnF-01	El sistema se construirá como un servicio REST externo e independiente a Crossroads.
RnF-02	El sistema deberá ser capaz de generar las respuestas en menos de 5 segundos.

Tabla 4.5: Requisitos no Funcionales del Recomendador

4.1.2. Explicación funcional del Recomendador

El Recomendador utiliza el modelo de agrupación construido. Cada posibles combinación de respuestas al cuestionario tienen asociado un único escenario característico (representante del clúster final al que pertenece la MTS generada por el proceso de simulación), de forma que el Recomendador utiliza estos escenarios en lugar de los escenarios simulados.

Como entrada, el Recomendador requiere los objetivos de temperatura y de PIB que el jugador se determinó, así como las doce respuestas dadas al cuestionario del juego (que permiten determinar el escenario de simulación y por extensión el clúster y e escenario característico). Además, se alimenta de un conjunto de datos estáticos (su estructura de almacenamiento, utilizando varios ficheros, se recoge en la **Sección 4.3**):

- Asociación de cada combinación de respuestas con su escenario característico.
- Descripción realizada por un experto en la materia de cada uno de los escenarios característicos.
- Los enunciados de cada preguntas del cuestionario y la relación de orden entre las respuestas, según el valor numérico dado para el último año.
- Valores numéricos asociados a cada opción posible de objetivos de temperatura y PIB.
- Un subconjunto de escenarios característicos que se han considerado medioambiental y económicamente adecuados (**Sección 4.2**).

El Recomendador sigue un proceso secuencial de cuatro fases, descrito en la **Figura 4.1**:

Fase 1 **Escenario característico.** La entrada al proceso de recomendación es una respuesta a cada una de las doce preguntas que componen el cuestionario. Con ellas, se identifica el escenario característico asociado, utilizado en las tres etapas restantes.

Fase 2 **Objetivos.** En esta fase se obtiene la primera parte del *feedback* sobre el cumplimiento dichos objetivos. Como entrada se recibe los objetivos definidos por el propio jugador y el escenario característico obtenido en la primera fase. Con este último se extraen los valores finales (para el último año simulado) de temperatura y PIB, que se compara con el valor numérico de cada objetivo.

Fase 3 **Valoración.** En esta fase, dado el escenario característico asociado en la primera etapa, se incluye la correspondiente descripción de este, realizada por un experto. Además, utilizando el subconjunto de escenarios finales considerados como adecuados en términos medioambientales y socioeconómicos, se genera una valoración de la proximidad del resultado obtenido con respecto a uno adecuado.

Fase 4 **Recomendación.** Cuando proceda, se propone al jugador una serie de cambios en sus respuestas. Esta fase requiere conocer el escenario característico asociado y el conjunto de escenarios considerados como adecuados.

a) Se seleccionan todas los posibles *inputs* que tengan asignado un clúster de los considerados adecuados.

b) Entre los *inputs* seleccionados, se comprueba si existe alguno con las mismas hipótesis que las respuestas dadas por el jugador. En caso afirmativo, se ignoran todos los *inputs* con hipótesis diferentes.

c) De las respuestas consideradas, se selecciona aquella que requiere el mínimo número de cambios:

1) **Por número de cambios**. En primer lugar, se valora el número de respuestas diferentes entre las dadas por el jugador y cada *inputs* seleccionado. Se mantienen solo los *inputs* que hacen mínimo dicho valor.
2) **Por severidad de cambios.** Teniendo en cuenta que existe un orden en las respuestas a una misma pregunta[1], se elige el *input* cuyos cambios sean más próximos a las respuestas del jugador.

d) Se sugieren los cambios.

En la salida que ofrece el Recomendador se diferencian cuatro componentes:

- Cercanía a los objetivos propuestos por el jugador (resultado de la *Fase 2*).

[1]Generalmente, el valor que representa la respuesta *a* es menor que el valor de *b* que es menor que el de *c*. Además, el valor de la respuesta *d* es menor que el de *e* que es menor que el de *a*. Por ejemplo, en el caso de la hipótesis **m1**, la población en el año 2050 es: a) 8500 millones, b) 9200 millones, c) 10000 millones, d) 5000 millones, e) 7000 millones

- Descripción del escenario incluyendo las consecuencias sobre la vida de producirse, y cercanía con respecto a los escenarios característicos considerados como adecuados (resultado de la *Fase 3*)
- Recomendación de cambios en las medidas políticas e hipótesis cuando procediese y fuere necesario (resultado de la *Fase 4*).

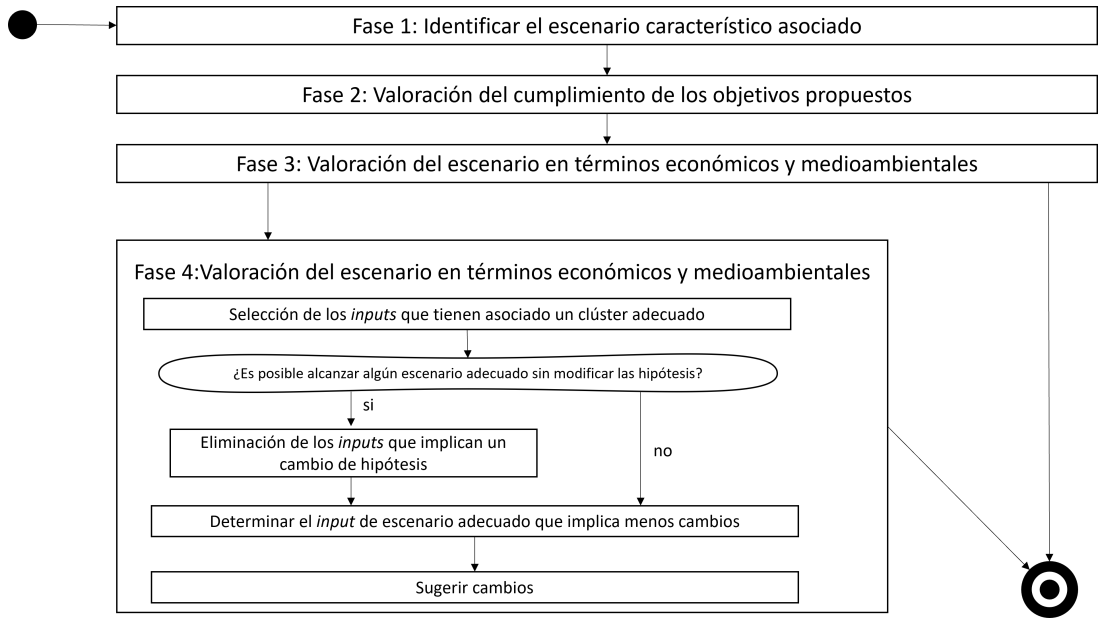


Figura 4.1: Actividades realizadas por el Recomendador

En la **Figura 4.2** se muestra un ejemplo con el funcionamiento del Recomendador. En primer lugar, se presentan las 12 respuestas que el jugador determino como sus hipótesis y medidas políticas. Posteriormente, se muestra el resultado obtenido con esta combinación de respuestas en términos de incrementos de temperatura media y evolución del PIB global. También se ilustran los objetivos de temperatura y PIB que se marco el jugador. Por último se incluye la retroalimentación que da el Recomendador:

- En rojo, la valoración con respecto a los objetivos propuestos (*Fase 1*).
- En verde, la descripción experta del escenario obtenido, donde se explica que ha ocurrido y por qué se ha producido (*Fase 1*).
- En azul, la valoración del escenario en términos económicos y medioambientales (*Fase 3*) y, dado que el escenario obtenido no cumple con los requisitos de idoneidad, un conjunto de modificaciones sobre las medidas políticas que permitan alcanzar un escenario más respetuoso con el medioambiente (*Fase 4*). Finalmente se presentan los cambios a los que habría conducido la recomendación dada.

H1: Disponibilidad de recursos energéticos no-renovables (petróleo, carbón, gas natural y uranio) hasta el año 2050. **Ilimitada (c)**

H2: Impactos producidos por el cambio climático en la economía global. **Ninguno (a)**

H3: Evolución de la población**. Menor crecimiento que el previsto por las tendencias históricas (a)**

M1: Evolución planeada del Producto Interno Bruto (PIB) per capita. **Continuación de las tendencias históricas (b)**

M2: Programa global de reforestación para capturar carbono desde el año 2020. **No (a)**

M3: Energía nuclear**. Mantener la capacidad instalada (b)**

M4: Producción planeada de energía renovable (electricidad y calor). **Mayor crecimiento que las tendencias históricas (c)**

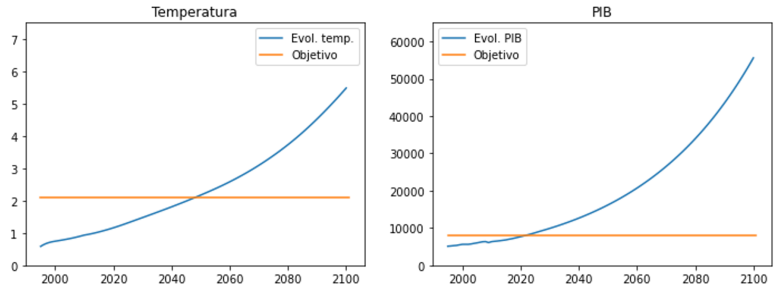
M5: Producción planeada de biocombustibles líquidos (bioetanol, biodiesel, etc.). **Continuar crecimiento histórico (b)**

M6: Transporte terrestre (coches, motos, autobuses, trenes, camiones, etc.). **Transición a vehículos de gas y eléctricos (b)**

M7: Tendencia en la evolución de la eficiencia energética. **Mayor crecimiento que las tendencias históricas (c)**

M8: Tasas de reciclado de minerales necesarios para la transición energética hacia un mayor uso de renovables. **Mejorar los porcentajes actuales (b)**

M9: Emisiones de gases de efecto invernadero no dependientes del consumo energético. **Aumento respecto de la tendencia actúa (c)**



La temperatura final está muy lejos de tu objetivo. El PIB está muy lejos de tu objetivo.

Se han escogido recursos ilimitados y ningún impacto del cambio climático en la economía. Aunque económicamente es muy prometedor, el escenario es preocupante, porque la temperatura no logra estabilizarse. Bajo este escenario ideal, la economía crece a placer del modelador, y las emisiones de esta solo pueden verse afectadas por la mejora en la eficiencia energética o sustitución de energías fósiles por otras neutras.

La temperatura final es elevada, y el PIB es adecuado. Para conseguir un resultado con niveles económicos más realista y respetuoso con el medio ambiente podrías:

- **Bajar el valor de la medida m1 relativa a "Evolución planeada del Producto Interno Bruto (PIB) per capita“**
- **Bajar el valor de la medida m9 relativa a "Emisiones de gases de efecto invernadero no dependientes del consumo energético”.**

Se cambia la medida m1 de “b” a “e” (Fijar el PIB per capita en 7.000$/persona/año para 2050 y mantenerlo constante) y m9 de “c” a “b” (Continuación de la tendencia actúal).

Figura 4.2: Ejemplo de aplicación del Recomendador.

4.2. El modelo del Recomendador

El Modelo Básico (**Sección 3.6**) permite identificar un conjunto muy reducido de escenarios característicos. Sin embargo, de forma experimental, se ha visto que no es adecuado para su uso en el Recomendador.

Esto es debido a que el juego utiliza la situación de los indicadores en el último año para valorar el cumplimiento de objetivos, y dicho modelo, centrado en agrupaciones por la forma, comente un error no despreciable. Además, dentro de ciertos clústeres se han mezclado situaciones aceptables e indeseadas, relacionadas con valores de temperatura próximos a los 2ºC que no se estabilizan o caídas bruscas del PIB a partir de cierto año.

En consecuencia, se recurre a un segundo modelo, diseñado *ad-hoc* para su uso en el Recomendador, que considera 6 clústeres para la temperatura y 8 clústeres para el PIB. Esta agrupación se ha realizado sobre los datos completos (**Sección 3.6.5**), con 200 clústeres en la fase de *reducción de datos*. El número de clústeres para la fase de *extracción de patrones* han sido elegidos de forma empírica, tras analizar los diferentes conjuntos formados en varios niveles del histograma.

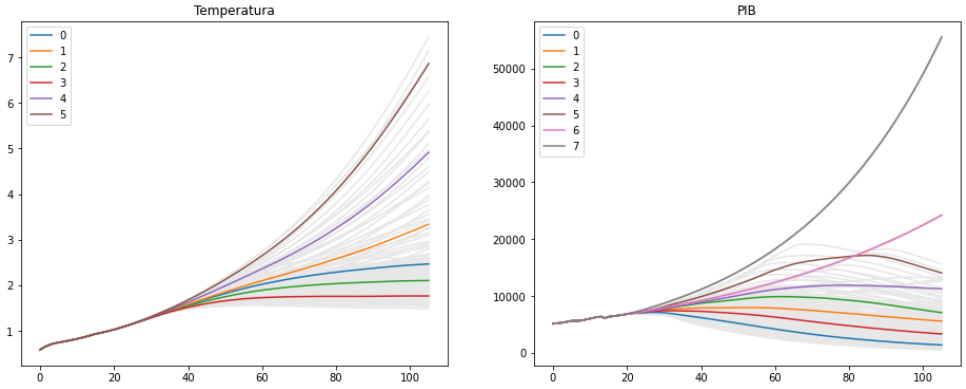


Figura 4.3: Patrones de temperatura y PIB para el modelo del Recomendador

En lo que respecta a la temperatura, los escenarios considerados como adecuados son los que se asocian a los patrones **0**, **1** y **2** representados en la **Figura 4.3**, en los cuales la temperatura se estabiliza en valores próximos a 2,5ºC. Cabe destacar que el objetivo de París, estabilizar la temperatura en 1,5ºC, solo se alcanza en el primero de los patrones.

Por otro lado, los patrones aceptables para el PIB son los denotados por **3**, **4**, **5**, **6** y **7**, donde este no decrece. En lo que respecta a los patrones **6** y **7**, son escenarios irreales resultado de combinar ciertas hipótesis ideales y poco probables (como recursos ilimitados o estricto control poblacional). Por tanto, pese a ser escenarios aceptables desde el punto de vista del PIB, el Recomendador los evitará por ser irreales.

En los patrones obtenidos se produce un cierto suavizado, consecuencia de aplicar la media de los centroides (calculado como la media de sus componentes), por lo que en el instante final existe una mayor variabilidad entre el escenario y el resultado simulado.

En la **Figura 4.4** se muestra la distribución de los resultados de simulación en los clústeres construidos en el modelo del Recomendador. Se puede observar que los escenarios considerados adecuados representan menos del 10 % del total (en azul en la **Figura 4.4**). Este número reducido implica que no todas las combinaciones de las hipótesis (**h1**, **h2** y **h3**) presentan alguna combinación de medidas políticas cuyo resultado de simulación sea un escenario adecuado. Por ello, en ciertos casos es necesario un cambio en las hipótesis

Por otro lado, el patrón de temperatura **2** podría considerarse como adecuado admitiendo cierto error en la "estabilización de la temperatura" (en gris en la **Figura 4.4**). En particular, la temperatura presenta valores inferiores a $2{,}75C$ pero no siempre logra la estabilización, alcanzando el último año con tendencia creciente. Sin embargo, el considerar dicho patrón de temperatura como adecuado aumenta notablemente el número de resultados de simulación válidos, lo que depende del error que se acepte.

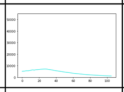
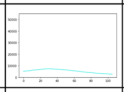
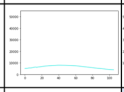
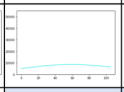
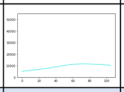
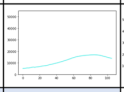
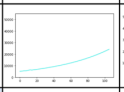
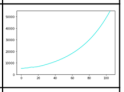
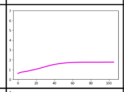
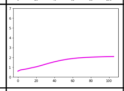
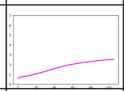
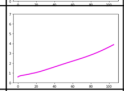
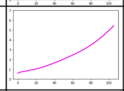
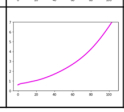
		0	1	2	3	4	5	6	7
									
0		0	2,6	1,1	4,4	0,7	0,1	0,2	0*
1		0	0,5	0	0,1	1,2	0*	1,2	0,4
2		10,3	5,78	3,60	15,90	1,5	0,1	0,1	0
3		31,5	3,7	0,7	11	0,5	0	0	0
4		0	0	0	0	0	0,8	0,6	1,2
5		0	0	0	0	0	0	0	0,5

Figura 4.4: Distribución de los resultados de simulación sobre para la agrupación utilizada en el Recomendador en porcentaje respecto del total. En color azul se diferencian los clústeres considerados adecuados, y en gris aquellos que podrían considerarse admitiendo cierto error.

4.3. Implementación e integración en Crossroads

El Recomendador se ha implementado en Python como un servicio REST externo al juego, accesible mediante una petición `HTTP-POST` al recurso `/recomendador` del servidor web. Además, se envía un archivo `json` con la siguiente estructura:

```
{
   "objetivo_temperatura": "d",
   "objetivo_pib":"a",
   "respuestas":["c","b","c","a","a","a","a","a","a","a","a","a","a"]
}
```

Donde `objetivo_temperatura` y `objetivo_pib` indican, respectivamente, el objetivo de temperatura y PIB que el jugador se marcó de un conjunto cerrado de opciones. Además, `respuestas` contiene una lista con las respuestas que el jugador ha dado a las 12 preguntas del cuestionario. El servicio retorna un *String* con toda la información generada por el Recomendador.

Para su implementación se ha utilizado el módulo `Flask` de Python, que permite montar el servidor web. Además de los módulos anteriormente descritos utilizados para cargar y estructurar la información (`pandas` y `numpy`). En la **Figura 4.5** se ilustra el diagrama de despliegue para Crossroads en una única maquina que reúne todos los servicios.

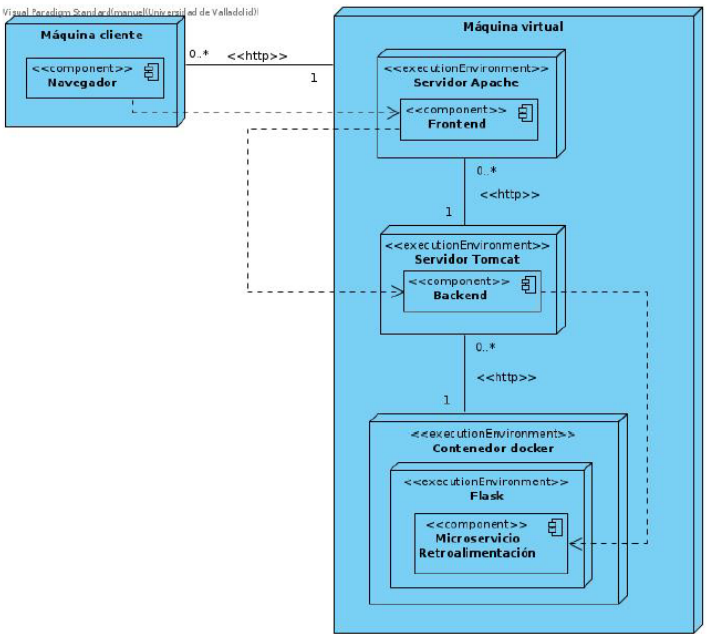


Figura 4.5: Diagrama de despliegue de Crossroads. Fuente [2]

Datos estáticos de entrada al Recomendador

Además de la información que recibe en la petición `HTTP`, el Recomendador requiere de un conjunto de datos e información estática para su funcionamiento. Esta información es invariable durante todo el proceso de recomendación, e idéntica para diferentes peticiones gestionadas. Dicha información se recoge en cuatro `csv` y tres `py` de configuración, y se divide entre información generada por el modelo e información externa:

- La información generada por el modelo se localiza en el directorio `./saves`, y se compone de:
  - `centers_pib`: patrones característicos construidos para el PIB.
  - `centers_tem`: patrones característicos construidos para la temperatura.
  - `descripción_escenarios.csv`: descripción realizada de los escenarios característicos, que presenta el primer *feedback* de estos.
  - `hipoteses_con_cluster.csv`: contiene todas las combinaciones de posibles respuestas dadas al cuestionario, y el clúster final (identificador del escenario característico) asociado.
- Los recursos necesarios para la configuración de la aplicación y la traducción de ciertas representaciones de datos se encuentran en la carpeta `Extras`:
  - `informacion_adicional.py`: permite recuperar el enunciado de cada pregunta del cuestionario a partir de su identificador.
  - `diccionario_objetivos.py`: permite obtener un valor numérico para los objetivos elegidos por el jugador, representados como una respuesta a una pregunta cerrada.
  - `config.py`: contiene los valores de configuración, donde se indican los patrones de temperatura y PIB considerados adecuados.

Arquitectura lógica

La implementación del servicio REST en Python la compone dos archivos diferentes:

- `app.py`: implementa las funcionalidades del servidor `Flask` y la parte del servicio REST que no está relacionada con la lógica interna (administración de las peticiones y captura de errores).
- `recomendador.py`: implementa la lógica interna del proceso de evaluación y recomendación. Su función principal es invocada en `app.py`.

Además, existe un tercer archivo, `dockerfile`, que se utiliza en la construcción de la imagen Docker. Determina la versión e imagen de Python a instalar en la imagen, instala los módulos de Python necesarios (especificados en `requirement.txt`) y configura como se ejecutará el servidor, incluyendo su visibilidad y acceso.

4.4. Evaluación del Recomendador

Con el fin de obtener una valoración del funcionamiento del Recomendador se ha desarrollado un cuestionario de 10 preguntas con un formato análogo al ejemplo representado en la **Figura 4.2**. En estas preguntas se plantean escenarios medioambientalmente inadecuados, donde el incremento de temperatura crece a valores superiores a 2,5ºC y no consigue estabilizarse.

De las diez cuestiones, cinco incluyen una recomendación generada de forma automática por el Recomendador, que mejora la evolución medioambiental manteniendo la economía en niveles aceptables y superiores a los actuales. Las otras cinco preguntas tienen recomendaciones inadecuadas, escritas arbitrariamente y que no mejoran la situación. Las preguntas se recogen en en **Apéndice D**

El participante valora según su experiencia (no se muestra el resultado final con los cambios recomendados) cada recomendación en una escala Likert de 1 a 5, donde uno implica que la recomendación es inadecuada y no logra mejorar la situación medioambiental en el último año, y 5 indica que la recomendación es correcta y supone una reducción en el incremento de temperatura. El Recomendador realizará buenas valoraciones si existe una clara diferencia entre las puntuaciones de las recomendaciones arbitrarias e incorrectas y las recomendaciones automáticas dadas por el Recomendador. El cuestionario se recoge en el **Apéndice D**.

Dado que este cuestionario solo puede ser resuelto por expertos en la materia, que sean capaces de evaluar subjetivamente como un cambio en las hipótesis y medidas políticas afecta al resultado simulado, la muestra de realizaciones que se espera es pequeña. Además, dadas las diferentes interpretaciones, se ha dado la opción de justificar brevemente las respuestas, para así descartar "malentendidos". En el caso que se presenta, se tiene una única respuesta realizado por un experto (Dr. Íñigo Capellan), donde las recomendaciones correctas han tenido una puntuación de 4,4 frente al 1,8 de las erróneas. Por tanto, se puede afirmar que existe una diferencia notable entre las recomendaciones automatizadas y las arbitrarias. Cabe destacar que ha errado en las preguntas 6 y 7, al dar una puntuación de 4 y 2 respectivamente, siendo la primera errónea y la segunda correcta.

4.5. Comparación de los modelos propuestos

La metodología de clasificación requiere determinar tres parámetros: el número de elementos en la reducción, el número de clústeres de temperatura y el número de clústeres de PIB. Manteniendo fijo el primer parámetro se han presentado dos clasificaciones diferentes: el modelo básico (**Sección 3.6**) y el modelo del Recomendador (**Sección 4.2**). Cabe destacar que el cuerpo de datos utilizado en el segundo modelo es mayor que el conjunto utilizado en el primer modelo, y se ha completado apoyándose en este.

En la **Tabla 4.6** se presentan diferentes métricas de error aplicadas a ambos modelos. En general, el modelo del Recomendador presenta una mejoría global en estas métricas[2], con una reducción de la desviación estándar (datos más agrupados en torno a una media de error menor).

Sin embargo, pese a que el segundo modelo presenta una mejoría respecto al primero, también supone un incremento en el número de clústeres construidos, de 17 a 26. Además, la diferencia entre los escenarios es menor. Esto se puede observar en la distancia entre los centroides de temperatura o PIB (**Tabla 4.7**), especialmente entre aquellos que son próximos entre sí.

Por otro lado, el coeficiente de la silueta para el modelo básico es de 0,23, frente a 0,26 obtenido para el modelo del Recomendador. Es decir, pese a haber aumentado el número de conjuntos, el valor del coeficiente s no ha aumentado en consecuencia, y por tanto las agrupaciones realizadas en el modelo del Recomendador son algo más "artificiales".

Por último, dado que para medir el cumplimiento de los objetivos propuestos por el jugador se consideran exclusivamente los valores de temperatura y PIB en el último año simulado, cobra especial importancia los errores cometidos al sustituir el último año de la MTS simulada por el de su escenario característico asociado (**Tabla 4.8**).

En general, el modelo del Recomendador presenta una reducción en el error medio cometido, así como en la desviación de los residuos. Es decir, de forma general el segundo modelo realiza mejores estimaciones del instante final, y las estimaciones son más próximas al valor real. Sin embargo, se puede apreciar cómo, puntualmente, esto no se cumple, como ocurre para la temperatura.

[2] aunque puntualmente puede empeorar, como en el caso del máximo error para la temperatura, pese a mejorar en el valor medio y tener una menor desviación estándar

		Temperatura (ºC)		**PIB *per Capita ($)***	
		Modelo Básico	**Modelo Recomendador**	**Modelo Básico**	**Modelo Recomendador**
RMSE	max	0,41	0,45	3344,73	3188,58
	min	$9{,}7 \cdot 10^{-4}$	$1{,}1 \cdot 10^{-3}$	18,32	19,11
	μ	0,082	0,065	1355,61	589,40
	σ	0,051	0,044	445,24	351
Error Máximo	max	1,08	1,32	7526,86	6891, 55
	min	$2{,}4 \cdot 10^{-3}$	$2{,}9 \cdot 10^{-3}$	24,43	26,72
	μ	0,17	0,13	1355,61	1072,14
	σ	0,12	0,044	971,63	654,76
Error Máx. Relativo	max	0,30	0,27	8,6	8,15
	min	$1{,}4 \cdot 10^{-3}$	$1{,}7 \cdot 10^{-3}$	$2{,}3 \cdot 10^{-3}$	$2{,}6 \cdot 10^{-3}$
	μ	0,081	0,064	0,35	0,30
	σ	0,047	0,038	0,34	0,34
Rango dinámico	max	7,715		55787,8	
	min	0,585		275,8	
	μ	1,58		6117,45	
	σ	0,53		3878,43	

Tabla 4.6: Errores cometido por los modelos

Modelo Básico: distancia entre centroides de temperatura

Clúster	0	1	2	3	4
0	–	2,49	7,53	13,5	20,22
1	2,49	–	5,11	11,09	17,82
2	7,53	5,11	–	5,98	12,71
3	13,5	11,09	5,98	–	6,73
4	20,22	17,82	12,71	6,73	–

Modelo Básico: Distancia entre centroides de PIB

Clúster	0	1	2	3	4	5
0	–	18421	45427	81782	110029	229435
1	18421	–	27327	63544	93273	213328
2	45427	27327	–	36560	67027	187428
3	81782	63544	36560	–	40376	157552
4	110029	93273	67027	40376	–	120439
5	229435	213328	187428	157552	120439	–

Modelo del Recomendador: Distancia entre centroides de temperatura

Clúster	0	1	2	3	4	5
0	–	1,95	3,79	6,74	12,81	20,35
1	1,95	–	1,84	4,85	10,94	18,49
2	3,79	1,84	–	3,07	9,14	16,69
3	6,74	4,85	3,07	–	6,09	13,64
4	12,81	10,94	9,14	6,09	–	7,55
5	20,35	18,49	16,69	13,64	7,55	–

Modelo del Recomendador: Distancia entre centroides de PIB

Clúster	0	1	2	3	4	5	6	7
0	–	16519	31501	47242	66023	97662	109825	229279
1	16519	–	15138	30804	49876	81489	94745	214777
2	31501	15138	–	15811	34763	66381	80280	200590
3	47242	30804	15811	–	19894	50967	67007	187393
4	66023	49876	34763	19894	–	31936	47577	167792
5	97662	81489	66381	50967	31936	–	28763	141586
6	109825	94745	80280	67007	47577	28763	–	120470
7	229279	214777	200590	187393	167792	141586	120470	–

Tabla 4.7: Distancia entre los centroies

		Temperatura (ºC)		PIB *per Capita ($)*	
		Modelo Básico	**Modelo Recomendador**	**Modelo Básico**	**Modelo Recomendador**
Diferencia absoluta último año	max	1,08	1,32	6840	6103,08
	min	$4{,}9 \cdot 10^{-6}$	$5{,}2 \cdot 10^{-7}$	$4{,}6 \cdot 10^{-3}$	$4 \cdot 10^{-3}$
	μ	0,15	0,11	883,79	663,63
	σ	0,13	0,11	987,97	588,18
Rango dinámico	max	7,71		55787,8	
	min	1,406		275,8	
	μ	2,1		5151,34	
	σ	0,64		8786,14	

Tabla 4.8: Errores cometido por los modelos al sustituir en el último instante de simulación

4.6. Conclusiones

En este capítulo se ha abordado la construcción de un sistema autónomo de evaluación y recomendación para los escenarios de simulación del videojuego *Crossroads*. El Recomendador propuesto genera un *feedback* para una respuesta a cada una de las preguntas del cuestionario del juego. Dicho *feedback* se divide en tres partes:

- Cumplimiento de los objetivos definidos por el propio jugador en términos del resultado de simulación obtenido.
- Cumplimiento de los requisitos medioambientales y económicos para el resultado de simulación obtenido.
- Cuando el resultado obtenido no cumpla con los requisitos del punto anterior (generalmente la temperatura se sitúa en valores peligrosos para la vida humana y/o no se logra su estabilización) se propone al jugador una serie de cambios, tratando de minimizar el impacto, sobre sus respuestas dadas.

Para el Recomendador se ha optado por utilizar un segundo modelo, cuyos parámetros se han determinado *ad-hoc*. Este nuevo modelo amplía los escenarios característicos con el objetivo de reducir los errores que puedan cometerse en el proceso de recomendación.

Para satisfacer la necesidades de tiempo de respuesta e independencia de Crossroads, el modelo y toda la información necesaria, se ha estructurado mediante un conjunto de ficheros estáticos. Con esto, se evita recurrir a la base de datos de Crossroads, desarrollada en `MongoDB`, cuyos tiempos de lectura excedían del tiempo de espera máximo permitido. Por contra, cuando se quiera hacer algún cambio sobre el juego (por ejemplo, nuevos escenarios u otro modelo) es necesario modificar su base de datos como reconstruir los ficheros que alimentan el Recomendador. Finalmente, con el objetivo de tener un sistema sencillo de desplegar que actúe como una "caja negra" para las personas ajenas, se ha decidido implementarlo bajo una imagen `Docker`.

Capítulo 5

Conclusiones y trabajo futuro

5.1. Conclusiones

Este trabajo se ha enfocado al problema de agrupación sobre conjuntos de datos de simulación, englobándose dentro del proyecto Locomotion H2020. Este proyecto tiene el objetivo de diseñar u sistema de modelado confiable y práctico para evaluar la viabilidad, efectividad costos y ramificaciones de diferentes medidas políticas sobre el medioambiente.

El cambio climático, y en especial la transición hacia una sociedad de bajas emisiones de carbono, es un gran desafío, donde se presentan una gran cantidad de alternativas y con diferentes políticas sociales y económicas que permitan alcanzar este objetivo. Este trabajo trata de contribuir a alcanzar dicho objetivo agrupando aquellos escenario futuros que son similares (tanto en mejora de la situación medioambiental como en un empeoramiento que pueda llevar a la humanidad a su extinción), y analizando las medidas necesarias para alcanzarlos. Finalmente, los resultados obtenidos se han utilizado para apoyar una aplicación de concienciación sobre el cambio climático, desarrollada como un videojuego educativo.

Se ha presentado un método original para el clúster de series temporales multivariables (MTS). El clústering sobre MTS es un problema abierto, donde existen múltiples alternativas dado el carácter de tiempo inherente a este tipo de datos. En el caso de estudio, dada la existencia de una homogeneidad temporal entre todas las series, dicho carácter ha pasado a un segundo plano.Sin embargo, la solución propuesta ha demostrado ser útil para agrupar y visualizar escenarios de cambio climático, extrayendo un conjunto reducido de representaciones a partir de un gran conjunto de muestras. Aunque se ha probado con este caso particular, donde se estudian evoluciones temporales de la temperatura (indicador medioambiental) y el PIB mundial (indicador de la economía) simuladas, el método se presenta de forma genérica con una formulación básica válida para otras dimensiones y problemas.

La agrupación del conjunto de escenarios simulados ha permitido el desarrollo de un Recomendado Recomendador, como sistema inteligente que conoce los escenarios posibles y que es capaz de hacer recomendaciones a los usuarios de la aplicación para que realicen nuevas propues-

tas más adecuadas a sus objetivos e intenciones. El método planteado se basa en establecer una asociación entre las múltiples entradas posibles que el usuario elije en el videojuego con uno de los escenarios característicos establecidos. Con ello, dada una selección de hipótesis y medidas políticas concretas, se determina el escenario característico correspondiente. Además, en caso de no cumplir los requerimientos medioambientales y socioeconómicos, se propone el mínimo número de cambios sobre las medidas dadas para alcanzar un escenario adecuado.

Es trabajo futuro presentar los resultados obtenidos en revistas y congresos de investigación especializados. Al haber definido un método genérico para el clústering de series temporales multivariables, se están redactando versiones de artículos para revistas sobre reconocimiento de patrones y *data mining*. Además, dado que se ha planteado un método automático de recomendación integrado en herramienta gamificada, se tiene como objetivo el envío de resultados a congresos especializados en el uso de inteligencia articial en videojuegos.

Para esto e hace necesario completar la implementación del algoritmo, permitiéndose su aplicación sobre series temporales multivariantes de más de dos componentes. Además, la medida de comparación utilizada relega a un segundo plano el carácter temporal de las series, por lo que sería de interés añadir nuevas métricas con carácter temporal (como la DTW), así como estudiar los resultados obtenidos con estas modificaciones. En cuanto al sistema de recomendación, se proponen mejoras en la generación de mensajes, actualmente realizado a través de plantillas sobre un conjunto acotado de opciones predefinidas, para obtener mensajes menos artificiales.

Finalmente, destacar que los resultados de este trabajo tienen un valor interdisciplinar. La notable reducción de escenarios futuros y su visualización en relación con las políticas energéticas y medioambientales pueden tener impacto en otras áreas de conocimiento relacionada con dominios diferentes a la informática. Entre otras posibles influencias, ha permitido obtener una descripción experta de los escenarios característicos, que se ha integrado en el sistema de recomendación.

Apéndice A

Contenido CD y Manuales

A.1. Contenido CD

En la documentación entregada junto a presente trabajo, se encuentran los diferentes archivos utilizados en el estudio del problema, generación de modelos e implementación del sistema autónomo de recomendaciones.

Dentro de la carpeta `modelos` se encuentran diferentes
textitNotebooks de Jupyter utilizados para el estudio del problema y del modelo:

`1-Determinar_Kmedias.ipynb`. Estudio mediante el método del codo para determinar el número de clústeres en la fase de reducción.

`2-Generación_del_modelo_e_información_asociada.ipynb`. Estudio del número de clústeres para la temperatura y para el PIB en función del resultado de la fase de reducción, generación del modelo y valoración inicial de los errores.

`3-Estudio_Errores_modelos.ipynb`. Estudio detallado de los errores cometidos por un modelo. Los archivos de errores necesarios se generan con la ejecución del segundo *Notebook*.

`4-Evaluación_consistencia_rand.ipynb`. Evaluación del coeficiente c definido en la **Sección 2.4.3**. Los archivos de cálculos necesarios se generan con la ejecución del segundo *Notebook*.

`5-Evaluación_asignacion.ipynb`. Comparación de la asignación según el representante de reducción frente al principio de mínima distancia. Los archivos necesarios se generan con la ejecución del segundo *Notebook*.

Además, se presentan los siguientes *script de Python*:

`clasesAgrupación.py`. Implementación de las clases que dan soporte al desarrollo del modelo (**Sección 3.5**).

`modelo2FasesMain.py`. Script principal para la generación de un modelo de agrupación y todos los recursos necesarios para su uso en el sistema autónomo de recomendación.

`config.py`. Configuración de los parámetros necesarios para la ejecución de `modelo2FasesMain.py`.

`requirements.txt`. Definición de los requisitos de paquetes de instalación en *Python*. Para ejecutar la generación del modelo basta con crear un entorno con los presentes requisitos. La versión de *Python* bajo la que se ha desarrollado es 3.8.x.

Además, en la carpeta de `datos` se presentan los datos originales y los datos completos (**Sección 3.6.5**). En la carpeta de `saves` se encuentran los recursos que se generan en el proceso de construcción del modelo básico (`TEM+PIB_200&5+6`) y en el modelo del Recomendador (`TEM+PIB_200&6+8`), así como los resultados de la fase de reducción en ambos casos.

Por otro lado, en la carpeta `Recomendador` se encuentra la implementación del Recomendador como contenedor `Docker`.

`app.py`. Implementación del servicio REST en *Flask* ara la recepción y gestión de las peticiones del Recomendador.

`recomendador.py`. Implementación de la lógica del Recomendador.

`Dockerfile`. Especificación de los requisitos del contenedor `Docker`, incluyendo el entorno *Python*.

`requirements.txt`. Definición de los requisitos de paquetes de instalación en *Python*.

Además, en la carpeta `Extras` se recoge la información de configuración del Recomendador (escenarios adecuados, traducción de los objetivos a valores numéricos, enunciado de las preguntas, etc.). En la carpeta `saves` se presentan todos los recursos necesarios relacionado con el modelo construido (representantes finales, asignación de respuestas al cuestionario con un clúster, descripción experta de los escenarios, etc.).

La información anterior se encuentra en
https://github.com/AdrianM97/TFM_ADRIAN_MANZANO (solo los scripts y notebooks sin datos de ingesta) y en
https://drive.google.com/file/d/1RmoyK6Qh8eILr6GQCjid8DUjaiCnl685/view?usp=sharing (con los datos de ingesta y resultados obtenidos con los modelos estudiados).

A.2. Instalación y ejecución del Recomendador

1. Instalar `Docker`

   ```
   sudo snap install docker
   ```

2. Crear el `Docker` del Recomendador (en el directorio raíz del Recomendador, donde se encuentra el documento `Dockerfile`)

   ```
   sudo docker build --tag recomendador_docker .
   ```

3. Ejecutar la imagen `Docker`.

   ```
   sudo docker run -p XXXX:5000 recomendador_docker
   ```

   donde XXXX indica el puerto de la máquina local al que se relizaran las peticiones. La aplicación corre en el puerto 5000 del servicio docker, donde se redirigen las solicitudes con `-p XXXX:5000`. Por último, podemos usar la opción `-d` para que se ejecute en segundo plano.

Apéndice B

Información Mutua por clústeres

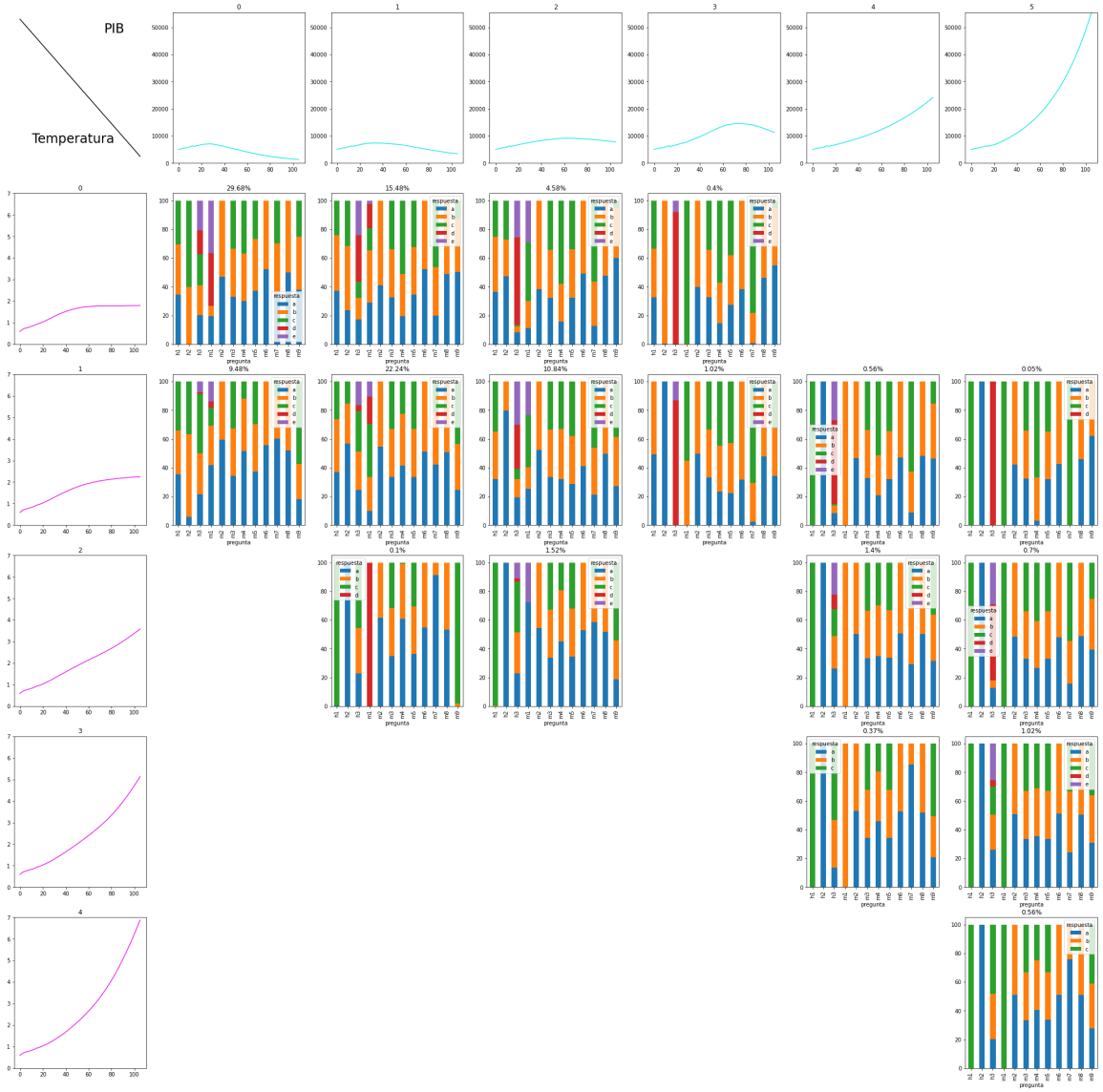


Figura B.1: Distribución de las hipótesis y decisiones políticas de los elementos que pertenecen a cada clúster, modelo de mínimos

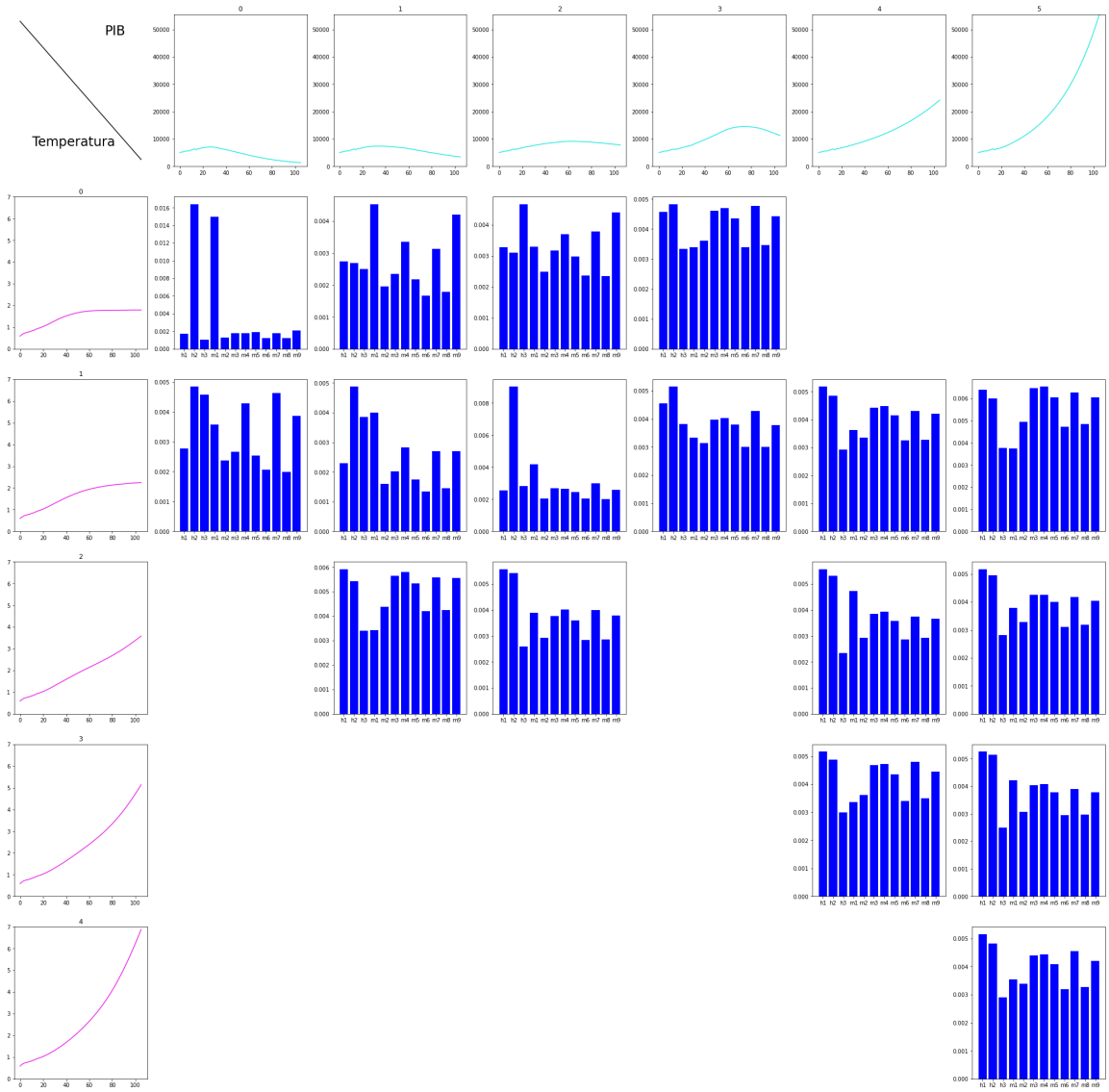
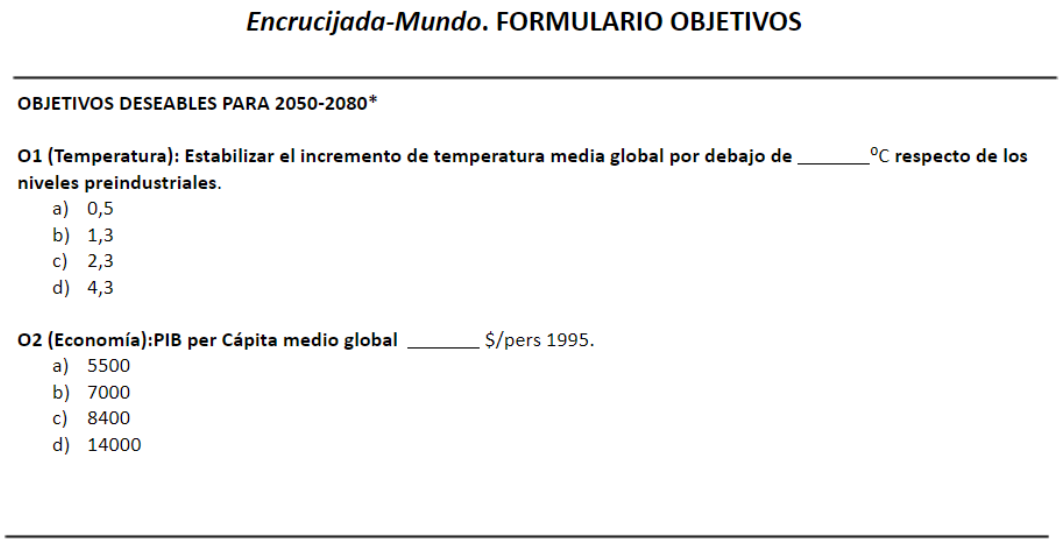


Figura B.2: Representación de la estimación dada por la Información Mutua utilizando el modelo de mínimos

Apéndice C

Preguntas del cuestionario

Este apéndice recoge las preguntas del videojuego Crossroads que permiten definir los objetivos de incrementos de temperatura y PIB, así como determinar las hipótesis y medidas política que a su vez establece el escenario futuro.



***Encrucijada-Mundo*. FORMULARIO OBJETIVOS**

OBJETIVOS DESEABLES PARA 2050-2080*

O1 (Temperatura): Estabilizar el incremento de temperatura media global por debajo de _____ºC respecto de los niveles preindustriales.

a) 0,5
b) 1,3
c) 2,3
d) 4,3

O2 (Economía):PIB per Cápita medio global _____ $/pers 1995.

a) 5500
b) 7000
c) 8400
d) 14000

Figura C.1: Preguntas para determinar los objetivos

Encrucijada-Mundo. HIPÓTESIS

Para cada opción H1-2 (hipótesis) marcar la opción seleccionada (a, b, c…).
Razonar las hipótesis seleccionadas: ¡éstas no se pueden modificar a lo largo del juego!

HIPÓTESIS

H1: Disponibilidad de recursos energéticos no-renovables (petróleo, carbón, gas natural y uranio) hasta el año 2050:

a) Media.
b) Alta.
c) Ilimitada.

H2: Impactos producidos por el cambio climático en la economía global:

a) Ninguno.
b) Medios.
c) Altos.

H3: Evolución de la población

a) Menor crecimiento que el previsto por las tendencias históricas.
b) Continuación de las tendencias históricas.
c) Mayor crecimiento que lo previsto en las tendencias históricas.
d) Limitar la población mundial en 5.000 millones de personas a partir de 2050.
e) Limitar la población mundial en 7.000 millones de personas a partir de 2050.

Encrucijada-Mundo. Medidas Políticas

Para cada meta (M_1-M_{10}) marcar la opción deseada (a, b, c…).
Es recomendable leer todo el formulario antes de empezar a seleccionar respuestas.

M_1: Evolución planeada del Producto Interno Bruto (PIB) *per cápita*:

a) Menor crecimiento que el histórico.
b) Continuación de las tendencias históricas.
c) Mayor crecimiento que el histórico.
d) Fijar el PIB *per cápita* en 5.000$/persona/año para 2050 y mantenerlo constante.
e) Fijar el PIB *per cápita* en 7.000$/persona/año para 2050 y mantenerlo constante.

M_2: Programa global de reforestación para capturar carbono desde el año 2020

a) No
b) Sí

M_3: Energía nuclear:

a) Disminución de la capacidad instalada.
b) Mantener constante la capacidad actual.
c) Aumento de la capacidad instalada.

Figura C.2: Preguntas para determinar las hipótesis y medidas políticas I

M_4: Producción planeada de energía renovable (electricidad y calor):
a) Menor crecimiento que las tendencias históricas.
b) Continuación de las tendencias históricas.
c) Mayor crecimiento que las tendencias históricas.

M_5: Producción planeada de biocombustibles líquidos (bioetanol, biodiesel, etc.):
a) Producción constante en los niveles actuales.
b) Continuar el crecimiento histórico.
c) Mayor crecimiento que la tendencia histórica.

M_6: Transporte terrestre (coches, motos, autobuses, trenes, camiones, etc.):
a) Continuación de la dependencia del petróleo.
b) Transición a vehículos a gas y eléctricos.

M_7: Tendencia en la evolución de la eficiencia energética:
a) Menor crecimiento que las tendencias históricas.
b) Continuar la tendencia de mejora histórica.
c) Mayor crecimiento que la tendencia histórica.

M_8: Tasas de reciclado de minerales necesarios para la transición energética hacia un mayor uso de renovables:
a) Mantener constantes los porcentajes actuales.
b) Mejorar los porcentajes actuales.

M_9: Emisiones de gases de efecto invernadero no dependientes del consumo energético:
a) Reducción respecto a la tendencia actual
b) Continuación de la tendencia actual
c) Aumento respecto a la tendencia actual

Figura C.3: Preguntas para determinar las hipótesis y medidas políticas II

Apéndice D

Cuestionario de Evaluación del Recomendador

H1: Disponibilidad de recursos energéticos no-renovables (petróleo, carbón, gas natural y uranio) hasta el año 2050. **Ilimitada (c)**

H2: Impactos producidos por el cambio climático en la economía global. **Ninguno (a)**

H3: Evolución de la población. **Limitar la población mundial en 7.000 millones de personas a partir de 2050 (e)**

M1: Evolución planeada del Producto Interno Bruto (PIB) per capita. **Mayor crecimiento que el histórico (c)**

M2: Programa global de reforestación para capturar carbono desde el año 2020. **Si (b)**

M3: Energía nuclear**. Mantener constante la capacidad actual (b)**

M4: Producción planeada de energía renovable (electricidad y calor). **Mayor crecimiento que las tendencias históricas (c)**

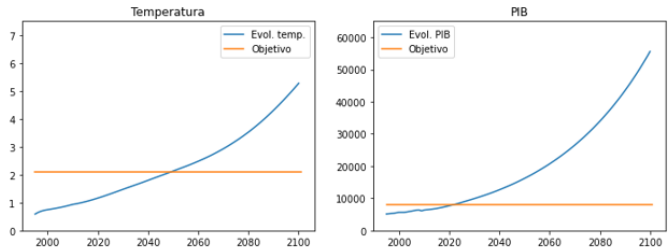
M5: Producción planeada de biocombustibles líquidos (bioetanol, biodiesel, etc.). **Producción constante en los niveles actuales (a)**

M6: Transporte terrestre (coches, motos, autobuses, trenes, camiones, etc.). **Transición a vehículos gas y eléctricos (a)**

M7: Tendencia en la evolución de la eficiencia energética. **Menor crecimiento que las tendencias históricas (a)**

M8: Tasas de reciclado de minerales necesarios para la transición energética hacia un mayor uso de renovables. **Mantener constantes los porcentajes actuales (a)**

M9: Emisiones de gases de efecto invernadero no dependientes del consumo energético. **Continuación de la tendencia actual (b)**



Para conseguir un resultado con niveles económicos más realista y respetuoso con el medio ambiente podrías bajar el valor de la medida m1 relativa a "Evolución planeada del Producto Interno Bruto (PIB) per capita".

Se cambia la medida m1 de "c" a "e" (Fijar el PIB per capita en 7.000$/persona/año para 2050 y mantenerlo constante)

Figura D.1: Cuestionario de Evaluación del Recomendador: Pregunta 1

H1: Disponibilidad de recursos energéticos no-renovables (petróleo, carbón, gas natural y uranio) hasta el año 2050. **Alta (b)**

H2: Impactos producidos por el cambio climático en la economía global. **Media (b)**

H3: Evolución de la población. **Continuar las tendencias históricas (b)**

M1: Evolución planeada del Producto Interno Bruto (PIB) per capita. **Continuación de las tendencias históricas (a)**

M2: Programa global de reforestación para capturar carbono desde el año 2020. **No (a)**

M3: Energía nuclear**. Aumentar la capacidad instalada (c)**

M4: Producción planeada de energía renovable (electricidad y calor). **Menor crecimiento que las tendencias históricas (a)**

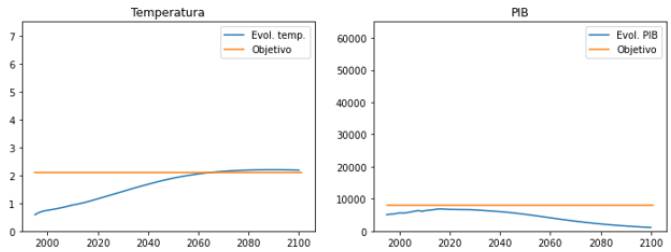
M5: Producción planeada de biocombustibles líquidos (bioetanol, biodiesel, etc.). **Producción constante en los niveles actuales (a)**

M6: Transporte terrestre (coches, motos, autobuses, trenes, camiones, etc.). **Transición a vehículos gas y eléctricos (b)**

M7: Tendencia en la evolución de la eficiencia energética. **Menor crecimiento que las tendencias históricas (a)**

M8: Tasas de reciclado de minerales necesarios para la transición energética hacia un mayor uso de renovables. **Mantener constantes los porcentajes actuales (a)**

M9: Emisiones de gases de efecto invernadero no dependientes del consumo energético. **Continuación de la tendencia actual (b)**



Para conseguir un resultado económicamente más equilibrado podrías:

- Subir el valor de la hipótesis h1 relativa a “Disponibilidad de recursos energéticos no-renovables”
- Bajar el valor de la hipótesis h2 relativa a “Impactos producidos por el cambio climático en la economía global”

Se cambia la hipótesis h1 de “b” a “c” (recursos ilimitados) y h2 de “b” a “a” (ningún efecto del cambio climático en la economía)

Figura D.2: Cuestionario de Evaluación del Recomendador: Pregunta 2

H1: Disponibilidad de recursos energéticos no-renovables (petróleo, carbón, gas natural y uranio) hasta el año 2050. **Ilimitada (c)**

H2: Impactos producidos por el cambio climático en la economía global. **Ninguno (a)**

H3: Evolución de la población**. Menor crecimiento que el previsto por las tendencias históricas (a)**

M1: Evolución planeada del Producto Interno Bruto (PIB) per capita. **Menor crecimiento que el histórico (a)**

M2: Programa global de reforestación para capturar carbono desde el año 2020. **Si (b)**

M3: Energía nuclear**. Disminución de la capacidad instalada (a)**

M4: Producción planeada de energía renovable (electricidad y calor). **Continuación del crecimiento histórico (b)**

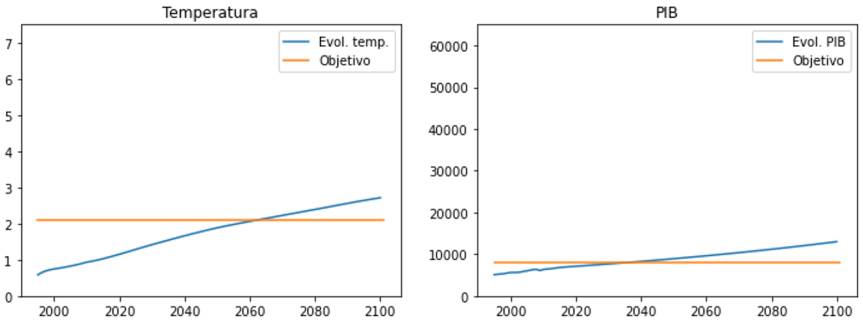
M5: Producción planeada de biocombustibles líquidos (bioetanol, biodiesel, etc.). **Continuar crecimiento histórico (b)**

M6: Transporte terrestre (coches, motos, autobuses, trenes, camiones, etc.). **Continuación dependencia del petróleo (a)**

M7: Tendencia en la evolución de la eficiencia energética. **Continuar la tendencia de mejora histórica (b)***

M8: Tasas de reciclado de minerales necesarios para la transición energética hacia un mayor uso de renovables. **Mantener constantes los porcentajes actuales (a)**

M9: Emisiones de gases de efecto invernadero no dependientes del consumo energético. **Reducción de la tendencia actual (a)***



Para conseguir un resultado más respetuoso con el medio ambiente podrías:

- Subir el valor de la medida m3 relativa a “Energía nuclear“.
- Subir el valor de la medida m8 relativa a “Tasas de reciclado de minerales necesarios para la transición energética”.

Se cambia la medida m3 de “a” a “b” (Mantener constante la capacidad actual) y m8 de “a” a “b” (Mejorar los porcentajes actuales)

Figura D.3: Cuestionario de Evaluación del Recomendador: Pregunta 3

H1: Disponibilidad de recursos energéticos no-renovables (petróleo, carbón, gas natural y uranio) hasta el año 2050. **Ilimitada (c)**

H2: Impactos producidos por el cambio climático en la economía global. **Ninguno (a)**

H3: Evolución de la población**. Limitar la población mundial en 7.000 millones de personas a partir de 2050 (e)**

M1: Evolución planeada del Producto Interno Bruto (PIB) per capita. **Continuación de las tendencias históricas (b)**

M2: Programa global de reforestación para capturar carbono desde el año 2020. **No (a)**

M3: Energía nuclear**. Aumento de la capacidad instalada (c)**

M4: Producción planeada de energía renovable (electricidad y calor). **Continuación del crecimiento histórico (b)**

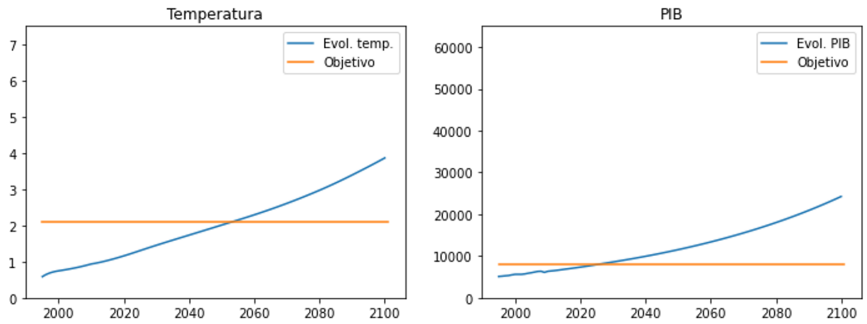
M5: Producción planeada de biocombustibles líquidos (bioetanol, biodiesel, etc.). **Continuar crecimiento histórico (b)**

M6: Transporte terrestre (coches, motos, autobuses, trenes, camiones, etc.). **Continuación dependencia del petróleo (a)**

M7: Tendencia en la evolución de la eficiencia energética. **Menor crecimiento que las tendencias históricas (a)**

M8: Tasas de reciclado de minerales necesarios para la transición energética hacia un mayor uso de renovables. **Mejorar los porcentajes actuales (b)**

M9: Emisiones de gases de efecto invernadero no dependientes del consumo energético. **Continuación de la tendencia actual (b)**



Para conseguir un resultado con niveles económicos más realistas y respetuoso con el medio ambiente podrías:

- Bajar el valor de la medida m1 relativa a "Evolución planeada del Producto Interno Bruto (PIB) per capita“
- Subir el valor de la medida m7 relativa a "Tendencia en la evolución de la eficiencia energética“

Se cambia la medida m1 de “b” a “e” (Fijar el PIB per capita en 7.000$/persona/año para 2050 y mantenerlo constante) y m7 de “a” a “b” (Continuar la tendencia de mejora histórica).

Figura D.4: Cuestionario de Evaluación del Recomendador: Pregunta 4

H1: Disponibilidad de recursos energéticos no-renovables (petróleo, carbón, gas natural y uranio) hasta el año 2050. **Ilimitada (c)**

H2: Impactos producidos por el cambio climático en la economía global. **Ninguno (a)**

H3: Evolución de la población**. Menor crecimiento que el previsto por las tendencias históricas (a)**

M1: Evolución planeada del Producto Interno Bruto (PIB) per capita. **Continuación de las tendencias históricas (b)**

M2: Programa global de reforestación para capturar carbono desde el año 2020. **No (a)**

M3: Energía nuclear**. Mantener la capacidad instalada (b)**

M4: Producción planeada de energía renovable (electricidad y calor). **Mayor crecimiento que las tendencias históricas (c)**

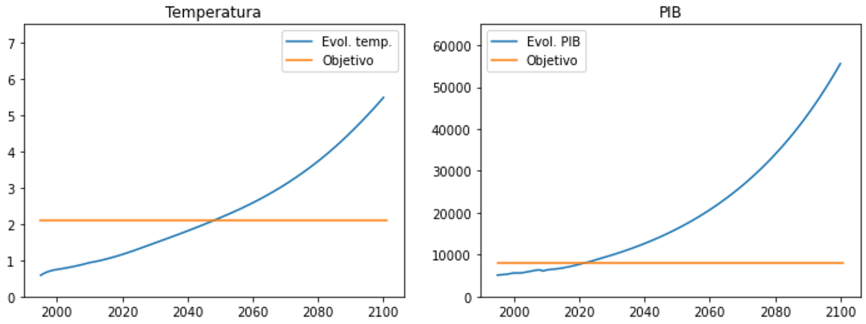
M5: Producción planeada de biocombustibles líquidos (bioetanol, biodiesel, etc.). **Continuar crecimiento histórico (b)**

M6: Transporte terrestre (coches, motos, autobuses, trenes, camiones, etc.). **Transición a vehículos de gas y eléctricos (b)**

M7: Tendencia en la evolución de la eficiencia energética. **Mayor crecimiento que las tendencias históricas (c)**

M8: Tasas de reciclado de minerales necesarios para la transición energética hacia un mayor uso de renovables. **Mejorar los porcentajes actuales (b)**

M9: Emisiones de gases de efecto invernadero no dependientes del consumo energético. **Aumento respecto de la tendencia actúa (c)**



Para conseguir un resultado con niveles económicos más realista y respetuoso con el medio ambiente podrías:

- Bajar el valor de la medida m1 relativa a "Evolución planeada del Producto Interno Bruto (PIB) per capita“
- Bajar el valor de la medida m9 relativa a "Emisiones de gases de efecto invernadero no dependientes del consumo energético”.

Se cambia la medida m1 de “b” a “e” (Fijar el PIB per capita en 7.000$/persona/año para 2050 y mantenerlo constante) y m9 de “c” a “b” (Continuación de la tendencia actúal).

Figura D.5: Cuestionario de Evaluación del Recomendador: Pregunta 5

H1: Disponibilidad de recursos energéticos no-renovables (petróleo, carbón, gas natural y uranio) hasta el año 2050. **Ilimitada (c)**

H2: Impactos producidos por el cambio climático en la economía global. **Ninguno (a)**

H3: Evolución de la población**. Limitar la población mundial en 7.000 millones de personas a partir de 2050 (e)**

M1: Evolución planeada del Producto Interno Bruto (PIB) per capita. **Mayor crecimiento que el histórico (c)**

M2: Programa global de reforestación para capturar carbono desde el año 2020. **No (a)**

M3: Energía nuclear**. Aumentar la capacidad instalada (c)**

M4: Producción planeada de energía renovable (electricidad y calor). **Continuar crecimiento histórico (b)**

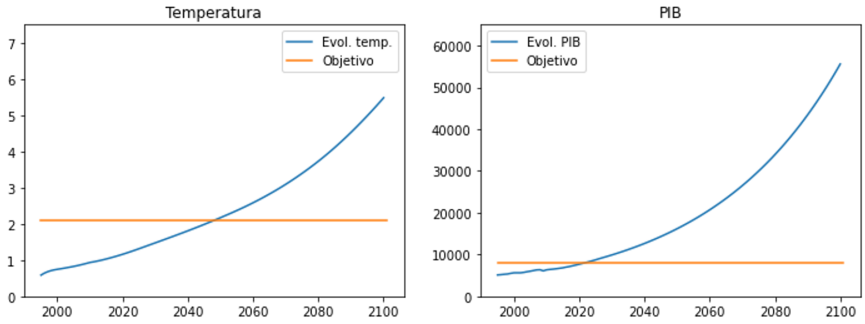
M5: Producción planeada de biocombustibles líquidos (bioetanol, biodiesel, etc.). **Continuar crecimiento histórico (b)**

M6: Transporte terrestre (coches, motos, autobuses, trenes, camiones, etc.). **Continuación de la dependencia del petróleo (a)**

M7: Tendencia en la evolución de la eficiencia energética. **Mayor crecimiento que las tendencias históricas (c)**

M8: Tasas de reciclado de minerales necesarios para la transición energética hacia un mayor uso de renovables. **Mejorar los porcentajes actuales (b)**

M9: Emisiones de gases de efecto invernadero no dependientes del consumo energético. **Continuación de la tendencia actual (b)**



Para conseguir un resultado con niveles económicos más realista y respetuoso con el medio ambiente podrías:

- Subir el valor de la medida m2 relativa a “Programa global de reforestación para capturar carbono desde el año 2020“
- Bajar el valor de la medida m9 relativa a “Emisiones de gases de efecto invernadero no dependientes del consumo energético”.

Se cambia la medida m2 de “a” a “b” (Si) y m9 de “b” a “a” (Reducir la tendencia actual).

Figura D.6: Cuestionario de Evaluación del Recomendador: Pregunta 6

H1: Disponibilidad de recursos energéticos no-renovables (petróleo, carbón, gas natural y uranio) hasta el año 2050. **Ilimitada (c)**

H2: Impactos producidos por el cambio climático en la economía global. **Ninguno (a)**

H3: Evolución de la población**. Limitar la población mundial en 5.000 millones de personas a partir de 2050 (d)**

M1: Evolución planeada del Producto Interno Bruto (PIB) per capita. **Mayor crecimiento que el histórico (c)**

M2: Programa global de reforestación para capturar carbono desde el año 2020. **No (a)**

M3: Energía nuclear**. Aumentar la capacidad instalada (c)**

M4: Producción planeada de energía renovable (electricidad y calor). **Continuar crecimiento histórico (b)**

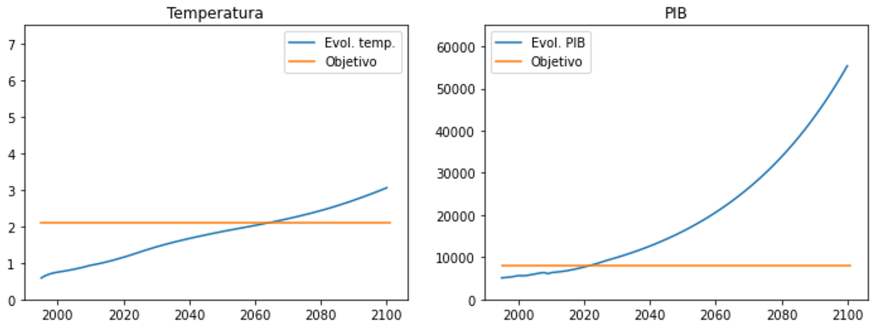
M5: Producción planeada de biocombustibles líquidos (bioetanol, biodiesel, etc.). **Producción constante en los niveles actuales (a)**

M6: Transporte terrestre (coches, motos, autobuses, trenes, camiones, etc.). **Continuación de la dependencia del petróleo (a)**

M7: Tendencia en la evolución de la eficiencia energética. **Mayor crecimiento que las tendencias históricas (c)**

M8: Tasas de reciclado de minerales necesarios para la transición energética hacia un mayor uso de renovables. **Mejorar los porcentajes actuales (b)**

M9: Emisiones de gases de efecto invernadero no dependientes del consumo energético. **Reducción respecto de la tendencia actual (a)**



Para conseguir un resultado con niveles económicos más realista y respetuoso con el medio ambiente podrías bajar el valor de la medida m1 relativa a "Evolución planeada del Producto Interno Bruto (PIB) per capita"

Se cambia la medida m1 de “c” a “a” (menor crecimiento que el histórico).

Figura D.7: Cuestionario de Evaluación del Recomendador: Pregunta 7

H1: Disponibilidad de recursos energéticos no-renovables (petróleo, carbón, gas natural y uranio) hasta el año 2050. **Media (a)**

H2: Impactos producidos por el cambio climático en la economía global. **Ninguno (a)**

H3: Evolución de la población**. Limitar la población mundial en 5.000 millones de personas a partir de 2050 (d)**

M1: Evolución planeada del Producto Interno Bruto (PIB) per capita. **Continuación de la tendencia histórica (b)**

M2: Programa global de reforestación para capturar carbono desde el año 2020. **Si (b)**

M3: Energía nuclear**. Aumentar la capacidad instalada (b)**

M4: Producción planeada de energía renovable (electricidad y calor). **Continuar crecimiento histórico (b)**

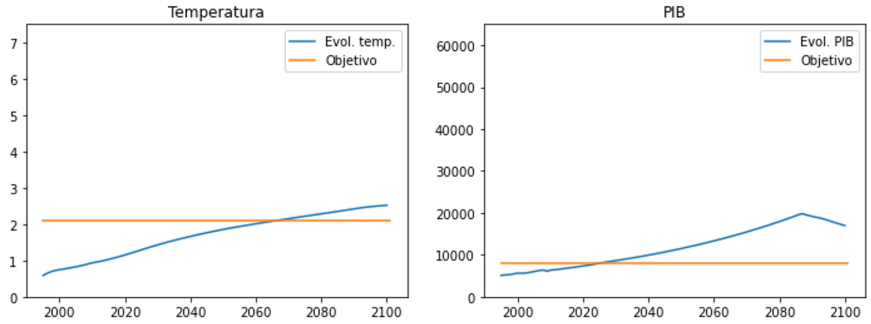
M5: Producción planeada de biocombustibles líquidos (bioetanol, biodiesel, etc.). **Producción constante en los niveles actuales (c)**

M6: Transporte terrestre (coches, motos, autobuses, trenes, camiones, etc.). **Continuación de la dependencia del petróleo (b)**

M7: Tendencia en la evolución de la eficiencia energética. **Mayor crecimiento que las tendencias históricas (c)**

M8: Tasas de reciclado de minerales necesarios para la transición energética hacia un mayor uso de renovables. **Mejorar los porcentajes actuales (b)**

M9: Emisiones de gases de efecto invernadero no dependientes del consumo energético. **Reducción respecto de la tendencia actual (c)**



Para conseguir un resultado más respetuoso con el medio ambiente podrías:

- Subir el valor de la medida m1 relativa a “Evolución planeada del Producto Interno Bruto (PIB) per capita”
- Subir el valor de la hipótesis h1 relativa a “Disponibilidad de recursos energéticos no-renovables”

Se cambia la medida m1 de “b” a “c” (Mayor crecimiento que lo previsto en las tendencias históricas) y la hipótesis h1 de “a” a “c” (Ilimitados).

Figura D.8: Cuestionario de Evaluación del Recomendador: Pregunta 8

H1: Disponibilidad de recursos energéticos no-renovables (petróleo, carbón, gas natural y uranio) hasta el año 2050. **Ilimitados (c)**

H2: Impactos producidos por el cambio climático en la economía global. **Ninguno (a)**

H3: Evolución de la población**. Mayor crecimiento que lo previsto en las tendencias históricas (c)**

M1: Evolución planeada del Producto Interno Bruto (PIB) per capita. **Menor crecimiento que el previsto por las tendencias históricas (a)**

M2: Programa global de reforestación para capturar carbono desde el año 2020. **Si (b)**

M3: Energía nuclear**. Mantener constante la capacidad instalada (b)**

M4: Producción planeada de energía renovable (electricidad y calor). **Menor crecimiento que las tendencias históricas (a)**

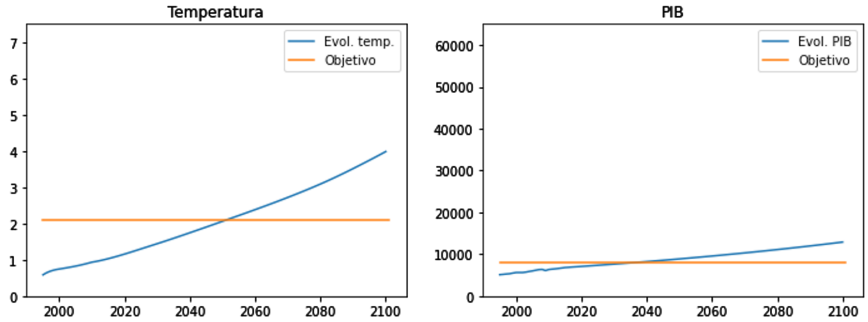
M5: Producción planeada de biocombustibles líquidos (bioetanol, biodiesel, etc.). **Mayor crecimiento que la tendencia histórica (c)**

M6: Transporte terrestre (coches, motos, autobuses, trenes, camiones, etc.). **Transición a vehículos a gas y eléctricos.(b)**

M7: Tendencia en la evolución de la eficiencia energética. **Menor crecimiento que las tendencias históricas (a)**

M8: Tasas de reciclado de minerales necesarios para la transición energética hacia un mayor uso de renovables. **Mantener constantes porcentajes actuales (a)**

M9: Emisiones de gases de efecto invernadero no dependientes del consumo energético. **Continuación de la tendencia actual (b)**



Para conseguir un resultado más respetuoso con el medio ambiente podrías:

- Bajar el valor de la medida m1 relativa a "Evolución planeada del Producto Interno Bruto (PIB) per capita".
- Subir el valor de la medida m7 relativa a "Tendencia en la evolución de la eficiencia energética".
- Bajar el valor de la medida m9 relativa a "Emisiones de gases de efecto invernadero no dependientes del consumo energético"

Se cambia la medida m1 de "a" a "e" (Fijar el PIB per capita en 7.000$/persona/año para 2050 y mantenerlo constante), la medida m7 de "a" a "c" (Mayor crecimiento que la tendencia histórica) y la medida m9 de "b" a "a" (Reducción respecto a la tendencia actual).

Figura D.9: Cuestionario de Evaluación del Recomendador: Pregunta 9

H1: Disponibilidad de recursos energéticos no-renovables (petróleo, carbón, gas natural y uranio) hasta el año 2050. **Media (c)**

H2: Impactos producidos por el cambio climático en la economía global. **Ninguno (a)**

H3: Evolución de la población**. Mayor crecimiento que lo previsto en las tendencias históricas (c)**

M1: Evolución planeada del Producto Interno Bruto (PIB) per capita. **Continuación de las tendencias históricas (b)**

M2: Programa global de reforestación para capturar carbono desde el año 2020. **No (a)**

M3: Energía nuclear**. Disminución de la capacidad instalada (a)**

M4: Producción planeada de energía renovable (electricidad y calor). **Menor crecimiento que las tendencias históricas (a)**

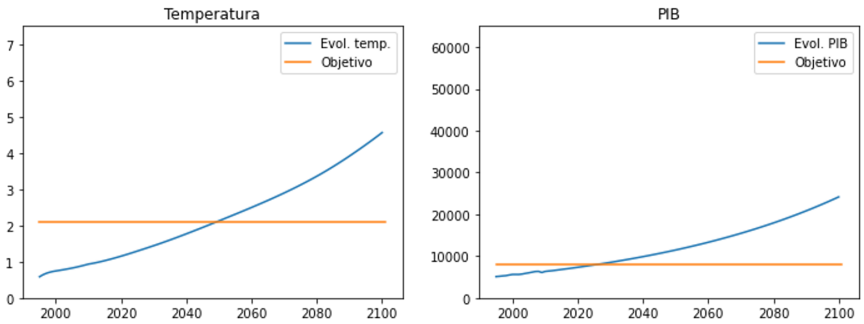
M5: Producción planeada de biocombustibles líquidos (bioetanol, biodiesel, etc.). **Continuar crecimiento histórico (b)**

M6: Transporte terrestre (coches, motos, autobuses, trenes, camiones, etc.). **Continuación de la dependencia del petróleo (a)**

M7: Tendencia en la evolución de la eficiencia energética. **Menor crecimiento que las tendencias históricas (b)**

M8: Tasas de reciclado de minerales necesarios para la transición energética hacia un mayor uso de renovables. **Mejorar porcentajes actuales (b)**

M9: Emisiones de gases de efecto invernadero no dependientes del consumo energético. **Continuación de la tendencia actual (b)**



Para conseguir un resultado económicamente más realista y respetuoso con el medio ambiente podrías:

- Subir el valor de la medida m3 relativa a “Energía nuclear”
- Bajar el valor de la medida m9 relativa a "Emisiones de gases de efecto invernadero no dependientes del consumo energético"

Se cambia la medida m3 de “a” a “c” (Aumentar la capacidad instalada), la medida m9 de “b” a “a” (Reducción respecto a la tendencia actual).

Figura D.10: Cuestionario de Evaluación del Recomendador: Pregunta 10

Preguntas Erróneas: 2, 3, 6, 8, 10.

Preguntas Correctas: 1, 4, 5, 7, 9.

Bibliografía

[1] Segmentación de imágenes médicas mediante agrupación de k-medias y algoritmo mejorado de cuencas hidrográficas. In *Simposio IEEE Southwest 2006 sobre análisis e interpretación de imágenes*.

[2] Alda Peñafiel, M. *Desarrollo del front-end y mejoras en el back-end de un juego didáctico multijugador de competición y consenso sobre el cambio climático*. 2021.

[3] Bennasar, M., Hicks, Y., and Setchi, R. Feature selection using joint mutual information maximisation. *Expert Systems with Applications 42*, 22 (2015), 8520–8532.

[4] Bholowalia, P., and Kumar, A. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications 105*, 9 (2014).

[5] Brockwell, P. J., Brockwell, P. J., Davis, R. A., and Davis, R. A. *Introduction to time series and forecasting*. Springer, 2016.

[6] Carpenter, G. A., and Grossberg, S. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer vision, graphics, and image processing 37*, 1 (1987), 54–115.

[7] Cheeseman, P. C., Stutz, J. C., et al. Bayesian classification (autoclass): theory and results. *Advances in knowledge discovery and data mining 180* (1996), 153–180.

[8] Corduas, M., and Piccolo, D. Time series clustering and classification by the autoregressive metric. *Computational statistics & data analysis 52*, 4 (2008), 1860–1872.

[9] Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Classification*, 2 ed. Wiley, New York, 2001.

[10] Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (1996), vol. 96, pp. 226–231.

[11] Estevez, P. A., Tesmer, M., Perez, C. A., and Zurada, J. M. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks 20*, 2 (2009), 189–201.

[12] Grimmett, G., and Stirzaker, D. *Probability and random processes*. Oxford university press, 2020.

[13] Han, J., Kamber, M., and Pei, J. *Data mining concepts and techniques third edition*, vol. 5. 2011.

[14] Kalpakis, K., Gada, D., and Puttagunta, V. Distance measures for effective clustering of arima time-series. In *Proceedings 2001 IEEE international conference on data mining* (2001), IEEE, pp. 273–280.

[15] Kohonen, T. *Self-organizing maps*, vol. 30. Springer Science & Business Media, 2012.

[16] Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. *Physical review E 69*, 6 (2004), 066138.

[17] Lloyd, S. Least squares quantization in pcm. *IEEE transactions on information theory 28*, 2 (1982), 129–137.

[18] Mahajan, M., Nimbhorkar, P., and Varadarajan, K. The planar k-means problem is np-hard. In *WALCOM: Algorithms and Computation* (Berlin, Heidelberg, 2009), S. Das and R. Uehara, Eds., Springer Berlin Heidelberg, pp. 274–285.

[19] Minnen, D., Starner, T., Essa, I., and Isbell, C. Discovering characteristic actions from on-body sensor data. In *2006 10th IEEE international symposium on wearable computers* (2006), IEEE, pp. 11–18.

[20] Petitjean, F., Ketterlin, A., and Gançarski, P. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition 44*, 3 (2011), 678–693.

[21] Rand, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association 66*, 336 (1971), 846–850.

[22] ROTHSCHILD, L. P. A bit of information theory. *San Diego, CA: UCSD Dept. of Mathematics* (2015).

[23] Sakoe, H., and Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing 26*, 1 (1978), 43–49.

[24] Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal 27*, 3 (1948), 379–423.

[25] Wang, W., Yang, J., Muntz, R., et al. Sting: A statistical information grid approach to spatial data mining. In *VLDB* (1997), vol. 97, pp. 186–195.

[26] Warren Liao, T. *Clustering of time series data—a survey*, vol. 38. 2005.